What Is Fast Matrix Multiplication?

Nicholas J. Higham*

September 13, 2022

The definition of matrix multiplication says that for $n \times n$ matrices A and B, the product $C = AB$ is given by $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$. Each element of C is an inner product of a row of A and a column of B, so if this formula is used then the cost of forming C is $n^2(2n-1)$ additions and multiplications, that is, $O(n^3)$ operations. For over a century after the development of matrix algebra in the 1850s by Cayley, Sylvester and others, all methods for matrix multiplication were based on this formula and required $O(n^3)$ operations.

In 1969 Volker Strassen showed that when $n = 2$ the product can be computed from the formulas

$$\begin{aligned} p_1 &= (a_{11} + a_{22})(b_{11} + b_{22}), \\ p_2 &= (a_{21} + a_{22})b_{11}, \qquad p_3 = a_{11}(b_{12} - b_{22}), \\ p_4 &= a_{22}(b_{21} - b_{11}), \qquad p_5 = (a_{11} + a_{12})b_{22}, \\ p_6 &= (a_{21} - a_{11})(b_{11} + b_{12}), \qquad p_7 = (a_{12} - a_{22})(b_{21} + b_{22}), \\ C &= \begin{bmatrix} p_1 + p_4 - p_5 + p_7 & p_3 + p_5 \\ p_2 + p_4 & p_1 + p_3 - p_2 + p_6 \end{bmatrix}. \end{aligned}$$

The evaluation requires 7 multiplications and 18 additions instead of 8 multiplications and 4 additions for the usual formulas.

At first sight, Strassen's formulas may appear simply to be a curiosity. However, the formulas do not rely on commutativity so are valid when the a_{ij} and b_{ij} are matrices, in which case for large dimensions the saving of one multiplication greatly outweighs the extra 14 additions. Assuming n is a power of 2, we can partition A and B into four blocks of size $n/2$, apply Strassen's formulas for the multiplication, and then apply the same formulas recursively on the half-sized matrix products.

Let us examine the number of multiplications for the recursive Strassen algorithm. Denote by $M(k)$ the number of scalar multiplications required to multiply two $2^k \times 2^k$ matrices. We have $M(k) = 7M(k-1)$, so

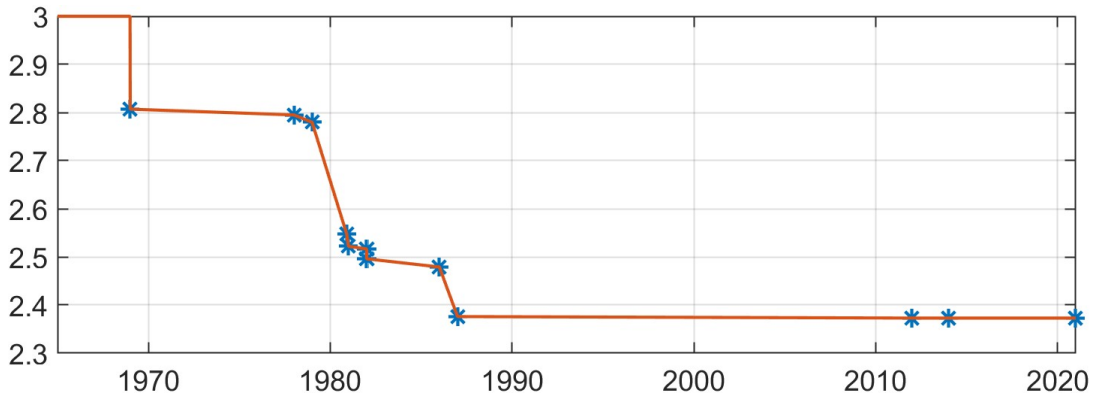
$$M(k) = 7M(k-1) = 7^2 M(k-2) = \cdots = 7^k M(0) = 7^k.$$

But $7^k = 2^{\log_2 7^k} = (2^k)^{\log_2 7} = n^{\log_2 7} = n^{2.807\ldots}$. The number of additions can be shown to be of the same order of magnitude, so the algorithm requires $O(n^{\log_2 7}) = O(n^{2.807\ldots})$ operations.

*Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (nick.higham@manchester.ac.uk).

Strassen's work sparked interest in finding matrix multiplication algorithms of even lower complexity. Since there are $O(n^2)$ elements of data, which must each participate in at least one operation, the exponent of n in the operation count must be at least 2.

The current record upper bound on the exponent is 2.37286, proved by Alman and Vassilevska Williams (2021) which improved on the previous record of 2.37287, proved by Le Gall (2014) The following figure plots the best upper bound for the exponent for matrix multiplication over time.



In the methods that achieve exponents lower than 2.775, various intricate techniques are used, based on representing matrix multiplication in terms of bilinear or trilinear forms and their representation as tensors having low rank. Laderman, Pan, and Sha (1993) explain that for these methods "very large overhead constants are hidden in the 'O' notation", and that the methods "improve on Strassen's (and even the classical) algorithm only for immense numbers N."

Strassen's method, when carefully implemented, can be faster than conventional matrix multiplication for reasonable dimensions. In practice, one does not recur down to 1×1 matrices, but rather uses conventional multiplication once $n_0 \times n_0$ matrices are reached, where the parameter n_0 is tuned for the best performance.

