

Asymptotic Rank Speedup Theorems, Revisited

Josh Alman 

Columbia University, New York, NY, USA

Baitian Li 

Columbia University, New York, NY, USA

Abstract

Motivated by fast matrix multiplication and recent connections between asymptotic tensor rank and fine-grained complexity, we revisit classical tools from the matrix multiplication literature and develop a framework for obtaining improved asymptotic rank upper bounds for tensors beyond matrix multiplication.

In the 1980s, Coppersmith–Winograd and Strassen discovered a series of *speedup theorems* for asymptotic rank: in certain regimes, one can extract additional terms from a border rank upper bound on a tensor T , and then use these terms to obtain an improved asymptotic rank of T . We establish general speedup theorems that subsume these results and enable quantitative improvements. Two representative applications are:

1. The asymptotic rank of the small Coppersmith–Winograd tensor cw_q is less than its border rank. For instance, we prove $\underline{R}(\text{cw}_2) < 3.931$, improving on $\underline{R}(\text{cw}_2) = 4$. It is known that $\underline{R}(\text{cw}_2) = 3$ would imply $\omega = 2$.
2. A general improvement over Strassen’s bound: we obtain an upper bound below $d^{2\omega/3}$ on the asymptotic rank of any $d \times d \times d$ tensor.

To make full use of speedups, we analyze degenerations in which both sides are nontrivial direct sums, a setting where the optimal quantitative bound one can achieve was previously unclear. We do so via an approach we call *Strassen calculus*: a systematic method for converting such degeneration data into explicit asymptotic rank bounds using Strassen’s theory of the asymptotic spectrum.

2012 ACM Subject Classification Theory of computation → Algebraic complexity theory

Keywords and phrases tensor rank, matrix multiplication, Coppersmith–Winograd tensor, asymptotic spectrum

Digital Object Identifier 10.4230/LIPIcs.CCC.2026.36

Funding *Josh Alman*: Supported in part by NSF Grant CCF-2238221 and a Packard Foundation Fellowship.

Baitian Li: Supported in part by NSF Grant CCF-2238221, a Packard Foundation Fellowship, and a Columbia SEAS Presidential Fellowship.

1 Introduction

The exponent of matrix multiplication, ω , is defined as the smallest real number such that two $n \times n$ matrices can be multiplied using at most $n^{\omega+o(1)}$ field operations. Because of the fundamental and widespread applications of matrix multiplication algorithms, determining the value of ω has become one of the most important open problems in computer science. Strassen’s famous algorithm [43] showed the first nontrivial bound $\omega \leq \log_2(7) < 2.81$. Since then, a long line of follow-up work has culminated in the current best bound $\omega < 2.371339$ [2]. The only known lower bound is the straightforward $\omega \geq 2$, since the input and output sizes are n^2 . In fact, it is frequently conjectured that $\omega = 2$.

It has been known since Strassen’s work [44] that determining ω is equivalent to understanding the *asymptotic rank* of the matrix multiplication tensor. In fact, there are three different rank notions for a (3-mode/3-dimensional) tensor T which have played a prominent



© Josh Alman and Baitian Li;
licensed under Creative Commons License CC-BY 4.0

41st Computational Complexity Conference (CCC 2026).

Editor: Dana Moshkovitz; Article No. 36; pp. 36:1–36:42



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



role in the study of matrix multiplication algorithms: the rank $R(T)$, the *border rank* $\underline{R}(T)$ (roughly, the smallest r such that T is the limit of a sequence of tensors of rank r), and the asymptotic rank $\underline{\underline{R}}(T)$, where one takes the ranks of large tensor powers of T :

$$\underline{\underline{R}}(T) := \lim_{n \rightarrow \infty} (R(T^{\otimes n})^{1/n}).$$

For instance, Strassen's algorithm [43] showed that the rank of the tensor for multiplying 2×2 matrices is at most 7. More generally, because of the recursive structure of matrix multiplication under tensor products, to upper bound ω it suffices to bound the asymptotic rank of matrix multiplication tensors. It also suffices to bound the border rank because of a connection due to Bini [8] which showed that, for any tensor T , we have

$$\underline{\underline{R}}(T) \leq \underline{R}(T) \leq R(T).$$

More recently, the asymptotic rank of tensors has found other algorithmic applications beyond matrix multiplication. Björklund and Kaski [10] and Pratt [34] showed that improved asymptotic rank bounds for certain tensors would give a faster algorithm for the *Set Cover* problem, which is conjectured in fine-grained complexity theory to be impossible. Follow-up work showed similar connections for other algorithmic problems including computing the chromatic number of a graph [9] and computing the permanent of a matrix [11]. Many exponential and parameterized algorithms rely on a (generalized) convolution problem, the complexity of which is determined by the asymptotic rank of the corresponding tensor [14].

Amazingly, Strassen conjectured that *every* tensor has the minimum possible asymptotic rank given its dimensions (as long as it is “tight”, a technical condition satisfied by most tensors of interest):

► **Conjecture 1** (Asymptotic Rank Conjecture [45]). *Every $d \times d \times d$ tensor T which is concise and tight has $\underline{\underline{R}}(T) = d$.*

This would imply $\omega = 2$ and lead to breakthrough algorithms for all the aforementioned problems.

Unfortunately, despite the importance of asymptotic rank, we don't know many techniques for bounding the asymptotic rank of a tensor T beyond bounding the border rank of T or a small tensor power of T . Even this approach, which often gives suboptimal bounds, is difficult to use, as bounding the rank or border rank of fairly small tensors has proven to be quite challenging. For example, it remains an open problem to determine the rank or border rank of $\langle 3, 3, 3 \rangle$, the tensor for multiplying 3×3 matrices (which is a $9 \times 9 \times 9$ tensor).

Beyond this, the two most successful techniques for bounding asymptotic rank for the purposes of bounding ω have been the laser method (a combinatorial technique for restricting tensor powers of structured tensors into direct sums of matrix multiplication tensors), and the *asymptotic sum inequality* (a.k.a. *Schönhage's τ theorem*) for taking advantage of such direct sums:

► **Theorem 2** (Asymptotic Sum Inequality [38]). *Suppose a direct sum of matrix multiplication tensors has a border rank bound*

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

and that this is nontrivial (i.e., $n_i m_i p_i > 1$ for at least one i). Let τ be the unique solution of the transcendental equation

$$\sum_{i=1}^{\ell} (n_i m_i p_i)^{\tau} = r.$$

Then, $\omega \leq 3\tau$.

It was then shown in follow-up work of Coppersmith and Winograd [22] that the asymptotic sum inequality is *never tight*: One always has the strict inequality $\omega < 3\tau$! Their proof is constructive, in the sense that they found a systematic way to convert a given direct sum construction $\underline{R}(T) \leq r$ into another one $\underline{R}(T') \leq r'$, which yields a *more effective* bound on the matrix multiplication exponent ω .

This type of result is reminiscent of Blum's classical *speedup* theorem for Turing machines [13], and so we refer to such results as speedup theorems in this paper. That said, Blum's speedup theorem concerns non-constructible time functions, whereas the tensor speedup theorems we discuss here concern explicit problems of great interest like matrix multiplication.

Our approach. In this paper, we revisit the approach of bounding the asymptotic ranks of tensors via speedup theorems. We establish a generalization of prior speedup theorems, and analyze them using a method we call *Strassen calculus* which makes use of Strassen's theory of the asymptotic spectrum, in order to give new quantitative asymptotic rank bounds. We will discuss this in much more detail in Section 2 below, but we first present some applications of our approach.

Small Coppersmith–Winograd tensor. As a main application, we give the first nontrivial upper bound on the asymptotic rank of the *small Coppersmith–Winograd tensor*, cw_q . This is, for any integer $q \geq 2$, the $(q+1) \times (q+1) \times (q+1)$ tensor given by

$$\text{cw}_q = \sum_{i=1}^q x_0 y_i z_i + x_i y_0 z_i + x_i y_i z_0.$$

The cw_q tensor was introduced in the celebrated work of Coppersmith and Winograd [23]. They further developed Strassen's laser method, and applied it first to cw_q as their easy (warmup) construction [23, §6], then later to the *big Coppersmith–Winograd tensor*, CW_q , the $(q+2) \times (q+2) \times (q+2)$ tensor which adds three “corner terms” to cw_q ,

$$\text{CW}_q = x_0 y_0 z_{q+1} + x_0 y_{q+1} z_0 + x_{q+1} y_0 z_0 + \text{cw}_q.$$

Coppersmith and Winograd showed that both cw_q and CW_q have border rank $q+2$ (and thus, asymptotic rank at most $q+2$). By virtue of the extra corner terms, they found CW_q to be more valuable in their laser method analysis, giving their bound $\omega < 2.376$. The big tensor CW_q has since been used for all record-holding bounds on ω , including the current bound $\omega < 2.371339$ [2].

However, the tensor CW_q is now known to be subject to a number of *barrier* results [6, 4, 18, 1] which show that it is *not* valuable enough to prove $\omega = 2$. (The exact numerical barrier depends on which techniques one allows to analyze CW_q , but the most general shows that even with the theoretically optimal asymptotic restrictions/degenerations, one cannot achieve a better bound on ω than 2.16 using CW_q .)

There is no known barrier like this to using cw_q . The key difference is in the dimensions of cw_q compared to CW_q . Since CW_q is a concise (roughly meaning “nondegenerate”) tensor of dimensions $(q+2) \times (q+2) \times (q+2)$, it follows that $\underline{R}(\text{CW}_q) \geq q+2$, and so the bound $\underline{R}(\text{CW}_q) \leq q+2$ from [23] is tight. By contrast, for cw_q there is a gap between the lower bound $\underline{R}(\text{cw}_q) \geq q+1$ from its dimensions, and the upper bound $\underline{R}(\text{CW}_q) \leq q+2$ from [23]. Using their laser method analysis, Coppersmith and Winograd [23, §11] observed that closing this gap could have major implications:

► **Theorem 3** ([23]). *If $\mathbb{R}(\text{cw}_2) = 3$ then $\omega = 2$.*

Prior to our work, no bound better than Coppersmith and Winograd’s border rank bound $\mathbb{R}(\text{cw}_2) \leq \underline{\mathbb{R}}(\text{cw}_2) = 4$ was known. But because of Theorem 3, many prior works have studied and asked about improved bounds on $\mathbb{R}(\text{cw}_q)$. See, for instance, [15, Remark 15.44 and Exercise 15.25], [12, Problem 9.8], and [21, Question 3].

Most recently, Conner, Landsberg, Gesmundo, and Ventura [21] studied whether the border ranks of small Kronecker powers $\text{cw}_q^{\otimes k}$ for $k = 2, 3$ could be used to improve the upper bound on $\mathbb{R}(\text{cw}_q)$, but they achieved a negative result: $\underline{\mathbb{R}}(\text{cw}_q^{\otimes 2}) = (q + 2)^2$ (for all $q > 2$) and $\underline{\mathbb{R}}(\text{cw}_q^{\otimes 3}) = (q + 2)^3$ (for all $q > 4$). This suggests that one may need to take large Kronecker powers of cw_q before any border rank submultiplicativity is observed.

As an application of our approach for bounding the asymptotic ranks of tensors, we achieve the first upper bound improvement:

► **Theorem 4.** $\mathbb{R}(\text{cw}_2) < 3.931$.

More generally, for any $q \geq 2$, we show there is a $\delta_q > 0$ such that $\mathbb{R}(\text{cw}_q) < q + 2 - \delta_q$; a table of these bounds for $2 \leq q \leq 10$ is given in Table 1 in Section 7.1 below.

We remark that, in hindsight, the existence of such a $\delta_q > 0$ actually follows from Strassen’s original work on the asymptotic spectrum [41, Theorem 3.14]. That said, we believe that a quantitative bound on how large δ_q is requires the new analysis we introduce here. We will discuss this more in the technique overview in the next section.

Improved general upper bound on asymptotic rank. Beyond Theorem 4, we prove additional upper bounds on the asymptotic rank of tensors, as well as new restrictions involving direct sums of tensors of interest. As one example, we obtain a new upper bound on the maximum possible asymptotic rank of a $d \times d \times d$ tensor.

For small d , the generic rank bounds [39, 32] already yield nontrivial estimates, and recent geometric methods can slightly improve them [30, 31]; however, these bounds remain $\Omega(d^2)$. For larger d , the best known general upper bound instead comes from a generic reduction of an arbitrary tensor to matrix multiplication, which implies that every $d \times d \times d$ tensor T satisfies $\mathbb{R}(T) \leq d^{\frac{2}{3}\omega}$ [41].

It is natural to wonder whether this can be improved, and recently Kaski [29, Open Problem 1] asked this explicitly in the context of further understanding sequences of tensors related to Strassen’s asymptotic rank conjecture. We improve it as follows:

► **Theorem 5.** *For any $d \geq 3$, any $d \times d \times d$ tensor T satisfies*

$$\mathbb{R}(T) \leq \sqrt{d^{\frac{4}{3}\omega} - \left\lfloor \frac{d}{3} \right\rfloor \cdot (2^{\frac{\omega}{3}} - 1)d^{\frac{2}{3}\omega}} < d^{\frac{2}{3}\omega}.$$

2 Technique overview

2.1 Strassen calculus: Bounds via asymptotic spectrum

Traditionally, techniques for bounding the matrix multiplication exponent ω proceed by constructing a chain of border-rank upper bounds. This requires carefully tracking the orders of magnitude of intermediate tensors, and typically culminates in a degeneration of the form $\langle n, n, n \rangle \leq \langle r \rangle$, which bounds ω via the basic inequality $\omega \leq \log_n \mathbb{R}(\langle n, n, n \rangle)$. A major drawback is that these arguments often rely on *ad hoc* derivation chains. A prominent example is the standard proof of the asymptotic sum inequality (cf. [15, §15.5], [12, §7]), which hinges on a specific and rather involved degeneration.

Furthermore, it is often unclear whether such a derivation yields the optimal bound theoretically obtainable from a given set of inequalities. For instance, in 1981, Schönhage [38] obtained the (record-holding at the time) bound $\omega < 2.521801$ by proving and analyzing the direct sum degeneration (specifically, setting $k = 5$ and $n = 10$):

$$\langle 1, k, 2n \rangle \oplus \langle n, 2, k \rangle \oplus \langle 2k, n, 1 \rangle \leq \langle 2(n+1)(k+1) \rangle \oplus k \odot \langle 1, 1, 2 \rangle.$$

Schönhage’s analysis, however, was not optimal. Romani [36] later improved this estimate to $\omega < 2.516649$ by exploiting the same identity more effectively. Even this is not tight: the best bound obtainable from the identity is $\omega \leq 3\tau < 2.516094$, where τ is the solution of the following equation (which looks like a natural generalization of asymptotic sum inequality):

$$3 \cdot (2kn)^\tau \leq 2(n+1)(k+1) + k \cdot 2^\tau, \quad k = 5, n = 10.$$

This was simultaneously noted by Pan [33, Appendix A] and Schönhage (and as claimed in Pan’s paper, a full proof is included in a unpublished manuscript of Schönhage). As far as we are aware, there is still no accessible proof of this analysis, likely because the community’s attention shifted to other methods to design matrix multiplication algorithms. We provide a proof in Appendix A below that $\omega \leq 2.516094$ can be achieved by analyzing this direct sum identity, and is the *optimal* bound one can achieve from it.

In Section 4, we introduce *Strassen calculus* – a systematic framework for deriving bounds from a fixed collection of degenerations between direct sums of tensors – via Strassen’s theory of the *asymptotic spectrum* [41, 42, 46].

Informally, the asymptotic spectrum provides a family of numerical “measures” for tensors. Concretely, an element of the asymptotic spectrum $\mathcal{X}$ is a function

$$\phi : \{\mathbb{F}\text{-tensors}\} \rightarrow \mathbb{R}_{\geq 0}$$

that behaves well with respect to the tensor operations. We refer to such a function ϕ as a *spectral point*; see Section 4 for the precise definitions.

The key structural statement underlying the theory of the asymptotic spectrum is Strassen duality (Theorem 6). Roughly speaking, it says that asymptotic degeneration relations are exactly the relations that hold after applying *all* spectral points: if T asymptotically degenerates to S , then $\phi(T) \leq \phi(S)$ for every ϕ , and conversely, if this inequality holds for all ϕ , then this implies the existence of the asymptotic degeneration. In particular, this characterizes asymptotic rank via a max formula,

$$\underline{R}(T) = \max_{\phi \in \mathcal{X}} \phi(T).$$

Conceptually, this turns questions about asymptotic rank into optimization problems over the asymptotic spectrum.

Of course, the space $\mathcal{X}$ of all spectral points is defined *nonconstructively* and may be *infinite-dimensional a priori*, so constructing explicit spectral points is generally difficult, although there has been significant progress in the past decade [19, 17, 37]. Our approach is therefore not to “parametrize” $\mathcal{X}$ directly. Instead, we start from a fixed collection of tensor degenerations, translate them into numerical inequalities that must be satisfied by every spectral point, and then use Strassen duality to turn the resulting constraints into explicit upper bounds on asymptotic rank. In favorable regimes one can solve the corresponding optimization problems analytically; more generally, one can obtain sharp bounds via computer-aided evaluation. Moreover, Strassen calculus is *complete*: for a fixed

set of degenerations, optimizing over these constraints yields the best bound obtainable from those relations. In Appendix A, we provide a rigorous treatment of the notion of “optimal usage”, and show the completeness under this formalism.

We emphasize that although the theory of the asymptotic spectrum was originally motivated by the study of tensor rank, its subsequent development has followed a curious trajectory. While the framework has flourished in other areas such as (quantum) information theory, its application to the *asymptotic rank* of tensors has remained largely conceptual. Prior to this work, the theory occupied a somewhat dormant role in its own founding domain, as to the best of our knowledge, *no* practical instances were identified to yield concrete, improved upper bounds.

The following examples illustrate the utility of the framework in other areas, and highlight the gap between its success elsewhere and its (so far) limited impact on tensor rank:

- *Strassen’s Foundation.* In his seminal paper [41], Strassen introduced the asymptotic spectrum to provide a new proof of the asymptotic sum inequality, setting the stage for a theory that, while foundational, remained difficult to apply to specific new bounds.
- *Barriers to Fast Matrix Multiplication.* Improvements to the exponent ω over the past thirty years have largely relied on analyzing the big Coppersmith–Winograd tensor CW_q . However, Christandl, Vrana, and Zuiddam [18] utilized spectral points to show that the laser method – even in a highly generalized form – applied on big CW tensors, cannot yield an upper bound better than 2.16. Subsequent work [16] has established analogous barriers for the rectangular case.
- *Graph Theory.* Zuiddam [49] applied the general theory of asymptotic spectrum (not confined to tensors) to study the Shannon capacity $\Theta(G)$ of graphs. Wigderson and Zuiddam [47, Theorem 4.1] utilized the asymptotic spectrum to show a striking equivalence: for graphs G and H , where $G \boxtimes H$ and $G \sqcup H$ denote the strong product and disjoint union respectively, $\Theta(G \boxtimes H) > \Theta(G)\Theta(H)$ if and only if $\Theta(G \sqcup H) > \Theta(G) + \Theta(H)$.¹
- *Circuit Complexity.* Robere and Zuiddam [35] developed an analogous theory to study and generalize some results in the complexity of catalytic branching programs. Alman and Li [3] applied asymptotic spectrum theory to derive a duality theorem for depth-2 circuits, resolving the problem of optimally combining fixed constructions in that context.

