2018 IEEE 59th Annual Symposium on Foundations of Computer Science

Limits on All Known (and Some Unknown) Approaches to Matrix Multiplication

Josh Alman* and Virginia Vassilevska Williams†
MIT CSAIL and EECS
Cambridge, MA, USA
{*jalman, virgi*}*@mit.edu*

***Abstract*—We study the known techniques for designing Matrix Multiplication algorithms. The two main approaches are the Laser method of Strassen, and the Group theoretic approach of Cohn and Umans. We define a generalization based on zeroing outs which subsumes these two approaches, which we call the Solar method, and an even more general method based on monomial degenerations, which we call the Galactic method.**

We then design a suite of techniques for proving lower bounds on the value of ω, the exponent of matrix multiplication, which can be achieved by algorithms using many tensors T and the Galactic method. Some of our techniques exploit 'local' properties of T, like finding a sub-tensor of T which is so 'weak' that T itself couldn't be used to achieve a good bound on ω, while others exploit 'global' properties, like T being a monomial degeneration of the structural tensor of a group algebra.

Our main result is that there is a universal constant $\ell > 2$ such that a large class of tensors generalizing the Coppersmith-Winograd tensor CW_q cannot be used within the Galactic method to show a bound on ω better than ℓ, for any q. We give evidence that previous lower-bounding techniques were not strong enough to show this. We also prove a number of complementary results along the way, including that for any group G, the structural tensor of $\mathbb{C}[G]$ can be used to recover the best bound on ω which the Coppersmith-Winograd approach gets using $CW_{|G|-2}$ as long as the asymptotic rank of the structural tensor is not too large.

***Keywords*-matrix multiplication; lower bound; Coppersmith-Winograd tensor; monomial degeneration;**

I. Introduction

A fundamental problem in theoretical computer science is to determine the time complexity of Matrix Multiplication (MM), one of the most basic linear algebraic operations. The question typically translates to determining the *exponent of matrix multiplication*: the smallest real number ω such that the product of two $n \times n$ matrices over a field $\mathbb{F}$ can be determined using $n^{\omega+o(1)}$ operations over $\mathbb{F}$. Trivially, $2 \leq \omega \leq 3$. Many have conjectured over the years that $\omega = 2$. This conjecture is extremely attractive: a near-linear time algorithm for MM would immediately imply near-optimal algorithms for many problems.

Almost 50 years have passed since Strassen [1] first showed that $\omega \leq 2.81 < 3$. Since then, an impressive toolbox of techniques has been developed to obtain faster MM algorithms, culminating in the current best bound $\omega < 2.373$ [2], [3]. Unfortunately, this bound is far from 2, and the current methods seem to have reached a standstill. Recent research has turned to proving limitations on the two main MM techniques: the Laser method of Strassen [4] and the Group theoretic method of Cohn and Umans [5].

Both Coppersmith and Winograd [6] and Cohn et al. [7] proposed conjectures which, if true, would imply that $\omega = 2$. The first conjecture works in conjunction with the Laser method, and the second with the Group-theoretic method. The first "technique limitation" result was by Alon, Shpilka and Umans [8] who showed that both conjectures would contradict the widely believed Sunflower conjecture of Erdös and Rado.

Ambainis, Filmus and Le Gall [9] formalized the specific implementation of the Laser method proposed by Coppersmith and Winograd [6] which is used in the recent papers on MM. They gave limitations of this implementation, and in particular showed that the exact approach used in [6], [10], [2], [3] cannot achieve a bound on ω better than 2.3078. The analyzed approach, the "Laser Method with Merging", is a bit more general than the approaches in [6], [10], [2], [3]: in a sense it corresponds to a dream implementation of the exact approach.

Blasiak et al. [11] considered the group theoretic framework for developing MM algorithms proposed by Cohn and Umans [5], and showed that this approach cannot prove $\omega = 2$ using any fixed abelian group. In follow-up work, Sawin [12] extended this to any fixed non-abelian group, and Blasiak et al. [13] extended it

*Supported by two NSF Career Awards.
†Partially supported by an NSF Career Award, a Sloan Fellowship, NSF Grants CCF-1417238, CCF-1528078 and CCF-1514339, and BSF Grant BSF:2012338.

2575-8454/18/$31.00 ©2018 IEEE
DOI 10.1109/FOCS.2018.00061

to a host of families of non-abelian groups.

Alman and Vassilevska W. [14] considered a generalization of the Laser method and proved limitations on this generalization when it is applied to any tensor which is a *monomial degeneration* of the structure tensor of the group algebra $\mathbb{C}[C_q]$ of the cyclic group C_q of order q. (See Section III for the definitions.) The bounds on ω achieved by known implementations of the Laser method [4], [6], [10], [2], [3] can all be obtained from tensors of this form. The formalization also subsumes the group theoretic approach applied to C_q. The main result of [14] is that this generalized approach cannot achieve $\omega = 2$ for any fixed q.

All limitations proven so far, however, suffer from several weaknesses:

- All three of [11], [13] and [14] show how *some* approach that can yield the current best bounds on ω cannot give $\omega = 2$. None of the three works actually prove that one cannot use the particular tensor CW_q used in recent work [6], [10], [3], [2] to show $\omega = 2$. [14] proved this limitation for a rotated version of CW_q, but only for small q. Although [11] and [13] do not say which version their proofs apply to, in this paper we give evidence that CW_q does not embed easily in a group tensor. Moreover, even for the Coppersmith-Winograd-like tensors for which the known limitations do apply, it is only shown that for a *fixed* q one cannot derive $\omega = 2$. In particular, the lower bounds ω_q on the ω one can achieve for a value q approached 2. This left open the possibility to prove $\omega = 2$ by analyzing CW_q in the limit as $q \to \infty$.
- All limitations proven so far are for very specific attacks on proving $\omega = 2$. While the proofs of [9] apply directly to CW_q, they only apply to the restricted Laser Method with Merging, and no longer apply to slight changes to this. The proofs in [11] and [13] are tailored to the group theoretic approach and do not apply (for instance) to the Laser method on "non-group" tensors. While the limits in [14] do apply to a more general method than both the group theoretic approach and the Laser method, they only work for specific types of tensors, which in particular do not include CW_q.

Our Results.

All known approaches to matrix multiplication follow the following outline. First, obtaining a bound on ω corresponds to determining the *asymptotic rank* of the matrix multiplication tensor $\langle n, n, n\rangle$ (see the Preliminaries for a formal definition). Because getting a handle on this asymptotic rank seems difficult, one typically works with a tensor t (or a tensor family) whose asymptotic rank r is known. Then, to analyze the asymptotic rank of matrix multiplication, one considers large tensor powers $t^{\otimes n}$ of t and attempts to embed $\langle N, N, N\rangle$ into $t^{\otimes n}$ for large N. In effect, one is showing that the recursive $O(r^n)$ time algorithm for computing $t^{\otimes n}$ can be used to multiply $N \times N$ matrices. This gives a bound on ω from $N^\omega \leq r^n$. The larger N is in terms of n, the smaller the bound on ω.

Now, an embedding of matrix multiplication into a tensor power $t^{\otimes n}$ produces a new tensor from t, and to get a bound on ω, this tensor should ideally have asymptotic rank no more than r^n. The most general type of embedding that preserves asymptotic rank is a so called *degeneration* of the tensor $t^{\otimes n}$. A more restricted type of rank-preserving embedding is a so called *monomial degeneration*. The embeddings used in all known approaches for upper bounding ω so far are even more restricted *zeroing outs*. The laser method is a restricted type of zeroing out that has only been applied so far to tensors that look like matrix multiplication tensors or to ones related to the Coppersmith-Winograd tensor. The group theoretic approach gives clean definitions that imply the existence of a zeroing out of a group tensor into a matrix multiplication tensor. (See the preliminaries for formal definitions.)

We define three very general methods of analyzing tensors. There are *no known* techniques to analyze tensors in this generality.

- The **Solar** Method applied to a tensor t of asymptotic rank r considers $t^{\otimes n}$ for large n, finds the largest N for which there is a *zeroing out* of $t^{\otimes n}$ into $\langle N, N, N\rangle$, and gives a bound $\omega \leq n \log_N r$. This method already subsumes both the group theoretic method and the laser method. It is also much more general, as it is unclear whether the two known techniques produce the best possible zeroing outs even for specific tensors.
- The **Galactic** Method applied to a tensor t of asymptotic rank r considers $t^{\otimes n}$ for large n, finds the largest N for which there is a *monomial degeneration* of $t^{\otimes n}$ into $\langle N, N, N\rangle$, and gives a bound $\omega \leq n \log_N r$.
- The **Universal** Method applied to a tensor t of asymptotic rank r considers $t^{\otimes n}$ for large n, finds the largest N for which there is **some** *degeneration* of $t^{\otimes n}$ into $\langle N, N, N\rangle$, and gives a bound $\omega \leq n \log_N r$.

