Efficient matrix multiplication

J.M. Landsberg

Texas A&M University

Supported by NSF grant CCF-1814254

Strassen's spectacular failure

Standard algorithm for matrix multiplication, row-column:

$$\begin{pmatrix} * & * & * \\ & & \\ & & \end{pmatrix} \begin{pmatrix} * & & \\ * & & \\ * & & \end{pmatrix} = \begin{pmatrix} * & & \\ & & \\ & & \end{pmatrix}$$

uses $O(n^3)$ arithmetic operations.

Strassen (1968) set out to prove this standard algorithm was indeed the best possible.

At least for 2×2 matrices. At least over $\mathbb{F}_2$.

He failed.

Strassen's algorithm

Let A, B be 2×2 matrices $A = \begin{pmatrix} a_1^1 & a_2^1 \\ a_1^2 & a_2^2 \end{pmatrix}$, $B = \begin{pmatrix} b_1^1 & b_2^1 \\ b_1^2 & b_2^2 \end{pmatrix}$. Set

$$
\begin{aligned}
I &= (a_1^1 + a_2^2)(b_1^1 + b_2^2), \\
II &= (a_1^2 + a_2^2) b_1^1, \\
III &= a_1^1 (b_2^1 - b_2^2) \\
IV &= a_2^2 (-b_1^1 + b_1^2) \\
V &= (a_1^1 + a_2^1) b_2^2 \\
VI &= (-a_1^1 + a_1^2)(b_1^1 + b_2^1), \\
VII &= (a_2^1 - a_2^2)(b_1^2 + b_2^2),
\end{aligned}
$$

If $C = AB$, then

$$
\begin{aligned}
c_1^1 &= I + IV - V + VII, \\
c_1^2 &= II + IV, \\
c_2^1 &= III + V, \\
c_2^2 &= I + III - II + VI.
\end{aligned}
$$

Astounding conjecture

Iterate: $\rightsquigarrow 2^k \times 2^k$ matrices using $7^k \ll 8^k$ multiplications,

and $n \times n$ matrices with $O(n^{2.81})$ arithmetic operations.

Bini 1978, Schönhage 1983, Strassen 1987, Coppersmith-Winograd 1988 $\rightsquigarrow O(n^{2.3755})$ arithmetic operations.

Astounding Conjecture

For all $\epsilon > 0$, $n \times n$ matrices can be multiplied using $O(n^{2+\epsilon})$ arithmetic operations.

$\rightsquigarrow$ *asymptotically, multiplying matrices is nearly as easy as adding them!*

1988-2011 no progress, 2011-14 Stouthers,Williams,LeGall $O(n^{2.373})$ arithmetic operations.

Tensor formulation of conjecture

Set $N = n^2$.
Matrix multiplication is a *bilinear* map

$$M_{\langle n\rangle} : \mathbb{C}^N \times \mathbb{C}^N \to \mathbb{C}^N,$$

Bilinear maps $\mathbb{C}^N \times \mathbb{C}^N \to \mathbb{C}^N$ may also be viewed as trilinear

maps $\mathbb{C}^N \times \mathbb{C}^N \times \mathbb{C}^{N*} \to \mathbb{C}$.

In other words

$$M_{\langle n\rangle} \in \mathbb{C}^{N*} \otimes \mathbb{C}^{N*} \otimes \mathbb{C}^N.$$

Exercise: As a trilinear map, $M_{\langle n\rangle}(X, Y, Z) = \mathrm{trace}(XYZ)$.

Tensor formulation of conjecture

A tensor $T \in \mathbb{C}^N \otimes \mathbb{C}^N \otimes \mathbb{C}^N =: A \otimes B \otimes C$ has *rank one* if it is of the form $T = a \otimes b \otimes c$, with $a \in A$, $b \in B$, $c \in C$. Rank one tensors correspond to bilinear maps that can be computed using one scalar multiplication.