These developments demonstrate that Strassen duality is a mature and productive framework in other contexts. Our work aims to *re-apply* this machinery to its original setting by identifying concrete tensor degenerations for which it provably yields new, improved upper bounds.

In Section 4 we provide two self-contained examples of Strassen calculus. They focus on how one can turn degenerations of the form

$$T \oplus \langle 1, t, 1 \rangle \leq \langle r \rangle \oplus \langle 1, s, 1 \rangle$$

into explicit upper bounds on asymptotic rank. On the one hand, from such identities – with rectangular matrix multiplication tensors on the right-hand side – it was not previously understood what the optimal resulting asymptotic rank bound is. On the other hand, relations of this type (and more complicated variants) arise naturally in our generalized speedup theorems. For this reason, it is important to use the Strassen calculus in conjunction with our speedups.

¹ To appreciate this result, note that examples of graphs G, H satisfying the first strict inequality were found in [26], yet it took decades before examples satisfying the second were discovered [5].

In particular, in Section 4.2, our first example is a *toy example* illustrating the basic principles of the framework, together with a detailed derivation. Despite its simplicity, it already yields a novel and interesting consequence: every $3 \times 3 \times 3$ tensor T satisfies $\underline{R}(T) \leq 3 + 2^{\omega/3}$ (Corollary 12), giving a first glimpse of the calculus in action. (Previously, to the best of our knowledge, the upper bound on $\underline{R}(T)$ for a generic $3 \times 3 \times 3$ tensor T was 5, which is the generic rank. This result improves it to $3 + 2^{\omega/3} < 4.7297$.)

2.2 Speedup theorems

In Section 5, we prove a general speedup theorem that subsumes the results of Coppersmith–Winograd [22] and Strassen [41, Theorem 3.14]. The Coppersmith–Winograd theorem fits the following paradigm: starting from a border rank bound $T \leq \langle r \rangle$, one can (in certain regimes) “speed up” the degeneration by taking a direct sum with a one-slice matrix multiplication tensor on both sides, obtaining

$$T \oplus \langle 1, t, 1 \rangle \leq \langle r \rangle \oplus \langle 1, s, 1 \rangle.$$

Strassen later generalized this by showing that the right-hand side does not necessarily need to be a diagonal tensor (i.e., a border rank bound). Specifically, a degeneration $T \leq S$ can also be accelerated to $T \oplus \langle 1, t, 1 \rangle \leq S \oplus \langle 1, s, 1 \rangle$. Our main theorem further generalizes this idea: under suitable hypotheses, it allows one to add *arbitrary* tensors on both sides of a degeneration, producing relations of the form $T \oplus T' \leq S \oplus S'$.

In Section 6, we present three concrete instantiations of the general speedup theorem. Each has different strengths and ranges of applicability; together, they form the main toolkit we use to derive explicit upper bounds on asymptotic rank.

- The first result relates a certain variant of the min-rank problem to one-slice speedups of the form $T \oplus \langle 1, t, 1 \rangle \leq S \oplus \langle 1, s, 1 \rangle$, thereby recovering Strassen’s speedup theorem.
- The second result shows that, in a suitable regime, a degeneration of the form $T \oplus \langle 1, t, 1 \rangle \leq S'$ can be further accelerated to $T \oplus \langle 1, t', 1 \rangle \leq S'$ for some $t' > t$, while keeping the right-hand side fixed. In particular, if one takes an identity obtained from the first result and takes a tensor power of both sides, this acceleration always applies; thus the first method can be *iterated*. This generalizes a result of Coppersmith and Winograd [22, Theorem B].
- The third result shows that, in a certain regime, one can add multiple one-slice tensors on both sides, obtaining degenerations of the form $T \oplus \bigoplus_i \langle 1, t_i, 1 \rangle \leq S \oplus \bigoplus_j \langle 1, s_j, 1 \rangle$.

As a motivating illustration, we use these methods to show that any non-minimal border rank bound $\underline{R}(T) \leq r$ can be strictly improved (a *qualitative* fact first established by Strassen [41, Theorem 3.14]). The three instantiations above are formalized as Theorem 26, Theorem 28, and Theorem 29, and we give a *quantitative* comparison of the improvement achieved by each method using Strassen calculus.

In Section 7, we apply these methods to obtain explicit bounds for concrete tensors of interest.

- In Section 7.1, we prove Theorem 4, improving the asymptotic rank of the small Coppersmith–Winograd tensors cw_q ; for instance, we improve the best known bound for $\underline{R}(\text{cw}_2)$ from 4 (coming from border rank) to 3.931 (see Table 1).
- In Section 7.2, we prove Theorem 5, improving general upper bounds on asymptotic rank.
- In Section 7.3, we use the speedup theorem to construct new degenerations between direct sums of matrix multiplication tensors in which the right-hand side is itself a direct sum of nontrivial matrix multiplication tensors. While this does not (at present) improve the matrix multiplication exponent, the resulting identities do not appear to follow easily from previously known ones, and may provide insights for future developments.

Finally, it is natural to wonder whether this method yields faster algorithms for problems such as matrix multiplication or set cover. We find that our techniques are still not effective enough, although we are optimistic that further improvements along these lines may succeed. See Remark 35 and Remark 40 for further discussion.

3 Preliminaries

For a positive integer n , let $[n] := \{1, \dots, n\}$. We use Iverson bracket notation: for a proposition P , let $\llbracket P \rrbracket = 1$ if P is true and $\llbracket P \rrbracket = 0$ otherwise.

3.1 Tensors

Our notation largely follows [24, 48, 2].

Fix a base field $\mathbb{F}$. Let U, V, W be finite-dimensional vector spaces with bases $\{x_1, \dots, x_n\}$ of U , $\{y_1, \dots, y_m\}$ of V and $\{z_1, \dots, z_p\}$ of W . A tensor is an element $T \in U \otimes V \otimes W$. With respect to the chosen bases, we write T via its coefficients $a_{ijk} \in \mathbb{F}$ as

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

We say U, V, W are the *modes* of the tensor.

For two tensors $S \in U \otimes V \otimes W$ and $T \in U' \otimes V' \otimes W'$, we write their direct sum $S \oplus T$ as the image of their sum inside the common space $(U \oplus U') \otimes (V \oplus V') \otimes (W \oplus W')$, and their tensor product as the natural image inside the tensor product space $(U \otimes U') \otimes (V \otimes V') \otimes (W \otimes W')$.

A *restriction* from $S \in U \otimes V \otimes W$ to $T \in U' \otimes V' \otimes W'$ is specified by linear maps $A: U \rightarrow U'$, $B: V \rightarrow V'$ and $C: W \rightarrow W'$ such that $T = (A \otimes B \otimes C)S$. Concretely,

$$T = \sum_{i,j,k} a_{ijk} A(x_i) B(y_j) C(z_k).$$

In this case we say S *restricts to* T , and write $T \leq S$.

A *degeneration* is a one-parameter family of linear maps $A_\lambda, B_\lambda, C_\lambda$ (with entries rational in λ) such that $(A_\lambda \otimes B_\lambda \otimes C_\lambda)S = T + O(\lambda^d)$ for some d . In other words, write each entry of $A_\lambda, B_\lambda, C_\lambda$ as a rational function in λ and then expand the resulting expression as a Laurent series in λ :

$$A_\lambda = \sum_{i \geq i_0} A_i \lambda^i,$$

and similarly for B_λ, C_λ , we have the λ^ℓ -coefficient of $(A_\lambda \otimes B_\lambda \otimes C_\lambda)S$ is

$$\sum_{i+j+k=\ell} (A_i \otimes B_j \otimes C_k)S.$$

We require that the λ^0 -coefficient is exactly T , and that the λ^ℓ -coefficient is zero for all $\ell < 0$. In this case we say T is a *degeneration* of S , write S *degenerates to* T , and denote this by $T \trianglelefteq S$.

Two central families of tensors are diagonal tensors and matrix multiplication tensors. For any $n \geq 1$, the diagonal tensor is

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

and for any $n, m, p \geq 1$, the matrix multiplication tensor is

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

The rank $R(T)$ of a tensor T is the minimum r such that $T \leq \langle r \rangle$. Similarly, the subrank $Q(T)$ is the maximum q such that $\langle q \rangle \leq T$ (equivalently, $T \geq \langle q \rangle$). The *border rank* $\underline{R}(T)$ and *border subrank* $\underline{Q}(T)$ are defined analogously by replacing $\leq$ with $\leqslant$.

We write the n -fold direct sum of T as $n \odot T$; note that $n \odot T$ is equivalent to $\langle n \rangle \otimes T$. The *asymptotic rank* and *subrank* are defined as

$$\begin{aligned} \widetilde{R}(T) &= \lim_{n \rightarrow \infty} R(T^{\otimes n})^{1/n}, \\ \widetilde{Q}(T) &= \lim_{n \rightarrow \infty} Q(T^{\otimes n})^{1/n}. \end{aligned}$$

By Fekete's lemma, for $\widetilde{R}$ ($\widetilde{Q}$ resp.), the $\lim_{n \rightarrow \infty}$ can also be replaced by $\inf_{n \geq 1}$ ($\sup_{n \geq 1}$ resp.). Therefore, the matrix multiplication exponent ω is characterized by $\widetilde{R}(\langle n, n, n \rangle) = n^\omega$.

4 Strassen calculus

We first briefly recall the notion of the asymptotic spectrum. The asymptotic spectrum of $\mathbb{F}$ -tensors is the set $\mathcal{X} = \mathcal{X}_{\mathbb{F}}$ of functions ϕ from tensors to nonnegative real numbers $\mathbb{R}_{\geq 0}$ satisfying:

- Additive under direct sum: $\phi(T \oplus S) = \phi(T) + \phi(S)$.
- Multiplicative under tensor product: $\phi(T \otimes S) = \phi(T)\phi(S)$.
- Normalized on diagonal tensors: $\phi(\langle n \rangle) = n$.
- Monotone under restriction (degeneration²): $\phi(T) \leq \phi(S)$ when $T \leq S$ ($T \leqslant S$).

From this definition, it is immediate that $\mathcal{X}$ only “remembers” asymptotic relations: if $T^{\otimes n} \leq S^{\otimes n+o(n)}$, then $\phi(T) \leq \phi(S)$ for all $\phi \in \mathcal{X}$. Surprisingly, the converse holds as well; this is the *Strassen duality theorem* (cf. [41, Theorem 3.2], [47, §3]).

► **Theorem 6 (Strassen duality).** *For two tensors T, S , the following are equivalent:*

1. $T^{\otimes n} \leq S^{\otimes n+o(n)}$,
2. $T^{\otimes n} \leqslant S^{\otimes n+o(n)}$,
3. For any $\phi \in \mathcal{X}$, have $\phi(T) \leq \phi(S)$.

Moreover,

$$\widetilde{R}(T) = \lim_{n \rightarrow \infty} R(T^{\otimes n})^{1/n} = \lim_{n \rightarrow \infty} \underline{R}(T^{\otimes n})^{1/n} = \max_{\phi \in \mathcal{X}} \phi(T),$$

and

$$\widetilde{Q}(T) = \lim_{n \rightarrow \infty} Q(T^{\otimes n})^{1/n} = \lim_{n \rightarrow \infty} \underline{Q}(T^{\otimes n})^{1/n} = \min_{\phi \in \mathcal{X}} \phi(T).$$

For our purposes, it is enough to know Strassen duality; no further background on the asymptotic spectrum is required. For a complete treatment – including further properties and applications – we refer to Strassen's founding paper [41] and the survey of Wigderson and Zuiddam [47].

² It turns out that *restriction* and *degeneration* give equivalent definitions, i.e., they yield the same collection of functions; cf. [41, Theorem 3.1].

4.1 Matrix multiplication

Before proceeding to the toy example, we first introduce a fundamental concept to facilitate our use of Strassen calculus.

► **Proposition 7** ([41, Proposition 4.3]). *For any $\phi \in \mathcal{X}$, there exist unique real numbers $\theta_1, \theta_2, \theta_3$ such that*

$$\phi(\langle n, m, p \rangle) = n^{\theta_1} m^{\theta_2} p^{\theta_3}$$

for any n, m, p .

Proof. Let $\theta_1 = \log_2 \phi(\langle 2, 1, 1 \rangle)$. We will show that $\phi(\langle n, 1, 1 \rangle) = n^{\theta_1}$ for all n . For any k , let $t = \lfloor \log_2(n^k) \rfloor$, thus $2^t \leq n^k < 2^{t+1}$ and we have the restrictions $\langle 2^t, 1, 1 \rangle \leq \langle n^k, 1, 1 \rangle \leq \langle 2^{t+1}, 1, 1 \rangle$. Since ϕ is monotone, we have

$$\phi(\langle 2^t, 1, 1 \rangle) \leq \phi(\langle n^k, 1, 1 \rangle) \leq \phi(\langle 2^{t+1}, 1, 1 \rangle).$$

Since $\langle n^k, 1, 1 \rangle = \langle n, 1, 1 \rangle^{\otimes k}$ and ϕ is multiplicative, we can pull out the exponent:

$$\phi(\langle 2, 1, 1 \rangle)^t \leq \phi(\langle n, 1, 1 \rangle)^k \leq \phi(\langle 2, 1, 1 \rangle)^{t+1}.$$

Thus

$$2^{-\theta_1} n^{k\theta_1} \leq 2^{\theta_1 t} \leq \phi(\langle n, 1, 1 \rangle)^k \leq 2^{\theta_1(t+1)} \leq 2^{\theta_1} n^{k\theta_1}.$$

Taking k th roots and letting $k \rightarrow \infty$, we conclude $\phi(\langle n, 1, 1 \rangle) = n^{\theta_1}$. Similarly, we can obtain θ_2, θ_3 such that $\phi(\langle 1, m, 1 \rangle) = m^{\theta_2}$ and $\phi(\langle 1, 1, p \rangle) = p^{\theta_3}$.

Finally, since $\langle n, m, p \rangle = \langle n, 1, 1 \rangle \otimes \langle 1, m, 1 \rangle \otimes \langle 1, 1, p \rangle$ and ϕ is multiplicative, we have

$$\begin{aligned} \phi(\langle n, m, p \rangle) &= \phi(\langle n, 1, 1 \rangle \otimes \langle 1, m, 1 \rangle \otimes \langle 1, 1, p \rangle) \\ &= \phi(\langle n, 1, 1 \rangle) \cdot \phi(\langle 1, m, 1 \rangle) \cdot \phi(\langle 1, 1, p \rangle) \\ &= n^{\theta_1} m^{\theta_2} p^{\theta_3}. \end{aligned}$$

◀

Therefore, the behavior of the asymptotic spectrum on matrix multiplication tensors can be parametrized just by θ_1, θ_2 and θ_3 . This leads to the notion of the *spectrum of matrix multiplication*.

► **Definition 8** ([41, §4]). *The (logarithmic) spectrum $\Delta_{\text{MM}} \subset \mathbb{R}^3$ of matrix multiplication are the points $(\theta_1, \theta_2, \theta_3)$ that satisfy Proposition 7.*

All the basic facts of matrix multiplication tensor transfer to statements of Δ_{MM} .

► **Proposition 9.** *Every θ_i lies in $[0, 1]$. The value $\theta_1 + \theta_2 + \theta_3$ over Δ_{MM} attains minimum 2 and maximum ω .*

Proof. For any positive integer n , since $\mathbb{Q}(\langle n, 1, 1 \rangle) = 1$ and $\mathbb{R}(\langle n, 1, 1 \rangle) = n$, we have $\min \theta_1 = 0$ and $\max \theta_1 = 1$, and similarly for θ_2, θ_3 . The value $\theta_1 + \theta_2 + \theta_3$, by definition, equals $\log_n \phi(\langle n, n, n \rangle)$. Thus by Theorem 6, we have

$$\max_{(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}} (\theta_1 + \theta_2 + \theta_3) = \max_{\phi \in \mathcal{X}} \log_n \phi(\langle n, n, n \rangle) = \log_n \mathbb{R}(\langle n, n, n \rangle),$$

and it follows from the definition of ω that $\mathbb{R}(\langle n, n, n \rangle) = n^\omega$. Similarly, we have

$$\min_{(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}} (\theta_1 + \theta_2 + \theta_3) = \log_n \mathbb{Q}(\langle n, n, n \rangle),$$

and matrix multiplication has optimal asymptotic subrank $\mathbb{Q}(\langle n, n, n \rangle) = n^2$ [40, Theorem 6.6].

◀

4.2 Toy example

► **Proposition 10.** *Suppose $\mathbb{F}$ is an algebraically closed field. For any $3 \times 3 \times 3$ tensor T , we have $T \leq \langle 3 \rangle \oplus \langle 1, 2, 1 \rangle$.*

Proof. It is known that degeneration can be characterized topologically (cf. [40, Theorem 5.8], [15, Theorem 20.24]). Let $C \subseteq \mathbb{F}^3 \otimes \mathbb{F}^3 \otimes \mathbb{F}^3$ be the set of tensors that can be restricted from $\langle 3 \rangle \oplus \langle 1, 2, 1 \rangle$; then $T \leq \langle 3 \rangle \oplus \langle 1, 2, 1 \rangle$ if and only if $T \in \overline{C}$, where $\overline{C}$ denotes the Zariski closure of C . Thus, it suffices to prove $T \leq \langle 3 \rangle \oplus \langle 1, 2, 1 \rangle$ for a *generic* T , and the statement follows.

If T is generic, we may first assume that the z_1 -slice of T is a nonsingular matrix. By a change of basis, we assume that the z_1 -slice is of the form $x_1y_1 + x_2y_2 + x_3y_3$. Furthermore, we can assume the z_2 -slice is a diagonalizable matrix; we then apply the corresponding similarity transformation which preserves the form of the first slice and transforms the second slice into $\alpha_1x_1y_1 + \alpha_2x_2y_2 + \alpha_3x_3y_3$. Now we can write the tensor T as

$$\sum_{i=1}^3 x_i y_i (z_1 + \alpha_i z_2) + (xMy)z_3$$

for some matrix M . Letting λ be one of the eigenvalues of M , then we have

$$T = \sum_{i=1}^3 x_i y_i (z_1 + \alpha_i z_2 + \lambda z_3) + x(M - \lambda I)y z_3$$

where the first summation can be restricted from $\langle 3 \rangle$ and the second term is a matrix of rank at most 2, which is a restriction of $\langle 1, 2, 1 \rangle$. ◀

► **Proposition 11.** *If tensor T satisfies the degeneration relations on three directions,*

$$T \leq \langle r \rangle \oplus \langle s, 1, 1 \rangle, \tag{4.1}$$

$$T \leq \langle r \rangle \oplus \langle 1, s, 1 \rangle, \tag{4.2}$$

$$T \leq \langle r \rangle \oplus \langle 1, 1, s \rangle, \tag{4.3}$$

then its asymptotic rank satisfies $\mathbb{R}(T) \leq r + s^{\omega/3}$.

Proof. In this proof, we demonstrate the general principle of *Strassen calculus*.

