THE COMPLEXITY OF MATRIX MULTIPLICATION: DEVELOPMENTS SINCE 2014 EXTENDED ABSTRACT OF 2018 OBERWOLFACH COMPLEXITY MEETING PLENARY LECTURE

J.M. LANDSBERG

The complexity of all operations in linear algebra is governed by the complexity of matrix multiplication. In 1968 V. Strassen [32] discovered the way we usually multiply matrices is not the most efficient one and initiated the central problem of determining the complexity of matrix multiplication. He defined a fundamental constant ω, called the *exponent of matrix multiplication*, that governs its complexity. For a tensor $T \in \mathbb{C}^m \otimes \mathbb{C}^m \otimes \mathbb{C}^m$, let $\mathbf{R}(T)$ denote its tensor rank, the smallest r such that T may be written as a sum of r rank one tensors, and $\underline{\mathbf{R}}(T)$ its tensor border rank, the smallest r such that T may be written as a limit of a sequence of rank r tensors. Bini [6] proved the border rank of the matrix multiplication tensor $\underline{\mathbf{R}}(M_{\langle \mathbf{n} \rangle})$ asymptotically determines ω. More precisely, considering $\underline{\mathbf{R}}(M_{\langle \mathbf{n} \rangle})$ as a function of $\mathbf{n}$, $\omega = \inf_\tau \{\underline{\mathbf{R}}(M_{\langle \mathbf{n} \rangle}) = O(\mathbf{n}^\tau)\}$.

This talk has two goals: (i) report on progress in the last four years regarding upper and lower bounds for the complexity of matrix multiplication and tensors in general, and (ii) to explain the utility of algebraic geometry and representation theory for matrix multiplication and complexity theory in general.

Lower bounds. Strassen-Lickteig (1983, 1985) [30, 26] showed $\underline{\mathbf{R}}(M_{\langle \mathbf{n} \rangle}) \geq \frac{3\mathbf{n}^2}{2} + \frac{\mathbf{n}}{2} - 1$. Then, after 25 years without progress, Landsberg-Ottaviani (2013) [24] showed $\underline{\mathbf{R}}(M_{\langle \mathbf{n} \rangle}) \geq 2\mathbf{n}^2 - \mathbf{n}$. Around 2014 several authors [16, 15, 17] independently proved that the existing lower bound methods would not go much further. In [22] the *border substitution method* was developed, which led to the current best lower bound $\underline{\mathbf{R}}(M_{\langle \mathbf{n} \rangle}) \geq 2\mathbf{n}^2 - \log_2(\mathbf{n}) - 1$ [23].

The geometric approach to lower bounds is as follows: let $\sigma_r := \{T \in \mathbb{C}^m \otimes \mathbb{C}^m \otimes \mathbb{C}^m \mid \underline{\mathbf{R}}(T) \leq r\}$, the set of tensors in $\mathbb{C}^m \otimes \mathbb{C}^m \otimes \mathbb{C}^m$ of border rank at most r. This set is an *algebraic variety*, i.e., it is the zero set of a collection of (homogeneous) polynomials. Naïvely expressed, to prove $\underline{\mathbf{R}}(M_{\langle \mathbf{n} \rangle}) > r$, or to prove lower border rank bounds for any tensor, one simply looks for a polynomial in the ideal of σ_r (that is a polynomial P such that $P(T) = 0$ for all $T \in \sigma_r$) such that $P(M_{\langle \mathbf{n} \rangle}) \neq 0$ (here $m = \mathbf{n}^2$). But how can one find such polynomials? This is where *representation theory* comes in. The variety σ_r is *invariant* under changes of bases in the three spaces. That is, write $\mathbb{C}^m \otimes \mathbb{C}^m \otimes \mathbb{C}^m = A \otimes B \otimes C$, and let $GL(A)$ etc.. denote the invertible $m \times m$ matrices. There is a natural action of $G := GL(A) \times GL(B) \times GL(C)$ on $A \otimes B \otimes C$: on rank one tensors $(g_A, g_B, g_C) \cdot (a \otimes b \otimes c) := (g_A a) \otimes (g_B b) \otimes (g_C c)$, and the action on $A \otimes B \otimes C$ is defined by extending this action linearly. Then for all $g \in G$ and $x \in \sigma_r$, one has $g \cdot x \in \sigma_r$. Whenever a variety is invariant under the action of a group, its ideal is invariant under the group as well

2010 *Mathematics Subject Classification.* 68Q17; 14L30, 15A69.

Key words and phrases. Matrix multiplication complexity, Tensor rank, Asymptotic rank, Laser method.

