Efficient matrix multiplication

J.M. Landsberg

Texas A&M University

ILAS July 2019

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.

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!*

Tensor formulation of conjecture: review of linear maps

Review: a linear map $L : \mathbb{C}^M \to \mathbb{C}^N$

- In bases is represented by an $M \times N$ matrix X_L. The map takes a column vector v to the column vector $X_L v$
- Equivalent to a linear map $L^t : \mathbb{C}^{N*} \to \mathbb{C}^{M*}$, where in bases, a row vector β maps to the row vector βX_L
- Equivalent to a bilinear map $\mathbb{C}^M \times \mathbb{C}^{N^*} \to \mathbb{C}$, where in bases, (v, β) maps to the number $\beta X_L v$

A linear map L has *rank one* if it may be written as $L = \alpha \otimes w$, with $\alpha \in \mathbb{C}^{M*}$, $w \in \mathbb{C}^N$. Then $v \mapsto \alpha(v)w$.

A linear map L has rank at most r if it may be written as the sum of r rank one linear maps.

The set of all linear maps $\mathbb{C}^M \to \mathbb{C}^N$ is a vector space of dimension MN and is denoted $\mathbb{C}^{M*} \otimes \mathbb{C}^N$.

Tensor formulation of conjecture

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

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

i.e., an element of

$$\mathbb{C}^{N*} \otimes \mathbb{C}^{N*} \otimes \mathbb{C}^N$$

Bilinear maps $\mathbb{C}^N \times \mathbb{C}^N \to \mathbb{C}^N$ may also be viewed as trilinear maps $\mathbb{C}^N \times \mathbb{C}^N \times \mathbb{C}^{N*} \to \mathbb{C}$.
Exercise: As such, $M_{\langle n \rangle}(X, Y, Z) = \mathrm{trace}(XYZ)$.

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.

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.

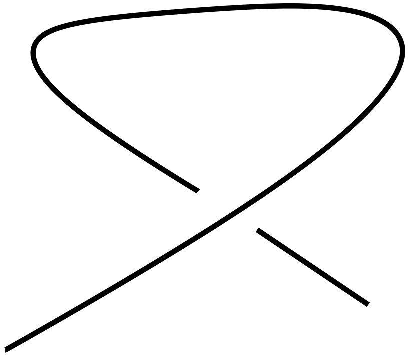
Classical: $\omega \leq 3$.

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

Astounding Conjecture

$\omega = 2$

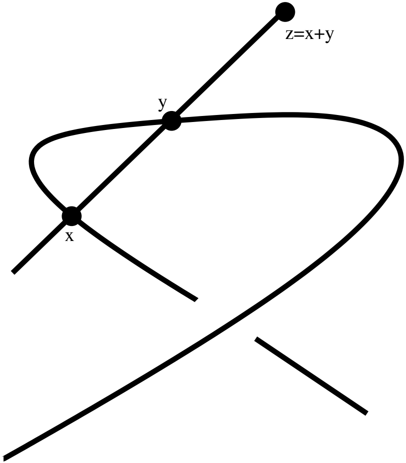
Geometric formulation of conjecture



Imagine this curve represents the set of tensors of rank one.

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 an r-plane to set of tensors of rank one.

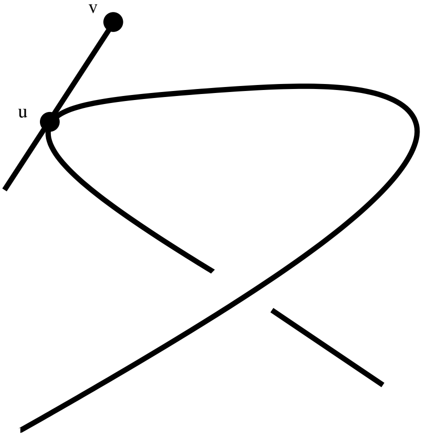
Bini's sleepless nights

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 $T \in \mathbb{C}^N \otimes \mathbb{C}^N \otimes \mathbb{C}^N$, let $\underline{\mathbf{R}}(T)$, the *border rank of T* denote 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 $N = 4$.