Step 1: Pass to asymptotic spectrum. Let $\phi \in \mathcal{X}$ be a point of the asymptotic spectrum. Recall that ϕ is *monotone* – it preserves degenerations – so we can apply it to (4.1), resulting in

$$\phi(T) \leq \phi(\langle r \rangle \oplus \langle s, 1, 1 \rangle).$$

Using the *additivity* of ϕ , we can pull out the direct sum:

$$\phi(T) \leq \phi(\langle r \rangle) + \phi(\langle s, 1, 1 \rangle).$$

Moreover, ϕ is *normalized*, so $\phi(\langle r \rangle) = r$. As shown in Proposition 7, the value of a matrix multiplication tensor can be parametrized by $(\theta_1, \theta_2, \theta_3)$ uniquely determined by ϕ , such that $\phi(\langle s, 1, 1 \rangle) = s^{\theta_1}$. We can apply the same argument to (4.2) and (4.3), obtaining

$$\phi(T) \leq r + s^{\theta_1}, \quad \phi(T) \leq r + s^{\theta_2}, \quad \phi(T) \leq r + s^{\theta_3}.$$

Step 2: Use known inequalities to estimate spectral values. Each of the three degenerations gives an upper bound of $\phi(T)$. To obtain the tightest bound, we take the minimum:

$$\phi(T) \leq r + \min\{s^{\theta_1}, s^{\theta_2}, s^{\theta_3}\}.$$

To translate this into an absolute value independent of the spectral point ϕ , we utilize Proposition 9: $\theta_1 + \theta_2 + \theta_3 \leq \omega$. We remind the reader that this inequality essentially stems from the fact that $\mathbb{R}(\langle n, n, n \rangle) = n^\omega$. Since the θ_i are all non-negative, at least one must be less than or equal to the average $\omega/3$, thus

$$\phi(T) \leq r + s^{(\theta_1 + \theta_2 + \theta_3)/3} \leq r + s^{\omega/3}.$$

Final step: Return to primal estimate. Now we have an absolute upper bound of $\phi(T)$ for any ϕ . By Strassen duality theorem (Theorem 6), the asymptotic rank $\mathbb{R}(T)$ of T is the maximum of $\phi(T)$ over $\mathcal{X}$, so we get $\mathbb{R}(T) \leq r + s^{\omega/3}$. ◀

► **Corollary 12.** Any $3 \times 3 \times 3$ tensor T has $\mathbb{R}(T) \leq 3 + 2^{\omega/3}$.

Proof. Since asymptotic rank is invariant under field extension [15, Proposition 15.17], we may assume $\mathbb{F}$ is algebraically closed. Then by Proposition 10, we have $T \leq \langle 3 \rangle \oplus \langle 1, 2, 1 \rangle$. Applying Proposition 10 on the tensor with modes relabeled yields the other two degenerations required by Proposition 11. Thus we have $\mathbb{R}(T) \leq 3 + 2^{\omega/3}$. ◀

Previously, to the best of our knowledge, the upper bound on $\mathbb{R}(T)$ for a generic $3 \times 3 \times 3$ tensor T was 5, which is the generic rank. We see that $3 + 2^{\omega/3} < 3 + 2 = 5$ represents an improvement. Numerically, using the current upper bound for ω , we have $3 + 2^{\omega/3} < 4.7297$.

The generic rank (Strassen–Lickteig [39, 32]) establishes that for any $d \times d \times d$ tensor T , $T \leq \langle r_d \rangle$, where

$$r_d = \begin{cases} 5, & \text{if } d = 3, \\ \left\lceil \frac{d^3}{3d-2} \right\rceil, & \text{otherwise.} \end{cases}$$

This bound works well for small d , while for large d the current best asymptotic rank bound comes from $T \leq \langle d, 1, d \rangle$ (though we will refine this in Section 7.2). This suggests an alternative approach to improving the asymptotic rank of general tensors:

► **Problem 13.** Find an intermediate tensor S_d that admits degenerations for all $d \times d \times d$ tensors (or significant subclasses) to obtain better asymptotic rank estimates.

We leave this question for future research.

► **Remark 14.** For this example, it is not difficult to derive the same upper bound via *traditional derivations*. We encourage the reader to consider such an alternative proof, which may help in understanding the interplay between classical deductions and the operations in Strassen calculus. However, as the given degenerations become more complicated, the complexity of classical deduction grows rapidly and becomes unintuitive. In contrast, the *Strassen calculus* remains systematic, typically adding only a few lines of derivation. In certain regimes, such as those analyzed in Appendix A, the optimum of the optimization problem may be difficult to describe by an easy criterion, making it reasonable to rely on computer-aided evaluation. In such cases, a traditional derivation would need to incorporate different regimes and sensitively detect shifts in behavior as parameters change, which is a prohibitively complicated task.

4.3 Complicated example

In this subsection we present another example of Strassen calculus, based on a slightly more complicated set of identities that will also be used in later applications.

► **Proposition 15.** *If tensor T satisfies the degeneration relations on three directions*

$$T \oplus \langle t, 1, 1 \rangle \leq \langle r \rangle \oplus \langle s, 1, 1 \rangle, \quad (4.4)$$

$$T \oplus \langle 1, t, 1 \rangle \leq \langle r \rangle \oplus \langle 1, s, 1 \rangle, \quad (4.5)$$

$$T \oplus \langle 1, 1, t \rangle \leq \langle r \rangle \oplus \langle 1, 1, s \rangle, \quad (4.6)$$

for some $s, t \geq 1$, then its asymptotic rank satisfies

$$\underline{R}(T) \leq \begin{cases} r + s^{2/3} - t^{2/3}, & \text{if } t > s, \\ r + s^{\omega/3} - t^{\omega/3}, & \text{if } t \leq s. \end{cases}$$

Proof. Fix an arbitrary spectral point $\phi \in \mathcal{X}$. Applying ϕ to the three degeneration relations yields:

$$\phi(T) \leq r + s^{\theta_1} - t^{\theta_1} \quad (\text{apply } \phi \text{ to (4.4)})$$

$$\phi(T) \leq r + s^{\theta_2} - t^{\theta_2} \quad (\text{apply } \phi \text{ to (4.5)})$$

$$\phi(T) \leq r + s^{\theta_3} - t^{\theta_3} \quad (\text{apply } \phi \text{ to (4.6)})$$

where $\phi(\langle r \rangle) = r$ by normalization, and $\phi(\langle t, 1, 1 \rangle) = t^{\theta_1}$, $\phi(\langle 1, t, 1 \rangle) = t^{\theta_2}$, $\phi(\langle 1, 1, t \rangle) = t^{\theta_3}$ by Proposition 7.

Taking the strongest bound among the three directions yields

$$\begin{aligned} \phi(T) &\leq r + \min_{i \in \{1,2,3\}} (s^{\theta_i} - t^{\theta_i}) \quad (\text{take the minimum over directions}) \\ &= r + \min_{i \in \{1,2,3\}} \tau(\theta_i) \quad (\text{set } \tau(x) = s^x - t^x). \end{aligned}$$

Now we split into two cases.

Case 1: $t > s$. Then τ is decreasing on $\mathbb{R}_{\geq 0}$, so $\min_i \tau(\theta_i) = \tau(\max_i \theta_i)$. Since $\theta_1 + \theta_2 + \theta_3 \geq 2$ (Proposition 9), we have $\max_i \theta_i \geq 2/3$, and thus

$$\phi(T) \leq r + \tau(\max_i \theta_i) \leq r + \tau(2/3) = r + s^{2/3} - t^{2/3}.$$

Case 2: $t \leq s$. Then τ is increasing on $\mathbb{R}_{\geq 0}$, so $\min_i \tau(\theta_i) = \tau(\min_i \theta_i)$. Since $\theta_1 + \theta_2 + \theta_3 \leq \omega$ (Proposition 9), we have $\min_i \theta_i \leq (\theta_1 + \theta_2 + \theta_3)/3 \leq \omega/3$, and therefore

$$\phi(T) \leq r + \tau(\min_i \theta_i) \leq r + \tau(\omega/3) = r + s^{\omega/3} - t^{\omega/3}.$$

Since the above inequality holds for every $\phi \in \mathcal{X}$, Strassen duality gives

$$\underline{R}(T) = \max_{\phi \in \mathcal{X}} \phi(T) \leq \begin{cases} r + s^{2/3} - t^{2/3}, & \text{if } t > s, \\ r + s^{\omega/3} - t^{\omega/3}, & \text{if } t \leq s. \end{cases} \quad \blacktriangleleft$$

► **Remark 16.** The proof illustrates a typical feature of Strassen calculus: it can *turn lower bounds into upper bounds*. In the case $t > s$ we used the nontrivial inequality $\theta_1 + \theta_2 + \theta_3 \geq 2$, which ultimately comes from the optimal asymptotic subrank bound $\underline{Q}(\langle n, n, n \rangle) = n^2$.

5 Speedup theorems

5.1 Notation and blockwise restrictions

To clarify our statements, we first introduce notation for handling tensor modes and restrictions. Let $T \in U' \otimes V' \otimes W'$ and $S \in U \otimes V \otimes W$ be tensors. If a tensor mode is identified as a direct sum, such as $U = U_1 \oplus U_2$ and $U' = U'_1 \oplus U'_2$, the linear map A in a restriction $T \leq S$ can be written in blockwise form:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{ij} is a linear map from U_j to U'_i . Correspondingly, we may decompose the tensor S as $S = S_1 + S_2$ where $S_j \in U_j \otimes V \otimes W$. In this context, the restriction $T = (A \otimes B \otimes C)S$ can be expanded as:

$$T = \sum_{i,j} (A_{ij} \otimes B \otimes C)(S_j).$$

We also need to clarify the relationship between subspaces of functional spaces and the linear maps used in restrictions. Let $U^\vee$ denote the dual space of U , consisting of all linear functions $f: U \rightarrow \mathbb{F}$. Suppose $A \subseteq U^\vee$ is a subspace of dimension d . We can identify the entire subspace A with a surjective linear map $U \rightarrow \mathbb{F}^d$: choosing a basis $\{f_1, \dots, f_d\}$ for A defines the coordinates of the image in $\mathbb{F}^d$. While this specific map depends on the choice of basis, any two such choices yield the same resulting tensor up to isomorphism. Thus, we may treat a subspace of $U^\vee$ as an “aggregate” linear map in a restriction $(A \otimes B \otimes C)S$ without ambiguity.

5.2 Free-lunch speedup theorem

We begin by describing a condition that allows one to obtain a direct sum on the left-hand side *for free*, without the need to add any additional terms to the right-hand side. For readers unfamiliar with these works, it is worth noting that the arguments of Coppersmith–Winograd [22, Theorem A] and Strassen [41, Lemma 3.11] likewise proceed in two distinct stages. Our approach follows this established logic, beginning here with the formulation of the first step.

► **Theorem 17 (Free-lunch speedup).** *Let $T \in U' \otimes V' \otimes W'$ and $S \in U \otimes V \otimes W$ be two tensors, with a restriction $T \leq S$ given by the maps (A, B, C) . Let $C^\perp \subseteq W^\vee$ be the subspace of linear functions $f: W \rightarrow \mathbb{F}$ such that $(A \otimes B \otimes f)S = 0$. For any chosen subspace $C' \subseteq C^\perp$, define the subspaces $A' \subseteq U^\vee$ and $B' \subseteq V^\vee$ as follows:*

$$\begin{aligned} A' &= \{u \in U^\vee : (u \otimes B \otimes C')S = 0\}, \\ B' &= \{v \in V^\vee : (A \otimes v \otimes C')S = 0\}. \end{aligned}$$

Then we have the direct sum degeneration:

$$T \oplus T' \leq S,$$

where $T' = (A' \otimes B' \otimes C')S$.

Proof. The construction of the degeneration proceeds in two steps.

Step 1. We first construct a restriction by augmenting the linear maps. We define the new maps as (A, A') , (B, B') , and (C, C') , where the block structure corresponds to the direct sum of the target spaces. Consider the resulting tensor:

$$R = \left[\begin{pmatrix} A \\ A' \end{pmatrix} \otimes \begin{pmatrix} B \\ B' \end{pmatrix} \otimes \begin{pmatrix} C \\ C' \end{pmatrix} \right] S.$$

We examine the blocks of R arising from the terms involving C' . By the construction of our subspaces:

- $(A \otimes B \otimes C')S = 0$ by the definition of C' .
- $(A' \otimes B \otimes C')S = 0$ by the definition of A' .
- $(A \otimes B' \otimes C')S = 0$ by the definition of B' .

Consequently, the only potentially non-zero block involving C' is $(A' \otimes B' \otimes C')S = T'$. In the overall block structure of R , the diagonal blocks are $R_{111} = T$ and $R_{222} = T'$, while the blocks R_{112} , R_{122} , and R_{212} are guaranteed to be zero.

Step 2. We now eliminate the remaining junk terms (namely R_{121} , R_{211} , and R_{221}) via a *monomial degeneration*. We apply the following diagonal transformation to the tensor R :

$$\begin{pmatrix} 1 & \\ & \lambda \end{pmatrix} \otimes \begin{pmatrix} 1 & \\ & \lambda \end{pmatrix} \otimes \begin{pmatrix} \lambda^2 & \\ & 1 \end{pmatrix} R = \lambda^2(T + T') + O(\lambda^3).$$

By scaling by λ^{-2} and taking the limit as $\lambda \rightarrow 0$, the junk terms are eliminated, yielding the direct sum $T \oplus T'$. In conclusion, we have obtained $T \oplus T' \leq R \leq S$, which implies $T \oplus T' \leq S$. $\blacktriangleleft$

One might be concerned about why Theorem 17 only speaks about starting with a *restriction* rather than *degeneration*. This is because the degeneration version is obtained by bootstrapping the result: It is not using any specific property of degeneration, but just regarding it as a restriction over the field extension $\mathbb{F}(\lambda)$.

► **Corollary 18 (Free-lunch speedup, degeneration version).** *Let $T \leq S$ be a degeneration of $\mathbb{F}$ -tensors. More specifically, let T_λ be a $\mathbb{F}(\lambda)$ -tensor before taking $\lambda \rightarrow 0$, i.e.,*

$$T + O(\lambda) = T_\lambda = (A_\lambda \otimes B_\lambda \otimes C_\lambda)S$$

is the restriction of $\mathbb{F}(\lambda)$ -tensors. Plug in $T \leq S$ with data A_λ, B_λ and C_λ into Theorem 17, for any choice of C' thereby, let the resulting $\mathbb{F}(\lambda)$ tensor be T'_λ , T' be the lowest order term of T'_λ (0 if T'_λ is completely zero). Then

$$T \oplus T' \leq S.$$

Proof. Firstly, by Theorem 17, we obtain a degeneration of $\mathbb{F}(\lambda)$ -tensors

$$T_\lambda \oplus T'_\lambda \leq S,$$

say this degeneration is parametrized by another symbol ε . We may scale by λ and ε , to assume that the degeneration is given by $A_{\lambda,\varepsilon}, B_{\lambda,\varepsilon}$ and $C_{\lambda,\varepsilon}$, and $T'_\lambda = T' + O(\lambda)$, such that

$$(T \oplus T') + O(\lambda) + O_\lambda(\varepsilon) = (A_{\lambda,\varepsilon} \otimes B_{\lambda,\varepsilon} \otimes C_{\lambda,\varepsilon})S.$$

The notation means that $O(\lambda)$ term is purely a λ -high order term tensor over $\mathbb{F}$, but $O_\lambda(\varepsilon)$ term is a ε -high order term tensor that can also contain λ . One can again reduce to the

regular definition of degeneration by substituting ε by a large enough power $\varepsilon = \lambda^k$, then the whole restriction again becomes

$$(T \oplus T') + O(\lambda) = (A_{\lambda, \lambda^k} \otimes B_{\lambda, \lambda^k} \otimes C_{\lambda, \lambda^k})S.$$

This concludes the degeneration. ◀

5.2.1 Isolated summand

We now introduce the notion of an *isolated* summand, a property defined by Coppersmith and Winograd [22, Definition 3.1] that is crucial for iterating speedup theorems.

Let $T \oplus T' \leq S$ be a degeneration defined by the polynomial maps A_λ, B_λ , and C_λ . Let $R_\lambda = (A_\lambda \otimes B_\lambda \otimes C_\lambda)S$ denote the $\mathbb{F}(\lambda)$ -tensor prior to taking the limit $\lambda \rightarrow 0$. By definition, we must have $(R_\lambda)_{111} = T + O(\lambda)$ and $(R_\lambda)_{222} = T' + O(\lambda)$, while all other blocks $(R_\lambda)_{ijk}$ (where indices are not all 1 or all 2) must satisfy $(R_\lambda)_{ijk} = O(\lambda)$. We say that the direct summand T' is an *isolated summand* in the third mode with respect to the degeneration $(A_\lambda, B_\lambda, C_\lambda)$ if $(R_\lambda)_{ij2}$ is identically zero for all $(i, j) \neq (2, 2)$.

One can verify that this definition essentially coincides with Coppersmith and Winograd's original formulation when $T' = \langle 1, t, 1 \rangle$. Moreover, the following facts are readily established.

► **Proposition 19.** *The direct summand T' of $T \oplus T' \leq S$ produced by Theorem 17 or Corollary 18 is isolated in the third mode.*

The following proposition generalizes [22, Proposition 4.1]:

► **Proposition 20.** *Let T'_1, T'_2 be isolated direct summands of degenerations*

$$T_1 \oplus T'_1 \leq S_1, \quad T_2 \oplus T'_2 \leq S_2.$$

Then, in the tensor product of these degenerations

$$(T_1 \otimes T_2) \oplus (T_1 \otimes T'_2) \oplus (T'_1 \otimes T_2) \oplus (T'_1 \otimes T'_2) \leq S_1 \otimes S_2,$$

the term $T'_1 \otimes T'_2$ remains an isolated summand in the third mode.

5.3 Speedup by one slice

In this subsection, we recover the classical speedup theorem of Coppersmith and Winograd, as well as Strassen's generalization, in the case where the tensors appended to both sides are one-slice matrix multiplication tensors. We also refine the notion of *fullness* originally introduced by Coppersmith and Winograd [22, Definition 3.1].

Usually, one cannot hope to directly use the free-lunch speedup theorem, because the first step requires choosing a subspace $C' \subseteq C^\perp$. However, for a reasonable degeneration $T \leq S$, the corresponding space $C^\perp$ is typically zero. Indeed, if $C^\perp$ were non-zero, then viewing S as an “upper bound” for T , the tensor S would contain redundant W -slices; removing these slices would essentially yield a more effective upper bound for T .

Therefore, the standard speedup strategy proceeds as follows: First, artificially add a tensor S' to the right-hand side. This new summand S' contributes nothing to the degeneration $T \leq S \oplus S'$ because its component in the C -map is set to zero. However, its components in the A and B maps can be carefully chosen to satisfy the conditions of Theorem 17 (or Corollary 18), enabling the extraction of a non-trivial summand T' .

We first examine the implications of the free-lunch speedup theorem when C' is a one-dimensional subspace.

► **Proposition 21.** *Assume the setting of Theorem 17. Let $C' \subseteq C^\perp$ be a one-dimensional subspace spanned by a linear function $f: W \rightarrow \mathbb{F}$. Let $r = R((\text{id}_U \otimes \text{id}_V \otimes f)S)$ be the rank of the resulting matrix in $U \otimes V$. Then we have $T' \cong \langle 1, t, 1 \rangle$, where*

$$t \geq r - n - m,$$

with $n = \dim U'$ and $m = \dim V'$.

