Pokroky matematiky, fyziky a astronomie

Jan Brandts; Michal Křížek

Geometrická interpretace Strassenova násobení matic

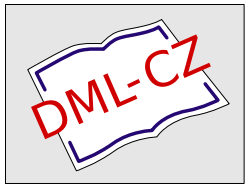
Pokroky matematiky, fyziky a astronomie, Vol. 70 (2025), No. 3, 141–148

Persistent URL: http://dml.cz/dmlcz/153540

Terms of use:

© Jednota českých matematiků a fyziků, 2025

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://dml.cz

Geometrická interpretace Strassenova násobení matic

Jan Brandts, Michal Křížek

Abstrakt. V roce 1969 Strassen ukázal, že součin dvou reálných matic typu 2×2 lze vypočítat pouze pomocí sedmi součinů reálných čísel. Do té doby se všeobecně věřilo, že je zapotřebí osm takových součinů. Výsledný algoritmus je však považován za neintuitivní a chybí také názorné a srozumitelné odvození Strassenovy metody. V tomto článku proto představíme geometrickou interpretaci této metody, podobnou té, která je základem Karatsubova algoritmu pro efektivní násobení celých čísel.

1. Úvod

Násobení dvou matic typu 2×2 se obvykle provádí podle definice

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & g \\ f & h \end{bmatrix} = \begin{bmatrix} ae+bf & ag+bh \\ ce+df & cg+dh \end{bmatrix}, \tag{1}$$

což vyžaduje vypočítat osm součinů dvojic reálných čísel. V roce 1969 však Volker Strassen v [8] ukázal, že maticový součin v (1) lze také získat pomocí pouhých sedmi součinů jako

$$\begin{bmatrix} m_1+m_4-m_5+m_7 & m_3+m_5 \\ m_2+m_4 & m_1-m_2+m_3+m_6 \end{bmatrix},$$

kde každé ze sedmi čísel $m_1, \ldots, m_7$ je součinem dvou reálných čísel:

$$m_1 = (a+d)(e+h), \quad m_2 = (c+d)e, \quad m_3 = a(g-h), \quad m_4 = d(f-e),$$

$$m_5 = (a+b)h, \quad m_6 = (c-a)(e+g), \quad m_7 = (b-d)(f+h). \tag{2}$$

Všimněme si, že Strassenova metoda vyžaduje celkem 18 sčítání reálných čísel, zatímco v (1) jsou potřeba pouze čtyři. Sčítání je ale z výpočetního hlediska mnohem levnější operace než násobení. Je ovšem zcela legitimní zabývat se otázkou, zda úspora pouhého jednoho součinu převáží náklady na tato dodatečná sčítání. Pokud výše uvedenou myšlenku aplikujeme na matici $2n \times 2n$, která je rozdělena do čtyř bloků o velikosti $n \times n$, je zisk zřejmý: sčítání matic totiž vyžaduje provedení $O(n^2)$ sčítání, zatímco náklady na násobení matic jsou (v dnešní době) stále podstatně vyšší. Rekurzivní použití Strassenova algoritmu skutečně vede k metodě, která potřebuje provedení $O(n^{\log_2 7}) \approx O(n^{2{,}81})$ násobení pro součin dvou matic $n \times n$, na rozdíl od $O(n^{\log_2 8}) = O(n^3)$ násobení pro rekurzivní použití vztahu (1).

Assoc. Prof. Jan Brandts, PhD., Korteweg-de Vries Institute, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, Nizozemsko, e-mail: J.H.Brandts@uva.nl, prof. RNDr. Michal Křížek, DrSc., Matematický ústav AV ČR, Žitná 25, 115 67 Praha 1, e-mail: krizek@math.cas.cz

Ačkoliv není těžké ověřit, že Strassenova metoda násobení matic je správně, není snadné si ji zapamatovat ani odvodit pomocí nějakého jednoduchého matematického principu. Z tohoto důvodu v dalším představíme její geometrickou interpretaci, která nám může napomoci k pochopení hlavní myšlenky.