We note that the methods only differ when they are applied to the same tensor t. Trivially, any one of the methods can find the best bound on ω if it is "applied"

to $t = \langle n, n, n \rangle$ itself. Starting with the same tensor t, however, the Universal method can in principle give much better bounds on ω than the Solar or Galactic methods applied to the same t.

For a tensor T, let $\omega_g(T)$ be the best bound on ω that one can obtain by applying the Galactic method to T. We define a class of *generalized* CW_q tensors that contain CW_q and many more tensors related to it, such as the rotated tensor used in [14]. Our **main result** is:

Theorem I.1 (Informal)**.** *There is a universal constant $\ell > 2$ independent of q so that for every one of the generalized CW_q tensors T, $\omega_g(T) \geq \ell$.*

Thus, if one uses a generalized CW tensor, even in the limit and even if one uses the Galactic method subsuming all known approaches, one cannot prove $\omega = 2$.

To prove this result, we develop several tools for proving lower bounds on $\omega_g(T)$ for structured tensors. Most are relatively simple combinatorial arguments but are still powerful enough to show strong lower bounds on $\omega_g(T)$.

We also study the relationship between the generalized CW tensors and the structure tensors of group algebras. We show several new results:

1) **A Limit on the Group-Theoretic Approach.** The original CW_q tensor is not a sub-tensor (and hence also not a monomial degeneration) of the structure tensor T_G of $\mathbb{C}[G]$ for any G of order $< 2q$ when (a) G is abelian and q arbitrary, or (b) G is non-abelian and $q \in \{3, 4, 5, 6, 7, 8, 9\}$. Note that CW_q for these small values of q are of particular interest: the best known bounds on ω have been proved using $q < 7$. This shows that lower bound techniques based on tri-colored sum-free sets and group tensors cannot be easily applied to CW_q.
2) **All Finite Groups Suffice for Current ω Bounds.** Every finite group G has a monomial degeneration to some generalized CW tensor of parameter $q = |G| - 2$. Thus, applying the Galactic method on T_G for *every* G (with sufficiently small asymptotic rank) can yield the current best bounds on ω.
3) **New Tri-Colored Sum-Free Set Constructions.** For every finite group G, there is a constant $c_{|G|} > 2/3$ depending only on $|G|$ such that its nth tensor power G^n has a tri-colored sum-free set of size at least $|G|^{c_{|G|}n - o(n)}$. For moderate $|G|$, the constant $c_{|G|}$ is quite a bit larger than $2/3$. To our knowledge, such a general result was not known until now.

For more details on our results, see Section II below.

II. Overview of Results and Proofs

In this section, we give an outline of our techniques which are used to prove our main result: that there exists a universal constant $c > 2$ such that the Galactic method, when applied to any generalized Coppersmith-Winograd tensor, cannot prove a better upper bound on ω than c. We will assume familiarity with standard notions and notation about tensors related to matrix multiplication algorithms in this section; we refer the reader to the Preliminaries, in Section III, where these are defined. For a tensor T, we will write $\omega_g(T)$ to denote the best upper bound on ω which can be achieved using the Galactic method applied to T.

Step 1: The Relationship Between Matrix Multiplication and Independent Tensors.

In Section IV, we begin by laying out the main framework for proving lower bounds on $\omega_g(T)$. The key is to consider a different property of T, the *asymptotic independence number of* T, denoted $\tilde{I}(T)$. Loosely, $\tilde{I}(T)$ gives a measure of how large of an independent tensor $T^{\otimes n}$ can monomial degenerate into for large n. From the definition, we will get a simple upper bound $\tilde{I}(T) \leq \tilde{R}(T)$, the *asymptotic rank* of T. By constructing upper bounds on $\tilde{I}(T)$, we will show in Corollary IV.1 that:

- For any tensor T, if $\omega_g(T) = 2$, then $\tilde{I}(T) = \tilde{R}(T)$, and moreover,
- For every constant $s < 1$, there is a constant $w > 2$ (which is increasing as s decreases), such that if $\tilde{I}(T) < \tilde{R}(T)^s$, then $\omega_g(T) \geq w$.

Hence, upper bounds on $\tilde{I}(T)$ give lower bounds on $\omega_g(T)$. We will thus present a number of different ways to prove upper bounds on $\tilde{I}(T)$ in the next steps.

Step 2: Partitioning Tools for Upper Bounding $\tilde{I}$.

In Section V, we present our first suite of tools for proving upper bounds on $\tilde{I}(T)$. These tools are based on finding 'local' combinatorial properties of the tensor T which imply that $\tilde{I}(T)$ can't be too large. They are loosely summarized as follows:

- Theorem V.1: Let S be any subset of the x-variables of T, and let A be the tensor T restricted to S (i.e. T with all the variables in $X \setminus S$ zeroed out). If $\tilde{I}(A)$ is sufficiently smaller than $|S|$, then $\tilde{I}(T) < |X| \leq \tilde{R}(T)$.
  In other words, if A has a sufficiently small $\tilde{I}(A)$ so that it is relatively far away from being able to prove $\omega_g(A) = 2$, then no matter how we complete A to get to T, the tensor T will still not be able to prove $\omega_g(T) = 2$.

- Theorem V.2: If T is a tensor such that $\tilde{I}(T)$ is close to $\tilde{R}(T)$, then there is a probability distribution on the terms of T such that each X, Y, and Z variable is assigned almost the same probability mass.
  For many tensors of interest, one or more of the variables 'behave differently' from the rest, and this can be used to prove that such a probability distribution cannot exist. For one example, we prove in Corollary V.1 that if T is a tensor with two 'corner terms' – terms $x_q y_1 z_1, x_1 y_q z_1 \in T$ such that no other term in T contains either x_q or y_q – then, $\tilde{I}(T) < \tilde{R}(T)$.
  These 'corner terms' are actually quite common in tensors which have been analyzed with the Laser Method. For instance, one of the main improvements of Coppersmith-Winograd [6] over Strassen [4] was noticing that the border rank expression of Strassen could be augmented by adding in three corner terms, resulting in the Coppersmith-Winograd tensor.
- Theorem V.3: For a tensor T over variables X, Y, Z, where each of these variables appears in the support of T, we define the *measure* of T, denoted $\mu(T)$, by $\mu(T) := |X| \cdot |Y| \cdot |Z|$. Suppose the terms of T can be partitioned[1] into tensors $T_1, \ldots, T_k$. Then, $\tilde{I}(T) \leq (\mu(T_1))^{2/3} + \cdots + (\mu(T_k))^{2/3}$.
  This gives a generalization of the basic inequality that $\tilde{I}(T) \leq \min\{|X|, |Y|, |Z|\}$. Whenever T can be partitioned up into parts which each do not have many of one or more type of variable, we can get a nontrivial upper bound on $\tilde{I}(T)$. Many natural border rank expressions naturally give rise to such partitions, as do the 'blockings' used in the Laser method.

As we will see, $\tilde{I}$ is neither additive nor multiplicative, i.e. there are tensors A and B such that $\tilde{I}(A+B) \gg \tilde{I}(A)+\tilde{I}(B)$, and tensors C and D such that $\tilde{I}(C \otimes D) \gg \tilde{I}(C) \cdot \tilde{I}(D)$. One of the main components of the proofs of correctness of each of the three tools above will be narrowing in on classes of tensors A and B such that $\tilde{I}(A+B)$ is not too much greater than $\tilde{I}(A) + \tilde{I}(B)$, or classes of tensors C and D such that $\tilde{I}(C \otimes D)$ is not too much greater than $\tilde{I}(C) \cdot \tilde{I}(D)$. Our proofs will then manipulate our tensors using partitionings so that they fall into these classes.

The Main Result.

The three partitioning tools are designed to be useful for proving nontrivial upper bounds on $\tilde{I}$ for general classes of tensors. They are especially well-suited to tensors which have structures that make them amenable to known techniques like the Laser Method. In particular, we will ultimately show that any generalized Coppersmith-Winograd tensor has *all three* of these properties. Indeed, our main result, Theorem VII.1, follows from these tools: For any generalized CW tensor T, a lower bound on $\omega_g(T)$ for small q will follow from Corollary V.1, and a lower bound on $\omega_g(T)$ as q gets large (but such that the bound gets larger as q increases, not smaller) will follow from either Theorem V.1 or Theorem V.3.