Proof. The restriction $T' = (A' \otimes B' \otimes f)S$ can be viewed as first contracting the W -mode: $M = (\text{id}_U \otimes \text{id}_V \otimes f)S$ yields a matrix of rank r . The subspace A' is defined as the kernel of the map $u \mapsto (u \otimes B \otimes f)S$. Since this map maps to V' , its kernel A' has codimension at most $\dim V' = m$. Thus, restricting M to A' reduces the rank by at most m . Similarly, restricting to B' reduces the rank by at most n . Consequently, $T' = (A' \otimes B')M$ has rank at least $t \geq r - n - m$. A rank- t matrix (viewed as a tensor) is isomorphic to $\langle 1, t, 1 \rangle$. ◀

We now explain how to append a one-slice tensor to both sides of the degeneration.

► **Proposition 22** (One-slice speedup [41, Lemma 3.12]). *Let $T \leq S$ be a restriction defined by maps A, B, C . For a linear function $f: W \rightarrow \mathbb{F}$, let $q = R((\text{id}_U \otimes \text{id}_V \otimes f)S)$ be the rank of the corresponding matrix in $U \otimes V$, and let $s = R((A \otimes B \otimes f)S)$ be the rank of the matrix in $U' \otimes V'$. Then there exists a restriction $T \leq S \oplus \langle 1, s, 1 \rangle$ extending A, B, C , such that there exists a linear function $f': (W \oplus \mathbb{F}) \rightarrow \mathbb{F}$ (where $W \oplus \mathbb{F}$ corresponds to the third mode of $S \oplus \langle 1, s, 1 \rangle$) satisfying the conditions of Proposition 21 with $r = q + s$.*

Proof. Since $M = (A \otimes B \otimes f)S$ has rank s , there exist linear maps $A': \mathbb{F}^s \rightarrow U'$ and $B': \mathbb{F}^s \rightarrow V'$ such that $M = (A' \otimes B')\langle 1, s, 1 \rangle$. It is then straightforward to verify that we can define the new restriction as

$$T = [(A \ A') \otimes (B \ B') \otimes (C \ 0)](S \oplus \langle 1, s, 1 \rangle),$$

with the new linear function $f' = (f \ -1)$. This function lies in the appropriate kernel, since

$$[(A \ A') \otimes (B \ B') \otimes (f \ -1)](S \oplus \langle 1, s, 1 \rangle) = M - (A' \otimes B')\langle 1, s, 1 \rangle = 0,$$

and $(\text{id} \otimes \text{id} \otimes f')(S \oplus \langle 1, s, 1 \rangle)$ is the direct sum of the matrices $(\text{id}_U \otimes \text{id}_V \otimes f)S$ and $\langle 1, s, 1 \rangle$, which has rank $r = q + s$. ◀

Finally, we generalize Coppersmith and Winograd's notion of *fullness*. Given a degeneration $T \oplus \langle 1, t, 1 \rangle \leq S$, let $f: W \rightarrow \mathbb{F}(\lambda)$ be the component of the C -map projecting to the slice $\langle 1, t, 1 \rangle$. We define the *fullness index* of $\langle 1, t, 1 \rangle$ (or f) as $q = R((\text{id}_U \otimes \text{id}_V \otimes f)S)$. If $q = \dim U = \dim V$, we say the slice is *full*.

Like the isolated property, the fullness index behaves well under speedup theorems and tensor products.

► **Proposition 23.** *The slice $\langle 1, t, 1 \rangle$ constructed in Proposition 21 has fullness index r .*

► **Proposition 24.** *Consider two degenerations*

$$T_1 \oplus \langle 1, t_1, 1 \rangle \leq S_1, \quad T_2 \oplus \langle 1, t_2, 1 \rangle \leq S_2,$$

wherein the summands $\langle 1, t_1, 1 \rangle$ and $\langle 1, t_2, 1 \rangle$ have fullness indices q_1 and q_2 , respectively. Then the direct summand $\langle 1, t_1 t_2, 1 \rangle$ in the tensor product of these two degenerations has fullness index $q_1 q_2$. In particular, if f_1 and f_2 are full, then $f_1 \otimes f_2$ is full.

5.4 Direct sum and compression of matrix multiplication

In the proof of Proposition 21, we used the following basic fact: for a one-slice matrix multiplication tensor $\langle 1, r, 1 \rangle$, restricting its first mode to a codimension- p subspace yields a tensor isomorphic to $\langle 1, r - p, 1 \rangle$. Although this observation looks elementary, it can be viewed as the simplest instance of a much more general phenomenon.

Strassen proved the following result (see [41, Proposition 6.4])³: if T is a direct sum of matrix multiplication tensors and $A \subseteq U^\vee$ is a subspace of small codimension, then the partially restricted tensor $(A \otimes \text{id}_V \otimes \text{id}_W)T$ still contains, as a restriction, a “valuable” tensor T' that is again a direct sum of matrix multiplication tensors. Informally, matrix multiplication is robust under losing a small amount of information in one mode. This is often referred to as the *compression* property of matrix multiplication tensors; see [47, Part III] for an exposition and generalizations. Algorithmic versions have also been developed recently [27, 28].

For our purposes, we only need a compression statement in the special case where all summands are one-slice tensors. This case is simple enough that we include a short proof for the reader’s convenience. We also note that the concrete applications in this paper do not require the full generality of this theorem: the tensors and linear maps we use enjoy additional structure that simplifies the analysis. Nonetheless, we expect that the compression statement in this general form may be useful for more complicated degenerations in the future, for instance when the relevant linear maps are found via computer search.

► **Proposition 25.** *Let $T = \bigoplus_{i=1}^k \langle 1, n_i, 1 \rangle$ be a direct sum of one-slice matrix multiplication tensors, and let $A \subseteq U^\vee$ be a subspace of codimension p . Then there exists a tensor*

$$T' \leq (A \otimes \text{id}_V \otimes \text{id}_W)T$$

that is still a direct sum of one-slice matrix multiplication tensors:

$$T' \cong \bigoplus_{i=1}^k \langle 1, n_i - p_i, 1 \rangle,$$

where the integers p_i satisfy $0 \leq p_i \leq n_i$ and $\sum_{i=1}^k p_i = p$.

Proof. Let $(x_j^{(i)})_{1 \leq j \leq n_i}$ and $(y_j^{(i)})_{1 \leq j \leq n_i}$ denote the bases of the U - and V -modes of T corresponding to the decomposition $T = \bigoplus_{i=1}^k \langle 1, n_i, 1 \rangle$. Identify A with a surjective linear map $A : U \rightarrow U'$ where $\dim U' = \dim U - p$.

Consider the set of vectors $\{A(x_j^{(i)})\} \subseteq U'$. Choose a subset $D \subseteq \{x_j^{(i)}\}$ of size $\dim U'$ such that $\{A(x) : x \in D\}$ is linearly independent (hence a basis of U'). Now restrict T by zeroing out every basis element $x_j^{(i)}$ (and the corresponding $y_j^{(i)}$) with $x_j^{(i)} \notin D$. This reduces the i th summand $\langle 1, n_i, 1 \rangle$ to $\langle 1, n_i - p_i, 1 \rangle$, where p_i is the number of removed basis elements from the i th block. By construction, $\sum_i p_i = p$.

Finally, the restriction of A to $\text{span } D$ is an isomorphism onto U' , so after the above zeroing-out operation the resulting tensor is indeed a restriction of $(A \otimes \text{id}_V \otimes \text{id}_W)T$ and has the claimed direct-sum form. ◀

³ Strassen’s original formulation and subsequent expositions (e.g., [41, Proposition 6.4] and [47, Part III]) often present a closely related variant in terms of an *additive* perturbation in the same tensor space: if $S = T + U$ where U has small flattening rank in the U -mode, then one can find a “valuable” restriction $T' \leq S$ that remains a direct sum of matrix multiplication tensors. The formulation we use here (starting from a direct sum and restricting one mode by a small-codimension subspace) is equivalent to that variant via the substitution method; we include this footnote to reconcile the two viewpoints.

6 Instantiations of speedup theorems

The first method is the most naive version of the speedup theorem, which generalizes [22, Corollary 3.1].

► **Theorem 26** (One-slice speedup for nonminimal border rank). *Let T be an $n \times n \times n$ tensor over $\mathbb{F}$, and suppose that $T \leq \langle r \rangle$. Let $\{a_i\}, \{b_i\}, \{c_i\}$ be the $\mathbb{F}(\lambda)$ -vectors representing the border rank decomposition, i.e.,*

$$\sum_{i=1}^r a_i b_i c_i = T + O(\lambda).$$

If there exist nonzero scalars $c'_i \in \mathbb{F}(\lambda) \setminus \{0\}$ such that the matrix

$$M = \sum_{i=1}^r a_i b_i c'_i \in U' \otimes V'$$

has rank at most s , then there exists a degeneration

$$T \oplus \langle 1, r + s - 2n, 1 \rangle \leq \langle r \rangle \oplus \langle 1, s, 1 \rangle.$$

When $r > 2n$, this yields the improved asymptotic rank bound

$$\underline{R}(T) \leq r + s^{2/3} - (r + s - 2n)^{2/3} < r.$$

In particular, if $r \geq n$, we have

$$T \oplus \langle 1, r - n, 1 \rangle \leq \langle r \rangle \oplus \langle 1, n, 1 \rangle,$$

and for $r > 2n$, this provides the strict improvement

$$\underline{R}(T) \leq r + n^{2/3} - (r - n)^{2/3} < r.$$

Proof. The hypothesis exactly matches the condition of Proposition 22. Applying that result, all the degeneration parameters align:

$$T \oplus \langle 1, r + s - 2n, 1 \rangle \leq \langle r \rangle \oplus \langle 1, s, 1 \rangle.$$

We refer to the setting of the three-direction speedup inequalities Equation (4.4)–Equation (4.6) with parameters $t = r + s - 2n$ and the same s . Since $r > 2n$, we have $t > s$, and thus Equation (4.4)–Equation (4.6) imply the asymptotic rank bound

$$\underline{R}(T) \leq r + s^{2/3} - t^{2/3} = r + s^{2/3} - (r + s - 2n)^{2/3}. \quad \blacktriangleleft$$

We can assess the efficacy of this method by the order of magnitude of the improvement. For $r \geq 3n$, the bound simplifies to

$$\underline{R}(T) \leq r - \Omega(r^{2/3}).$$

Although this method appears to yield an improvement only when $r > 2n$, we can bootstrap this result to show that any non-minimal border rank upper bound for T is not tight for its asymptotic rank.

► **Corollary 27.** *Let T be an $n \times n \times n$ tensor with $\underline{R}(T) \leq r$. If $r > n$, then $\underline{R}(T) < r$.*

Proof. Since $T \leq \langle r \rangle$, taking tensor powers yields $T^{\otimes k} \leq \langle r^k \rangle$. By choosing k sufficiently large such that $r^k > 2n^k$, Theorem 26 implies $\mathbb{R}(T)^k = \mathbb{R}(T^{\otimes k}) < r^k$. ◀

The second method demonstrates that the identity obtained via the first method can essentially always be *iterated* to yield a superior identity. This generalizes [22, Theorem B].

► **Theorem 28.** *Under the setting of Theorem 26, let $t = r + s - 2n$. There exists a degeneration*

$$T^{\otimes 2} \oplus 2 \odot T \otimes \langle 1, t, 1 \rangle \oplus \langle 1, t^2 + 2n^2, 1 \rangle \leq \langle r^2 \rangle \oplus 2r \odot \langle 1, s, 1 \rangle \oplus \langle 1, s^2, 1 \rangle.$$

This effectively replaces the summand $\langle 1, t^2, 1 \rangle$ in the tensor square of the result from Theorem 26 with $\langle 1, t^2 + 2n^2, 1 \rangle$.

Proof. Since the summand $\langle 1, t, 1 \rangle$ appearing in the degeneration

$$T \oplus \langle 1, t, 1 \rangle \leq \langle r \rangle \oplus \langle 1, s, 1 \rangle$$

is produced via the one-slice speedup theorem (Theorem 26), Proposition 19 and Proposition 23 guarantee that this summand is full and isolated w.r.t. the degeneration. Upon taking the tensor square, Proposition 20 and Proposition 24 guarantee that the summand $\langle 1, t^2, 1 \rangle$ in the degeneration

$$T^{\otimes 2} \oplus 2 \odot T \otimes \langle 1, t, 1 \rangle \oplus \langle 1, t^2, 1 \rangle \leq \langle r^2 \rangle \oplus 2r \odot \langle 1, s, 1 \rangle \oplus \langle 1, s^2, 1 \rangle$$

remains full and isolated. Let this summand be projected by a linear functional f in the third mode. First, we zero out the variables of this summand in all three directions, so that for the remaining degeneration

$$T^{\otimes 2} \oplus 2 \odot T \otimes \langle 1, t, 1 \rangle \leq \langle r^2 \rangle \oplus 2r \odot \langle 1, s, 1 \rangle \oplus \langle 1, s^2, 1 \rangle,$$

f satisfies the condition of Proposition 21, as f is isolated. The parameter r corresponds to the fullness of f , which in our case (since f is full) is the dimension of the right-hand side, i.e., $(r + s)^2$. Thus the replaced direct summand

$$T^{\otimes 2} \oplus 2 \odot T \otimes \langle 1, t, 1 \rangle \oplus \langle 1, t', 1 \rangle \leq \langle r^2 \rangle \oplus 2r \odot \langle 1, s, 1 \rangle \oplus \langle 1, s^2, 1 \rangle$$

has

$$t' = (r + s)^2 - 2(n^2 + 2nt) = t^2 + 2n^2. \quad \blacktriangleleft$$

Naturally, this iterative process can be continued indefinitely, but its improvement on the asymptotic rank would be marginal. Also, after some analysis, it turns out that Theorem 28 still gives an asymptotic rank bound of type $r - \Omega(r^{2/3})$ when $r \geq 3n$.

The third method involves appending not merely a single one-slice tensor to both sides, but a direct sum of such tensors.

► **Theorem 29.** *Let $T \leq \langle r \rangle$, and let $\{a_i\}, \{b_i\}, \{c_i\}$ be $\mathbb{F}(\lambda)$ -vectors representing the degeneration. Suppose there exist nonzero scalars c'_i and a partition of $[r]$ into p groups $[r] = I_1 \sqcup \dots \sqcup I_p$. For each $1 \leq \alpha \leq p$, let $r_\alpha = |I_\alpha|$, let $M_\alpha = \sum_{i \in I_\alpha} a_i b_i c'_i$, and let $s_\alpha \geq \mathbb{R}(M_\alpha)$. If $r_\alpha + s_\alpha \geq 2n$ for all α , then*

$$T \oplus \bigoplus_{\alpha=1}^p \langle 1, r_\alpha + s_\alpha - 2n, 1 \rangle \leq \langle r \rangle \oplus \bigoplus_{\alpha=1}^p \langle 1, s_\alpha, 1 \rangle.$$

We provide two proofs of this result. They are essentially equivalent; the first is simpler, while the second employs the concept of the compression theorem and generalizes to cases where the added terms are not direct sums of one-slice tensors (Theorem 42) or where the starting upper bound S is not a diagonal tensor (Section 7.2).

First proof. In the spirit of Corollary 18, it suffices to prove the restriction case, i.e., working over $\mathbb{F}$ instead of $\mathbb{F}(\lambda)$. In that setting, we can write T as a sum of tensors $T = \sum_{\alpha=1}^p T_\alpha$, where $T_\alpha = \sum_{i \in I_\alpha} a_i b_i c_i$. Then $R(T_\alpha) \leq r_\alpha$. We can apply Theorem 26 to each T_α individually and sum the results:

$$\bigoplus_{\alpha=1}^p (T_\alpha \oplus \langle 1, r_\alpha + s_\alpha - 2n, 1 \rangle) \leq \langle r \rangle \oplus \bigoplus_{\alpha=1}^p \langle 1, s_\alpha, 1 \rangle.$$

The conclusion follows from $T \leq \bigoplus_{\alpha} T_\alpha$. ◀

Second proof. As in the first proof, we focus on the restriction case. We begin with the redundant restriction

$$T \leq \langle r \rangle \oplus \bigoplus_{\alpha=1}^p \langle 1, s_\alpha, 1 \rangle.$$

Write

$$S := \langle r \rangle \oplus \bigoplus_{\alpha=1}^p \langle 1, s_\alpha, 1 \rangle = \sum_{i=1}^r u_i v_i w_i - \sum_{\alpha=1}^p \sum_{j=1}^{s_\alpha} u'_{\alpha j} v'_{\alpha j} w'_\alpha,$$

where the minus sign is introduced to simplify subsequent cancellations. Since M_α has rank at most s_α , there exist linear maps A_α, B_α such that $(A_\alpha \otimes B_\alpha) \langle 1, s_\alpha, 1 \rangle = M_\alpha$. The restriction is defined by the linear maps

$$\begin{aligned} A : \quad & a_i \mapsto u_i, \quad A_\alpha(u'_{\alpha j}) \mapsto u'_{\alpha j}, \\ B : \quad & b_i \mapsto v_i, \quad B_\alpha(v'_{\alpha j}) \mapsto v'_{\alpha j}, \\ C : \quad & c_i \mapsto w_i, \quad 0 \mapsto w'_\alpha. \end{aligned}$$

To apply the free-lunch speedup theorem (Theorem 17), define a subspace $C' \subseteq W^\vee$ via the linear map defined as follows, where $\alpha(i)$ is the index such that $i \in I_\alpha$:

$$C' : \quad c'_i w'_{\alpha(i)} \mapsto w_i, \quad w'_\alpha \mapsto w'_\alpha.$$

A direct computation shows that $(A \otimes B \otimes C')S = 0$.

Applying Theorem 17 yields a direct summand $T' = (A' \otimes B' \otimes C')S$ on the left-hand side. Let us first unwind the form of $(\text{id} \otimes \text{id} \otimes C')S$:

$$\begin{aligned} (\text{id} \otimes \text{id} \otimes C')S &= \sum_{i=1}^r u_i v_i \cdot c'_i w'_{\alpha(i)} - \sum_{\alpha=1}^p \sum_{j=1}^{s_\alpha} u'_{\alpha j} v'_{\alpha j} w'_\alpha \\ &= \sum_{\alpha=1}^p \left(\sum_{i \in I_\alpha} c'_i u_i v_i - \sum_{j=1}^{s_\alpha} u'_{\alpha j} v'_{\alpha j} \right) w'_\alpha, \end{aligned}$$

which is evidently $\bigoplus_{\alpha=1}^p \langle 1, r_\alpha + s_\alpha, 1 \rangle$. Recall that $A' \subseteq U^\vee$, the space of linear functionals, is given by those $f \in U^\vee$ such that the following expression vanishes:

$$(f \otimes B \otimes C')S = \sum_{\alpha=1}^p \left(\sum_{i \in I_\alpha} c'_i f(u_i) b_i - \sum_{j=1}^{s_\alpha} f(u'_{\alpha j}) B_\alpha(v'_{\alpha j}) \right) w'_\alpha$$

The image, viewed as a vector in an np -dimensional space, defines np linear equations on f , meaning A' is a subspace of codimension $\leq np$. Similarly, B' restricts to a subspace of codimension $\leq np$. Proposition 25 applies, which allows us to reduce the direct sum of matrix multiplication tensors by $2np$ terms. Notice that each equation associated with a component in w'_α only concerns the values of f on $\{u_i\}_{i \in I_\alpha}$ and $u'_{\alpha j}$. Thus, we do not need the full generality of Proposition 25, but can instead perform compression on each direct summand separately. Consequently, T' can be compressed to $T' \geq \bigoplus_\alpha \langle 1, r_\alpha + s_\alpha - 2n, 1 \rangle$, proving the claim. $\blacktriangleleft$

We now discuss the quantitative advantage of this method compared with previous ones. Using the naive asymptotic assumption $s \leq n$, one can take $p = \Omega(r/n)$ and let every $r_\alpha \geq 3n$. This gives the degeneration

$$T \oplus p \odot \langle 1, 2n, 1 \rangle \leq \langle r \rangle \oplus p \odot \langle 1, n, 1 \rangle,$$

and similar statements hold for the other two directions. Strassen calculus then yields an upper bound $\mathbb{R}(T) \leq r - (2^{2/3} - 1)pn^{2/3} = r - \Omega(r/n^{1/3})$, which is a superior bound when $r = \omega(n)$.