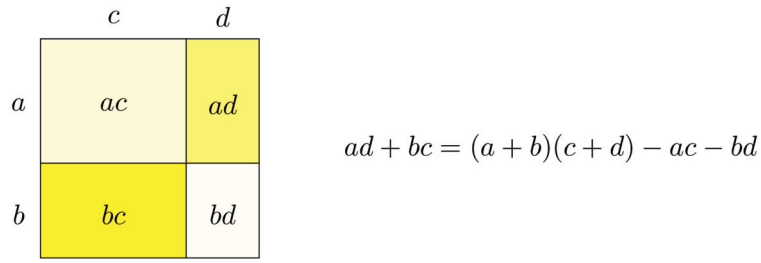
Nedostatečná názornost Strassenova algoritmu je v ostrém kontrastu s metodou, kterou objevil Anatolij A. Karatsuba [7] pro výpočet součinu dvou n-místných čísel, a to nikoli v čase $O(n^{\log_2 4}) = O(n^2)$, ale v $O(n^{\log_2 3}) \approx O(n^{1{,}58})$. Jak naznačují tyto rozdílné řády v exponentech, Karatsuba ukázal, jak lze vynásobit dvě dvouciferná čísla, aniž by použil čtyři, ale pouze tři násobení dvou číslic. Navíc podrobně popsal, jak lze tuto myšlenku rekurzivně použít na n-místná čísla. Jak se později ukázalo, stejný nápad měl již Carl Friedrich Gauss (1777–1855) v souvislosti s násobením komplexních čísel.

2. Znázornění Gaussova–Karatsubova násobení

Gauss si povšiml, že součin dvou komplexních čísel $a + b\mathrm{i}$ a $c + d\mathrm{i}$ je roven

$$(a + b\mathrm{i})(c + d\mathrm{i}) = ac + ad\mathrm{i} + bc\mathrm{i} + bd(\mathrm{i}^2) = ac + [(a + b)(c + d) - ac - bd]\mathrm{i} + bd(\mathrm{i}^2), \quad (3)$$

a proto jeho výpočet vyžaduje jen 3 násobení dvou reálných čísel. Tento výpočet lze geometricky vizualizovat jako problém znázornění příslušné plochy. Přitom je důležité si uvědomit, že nejsou potřeba jednotlivé součiny ad a bc, ale jen jejich součet. Tento součet se rovná ploše celého obdélníku na obrázku 1 zmenšené o již vypočítané plochy jeho levého horního a pravého dolního diagonálního bloku.



Obr. 1. Geometrická interpretace Gaussova–Karatsubova násobení pomocí ploch

Když se tato idea použije pro násobení dvou komplexních matic typu $n \times n$, tak zisk z jednoho uspořeného násobení dvou reálných matic typu $n \times n$ snadno převáží tři další sčítání reálných matic typu $n \times n$.

2.1. Karatsuba versus Kolmogorov

Výše uvedenou ideu Gausse použil Karatsuba [7] k efektivnímu násobení celých čísel. Na semináři z výpočetní složitosti na Moskevské státní univerzitě, který v roce 1960 organizoval Andrej N. Kolmogorov, Karatsuba opravil následující Kolmogorovovo tvrzení, že násobení dvou n-místných celých čísel vyžaduje n^2 součinů dvou číslic. Karatsuba si povšiml, že dvouciferné číslo $d_1 d_2$ se rovná $d_2 + d_1 \cdot 10$, a proto jeho vynásobení

jiným dvouciferným číslem vyžaduje pouze tři násobení dvou číslic. Přitom použil (3), kde imaginární jednotku i zaměnil za základ 10. Námitku, že součet dvou číslic nemusí být nutně jednociferný, lze snadno obejít. Povšimněme si, že pokud má výsledný součet dvě číslice, tak jeho levá číslice je nutně 1, což částečně ušetří prováděné operace. Důležitější ale je, že při rekurzivním použití této myšlenky na vícemístná čísla má součet dvou n-místných čísel nejvýše $n + 1$ číslic.