Supported by NSF grants DMS-1405348 and AF-1814254. This work was partially supported by the grant 346300 for IMPAN from the Simons Foundation and the matching 2015-2019 Polish MNiSW fund as well as an Simons Visiting Professor grant supplied by the Simons Foundation and by the Mathematisches Forschungsinstitut Oberwolfach.

via the induced action on polynomials. One can then attempt to use representation theory to decompose the space of all polynomials and systematically check which irreducible modules are in the ideal. This works well in small dimensions, e.g., to show $\underline{\mathbf{R}}(M_{\langle 2\rangle}) = 7$ [18], but in general one must use additional methods. A classical approach is to try to embed $A\otimes B\otimes C$ into a space of matrices, and then take minors, which (in a slightly different context) dates back at least to Sylvester. The advance here was to look for *G-equivariant* (*G-homomorphic*) embeddings. This idea led to the 2013 advance, but the limits described in [16, 15, 17] exactly apply to such embeddings, so to advance further one must find new techniques.

At this point I should mention the general *hay in a haystack* problem of finding explicit sequences of tensors $T_m \in \mathbb{C}^m\otimes\mathbb{C}^m\otimes\mathbb{C}^m$ of high rank and border rank. The maximum border rank is $\lceil \frac{m^3}{3m-2}\rceil$ for $m > 3$, and the maximum possible rank is *not known*. The current state of the art are explicit tensors with $\mathbf{R}(T_m) \geq 3m - o(m)$ [1, 19] and others with $\underline{\mathbf{R}}(T_m) \geq 2m - 2$ [20]. The border substitution method promises to at least improve the state of the art on these problems.

A last remark on lower bounds: in the past two months there has been a very exciting breakthrough due to Buczynska-Buczynski (personal communication), that avoids the above-mentioned barriers in the polynomial situation. The method is a combination of the classical apolarity method with the border substitution method. Buczynski and I are currently working to extend these methods to the tensor situation.

Upper bounds. The *Kronecker product* of $T \in A\otimes B\otimes C$ and $T' \in A'\otimes B'\otimes C'$, denoted $T \boxtimes T'$, is the tensor $T\otimes T' \in (A\otimes A')\otimes(B\otimes B')\otimes(C\otimes C')$ regarded as 3-way tensor. The Kronecker powers of T, $T^{\boxtimes N} \in (A^{\otimes N})\otimes(B^{\otimes N})\otimes(C^{\otimes N})$ are defined similarly. Also let $T \oplus T' \in (A \oplus A')\otimes(B \oplus B')\otimes(C \oplus C')$ denote the direct sum. Let $M_{\langle \mathbf{l},\mathbf{m},\mathbf{n}\rangle}$ denote the rectangular matrix multiplication tensor. The matrix multiplication tensor has the remarkable property that $M_{\langle \mathbf{l},\mathbf{m},\mathbf{n}\rangle} \boxtimes M_{\langle \mathbf{l}',\mathbf{m}',\mathbf{n}'\rangle} = M_{\langle \mathbf{l}\mathbf{l}',\mathbf{m}\mathbf{m}',\mathbf{n}\mathbf{n}'\rangle}$ which is the key to Strassen's *laser method*.

Following work of Strassen, Bini [6] showed that for all $\mathbf{l}, \mathbf{m}, \mathbf{n}$, setting $q = (\mathbf{lmn})^{\frac{1}{3}}$, that

$$\omega \leq \frac{\log(\underline{\mathbf{R}}(M_{\langle \mathbf{l},\mathbf{m},\mathbf{n}\rangle}))}{\log(q)}.$$

This was generalized by Schönhage [28]. A special case is as follows: say that for $1 \leq i \leq s$, $\mathbf{l}_i\mathbf{m}_i\mathbf{n}_i = q^3$, then

$$\omega \leq \frac{\log(\frac{1}{s}\underline{\mathbf{R}}(\bigoplus_{i=1}^s M_{\langle \mathbf{l}_i,\mathbf{m}_i,\mathbf{n}_i\rangle}))}{\log(q)}.$$

Schönhage also showed that border rank can be strictly sub-additive, so the result is nontrivial. This immediately implies that if T is a tensor such that $\bigoplus_{i=1}^s M_{\langle \mathbf{l}_i,\mathbf{m}_i,\mathbf{n}_i\rangle} \in \overline{G\cdot T}$, i.e., T *degenerates* to $\bigoplus_{i=1}^s M_{\langle \mathbf{l}_i,\mathbf{m}_i,\mathbf{n}_i\rangle}$, then

$$\omega \leq \frac{\log(\frac{1}{s}\underline{\mathbf{R}}(T))}{\log(q)}.$$

This can be useful if the border rank of T is easier to estimate than that of the direct sum of matrix multiplication tensors. One should think of $\underline{\mathbf{R}}(T)$ as the *cost* of T and s and $\log(q)$ as determining the *value* of T. One gets a good upper bound if cost is low and value is high. Strassen [31] then showed that the same result holds if the matrix multiplication tensors are nearly disjoint (i.e., nearly direct sums) by taking Kronecker powers of T and degenerating the powers to disjoint matrix multiplication tensors. This method was used by Coppersmith and