► **Remark 30.** An intriguing question is how far we can improve the order of magnitude in the speedup theorem. For example, can we prove that any border rank upper bound $\mathbb{R}(T) \leq r$ with $r > n^{1.0001}$ implies $\mathbb{R}(T) \leq 0.9999r$? Any mild strengthening of this question would have surprising consequences in fine-grained complexity theory: the *balanced partition tensor* T_n proposed by Pratt [34] is a family of tensors for which the currently known border rank upper bound is $r_n \approx N^{1.08}$ [25, Proposition 6.18], where N is the dimension of T_n . If one could improve the asymptotic rank of T_n to $o(r_n/\log r_n)$, then Pratt showed that this would refute the set cover conjecture.

7 Applications

7.1 Small CW tensors

Recall that the small Coppersmith–Winograd (CW) tensor cw_q has $q + 1$ variables in each mode. It is commonly written as

$$\text{cw}_q = \sum_{i=1}^q x_0 y_i z_i + x_i y_0 z_i + x_i y_i z_0.$$

The following identity, originally due to Coppersmith–Winograd, gives a border-rank upper bound of $q + 2$.

$$\begin{aligned} & \sum_{i=1}^q \lambda^{-2} (x_0 + \lambda x_i)(y_0 + \lambda y_i)(z_0 + \lambda z_i) \\ & - \lambda^{-3} \left(x_0 + \lambda^2 \sum_{i=1}^q x_i \right) \left(y_0 + \lambda^2 \sum_{i=1}^q y_i \right) \left(z_0 + \lambda^2 \sum_{i=1}^q z_i \right) \\ & + (\lambda^{-3} - q\lambda^{-2}) x_0 y_0 z_0 \\ & = \text{cw}_q + O(\lambda). \end{aligned}$$

It is known that this border-rank bound is tight, i.e. $\underline{R}(\text{cw}_q) = q + 2$. Moreover, Conner, Gesmundo, Landsberg, and Ventura [20, Theorem 1.4] proved that the border rank does not drop for small powers: when $q > 2$ and $n = 2$, one has $\underline{R}(\text{cw}_q^{\otimes n}) = (q + 2)^n$, and likewise when $q > 4$ and $n = 3$. We show that the asymptotic rank of cw_q is strictly smaller than $q + 2$, and hence this phenomenon must eventually fail for larger n .

We begin by finding low-rank matrices in order to apply Theorem 26.

► **Lemma 31.** *Let a_i, b_i, c_i be the data of the border-rank upper bound $\text{cw}_q \leq \langle q + 2 \rangle$ given above. Then there exist nonzero scalars c'_i such that the matrix $M = \sum_{i=1}^{q+2} c'_i \cdot a_i b_i \in U' \otimes V'$ has rank q .*

Proof. We first write the linear combination as

$$M = \sum_{i=1}^q c'_i (x_0 + \lambda x_i)(y_0 + \lambda y_i) + c'_{q+1} \left(x_0 + \lambda^2 \sum_{i=1}^q x_i \right) \left(y_0 + \lambda^2 \sum_{i=1}^q y_i \right) + c'_0 x_0 y_0.$$

For notational convenience, we apply the change of basis $x'_i := x_0 + \lambda x_i$ for $i > 0$ (and similarly for y). Then

$$M = c'_0 x_0 y_0 + \sum_{i=1}^q c'_i x'_i y'_i + c'_{q+1} \left((1 - q\lambda)x_0 + \lambda \sum_{i=1}^q x'_i \right) \left((1 - q\lambda)y_0 + \lambda \sum_{i=1}^q y'_i \right).$$

Assume $c'_i = 1$ for all $i \leq q$. We choose c'_{q+1} so that M becomes singular. Writing M in the form $M = I + c'_{q+1} ab^\top$, we have

$$\det(M) = \det(I + c'_{q+1} ab^\top) = 1 + c'_{q+1} b^\top a,$$

and a direct calculation gives $b^\top a = (1 - q\lambda)^2 + q\lambda^2 \neq 0$. Taking $c'_{q+1} = -1/(b^\top a)$ yields $\underline{R}(M) < q + 1$. On the other hand,

$$\underline{R}(M) = \underline{R}(I + c'_{q+1} ab^\top) \geq \underline{R}(I) - \underline{R}(c'_{q+1} ab^\top) = q.$$

Therefore, this choice produces a rank- q matrix. ◀

► **Proposition 32.** *For any $q \geq 2$ and $n \geq 1$, we have the tensor degeneration*

$$\text{cw}_q^{\otimes n} \oplus \langle 1, (q + 2)^n - 2(q + 1)^n + q^n, 1 \rangle \leq \langle (q + 2)^n \rangle \oplus \langle 1, q^n, 1 \rangle.$$

Proof. Let f be the linear functional on the W -mode given by the previous lemma. Then $M = (A \otimes B \otimes f) \langle q + 2 \rangle$ satisfies $\underline{R}(M) = q$. Tensoring these maps gives a degeneration $\text{cw}_q^{\otimes n} \leq \langle (q + 2)^n \rangle$, and $f^{\otimes n}$ is a linear functional such that

$$M^{\otimes n} = (A^{\otimes n} \otimes B^{\otimes n} \otimes f^{\otimes n}) \langle (q + 2)^n \rangle$$

has rank q^n . Applying Theorem 26 yields the desired degeneration. ◀

Combining the above degeneration with Proposition 15 immediately yields the following asymptotic-rank bound.

► **Corollary 33.** *For any $q \geq 2$ and $n \geq 1$, define*

$$\gamma_{q,n} := \left((q + 2)^n + q^{2n/3} - ((q + 2)^n - 2(q + 1)^n + q^n)^{2/3} \right)^{1/n}$$

and

$$\gamma_q^{(33)} := \min_{n \geq 1} \gamma_{q,n}.$$

Then the small CW tensor satisfies

$$\underline{R}(\text{cw}_q) \leq \gamma_q^{(33)} < q + 2.$$

We let $n^{(33)}$ denote the n that reaches the minimum in Corollary 33.

We can further apply the second method to iteratively speed up the degeneration. By Theorem 28, we obtain

$$\begin{aligned} & \text{cw}_q^{\otimes 2n} \oplus 2 \odot \text{cw}_q^{\otimes n} \otimes \langle 1, (q+2)^n - 2(q+1)^n + q^n, 1 \rangle \\ & \oplus \langle 1, [(q+2)^n - 2(q+1)^n + q^n]^2 + 2(q+1)^{2n}, 1 \rangle \\ & \leq (\langle (q+2)^n \rangle \oplus \langle 1, q^n, 1 \rangle)^{\otimes 2}. \end{aligned}$$

Strassen calculus then gives the upper bound

$$\underline{R}(\text{cw}_q) \leq \max_{\theta \in [2/3, 1]} F(\theta)$$

where

$$F(\theta) := \sqrt{(r + s^\theta)^2 - (t^2 + 2(q+1)^{2n})^\theta + t^{2\theta} - t^\theta}$$

with $r = (q+2)^n$, $s = q^n$, and $t = r + s - 2(q+1)^n$. We denote the value obtained by taking $n = n^{(33)}$ by $\gamma_q'^{(33)}$.

The values of $\gamma_q^{(33)}$, $n^{(33)}$ and $\gamma_q'^{(33)}$ for the first few values of q are shown in Table 1 (computed via computer-aided evaluation; in fact, the maximizing θ in the definition of $\gamma_q'^{(33)}$ equals $\theta = 2/3$).

■ **Table 1** Asymptotic rank upper bound for small CW tensors.

q	$\underline{R}(\text{cw}_q)$	$\gamma_q^{(33)}$	$n^{(33)}$	$\gamma_q'^{(33)}$
2	4	3.933484	4	3.930872
3	5	4.970152	4	4.967092
4	6	5.986817	5	5.986293
5	7	6.994597	6	6.994516
6	8	7.997665	6	7.997596
7	9	8.998978	7	8.998969
8	10	9.999595	8	9.999593
9	11	10.999850	9	10.999850
10	12	11.999929	9	11.999929

Note that this degeneration can be iterated indefinitely. We do not pursue further iterations here, as the resulting numerical improvements are marginal.

We also remark that the following variant of the cw_2 tensor, which first appears in [23, §11], is equally useful from the perspective of the laser method:

$$\text{cw}'_2 := x_1 y_2 z_3 + x_1 y_3 z_2 + x_2 y_1 z_3 + x_2 y_3 z_1 + x_3 y_1 z_2 + x_3 y_2 z_1,$$

In particular, if one could prove $\mathbb{R}(\text{cw}'_2) = 3$, then it would follow that $\omega = 2$. Currently, the best known upper bound is $\mathbb{R}(\text{cw}'_2) = 4$ (over fields of characteristic $\neq 2$), via

$$\begin{aligned} \text{cw}'_2 = & \frac{1}{4}[(x_1 + x_2 + x_3)(y_1 + y_2 + y_3)(z_1 + z_2 + z_3) \\ & + (-x_1 + x_2 - x_3)(-y_1 + y_2 - y_3)(-z_1 + z_2 - z_3) \\ & + (x_1 - x_2 - x_3)(y_1 - y_2 - y_3)(z_1 - z_2 - z_3) \\ & + (-x_1 - x_2 + x_3)(-y_1 - y_2 + y_3)(-z_1 - z_2 + z_3)]. \end{aligned}$$

Let the four rank-one summands in the decomposition above be $a_i b_i c_i$. Then

$$a_1 b_1 + a_2 b_2 - a_3 b_3 - a_4 b_4 = 4(x_1 y_3 + x_3 y_1)$$

has rank 2, matching the same parameter regime as in Lemma 31. Therefore, the preceding analysis applies to cw'_2 as well.

► **Corollary 34.** *If $\mathbb{F}$ has characteristic $\neq 2$, then we also have $\mathbb{R}(\text{cw}'_2) \leq \gamma_2'^{(33)}$.*

► **Remark 35.** If the asymptotic rank of the small CW tensor were optimal, i.e., $\mathbb{R}(\text{cw}_q) = q+1$, then applying the laser method directly to cw_q would in fact yield a better upper bound on ω than the standard analysis based on the big CW tensors CW_q . Since we improve the best known upper bounds on $\mathbb{R}(\text{cw}_q)$, it is natural to ask whether these improvements already translate into a better bound on ω .

Starting from the degeneration above, the most straightforward way to bound ω combines the laser method with our new asymptotic rank bounds. Interestingly, we remark that this incurs a “composition loss”: the asymptotic rank bound transfers some matrix multiplication structure into diagonal tensors, which the laser method cannot exploit. A more faithful approach would start from the degeneration used in our proof and combine Strassen calculus with the laser method at that level, and this indeed achieves a slightly better bound than the straightforward approach.

Nonetheless, our experiments indicate that, with the identities currently produced by our speedup theorems, this hybrid approach still performs worse than Coppersmith–Winograd’s analysis of the laser method applied to the first power $\text{CW}_q^{\otimes 1}$.

7.2 General tensors

A nontrivial general upper bound on the asymptotic rank of tensors, due to Strassen (implicit in [41, Proposition 3.6]), states that every $d \times d \times d$ tensor satisfies $\mathbb{R}(T) \leq d^{\frac{2}{3}\omega}$. In this subsection, we show how to slightly improve this bound.

To explain our improvement, we first recall Strassen’s $\frac{2}{3}\omega$ bound. The following straightforward rewriting shows the restriction $T \leq \langle d, 1, d \rangle$:

$$T = \sum_{i,j,k=1}^d a_{ijk} x_i y_j z_k = \sum_{i,j=1}^d x_i y_j \left(\sum_{k=1}^d a_{ijk} z_k \right) \leq \sum_{i,j,k=1}^d x_i y_j w_{ij} = \langle d, 1, d \rangle.$$

Applying the same argument to the other modes yields $T \leq \langle d, d, 1 \rangle$ and $T \leq \langle 1, d, d \rangle$. Tensoring these three restrictions gives $T^{\otimes 3} \leq \langle d^2, d^2, d^2 \rangle$, and hence $\mathbb{R}(T)^3 \leq d^{2\omega}$.

We first present a one-slice speedup of the above argument.

► **Theorem 36.** *When $d \geq 3$, any $d \times d \times d$ tensor T satisfies*

$$\underline{R}(T) \leq \sqrt{d^{\frac{4}{3}\omega} + d^{\frac{2}{3}\omega} - (d^3 - d^2)^{\frac{\omega}{3}}} < d^{\frac{2}{3}\omega}.$$

Proof. We start with the restriction $T^{\otimes 2} \leq \langle d, d^2, d \rangle$.

We wish to apply Proposition 22 to append a slice to both sides of this restriction. To do so, we first need a linear functional $f: W \rightarrow \mathbb{F}$ for $S = \langle d, d^2, d \rangle$ such that $(\text{id} \otimes \text{id} \otimes f)S$ has large rank. Write

$$S = \sum_{i=1}^{d^2} \sum_{j=1}^d \sum_{k=1}^{d^2} x_{ij} y_{jk} z_{ki}.$$

Let f be defined by $f(z_{ki}) = \llbracket i = k \rrbracket$. Then

$$(\text{id} \otimes \text{id} \otimes f)S = \sum_{i=1}^{d^2} \sum_{j=1}^d x_{ij} y_{ji}$$

is diagonal, hence has rank $s = d^3$. Applying Proposition 22 and Proposition 21, we obtain the degeneration

$$T^{\otimes 2} \oplus \langle 1, d^3 - d^2, 1 \rangle \leq \langle d, d^2, d \rangle \oplus \langle 1, d^2, 1 \rangle.$$

Passing to the asymptotic spectrum, this means

$$\phi(T)^2 \leq d^{\theta_1 + 2\theta_2 + \theta_3} + d^{2\theta_2} - (d^3 - d^2)^{\theta_2}.$$

Writing $\varpi := \theta_1 + \theta_2 + \theta_3$, by permuting modes we obtain analogous inequalities, we have

$$\phi(T)^2 \leq d^{\varpi + \theta_i} + d^{2\theta_i} - (d^3 - d^2)^{\theta_i}.$$

Since $\theta_1 + \theta_2 + \theta_3 = \varpi$ and $\theta_1, \theta_2, \theta_3 \geq 0$, there exists i such that $\theta_i \leq \varpi/3$. One can show that the function

$$F(\theta) = d^{\varpi + \theta} + d^{2\theta} - (d^3 - d^2)^\theta, \quad \theta \in [0, \varpi/3]$$

is maximized at $\theta = \varpi/3$ (the proof is deferred to Lemma 48). Furthermore, by Lemma 49, the quantity $F(\varpi/3)$ is increasing in ϖ on $[2, \omega]$. Since $\varpi \leq \omega$ for all $(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}$, we obtain

$$\phi(T)^2 \leq d^{\frac{4}{3}\omega} + d^{\frac{2}{3}\omega} - (d^3 - d^2)^{\omega/3}.$$

The claim then follows from Strassen duality. ◀

By appending direct sums rather than a single slice, we obtain a stronger bound.

► **Lemma 37.** *Let $\{d_\alpha\}_{1 \leq \alpha \leq p}$ be positive integers such that $\sum_\alpha d_\alpha \leq d$. Then we have degeneration*

$$T^{\otimes 2} \oplus \bigoplus_{\alpha=1}^p \langle 1, (d_\alpha - 1)d^2, 1 \rangle \leq \langle d, d^2, d \rangle \oplus p \odot \langle 1, d^2, 1 \rangle.$$

Proof. We start from the redundant restriction $T^{\otimes 2} \leq \langle d, d^2, d \rangle \oplus p \odot \langle 1, d^2, 1 \rangle$. Write

$$S := \langle d, d^2, d \rangle \oplus p \odot \langle 1, d^2, 1 \rangle = \sum_{i=1}^d \sum_{j=1}^{d^2} \sum_{k=1}^d u_{ij} v_{jk} w_{ki} - \sum_{\alpha=1}^p \sum_{\ell=1}^{d^2} u'_{\alpha\ell} v'_{\alpha\ell} w'_{\alpha},$$

where the minus sign is introduced to simplify later cancellations. Fix a partition $[\sum_{\alpha} d_{\alpha}] = I_1 \sqcup \cdots \sqcup I_p$ where $|I_{\alpha}| = d_{\alpha}$. Define the linear map C by augmenting the corresponding linear map of $T^{\otimes 2} \leq \langle d, d^2, d \rangle$ by zero on new components, and C' by

$$C' : w_{ki} \mapsto \begin{cases} w'_{\alpha} & i = k \in I_{\alpha}, \\ 0 & \text{otherwise,} \end{cases} \quad w'_{\alpha} \mapsto w'_{\alpha},$$

and choose linear maps A, B so that

$$\begin{aligned} 0 &= (A \otimes B \otimes C')S \\ &= \sum_{\alpha=1}^p \left(\sum_{i \in I_{\alpha}} \sum_{j=1}^{d^2} A(u_{ij}) B(v_{ji}) - \sum_{\ell=1}^{d^2} A(u'_{\alpha\ell}) B(v'_{\alpha\ell}) \right) w'_{\alpha}, \end{aligned}$$

where the action on the u - and v -variables is already determined (since it stems from $T^{\otimes 2} \leq \langle d, d^2, d \rangle$). Since $T^{\otimes 2}$ has dimension d^2 in each mode, we can choose $A(u'_{\alpha\ell})$ and $B(v'_{\alpha\ell})$ so that the terms cancel.

Applying Theorem 17 yields a direct summand $T' = (A' \otimes B' \otimes C')S$ on the left-hand side. We first expand the form of $(\text{id} \otimes \text{id} \otimes C')S$:

$$(\text{id} \otimes \text{id} \otimes C')S = \sum_{\alpha=1}^p \underbrace{\left(\sum_{i \in I_{\alpha}} \sum_{j=1}^{d^2} u_{ij} v_{ji} - \sum_{\ell=1}^{d^2} u'_{\alpha\ell} v'_{\alpha\ell} \right)}_{\langle 1, (d_{\alpha}+1)d^2, 1 \rangle} w'_{\alpha}.$$

Then, as in the second proof of Theorem 29, after applying the contraction $A' \otimes B'$, each summand $\langle 1, (d_{\alpha}+1)d^2, 1 \rangle$ restricts to $\langle 1, (d_{\alpha}-1)d^2, 1 \rangle$. This completes the proof. $\blacktriangleleft$

Compared with Theorem 36, this yields an improved bound in which the subtractive term is asymptotically stronger.