Bounds on $\tilde{I}$ for Group Tensors.

In addition to the above, we also study group tensors. For a finite group G, we call the structural tensor $T_{\mathbb{C}[G]}$ of the group algebra $\mathbb{C}[G]$ the *group tensor* T_G *of* G. We are able to achieve both nontrivial upper bounds and lower bounds on $\tilde{I}(T_G)$ for any finite group G, including non-abelian groups.

Upper Bounds on $\tilde{I}(T_G)$.

We first show that for any finite group G, we have $\tilde{I}(T_G) < |G| \leq \tilde{R}(T_G)$, and hence $\omega_g(T_G) > 2$. In other words, no fixed group G can yield $\omega = 2$ by using the Galactic method applied to T_G. By comparison, the Group Theoretic approach for G can be viewed as analyzing T_G using a particular technique within the Solar method (see Section III-D for more details). This therefore generalizes a remark which is already known within the Group Theoretic community [13]: that the Group Theoretic approach (using the so-called 'Simultaneous Triple Product Property') cannot yield $\omega = 2$ using any fixed finite group G. It does not, however, rule out using a sequence of groups whose lower bounds approach 2.

Our proof begins by proving a generalization of a remark from [14]: that lower bounds on $\tilde{I}(T_G)$ give rise to constructions of 'tri-colored sum-free sets' in G^n for sufficiently large integer n ([14] proved this when G is a cyclic group, although our proof is almost identical). Tri-colored sum-free sets are objects from extremal combinatorics which have been studied extensively recently. We will, in particular, use a recent result of Sawin [12], who showed that for any finite group G, there is a sufficiently large n such that G^n does not have particularly large tri-colored sum-free sets.

We give this proof in Section VI. In that section, we also show that there are natural tensors, like the Coppersmith-Winograd tensors used to give the best known upper bounds on ω, which *cannot* even be written as sub-tensors of relatively small group tensors. In other

[1]We mean 'partitioned' as in a set partition, not any restricted notion like the 'block partitions' of the Laser Method.

words, the high-powered hammer that $\omega_g(T_G) > 2$ cannot be used to give lower bound for every tensor of interest, and other techniques like the combinatorial partitioning techniques from step 2 above are needed.

Lower Bounds on $\tilde{I}(T_G)$

Although our main framework involves proving upper bounds on $\tilde{I}(T)$ for tensors T in order to prove lower bounds on $\omega_g(T)$, step 1 of our proof actually involves constructing lower bounds on $\tilde{I}(T)$ when T has a monomial degeneration to a matrix multiplication tensor. In the full version of this paper, we use this to give lower bounds on $\tilde{I}(T_G)$ for any finite group G.

We show there that for *any* finite group G, there is a monomial degeneration of T_G into a *generalized* Coppersmith-Winograd tensor of parameter $|G| - 2$. We will see that the Laser method applies just as well to any generalized Coppersmith-Winograd tensor of parameter $|G| - 2$ as it does to the original $CW_{|G|-2}$, and so the best-known approach for finding matrix multiplication tensors as monomial degenerations of a tensor can be applied to *any* group tensor T_G as well. Two important consequences of this are:

1) For any group G such that $\tilde{R}(T_G) = |G|$, we can use the Galactic method to achieve the best known upper bound on ω (that is known from $CW_{|G|-2}$) by using T_G as the underlying tensor instead of the Coppersmith-Winograd tensor. We think this has exciting prospects for designing new matrix multiplication algorithms; see the full version of this paper for further discussion about this.
2) Once T_G has been monomial degenerated into a Coppersmith-Winograd tensor, and thus a matrix multiplication tensor, we can then apply the tools from step 1 above to show that T_G has a monomial degeneration to a relatively large independent tensor. In particular, we show that for any group G, $\tilde{I}(T_G) \geq |G|^{c_{|G|}}$ for some constant $c_{|G|} > 2/3$ which depends only on $|G|$. Combining this with the connection between $\tilde{I}(T_G)$ and tri-colored sum-free sets in G, we see that for any finite group G, G^n has a tri-colored sum-free set of size at least $|G|^{c_{|G|}n-o(n)}$. See the full version of this paper for the details. We will find that $c_{|G|}$ is much bigger than $2/3$ for reasonable $|G|$; for instance, that $c_{|G|} > 3/4$ for $|G| < 250$.

A. Comparison with Full Version

In this extended abstract, we focus on proving our main result, Theorem I.1. The full version of this paper contains all the additional results, as well all the proof details which are omitted here, and a more detailed preliminaries section.

III. Preliminaries

A. Tensor Notation and Definitions

Let $X = \{x_1, \ldots, x_q\}$, $Y = \{y_1, \ldots, y_r\}$, and $Z = \{z_1, \ldots, z_s\}$ be three sets of formal variables. A *tensor over X, Y, Z* is a trilinear form

$$T = \sum_{x_i \in X, y_j \in Y, z_k \in Z} T_{ijk} x_i y_j z_k,$$

where the T_{ijk} coefficients come from an underlying field $\mathbb{F}$. One writes $T \in \mathbb{F}^q \otimes \mathbb{F}^r \otimes \mathbb{F}^s$, and the triads $x_i y_j z_k$ are typically written $x_i \otimes y_j \otimes z_k$; we omit the $\otimes$ for ease of notation. When X, Y, and Z are clear from context, we will just call T a *tensor*. The *support* of a tensor T are all triples (i, j, k) for which $T_{ijk} \neq 0$. The *size* of a tensor T, denoted $|T|$, is the size of its support. We will write $x_i y_j z_k \in T$ to denote that (i, j, k) is in the support of T, and in this case we call $x_i y_j z_k$ a *term* of T. We will call elements of X the 'x-variables of T', and similarly for Y and Z.

If $A \in \mathbb{F}^k \otimes \mathbb{F}^m \otimes \mathbb{F}^n$ and $B \in \mathbb{F}^{k'} \otimes \mathbb{F}^{m'} \otimes \mathbb{F}^{n'}$, then the *tensor product of A and B*, denoted $A \otimes B$, is a tensor in $\mathbb{F}^{k \times k'} \otimes \mathbb{F}^{m \times m'} \otimes \mathbb{F}^{n \times n'}$ over new variables $\bar{X}, \bar{Y}, \bar{Z}$ given by

$$A \otimes B = \sum_{\substack{(i,i') \in [k] \times [k'] \\ (j,j') \in [m] \times [m'] \\ (k,k') \in [n] \times [n']}} A_{ijk} B_{i'j'k'} \bar{x}_{ii'} \bar{y}_{jj'} \bar{z}_{kk'}.$$

The nth tensor power of a tensor A, denoted $A^{\otimes n}$, is the result of tensoring n copies of A together, so $A^{\otimes 1} = A$, and $A^{\otimes n} = A \otimes A^{\otimes n-1}$.

Intuitively, if A is over X, Y, Z and B is over X', Y', Z', then the variables $\bar{x}_{ii'}, \bar{y}_{jj'}, \bar{z}_{kk'}$ of $A \otimes B$ can be viewed as pairs of the original variables $(x_i, x'_{i'})(y_j, y'_{j'})(z_k, z'_{k'})$. We will use this view in some of our proofs. For instance, when considering $A^{\otimes n}$ we will often view the x,y and z variables of $A^{\otimes n}$ as ordered n-tuples of x,y and z variables of A. Then we can discuss for instance, in how many positions of an x variable of $A^{\otimes n}$, the variable x_i of A appears.

1) Tensor Rank: A tensor T has *rank one* if there are values $a_i \in \mathbb{F}$ for each $x_i \in X$, $b_j \in \mathbb{F}$ for each $y_j \in Y$, and $c_k \in \mathbb{F}$ for each $z_k \in Z$, such that $T_{ijk} = a_i b_j c_k$, or in other words,

$$\begin{aligned} T &= \sum_{x_i \in X, y_j \in Y, z_k \in Z} a_i b_j c_k \cdot x_i y_j z_k \\ &= \left(\sum_{x_i \in X} a_i x_i \right) \left(\sum_{y_j \in Y} b_j y_j \right) \left(\sum_{z_k \in Z} c_k z_k \right). \end{aligned}$$

More generally, the *rank of* T, denoted $R(T)$, is the smallest nonnegative integer m such that T can be written as the sum of m rank-one tensors.