Winograd [14] with the now named "little Coppersmith-Winograd tensor":

$$T_{cw,q} := \sum_{j=1}^{q} a_0 \otimes b_j \otimes c_j + a_j \otimes b_0 \otimes c_j + a_j \otimes b_j \otimes c_0,$$

to show

$$\omega \leq \frac{\log(\frac{4}{27}\underline{\mathbf{R}}(T_{cw,q})^3)}{\log(q)}.$$

Since $\underline{\mathbf{R}}(T_{cw,q}) = q + 2$, this implies $\omega < 2.41$ when $q = 8$. They improved this to $\omega \leq 2.3755$ using a slightly more complicated tensor (the Kronecker square of the big Coppersmith-Winograd tensor $T_{CW,q}$) which held the world record until 2012-3, when it was lowered to $\omega \leq 2.373$ [29, 33, 25] using higher Kronecker powers of $T_{CW,q}$. Then in 2014, Ambainus, Filmus and LeGall [4] proved that coordinate restrictions of Kronecker powers of the big Coppersmith-Winograd tensor could never be used to prove $\omega < 2.3$. This was generalized in [3, 2] to a larger class of tensors and degenerations, albeit with weaker bounds on the limitations. Regarding Kronecker powers, one also has

$$\omega \leq \frac{\log(\frac{4}{27}\underline{\mathbf{R}}(T_{cw,q}^{\boxtimes k})^{\frac{3}{k}})}{\log(q)}, \tag{1}$$

for any k.

I point out that all the above has little to do with practical matrix multiplication, the only better practical decomposition to arise since Strassen's 1968 work is due to V. Pan [27].

Given the barriers from [4, 3, 2], it makes sense to ask what geometry can do for upper bounds.

First idea: study known rank decompositions of $M_{\langle \mathbf{n} \rangle}$ to obtain new ones. For a tensor T, let $G_T := \{g \in G \mid g \cdot T = T\}$, denote the *symmetry group* of T. For example $G_{M_{\langle \mathbf{n} \rangle}}$ is the image of $GL_{\mathbf{n}}^{\times 3}$ in $GL_{\mathbf{n}^2}^{\times 3}$. In [9, 5] we studied decompositions and noticed that many of the decompositions had large *symmetry groups*, where if $\mathcal{S}_T$ is a rank decomposition of a tensor T, if one applies an element of G_T to the decomposition, it takes it to another rank decomposition of T, which sometimes is the same as the original. Let $\Gamma_{\mathcal{S}_T} := \{g \in G_T \mid g\mathcal{S}_T = \mathcal{S}_T\}$, denote the symmetry group of the decomposition. Then the decomposition may be expressed in terms of the orbit structure. For example, Strassen's 1968 rank 7 decomposition of $M_{\langle 2 \rangle}$ may be written

$$M_{\langle 2 \rangle} = \mathrm{Id}^{\otimes 3} + \Gamma \cdot \left[\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \right]$$

where $\Gamma \simeq \mathfrak{S}_3 \rtimes \mathbb{Z}_2$, and $\Gamma \cdot$ denotes the sum of the terms in the Γ-orbit. Here the orbit consists of six terms. The identity is acted on trivially by Γ, so the decomposition is a union of two orbits. This is work in progress.

Second idea: expand the playing field. Given any tensor, one can symmetrize it to get a cubic polynomial. Let $sM_{\langle \mathbf{n} \rangle}$ denote the symmetrized matrix multiplication tensor, the polynomial $X \mapsto \mathrm{trace}(X^3)$. In [8] we showed that the Waring rank of $sM_{\langle \mathbf{n} \rangle}$ also governs the exponent of matrix multiplication, where the Waring rank of a cubic polynomial is the smallest r such that the polynomial may be written as a sum of r cubes.