► **Theorem 38.** *When $d \geq 3$, any $d \times d \times d$ tensor T satisfies*

$$\mathbb{R}(T) \leq \sqrt{d^{\frac{4}{3}}\omega - \left\lfloor \frac{d}{3} \right\rfloor \cdot (2^{\frac{\omega}{3}} - 1)d^{\frac{2}{3}\omega}}.$$

Proof. Apply the previous lemma with $p = \lfloor d/3 \rfloor$ and $d_{\alpha} = 3$ for all α . It gives the degeneration

$$T^{\otimes 2} \oplus p \odot \langle 1, 2d^2, 1 \rangle \leq \langle d, d^2, d \rangle \oplus p \odot \langle 1, d^2, 1 \rangle.$$

Fix an arbitrary spectral point $\phi \in \mathcal{X}$, and let $(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}$ be its parameters. Passing to the asymptotic spectrum yields

$$\phi(T)^2 \leq d^{\theta_1+2\theta_2+\theta_3} - (2^{\theta_2} - 1)p d^{2\theta_2}.$$

By permuting tensor modes, we obtain the analogous bounds for $i \in \{1, 2, 3\}$:

$$\phi(T)^2 \leq d^{\varpi+\theta_i} - (2^{\theta_i} - 1)p d^{2\theta_i}, \quad \text{where } \varpi := \theta_1 + \theta_2 + \theta_3.$$

Let $\theta := \min\{\theta_1, \theta_2, \theta_3\}$, so $\theta \leq \varpi/3$. Define

$$h_\varpi(x) := d^{\varpi+x} - (2^x - 1)p d^{2x}.$$

Then $\phi(T)^2 \leq h_\varpi(\theta)$. By Lemma 50, h_ϖ is increasing on $[0, \varpi/3]$, hence

$$\phi(T)^2 \leq h_\varpi(\varpi/3) = d^{4\varpi/3} - (2^{\varpi/3} - 1)p d^{2\varpi/3}.$$

By Lemma 51, the right-hand side is increasing in ϖ on $[2, \omega]$; since $\varpi \leq \omega$ for every $(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}$, we obtain

$$\phi(T)^2 \leq d^{4\omega/3} - (2^{\omega/3} - 1)p d^{2\omega/3}.$$

Finally, taking the maximum over $\phi \in \mathcal{X}$ and using Strassen duality yields

$$\mathfrak{R}(T)^2 \leq d^{4\omega/3} - (2^{\omega/3} - 1)p d^{2\omega/3},$$

i.e.,

$$\mathfrak{R}(T) \leq \sqrt{d^{4\omega/3} - \left\lfloor \frac{d}{3} \right\rfloor \cdot (2^{\omega/3} - 1)d^{2\omega/3}}. \quad \blacktriangleleft$$

Let $\sigma(d)$ denote the exponent of $d \times d \times d$ tensors [30], i.e., the smallest σ such that every $d \times d \times d$ tensor T satisfies $\mathfrak{R}(T) \leq d^\sigma$. Then the above results imply the following.

► **Corollary 39.** *For any $d > 0$, $\sigma(d) < \frac{2}{3}\omega$.*

► **Remark 40.** This answers an open problem posed by Kaski [29, Open Problem 1], which asks whether Strassen's bound $\sigma(d) \leq \frac{2}{3}\omega$ can be improved. It is worth emphasizing, however, that our improvement has the form $\sigma(d) \leq \frac{2}{3}\omega - \varepsilon_d$, where $\varepsilon_d \rightarrow 0$ rapidly as $d \rightarrow \infty$.

In particular, to make progress on fine-grained questions such as the set cover conjecture [10, 34], one would need to achieve $\sigma(d) \leq 1.08$. While in the most ideal case $\omega = 2$, currently we only have an exponent bound $\sigma(d) \leq 4/3 - o(1)$. On the other hand, this current approach also seems ill-suited for improving ω : the strategy is to reduce a generic tensor T to matrix multiplication, but when T is itself a matrix multiplication tensor, this reduction becomes too indirect to be effective.

To make progress toward the (extended; see [30, Conjecture 2]) asymptotic rank conjecture, we propose the following question, which seems to require ideas beyond a naive application of the techniques developed in this paper.

► **Problem 41.** Prove that there exists $\delta > 0$ such that $\sigma(d) \leq \frac{2}{3}\omega - \delta$ for all d .

7.3 A direct sum identity

In this subsection, we use the speedup theorems developed above to derive an intriguing direct sum identity that generalizes Schönhage's classical construction [38, Lemma 6.1]:

$$\langle n, 1, m \rangle \oplus \langle 1, (n-1)(m-1), 1 \rangle \leq \langle nm+1 \rangle.$$

► **Theorem 42.** *Let p, q be positive integers, and let $\{n_\alpha\}_{1 \leq \alpha \leq p}$ and $\{m_\beta\}_{1 \leq \beta \leq q}$ be sequences of positive integers. Set*

$$n = \sum_{\alpha=1}^p n_\alpha, \quad m = \sum_{\beta=1}^q m_\beta.$$

Then there exists a degeneration between direct sums of matrix multiplication tensors:

$$\langle n, 1, m \rangle \oplus \bigoplus_{\substack{1 \leq \alpha \leq p \\ 1 \leq \beta \leq q}} \langle 1, (n_\alpha - 1)(m_\beta - 1), 1 \rangle \leq \langle nm \rangle \oplus \langle p, 1, q \rangle.$$

Before proving Theorem 42, we briefly compare it with Schönhage's original direct sum identity. Taking the direct sum over all (α, β) gives

$$\bigoplus_{\alpha, \beta} \langle n_\alpha, 1, m_\beta \rangle \oplus \bigoplus_{\alpha, \beta} \langle 1, (n_\alpha - 1)(m_\beta - 1), 1 \rangle \leq \langle nm \rangle \oplus \langle pq \rangle.$$

Compared with Theorem 42, the direct-sum tensor $\bigoplus_{\alpha, \beta} \langle n_\alpha, 1, m_\beta \rangle$ on the left-hand side is more valuable than $\langle n, 1, m \rangle$. However, this also comes with a bigger cost on the right-hand side: the slack term $\langle pq \rangle$ replaces the one-slice tensor $\langle p, 1, q \rangle$.

In the special case $n_\alpha = n$ and $m_\beta = m$ for all α, β , our identity becomes

$$\langle pn, 1, qm \rangle \oplus pq \odot \langle 1, (n - 1)(m - 1), 1 \rangle \leq \langle pqnm \rangle \oplus \langle p, 1, q \rangle.$$

If instead one tensors Schönhage's identity for (n, m) with $\langle p, 1, q \rangle$, one obtains

$$\langle pn, 1, qm \rangle \oplus \langle p, (n - 1)(m - 1), q \rangle \leq nm \odot \langle p, 1, q \rangle \oplus \langle p, 1, q \rangle.$$

Here the direct summand $nm \odot \langle p, 1, q \rangle$ on the right-hand side is cheaper than $\langle pqnm \rangle$, but the tensor $\langle p, (n - 1)(m - 1), q \rangle$ on the left-hand side is also less valuable than $pq \odot \langle 1, (n - 1)(m - 1), 1 \rangle$.

In sum, our new identity does not seem to be easily derived from Schönhage's identity alone.

Proof of Theorem 42. We start from a redundant restriction

$$\langle n, 1, m \rangle \leq \langle nm \rangle \oplus \langle p, 1, q \rangle.$$

Write

$$\langle n, 1, m \rangle = \sum_{i=1}^n \sum_{j=1}^m x_i y_j z_{ij}, \quad S := \langle nm \rangle \oplus \langle p, 1, q \rangle = \sum_{i=1}^n \sum_{j=1}^m u_{ij} v_{ij} w_{ij} - \sum_{\alpha=1}^p \sum_{\beta=1}^q u'_\alpha v'_\beta w'_{\alpha\beta},$$

where the minus sign is introduced to simplify later cancellations.

Fix partitions $[n] = I_1 \sqcup \cdots \sqcup I_p$ and $[m] = J_1 \sqcup \cdots \sqcup J_q$ with $|I_\alpha| = n_\alpha$ and $|J_\beta| = m_\beta$. Define linear maps A, B, C by

$$\begin{aligned} A : \quad x_i &\leftarrow u_{ij}, & \sum_{i \in I_\alpha} x_i &\leftarrow u'_\alpha, \\ B : \quad y_j &\leftarrow v_{ij}, & \sum_{j \in J_\beta} y_j &\leftarrow v'_\beta, \\ C : \quad z_{ij} &\leftarrow w_{ij}, & 0 &\leftarrow w'_{\alpha\beta}. \end{aligned}$$

These maps define a restriction $\langle n, 1, m \rangle \leq S$.

To apply the free-lunch speedup theorem (Theorem 17), define a subspace $C' \subseteq W^\vee$ via the linear map

$$C' : \quad w'_{\alpha(i)\beta(j)} \leftarrow w_{ij}, \quad w'_{\alpha\beta} \leftarrow w'_{\alpha\beta},$$

36:30 Asymptotic Rank Speedup Theorems, Revisited

where $\alpha(i)$ (resp. $\beta(j)$) is the unique index such that $i \in I_{\alpha(i)}$ (resp. $j \in J_{\beta(j)}$). A direct computation shows that $(A \otimes B \otimes C')S = 0$.

Applying Theorem 17 yields a direct summand $T' = (A' \otimes B' \otimes C')S$ on the left-hand side.

Let us make the resulting subspaces $A' \subseteq U^\vee$ and $B' \subseteq V^\vee$ explicit. By definition,

$$A' = \{u \in U^\vee : (u \otimes B \otimes C')S = 0\}, \quad B' = \{v \in V^\vee : (A \otimes v \otimes C')S = 0\}.$$

Unwinding the maps A, B and the definition of C' , one checks that A' is the annihilator of the subspace spanned by the vectors

$$u'_\alpha - \sum_{i \in I_\alpha} u_{ij} \quad (\text{for all } \alpha \in [p], j \in [m]),$$

and similarly B' is the annihilator of the subspace spanned by

$$v'_\beta - \sum_{j \in J_\beta} v_{ij} \quad (\text{for all } \beta \in [q], i \in [n]).$$

In particular, in T' the relations

$$u'_\alpha = \sum_{i \in I_\alpha} u_{ij} \quad (\text{for all } \alpha \in [p], j \in [m]) \quad \text{and} \quad v'_\beta = \sum_{j \in J_\beta} v_{ij} \quad (\text{for all } \beta \in [q], i \in [n])$$

hold simultaneously. For each α choose a distinguished index $i_\alpha^* \in I_\alpha$ and for each β choose a distinguished index $j_\beta^* \in J_\beta$. These relations allow us to eliminate the variables $u_{i_\alpha^* j}$ and $v_{ij j_\beta^*}$ via

$$u_{i_\alpha^* j} = u'_\alpha - \sum_{i \in I_\alpha \setminus \{i_\alpha^*\}} u_{ij}, \quad v_{ij j_\beta^*} = v'_\beta - \sum_{j \in J_\beta \setminus \{j_\beta^*\}} v_{ij}.$$

We now impose a further restriction by applying suitable projections on the first two modes. Define linear maps π_U and π_V by

$$\pi_U(u'_\alpha) = 0, \quad \pi_U(u_{i j_\beta^*}) = 0 \quad (\forall i, \beta), \quad \pi_U(u_{ij}) = u_{ij} \quad (j \neq j_\beta^*),$$

and

$$\pi_V(v'_\beta) = 0, \quad \pi_V(v_{i_\alpha^* j}) = 0 \quad (\forall \alpha, j), \quad \pi_V(v_{ij}) = v_{ij} \quad (i \neq i_\alpha^*).$$

Applying π_U and π_V to the first two modes of T' yields a further restriction

$$T^* := (\pi_U \otimes \pi_V \otimes \text{id})T' \leq T'.$$

By construction, π_U kills all u'_α and all variables in the distinguished columns $\{u_{i j_\beta^*}\}$, and π_V kills all v'_β and all variables in the distinguished rows $\{v_{i_\alpha^* j}\}$. After eliminating $u_{i_\alpha^* j}$ and $v_{ij j_\beta^*}$ using the identities above, every term that contains an eliminated variable involves either u'_α (or a distinguished column) or v'_β (or a distinguished row), and hence is annihilated by $\pi_U \otimes \pi_V$. Therefore, the only surviving terms are those with $i \in I_\alpha \setminus \{i_\alpha^*\}$ and $j \in J_\beta \setminus \{j_\beta^*\}$.

The resulting tensor is

$$T^* = \sum_{\alpha=1}^p \sum_{\beta=1}^q \left(\sum_{i \in I_\alpha \setminus \{i_\alpha^*\}} \sum_{j \in J_\beta \setminus \{j_\beta^*\}} u_{ij} v_{ij} \right) w_{\alpha\beta} \cong \bigoplus_{\alpha, \beta} \langle 1, (n_\alpha - 1)(m_\beta - 1), 1 \rangle.$$

Composing the restrictions and degenerations yields the claimed identity. $\blacktriangleleft$

References

- 1 Josh Alman. Limits on the universal method for matrix multiplication. *Theory Comput.*, 17:Paper No. 1, 30, 2021. doi:10.4086/toc.2021.v017a001.
- 2 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. SIAM, 2025. doi:10.1137/1.9781611978322.
- 3 Josh Alman and Baitian Li. Kronecker powers, orthogonal vectors, and the asymptotic spectrum. In *2025 IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1411–1441. IEEE, 2025. doi:10.1109/FOCS63196.2025.00074.
- 4 Josh Alman and Virginia Vassilevska Williams. Limits on all known (and some unknown) approaches to matrix multiplication. *SIAM J. Comput.*, 52(6):FOCS18–285–FOCS18–315, 2023. doi:10.1137/19M124695X.
- 5 Noga Alon. The Shannon capacity of a union. *Combinatorica*, 18(3):301–310, 1998. doi:10.1007/PL00009824.
- 6 Andris Ambainis, Yuval Filmus, and François Le Gall. Fast matrix multiplication: limitations of the coppersmith-winograd method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, pages 585–593, 2015. doi:10.1145/2746539.2746554.
- 7 Theodore Baker, John Gill, and Robert Solovay. Relativizations of the $\mathcal{P} = ?\mathcal{NP}$ question. *SIAM J. Comput.*, 4(4):431–442, 1975. doi:10.1137/0204037.
- 8 D. Bini. Relations between exact and approximate bilinear algorithms. Applications. *Calcolo*, 17(1):87–97, 1980. doi:10.1007/BF02575865.
- 9 Andreas Björklund, Radu Curticapean, Thore Husfeldt, Petteri Kaski, and Kevin Pratt. Fast deterministic chromatic number under the asymptotic rank conjecture. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2804–2818. SIAM, 2025. doi:10.1137/1.9781611978322.91.
- 10 Andreas Björklund and Petteri Kaski. The asymptotic rank conjecture and the set cover conjecture are not both true. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 859–870, 2024. doi:10.1145/3618260.3649656.
- 11 Andreas Björklund, Petteri Kaski, Tomohiro Koana, and Jesper Nederlof. Kronecker scaling of tensors with applications to arithmetic circuits and algorithms. In *53rd International Colloquium on Automata, Languages, and Programming (ICALP 2026)*, Leibniz International Proceedings in Informatics (LIPIcs), pages 36:1–36:24. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi:10.4230/LIPIcs.ICALP.2026.36.
- 12 Markus Bläser. *Fast Matrix Multiplication*. Number 5 in Graduate Surveys. Theory of Computing Library, 2013. doi:10.4086/toc.gs.2013.005.
- 13 Manuel Blum. A machine-independent theory of the complexity of recursive functions. *J. Assoc. Comput. Mach.*, 14:322–336, 1967. doi:10.1145/321386.321395.
- 14 Cornelius Brand, Radu Curticapean, Baitian Li, and Kevin Pratt. Faster convolutions: Yates and strassen revisited. In *2026 SIAM Symposium on Simplicity in Algorithms (SOSA)*, pages 328–339. SIAM, 2026. doi:10.1137/1.9781611978964.25.
- 15 Peter Bürgisser, Michael Clausen, and Mohammad A Shokrollahi. *Algebraic complexity theory*, volume 315. Springer Science & Business Media, 2013. doi:10.1007/978-3-662-03338-8.
- 16 Matthias Christandl, François Le Gall, Vladimir Lysikov, and Jeroen Zuiddam. Barriers for rectangular matrix multiplication. *Comput. Complexity*, 34(1):Paper No. 4, 38, 2025. doi:10.1007/s00037-025-00264-9.
- 17 Matthias Christandl, Vladimir Lysikov, and Jeroen Zuiddam. Weighted slice rank and a minimax correspondence to Strassen’s spectra. *J. Math. Pures Appl. (9)*, 172:299–329, 2023. doi:10.1016/j.matpur.2023.02.006.
- 18 Matthias Christandl, Péter Vrana, and Jeroen Zuiddam. Barriers for fast matrix multiplication from irreversibility. *Theory Comput.*, 17:Paper No. 2, 32, 2021. doi:10.4086/toc.2021.v017a002.

- 19 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.
- 20 Austin Conner, Fulvio Gesmundo, Joseph M. Landsberg, and Emanuele Ventura. Rank and border rank of Kronecker powers of tensors and Strassen’s laser method. *Comput. Complexity*, 31(1):Paper No. 1, 40, 2022. doi:10.1007/s00037-021-00217-y.
- 21 Austin Conner, Joseph M. Landsberg, Fulvio Gesmundo, and Emanuele Ventura. Kronecker powers of tensors and strassen’s laser method. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151, pages 10:1–10:28, 2020. doi:10.4230/LIPIcs.ITCS.2020.10.
- 22 D. Coppersmith and S. Winograd. On the asymptotic complexity of matrix multiplication. *SIAM J. Comput.*, 11(3):472–492, 1982. doi:10.1137/0211038.
- 23 Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *J. Symbolic Comput.*, 9(3):251–280, 1990. doi:10.1016/S0747-7171(08)80013-2.
- 24 Ran Duan, Hongxun Wu, and Renfei Zhou. Faster matrix multiplication via asymmetric hashing. In *2023 IEEE 64th annual symposium on Foundations of Computer Science (FOCS)*, pages 2129–2138. IEEE, 2023. doi:10.1109/FOCS57990.2023.00130.
- 25 Cosimo Flavi, Joachim Jelisiejew, and Mateusz Michałek. Symmetric powers: structure, smoothability, and applications. *Int. Math. Res. Not. IMRN*, 2025(18):Paper No. rnaf277, 40, 2025. doi:10.1093/imrn/rnaf277.
- 26 Willem Haemers. On some problems of Lovász concerning the Shannon capacity of a graph. *IEEE Trans. Inform. Theory*, 25(2):231–232, 1979. doi:10.1109/TIT.1979.1056027.
- 27 Shuichi Hirahara and Nobutaka Shimizu. Error-correction of matrix multiplication algorithms. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 785–794, 2025. doi:10.1145/3717823.3718244.
- 28 Shuichi Hirahara and Nobutaka Shimizu. An optimal error-correcting reduction for matrix multiplication. In *52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025)*, pages 97–1, 2025. doi:10.4230/LIPIcs.ICALP.2025.97.
- 29 Petteri Kaski. Tensor exponents and tensor sequences beyond matrix multiplication. Talk at the Simons Institute for the Theory of Computing, October 2025. URL: <https://simons.berkeley.edu/talks/petteri-kaski-aalto-university-2025-10-31>.
- 30 Petteri Kaski and Mateusz Michałek. A universal sequence of tensors for the asymptotic rank conjecture. In *16th Innovations in Theoretical Computer Science Conference (ITCS 2025)*, pages 64:1–64:24, 2025. doi:10.4230/LIPIcs.ITCS.2025.64.
- 31 Kisun Lee. Asymptotic rank bounds: a numerical census, 2026. arXiv:2601.08119.
- 32 Thomas Lickteig. Typical tensorial rank. *Linear Algebra Appl.*, 69:95–120, 1985. doi:10.1016/0024-3795(85)90070-9.
- 33 V. Ya. Pan. New combinations of methods for the acceleration of matrix multiplication. *Comput. Math. Appl.*, 7(1):73–125, 1981. doi:10.1016/0898-1221(81)90009-2.
- 34 Kevin Pratt. A stronger connection between the asymptotic rank conjecture and the set cover conjecture. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 871–874, 2024. doi:10.1145/3618260.3649620.
- 35 Robert Robere and Jeroen Zuiddam. Amortized circuit complexity, formal complexity measures, and catalytic algorithms. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 759–769. IEEE, 2021. doi:10.1109/FOCS52979.2021.00079.
- 36 Francesco Romani. Some properties of disjoint sums of tensors related to matrix multiplication. *SIAM J. Comput.*, 11(2):263–267, 1982. doi:10.1137/0211020.
- 37 Keiya Sakabe, M. Levent Doğan, and Michael Walter. Strassen’s support functionals coincide with the quantum functionals, 2026. arXiv:2601.21553.
- 38 A. Schönhage. Partial and total matrix multiplication. *SIAM J. Comput.*, 10(3):434–455, 1981. doi:10.1137/0210032.
- 39 V. Strassen. Rank and optimal computation of generic tensors. *Linear Algebra Appl.*, 52/53:645–685, 1983. doi:10.1016/0024-3795(83)80041-X.

