Matrix Multiplication II

Yuval Filmus

March 6, 2014

These notes started their life as a lecture given at the Toronto Student Seminar on February 9, 2012. The notes were updated for a seminar given at the IAS on March 4, 2014. The material is taken mostly from Davie and Stothers [DS] (the combinatorial construction, properties of the value) and Le Gall's recent paper [Gal] (everything else). Other sources are the classic paper by Coppersmith and Winograd [CW], §15.7 of *Algebraic Complexity Theory* [ACT], Stothers's thesis [Sto], V. Williams's recent paper [Wil], and the paper by Cohn at al. [CKSU].

1 Recap

Last week culminated in Schönhage's asymptotic sum inequality:

$$\sum_i (n_i m_i p_i)^{\omega/3} \leq \underline{\mathrm{R}}\left(\bigoplus_i \langle n_i, m_i, p_i\rangle\right).$$

Here ω is the exponent of matrix multiplicated, $\langle n, m, p\rangle$ is the matrix multiplication tensor

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

$\oplus$ is direct sum (meaning that the tensors have disjoint support), and $\underline{\mathrm{R}}(T)$ is the *border rank* of the tensor T, which is the minimal ρ such that some sequence of tensors of rank ρ tend to T. (The rank of a tensor is the minimal number of *rank one tensors* $(\sum_i \alpha_i x_i)(\sum_j \beta_j y_j)(\sum_k \gamma_k z_k)$ that sum to it.)

One of the main tools in proving the asymptotic sum inequality was the *tensor product*, which generalizes the Kronecker product of matrices. It satisfies the following two crucial properties: $\langle n_1, m_1, p_1\rangle \otimes \langle n_2, m_2, p_2\rangle = \langle n_1 n_2, m_1 m_2, p_1 p_2\rangle$ and $\underline{\mathrm{R}}(T_1 \otimes T_2) \leq \underline{\mathrm{R}}(T_1)\underline{\mathrm{R}}(T_2)$. Another property which will be useful is that the border rank of a tensor is invariant under the six permutations over the roles of x, y, z. When applied to matrix multiplication tensors, this yields $\underline{\mathrm{R}}(\langle n, m, p\rangle) = \underline{\mathrm{R}}(\langle m, n, p\rangle)$ and so on.

The asymptotic sum inequality answers the following question: Suppose that we have an identity bounding the border rank of a direct sum of matrix multiplication tensors; what is the best bound on ω that can be derived? In this talk we will describe Strassen's *laser method*, which allows us to obtain bounds on ω even if the matrix multiplication tensors are *not* disjoint.

2 Warmup

We start our exploration with the warmup identity of Coppersmith and Winograd [CW]:

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

This identity concerns the following sets of variables:

$$x_0^{[0]}, x_1^{[1]}, \ldots, x_q^{[1]},$$
$$y_0^{[0]}, y_1^{[1]}, \ldots, y_q^{[1]},$$
$$z_0^{[0]}, z_1^{[1]}, \ldots, z_q^{[1]}.$$

We have partitioned the x-variables into two different sets, one consisting of $x_0^{[0]}$, and the other of $x_1^{[1]}, \ldots, x_q^{[1]}$. The y-variables and z-variables are similarly partitioned. We can summarize this identity as follows:

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

We have annotated each matrix multiplication tensor with the labels of the x-variables, y-variables and z-variables involved. This tells us, for example, that $\langle 1,1,q\rangle^{0,1,1}$ and $\langle q,1,1\rangle^{1,0,1}$ share z-variables but have disjoint x-variables and y-variables.

We cannot apply the asymptotic sum inequality to this identity, since the matrix multiplication tensors are not disjoint. Instead, our strategy will be different: we will take a high tensor power and then zero variables in such a way that the tensors that remain are disjoint, and then apply the asymptotic sum inequality. We will always zero variables in groups, according to the indices describing them.

After taking an Nth tensor power, we get an identity bounding the border rank of a sum of 3^N matrix multiplication tensors by $(q+2)^N$. The tensors have different *formats* (dimensions) $\langle n,m,p\rangle$, but the asymptotic sum inequality only cares about their *volume* $(nmp)^{1/3}$, which is the same for all of them, namely $q^{N/3}$. Suppose that we could zero some of the variables so that what remains is K_N disjoint tensors. The asymptotic sum inequality then gives $K_N q^{N\omega/3} \leq (q+2)^N$ or $K_N^{1/N} q^{\omega/3} \leq (q+2)$. How large can K_N be?

Each tensor in the Nth tensor power involves certain groups of variables. For example, when $N=1$, the tensor $\langle 1,1,q\rangle^{0,1,1}$ involves the variables $x^{[0]}, y_i^{[1]}, z_i^{[1]}$. When $N=2$, the tensor $\langle 1,1,q\rangle^{0,1,1} \otimes \langle q,1,1\rangle^{1,0,1}$ involves the variables

$$x^{[0]} \otimes x_j^{[1]}, y_i^{[1]} \otimes y^{[0]}, z_i^{[1]} \otimes z_j^{[1]}.$$

We can write this more concisely as $x_j^{[01]}, y_i^{[10]}, z_{ij}^{[11]}$, and the tensor itself can be described concisely as $\langle q,1,q\rangle^{01,10,11}$. When zeroing variables, we will always zero variables in groups: for example, we

could zero all variables $z_{ij}^{[11]}$. We call these groups x-, $y-$ and z-*groups*. The annotations themselves (say 11) we call *indices*, so the annotation of each tensor is an *index triple*.

Easy upper bounds. We start with some rather trivial upper bounds on K_N. After zeroing out variables, all the x-groups that remain appear in a unique matrix multiplication tensor. There are 2^N different x-groups, so $K_N \leq 2^N$.

A slightly less trivial upper bound on K_N, due to Cohn et al. [CKSU], arises from the idea of *source distribution*. The set of index triples has the product structure $((0,1,1)+(1,0,1)+(1,1,0))^N$. We can decompose this set into types depending on how many times each factor was used: for each a, b, c summing to N, there are $\binom{N}{a,b,c}$ index triples of type (a, b, c). An x-index in any such triple contains a zeroes, a y-index contains b zeroes, and a z-index contains c zeroes. There are $\binom{N+2}{2}$ different types. The source distribution, a scale-invariant quantity, is $(a/N, b/N, c/N)$.