Strassen's method has the drawback that it satisfies a weaker form of rounding error bound that conventional multiplication. For conventional multiplication $C = AB$ of $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ we have the componentwise bound

$$|C - \widehat{C}| \le \gamma_n |A|@|B|, \qquad (1)$$

where $\gamma_n = nu/(1-nu)$ and u is the unit roundoff. For Strassen's method we have only a normwise error bound. The following result uses the norm $\|A\| = \max_{i,j} |a_{ij}|$, which is not a consistent norm.

Theorem 1 (Brent). *Let $A, B \in \mathbb{R}^{n\times n}$, where $n = 2^k$. Suppose that $C = AB$ is computed by Strassen's method and that $n_0 = 2^r$ is the threshold at which conventional multiplication is used. The computed product $\widehat{C}$ satisfies*

$$\|C - \widehat{C}\| \le \left[\left(\frac{n}{n_0}\right)^{\log_2 12} (n_0^2 + 5n_0) - 5n \right] u \|A\| \, \|B\| + O(u^2). \qquad (2)$$

With full recursion ($n_0 = 1$) the constant in (2) is $6n^{\log_2 12} - 5n \approx 6n^{3.585} - 5n$, whereas with just one level of recursion ($n_0 = n/2$) it is $3n^2 + 25n$. These compare with $n^2 u + O(u^2)$

for conventional multiplication (obtained by taking norms in (1)). So the constant for Strassen's method grows at a faster rate than that for conventional multiplication no matter what the choice of n_0.

The fact that Strassen's method does not satisfy a componentwise error bound is a significant weakness of the method. Indeed Strassen's method cannot even accurately multiply by the identity matrix. The product

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix}, \quad 0 < \epsilon \ll 1$$

is evaluated exactly in floating-point arithmetic by conventional multiplication, but Strassen's method computes

$$c_{22} = 2(1+\epsilon^2) + (\epsilon - \epsilon^2) - 1 - (1+\epsilon).$$

Because c_{22} involves subterms of order unity, the error $c_{22} - \widehat{c}_{22}$ will be of order u. Thus the relative error $|c_{22} - \widehat{c}_{22}|/|c_{22}| = O(u/\epsilon^2) \gg O(u)$,

Another weakness of Strassen's method is that while the scaling $AB \rightarrow (AD)(D^{-1}B)$, where D is diagonal, leaves (1) unchanged, it can alter (2) by an arbitrary amount. Dumitrescu (1998) suggests computing $D_1(D_1^{-1}A \cdot BD_2^{-1})D_2$, where the diagonal matrices D_1 and D_2 are chosen to equilibrate the rows of A and the columns of B in the ∞-norm; he shows that this scaling can improve the accuracy of the result. Further investigations along these lines are made by Ballard et al. (2016).

Should one use Strassen's method in practice, assuming that an implementation is available that is faster than conventional multiplication? Not if one needs a componentwise error bound, which ensures accurate products of matrices with nonnegative entries and ensures that the column scaling of A and row scaling of B has no effect on the error. But if a normwise error bound with a faster growing constant than for conventional multiplication is acceptable then the method is worth considering.

Notes

For recent work on high-performance implementation of Strassen's method see Huang et al. (2016, 2020).

Theorem 1 is from an unpublished technical report of Brent (1970). A proof can be found in Higham (2002, §23.2.2).

For more on fast matrix multiplication see Bini (2014) and Higham (2002, Chapter 23).

References

This is a minimal set of references, which contain further useful references within.

- Josh Alman and Virginia Vassilevska Williams. A refined laser method and faster matrix multiplication. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Society for Industrial and Applied Mathematics, January 2021, pages 522–539.
- Grey Ballard, Austin R. Benson, Alex Druinsky, Benjamin Lipshitz, and Oded Schwartz. Improving the numerical stability of fast matrix multiplication. *SIAM J. Matrix Anal. Appl*, 37(4):1382–1418, 2016.

- Benson, Alex Druinsky, Benjamin Lipshitz, and Oded Schwartz. Improving the numerical stability of fast matrix multiplication. *SIAM J. Matrix Anal. Appl.*, 37(4):1382–1418, 2016.
- Dario A. Bini. Fast matrix multiplication. In *Handbook of Linear Algebra*, Leslie Hogben, editor, second edition, Chapman and Hall/CRC, Boca Raton, FL, USA, 2014, pages 61.1–61.17.
- Bogdan Dumitrescu. Improving and estimating the accuracy of Strassen's algorithm. *Numer. Math.*, 79:485–499, 1998.
- Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.
- Jianyu Huang, Tyler M. Smith, Greg M. Henry, and Robert A. van de Geijn. Strassen's algorithm reloaded. In *SC16: International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE, November 2016.
- Jianyu Huang, Chenhan D. Yu, and Robert A. van de Geijn. Strassen's algorithm reloaded on GPUs, *ACM Trans. Math. Software*, 46(1):1:1–1:22, 2020.
- Julian Laderman, Victor Pan, and Xuan-He Sha. On practical algorithms for accelerated matrix multiplication. Linear Algebra Appl., 162–164:557–588, 1992.
- François Le Gall. Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, 2014, pages 296–303.

This article is part of the "What Is" series, available from `https://nhigham.com/category/what-is` and in PDF form from the GitHub repository `https://github.com/higham/what-is`.