- 40 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.
- 41 V. Strassen. The asymptotic spectrum of tensors. *J. Reine Angew. Math.*, 384:102–152, 1988. doi:10.1515/crll.1988.384.102.
- 42 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.
- 43 Volker Strassen. Gaussian elimination is not optimal. *Numer. Math.*, 13:354–356, 1969. doi:10.1007/BF02165411.
- 44 Volker Strassen. Vermeidung von Divisionen. *J. Reine Angew. Math.*, 264:184–202, 1973. doi:10.1515/crll.1973.264.184.
- 45 Volker Strassen. Algebra and complexity. In *First European Congress of Mathematics Paris, July 6–10, 1992: Vol. II: Invited Lectures (Part 2)*, pages 429–446. Springer, 1994.
- 46 Volker Strassen. Komplexität und Geometrie bilinear Abbildungen. *Jahresber. Deutsch. Math.-Verein.*, 107(1):3–31, 2005.
- 47 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.
- 48 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. SIAM, 2024. doi:10.1137/1.9781611977912.
- 49 Jeroen Zuiddam. The asymptotic spectrum of graphs and the Shannon capacity. *Combinatorica*, 39(5):1173–1184, 2019. doi:10.1007/s00493-019-3992-5.

A

 Degenerations between direct sums of matrix multiplication

The goal of this appendix is to clarify how the direct-sum degeneration used by Schönhage [38, §8],

$$\langle 1, 5, 20 \rangle \oplus \langle 10, 2, 5 \rangle \oplus \langle 10, 10, 1 \rangle \leq \langle 132 \rangle \oplus 5 \odot \langle 1, 1, 2 \rangle, \quad (\text{A.1})$$

can be analyzed (via Strassen calculus) to obtain the inequality

$$3 \cdot 100^{\omega/3} \leq 132 + 5 \cdot 2^{\omega/3}. \quad (\text{A.2})$$

Moreover, this analysis is *optimal* under a suitable formalization.

More generally, given a single degeneration between direct sums of matrix multiplication tensors, one may ask for the best upper bound on ω that can be extracted from it:

$$\bigoplus_{i=1}^t \langle n_i^{(1)}, n_i^{(2)}, n_i^{(3)} \rangle \leq \bigoplus_{j=1}^s \langle m_j^{(1)}, m_j^{(2)}, m_j^{(3)} \rangle. \quad (\text{A.3})$$

We begin with a Strassen-calculus proof of the classical asymptotic sum inequality, which yields an upper bound in the special case where the right-hand side is diagonal, i.e., $m_j^{(1)} = m_j^{(2)} = m_j^{(3)} = 1$. This proof is due to Strassen [41, §4].

► **Theorem 43** (Asymptotic sum inequality [38, Theorem 7.1]). *If there exists a degeneration,*

$$\bigoplus_{i=1}^t \langle n_i^{(1)}, n_i^{(2)}, n_i^{(3)} \rangle \leq \langle s \rangle,$$

let τ be the (unique) solution to the equation

$$\sum_{i=1}^t (n_i^{(1)} n_i^{(2)} n_i^{(3)})^\tau = s,$$

then $\omega \leq 3\tau$.

Proof.

Step 1: Pass to the asymptotic spectrum. Fix $\phi \in \mathcal{X}$ with parameters $(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}$. Applying ϕ to the degeneration yields

$$F(\theta_1, \theta_2, \theta_3) := \sum_{i=1}^t (n_i^{(1)})^{\theta_1} (n_i^{(2)})^{\theta_2} (n_i^{(3)})^{\theta_3} \leq s.$$

Step 2: Symmetry of tensors. We can permute the modes and obtain degenerations that permute the parameters of matrix multiplication. For any $\sigma \in \mathfrak{S}_3$, we also have a degeneration

$$\bigoplus_{i=1}^t \langle n_i^{(\sigma^{-1}(1))}, n_i^{(\sigma^{-1}(2))}, n_i^{(\sigma^{-1}(3))} \rangle \leq \langle s \rangle,$$

hence $F(\theta_{\sigma(1)}, \theta_{\sigma(2)}, \theta_{\sigma(3)}) \leq s$.

Step 3: Average by convexity. The function F is convex (it is a sum of exponentials), so by Jensen's inequality,

$$F(\mathbb{E}_{\sigma}[(\theta_{\sigma(1)}, \theta_{\sigma(2)}, \theta_{\sigma(3)})]) \leq \mathbb{E}_{\sigma}[F(\theta_{\sigma(1)}, \theta_{\sigma(2)}, \theta_{\sigma(3)})] \leq s.$$

Let $\varpi := \theta_1 + \theta_2 + \theta_3$. Averaging over $\sigma \in \mathfrak{S}_3$ gives

$$\mathbb{E}_{\sigma}[(\theta_{\sigma(1)}, \theta_{\sigma(2)}, \theta_{\sigma(3)})] = (\varpi/3, \varpi/3, \varpi/3).$$

Since $\varpi \leq \omega$ for all $(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}$ and F is coordinatewise nondecreasing, we obtain $F(\omega/3, \omega/3, \omega/3) \leq s$, which implies the claim. $\blacktriangleleft$

We emphasize that the first two steps in the proof above amount to a simple geometric relaxation, which extends verbatim to the case where the right-hand side is also a direct sum of matrix multiplication tensors:

Step 1: Pass to the asymptotic spectrum. Every point $(\theta_1, \theta_2, \theta_3) \in \Delta_{\text{MM}}$ must lie in the feasible region

$$\Delta_{\text{MM}} \subseteq \Delta_{(\text{A.3})} := \left\{ (\theta_1, \theta_2, \theta_3) : \sum_{i=1}^t (n_i^{(1)})^{\theta_1} (n_i^{(2)})^{\theta_2} (n_i^{(3)})^{\theta_3} \leq \sum_{j=1}^s (m_j^{(1)})^{\theta_1} (m_j^{(2)})^{\theta_2} (m_j^{(3)})^{\theta_3} \right\}.$$

Step 2: Symmetry. Permuting tensor modes yields the containment

$$\Delta_{\text{MM}} \subseteq \Delta_{(\text{A.3})}^{\sigma} = \{(\theta_1, \theta_2, \theta_3) : (\theta_{\sigma(1)}, \theta_{\sigma(2)}, \theta_{\sigma(3)}) \in \Delta_{(\text{A.3})}\}$$

for all $\sigma \in \mathfrak{S}_3$.

Step 3: Optimize over the relaxation. Intersecting all six permuted regions gives the strongest symmetric relaxation: $\Delta_{\text{MM}} \subseteq \Delta_{(\text{A.3})}^{\mathfrak{S}_3} := \bigcap_{\sigma \in \mathfrak{S}_3} \Delta_{(\text{A.3})}^{\sigma}$. By Strassen duality, we then obtain an upper bound on ω :

$$\omega = \max_{\Delta_{\text{MM}}} (\theta_1 + \theta_2 + \theta_3) \leq \max_{\Delta_{(\text{A.3})}^{\mathfrak{S}_3}} (\theta_1 + \theta_2 + \theta_3).$$

From this viewpoint, the classical asymptotic sum inequality says that, in the diagonal case, the region $\Delta_{(A.3)}$ is convex, so the maximum of $\theta_1 + \theta_2 + \theta_3$ over $\Delta_{(A.3)}^\mathfrak{S}$ is attained on the principal diagonal $\theta_1 = \theta_2 = \theta_3$, which simplifies the analysis.

Therefore, the analysis of the direct-sum identity (A.1) can be carried out via computer-aided optimization. The inequality induced by the degeneration is

$$\Delta_{(A.1)} := \{(\theta_1, \theta_2, \theta_3) : 5^{\theta_2} \cdot 20^{\theta_3} + 10^{\theta_1} \cdot 2^{\theta_2} \cdot 5^{\theta_3} + 10^{\theta_1} \cdot 10^{\theta_2} \leq 132 + 5 \cdot 2^{\theta_3}\}.$$

It turns out (and can be verified in standard mathematical software packages) that the maximum of $\theta_1 + \theta_2 + \theta_3$ over $\Delta_{(A.1)}^\mathfrak{S}$ is attained at $\theta_1 = \theta_2 = \theta_3$. This justifies (A.2) as a valid upper bound on ω .

One may then ask whether an asymptotic-sum-inequality statement holds in *full generality*. Namely, can one conclude

$$(A.3) \implies \sum_{i=1}^t (n_i^{(1)} n_i^{(2)} n_i^{(3)})^{\omega/3} \leq \sum_{j=1}^s (m_j^{(1)} m_j^{(2)} m_j^{(3)})^{\omega/3}. \quad (A.4)$$

The answer is nuanced: in a certain relativized sense the answer is *no*, but it is also plausible that the answer is *yes* for tensors.

We first explain the *no* side. A natural first attempt is to let τ be the largest real satisfying

$$\sum_{i=1}^t (n_i^{(1)} n_i^{(2)} n_i^{(3)})^\tau \leq \sum_{j=1}^s (m_j^{(1)} m_j^{(2)} m_j^{(3)})^\tau,$$

Then $(\tau, \tau, \tau) \in \Delta_{(A.3)}^\mathfrak{S}$, so 3τ provides a lower bound on the best upper bound on ω obtainable from the identity. However, the converse need not hold: the maximum of $\theta_1 + \theta_2 + \theta_3$ over $\Delta_{(A.3)}^\mathfrak{S}$ may be attained away from the principal diagonal $\theta_1 = \theta_2 = \theta_3$. In such cases, the optimal bound on ω extracted from the degeneration can be strictly larger than 3τ .

A cautious reader may notice that we have been deliberately vague about what it means to obtain “the best bound on ω from a single identity”. In particular, one should clarify what is meant by “using only the identity” as opposed to using additional tensor-specific facts.

The underlying intuition is that the classical asymptotic sum inequality – and the bound (A.2) derived from (A.1) – is, in a sense, *blind* to the arithmetic structure of tensors. Once one abstracts the basic axioms governing rank and degeneration, the same argument goes through without using any additional tensor-specific structure. From the perspective of complexity theory, this can be phrased as a form of *relativization* [7]: in any mathematical structure that satisfies the same axioms and “knows” only the single identity (A.3), an asymptotic-sum-inequality type bound still holds. In particular, one can show that there exists a relativized world in which ω cannot be smaller than $\max_{\Delta_{(A.3)}^\mathfrak{S}} (\theta_1 + \theta_2 + \theta_3)$. One way to formalize this is to work in a *universal commutative semiring* that encodes the axioms together with the given identity. We give a rigorous treatment in the subsequent subsection.

We briefly indicate why the answer *might* nevertheless be yes. The discussion above only shows that (A.4) is too strong to hold in all relativized settings; additional structure specific to tensors might still imply it. A natural candidate is *Strassen’s convexity conjecture*, which asserts that Δ_{MM} is a convex subset of $\mathbb{R}^3$. If Δ_{MM} were convex, then (A.4) would follow almost immediately. Moreover, the conjecture is not entirely speculative: Strassen proved partial results showing that Δ_{MM} is star-convex with respect to the points $(0, 1, 1)$, $(1, 0, 1)$, and $(1, 1, 0)$; see [41, §5]. For further discussion, see [41, §4, 5] or the more detailed exposition in [47, Part II, III].

A.1 Strassen duality and asymptotic sum inequality in general rings

To make the preceding discussion precise, we briefly recall the general notion of rank and Strassen duality in the abstract setting of commutative semirings. This framework is explained in detail by Wigderson and Zuydam [47, Part I].

General framework of Strassen duality. Let $\mathcal{R}$ be a commutative semiring with additive identity 0 and multiplicative identity 1. We think of elements of $\mathcal{R}$ as the “objects” of interest. For example, in the $\mathbb{F}$ -tensor world, $\mathcal{R} = \mathcal{R}_{\mathbb{F}}$ can be taken to be the set of all $\mathbb{F}$ -tensors (up to isomorphism), with addition given by direct sum $\oplus$ and multiplication given by tensor product $\otimes$; the zero tensor plays the role of 0, and the unit diagonal tensor $\langle 1 \rangle$ plays the role of 1.

We also equip $\mathcal{R}$ with a preorder P (written $\leq_P$), which encodes the basic comparison relation between objects (e.g., restriction or degeneration). In this setting, the rank of an element $a \in \mathcal{R}$ is defined as

$$R_P(a) := \min\{n \in \mathbb{N} : a \leq_P n\},$$

where $n \in \mathcal{R}$ denotes the sum of n copies of the unit 1. To ensure that this notion behaves as expected, one assumes the following axioms; a preorder satisfying them will be called a *Strassen preorder*:

(Preorder) Reflexivity ($a \leq_P a$) and transitivity ($a \leq_P b, b \leq_P c \implies a \leq_P c$).

(Compatibility) $0 \leq_P a$, and the preorder respects addition and multiplication: $a \leq_P b \implies a + c \leq_P b + c$ and $ac \leq_P bc$.

(Faithfulness on natural numbers) For $n, m \in \mathbb{N}$, one has $n \leq_P m$ if and only if $n \leq m$ as natural numbers.

(Strong Archimedean property) For any $a \neq 0$, there exists $n \in \mathbb{N}$ such that $1 \leq_P a \leq_P n$. From now on, we omit P when unambiguous. Under these axioms, the rank function is subadditive and submultiplicative, and one can define the asymptotic rank $\underline{R}(a)$ in the usual way. For example, in the tensor setting, both restriction $\leq$ and degeneration $\preceq$ satisfy these axioms.

One can then define the asymptotic spectrum $\mathcal{X}$ of $(\mathcal{R}, \leq)$: it consists of all functions $\phi: \mathcal{R} \rightarrow \mathbb{R}_{\geq 0}$ that

(Semiring homomorphism) $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(ab) = \phi(a)\phi(b)$.

(Normalization) $\phi(n) = n$ for all $n \in \mathbb{Z}_{\geq 0}$.

(Monotonicity) $a \leq b \implies \phi(a) \leq \phi(b)$.

Strassen duality (as discussed in Section 4) remains valid in this abstract setting. For our purposes, we mainly use the identity

$$\underline{R}(a) = \max_{\phi \in \mathcal{X}} \phi(a).$$

Relativization of matrix multiplication. We now formulate an abstract semiring *representing* matrix multiplication, in order to study the asymptotic sum inequality. We define $\mathcal{R}_{\text{MM}}$ to be the commutative semiring of finite formal $\mathbb{N}$ -linear combinations of symbols $\langle n, m, p \rangle$:

$$\mathcal{R}_{\text{MM}} = \left\{ \sum_{n,m,p} a_{nmp} \langle n, m, p \rangle : a_{nmp} \in \mathbb{N}, \sum_{n,m,p} a_{nmp} < \infty \right\}.$$

Addition in $\mathcal{R}_{\text{MM}}$ is defined by adding coefficients coordinatewise, and multiplication is defined on generators by $\langle n, m, p \rangle \cdot \langle n', m', p' \rangle = \langle nn', mm', pp' \rangle$ and extended bilinearly:

$$\left(\sum_{n,m,p} a_{nmp} \langle n, m, p \rangle \right) \left(\sum_{n',m',p'} b_{n'm'p'} \langle n', m', p' \rangle \right) = \sum_{\substack{n,m,p \\ n',m',p'}} a_{nmp} b_{n'm'p'} \langle nn', mm', pp' \rangle.$$

Note that $\langle 1, 1, 1 \rangle$ is the multiplicative identity, so we will identify $\langle 1, 1, 1 \rangle$ with 1.

One may view elements of $\mathcal{R}_{\text{MM}}$ as formal analogues of direct sums of matrix multiplication tensors, e.g., $\sum_{n,m,p} a_{nmp} \langle n, m, p \rangle$ corresponds heuristically to $\bigoplus_{n,m,p} a_{nmp} \odot \langle n, m, p \rangle$. However, elements of $\mathcal{R}_{\text{MM}}$ are purely formal: they do not “know” anything about tensors beyond the axioms we impose below.

We next define a preorder by specifying a set of axioms. Let Σ be a collection of inequalities $a \leq b$ with $a, b \in \mathcal{R}_{\text{MM}}$, encoding a minimal set of properties of matrix multiplication. Concretely, Σ contains the following families of inequalities (for all $n \in \mathbb{Z}_+$):

(Monotonicity) For any $n \in \mathbb{Z}_+$, we have $\langle n, 1, 1 \rangle \leq \langle n+1, 1, 1 \rangle$, $\langle 1, n, 1 \rangle \leq \langle 1, n+1, 1 \rangle$ and $\langle 1, 1, n \rangle \leq \langle 1, 1, n+1 \rangle$.

(Trivial upper bound) For any $n \in \mathbb{Z}_+$, we have $\langle n, 1, 1 \rangle \leq n$, $\langle 1, n, 1 \rangle \leq n$ and $\langle 1, 1, n \rangle \leq n$. Given a set of axioms Σ , we write $\Sigma \vdash a \leq b$ if the inequality $a \leq b$ can be *derived* from Σ in finitely many steps using only (Preorder) and (Compatibility).

We define a preorder $P(\Sigma)$ on $\mathcal{R}_{\text{MM}}$ by declaring $a \leq_{P(\Sigma)} b$ if and only if $\Sigma \vdash a \leq b$.