Let λ be a formal variable, and suppose T is a tensor over X, Y, Z. The *border rank of* T, denoted[2] $\bar{R}(T)$, is the smallest r such that there is a tensor $\mathcal{T}$ with coefficients $\mathcal{T}_{ijk}$ in $\mathbb{F}[\lambda]$ (polynomials in λ), so that for every setting of $\lambda \in \mathbb{F}$, $\mathcal{T}$ evaluated at λ has rank r, and so that there is an integer $h \geq 0$ for which:

$$\lambda^h T = \mathcal{T} + O(\lambda^{h+1}).$$

The above notation means that for every i, j, k, the polynomial $\mathcal{T}_{ijk}$ over λ has no monomials with λ^j with $j < h$, and the coefficient in front of λ^h in $\mathcal{T}_{ijk}$ is exactly T_{ijk}. In a sense, the family of rank r tensors $\mathcal{T}\lambda^{-h}$ for $\lambda \neq 0$ can get arbitrarily close to T – if $\mathbb{F} = \mathbb{R}$, then we could think of taking $\lambda \to 0$ and then $\mathcal{T}\lambda^{-h} \to T$.

The *asymptotic rank* of a tensor T is defined as $\tilde{R}(T) := \lim_{n\to\infty}(R(T^{\otimes n}))^{1/n}$. The limit exists and equals $\inf_{n\to\infty}(R(T^{\otimes n}))^{1/n}$. It is known that for any tensor T,

$$R(T) \geq \bar{R}(T) \geq \tilde{R}(T),$$

and that each of these inequalities can be strict[3]. One of the most common ways to show asymptotic rank upper bounds is to give border rank upper bounds, frequently using a tool called a 'monomial degeneration' which we will define shortly.

The tensor $\langle r \rangle$ in $\mathbb{F}^r \otimes \mathbb{F}^r \otimes \mathbb{F}^r$ is defined as follows: for all $i \in \{1, \ldots, r\}$, $\langle r \rangle_{i,i,i} = 1$ and for all other entries $\langle r \rangle_{i,j,k} = 0$. $\langle r \rangle$ clearly has rank r; it is the natural generalization of an identity matrix. If a tensor T is equivalent to $\langle r \rangle$ up to permutation of the indices, we say that T is an independent tensor of size $|T| = r$.

2) Sub-Tensors and Degenerations: We call a tensor t a *sub-tensor* of a tensor t', denoted by $t \subseteq t'$, if t can be obtained from t' by removing triples from its support, i.e. for every i, j, k, either $t_{i,j,k} = t'_{i,j,k}$, or $t_{i,j,k} = 0$.

A tensor $t \in \mathbb{F}^k \otimes \mathbb{F}^m \otimes \mathbb{F}^n$ is a *restriction* of a tensor $t' \in \mathbb{F}^{k'} \otimes \mathbb{F}^{m'} \otimes \mathbb{F}^{n'}$, written $t \leq t'$, if there are homomorphisms $\alpha : \mathbb{F}^k \mapsto \mathbb{F}^{k'}$, $\beta : \mathbb{F}^m \mapsto \mathbb{F}^{m'}$, and $\gamma : \mathbb{F}^n \mapsto \mathbb{F}^{n'}$, so that $t = (\alpha \otimes \beta \otimes \gamma)t'$.[4] The rank of t is $\leq r$ if and only if $t \leq \langle r \rangle$.

A special type of restriction is the so called *zeroing out* (also called *combinatorial restriction*): let t be a tensor over X, Y, Z; t' is a zeroing out of t if it is obtained by selecting $X' \subseteq X, Y' \subseteq Y, Z' \subseteq Z$ and setting to zero all $x_i \in X \setminus X', y_j \in Y \setminus Y', z_k \in Z \setminus Z'$; thus, t' is a tensor over X', Y', Z' and it equals t on all triples over these sets.

A *degeneration* $t' \in \mathbb{F}^{k'} \otimes \mathbb{F}^{m'} \otimes \mathbb{F}^{n'}$ of a tensor $t \in \mathbb{F}^k \otimes \mathbb{F}^m \otimes \mathbb{F}^n$, written $t' \trianglelefteq t$, is obtained as follows. Similarly to the definition of border rank, let λ be a formal variable. We say that $t' \trianglelefteq t$ if there exist $q \in \mathbb{N}$, $A(\lambda) \in \mathbb{F}^{k'\times k}, B(\lambda) \in \mathbb{F}^{m'\times m}, C(\lambda) \in \mathbb{F}^{n'\times n}$ matrices with entries which are polynomials in λ (i.e. in $\mathbb{F}[\lambda]$), so that

$$\lambda^q t' = (A(\lambda) \otimes B(\lambda) \otimes C(\lambda))t + O(\lambda^{q+1}).$$

Similarly to the relationship between rank and restriction, the border rank of t is at most r if and only if $t \trianglelefteq \langle r \rangle$.

A special type of degeneration is the so called *monomial degeneration* (also called *combinatorial degeneration* or *toric degeneration*), in which the matrices $A(\lambda), B(\lambda), C(\lambda)$ have entries that are monomials in λ. An equivalent definition of monomial degeneration [14] is as follows: suppose that t' is a tensor over $\mathbb{F}^k \otimes \mathbb{F}^m \otimes \mathbb{F}^n$, $t \subseteq t'$ is a sub-tensor, and there are functions $a : [k] \to \mathbb{Z}$, $b : [m] \to \mathbb{Z}$, and $c : [n] \to \mathbb{Z}$ such that (1) whenever $t'_{ijk} \neq 0$, $a(i) + b(j) + c(k) \geq 0$, (2) if $a(i) + b(j) + c(k) = 0$, then $t_{i,j,k} = t'_{i,j,k}$, and (3) if $t_{ijk} \neq 0$, then $a(i) + b(j) + c(k) = 0$.

3) Structural Properties of Tensors: We say that a tensor T is *partitioned* into tensors $T_1, \ldots, T_\ell$, if $T = T^1 + \ldots + T^\ell$, and for every triple i, j, k, there is a w such that $T^w_{i,j,k} = T_{i,j,k}$ and for all $w' \neq w$, $T^w_{i,j,k} = 0$. In other words, the triples in the support of T are partitioned into ℓ parts, forming ℓ tensors summing to T.[5]

A *direct sum* of two tensors t and t' over disjoint variable sets X, Y, Z and X', Y', Z', $t \oplus t'$ is the tensor on variable sets $X \cup X', Y \cup Y', Z \cup Z'$ which is exactly t on triples in $X \times Y \times Z$, exactly t' on triples in $X' \times Y' \times Z'$, and is 0 on all other triples. In contrast, a regular sum $t + t'$ could have t and t' share variables.

All the explicit tensors t we will discuss throughout this paper, including the tensor of matrix multiplication, and the Coppersmith-Winograd tensor, are *concise*, which implies that $\bar{R}(t) \geq \max\{|X|, |Y|, |Z|\}$. We defer to the full version of this paper for the technical definition of a concise tensor.

[2]Much of the literature uses $\underline{R}$ for border rank; we instead use $\bar{R}$ for ease of notation.

[3]For example, the first inequality is strict for the Coppersmith-Winograd tensor, and the second inequality is strict for the $2 \times 2 \times 2$ matrix multiplication tensor. Both of these tensors will be defined shortly.

[4]The notation $(\alpha \otimes \beta \otimes \gamma)t$ means the following. Let $t = \sum_{\ell=1}^r (\sum_i a_i^\ell x_i)(\sum_j b_j^\ell y_j)(\sum_k c_k^\ell z_k) = \sum_{\ell=1}^r (a^\ell \cdot x)(b^\ell \cdot y)(c^\ell \cdot z)$ be any decomposition of t into a sum of rank 1 tensors, where $a^\ell = (a_1^\ell, \ldots, a_k^\ell) \in \mathbb{F}^k, b^\ell = (b_1^\ell, \ldots, b_m^\ell) \in \mathbb{F}^m, c^\ell = (c_1^\ell, \ldots, c_n^\ell) \in \mathbb{F}^n$. Then $(\alpha \otimes \beta \otimes \gamma)t := \sum_{\ell=1}^r (\alpha(a^\ell) \cdot x)(\beta(b^\ell) \cdot y)(\gamma(c^\ell) \cdot z)$ is well-defined.

[5]Note that this notion of partitioning is more general than 'block partitioning' from the Laser Method (which we define shortly), although 'block partitioning' is occasionally referred to as just 'partitioning' in the literature.