Third idea: combine the first two. A Conner [11] found a remarkable Waring rank 18 decomposition of $sM_{\langle 3 \rangle}$, with symmetry group that of the Hasse diagram, namely $(\mathbb{Z}_3^{\times 2} \rtimes SL_2(\mathbb{F}_3)) \rtimes \mathbb{Z}_2$. He also found a Waring rank 40 decomposition of $sM_{\langle 4 \rangle}$ with symmetry group that of the cube. For comparison, the best known rank decompositions of $M_{\langle 3 \rangle}, M_{\langle 4 \rangle}$ respectively are of ranks 23 and 49. This launched his program to find explicit sequences of finite groups $\Gamma_{\mathbf{n}} \subset G_{sM_{\langle \mathbf{n} \rangle}}$ such that the space of $\Gamma_{\mathbf{n}}$-invariants in the space of cubic polynomials in $\mathbf{n}^2$ variables only contains polynomials of low Waring rank, translating the study of upper bounds on ω to a study of properties of sequences of finite groups, in the spirit of (but very different from) the Cohn-Umans program [10].

Fourth idea: the Ambainus-Filmus-LeGall challenge: find new tensors useful for the laser method. Michalek and I [21] had the idea to isolate *geometric* properties of the Coppersmith-Winograd tensors and to find other tensors with similar geometric properties, in the hope that they might also be useful for the laser method. We succeeded in isolating many interesting geometric properties. Unfortunately, we then proved that the Coppersmith-Winograd tensors were the *unique* tensors with such properties.

Fourth idea, second try: In [13] we examine the symmetry groups of the Coppersmith-Winograd tensors. We found that the big Coppersmith-Winograd tensor has the largest dimensional symmetry group among 1-*generic tensors* in odd dimensions (1-genericity is a natural condition for implementing the laser method), but that in even dimensions, there is an even better tensor, which we call the *skew big Coppersmith-Winograd tensor*. We also found other tensors with large symmetry groups. Unfortunately, none of the new tensors with maximal or near maximal symmetry groups are better for the laser method than $T_{CW,q}$.

Fifth idea: Go back to the inequality (1) and upper bound Kronecker powers of $T_{cw,q}$ (this was posed as an open question for the square as early as [7]), which could even (when $q = 2$) potentially show $\omega = 2$. Unfortunately, we show in [12] that $15 \leq \underline{\mathbf{R}}(T_{cw,2}^{\boxtimes 2}) \leq 16$, and we expect 16, and for $q > 2$, that $\underline{\mathbf{R}}(T_{cw,q}^{\boxtimes 2}) = (q+2)^2$.

Sixth idea: Combine the last two ideas. The skew cousin of the little Coppersmith-Winograd tensor also satisfies (1). It has the same value, but unfortunately, it has higher cost. For example, $\underline{\mathbf{R}}(T_{skew-cw,2}) = 5 > 4 = \underline{\mathbf{R}}(T_{cw,2})$. However, we show it satisfies $\underline{\mathbf{R}}(T_{skew-cw,2}^{\boxtimes 2}) = 17 \ll 25 = \underline{\mathbf{R}}(T_{skew-cw,2})^2$. This is one of the few explicit tensors known to have strictly submultiplicative border rank under Kronecker square (the first known being $M_{\langle 2\rangle}$), and if this drop continues, it would be very good indeed for the laser method.

References

1. Boris Alexeev, Michael A. Forbes, and Jacob Tsimerman, *Tensor rank: some lower and upper bounds*, 26th Annual IEEE Conference on Computational Complexity, IEEE Computer Soc., Los Alamitos, CA, 2011, pp. 283–291. MR 3025382
2. J. Alman and V. Vassilevska Williams, *Limits on All Known (and Some Unknown) Approaches to Matrix Multiplication*, ArXiv e-prints (2018).
3. Josh Alman and Virginia Vassilevska Williams, *Further limitations of the known approaches for matrix multiplication*, 9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA, 2018, pp. 25:1–25:15.
4. Andris Ambainis, Yuval Filmus, and François Le Gall, *Fast matrix multiplication: limitations of the Coppersmith-Winograd method (extended abstract)*, STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing, ACM, New York, 2015, pp. 585–593. MR 3388238
5. Grey Ballard, Christian Ikenmeyer, J.M. Landsberg, and Nick Ryder, *The geometry of rank decompositions of matrix multiplication ii: 3x3 matrices*, to appear in JPAA.
6. D. Bini, *Relations between exact and approximate bilinear algorithms. Applications*, Calcolo **17** (1980), no. 1, 87–97. MR 605920 (83f:68043b)
7. Markus Bläser, *Fast matrix multiplication*, Graduate Surveys, no. 5, Theory of Computing Library, 2013.
8. Luca Chiantini, Jonathan D. Hauenstein, Christian Ikenmeyer, Joseph M. Landsberg, and Giorgio Ottaviani, *Polynomials and the exponent of matrix multiplication*, Bull. Lond. Math. Soc. **50** (2018), no. 3, 369–389. MR 3829726
9. Luca Chintini, Christian Ikenmeyer, J.M. Landsberg, and Giorgio Ottaviani, *The geometry of rank decompositions of matrix multiplication i: Strassen's algorithm*, to appear in Exper. Math.