► **Proposition 44.** *The preorder $P(\Sigma)$ is a Strassen preorder.*

Proof. By construction, $P(\Sigma)$ is a preorder and is compatible with addition and multiplication. We verify the remaining axioms.

For (Strong Archimedean property), let $a = \sum_{n,m,p} a_{nmp} \langle n, m, p \rangle \in \mathcal{R}_{\text{MM}}$. From Σ one can derive

$$\sum_{n,m,p} a_{nmp} \leq \sum_{n,m,p} a_{nmp} \langle n, m, p \rangle \leq \sum_{n,m,p} a_{nmp} \cdot nmp.$$

If $a \neq 0$, then $\sum_{n,m,p} a_{nmp} \geq 1$, so in particular $1 \leq a \leq N$ for some $N \in \mathbb{N}$.

Finally, for (Faithfulness on natural numbers), observe that each inequality in Σ holds when interpreting elements of $\mathcal{R}_{\text{MM}}$ in the tensor semiring $\mathcal{R}_{\mathbb{F}}$ (with restriction preorder). Hence, if $\Sigma \vdash n \leq m$, then $\langle n \rangle \leq \langle m \rangle$ in the tensor world, which forces $n \leq m$ as natural numbers. ◀

Since $P(\Sigma)$ is a Strassen preorder, Strassen duality applies, and it becomes meaningful to describe its asymptotic spectrum. In particular, the parametrization in Proposition 7 still holds: re-running the proof shows that it relies only on (Monotonicity), which is already contained in Σ . Moreover, in this setting the asymptotic spectrum is naturally parametrized by triples $(\theta_1, \theta_2, \theta_3)$, since every element of $\mathcal{R}_{\text{MM}}$ is built from generators $\langle n, m, p \rangle$.

Now fix a collection Γ of additional identities of interest (i.e., inequalities $a \leq b$ in $\mathcal{R}_{\text{MM}}$). In principle, Γ may be infinite – for instance, it could encode an abstracted family of degenerations arising from the laser method – but the reader may keep in mind the case where Γ consists of a single identity, such as Schönhage’s.

We analogously define the preorder $P(\Sigma \cup \Gamma)$ by closing $\Sigma \cup \Gamma$ under derivations. It is immediate that $P(\Sigma \cup \Gamma)$ satisfies the axioms of a Strassen preorder except, potentially, (Faithfulness on natural numbers): for an arbitrary choice of Γ , one could force a contradiction (e.g., by including $1 \leq 0$). In the applications we have in mind, Γ comes from genuine tensor identities, and faithfulness holds.

From now on, we assume $P(\Sigma \cup \Gamma)$ is indeed a Strassen preorder, and now try to describe the structure of the corresponding Δ_{MM} in this case, denoted as $\Delta_{\text{MM}}(\Gamma)$.

For an element in $\mathcal{R}_{\text{MM}}$:

$$a = \sum_{n,m,p} a_{nmp} \langle n, m, p \rangle,$$

and for $(\theta_1, \theta_2, \theta_3)$ we write

$$a(\theta_1, \theta_2, \theta_3) := \sum_{n,m,p} a_{nmp} n^{\theta_1} m^{\theta_2} p^{\theta_3}.$$

► **Proposition 45.** *Let Δ_Γ be the feasible region cut out by the inequalities in Γ :*

$$\Delta_\Gamma := \left\{ (\theta_1, \theta_2, \theta_3) : a(\theta_1, \theta_2, \theta_3) \leq b(\theta_1, \theta_2, \theta_3) \text{ for all } (a \leq b) \in \Gamma \right\}.$$

If $P(\Sigma \cup \Gamma)$ is a Strassen preorder, then $\Delta_{\text{MM}}(\Gamma) = [0, 1]^3 \cap \Delta_\Gamma$.

Proof. The spectrum $\Delta_{\text{MM}}(\Gamma)$, by definition, should be the set of $(\theta_1, \theta_2, \theta_3)$ that the function defined by

$$\phi(a) = a(\theta_1, \theta_2, \theta_3)$$

satisfies all the axioms of an asymptotic-spectrum point. It is straightforward to verify that ϕ is nonnegative and satisfies (Semiring homomorphism) and (Normalization).

The only nontrivial condition is (Monotonicity). If $\Sigma \cup \Gamma \vdash a \leq b$, then we must have $a(\theta_1, \theta_2, \theta_3) \leq b(\theta_1, \theta_2, \theta_3)$. Since the derivation rules are compatible with passing to nonnegative real numbers, it suffices to enforce these inequalities for the generating set $\Sigma \cup \Gamma$.

The inequalities in Γ impose the region Δ_Γ , while the axioms in Σ impose:

- (Monotonicity) implies $\theta_i \geq 0$.
- (Trivial upper bound) implies $\theta_i \leq 1$.

Thus Σ forces $(\theta_1, \theta_2, \theta_3) \in [0, 1]^3$, and the claim follows. ◀

Define

$$\omega_\Gamma := \log_2(\mathbb{R}_{P(\Sigma \cup \Gamma)}(\langle 2, 2, 2 \rangle)).$$

Since Proposition 9 can also be checked to relativize⁴, we obtain the characterization

$$\omega_\Gamma = \max_{\Delta_{\text{MM}}(\Gamma)} (\theta_1 + \theta_2 + \theta_3).$$

On the one hand, ω_Γ is the optimum of $\theta_1 + \theta_2 + \theta_3$ over the feasible region $[0, 1]^3 \cap \Delta_\Gamma$. On the other hand, unwinding the definition yields the following “derivation-only” interpretation:

- Start from the axioms Σ together with the auxiliary identities Γ .
- Derive (in finitely many steps, using only (Preorder) and (Compatibility)) an inequality of the form $\langle n, n, n \rangle \leq r$.
- Conclude the bound $\omega \leq \log_n r$.
- Take the infimum over all such bounds (using $\omega = \inf_{n \geq 1} \log_n \mathbb{R}(\langle n, n, n \rangle)$ in this abstract setting).

This provides a precise meaning of “using only the identities in Γ ”. Under this notion, Strassen calculus is *complete*: the optimum over $[0, 1]^3 \cap \Delta_\Gamma$ is not merely an upper bound on ω_Γ ; it equals ω_Γ .

⁴ Strictly speaking, something breaks down for asymptotic subrank $\mathbb{Q}$, but we will not use it.

Asymptotic sum inequality. We are now ready to formulate a relativized version of the asymptotic sum inequality. There is, however, one remaining ingredient that we have implicitly used but have not yet formalized in $\mathcal{R}_{\text{MM}}$: the symmetry under permuting tensor modes.

Let a be an element

$$a = \sum_{n^{(1)}, n^{(2)}, n^{(3)}} a_{n^{(1)} n^{(2)} n^{(3)}} \langle n^{(1)}, n^{(2)}, n^{(3)} \rangle,$$

and let $\sigma \in \mathfrak{S}_3$ be a permutation. We write

$$a^\sigma := \sum_{n^{(1)}, n^{(2)}, n^{(3)}} a_{n^{(1)} n^{(2)} n^{(3)}} \langle n^{(\sigma^{-1}(1))}, n^{(\sigma^{-1}(2))}, n^{(\sigma^{-1}(3))} \rangle.$$

(Symmetry) For any inequality $a \leq b$ and any $\sigma \in \mathfrak{S}_3$, the permuted inequality $a^\sigma \leq b^\sigma$ also holds.

At first glance, (Symmetry) appears to be an additional derivation rule, which would force us to rebuild the preceding formalism. This is unnecessary: we can instead close the axiom set under permutations once and for all.

► **Lemma 46.** Let $\Gamma^\sigma := \{a^\sigma \leq b^\sigma : (a \leq b) \in \Gamma\}$, and let $\Gamma^\mathfrak{S} := \bigcup_{\sigma \in \mathfrak{S}_3} \Gamma^\sigma$. Then $P(\Sigma \cup \Gamma^\mathfrak{S})$ coincides with the preorder obtained from $\Sigma \cup \Gamma$ by allowing derivations using (Symmetry) in addition to (Preorder) and (Compatibility).

Proof. Let P' denote the preorder obtained by closing $\Sigma \cup \Gamma$ under derivations using (Preorder), (Compatibility), and (Symmetry). Since P' is closed under (Symmetry) by definition, it contains every permuted axiom in $\Gamma^\mathfrak{S}$, and hence $\Sigma \cup \Gamma^\mathfrak{S} \subseteq P'$. Therefore $P(\Sigma \cup \Gamma^\mathfrak{S}) \subseteq P'$.

Conversely, we claim that $P(\Sigma \cup \Gamma^\mathfrak{S})$ is closed under (Symmetry). Indeed, if $\Sigma \cup \Gamma^\mathfrak{S} \vdash a \leq b$, then there is a derivation of $a \leq b$ from axioms in $\Sigma \cup \Gamma^\mathfrak{S}$ using only (Preorder) and (Compatibility). Applying a fixed permutation $\sigma \in \mathfrak{S}_3$ to each line of the derivation yields a valid derivation of $a^\sigma \leq b^\sigma$, because $\Sigma \cup \Gamma^\mathfrak{S}$ is stable under permutation and the preorder/compatibility rules commute with permuting modes. Thus $P(\Sigma \cup \Gamma^\mathfrak{S})$ is closed under (Symmetry), so $P' \subseteq P(\Sigma \cup \Gamma^\mathfrak{S})$.

Combining the inclusions gives $P' = P(\Sigma \cup \Gamma^\mathfrak{S})$. ◀

Putting the preceding formalization together, we obtain a clean “complete” statement for the generalized asymptotic sum inequality in this relativized setting.

► **Proposition 47** (Generalized asymptotic sum inequality). Let Γ be a collection of inequalities in $\mathcal{R}_{\text{MM}}$. Allowing derivations using (Preorder), (Compatibility), and (Symmetry), the optimal bound on ω obtainable from Γ equals $\omega_{\Gamma^\mathfrak{S}}$. Equivalently, it is the maximum of $\theta_1 + \theta_2 + \theta_3$ over the feasible region $[0, 1]^3 \cap \Delta_{\Gamma^\mathfrak{S}}$.

B Analysis facts

► **Lemma 48.** Let $d \geq 3$ and $\varpi \geq 2$. Define

$$f(\theta) = d^{\varpi+\theta} + d^{2\theta} - (d^3 - d^2)^\theta, \quad \theta \in [0, \varpi/3].$$

Then f is strictly increasing on $[0, \varpi/3]$, hence

$$\max_{0 \leq \theta \leq \varpi/3} f(\theta) = f(\varpi/3).$$

Proof. Differentiate:

$$f'(\theta) = (\ln d) d^{\varpi+\theta} + 2(\ln d) d^{2\theta} - \ln(d^3 - d^2) (d^3 - d^2)^\theta.$$

Write $d^3 - d^2 = d^3(1 - 1/d)$ and set

$$\delta_d := -\frac{\ln(1 - 1/d)}{\ln d} > 0, \quad \text{so} \quad \ln(d^3 - d^2) = (3 - \delta_d) \ln d.$$

Also $(d^3 - d^2)^\theta = d^{3\theta}(1 - 1/d)^\theta \leq d^{3\theta}$. Hence

$$\frac{f'(\theta)}{\ln d} \geq d^{\varpi+\theta} + 2d^{2\theta} - (3 - \delta_d)d^{3\theta} = d^{3\theta} \left(d^{\varpi-2\theta} + 2d^{-\theta} - (3 - \delta_d) \right).$$

For $\theta \in [0, \varpi/3]$,

$$d^{\varpi-2\theta} + 2d^{-\theta} \geq d^{\varpi/3} + 2d^{-\varpi/3} \geq d^{2/3} + 2d^{-2/3},$$

where the last inequality uses that $x \mapsto d^x + 2d^{-x}$ is increasing for $x \geq \frac{\ln \sqrt{2}}{\ln d}$ and $\frac{\ln \sqrt{2}}{\ln d} < \frac{2}{3} \leq \varpi/3$. Therefore,

$$\frac{f'(\theta)}{\ln d} \geq d^{3\theta} (d^{2/3} + 2d^{-2/3} - 3).$$

Let $g(d) = d^{2/3} + 2d^{-2/3}$. Since

$$g'(d) = \frac{2}{3}d^{-5/3}(d^{4/3} - 2) > 0 \quad (d \geq 3)$$

and $g(3) = 3^{2/3} + 2 \cdot 3^{-2/3} > 3$, we have $d^{2/3} + 2d^{-2/3} - 3 > 0$ for all $d \geq 3$. Thus $f'(\theta) > 0$ on $[0, \varpi/3]$, so f is strictly increasing there and the maximum is attained at $\theta = \varpi/3$. ◀

► **Lemma 49.** Let $d \geq 3$. For $\varpi \geq 2$, define

$$F(\varpi) := d^{4\varpi/3} + d^{2\varpi/3} - (d^3 - d^2)^{\varpi/3}.$$

Then F is strictly increasing on $[2, 3]$.

Proof. Differentiate:

$$\frac{d}{d\varpi} F(\varpi) = \frac{4}{3}(\ln d) d^{4\varpi/3} + \frac{2}{3}(\ln d) d^{2\varpi/3} - \frac{1}{3} \ln(d^3 - d^2) (d^3 - d^2)^{\varpi/3}.$$

Write $d^3 - d^2 = d^3(1 - 1/d)$ and set

$$\delta_d := -\frac{\ln(1 - 1/d)}{\ln d} > 0, \quad \text{so that} \quad \ln(d^3 - d^2) = (3 - \delta_d) \ln d.$$

Moreover, $(d^3 - d^2)^{\varpi/3} = d^\varpi(1 - 1/d)^{\varpi/3} \leq d^\varpi$. Hence

$$\frac{3}{\ln d} \frac{d}{d\varpi} F(\varpi) \geq 4d^{4\varpi/3} + 2d^{2\varpi/3} - (3 - \delta_d)d^\varpi.$$

Factor out $d^{2\varpi/3}$ to obtain

$$\frac{3}{\ln d} \frac{d}{d\varpi} F(\varpi) \geq d^{2\varpi/3} \left(4d^{2\varpi/3} + 2 - (3 - \delta_d)d^{\varpi/3} \right).$$

For $\varpi \in [2, 3]$, we have $d^{\varpi/3} \in [d^{2/3}, d]$, hence

$$4d^{2\varpi/3} + 2 - (3 - \delta_d)d^{\varpi/3} \geq 4d^{4/3} + 2 - (3 - \delta_d)d.$$

Using $\delta_d > 0$ and $d \geq 3$, the right-hand side satisfies

$$4d^{4/3} + 2 - (3 - \delta_d)d > 4d^{4/3} + 2 - 3d = d(4d^{1/3} - 3) + 2 > 0.$$

Therefore $\frac{d}{d\varpi} F(\varpi) > 0$ for all $\varpi \in [2, 3]$, and F is strictly increasing on this interval. ◀

► **Lemma 50.** Let $d \geq 3$, $p \in \mathbb{R}_{\geq 0}$, and $\varpi \geq 2$. Define

$$h_{\varpi}(\theta) := d^{\varpi+\theta} - (2^{\theta} - 1)p d^{2\theta}, \quad \theta \in [0, \varpi/3].$$

If $p \leq \lfloor d/3 \rfloor$, then h_{ϖ} is strictly increasing on $[0, \varpi/3]$.

Proof. Differentiate:

$$h'_{\varpi}(\theta) = (\ln d) d^{\varpi+\theta} - p \frac{d}{d\theta} ((2^{\theta} - 1)d^{2\theta}).$$

Using

$$\frac{d}{d\theta} ((2^{\theta} - 1)d^{2\theta}) = d^{2\theta} ((\ln 2)2^{\theta} + 2(\ln d)(2^{\theta} - 1)),$$

we obtain

$$h'_{\varpi}(\theta) = d^{2\theta} ((\ln d) d^{\varpi-\theta} - p((\ln 2)2^{\theta} + 2(\ln d)(2^{\theta} - 1))).$$

For $\theta \in [0, \varpi/3]$ we have $d^{\varpi-\theta} \geq d^{2\varpi/3}$ and, since $\varpi \leq \omega < 3$ in our applications, also $\theta \leq 1$ and hence $2^{\theta} \leq 2$ and $2^{\theta} - 1 \leq 1$. Therefore,

$$h'_{\varpi}(\theta) \geq d^{2\theta} ((\ln d) d^{2\varpi/3} - p(2 \ln 2 + 2 \ln d)).$$

Using $p \leq \lfloor d/3 \rfloor \leq d/3$ and $\varpi \geq 2$ (so $d^{2\varpi/3} \geq d^{4/3}$), the bracket is at least

$$(\ln d) d^{4/3} - \frac{d}{3}(2 \ln 2 + 2 \ln d) = d \left((\ln d) d^{1/3} - \frac{2}{3}(\ln 2 + \ln d) \right).$$

For $d \geq 3$, we have $d^{1/3} > 1$ and thus

$$(\ln d) d^{1/3} - \frac{2}{3}(\ln 2 + \ln d) \geq (\ln 3) 3^{1/3} - \frac{2}{3}(\ln 2 + \ln 3) > 0.$$

Hence $h'_{\varpi}(\theta) > 0$ on $[0, \varpi/3]$, so h_{ϖ} is strictly increasing there. ◀

► **Lemma 51.** Let $d \geq 3$, $p \in \mathbb{R}_{\geq 0}$, and define for $\varpi \geq 2$,

$$g(\varpi) := d^{4\varpi/3} - (2^{\varpi/3} - 1)p d^{2\varpi/3}.$$

If $p \leq \lfloor d/3 \rfloor$, then g is strictly increasing on $[2, 3]$.

Proof. Differentiate:

$$g'(\varpi) = \frac{4}{3}(\ln d) d^{4\varpi/3} - \frac{p}{3} \frac{d}{d\varpi} ((2^{\varpi/3} - 1)d^{2\varpi/3}).$$

Since

$$\frac{d}{d\varpi} ((2^{\varpi/3} - 1)d^{2\varpi/3}) = d^{2\varpi/3} \left(\frac{\ln 2}{3} 2^{\varpi/3} + \frac{2 \ln d}{3} (2^{\varpi/3} - 1) \right),$$

we can write

$$g'(\varpi) = \frac{1}{3} d^{2\varpi/3} \left(4(\ln d) d^{2\varpi/3} - p((\ln 2)2^{\varpi/3} + 2(\ln d)(2^{\varpi/3} - 1)) \right).$$

For $\varpi \in [2, 3]$ we have $2^{\varpi/3} \leq 2$ and $2^{\varpi/3} - 1 \leq 1$, hence

$$g'(\varpi) \geq \frac{1}{3} d^{2\varpi/3} \left(4(\ln d) d^{2\varpi/3} - p(2 \ln 2 + 2 \ln d) \right).$$

36:42 Asymptotic Rank Speedup Theorems, Revisited

Using $p \leq d/3$ and $\varpi \geq 2$ (so $d^{2\varpi/3} \geq d^{4/3}$), it suffices to check

$$4(\ln d)d^{4/3} - \frac{d}{3}(2 \ln 2 + 2 \ln d) > 0.$$

This is equivalent to

$$d \left(4(\ln d)d^{1/3} - \frac{2}{3}(\ln 2 + \ln d) \right) > 0,$$

which holds for all $d \geq 3$ (e.g., already at $d = 3$ the bracket is positive, and it increases with d). Therefore $g'(\varpi) > 0$ on $[2, 3]$, and g is strictly increasing there. ◀