Karatsubova metoda je konkrétním příkladem násobicích schémat A. L. Tooma [9].

2.2. Konvoluce a rychlá Fourierova transformace

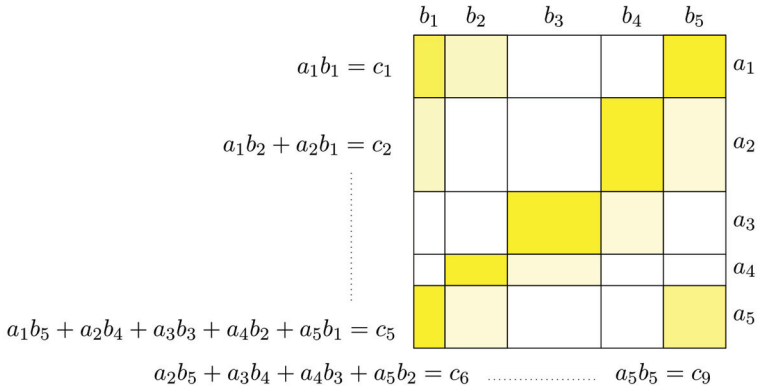
Číslice součinu dvou celých čísel lze získat pomocí konvoluce vektorů $\mathbf{a} = (a_1, \ldots, a_n)$ a $\mathbf{b} = (b_1, \ldots, b_n)$, jejichž prvky jsou číslice. Konvoluce $\mathbf{a} * \mathbf{b} = \mathbf{c} = (c_1, \ldots, c_{2n-1})$ je definována vztahem

$$c_\ell = \sum_{i+j=\ell+1} a_i b_j, \qquad \ell \in \{1, \ldots, 2n-1\}. \tag{4}$$

Poznamenejme, že obrázek 1 vlastně ilustruje, jak vypočítat konvoluci dvou dvousložkových vektorů

$$(a_1, a_2) * (b_1, b_2) = (a_1 b_1, a_1 b_2 + a_2 b_1, a_2 b_2) \tag{5}$$

při použití minimálního počtu tří násobení. Obecně platí, že konvoluce dvou n-složkových vektorů $\mathbf{a}$ a $\mathbf{b}$ se rovná vektoru diagonálních součtů obdélníku, jehož strany mají délky $a_1 + \ldots + a_n$ a $b_1 + \ldots + b_n$, jak je znázorněno na obrázku 2.



Obr. 2. Geometrická interpretace konvoluce jakožto součtu ploch obdélníků na jednotlivých diagonálách

Po provedení konvoluce (4), resp. (5), potřebujeme k nalezení vektoru číslic součinu n-místných čísel pouze sčítání. Tento proces je všeobecně známý jako *přenos číslic*.

Také pro $n > 2$ byly vyvinuty metody výpočtu konvoluce s použitím nízkého nebo dokonce minimálního počtu součinů celých čísel založené na práci Tooma a Winograda [2], [9], [10]. V některých z těchto metod se násobení pevnou konstantou odlišnou od prvků vektorů $\mathbf{a}$ a $\mathbf{b}$ nepočítá do počtu násobení potřebných k výpočtu konvoluce.

Hlavní myšlenku většiny rychlých metod výpočtu konvolucí lze vysvětlit tak, že vektory **a**, **b** považujeme za koeficienty polynomů

$$p(x) = a_1 + a_2 x + \ldots + a_n x^{n-1} \quad \text{a} \quad q(x) = b_1 + b_2 x + \ldots + b_n x^{n-1}.$$

Potom konvoluce $\mathbf{c} = \mathbf{a} * \mathbf{b}$ je vektor koeficientů polynomu $r(x) = p(x)q(x)$. Protože r je jednoznačně určeno svými hodnotami v $m = 2n - 1$ bodech, lze z těchto hodnot získat **c**, které lze zjistit výpočtem součinu hodnot p a q v těchto bodech.