B. The Matrix Multiplication Tensor and Methods for Analyzing ω

Let $m, n, p \geq 1$ be integers. The tensor of $m \times n$ by $n \times p$ matrix multiplication over a field $\mathbb{F}$, denoted by $\langle m, n, p \rangle$, lies in $\mathbb{F}^{m \times n} \otimes \mathbb{F}^{n \times p} \otimes \mathbb{F}^{p \times m}$, and in trilinear notation looks like this:

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

The theory of matrix multiplication algorithms is concerned with determining the value ω, defined as $\omega := \inf\{c \in \mathbb{R} \mid R(\langle n, n, n \rangle) \leq O(n^c)\}$. (As shown by Coppersmith and Winograd [15], ω is a limit point that cannot be achieved by any single algorithm.)

Getting a handle on ω has been difficult. Over the years various methods have been developed to obtain better understanding of the rank of $\langle n, n, n \rangle$. The basic idea of all methods is as follows: Although we do not know what the true rank of $\langle n, n, n \rangle$ is, as n grows, there are many other tensors for which we know their rank and even their asymptotic rank exactly. Hence, the approach is, take a tensor t whose asymptotic rank $\tilde{R}(t)$ we understand, take a large tensor power $t^{\otimes N}$ of t, and "embed" $\langle f(N), f(N), f(N) \rangle$ into $t^{\otimes N}$ so that the embedding shows that $\tilde{R}(\langle f(N), f(N), f(N) \rangle) \leq \tilde{R}(t)^N$. From this inequality we can get a bound on ω.

The way in which the approaches differ is mainly in how the embedding into $t^{\otimes N}$ is obtained. All known approaches to embed a matrix multiplication tensor into a tensor power $t^{\otimes N}$ of some other tensor t actually all zero out variables in $t^{\otimes N}$ and argue that after the zeroing out, the remaining tensor is a matrix multiplication tensor.

There are two main approaches for obtaining good bounds on ω via zeroing out $t^{\otimes N}$: the laser method and the group theoretic approach. We will describe them both shortly.

Zeroing out is a very restricted border-rank preserving operation on a tensor. The most general embedding of a matrix multiplication tensor into $t^{\otimes N}$ would be a potentially complicated degeneration of $t^{\otimes N}$. In fact, in this case, since every border rank q tensor is a degeneration[6] of the structure tensor for addition modulo q, $T_q = \sum_{i=0}^{q-1} \sum_{j=0}^{q-1} x_i y_j z_{i+j \bmod q}$, it would suffice to find a degeneration of $T_q^{\otimes n}$ into a large matrix multiplication tensor, for large n. Unfortunately, we currently do not have techniques to find good degenerations. We call this hypothetical method the **Universal** method.

Instead of considering arbitrary degenerations of $t^{\otimes n}$, we could instead consider monomial degenerations of $t^{\otimes n}$ into a large matrix multiplication tensor. This approach would subsume both the Laser Method and the Group Theoretic approach. Although again there are no known techniques to obtain better monomial degenerations than zeroing outs, monomial degenerations seem easier to argue about than arbitrary degenerations. We call the method of finding the optimal (with respect to bounding ω) monomial degeneration of a tensor power into a matrix multiplication tensor, the **Galactic** method. (Reaching the end of our Galaxy is more feasible than seeing the entire Universe.) To complete the analogy, we can call the method using zeroing outs the **Solar** method (i.e. exploring the Solar System).

The Solar method subsumes the Group Theoretic Approach and the Laser Method, but is more general, and current techniques do not suffice to find the optimal zeroing-out of $t^{\otimes n}$ into matrix multiplication even for simple tensors. Our lower bounds will be not only for the Solar method, but also for the Galactic method which is even more out of reach for the current matrix multiplication techniques.

To be clear, the Solar method, Galactic method, and Universal method, give us successively more power when analyzing specific tensors. For example, it may be the case that for a specific tensor T, the Solar method applied to T cannot get as low an upper bound on ω as the Universal method applied to T can. This captures the known methods to get bounds on ω by using tensors like the Coppersmith-Winograd tensor or a group tensor, which we will define shortly. The three different methods will trivially give the same bound, ω, when applied to matrix multiplication tensors themselves, but this is not particularly interesting: the entire point of these different methods is that the asymptotic rank of matrix multiplication tensors is not well-understood, and applying the methods to other tensors can help us get better bounds on it.

We will now describe the two approaches that follow the Solar method.

C. The Laser Method

We give a very brief overview of the Laser Method here; a much more detailed overview is given in the full version of this paper. Strassen [4] proposed a method for embedding a matrix multiplication tensor into a large tensor power of a starting tensor. He called it the *Laser Method*. This method is particularly effective when the starting tensor can be partitioned into 'blocks' which are each smaller matrix multiplication tensors.

In a large power of the starting tensor, products of these blocks make up larger matrix multiplication tensors. Then one uses a clever zeroing out to remove blocks

[6]This folklore fact follows from inverting the DFT over cyclic groups; see eg. [14, Section 3.1].

which share variables with each other, combined with the *asymptotic sum inequality* of Schönhage [16] to obtain a bound on ω:

Theorem III.1 (Asymptotic Sum Inequality [16])**.** *If $\bigoplus_{i=1}^{p}\langle k_i, m_i, n_i\rangle$ has border rank $\leq r$, and $r > p$, then $\omega \leq 3\tau$, where $\sum_{i=1}^{p}(k_i m_i n_i)^{\tau} = r$.*

We now turn to the most successful implementation of the Laser Method: the **Coppersmith-Winograd** approach. The Coppersmith-Winograd (CW) family of tensors is as follows: Let $q \geq 1$ be an integer.

$$CW_q = x_0y_0z_{q+1} + x_{q+1}y_0z_0 + x_0y_{q+1}z_0 + \sum_{i=1}^{q}(x_iy_0z_i + x_0y_iz_i + x_iy_iz_0).$$

CW_q is a concise tensor over $\mathbb{F}^{q+2} \otimes \mathbb{F}^{q+2} \otimes \mathbb{F}^{q+2}$, of border rank (and hence also asymptotic rank) $q+2$.

Coppersmith and Winograd [6], as well as the later improvements by Stothers [10], Vassilevska W. [3] and Le Gall [2], all apply the Laser Method to powers of CW_q for $q = 5$ or $q = 6$.

Since the Laser Method only relies on certain subtensors of CW_q being matrix multiplication tensors, we define a family of *generalized* CW tensors, $\underline{CW}_q$ as follows, to which the Laser Method applies equally well.

Definition III.1. *The family $\underline{CW}_q$ of tensors includes, for every permutation $\sigma \in S_q$, the tensor*

$$CW_q^{\sigma} = (x_0y_0z_{q+1} + x_0y_{q+1}z_0 + x_{q+1}y_0z_0) + \sum_{i=1}^{q}(x_iy_{\sigma(i)}z_0 + x_iy_0z_i + x_0y_iz_i).$$

We remark that the family above contains all tensors obtained from CW_q by replacing $\sum_{i=1}^{q}(x_iy_iz_0 + x_iy_0z_i + x_0y_iz_i)$ with $\sum_{i=1}^{q}(x_{\tau(i)}y_{\sigma(i)}z_0 + x_{\alpha(i)}y_0z_{\beta(i)} + x_0y_{\gamma(i)}z_{\delta(i)})$ for any choice of $\alpha, \beta, \gamma, \delta, \sigma, \tau \in S_q$.

For any such tensor from the family $\underline{CW}_q$, if its border rank is $q+2$, the Coppersmith-Winograd approach would give exactly the same bound on ω, as with CW_q.

D. Group-theoretic approach

We now give a very brief overview of the Group-theoretic approach; again, a more detailed overview is given in the full version of this paper. Cohn and Umans [5] pioneered a new group-theoretic approach for matrix multiplication. The idea is as follows. Take a group G and consider its group tensor defined below. (Throughout this paper, we write groups in multiplicative notation.)

Definition III.2. *For any finite group G, the* group tensor *of G, denoted T_G, is a tensor over X_G, Y_G, Z_G where $X_G := \{x_g \mid g \in G\}$, $Y_G := \{y_g \mid g \in G\}$, and $Z_G := \{z_g \mid g \in G\}$, given by*

$$T_G := \sum_{g,h \in G} x_g y_h z_{gh}.$$

(Note that the group tensor of G is really the structure tensor of the group algebra $\mathbb{C}[G]$, often written as $T_{\mathbb{C}[G]}$. We use T_G for ease of notation.)