The *rank* of a tensor T, $\mathbf{R}(T)$, is the smallest r such that T may be written as a sum of r rank one tensors. The rank is essentially the number of scalar multiplications needed to compute the corresponding bilinear map.

standard presentation is $M_{\langle n\rangle} = \sum_{i,j,k=1}^n x^i_j \otimes y^j_k \otimes z^k_i$
Strassen's presentation is

$$\begin{aligned}
M_{\langle 2\rangle} =& x^1_1 \otimes y^1_1 \otimes z^1_1 \\
&+ (-x^1_2 + x^2_1 - x^2_2) \otimes (-y^1_2 + y^2_1 - y^2_2) \otimes (-z^1_2 + z^2_1 - z^2_2) \\
&+ (x^1_2 + x^2_2) \otimes (y^1_2 + y^2_2) \otimes (z^1_2 + z^2_2) \\
&+ (-x^2_1 + x^2_2) \otimes (-y^2_1 + y^2_2) \otimes (-z^2_1 + z^2_2) \\
&+ \mathbb{Z}_3 \cdot [x^1_2 \otimes y^2_1 \otimes (z^1_1 - z^1_2 + z^2_1 - z^2_2)]
\end{aligned}$$

Tensor formulation of conjecture

Theorem (Strassen): $M_{\langle n \rangle}$ can be computed using $O(n^\tau)$ arithmetic operations $\Leftrightarrow$ $\mathbf{R}(M_{\langle n \rangle}) = O(n^\tau)$

Let $\omega := \inf_\tau \{\mathbf{R}(M_{\langle n \rangle}) = O(n^\tau)\}$

ω is called the *exponent* of matrix multiplication.

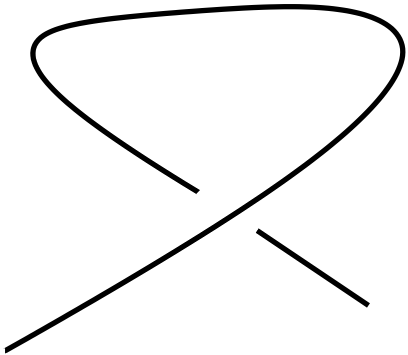
Classical: $\omega \leq 3$.

Corollary of Strassen's algorithm: $\omega \leq \log_2(7) \simeq 2.81$.

Astounding Conjecture

$\omega = 2$

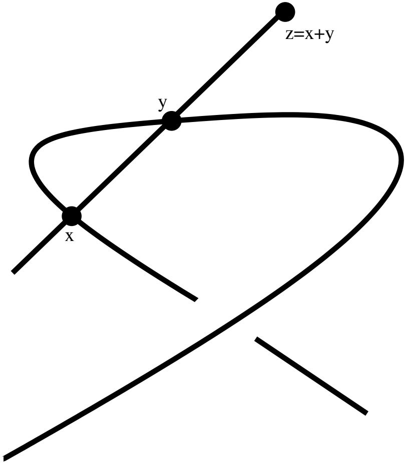
Geometric formulation of conjecture



Imagine this curve represents the set of tensors of rank one sitting in the N^3 dimensional space of tensors.

Geometric formulation of conjecture

$\{$ tensors of rank at most two$\}$ $=$
$\{$ points on a secant line to set of tensors of rank one$\}$



Conjecture is about a point (matrix multiplication) lying on a secant r-plane to set of tensors of rank one.

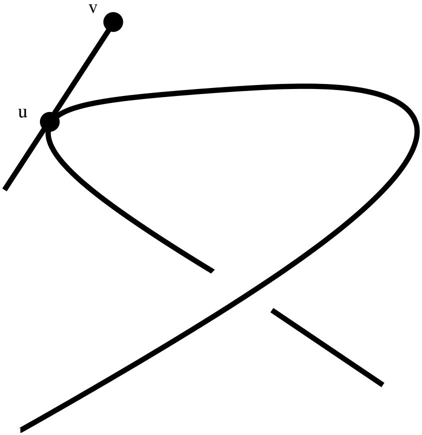
Bini's sleepless nights

Bini-Capovani-Lotti-Romani (1979) investigated if $M_{\langle 2 \rangle}$, with one matrix entry set to zero, could be computed with five multiplications (instead of the naïve 6), i.e., if this reduced matrix multiplication tensor had rank 5.

They used numerical methods.

Their code appeared to have a problem.

The limit of secant lines is a tangent line!