Tedy konkrétně nechť $x_1, \ldots, x_m$ jsou navzájem různé body. Pak **c** lze získat z hodnot r pomocí vztahu

$$Vc = \begin{bmatrix} x_1^0 & \cdots & x_1^{2n-2} \\ \vdots & & \vdots \\ x_m^0 & \cdots & x_m^{2n-2} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_{2n-1} \end{bmatrix} = \begin{bmatrix} r(x_1) \\ \vdots \\ r(x_m) \end{bmatrix}, \tag{6}$$

zatímco naopak hodnoty p a q v bodech $x_1, \ldots, x_m$ lze získat vynásobením vektorů koeficientů **a** a **b** Vandermondovou maticí V v (6) po rozšíření **a** i **b** o $n - 1$ nulových prvků. Přitom násobení maticí V typu $m \times m$ vyžaduje obecně vyčíslení $O(n^2) = O(m^2)$ součinů. Řešení (6) také vyžaduje jen $O(n^2)$ násobení, protože pro danou množinu bodů počítáme V^{-1} pouze jednou, a pak každé násobení V^{-1} spotřebuje jen $O(n^2)$ násobení.

Pokud zvolíme jako interpolační body $2n - 1$ kořenů jedničky $\mathrm{e}^{2k\pi\mathrm{i}/(2n-1)}$ ($k = 0, 1, \ldots, 2n-2$) v komplexní rovině, potom transformace $\mathcal{F}\colon \mathbb{R}^{2n-1} \to \mathbb{R}^{2n-1}$ mezi koeficienty a hodnotami polynomů v (6) se nazývá *diskrétní Fourierova transformace*. Rychlá Fourierova transformace (FFT) [3] je vlastně velice sofistikovaná implementace, která využívá grupovou strukturu kořenů jedničky takovým způsobem, že jak $\mathcal{F}(x)$ tak i $\mathcal{F}^{-1}(y)$ potřebují pouze $O(n \log_2 n)$ součinů dvou komplexních čísel. Tedy pokud jsou $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, pak jejich konvoluci $\mathbf{a} * \mathbf{b}$ lze vypočítat pomocí vztahu

$$\mathbf{a} * \mathbf{b} = \mathcal{F}^{-1}\left(\mathcal{F}(\mathbf{a}) \cdot \mathcal{F}(\mathbf{b})\right). \tag{7}$$

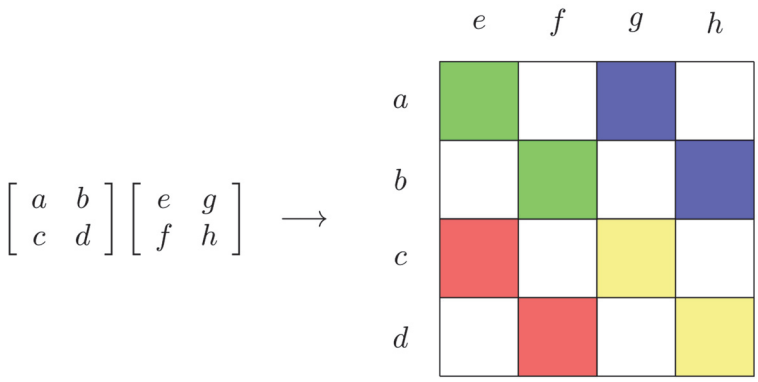
Symbol $\cdot$ v (7) označuje násobení vektorů po jednotlivých prvcích s výpočetní složitostí $O(n)$. Výpočet celé konvoluce pak s ohledem na FFT a její inverzi spotřebuje čas $O(n \log_2 n)$, viz [3].