Cohn and Umans show how the asymptotic rank of T_G can be expressed using representation theory. They then defined a property of subsets of G, the 'sumultaneous triple product property', such that any subset of G satisfying this property leads to a zeroing out of T_G into a direct sum of matrix multiplication tensors, after which the Asumptotic Sum Inequality can be applied. In summary, they give extremely clean group-theoretic definitions for how to upper bound ω using T_G.

In addition to the full version of this paper, we refer the reader to [17, Section 3.5] for more exposition on the Group-theoretic approach and its interpretation as finding a zeroing out of group tensors.

E. Independent Tensors

In this paper, we will be especially interested in zeroing outs and monomial degenerations from tensors T to independent tensors $\langle r\rangle$. We give a few relevant definitions here.

For a tensor T over X, Y, Z, its *independence number*, $I(T)$, is the maximum size of an independent tensor which can result from a zeroing out of T. We similarly can define the *asymptotic independence number* of T by

$$\tilde{I}(T) := \limsup_{n \in \mathbb{N}} \left[I(T^{\otimes n})\right]^{1/n}.$$

Since a zeroing out cannot increase the number of x-variables, y-variables, or z-variables, we get a simple upper bound $I(T) \leq \min\{|X|, |Y|, |Z|\}$. It similarly follows that $\tilde{I}(T) \leq \min\{|X|, |Y|, |Z|\}$. Throughout this paper, we will see a number of tensors which achieve equality in this bound, including all matrix multiplication tensors. In Section IV, we will prove this and many other properties of $\tilde{I}$.

IV. Matrix Multiplication and Independent Tensors

In this section, we will lay out our main framework for proving lower bounds on what values of ω can be achieved using different tensors T in the Galactic Method. The main idea is that, to prove such a lower bound for tensor T, it is sufficient to give an upper bound on $\tilde{I}(T)$.

Definition IV.1. *For a tensor T, let $\omega_g(T) \geq 2$ denote the best bound on ω that one can achieve using the Galactic Method with T. Hence, for all tensors T, we have $\omega \leq \omega_g(T)$.*

Lemma IV.1. *Let T be any tensor. For each positive integer n, let $\langle a_n, b_n, c_n \rangle$ be the dimensions of the largest matrix multiplication tensor which can result from a monomial degeneration of $T^{\otimes n}$ (i.e. the one which maximizes $a_n b_n c_n$). Then,*

$$\omega_g(T) = 3\log(\tilde{R}(T)) \cdot \liminf_{n \in \mathbb{N}} \frac{n}{\log(a_n b_n c_n)}.$$

Proof: If $T^{\otimes n}$ has a monomial degeneration to $\langle a_n, b_n, c_n \rangle$, this shows that $\tilde{R}(\langle a_n, b_n, c_n \rangle) \leq \tilde{R}(T^{\otimes n}) = (\tilde{R}(T))^n$, which yields $\omega_g(T) \leq 3\log((\tilde{R}(T))^n)/\log(a_n b_n c_n)$, as desired. ∎

We use the following monomial degeneration of matrix multiplication tensors which slightly generalizes Strassen's (from [4, Theorem 4]).

Lemma IV.2. *For any positive integers a, b, c, there is a monomial degeneration of $\langle a, b, c \rangle$ into an independent tensor of size $\frac{3}{4} \cdot \frac{abc}{\max\{a,b,c\}}$.*

Using this, we can prove the main idea behind our lower bound framework:

Theorem IV.1. *For any concise tensor T,*

$$\tilde{I}(T) \geq \tilde{R}(T)^{\frac{6}{\omega_g(T)} - 2}.$$

Corollary IV.1. *For any tensor T, if $\omega_g(T) = 2$, then $\tilde{I}(T) = \tilde{R}(T)$. Moreover, for every constant $s < 1$, there is a constant $w > 2$ such that every tensor T with $\tilde{I}(T) \leq \tilde{R}(T)^s$ must have $\omega_g(T) \geq w$.*

We defer the proof details of Theorem IV.1 and the intermediate results to the full version of this paper.

V. Partitioning Tools for Proving Lower Bounds

The goal of this section is to show some 'local' properties of tensors T which imply upper bounds on $\tilde{I}(T)$ (and hence, they will be ultimately used to prove lower bounds on $\omega_g(T)$). The general idea is that we will be finding partitions $T = A + B$ of our tensors, such that at least one of $\tilde{I}(A)$ and $\tilde{I}(B)$ is low, and using this to show that $\tilde{I}(T)$ is itself low. If $\tilde{I}$ were additive, i.e. if it were the case that $\tilde{I}(T) = \tilde{I}(A) + \tilde{I}(B)$ for any partition $T = A + B$, then this would be relatively straightforward. Unfortunately, $\tilde{I}$ is not additive in general, and even in many natural situations:

Example V.1. Let q be any positive integer, and define the tensors $T_1 := \sum_{i=0}^{q} x_0 y_i z_i$, $T_2 := \sum_{i=1}^{q+1} x_i y_0 z_i$, and $T_3 := \sum_{i=1}^{q+1} x_i y_i z_{q+1}$. We can see that T_1 has only one x-variable, T_2 has only one y-variable, and T_3 has only one z-variable, and so $\tilde{I}(T_1) = \tilde{I}(T_2) = \tilde{I}(T_3) = 1$. However, $T_1 + T_2 + T_3 = CW_q$, so the three tensors give a partition of the Coppersmith-Winograd tensor! Since $CW_q^{\otimes n}$ is known to zero out into fairly large matrix multiplication tensors for a large enough constant n, we see that $\tilde{I}(T_1 + T_2 + T_3)$ can grow unboundedly large as we increase q (in particular, we will see in the full version of this paper that $\tilde{I}(T_1 + T_2 + T_3) \geq (q+2)^{2/3}$). We can similarly see that $\tilde{I}(T_1 \otimes T_2 \otimes T_3)$ grows unboundedly with q, and so $\tilde{I}$ is not *multiplicative* either.

Throughout this section, we will nonetheless describe a number of general situations where, if T is partitioned into $T = A + B$, then bounds on $\tilde{I}(A)$ and $\tilde{I}(B)$ are sufficient to give bounds on $\tilde{I}(T)$.

We begin with some useful terminology and notation about partitioning tensors. Let D be a sub-tensor of a tensor T, that is, it is obtained by removing triples from the support of T. If T is over variable sets $X = \{x_1, \ldots, x_a\}, Y = \{y_1, \ldots, y_b\}, Z = \{z_1, \ldots, z_c\}$, then $T^{\otimes n}$, and hence $D^{\otimes n}$, is over variable sets $\bar{X}, \bar{Y}, \bar{Z}$, where the variables in $\bar{X}$ are indexed by n-length sequences over $[a]$, the variables in $\bar{Y}$ are indexed by n-length sequences over $[b]$, the variables in $\bar{Z}$ are indexed by n-length sequences over $[c]$.

Definition V.1. *Let T be a partitioned tensor $T = \sum_i P_i$, and let D be a sub-tensor of $T^{\otimes n}$.Consider some $j \in \{1, \ldots, n\}$. We say that D has an entry of P_i in the jth coordinate if there is a triple (α, β, γ) in the support of D for which $(\alpha_j, \beta_j, \gamma_j)$ is in the support of P_i.*

Since the P_i partition the triples in the support of T, this is well-defined.

We begin with our first partitioning tool, which we interpret after the Theorem statement.

Theorem V.1. *Suppose T is a tensor over X, Y, Z with $|X| = q$, and $x_1 \in X$ is any x-variable such that x_1 is in at most q terms in T. Let $B := T|_{X \setminus \{x_1\}}$ be the tensor over $X \setminus \{x_1\}, Y, Z$ from zeroing out x_1 in T, and suppose that $c := \tilde{I}(B)$ satisfies*

$$c \leq \frac{q-1}{q^{1/(q-1)}}.$$

Then,

$$\tilde{I}(T) \leq \left(\frac{q-1}{1-p}\right)^{1-p} \cdot \frac{1}{p^p},$$

where $p \in [0,1]$ is given by

$$p := \frac{\log\left(\frac{q-1}{c}\right)}{\log(q) + \log\left(\frac{q-1}{c}\right)}.$$

Remark V.1. Before we prove Theorem V.1, let us briefly interpret its meaning. Since B has only $q-1$ different x-variables, we know that $\tilde{I}(B) \leq q-1$. The theorem tells us that if, in fact, $\tilde{I}(B)$ is mildly smaller than this, then regardless of what terms in T involve x_1, we still get a nontrivial upper bound on $\tilde{I}(T)$. One can verify that $p = 1/q$ when $c = (q-1)/q^{1/(q-1)}$, and for every c less than this, $p > 1/q$, which gives a resulting bound on $\tilde{I}(T)$ which is strictly less than q.