Consider any type (a, b, c), and assume without loss of generality that $a \leq N/3$. Focus on surviving index triples of type (a, b, c). Each such index triple must have a unique x-index, and there are $\binom{N}{a} \leq \binom{N}{N/3}$ of these. Accounting for all possible source distributions, we obtain $K_N \leq \binom{N+2}{2}\binom{N}{N/3} \approx 2^{Nh(1/3)}$. (Here $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function.)

The bound. Surprisingly, this bound is tight! There is a construction achieving $K_N \geq 2^{Nh(1/3)-o(N)}$. This implies that $2^{h(1/3)}q^{\omega/3} \leq (q+2)$. Optimizing over q, we find out that when $q = 8$ we get the bound $\omega \leq 2.404$.

3 The Construction

We now describe the construction giving $K_N \approx 2^{Nh(1/3)}$. It will be advantageous to describe it in greater generality. We start by describing the *parameters* of the construction, instantiating them for our present case. First, the basic index set, consisting of triples of integers, which we assume to be rotation-invariant (with respect to permutations of the three indices), and *tight*: every basic index triple sums to the same value. Tightness is crucial for the construction. The basic index set in our case is

$$(0,1,1), (1,0,1), (1,1,0).$$

Second, a rotation-invariant source distribution π, prescribing the type of tensors we're looking at:

$$\pi = (1/3, 1/3, 1/3).$$

These are the inputs. Given N (the tensor power), we will describe the results in terms of three quantities P, F, G, defined below. We will use $N\pi$ to mean the type of an index triple induced by π, in our case $(N/3, N/3, N/3)$ (we will not worry about these being integers). We are aiming at index triples of type $N\pi$. An x-index has type $N\pi$ if it appears in an index triple of type $N\pi$. An index triple (i, j, k) has *projection type* $N\pi$ if i, j, k all have type $N\pi$. In our case, an index triple (i, j, k) has projection type $N\pi = (N/3, N/3, N/3)$ if and only if it has type $N\pi$, but in some cases there might be more index triples of projection type $N\pi$ than index triples of (strict) type $N\pi$ (see discussion below). The parameters P, F, G are:

- P is the number of x-indices of type $N\pi$. (Because of rotation invariance, it is also the number of y-indices and z-indices of type $N\pi$.)

- F is the number of index triples of type $N\pi$.
- G is the number of index triples of projection type $N\pi$.

The upper bound we described above is P. Our construction will give $(F/G)P^{1-o(1)}$, which is tight when $F = G$.

The applications below will actually require a slightly more general result: instead of considering all index triples of projection type $N\pi$ as possible obstructions, we will allow a pre-filtering step which eliminates some of them at the outset. In that case, G will be the number of index triples after the pre-filtering, and what we require is that after zeroing variables, no index repeats in the pre-filtered index triples.

Mock construction. We start with a mock construction. Suppose that instead of zeroing variables, we are allowed to select index triples, under the constraint that no x-index, y-index or z-index repeats. In this case, we will be able to achieve $\Omega(P)$ using a very simple construction. We can focus on index triples of type $N\pi$. The idea is to select each index triple with some probability ϵ, and to filter out bad index triples, which are surviving index triples that have a matching index with some other surviving index triple. Call the remaining index triples *good*.

Each index triple (i, j, k) could potentially conflict with $3(F/P - 1)$ other index triples: symmetry considerations show that there are F/P index triples each of the form $(i, \cdot, \cdot)$, of the form $(\cdot, j, \cdot)$, and of the form $(\cdot, \cdot, k)$. Therefore the probability that an index triple is bad is at most $\epsilon^2 3F/P$, and the probability that it is good is at least $\epsilon - \epsilon^2 3F/P$. The choice of ϵ maximizing this expression is $\epsilon = P/6F$, giving a probability of $P/12F$. Therefore the expected number of good index triples is $F \cdot (P/12F) = P/12 = \Omega(P)$.

The problem. For the actual construction, we need to zero variables rather than eliminate triples. We start by zeroing all variables other than those of type $N\pi$, leaving only index triples of projection type $N\pi$ in play (it is this step that will be replaced by a different pre-filtering in some of the applications). We are going to follow the same basic approach: we select each index triple of type $N\pi$ with some probability ϵ, and then eliminate bad index triples (these will be defined in a bit). Following that, we will restrict the variables to those variables appearing in good index triples. This gives us a set of *resulting index triples*, those index triples all whose variables have survived. We want this set to have no repeated x-, y- or z-indices.

We will define bad index triples in such a way that no index repeats in the good index triples. But this doesn't guarantee that the *resulting* index triples have no repeated indices. We could have the following situation:

$$\begin{array}{ccc} \textcircled{i_1} & j_1 & k_1 \\ i_2 & \textcircled{j_2} & k_2 \\ i_3 & j_3 & \textcircled{k_3} \end{array}$$

Here $(i_1, j_1, k_1), (i_2, j_2, k_2), (i_3, j_3, k_3)$ are good index triples, and the diagonal (i_1, j_2, k_3) is a potential resulting index triple that conflicts with them. (This should remind the knowledgeable reader of the uniquely solvable puzzles (USPs) in Cohn et al. [CKSU].) Notice that while the former index triples have type $N\pi$, the diagonal (i_1, j_2, k_3) is only known to have *projection* type $N\pi$.

Given the first index triple (i_1, j_1, k_1), there are G/P choices for j_2, k_3, and then F/P choices each for i_2, k_2 and i_3, j_3. If we define a bad index triple as one which participates in such a constellation, then the probability that an index triple is good is at least $\epsilon - 3F/P\epsilon^2 - 3GF^2/P^3\epsilon^3$ (going

over all possible choices of the common index). For simplicity, remove the term $3F/P\epsilon^2$ from this expression. The choice of ϵ maximizing the expression $\epsilon - 3GF^2/P^3\epsilon^3$ is $\epsilon = P^{3/2}/3G^{1/2}F$, giving a probability of $2P^{3/2}/9G^{1/2}F$, and an expected number of resulting triples which is $2P^{3/2}/9G^{1/2}$. In our running example, $P^{3/2}/G^{1/2} = \binom{n}{n/3}^{3/2}/\binom{n}{n/3,n/3,n/3}^{1/2} \approx 2^{n(\log_2 3-1)}$, whereas $P = \binom{n}{n/3} \approx 2^{n(\log_2 3-2/3)}$. In other words, the bound is not asymptotically tight.