3. Strassenovo násobení

Matice na levé straně v (1) v zásadě nevyžaduje žádný z jednotlivých součinů dvou reálných čísel, ale pouze čtyři součty příslušných dvojic. Proto lze snížit počet součinů dvou reálných čísel podobným způsobem jako v (3) a na obrázku 1, a to vhodnou geometrickou interpretací součinů reálných čísel jakožto ploch.

3.1. Prvotní zjištění

Násobení dvou matic typu 2×2 ve vztahu (1) z hlediska ploch podobných obrázku 1 je znázorněno na obrázku 3. Každý ze čtyř prvků součinu matic se rovná součtu ploch dvou obdélníků stejné barvy. Na rozdíl od konvolucí znázorněných na obrázku 2:



Obr. 3. Geometrická interpretace součinu dvou matic typu 2×2 je znázorněna různě obarvenými plochami

(i) je třeba uvažovat jen diagonály, které jsou rovnoběžné s hlavní diagonálou;

(ii) stačí využít prvky pouze na třech ze sedmi diagonál;

(iii) nestačí ale získat jen celkové součty těchto tří diagonál.

Bod (i) lze snadno vyřešit obrácením pořadí prvků jednoho ze dvou činitelů součinu. Bod (iii) je však zásadní: rychlé metody výpočtu konvolucí ve skutečnosti nevyžadují výpočet jednotlivých diagonálních prvků nebo částí diagonál.

Prvky součinu Z dvou matic X a Y typu $n \times n$ pro $n > 2$ lze znázornit podobně jako na obrázku 3. Takový prvek z_{ij} se skutečně rovná skalárnímu součinu $x_i^\top y_j$, kde $x_i^\top$ je i-tý řádek matice X a y_j je j-tý sloupec matice Y. Díky vztahu

$$z_{ij} = x_i^\top y_j = \operatorname{trace}(x_i^\top y_j) = \operatorname{trace}(y_j x_i^\top) = \operatorname{trace}(x_i y_j^\top) \tag{8}$$

lze každý prvek z_{ij} matice Z nahradit maticí $x_i y_j^\top \in \mathbb{R}^{n \times n}$ s hodností 1, jak je znázorněno na obrázku 3 pro $n = 2$. Z výsledné matice hodnosti 1 a velikosti $n^2 \times n^2$ je relevantních pouze $2n - 1$ diagonál, ale důležité jsou všechny dílčí součty přes každých n prvků na takové diagonále, a nikoli pouze jejich celkový součet. Ten lze vypočítat použitím FFT pomocí $O(n^2 \log_2 n^2)$ násobení. Není jasné, zda lze FFT nebo jinou metodu výpočtu konvolucí pomocí malého počtu násobení přizpůsobit tak, aby se příslušné dílčí součty daly efektivně vypočítat.

3.2. Geometrická interpretace Strassenovy metody

Otázka, zda lze každou ze čtyř barevných ploch na obrázku 3 získat sčítáním a odečítáním ploch méně než osmi obdélníkových částí $ac, \ldots, dh$ ze vztahu (1), je v (2) zodpovězena pouze částečně, protože některé z těchto částí, například $m_3 = a(g - h)$, jsou samy o sobě rozdíly ploch. Ve skutečnosti je v každém ze součinů m_3, m_4, m_6 a m_7 v (2) jedním z činitelů rozdíl mezi mimodiagonálním prvkem a diagonálním prvkem ve stejném sloupci.