We defer the proof of Theorem V.1 to the full version of this paper.

We next move on to our second tool. We show that if a tensor T has a large asymptotic independence number, then there must be a way to define a probability distribution on the terms of T such that each variable is assigned approximately the same probability mass.

Theorem V.2. *Suppose $q \geq 2$ is an integer, and T is a tensor over X, Y, Z with $|X| = |Y| = |Z| = q$, and $\delta \geq 0$ is such that $\tilde{I}(T) = q^{1-\delta}$. Then, for every $\kappa > 0$, there is a map $p : X \otimes Y \otimes Z \to [0,1]$ such that:*

- $\sum_{x_i y_j z_k \in T} p(x_i y_j z_k) = 1$, *and*
- *For each fixed i, fixed j, or fixed k, $\sum_{x_i y_j z_k \in T} p(x_i y_j z_k) \geq \frac{1}{q} - \sqrt{(\delta + \kappa)\ln(q)}$.*

Before proving Theorem V.2, we first prove a key Lemma:

Lemma V.1. *For any integers $n \geq 1$ and $q \geq 2$, any real $\delta \geq 0$, and any tensor T over X, Y, Z with $|X| = q$ and $x_1 \in X$, suppose $T^{\otimes n}$ has a zeroing out into an independent tensor D of size $|D| = q^{(1-\delta)n}$. Let $S_X \subseteq X^n$ be the set of all x-variables used in terms in D, and let $\varepsilon = \sqrt{\delta \ln(q)}$. Then, at least $q^{(1-\delta)n} - q^{(1-2\delta)n}$ of the elements $x \in S_X$ have x_1 appear in between $(1/q - \varepsilon)n$ and $(1/q + \varepsilon)n$ of the entries of x.*

Proof: Notice that the number of different n-tuples of variables of X which contain x_1 exactly i times is $\binom{n}{i} \cdot (q-1)^{n-i}$. Hence, the number of elements $x \in X^n$ which do not have x_1 appear in between $\frac{1-\varepsilon}{q}n$ and $\frac{1+\varepsilon}{q}n$ of the entries of x is

$$\sum_{i=0}^{\frac{1-\varepsilon}{q}n} \binom{n}{i}(q-1)^{n-i} + \sum_{i=\frac{1+\varepsilon}{q}n}^{n} \binom{n}{i}(q-1)^{n-i}. \quad (1)$$

We will bound the sum (1) using Hoeffding's inequality[7]. Let $A_1, \ldots, A_n$ be n independent random variables taking on the value 1 with probability $1/q$ and 0 otherwise, and let $A = \sum_{i=1}^{n} A_i$. We can see that (1) is equal to $q^n \cdot \Pr[|A - n/q| \geq \varepsilon n]$. By Hoeffding's inequality, if we pick $\varepsilon = \sqrt{\delta \ln(q)}$, then $\Pr[|A - n/q| \geq \varepsilon n] \leq q^{-2\delta n}$. Thus, (1) is at most $q^n \cdot q^{-2\delta n} = q^{(1-2\delta)n}$, and the result follows. ∎

Theorem V.2 can be proved using Lemma V.1; we defer the proof details to the full version of this paper.

For one simple but interesting Corollary, we will show that in any tensor T which has two 'corner terms' (see the Corollary statement for the precise meaning; we will see later that many important tensors have these corner terms), then no matter what the remainder of T looks like, T still does not have too large of an asymptotic independence number.

Corollary V.1. *Suppose $q \geq 2$ is an integer, and T is a tensor over X, Y, Z with $|X| = |Y| = |Z| = q$, such that $x_1, x_q \in X$, $y_1, y_q \in Y$, $z_1 \in Z$, and T contains the triples $x_q y_1 z_1$ and $x_1 y_q z_1$, and neither x_q nor y_q appears in any other triples in T. Then, there is a constant $c_q < q$ depending only on q such that $\tilde{I}(T) \leq c_q$.*

Proof: Suppose $\tilde{I}(T) = q^{1-\delta}$, and for any $\kappa > 0$, let p be the probability distribution on the terms of T which is guaranteed by Theorem V.2. For any fixed i, define $p(x_i) := \sum_{x_i y_j z_k \in T} p(x_i y_j z_k)$, and define $p(y_j)$ and $p(z_k)$ similarly. Since $x_q y_1 z_1$ and $x_1 y_q z_1$ are the only terms containing x_q or y_1, and they each contain z_1, it follows that $p(z_1) \geq p(x_q) + p(y_q)$. This, combined with Theorem V.2, implies the desired result; we defer the remaining details to the full version of this paper. ∎

Finally, we move on to our third partitioning tool. This tool is a substantial generalization of the fact that if T is a tensor over X, Y, Z, then $\tilde{I}(T) \leq \min\{|X|, |Y|, |Z|\}$, i.e. $\tilde{I}(T)$ must be small if T does not have many of one type of variable. We will show that, even if T can be partitioned into tensors which each do not have many of one type of variable, then $\tilde{I}(T)$ must be small. We will formalize this idea by introducing the notion of the measure of a tensor:

Definition V.2. *Let T be a tensor over X, Y, Z. We say that $X' \subseteq X$, $Y' \subseteq Y$, $Z' \subseteq Z$ are* minimal *for T if X' is the minimal (by inclusion) subset of X such that for each $x_i \in X \setminus X'$, for all j, k, $T_{i,j,k} = 0$, and similarly, Y' is the minimal subset of Y such that for each $y_j \in Y \setminus Y'$, for all i, k, $T_{i,j,k} = 0$ and Z' is the*

[7]Hoeffding's inequality states that if $X_1, \ldots, X_n$ are independent random variables taking on values in $[0,1]$, then for any $t \in [0,1]$, we have $\Pr[\sum_{i=1}^{n} X_i - \mathbb{E}[\sum_{i=1}^{n} X_i] \geq tn] \leq e^{-2nt^2}$.

minimal subset of Z such that for each $z_k \in Z \setminus Z'$, for all i, j, $T_{i,j,k} = 0$.

If T is a tensor, then the measure of T, *denoted $\mu(T)$, is given by $\mu(T) := |X| \cdot |Y| \cdot |Z|$, where X, Y, Z are minimal for T.*

Lemma V.2. *For any tensor T, we have $\tilde{I}(T) \leq \mu(T)^{1/3}$.*

Proof: Suppose X, Y, Z are the smallest sets of variables such that T is a tensor over X, Y, Z. Hence,

$$\tilde{I}(T) \leq \min(|X|, |Y|, |Z|) \leq (|X| \cdot |Y| \cdot |Z|)^{1/3} = \mu(T)^{1/3}.$$

■

For our main tool, we can generalize this to partitioned tensors:

Theorem V.3. *Suppose T is a tensor which is partitioned into k parts $T = P_1 + P_2 + \cdots + P_k$ for any positive integer k. Then, $\tilde{I}(T) \leq \sum_{i=1}^{k} (\mu(P_i))^{1/3}$.*

Proof: Let $s := \sum_{i=1}^{k} (\mu(P_i))^{1/3}$, and for each $i \in \{1, 2, \ldots, k\}$, let $p_i := (\mu(P_i))^{1/3}/s$, so that $p_i \in [0, 1]$ and $\sum_{i=1}^{k} p_i = 1$. For any positive integer n, let D_n be the biggest independent tensor which can result from a zeroing out of $T^{\otimes n}$.

Set $T' = T^{\otimes n}$, and $D' = D_n$, and then for j from 1 to n do the following process:

Currently $T' = Q_1 \otimes Q_2 \otimes \cdots \otimes Q_{j-1} \otimes T^{n-j+1}$, and $|D'| \geq q_1 q_2 \cdots q_{j-1} \cdot |D_n|$, and moreover, D' is a zeroing out of T'. Pick an i such that at least a p_i fraction of the independent triples in D' have an entry of P_i in their jth coordinate; since $\sum_j p_j = 1$, such an i exists. Set $Q_j = P_i$ and $q_j = p_i$. Recall that there is a zeroing out z such that $z(T') = D'$. Now, replace the jth tensor in the product defining T' by Q_j, i.e. set $T' = Q_1 \otimes Q_2 \otimes \cdots \otimes Q_j \otimes T^{n-j}$. By our choice of Q_j, we know that if we apply the same zeroing out z to the new T', we get at least a q_j fraction of the number of independent triples we had before, i.e. $|z(T')| \geq q_j |D'|$. Let D' be this new independent tensor $z(T')$.

Once we have done this for all j, we are left with a tensor $\bigotimes_{j=1}^{n} Q_j$ which has a zeroing out into $|D| \cdot \prod_{j=1}^{n} q_j$ independent triples. Note that measure is multiplicative, and so in particular, $\mu(\bigotimes_{j=1}^{n} Q_j) = \prod_{j=1}^{n} \mu(Q_j)$. Hence, by Lemma V.2,

$$\tilde{I}(\bigotimes_{j=1}^{n} Q_j) \leq \prod_{j=1}^{n} \mu(Q_j)^{1/3} = \prod_{j=1}^{n} (s \cdot q_j) = s^n \cdot \prod_{j=1}^{n} q_j.$$

Since D' is a zeroing out of $\bigotimes_{j=1}^{n} Q_j$, it follows that $|D'| \leq s^n \cdot \prod_{j=1}^{n} q_j$. But, $|D'| \geq |D| \cdot \prod_{j=1}^{n} q_j$. Combining the two, we get that $|D_n| \leq s^n$, as desired. ■

VI. Lower Bounds for Group Tensors

In contrast with the previous section, in this section we will show a 'global' property of tensors T which imply upper bounds on $\tilde{I}(T)$ (and hence lower bounds on $\omega_g(T)$). In particular, we will see that if T is the group tensor of any finite group G, or a monomial degeneration of any such group tensor with the same measure, then $\tilde{I}(T) < \tilde{R}(T)$ and so $\omega_g(T) > 2$. In this section, we also study the Coppersmith-Winograd tensor CW_q, and show that it cannot be found as a sub-tensor of group tensors of relatively small groups, giving evidence that lower bounds on the group-theoretic approach are insufficient to imply strong lower bounds on $\omega_g(CW_q)$. We defer the details to the full version of this paper.

VII. Applications of our Lower Bound Techniques

In this section, we use the lower bounding techniques that we have developed throughout the paper for a number of applications to tensors of interest.

A. Generalized CW tensors

We begin by proving our main result:

Theorem VII.1. *There is a universal constant $c > 2$ such that for any generalized Coppersmith-Winograd tensor T (with any parameter q), we have $\omega_g(T) \geq c$.*

Proof: This follows from Lemmas VII.1 and VII.2, which we state and prove below. ■

Lemma VII.1. *For every nonnegative integer q, there is a constant $c_q > 2$ such that for any generalized Coppersmith-Winograd tensor T with parameter q, we have $\omega_g(T) \geq c_q$.*

Lemma VII.2. *There is a constant $c' > 2$ and a positive integer q' such that for any integer $q \geq q'$, and any generalized Coppersmith-Winograd tensor T with parameter q, we have $\omega_g(T) \geq c'$.*

Proof of Lemma VII.1: For each q, and each generalized Coppersmith-Winograd tensor T with parameter q, the tensor T is of the form described by Corollary V.1, which says that $\tilde{I}(T) < s_{q+2}$ for some constant $s_{q+2} < q + 2$ which depends only on q. It then follows from Corollary IV.1 that $\omega_g(T) > c_q$ for some constant $c_q > 2$ determined by s_q, as desired. ■

The proof above of Lemma VII.1 used Corollary V.1, which follows from Theorem V.2, as its main tool. We will next give two different proofs of Lemma VII.2; the first will showcase Theorem V.3, and the second will showcase Theorem V.1. Each of Theorems V.1, V.2, and V.3 describes a different property of a tensor T which is

enough to imply that $\omega_g(T) > 2$. Throughout these three proofs, we are showing that the Coppersmith-Winograd tensor has *all three* of these properties!

First proof of Lemma VII.2: Suppose T is a generalized Coppersmith-Winograd tensor with parameter q. Hence, T can be written as

$$\begin{aligned} T =& x_0y_0z_0 + x_0y_{q+1}z_{q+1} + x_{q+1}y_0z_{q+1} \\ &+ \sum_{i=1}^{q}(x_0y_iz_i + x_iy_0z_i + x_iy_{\sigma(i)}z_{q+1}), \end{aligned}$$

for some permutation σ on $\{1, 2, \ldots, q\}$. We partition T into three parts T_1, T_2, T_3 as follows:

$$T_1 = \sum_{i=0}^{q} x_0y_iz_i, \qquad T_2 = \sum_{i=1}^{q+1} x_iy_0z_i,$$

$$T_3 = x_0y_{q+1}z_{q+1} + \sum_{i=1}^{q} x_iy_{\sigma(i)}z_{q+1}.$$

Note that T_1 has only one x-variable, T_2 has only one y-variable, and T_3 has only one z-variable. Hence, $\mu(T_1) = \mu(T_2) = \mu(T_3) = q^2$. It follows from Theorem V.3 that $\tilde{I}(T) \leq 3q^{2/3}$. When $q \geq 28$, we have $3q^{2/3} < q^{0.997}$, and so by Corollary IV.1, there is a fixed constant $c' > 2$ independent of q such that $\omega_g(T) \geq c'$, as desired. ■

Our second proof, which will use Theorem V.1 instead of Theorem V.3 as our primary tool, can be found in the full version of this paper.

B. Further Applications

Further applications can be found in the full version of this paper, including

- For every finite group G, a monomial degeneration of T_G into a generalized Coppersmith-Winograd tensor with parameter $|G| - 2$, and
- A characterization of $\tilde{I}(T)$ for lower triangular tensors T.

ACKNOWLEDGMENT

The authors are extremely grateful to JM Landsberg for answering their many questions, and to Ryan Williams for his many useful suggestions.

REFERENCES

[1] V. Strassen, "Gaussian elimination is not optimal," *Numerische mathematik*, vol. 13, no. 4, pp. 354–356, 1969.

[2] F. Le Gall, "Powers of tensors and fast matrix multiplication," in *ISSAC*, 2014, pp. 296–303.

[3] V. V. Williams, "Multiplying matrices faster than Coppersmith-Winograd," in *STOC*, 2012, pp. 887–898.

[4] V. Strassen, "The asymptotic spectrum of tensors and the exponent of matrix multiplication," in *FOCS*, 1986, pp. 49–54.

[5] H. Cohn and C. Umans, "A group-theoretic approach to fast matrix multiplication," in *FOCS*, 2003, pp. 438–449.

[6] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Journal of symbolic computation*, vol. 9, no. 3, pp. 251–280, 1990.

[7] H. Cohn, R. Kleinberg, B. Szegedy, and C. Umans, "Group-theoretic algorithms for matrix multiplication," in *FOCS*, 2005, pp. 379–388.

[8] N. Alon, A. Shpilka, and C. Umans, "On sunflowers and matrix multiplication," *Computational Complexity*, vol. 22, no. 2, pp. 219–243, 2013.

[9] A. Ambainis, Y. Filmus, and F. Le Gall, "Fast matrix multiplication: limitations of the Coppersmith-Winograd method," in *STOC*, 2015, pp. 585–593.

[10] A. Davie and A. J. Stothers, "Improved bound for complexity of matrix multiplication," *Proceedings of the Royal Society of Edinburgh, Section: A Mathematics*, vol. 143, pp. 351–369, 4 2013.

[11] J. Blasiak, T. Church, H. Cohn, J. A. Grochow, E. Naslund, W. F. Sawin, and C. Umans, "On cap sets and the group-theoretic approach to matrix multiplication," *Discrete Analysis*, vol. 2017, no. 3, pp. 1–27, 2017.

[12] W. Sawin, "Bounds for matchings in nonabelian groups," *arXiv preprint arXiv:1702.00905*, 2017.

[13] J. Blasiak, T. Church, H. Cohn, J. A. Grochow, and C. Umans, "Which groups are amenable to proving exponent two for matrix multiplication?" *arXiv preprint arXiv:1712.02302*, 2017.

[14] J. Alman and V. V. Williams, "Further limitations of the known approaches for matrix multiplication," in *Proc. of ITCS*, 2018, pp. 25:1–25:15.

[15] D. Coppersmith and S. Winograd, "On the asymptotic complexity of matrix multiplication," *SIAM J. Comput.*, vol. 11, no. 3, pp. 472–492, 1982.

[16] A. Schönhage, "Partial and total matrix multiplication," *SIAM J. Comput.*, vol. 10, no. 3, pp. 434–455, 1981.

[17] J. M. Landsberg, *Geometry and complexity theory*. Cambridge University Press, 2017, vol. 169.