For $T \in \mathbb{C}^N \otimes \mathbb{C}^N \otimes \mathbb{C}^N$, the *border rank of T* $\underline{\mathbf{R}}(T)$ denotes the smallest r such that T is a limit of tensors of rank r.
Theorem (Bini 1980) $\underline{\mathbf{R}}(M_{\langle n \rangle}) = O(n^\omega)$, so border rank is also a legitimate complexity measure.

Wider geometric perspective

Let $X \subset \mathbb{CP}^M$ be a projective variety.

Our case: $M = N^3 - 1$,
$X = Seg(\mathbb{P}^{N-1} \times \mathbb{P}^{N-1} \times \mathbb{P}^{N-1}) \subset \mathbb{P}(\mathbb{C}^N \otimes \mathbb{C}^N \otimes \mathbb{C}^N)$.

Stratify $\mathbb{CP}^M$ by a sequence of nested varieties

$$X \subset \sigma_2(X) \subset \sigma_3(X) \subset \cdots \subset \sigma_f(X) = \mathbb{CP}^M$$

where

$$\sigma_r(X) := \overline{\cup_{x_1,\ldots,x_r \in X} \mathrm{span}\{x_1, \ldots, x_r\}}$$

is the variety of secant $\mathbb{P}^{r-1}$'s to X.

Secant varieties have been studied for a long time.

In 1911 Terracini could have predicted Strassen's discovery:
$\sigma_7(Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)) = \mathbb{P}^{63}$.

How to *disprove* astounding conjecture?

Let $\sigma_r \subset \mathbb{C}^N \otimes \mathbb{C}^N \otimes \mathbb{C}^N = \mathbb{C}^{N^3}$: tensors of border rank at most r.

Find a polynomial P (in N^3 variables) in the ideal of σ_r, i.e., such that $P(T) = 0$ for all $T \in \sigma_r$.

Show that $P(M_{\langle n \rangle}) \neq 0$.

Embarassing (?): had not been known even for $M_{\langle 2 \rangle}$, i.e., for σ_6 when $N = 4$.

Arora and Barak: lower bounds are "**complexity theory's Waterloo** "

Why did I think this would be easy?: Representation Theory

Matrices of rank at most r: zero set of size $r+1$ minors.

Tensors of border rank at most 1: zero set of size 2 minors of flattenings tensors to matrices: $A \otimes B \otimes C = (A \otimes B) \otimes C$.

Tensors of border rank at most 2: zero set of degree 3 polynomials.

Representation theory: systematic way to search for polynomials.

2004 L-Manivel: No polynomials in ideal of σ_6 of degree less than 12

2013 Hauenstein-Ikenmeyer-L: No polynomials in ideal of σ_6 of degree less than 19. However there are polynomials of degree 19. Caveat: too complicated to evaluate on $M_{\langle 2\rangle}$. Good news: easier polynomial of degree 20 (trivial representation) $\rightsquigarrow$ (L 2006, Hauenstein-Ikenmeyer-L 2013) $\underline{\mathbf{R}}(M_{\langle 2\rangle}) = 7$.

Polynomials via a retreat to linear algebra

$T \in A \otimes B \otimes C = \mathbb{C}^N \otimes \mathbb{C}^N \otimes \mathbb{C}^N$ may be recovered from the linear space $T(C^*) \subset A \otimes B$.
tensors up to changes of bases $\sim$ linear subspaces of spaces of matrices up to changes of bases.
Even better than linear maps are endomorphisms. Assume $T(C^*) \subset A \otimes B$ contains an element of full rank. Use it to obtain an isomorphism $A \otimes B \simeq \mathrm{End}(A) \rightsquigarrow$ space of endomorphisms.
$\mathbf{R}(T) = N \Leftrightarrow N$-dimensional space of simultaneously diagonalizable matrices

Good News: Classical linear algebra!

Bad News: Open question.

Retreat to linear algebra, cont'd

Simultaneously diagonalizable matrices $\Rightarrow$ commuting matrices

Good news: Easy to Test.

Better news (Strassen): Can upgrade to tests for higher border rank than N: $\underline{\mathbf{R}}(T) \geq N + \frac{1}{2}$(rank of commutator)

$\rightsquigarrow$ (Strassen 1983) $\underline{\mathbf{R}}(M_{\langle n \rangle}) \geq \frac{3}{2}\mathbf{n}^2$

Variant: (Lickteig 1985) $\underline{\mathbf{R}}(M_{\langle n \rangle}) \geq \frac{3}{2}\mathbf{n}^2 + \frac{\mathbf{n}}{2} - 1$

1985-2012: no further progress other than for $M_{\langle 2 \rangle}$.