To naznačuje, že mimodiagonální prvky lze získat pomocí diagonálních prvků. Vskutku, aniž bychom počítali jakékoli součiny, můžeme nejprve zapsat jednotlivé činitele součinu ve tvaru

$$\begin{bmatrix} a & r+d \\ a+s & d \end{bmatrix} \begin{bmatrix} e & u+h \\ e+v & h \end{bmatrix}, \tag{9}$$

pro něž podle (1) platí $b = r + d$, $c = a + s$, $f = e + v$ a $g = u + h$. Odtud a podle vztahů (2) dostáváme

$$m_1 = (a+d)(e+h), \quad m_2 = (a+d+s)e, \quad m_3 = au, \quad m_4 = dv,$$

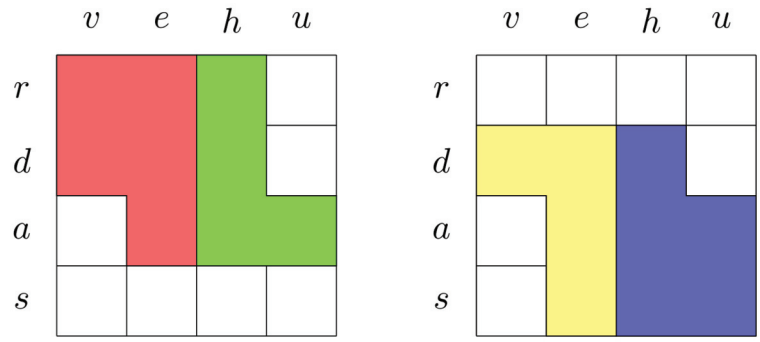
$$m_5 = (a+d+r)h, \quad m_6 = s(e+u+h), \quad m_7 = r(e+v+h). \tag{10}$$

Jádrem součinu (9) je maticové násobení

$$\begin{bmatrix} a & d \\ a & d \end{bmatrix} \begin{bmatrix} e & h \\ e & h \end{bmatrix}, \tag{11}$$

ve kterém se objevují čtyři z celkového počtu osmi maticových prvků. Vyžaduje to výpočet pouze dvou součinů, například $(a + d)e$ a $(a + d)h$. Nyní uvažujme (9) jako perturbaci (11).

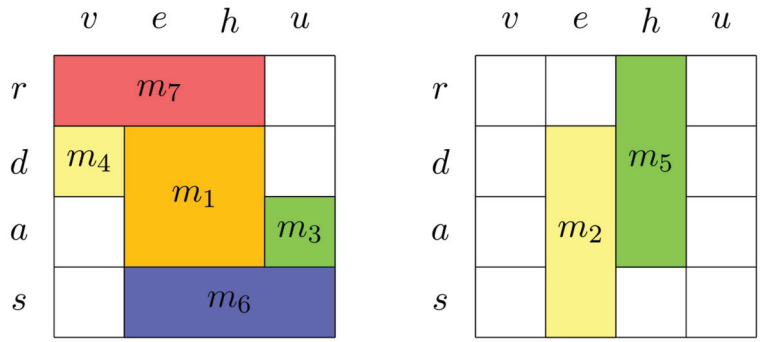
Prvky součinu (9) lze vizualizovat jako na obrázku 4. Protože prvky ve druhém řádku součinů překrývají prvky v prvním řádku, zobrazujeme je zvlášť, tj. levá část obrázku odpovídá prvnímu řádku součinu, pravá část druhému řádku.



Obr. 4. Prvky součinu matic ve vztahu (9) jsou znázorněny jako barevné plochy

Konečně na obrázku 5 znázorňujeme plochy sedmi obdélníků získané přizpůsobením vztahů (2) na matice v součinu (9). Necháváme na čtenáři, aby si vizuálně ověřil, že každou z ploch na obrázku 4 lze získat jako součet a rozdíl obdélníků na obrázku 5. Jinými slovy, tímto způsobem (pomocí sčítání a odčítání ploch) získáme prvky Strassenova součinu uvedeného v (10), který je ekvivalentní s (2).

Všimněme si ještě, že střední část $(a + d)(e + h)$ úzce souvisí s jednoduchým součinem uvedeným v (11).