The trick. The idea is to come up with a way of choosing the initial index triples with the following property. In the situation described above, if $(i_1, j_1, k_1), (i_2, j_2, k_2), (i_3, j_3, k_3)$ are all selected, then so is (i_1, j_2, k_3). This means that our method of choosing index triples will choose not only index triples of type $N\pi$ but also index triples of projection type $N\pi$; we will only be interested in the number of resulting index triples of (strict) type $N\pi$. In order to guarantee that no bad constellation arises, it will be enough to guarantee that all the chosen index triples will have distinct indices. Retracing our footsteps, we expect that an index triple be good with probability roughly $\epsilon - \epsilon^2 3G/P$, assuming that our method of choosing index triples is 2-wise independent (at least approximately). Optimizing over ϵ, this probability should be roughly $P/6G$, and so the number of good index triples of (strict) type $N\pi$ should be around $(F/6G)P$ in expectation.

We now reach the surprising part of the construction, which is the new method of choosing index triples. Let M be a large prime roughly equal to $1/\epsilon$. We will define three random hash functions h_x, h_y, h_z mapping the individual indices to $\mathbb{Z}_M$. For some appropriately defined set $A \subseteq \mathbb{Z}_M$ of cardinality $M^{1-o(1)}$, we will choose all index triples (i, j, k) such that $h_x(i), h_y(j), h_z(k) \in A$. Our random hash functions and set A will be chosen so that the following properties hold:

(P1) For all x-indices i_1 and y-indices $j_1 \neq j_2$, the hash triple $(h_x(i_1), h_y(j_1), h_y(j_2))$ is distributed uniformly over $\mathbb{Z}_M^3$. (And similarly for all other choices of positions among x, y, z.)

(P2) For all index triples (i, j, k),

$$h_x(i), h_y(j), h_z(k) \in A \text{ if and only if } h_x(i) = h_y(j) = h_z(k) \in A.$$

(P3) For all index triples (i, j, k), if any two of $h_x(i), h_y(j), h_z(k)$ are equal then all are equal.

Given these properties, we show how to complete the proof. An index triple (i_1, j_1, k_1) is chosen if for some $a \in A$, $h_x(i_1) = h_y(j_1) = h_z(k_1) = a$, or equivalently $h_x(i_1) = h_y(j_1)$; this happens with probability $|A|M^{-2} = M^{-1-o(1)} = \epsilon^{1+o(1)}$. A chosen index triple (i_1, j_1, k_1) is *bad* if it shares an index with another chosen index triple. There are at most $3G/P$ such index triples which could conflict with (i_1, j_1, k_1). A pair of conflicting index triples, say (i_1, j_1, k_1) and (i_1, j_2, k_2), are both chosen if for some $a \in A$, $h_x(i_1) = h_y(j_1) = h_y(j_2) = a$. This happens with probability $|A|M^{-3} = M^{-2-o(1)} = \epsilon^{2+o(1)}$. Hence an index triple is good with probability at least $\epsilon^{1+o(1)} - (3G/P)\epsilon^{2+o(1)}$, which for an appropriate choice of ϵ becomes $(P/G)^{1-o(1)}$. So the expected number of good triples of (strict) type $N\pi$ is $F(P/G)^{1-o(1)} \geq (F/G)P^{1-o(1)}$. As commented above, the resulting index triples have no repeated indices, completing the proof.

Defining the hash function. We will construct h_x, h_y, h_z so that for all index triples i, j, k, $h_x(i) + h_y(j) = 2h_z(k)$ (property P4). We will be able to guarantee this property since the basic index triples are tight by assumption: all indices sum to the same constant, which without loss of generality is zero. This already shows that for every index triple (i, j, k), if any two of

$h_x(i), h_y(j), h_z(k)$ are equal, then all are equal (property P3). Property P2 would hold exactly if A contains no three-term arithmetic progressions other than constant ones. We're very lucky that Salem and Spencer [SS] constructed such a set for us of size $M^{1-o(1)}$. (Better constructions exist [Beh, Mos, Elk], but the difference doesn't matter here.) It is now a simple exercise to define h_x, h_y, h_z so that properties P1 and P4 are satisfied: let $\alpha \in \mathbb{Z}_M^N$ and $X, Y, Z \in \mathbb{Z}_M$ be chosen randomly under the constraint $X + Y + Z = 0$, and define (over $\mathbb{Z}_M$)

$$\begin{aligned} h_x(i) &= 2(\langle \alpha, i \rangle + X), \\ h_y(j) &= 2(\langle \alpha, j \rangle + Y), \\ h_z(k) &= -(\langle \alpha, k \rangle + Z). \end{aligned}$$

To see that property P4 holds, use $i + j + k = \mathbf{0}$ to get

$$h_x(i) + h_y(j) = 2(\langle \alpha, i + j \rangle + X + Y) = -2(\langle \alpha, k \rangle + Z) = 2h_z(k).$$

To see that property P1 holds, let $D = j_2 - j_1 \neq \mathbf{0}$. We have

$$\tfrac{1}{2}(h_x(i_1), h_y(j_1), h_y(j_2)) = (\langle \alpha, i_1 \rangle + X, \langle \alpha, j_1 \rangle + Y, \langle \alpha, j_1 \rangle + Y + \langle \alpha, D \rangle).$$

The three quantities $X, Y, \langle \alpha, D \rangle$ are distributed uniformly in $\mathbb{Z}_M^3$, implying that property P1 holds.

Strassen's version. Strassen [S2] came up with an alternative construction which trades Salem–Spencer sets for tensor *degeneration*. This is the process in which, in addition to zeroing some of the variables, we multiply the rest by some powers of ϵ, the result being those tensors with the minimal coefficient of ϵ. This gives an upper bound on the border rank of the adjusted tensor, and we can apply the asymptotic sum inequality as before. We start by zeroing all variables other than those of projection type $N\pi$. We then multiply variables according to the hash of their index, as follows:

$$\begin{aligned} x^{[i]} &\to \epsilon^{2h_x(i)^2}, \\ y^{[j]} &\to \epsilon^{2h_y(j)^2}, \\ z^{[k]} &\to \epsilon^{-4h_z(k)^2}. \end{aligned}$$

Here we think of the hashes as integers in $\{0, \ldots, M-1\}$ rather than elements of $\mathbb{Z}_M$. The analog of property P2 is this: an index triple (i, j, k) has minimal ϵ power if and only if $h_x(i) = h_y(j) = h_z(k) \in B$, where $B \subseteq \{0, \ldots, M-1\}$ will turn out to have size $\Omega(M)$ (rather than just $M^{1-o(1)}$).