Retreat to linear algebra, cont'd

Perspective: Strassen mapped space of tensors to space of matrices, found equations by taking minors.

Classical trick in algebraic geometry to find equations via minors.
$\rightsquigarrow$ (L-Ottaviani 2013) $\underline{\mathbf{R}}(M_{\langle n \rangle}) \geq 2\mathbf{n}^2 - \mathbf{n}$

For those familiar with representation theory: found via a $G = GL(A) \times GL(B) \times GL(C)$ module map from $A \otimes B \otimes C$ to a space of matrices (systematic search possible).

Punch line: Found equations by exploiting symmetry of σ_r

Note: Only gives good bounds if $\dim B \sim \dim C \geq \dim A$.
E.g. says nothing new for $M_{\langle 2nn \rangle}$, $M_{\langle 3nn \rangle}$ for $n > 3$.

Bad News: Barriers

Theorem (Bernardi-Ranestad,Buczynski-Galcazka,Efremenko-Garg-Oliviera-Wigderson): Game (almost) over for determinantal methods.

For the experts: Variety of zero dimensional schemes of length r is not irreducible $r > 13$.

Determinantal methods detect zero dimensional schemes (want zero dimensional smoothable schemes).

$$\sigma_r(X) := \overline{\bigcup \{\langle R\rangle \mid \text{length}(R) = r,\ \text{support}(R) \subset X,\ R : \text{smoothable}\}}$$

secant variety.
Let

$$\kappa_r(X) := \overline{\bigcup \{\langle R\rangle \mid \text{length}(R) = r,\ \text{support}(R) \subset X\}}$$

cactus variety.

Determinantal equations are equations for the cactus variety.

Punch line: **Barrier** to progress.

How to go further?

So far, lower bounds via symmetry of σ_r.

The matrix multiplication tensor also has symmetry:

$T \in A \otimes B \otimes C$, define *symmetry group of T*
$G_T := \{g \in GL(A) \times GL(B) \times GL(C) \mid g \cdot T = T\}$

$GL_{\mathbf{n}}^{\times 3} \subset G_{M_{\langle n \rangle}} \subset GL_{\mathbf{n}^2}^{\times 3} = GL(A) \times GL(B) \times GL(C)$:

Proof: $(g_1, g_2, g_3) \in GL_{\mathbf{n}}^{\times 3}$

$$\mathrm{trace}(XYZ) = \mathrm{trace}((g_1 X {g_2}^{-1})(g_2 Y {g_3}^{-1})(g_3 Z {g_1}^{-1}))$$

How to exploit G_T?

Given $T \in A \otimes B \otimes C$
$\underline{\mathbf{R}}(T) \leq r \Leftrightarrow \exists$ curve $E_t \subset G(r, A \otimes B \otimes C)$ such that
i) For $t \neq 0$, E_t is spanned by r rank one elements.
ii) $T \in E_0$.

For all $g \in G_T$, gE_t also works.
$\rightsquigarrow$ (L-Michalek 2017) can insist on normalized curves (for $M_{\langle n \rangle}$, those with E_0 Borel fixed).

$\rightsquigarrow \underline{\mathbf{R}}(M_{\langle n \rangle}) \geq 2\mathbf{n}^2 - \log_2 \mathbf{n} - 1$

More bad news: this method cannot go much further.

New idea, following Buczynska-Buczynski

Use more algebraic geometry: Consider not just curve of r points, but the curve of **ideals** $I_t \in Sym(A^* \oplus B^* \oplus C^*)$ it gives rise to: *border apolarity method*

$T = \lim_{t \to 0} \sum_{j=1}^r T_{j,t}$

I_t ideal of $[T_{1,t}] \cup \cdots \cup [T_{r,t}] \subset \mathbb{P}A \times \mathbb{P}B \times \mathbb{P}C$

Can insist that limiting ideal I_0 is Borel fixed: reduces to small search in each multi-degree.