Obr. 5. Sedm obdélníků, ze kterých lze sestavit jednotlivé části na obrázku 4

4. Závěrečné poznámky

Od doby, co Strassen prokázal, že násobení matic nevyžaduje, jak se všeobecně předpokládalo, $O(n^3)$ aritmetických operací, řada autorů přispěla ke snížení exponentu u n směrem k hodnotě dvě. V roce 1990 Coppersmith a Winograd [4] snížili hodnotu tohoto exponentu na úchvatných 2,3755. Od té doby nedošlo k žádnému výraznému kvantitativnímu pokroku: v roce 2024 zůstává hledaný optimální exponent na hodnotě

$$2{,}371\,339,$$

viz [1]. Mnohé z těchto algoritmů se nazývají *galaktické*, což znamená, že multiplikativní konstanta ukrytá v horním odhadu $O(n^\omega)$ je příliš velká na to, aby vyšetřovaná metoda měla jakýkoli praktický význam.

S rostoucí použitelností umělé inteligence při hledání nových matematických výsledků vývojáři šachového programu AlphaZero a programu AlphaGo pro starodávnou čínskou hru go proměnili otázku hledání způsobů, jak násobit matice různých malých rozměrů pomocí malého počtu součinů, v hru nazvanou AlphaTensor [6]. AlphaTensor byl schopen dosáhnout většiny současně známých dolních odhadů počtu potřebných násobení, ale v některých případech je dokonce překonal.

Na závěr poznamenejme, že základní problémy z lineární algebry, jako je inverze matice a výpočet determinantu, mají stejnou výpočetní složitost jako násobení matic, viz [8]. Je také známo, že Strassenovo násobení není zpětně stabilní algoritmus na počítači s konečnou přesností aritmetiky, viz [5]. To je často považováno za hlavní nevýhodu Strassenovy metody a dalších rychlých metod pro násobení matic.

Poděkování. Druhý autor byl podpořen RVO 67985840 AV ČR a grantem GAČR 24-10586S.

L i t e r a t u r a

[1] Alman, J., Duan, R., Williams, V. V., Xu, Y., Xu, Z., Zhou, R.: *More asymmetry yields faster matrix multiplication.* Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, 2024, 2005–2039.

[2] Cariow, A., Paplinski, J. P.: *Some algorithms for computing short-length linear convolutions*. Electronics 9 (2020), 1–22.

[3] Cooley, J. W., Tukey, J. W.: *An algorithm for the machine calculation of complex Fourier series*. Math. Comp. 19 (1965), 297–301.

[4] Coppersmith, D., Winograd, S.: *Matrix multiplication via arithmetic progressions*. J. Symbolic Comput. 9 (1990), 251–280.

[5] Dumas, J.-G., Pernet, C., Sedoglavic, A.: *Strassen's algorithm is not optimally accurate*. ISSAC '24: Proceedings of the 2024 International Symposium on Symbolic and Algebraic Computation, ACM, 2024, 254–263.

[6] Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., Ruiz, F. J. R., Schrittwieser, J., Swirszcz, G., Silver, D., Hassabis, D., Kohli, P.: *Discovering faster matrix multiplication algorithms with reinforcement learning*. Nature 610 (2022), 47–53.

[7] Karatsuba, A. A., Ofman, Y. P.: *Multiplication of many-digital numbers by automatic computer (in Russian)*. Dokl. Akad. Nauk SSSR 145 (1962), 293–294. Engl. transl. Physics-Doklady 7 (1963), 595–596.

[8] Strassen, V.: *Gaussian elimination is not optimal*. Numer. Math. 13 (1969), 354–356.

[9] Toom, A. L.: *The complexity of a scheme of functional elements simulating the multiplication of integers (in Russian)*. Dokl. Akad. Nauk SSSR 150 (1963), 496–498. Engl. transl. Soviet Math. 3 (1963), 714–716.

[10] Winograd, S.: *Arithmetic complexity of computations*. SIAM, 1980.