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

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

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

Tensors of rank at most 3 $\Leftrightarrow$ 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 $T(C^*) \subset A \otimes B$.
I.e., study of tensors up to changes of bases is equivalent to study of 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: $\underline{\mathbf{R}}(T) \geq N + \frac{1}{2}(\text{rank of commutator})$

$\rightsquigarrow$ (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 for general $\mathbf{n}$.

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 to a space of matrices (systematic search possible).

Punch line: Found equations by exploiting symmetry of σ_r

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).

How to go further?

So far, lower bounds via symmetry of σ_r.

The matrix multiplication tensor also has symmetry:

$T \in \mathbb{C}^N \otimes \mathbb{C}^N \otimes \mathbb{C}^N$, define *symmetry group of T*
$G_T := \{g \in GL_N^{\times 3} \mid g \cdot T = T\}$

$GL_{\mathbf{n}}^{\times 3} \subset G_{M_{\langle n \rangle}} \subset GL_{\mathbf{n}^2}^{\times 3}$:

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

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

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.
$\rightsquigarrow$ can insist on normalized curves (for $M_{\langle n \rangle}$, those with E_0 Borel stable).

$\rightsquigarrow$ (L-Michalek 2017) $\underline{\mathbf{R}}(M_{\langle n \rangle}) \geq 2\mathbf{n}^2 - \log_2 \mathbf{n} - 1$

More bad news: this method cannot go much further.

Upper bounds: How to *prove* conjecture?

History: $M_{\langle n \rangle}$ too complicated, look for something simpler.
Schönhage:

$$\underline{\mathbf{R}}(M_{\langle 1,(\mathbf{n}-1)^2,(\mathbf{n}-1)^2 \rangle} \oplus M_{\langle \mathbf{n},\mathbf{n},1 \rangle}) = \mathbf{n}^2 + 1$$
$$<< \underline{\mathbf{R}}(M_{\langle 1,(\mathbf{n}-1)^2,(\mathbf{n}-1)^2 \rangle}) + \underline{\mathbf{R}}(M_{\langle \mathbf{n},\mathbf{n},1 \rangle})$$

Can upper bound ω by upper bounding border ranks of such sums.
Showed $\omega < 2.55$

Upper bounds

$T \in A \otimes B \otimes C$, define Kronecker power
$T^{\boxtimes k} \in (A^{\otimes k}) \otimes (B^{\otimes k}) \otimes (C^{\otimes k})$.

Exercise: $M_{\langle n \rangle}^{\boxtimes k} = M_{\langle \mathbf{n}^k \rangle}$ (self-reproducing property!)

Strassen: Look for "simple" T such that $T^{\boxtimes k}$ degenerates to direct sum of many disjoint rectangular matrix multiplications.
$\rightsquigarrow \omega < 2.38$ (Strassen, Coppersmith-Winograd 1989) via "big Coppersmith-Winograd tensor"

1989-2011 no improvment
2011-2013 $\omega < 2.373$ (Stouthers-Williams-LeGall)
2014 Bad news: Ambainus-Filimus-LeGall: game over for big Coppersmith-Winograd tensor — need new tensors!

2019: Conner-Gesmundo-L-Ventura: New tensors, potentially better for Strassen's method — stay tuned.

New idea, following Buczynska-Buczynski: Upper and Lower bounds

Use more algebraic geometry: Consider not just curve of r points, but the curve of ideals it gives rise to: *border apolarity method*

Can insist that limiting ideal is Borel fixed: reduces to small search.

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: $\underline{\mathbf{R}}(M_{\langle 3\rangle}) \geq 17$ (so far - work in progress).

Bonus 1: Method in principle can overcome lower bound barriers.

Bonus 2: Method also can be used for *upper* bounds.

Thank you for your attention

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