Instead of single curve $E_t \subset G(r, A{\otimes}B{\otimes}C)$ limiting to Borel fixed point, for each (i, j, k) get curve in $G(r, S^iA^*{\otimes}S^jB^*{\otimes}S^kC^*)$, each limiting to Borel fixed point *and* satisfying compatibility conditions.

Upshot: algorithm that either produces all normalized candidate I_0's or proves border rank $> r$.

The border apolarity method

If $\underline{\mathbf{R}}(T) \leq r$, there exists a multi-graded ideal I satisfying:

1. I is contained in the annihilator of T. This condition says $I_{110} \subset T(C^*)^\perp$, $I_{101} \subset T(B^*)^\perp$, $I_{011} \subset T(A^*)^\perp$ and $I_{111} \subset T^\perp \subset A^* \otimes B^* \otimes C^*$.
2. For all (ijk) with $i+j+k>1$, $\operatorname{codim} I_{ijk} = r$.
3. each I_{ijk} is Borel-fixed.
4. I is an ideal, so the multiplication maps $I_{i-1,j,k} \otimes A^* \oplus I_{i,j-1,k} \otimes B^* \oplus I_{i,j,k-1} \otimes C^* \to S^i A^* \otimes S^j B^* \otimes S^k C^*$ have image contained in I_{ijk}.

Matrix multiplication and border apolarity

Here $A = U^* \otimes V$, $B = V^* \otimes W$, $C = W^* \otimes U$,

$M_{\langle n \rangle}$ reordering of $\mathrm{Id}_U \otimes \mathrm{Id}_V \otimes Id_W$, $\mathrm{Id}_U \in U^* \otimes U$.

$M_{\langle n \rangle}(C^*) = U^* \otimes \mathrm{Id}_V \otimes W$

$$\begin{aligned} \subset A \otimes B &= (U^* \otimes V) \otimes (V^* \otimes W) \\ &= M_{\langle n \rangle}(C^*) \oplus [U^* \otimes \mathfrak{sl}(V) \otimes W] \end{aligned}$$

Need to understand Borel fixed subspaces in $U^* \otimes \mathfrak{sl}(V) \otimes W$.
Borel: upper triangular invertible matrices in
$SL(U) \times SL(V) \times SL(W) = SL_n \times SL_n \times SL_n$.

Borel fixed subspaces for $U^*{\otimes}\mathfrak{sl}(V){\otimes}W$

Candidate I_{110} codim$=r$ Equivalently, $I_{110}^{\perp}$, dim$=r$ containing $T(C^*) = U^*{\otimes}\,\mathrm{Id}_V\,{\otimes}W$ need to add $r-n^2$ dimensional Borel fixed subspace Case $M_{\langle 2\rangle}$: $r=6$, $n^2=4$, $r-n^2=2$

$$\begin{tikzcd}
 & x_1^2\otimes y_1^2 \arrow[d] \arrow[ddl] \arrow[ddr] & \\
 & x_1^2\otimes y_1^1 - x_2^2\otimes y_1^2 \arrow[dd] \arrow[dddl] \arrow[dddr] & \\
x_1^2\otimes y_2^2 \arrow[dd] \arrow[dddr] & & x_1^1\otimes y_1^2 \arrow[dd] \arrow[dddl] \\
 & x_2^2\otimes y_1^1 \arrow[dddl] \arrow[dddr] & \\
x_1^2\otimes y_2^1 - x_2^2\otimes y_2^2 \arrow[dd] \arrow[dddr] & & x_1^1\otimes y_1^1 - x_2^1\otimes y_1^2 \arrow[dd] \arrow[dddl] \\
 & x_1^1\otimes y_2^2 \arrow[dd] & \\
x_2^2\otimes y_2^1 \arrow[ddr] & & x_2^1\otimes y_1^1 \arrow[ddl] \\
 & x_1^1\otimes y_2^1 - x_2^1\otimes y_2^2 \arrow[d] & \\
 & x_2^1\otimes y_2^1 &
\end{tikzcd}$$

$x_j^i = u^i{\otimes}v_j$ etc..