10. H Cohn and C. Umans, *A group theoretic approach to fast matrix multiplication*, Proceedings of the 44th annual Symposium on Foundations of Computer Science (2003), no. 2, 438–449.
11. A. Conner, *A rank 18 Waring decomposition of $sM_\langle 3\rangle$ with 432 symmetries*, ArXiv e-prints, to appear in Exper. Math. (2017).
12. Austin Conner, Fulvio Gesmundo, J.M. Landsberg, and Emanuele Ventura, *Kronecker powers of tensors with symmetry*, preprint.
13. ———, *Tensors with maximal symmetries*, preprint.
14. Don Coppersmith and Shmuel Winograd, *Matrix multiplication via arithmetic progressions*, J. Symbolic Comput. **9** (1990), no. 3, 251–280. MR 91i:68058
15. Harm Derksen and Visu Makam, *On non-commutative rank and tensor rank*, Linear Multilinear Algebra **66** (2018), no. 6, 1069–1084. MR 3781583
16. K. Efremenko, A. Garg, R. Oliveira, and A. Wigderson, *Barriers for Rank Methods in Arithmetic Complexity*, ArXiv e-prints (2017).
17. Maciej Galcazka azka, *Vector bundles give equations of cactus varieties*, Linear Algebra Appl. **521** (2017), 254–262. MR 3611482
18. Jonathan D. Hauenstein, Christian Ikenmeyer, and J. M. Landsberg, *Equations for lower bounds on border rank*, Exp. Math. **22** (2013), no. 4, 372–383. MR 3171099
19. J. M. Landsberg, *New lower bounds for the rank of matrix multiplication*, SIAM J. Comput. **43** (2014), no. 1, 144–149. MR 3162411
20. ———, *Nontriviality of equations and explicit tensors in $\mathbb{C}^m \otimes \mathbb{C}^m \otimes \mathbb{C}^m$ of border rank at least $2m-2$*, J. Pure Appl. Algebra **219** (2015), no. 8, 3677–3684. MR 3320240
21. J. M. Landsberg and Mateusz Michalek, *Abelian tensors*, J. Math. Pures Appl. (9) **108** (2017), no. 3, 333–371. MR 3682743
22. ———, *On the geometry of border rank decompositions for matrix multiplication and other tensors with symmetry*, SIAM J. Appl. Algebra Geom. **1** (2017), no. 1, 2–19. MR 3633766
23. Joseph M. Landsberg and Mateusz Michał ek, *A $2n^2 - \log_2(n) - 1$ lower bound for the border rank of matrix multiplication*, Int. Math. Res. Not. IMRN (2018), no. 15, 4722–4733. MR 3842382
24. Joseph M. Landsberg and Giorgio Ottaviani, *New lower bounds for the border rank of matrix multiplication*, Theory Comput. **11** (2015), 285–298. MR 3376667
25. François Le Gall, *Powers of tensors and fast matrix multiplication*, Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation (New York, NY, USA), ISSAC '14, ACM, 2014, pp. 296–303.
26. Thomas Lickteig, *A note on border rank*, Inform. Process. Lett. **18** (1984), no. 3, 173–178. MR 86c:68040
27. Victor Pan, *How to multiply matrices faster*, Lecture Notes in Computer Science, vol. 179, Springer-Verlag, Berlin, 1984. MR MR765701 (86g:65006)
28. A. Schönhage, *Partial and total matrix multiplication*, SIAM J. Comput. **10** (1981), no. 3, 434–455. MR MR623057 (82h:68070)
29. A. Stothers, *On the complexity of matrix multiplication*, PhD thesis, University of Edinburgh, 2010.
30. V. Strassen, *Rank and optimal computation of generic tensors*, Linear Algebra Appl. **52/53** (1983), 645–685. MR 85b:15039
31. ———, *Relative bilinear complexity and matrix multiplication*, J. Reine Angew. Math. **375/376** (1987), 406–443. MR MR882307 (88h:11026)
32. Volker Strassen, *Gaussian elimination is not optimal*, Numer. Math. **13** (1969), 354–356. MR 40 #2223
33. Virginia Williams, *Breaking the coppersimith-winograd barrier*, preprint.

Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, USA
E-mail address: jml@math.tamu.edu