What is the ϵ power corresponding to a tensor with index triple (i, j, k)? Recall that $h_x(i) + h_y(j) \equiv 2h_z(k) \pmod{M}$. The left-hand side is strictly smaller than $2M$, and so there are two cases: either $2h_z(k) = h_x(i) + h_y(j)$ or $2h_z(k) = h_x(i) + h_y(j) - M$. In the first case, the exponent of ϵ is

$$2h_x(i)^2 + 2h_y(j)^2 - (2h_z(k))^2 = 2h_x(i)^2 + 2h_y(j)^2 - (h_x(i) + h_y(j))^2 = (h_x(i) - h_y(j))^2.$$

In the second case, the exponent of ϵ is

$$\begin{aligned} 2h_x(i)^2 + 2h_y(j)^2 - (2h_z(k))^2 &= 2h_x(i)^2 + 2h_y(j)^2 - (h_x(i) + h_y(j) - M)^2 \\ &= (h_x(i) - h_y(j))^2 + M(2h_x(i) + 2h_y(j) - M) \geq M^2. \end{aligned}$$

We conclude that the minimum power of ϵ is attained at index triples (i, j, k) satisfying $h_x(i) = h_y(j)$ and $2h_x(i) < M$, that is, $h_x(i) = h_y(j) = h_z(k) \in B = \{0, \ldots, (M-1)/2\}$, where $|B| > |M|/2$.

Strassen's identity. Strassen [S1] originally came up with this method of tensoring followed by zeroing (or degeneration) in the context of another identity:

$$\epsilon \sum_{i=1}^{q} (x_i^{[1]} y_0^{[0]} z_i^{[1]} + x_0^{[0]} y_i^{[1]} z_i^{[1]}) + O(\epsilon^2) =$$
$$\sum_{i=1}^{q} (x_0^{[0]} + \epsilon x_i^{[1]})(y_0^{[0]} + \epsilon y_i^{[1]}) z_i - x_0^{[0]} y_0^{[0]} \sum_{i=1}^{q} z_i.$$

This identity shows that $\underline{\mathrm{R}}(\langle q,1,1\rangle^{1,0,1} + \langle 1,1,q\rangle^{0,1,1}) \leq q+1$. Strassen used a direct construction showing $4q^\omega \leq (q+1)^3$, which for $q=5$ gives $\omega < 2.48$. See Strassen's original paper or §15.6 of [ACT] for details.

Tightness of the construction. So far we have only considered one application of the construction, in which $G = F$. As we remarked above, when $G = F$ the construction is optimal in the sense that it achieves the best possible $\lim_{N\to\infty} K_N^{1/N}$ under the given combinatorial constraints. However, in some cases $G \neq F$. This can happen when there are two rotation-invariant source distributions whose projections to the x-index are the same. Here is an example:

Index triple	Probability
$(0,2,4)$	$1/9$
$(2,4,0)$	$1/9$
$(4,0,2)$	$1/9$
$(1,2,3)$	$2/9$
$(2,3,1)$	$2/9$
$(3,1,2)$	$2/9$

Index triple	Probability
$(0,3,3)$	$1/9$
$(3,0,3)$	$1/9$
$(3,3,0)$	$1/9$
$(1,1,4)$	$1/9$
$(1,4,1)$	$1/9$
$(4,1,1)$	$1/9$
$(2,2,2)$	$1/3$

The projections to the x-index are both

x-index	Probability
0	$1/9$
1	$2/9$
2	$1/3$
3	$2/9$
4	$1/9$

This is the smallest example in the sense that there is no such example with only $\{0,1,2,3\}$.

For given N, the answer lies somewhere between P and $(F/G)P^{1-o(1)}$. We can estimate these quantities asymptotically. It is well-known that $F \approx 2^{NH(\pi)}$, where $H(\pi) = -\sum_i \pi(i)\log_2 \pi(i)$ is the binary entropy function, and the approximation is up to polynomial terms (in N). Similarly, if π_x is the projection of π into the x-index, then $P \approx 2^{NH(\pi_x)}$. Finally, $G \approx \sum_{\sigma:\, \pi_x=\sigma_x} 2^{NH(\sigma)}$, for all source distributions σ realizable exactly on N coordinates, that is, each probability is an integer multiple of $1/N$. There are only polynomially many of these, and so we can approximate G by the maximal one. As N tends to infinity, we can disregard the realizability constraint and consider arbitrary distributions, and so $G \approx \max_{\sigma:\, \pi_x=\sigma_x} 2^{NH(\sigma)}$, where σ now is any probability distribution (technically this should be a supremum, but it is easy to check that the supremem is

always achieved). In fact, the concavity of the entropy function shows that the maximizing σ must be rotation-invariant. Putting everything together, we get the following bounds:

$$H(\pi_x) + H(\pi) - \max_{\sigma\colon \sigma_x = \pi_x} H(\sigma) \leq \log_2 \lim_{N\to\infty} K_N^{1/N} \leq H(\pi_x).$$

(It is not hard to check that $K_{N_1 N_2} \geq K_{N_1} K_{N_2}$, implying that the limit $\lim_{N\to\infty} K_N^{1/N}$ in fact exists.)

This leads to the following open question.

Open question 1. What is the best possible $\lim_{N\to\infty} K_N^{1/N}$ achievable given basic index triples and a source distribution? Does using degeneration (see Strassen's version above) lead to better results?

We conjecture that the construction is optimal.

4 A better identity

The identity considered up to now can be slightly tweaked to yield even more:

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

This shows that

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

The new factors $\langle 1,1,1\rangle$ come at absolutely no cost! In fact, more can be added in the same way, but this doesn't result in any further improvement.

In contrast with the previous identity, the basic index triples here have different volumes (recall that the volume of $\langle n,m,p\rangle$ is $(nmp)^{\omega/3}$). For a given (rotation-invariant) source distribution $\pi = (\alpha/3, \alpha/3, \alpha/3, \beta/3, \beta/3, \beta/3)$, the volume of each of the corresponding tensors in the Nth tensor power of the basic identity is $q^{(\alpha/3)N}$. The corresponding distribution of x-indices is $\pi_1 = (\alpha/3 + 2\beta/3, 2\alpha/3, \beta/3)$, from which we can recover π; this shows that we can recover π from π_1 (as long as we are restricted to rotation-invariant distributions), and so $G = F$ in the combinatorial construction. The combinatorial construction therefore shows how to zero variables so as to obtain T_N tensors, where $T_N \approx 2^{H(\pi_1)N}$. Feeding this to the asymptotic sum inequality, we obtain

$$2^{H(\alpha/3+2\beta/3, 2\alpha/3, \beta/3)} q^{(\alpha/3)\omega} \leq q+2.$$

For each given value of q, we can minimize ω over the choice of $\alpha + \beta = 1$ numerically (we will describe how this can be done systematically later on). We find out that the smallest value of ω is obtained for $q = 6$ and $\alpha \approx 0.952$, which gives the bound $\omega \leq 2.388$.

5 Milking the identity

All further improvements rely on the same identity just analyzed, dating to the STOC '87 paper if not earlier. How can this identity yield better bounds? So far we have considered the following process: starting with the basic identity, take a high tensor power, zero some of the variables so that whatever remains is a collection of disjoint matrix multiplication tensors, and apply the asymptotic sum inequality. We can improve on this by allowing non-disjoint matrix multiplication tensors, assuming that we can merge them together[1]. For example, consider the two tensors

$$\langle 1,1,q_1\rangle^{i,j_1,k_1} = \sum_{t=1}^{q_1} x^{[i]} y_t^{[j_1]} z_t^{[k_1]},$$
$$\langle 1,1,q_2\rangle^{i,j_2,k_2} = \sum_{t=1}^{q_2} x^{[i]} y_t^{[j_2]} z_t^{[k_2]},$$

having an identical x-index. If we sum them together then we get the tensor

$$x^{[i]} \left(\sum_{t=1}^{q_1} y_t^{[j_1]} z_t^{[k_1]} + \sum_{t=1}^{q_2} y_t^{[j_2]} z_t^{[k_2]} \right),$$

which is simply a manifestation of $\langle 1,1,q_1+q_2\rangle$. Without this merging, we can only keep one of the tensors above. With this merging, we can pick both of them together, though the merged version is not worth as much as two separate ones since $(q_1+q_2)^{\omega/3} < q_1^{\omega/3} + q_2^{\omega/3}$ (assuming $\omega < 3$). We gain since $(q_1+q_2)^{\omega/3} > \max(q_1^{\omega/3}, q_2^{\omega/3})$.

One principled way of applying this idea is by merging tensors as soon as possible: we take a small tensor power of the basic identity, merge some tensors, and then apply the construction. After taking the small tensor power, each index is now a *vector* of integers. However, for the combinatorial construction to work, we need integer indices whose sum is constant. One natural choice of such "coupling" is by letting the new index be the sum of the vector of indices. This folds some of the tensors together. When the folded tensors can be put together into one larger matrix multiplication tensor, we have gained something. However, the sum could also be a new kind of tensor which isn't a matrix multiplication tensor. We will handle such tensors by applying the construction recursively.

The smallest tensor power that can be considered is the square, and this case has been considered already by Coppersmith and Winograd. Successively larger power have been considered by other authors: Stothers [Sto, DS] considered the fourth power, Vassilevska-Williams [Wil] considered (independently) the fourth and eighth powers, and very recently Le Gall [Gal] considered the sixteenth and thirty-second power. The following table, taken from Le Gall's paper, summarizes the results obtained in this way:

Who	Power	Bound
C.-W.	1	$\omega < 2.3871900$
C.-W.	2	$\omega < 2.3754770$
Stothers	4	$\omega < 2.3729269$
V.-Williams	8	$\omega < 2.3728642$
Le Gall	16	$\omega < 2.3728640$
Le Gall	32	$\omega < 2.3728639$

[1] We thank Andris Ambainis for this observation.

6 Squared identity

We start by considering the simplest case, the square of the basic identity:

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

Opening up the square, we get 36 terms on the left-hand side, starting with $\langle 1,1,q\rangle^{0,1,1} \otimes \langle 1,1,q\rangle^{0,1,1} = \langle 1,1,q^2\rangle^{00,11,11}$. Each of these 36 terms are matrix multiplication tensors. In order to apply the combinatorial edifice, we will need to divide the new variables into groups with *integer* indices, in such a way that the indices in each tensor add up to the same number. One way of doing this is by putting variables $x^{[i_1,i_2]}$ in group $i_1 + i_2$, and so on: this guarantees that the sum is always 4. Continuing our example, the variables in $\langle 1,1,q^2\rangle^{00,11,11}$ are now going to belong to the groups $0,2,2$. There are two other tensors whose new grouping will be $0,2,2$: $\langle 1,1,1\rangle^{0,2,0} \otimes \langle 1,1,1\rangle^{0,0,2} = \langle 1,1,1\rangle^{00,20,02}$ and $\langle 1,1,1\rangle^{0,0,2} \otimes \langle 1,1,1\rangle^{0,2,0} = \langle 1,1,1\rangle^{00,02,20}$. Notice that all these three tensors share the same x variable, but otherwise the variables are disjoint. Therefore we can merge all three of them to a tensor $\langle 1,1,q^2+2\rangle^{0,2,2}$.

By employing a similar analysis to all the terms, we come up with the following terms:

(a) 3 terms similar to $\langle 1,1,1\rangle^{0,0,4}$, coming from

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

(b) 6 terms similar to $\langle 1,1,2q\rangle^{0,1,3}$, coming from

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

(c) 3 terms similar to $\langle 1,1,q^2+2\rangle^{0,2,2}$, coming from

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

(d) 3 terms similar to $T_4^{1,1,2}$, coming from

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

There is a problem here: if we group together all terms involving the new groups $1,1,2$, we don't get a matrix multiplication tensor, but rather

$$T_4 = \langle 1,q,1\rangle^{10,10,02} + \langle 1,q,1\rangle^{01,01,20} + \langle q,1,q\rangle^{10,01,11} + \langle q,1,q\rangle^{01,10,11}.$$

There are two other terms $\rho(T_4), \rho^2(T_4)$ similar to T_4 which differ only by rotation of the indices. (In general, the operator ρ rotates the tensor by applying the transformation $(x,y,z) \mapsto (y,z,x)$; for example $\rho(\langle n,m,p\rangle) = \langle m,p,n\rangle$.) We have retained the original indices to show which variables are the same and which are disjoint. We can retain the same information by considering just the first half of each index, because the two halves sum to a constant. The result is

$$T_4 = \langle 1,q,1\rangle^{1,1,0} + \langle 1,q,1\rangle^{0,0,2} + \langle q,1,q\rangle^{1,0,1} + \langle q,1,q\rangle^{0,1,1}.$$

The indices now sum to the original value 2.

Following the recipe, we should now take a high tensor power, zero some variables, and use the asymptotic sum inequality to get a bound on ω. This will be problematic since T_4 is not a matrix multiplication tensor: after zeroing the variables, we will have terms of the form $\langle n, m, p\rangle \otimes T_4^q \otimes \rho(T_4)^q \otimes \rho^2(T_4)^q$ (ostensibly, the three tensors $T_4, \rho(T_4), \rho^2(T_4)$ could have different powers, but this does not happen since the source distribution we choose is always rotation-invariant). The idea now is to zoom in on $(T_4 \otimes \rho(T_4) \otimes \rho^2(T_4))^q$, and zero variables so that disjoint matrix multiplication tensors are obtained. We already know how to do this: just apply the very same construction that got us here!

Value of a tensor. We now formalize this idea through the notion of *value*, which is what a given tensor gives you in view of applying the asymptotic sum inequality. Our goal is to generalize the asymptotic sum inequality to

$$\sum_i V_\omega(T_i) \leq \underline{\mathrm{R}}(\bigoplus_i T_i).$$

If all tensors T_i are matrix multiplication tensors $T_i = \langle n_i, m_i, p_i\rangle$, then putting $V_\omega(T_i) = (n_i m_i p_i)^{\omega/3}$ recovers the asymptotic sum inequality.

We define the value of a tensor in two steps. Suppose first that T is rotation-invariant: $T = \rho(T)$. For each N, we define $V_{\omega,N}$ as the maximum of $\sum_i (n_i m_i p_i)^{\omega/3}$ over all ways of zeroing variables to obtain $\bigoplus_i \langle n_i, m_i, p_i\rangle$ from $T^{\otimes N}$. It is easy to check that $V_{\omega,N_1N_2} \geq V_{\omega,N_1} V_{\omega,N_2}$, and so $V_{\omega,N}^{1/N}$ tends to a limit, which is the value $V_\omega(T)$.

When T is not rotation-invariant, we define $V_\omega(T) = (V_\omega(T \otimes \rho(T) \otimes \rho^2(T)))^{1/3}$. When T is rotation-invariant, both definitions coincide. The reason we consider $T \otimes \rho(T) \otimes \rho^2(T)$ is that the tensor T in fact appears in such a fashion in the construction, as we hinted above.

The generalized asymptotic sum inequality. The proof of the generalized asymptotic sum inequality isn't too hard. We start with a preliminary result.

Lemma 1. *For any two tensors T_1, T_2, $V_\omega(T_1 \otimes T_2) \geq V_\omega(T_1)V_\omega(T_2)$ and $V_\omega(T_1 \bigoplus T_2) \geq V_\omega(T_1) + V_\omega(T_2)$.*

Proof. The first part follows from the simple observation $V_{\omega,N}(T_1 \otimes T_2) \geq V_{\omega,N}(T_1)V_{\omega,N}(T_2)$.

For the second part, assume for simplicity that the tensors T_1, T_2 are rotation-invariant: otherwise, the symmetrized tensors $T_i \otimes \rho(T_i) \otimes \rho^2(T_i)$ should be considered instead. For every two rotation-invariant tensors S_1, S_2, $V_{\omega,N}(S_1 \bigoplus S_2) = V_{\omega,N}(S_1) + V_{\omega,N}(S_2)$. Therefore

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

When N_1, N_2 are large, which accounts for most of the sum, $V_{\omega,N_1}(T_1) \approx V_\omega(T_1)^{N_1}$ and $V_{\omega,N_2}(T_2) \approx V_\omega(T_2)^{N_2}$, and so the binomial theorem gives $V_{\omega,N}(T_1 + T_2) \approx (V_\omega(T_1) + V_\omega(T_2))^N$.

This argument can be formalized; the only step we need to justify is that large N_1, N_2 account for most of the sum. Indeed, the largest term in the sum accounts for a polynomial fraction of the sum, and this term is the one satisfying $N_1/N_2 \approx V_\omega(T_1)/V_\omega(T_2)$. Assuming the values are non-zero, this shows that a polynomial fraction of the sum is concentrated at a term with $N_1, N_2 = \Omega(N)$. □

In order to prove the generalized asymptotic sum inequality, it will be more convenient to state the asymptotic sum inequality in the following equivalent form:

$$\sum_i (n_i m_i p_i)^{\alpha/3} \geq \underline{\mathrm{R}}(\bigoplus_i T_i) \Longrightarrow \omega \leq \alpha,$$

and prove a similar statement for the value. Let $T = \bigoplus_i T_i$ and $S = T \otimes \rho(T) \otimes \rho^2(T)$. Suppose that

$$\sum_i V_\alpha(T_i) \geq \underline{\mathrm{R}}(T).$$

The lemma shows that $V_\alpha(T) \geq \underline{\mathrm{R}}(T)$, and so for large N,

$$V_{\alpha,N}(S) \gtrsim V_\alpha(S)^n = V_\alpha(T)^{3n} \geq \underline{\mathrm{R}}(T)^{3N} \geq \underline{\mathrm{R}}(S)^N.$$

The asymptotic inequality immediately implies that $\omega \lesssim \alpha$. In the limit, we obtain $\omega \leq \alpha$.

Lower bounding the value of a tensor. Having shown how the value can be used to deduce upper bounds on ω, we proceed to interpret the combinatorial construction as a lower bound on the value. When $T = \langle n, m, p \rangle$, we immediately get $V_\omega(T) \geq (nmp)^{\omega/3}$ (in fact there is equality). Suppose now that we have a tensor of the form

$$T = \sum_t T_t^{i_t, j_t, k_t}.$$

Each individual tensor $T_t^{i_t,j_t,k_t}$ is supported by x-variables having the (integer) index i_t, y-variables having the index j_t, and z_t-variables having the index k. The crucial property of the indices is that for $i_t \neq i_s$, the x-variables having index i_t are disjoint from the x-variables having index i_s. We also need the index triples to be tight: $i_t + j_t + k_t$ must be a constant independent of t.

Suppose first that T is rotation-invariant. We need to estimate $V_\omega(T)$. The definition of value easily implies that $V_\omega(T) = V_\omega(T^{\otimes N})^{1/N}$. For each source distribution π, the combinatorial construction shows that starting with $T^{\otimes N}$, there is a way of zeroing variables so as to obtain K_N copies the tensor $\bigotimes_t T_t^{\otimes N\pi(t)}$, where $K_N \approx 2^{N\kappa}$ for

$$\kappa = H(\pi_x) + H(\pi) - \max_{\sigma\colon \sigma_x = \pi_x} H(\sigma).$$

(Recall that π_x is the projection of π into the x-index.) Since the value is super-additive and super-multiplicative, this shows that

$$V_\omega(T)^N = V_\omega(T^{\otimes N}) \gtrsim 2^{N\kappa} \prod_t V_\omega(T_t)^{N\pi(t)}.$$

Therefore for each source distribution π, we obtain the following bound on $V_\omega(T)$:

$$\log_2 V_\omega(T) \geq H(\pi_x) + H(\pi) - \max_{\sigma\colon \sigma_x = \pi_x} H(\sigma) + \sum_t \pi(t) \log_2 V_\omega(T_t).$$

Maximizing this over π (a process which we address later on) gives us a lower bound on $V_\omega(T)$.

When T is not rotation-invariant, the definition requires us to examine $S = T \otimes \rho(T) \otimes \rho^2(T)$ instead. Every source distribution π for T lifts naturally to a source distribution π^S on S, given by

$\pi^S(t_1,t_2,t_3) = \pi(t_1)\pi(t_2)\pi(t_3)$ (here it is important to use the same indexing for the constituent tensors of $T, \rho(T), \rho^2(T)$). It is easy to check that $H(\pi^S) = 3H(\pi)$ and $H(\pi^S_x) = H(\pi_x) + H(\pi_y) + H(\pi_z)$. Similarly,

$$\begin{aligned}
&\sum_{t_1,t_2,t_3} \pi^S(t_1,t_2,t_3) \log_2 V_\omega(T_{t_1} \otimes \rho(T_{t_2}) \otimes \rho^2(T_{t_3})) \\
\geq &\sum_{t_1,t_2,t_3} \pi(t_1)\pi(t_2)\pi(t_3)(\log_2 V_\omega(T_{t_1}) + \log_2 V_\omega(T_{t_2}) + \log_2 V_\omega(T_{t_3})) \\
= &3\sum_t \pi(t) \log_2 V_\omega(T_t).
\end{aligned}$$

One last thing we need to determine is when a distribution σ^S satisfies $\sigma^S_x = \pi^S_x$. Suppose for the moment that we only considered lifted distributions. In that case, it is not hard to check that the requied condition is $\sigma_x = \pi_x$, $\sigma_y = \pi_y$ and $\sigma_z = \pi_z$, which we summarize by writing $\sigma \equiv \pi$, in words, σ is compatible with π. Assuming this, we get the lower bound

$$\log_2 V_\omega(T) = \frac{1}{3}\log_2 V_\omega(S) \geq \frac{H(\pi_x) + H(\pi_y) + H(\pi_z)}{3} + H(\pi) - \max_{\sigma \equiv \pi} H(\sigma) + \sum_t \pi(t) \log_2 V_\omega(T_t).$$

Unfortunately, while this lower bound indeed holds, its proof is slightly more complicated, involving a pre-filtering step. Instead of taking an Nth tensor power of S, we take an Nth tensor power of T, and zero out all variables other than those having projection type $N\pi$, obtaining a tensor T'. The number of index triples having projection type $N\pi$ is roughly $\max_{\sigma \equiv \pi} 2^{NH(\sigma)}$. We then consider the tensor $S' = T' \otimes \rho(T') \otimes \rho^2(T)$, which can be obtained from S by zeroing out variables. All index triples have projection type $N^3\pi^S$, and their number is approximately $2^{3NH(\sigma)}$. The construction now applies, and gives the stated lower bound.

Wrapping things up. For each value of α, we can now obtain a lower bound on $V_\alpha(T_4)$, since all the component tensors T_t are matrix multiplication tensors and so have known value. In fact, we can solve the required optimization explicitly, obtaining the lower bound $4^{1/3} q^\alpha (2 + q^{3\alpha})^{1/3}$. Given $V_\alpha(T_4)$, we can now optimize over the value of the folded squared identity. If the resulting value is at least $(q+2)^2$, then $\omega \leq \alpha$. Using binary search, we can find the best bound on ω that can be obtained in this way, which is $\omega \leq 2.376$, obtained for $q = 6$.

7 Optimizing the value lower bound

Our work so far culminated with the following lower bound on the value:

$$\log_2 V_\alpha(T) \geq \frac{H(\pi_x) + H(\pi_y) + H(\pi_z)}{3} + H(\pi) - \max_{\sigma \equiv \pi} H(\sigma) + \sum_t \pi(t) \log_2 V_\alpha(T_t).$$

Fix α, and suppose we have somehow calculated $V_\alpha(T_t)$. It remains to maximize the right-hand side of the inequality. We start by analyzing the condition $\sigma \equiv \pi$. Let the index triples in question be (i_t, j_t, k_t), and suppose that the possible values for i, j, k are X, Y, Z (respectively). We can write

the condition $\sigma \equiv \pi$ as a linear system:

$$\begin{aligned}
\sum_{t\colon i_t = x} (\sigma(t) - \pi(t)) &= 0 && \text{for all } x \in X,\\
\sum_{t\colon j_t = y} (\sigma(t) - \pi(t)) &= 0 && \text{for all } y \in Y,\\
\sum_{t\colon k_t = z} (\sigma(t) - \pi(t)) &= 0 && \text{for all } z \in Z,\\
\sum_{t} (\sigma(t) - \pi(t)) &= 0.
\end{aligned}$$

This is a linear system in $\sigma(t) - \pi(t)$. The easy case is when this linear system has no non-trivial solutions. This happens, for example, when computing $V_\alpha(T_4)$ and when computing the value of the folded squared identity. In this case, we can write

$$\log_2 V_\alpha(T) \geq \frac{H(\pi_x) + H(\pi_y) + H(\pi_z)}{3} + \sum_t \pi(t) \log_2 V_\alpha(T_t).$$

We want to maximize the right-hand side under the constraint that π is a probability distribution. Since the entropy function is concave, this is a convex optimization program, which can be solved efficiently using known methods such as gradient descent.

When the linear system in $\sigma(t) - \pi(t)$ has non-trivial solutions, we are not so lucky. In fact, it seems hard to compute the exact maximum. Instead, Stothers/Vassilevska-Williams [Sto, Wil] and Le Gall [Gal] offer two different methods to compute good lower bounds, which seem to perform quite well in practice. Both methods are optimal when the linear system has no non-trivial solutions.

Le Gall's method. While appearing after the other method, Le Gall's method is simpler and much more efficient. He suggests first computing

$$\pi^* = \operatorname*{argmax}_\pi \frac{H(\pi_x) + H(\pi_y) + H(\pi_z)}{3} + \sum_t \pi(t) \log_2 V_\alpha(T_t),$$

and then computing

$$\sigma^* = \operatorname*{argmax}_{\sigma \equiv \pi^*} H(\sigma).$$

Both maximizations are convex optimization programs, and so can be solved efficiently. The resulting lower bound is

$$\log_2 V_\alpha(T) \geq \frac{H(\pi^*_x) + H(\pi^*_y) + H(\pi^*_z)}{3} + H(\pi^*) - H(\sigma^*) + \sum_t \pi^*(t) \log_2 V_\alpha(T_t).$$

Stothers/Vassilveska-Williams' method. The idea of this method is to add the constraint $\pi = \operatorname{argmax}_{\sigma \equiv \pi} H(\sigma)$. We can express this constraint using Lagrange multipliers. Let V be the space of solutions of the linear system considered above. Given π, we want π to be a local maximum of $H(\pi)$ in a small neighborhood of the form $H(\pi + \epsilon V)$; concavity of H implies that a local maximum is in fact a global maximum. This means that $\nabla H(\pi) \in V^\perp$. Since $\nabla H(\pi)$ is the

vector $-1-\log\pi(t)$ and the all one vector is in $V^\perp$, this condition can be stated as $(\log\pi(t)) \in V^\perp$. (We are skipping a small technicality here, showing that all relevant probabilities must be positive so that this condition is well-defined.) The resulting optimization program is

$$\max_{\pi\colon (\log\pi(t))\in V^\perp} \frac{H(\pi_x)+H(\pi_y)+H(\pi_z)}{3} + \sum_t \pi(t)\log_2 V_\alpha(T_t).$$

This is not a convex optimization program, but can nevertheless be solved in practice using non-linear solvers. Compared to Le Gall's method, this method tends to give slightly better lower bounds on values, but runs much slower since the optimization problem that needs to be solved is more complicated.

8 Higher powers

It is time now to consider higher powers of the basic identity. Empirically it seems that it is best to consider powers which are themselves powers of two. The idea is to iterate the squaring operation, computing the values of all the tensors encountered along the way. For example, consider the squared identity, which has tensors of the forms $\langle 1,1,1\rangle^{0,0,4}$, $\langle 1,1,2q\rangle^{0,1,3}$, $\langle 1,1,q^2+2\rangle^{0,2,2}$, $T_4^{1,1,2}$, and their rotations. Squaring the squared identity, we can couple the indices again, and consider the folded tensor. Most of the tensors appearing in the new expression are complicated, but can be analyzed by calculating lower bounds on their values, since we already know the values of the constituent tensors. The advantage we gain is by merging some matrix multiplication tensors to larger matrix multiplication tensors.

In the case of the identity squared, the tensors $\langle 1,1,2q\rangle^{0,1,3}$ and $\langle 1,1,q^2+2\rangle^{0,2,2}$ resulted from such merging. Similarly, every index triple of the form $(0,\cdot,\cdot)$ comes from tensor products of the form $\langle 1,1,\cdot\rangle^{0,\cdot,\cdot} \times \langle 1,1,\cdots\rangle^{0,\cdot,\cdot}$, and so just like before, these can be merged. Vassilevska-Williams [Wil] calculates that for an index triple $(0,b,2^{r+1}-b)$, this merging results in a matrix multiplication tensor $\langle 1,1,B\rangle$ for

$$B = \sum_{\substack{e\in\{0,\ldots,b\}\\ e\equiv b \pmod 2}} \frac{2^r!}{e!(\frac{b-e}{2})!(2^r-\frac{b+e}{2})!} q^e.$$

For each α, given the values V_α for all the tensors appearing in the new expression (which can be calculated using the values of their constituent tensors), we can calculate the value of the entire expression. Using binary search, we find α such that this value exceeds $(q+2)^4$, and this gives an upper bound on ω. For $q=6$, we obtain $\omega < 2.374$ this way. This process can be repeated to obtain even better bounds on ω, resulting in the bounds reported above.

Open question 2. What is the bound on ω obtained by taking the limit of this construction?

We suspect that there is a different way of obtaining this limiting bound that combines tensors directly rather than by merging tensors in a small tensor power of the basic identity.

Open question 3. Are there other identities giving better bounds on ω?

Computer search might be helpful here.

References

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

[Beh] Felix Behrend, *On sets of integers which contain no three terms in arithmetic progression*, Proc. Nat. Acad. Sci. 32:331–332, 1946.

[Elk] Michael Elkin, *An improved construction of progression-free sets*, Israeli J. Math. 184:93–128, 2011.

[CKSU] Henry Cohn, Robert Kleinberg, Balázs Szegedy and Chris Umans, *Group-theoretic algorithms for matrix multiplication*, FOCS 2005.

[CW] Dan Coppersmith and Shmuel Winograd, *Matrix multiplication via arithmetic progressions*, J. Symb.Comput. 9(3):251–280, 1990.

[DS] A. M. Davie and A. J. Stothers, *Improved bound for complexity of matrix multiplication*, Proc. Roy. Soc. Edinburgh 143A:351–370, 2013.

[Gal] François Le Gall, *Powers of tensors and fast matrix multiplication*, arXiv:1401.771, 2014.

[Mos] Leo Moser, *On non-averaging sets of integers*, Canad. J. Math. 5:245–253, 1953.

[SS] Raphaël Salem and Donald Spencer, *On sets of integers which contain no three in arithmetic progressions*, Proc. Nat. Acad. Sci. 28:561–563, 1942.

[Sto] Andrew James Stothers, *On the complexity of matrix multiplication*, Ph.D. thesis (U. of Edinburgh), 2010.

[S1] Volker Strassen, *The asymptotic spectrum of tensors*, Crelles J. Reine Angew. Math. 384:102–152, 1988.

[S2] Volker Strassen, *Degeneration and complexity of bilinear maps: some asymptotic spectra*, Crelles. J. Reine Angew. Math. 413:127–180, 1991.

[Wil] Virginia Vassilevska-Williams, *Breaking the Coppersmith-Winograd barrier*, STOC 2012.