Border apolarity: results

Conner-Harper-L May 2019:
$\rightsquigarrow$ very easy algebraic proof $\underline{\mathbf{R}}(M_{\langle 2\rangle}) = 7$

Recall: Strassen $\underline{\mathbf{R}}(M_{\langle 3\rangle}) \geq 14$, L-Ottaviani $\underline{\mathbf{R}}(M_{\langle 3\rangle}) \geq 15$,
L-Michalek $\underline{\mathbf{R}}(M_{\langle 3\rangle}) \geq 16$.

Conner-Harper-L June 2019: $\underline{\mathbf{R}}(M_{\langle 3\rangle}) \geq 17$

Recall: so far only $\underline{\mathbf{R}}(M_{\langle 2\rangle})$ known among nontrivial matrix multiplication tensors.

Conner-Harper-L:
August 2019: $\underline{\mathbf{R}}(M_{\langle 223\rangle}) = 10$
August 2019: $\underline{\mathbf{R}}(M_{\langle 233\rangle}) = 14$

September 2019: For all $\mathbf{n} > 25$, $\underline{\mathbf{R}}(M_{\langle 2\mathbf{nn}\rangle}) \geq \mathbf{n}^2 + 1.32\mathbf{n} + 1$.
Previously, only $\underline{\mathbf{R}}(M_{\langle 2\mathbf{nn}\rangle}) \geq \mathbf{n}^2 + 1$ known.
November 2019: For all $\mathbf{n} > 14$, $\underline{\mathbf{R}}(M_{\langle 3\mathbf{nn}\rangle}) \geq \mathbf{n}^2 + 2\mathbf{n}$.
Previously, only $\underline{\mathbf{R}}(M_{\langle 3\mathbf{nn}\rangle}) \geq \mathbf{n}^2 + 2$ known.

Other results

Strassen laser method: bound ω indirectly via other tensors.

Prop. (Conner-Gesmundo-L-Ventura) det_3, perm_3 potentially could be used to prove $\omega = 2$.

$\underline{\mathbf{R}}(\mathrm{det}_3) = 17$ (Conner-Harper-L). Nov. 2019

$\underline{\mathbf{R}}(\mathrm{perm}_3) = 16$ (Conner-Huang-L). May 2020

Idea of proof for asymptotic results

How to prove lower bounds for all n?

Candidate $I_{110}^{\perp}$:
$U^* \otimes \mathrm{Id}_V \otimes W \subset I_{110}^{\perp} \subset B \otimes C = U^* \otimes \mathfrak{sl}(V) \otimes W \oplus U^* \otimes \mathrm{Id}_V \otimes W$

To prove $\underline{\mathbf{R}}(M_{\langle mnn \rangle}) \geq n^2 + \rho$, we show:

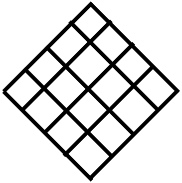
$\forall\ E \in G(\rho, U^* \otimes \mathfrak{sl}(V) \otimes W)^{\mathbb{B}}$, (210) or (120) test fails.

Idea of proof for asymptotic results

Set of $U^* \otimes W$ weights of $I_{110}^{\perp}$ "outer structure"

Given $U^* \otimes W$ weight (s, t), set of $\mathfrak{sl}(V)$-weights appearing with it "inner structure"

$\rightsquigarrow$ $n \times n$ grid, attach to each vertex a $\mathbb{B}$-closed subspace of $\mathfrak{sl}(V)$.
Split calculation of the kernel into a local and global computation.
Bound local (grid point) contribution to kernel by function of s, t and dimension of subspace of $\mathfrak{sl}(V)$.



Solve a nearly convex optimization problem over all possible outer structures.
Show extremal values fail test $\rightsquigarrow$ all choices fail test.

What about the barrier?

Bad news: (ijk)-tests are determinantal equations— subject to barrier, i.e., candidate ideals may be candidate cactus border rank decompositions.

How to tell if zero dimensional scheme is smoothable?

In general, hopeless. But: algorithm produces Borel fixed ideals $\rightsquigarrow$ schemes supported at a point.

Here there are recent techniques using deformation theory (Jelisejew).

Extremely recent: an elementary technique that distinguishes border rank ideals from impostors – the flag conditions.

Thank you for your attention

For more on **tensors**, their geometry and applications, resp. **geometry and complexity**, resp. **recent developments**: