Algebraische Komplexitätstheorie II

Schnelle Matrixmultiplikation und Kombinatorik

Peter Bürgisser, Universität Zürich

Dies ist die Ausarbeitung des zweiten Vortrags unserer Reihe, in dem die bilineare Komplexität, ein Teilgebiet der algebraischen Komplexitätstheorie, näher vorgestellt wurde. Sie hat ihren Ursprung in Strassens erstaunlicher Entdeckung aus dem Jahre 1969, wonach Gausselimination kein optimaler Algorithmus zur Lösung zahlreicher Berechnungsprobleme der linearen Algebra ist. Dieses Ergebnis wirkte damals äusserst stimulierend: die Optimalität verschiedenster Berechnungsverfahren, über die man sich bis anhin kaum systematisch Gedanken gemacht hatte, war damit in Frage gestellt. Tatsächlich wurden in einer stürmischen Entwicklung in der ersten Hälfte der 70er Jahre viele neue, überraschend schnelle Algorithmen gefunden. Parallel dazu entdeckte man die ersten unteren Schranken und Optimalitätsbeweise. Das Problem der Matrixmultiplikation erwies sich jedoch als besonders hartnäckig: es vergingen 9 Jahre bis Strassens Algorithmus durch Pan [9] erstmals verbessert wurde. Mittlerweile hat man ein deutlich tieferes Verständnis des Problems gewonnen. Es ist das Ziel dieses Vortrags, einige der Resultate zu präsentieren, sowie die Ideen und Methoden zu skizzieren, die zu den neuesten Fortschritten in diesem Gebiet geführt haben. Wie der Titel erahnen lässt, spielten kombinatorische Methoden bei den jüngsten Fortschritten eine erhebliche Rolle.

1 Der Exponent

Das der Definition folgende Verfahren zur Multiplikation zweier n-reihiger Matrizen über einem Köper k benötigt Grössenordnung n^3 arithmetische Ope-

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

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

Abbildung 1: Strassens Algorithmus für die Multiplikation zweireihiger Matrizen.

rationen. Bekanntlich ist dies keineswegs optimal: Strassen [11] entdeckte im Jahre 1969, dass Grössenordnung $n^{2.81}$ Operationen ausreichen! Der Entwurf seines Algorithmus verwendet Argumente, die bei der Untersuchung der Komplexität der Matrixmultiplikation in abgewandelter Form immer wieder vorkommen. Wir beginnen deshalb mit einer Beschreibung von Strassens Algorithmus.

Im ersten Schritt wird ein Verfahren zur Multiplikation von zweireihigen Matrizen entwickelt, das gegenüber dem gewöhnlichen Vorgehen *eine* Multiplikation auf Kosten von einigen Additionen oder Subtraktionen einspart. Dies geschieht mittels der Formeln in Abb. 1, die hier etwas vom Himmel fallen. (Man kann diese übrigens mit darstellungstheoretischen Argumenten herleiten [5].) Nebenbei sei auch erwähnt, dass man nicht mit weniger als 7 Multiplikationen auskommen kann.

Diese bescheiden anmutende Einsparung einer Multiplikation entfaltet ihre volle Wirkung erst zusammen mit der Idee der Rekursion. Die Multiplikation von 2^N-reihigen Matrizen kann via Blockzerlegung als Multiplikation zweireihiger Matrizen über dem Ring der 2^{N-1}-reihigen Matrizen aufgefasst werden. Das oben beschriebene Verfahren zur zweireihigen Matrixmultiplikation funktioniert jedoch über einem beliebigen, nicht notwendig kommutativen Ring! Also bekommt man ein rekursives Verfahren mit einem Aufwand $T(n)$, $n = 2^N$, das folgender Rekursionsgleichung genügt

$$T(n) \leq 7 \cdot T(n/2) + 18(n/2)^2,$$

was die Abschätzung $T(n) \leq 7\,n^{\log_2 7} \leq 7\,n^{2.81}$ impliziert. Bemerkenswert ist die Tatsache, dass die Anzahl Additionen und Subtraktionen durch die Anzahl Multiplikationen dominiert wird: Wenn Strassens Verfahren zur zweireihigen Matrixmultiplikation 100 statt 18 Additionen verwendete, so erhielte man trotzdem $T(n) = O(n^{2.81})$.

Man definiert nun den sogenannten *Exponenten* ω der Matrixmultiplikation über k als das Infimum aller Exponenten τ, sodass sich n-reihige Matrizen mit Aufwand $O(n^\tau)$ multiplizieren lassen. (Man kann zeigen, dass der Exponent höchstens von der Charakteristik des Grundkörpers k abhängt [10].)

Wir haben gerade gesehen, dass $\omega < 2.81$ ist. Ausserdem sieht man leicht, dass ω mindestens zwei sein muss. Dies ist aber auch die einzige untere Schranke für ω, die man kennt!

Die beste bekannte obere Schranke für den Exponenten ist $\omega < 2.38$. Dieses Ergebnis aus dem Jahre 1987 wurde von Coppersmith und Winograd [7, 8] erzielt. Deren bemerkenswerter Beweis basiert auf einer von Strassen [13] entwickelten Technik, die von ihm *Lasermethode* genannt wurde. Das Kernstück in der Argumentation von Coppersmith und Winograds ist ein nichtkonstruktiver Existenzbeweis für eine kombinatorische Struktur, der mittels der probabilistischen Methode geführt wird. Es ist eines der Ziele dieses Vortrags, Ihnen die Hauptideen einer geglätteten und vereinfachten Version dieses Beweises zu erklären; für Details verweisen wir auf Kapitel 15 unseres Buchs [4]. Wir möchten auch nicht verschweigen, dass Coppersmith und Winograds Algorithmus von rein theoretischem Interesse ist, weil sich der asymptotische Gewinn erst bei astronomischen Matrixformaten bemerkbar machen würde.

Bevor wir weiterfahren, motivieren wir die Definition des Exponenten noch etwas weiter. Eine genaue Bestimmung des Aufwands der Matrixmultiplikation liegt weit ausserhalb der Reichweite der heutigen, bekannten Methoden. Sogar für die Multiplikation dreireihiger Matrizen ist die multiplikative Komplexität nicht genau bekannt! Es ist deshalb naheliegend, sich auf asymptotische Aussagen zu beschränken.

Die Subroutine Matrixmultiplikation wird von fast allen Algorithmen der linearen Algebra benutzt, so z.B. für die Berechnung der Determinante, des charakteristischen Polynoms, für Matrixinversion und für die Lösung linearer Gleichungssysteme. Man kann zeigen, dass schnelle Algorithmen für die Matrixmultiplikation schnelle Algorithmen für alle obigen Probleme liefern. Aber auch die Umkehrung gilt! Zum Beispiel kann man zeigen, dass sich jeder Algorithmus zur Determinantenberechnung in einen nicht viel langsameren zur Multiplikation von Matrizen transformieren lässt [1]. Insgesamt gilt, dass all diese Probleme die gleiche asymptotische Berechnungskomplexität wie die Matrixmultiplikation haben. Deshalb beschreibt der Exponent ω die asymptotische Komplexität vieler Berechnungsprobleme der linearen Algebra.

2 Rang von Tensoren

Wir entwickeln hier einen formalen Rahmen, um die Komplexität der Matrixmultiplikation und allgemeinerer bilinearer Abbildungen zu diskutieren.

Ein *(Koordinaten-) Tensor* t über k ist definiert als ein Element des Tensorprodukts dreier endlichdimensionaler (Koordinaten-) Vektorräume

$$t = (t_{ij\ell})_{i,j,\ell} \in k^{m\times n\times p} \simeq k^m \otimes k^n \otimes k^p,$$

den man sich am besten als eine dreidimensionale $m \times n \times p$-Matrix vorstellt. Eine *Triade* ist ein Tensor der Gestalt

$$u \otimes v \otimes w := (u_i v_j w_\ell)_{i,j,\ell},$$

wobei $u \in k^m$, $v \in k^n$, $w \in k^p$. Man definiert nun den *Rang* $R(t)$ eines Tensors t als die minimiale Anzahl Triaden, sodass sich der Tensor als Summe von diesen schreiben lässt:

$$\begin{aligned} R(t) := \min \{ r \in \mathbb{N} \quad | \quad & \exists u_\rho \in k^m,\, v_\rho \in k^n,\, w_\rho \in k^p : \\ & t = \textstyle\sum_{\rho=1}^r u_\rho \otimes v_\rho \otimes w_\rho \}. \end{aligned}$$

Im Spezialfall $p = 1$, wenn also t eine Matrix ist, stimmt diese Grösse mit dem Matrixrang überein.

Wir stellen nun den Bezug zur Komplexität her. Eine bilineare Abbildung

$$\varphi : k^m \times k^n \to k^p,\ \varphi_\ell(x,y) = \sum_{i,j} t_{ij\ell} x_i y_j$$

korrespondiert in bijektiver Weise mit ihrem Koordinatentensor $t = (t_{ij\ell})$. Unter der *multiplikativen Komplexität* von φ verstehen wir die Ostrowski-Komplexität $L(\varphi_1(x,y),\ldots,\varphi_p(x,y))$, wobei hier die x_i, y_j als Unbestimmte über k interpretiert werden (vgl. den ersten Vortrag).

Es gilt nun das folgende wichtige Ergebnis.

Satz 1 (Strassen [12]) *Die multiplikative Komplexität der bilinearen Abbildung φ unterscheidet sich vom Rang $R(t)$ ihres Koordinatentensors t um höchstens den Faktor zwei.*

Aufgrund dieses Resultats kann man sich – zumindest für asymptotische Untersuchungen – auf den Rang konzentrieren. Die Definition des Rangs ist mathematisch so klar und einfach, dass man versucht ist zu denken, dass mit obigem Satz die Hauptarbeit geleistet ist. Der Schein trügt! Es zeigte sich, dass die Bestimmung des Rangs eines konkreten Tensors – im Unterschied zum Matrixrang – ein sehr schwieriges Problem ist.

Es seien $t \in k^m \otimes k^n \otimes k^p$ und $t' \in k^{m'} \otimes k^{n'} \otimes k^{p'}$ zwei Tensoren. Wir nennen t und t' *isomorph*, $t \simeq t'$, falls es lineare Isomorphismen $\alpha\colon k^m \to k^{m'}$, $\beta\colon k^n \to k^{n'}$, $\gamma\colon k^p \to k^{p'}$ gibt, sodass $t' = (\alpha \otimes \beta \otimes \gamma)(t)$. In naheliegender Weise definieren wir die *direkte Summe* $t \oplus t' \in k^{m+m'} \otimes k^{n+n'} \otimes k^{p+p'}$ von t und t':

$$(t \oplus t')_{ij\ell} := \begin{cases} t_{ij\ell} & \text{falls } i \le m_1,\ j \le n_1,\ \ell \le p_1, \\ t'_{i-m_1,j-n_1,\ell-p_1} & \text{falls } i > m_1,\ j > n_1,\ \ell > p_1, \\ 0 & \text{sonst.} \end{cases}$$

Das *Tensorprodukt* $t \otimes t' \in k^{m \times m'} \otimes k^{n \times n'} \otimes k^{p \times p'}$ von t und t' wird erklärt durch

$$(t \otimes t')_{(i,i')(j,j')(\ell,\ell')} := t_{ij\ell}\, t'_{i'j'\ell'}.$$

Die Rangfunktion verhält sich bezüglich dieser Bildungen angenehm, denn sie ist *subadditiv* und *submultiplikativ*

$$\begin{aligned} R(t \oplus t') &\leq R(t) + R(t'), \\ R(t \otimes t') &\leq R(t)\, R(t'), \end{aligned}$$

sowie invariant unter Isomorphie.

Wir bezeichnen in der Folge den Tensor der Matrixmultiplikation

$$k^{e\times h} \times k^{h\times \ell} \to k^{e\times \ell},\ (A,B) \mapsto AB$$

mit $\langle e,h,\ell\rangle$. Man kann sich leicht überlegen, dass

$$\langle e,h,\ell\rangle = \sum_{i,j,m} u_{ij} \otimes v_{jm} \otimes w_{mi} \in k^{e\times h} \otimes k^{h\times \ell} \otimes k^{\ell\times e},$$

wo u_{ij}, v_{jm}, bzw. w_{mi} die kanonischen Basisvektoren von $k^{e\times h}$, $k^{h\times \ell}$, bzw. $k^{\ell\times e}$ bezeichnen. Die Tatsache, dass sich die Multiplikation hh'-reihiger Matrizen via Blockzerlegung zurückführen lässt auf die Multiplikation h-reihiger Matrizen, deren Einträge h'-reihige Matrizen sind, führt auf die folgende fundamentale Beziehung

$$\langle e,h,\ell\rangle \otimes \langle e',h',\ell'\rangle \simeq \langle ee', \ell\ell', hh'\rangle.$$

Insbesondere ist das Tensorprodukt von Matrixtensoren wieder ein Matrixtensor.

Wir können nun die frühere Überlegung, welche $\omega < 2.81$ lieferte, mit unseren formalen Begriffen so beschreiben: $R(\langle 2,2,2\rangle) \leq 7$ impliziert

$$R(\langle 2^m, 2^m, 2^m\rangle) = R(\langle 2,2,2\rangle^{\otimes m}) \leq 7^m \leq (2^m)^{\log_2 7},$$

was $\omega < \log_2 7$ oder $2^\omega \leq 7$ liefert.

Diese Überlegung lässt sich verallgemeinern zu

$$R(\langle e,h,\ell\rangle) \leq r \Longrightarrow (eh\ell)^{\omega/3} \leq r. \tag{1}$$

Jeder Algorithmus, um Matrizen von speziellem Format zu multiplizieren, führt also auf Algorithmen, um Matrizen von beliebigem Format zu multiplizieren und damit zu einer oberen Schranke für den Exponenten.

3 Grenzrang

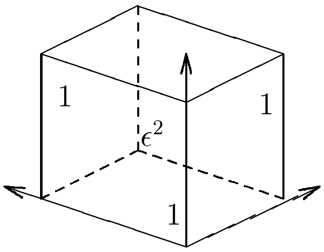
Ein signifikanter Schritt im Unterfangen, obere Schranken für den Exponenten zu gewinnen, ist das Konzept des Grenzrangs, welches durch Bini, Capovani, Lotti und Romani [3] eingeführt wurde. Ausgangspunkt war die Beobachtung, dass es, etwa über $\mathbb{C}$, Tensoren gibt, die sich mit beliebiger Genauigkeit durch Tensoren kleineren Rangs approximieren lassen.

Die algebraische Definition des Grenzrangs lautet so:

Definition 2 *Sei $t \in k^{m\times n\times p}$ und ϵ eine Unbestimmte über k. Der* Grenzrang *$\underline{R}(t)$ von t ist die kleinste natürliche Zahl $r \in \mathbb{N}$, für die ein Tensor $t_1 \in k[\epsilon]^{m\times n\times p}$ existiert so, dass $R(t+\epsilon t_1) \leq r$. (Hierbei wird $t+\epsilon t_1$ als ein Tensor über dem rationalen Funktionenkörper $k(\epsilon)$ interpretiert.)*

Um sich diese Definition zu veranschaulichen, stellt man sich ϵ am besten als ein bezüglich k infinitesimales Element vor. Dann hat ein Tensor einen Grenzrang kleiner als r genau dann, wenn eine infinitesimale Störung dieses Tensors einen Rang kleiner als r hat. Wir bemerken noch, dass man den Grenzrang über algebraisch abgeschlossenen Körpern k auch mit Hilfe der Zariski-Topologie charakterisieren kann (vgl. etwa [4, Chap. 20]). Genauso wie der Rang ist auch der Grenzrang subadditiv, submultiplikativ und invariant unter Isomorphie.

Das folgende Bild zeigt einen $2\times 2\times 2$-Tensor vom Rang drei, dessen Grenzrang zwei ist.



$t = e_{000} + e_{011} + e_{101} \in k^{2\times 2\times 2}$
$R(t) = 3$, aber
$\underline{R}(t) \leq 2$, weil
$R(t + \epsilon^2 e_{110}) \leq 2$

Bini, Capovani, Lotti und Romani [3, 2] zeigten nun mit einer direkten Konstruktion $\underline{R}(\langle 3,2,2\rangle \leq 10$. Ausserdem bewiesen sie, dass die Implikation (1) auch für den Grenzrang gilt, woraus sie die Abschätzung $\omega < 2.78$ folgerten. Die Idee ist dabei, grob gesprochen, aus einem gegebenen "approximativen Algorithmus" für ein spezielles Matrixformat zunächst via Tensorproduktbildung approximative Algorithmen für beliebig grosse Formate zu konstruieren, woraus man dann mittels Interpolation exakte Algorithmen gewinnt. Der für die Interpolation zusätzlich benötigte Aufwand ist asymptotisch vernachlässigbar.

Eines der wichtigsten Werkzeuge, um effiziente Algorithmen für die Matrixmultiplikation zu entwickeln, ist die Entdeckung von Schönhage, dass sich die Implikation (1) auf direkte Summen übertragen lässt.

Satz 3 (Asymptotische Summenungleichung, Schönhage [10])

$$\underline{R}\Big(\bigoplus_{i=1}^{s}\langle e_i, h_i, \ell_i\rangle\Big) \leq r \Longrightarrow \sum_{i=1}^{s}(e_i h_i \ell_i)^{\omega/3} \leq r.$$

Wäre der Grenzrang additiv, so würde sich dies als eine unmittelbare Folgerung aus dem Ergebnis von Bini, Capovani, Lotti und Romani ergeben. Leider ist dem aber nicht so, wie Schönhage in derselben Arbeit nachwies.

Die asymptotische Summenungleichung besagt, etwas ungenau, in Worten:

> *Jeder Algorithmus, um simultan mehrere unabhängige Matrixmultiplikationen von speziellem Format "approximativ" durchzuführen, führt auf Algorithmen, um Matrizen von beliebigem Format zu multiplizieren und damit zu einer oberen Schranke für den Exponenten.*

Schönhage gab folgende Anwendung: er zeigte

$$\underline{R}(\langle 4,1,4\rangle \oplus \langle 1,9,1\rangle) \leq 17,$$

woraus er mit der asymptotischen Summenungleichung zu der Abschätzung $\omega < 2.55$ gelangte.

4 Kombinatorische Degeneration

Bevor wir Strassens [13] Strategie der Lasermethode erklären können, müssen wir noch weitere Begriffe definieren.

Gegeben sei ein Tensor $t \in U \otimes V \otimes W$ bei dem die Vektorräume U, V, W die folgenden endlichen direkten Summenzerlegungen aufweisen:

$$D: \quad U = \bigoplus_{i \in I} U_i, \quad V = \bigoplus_{j \in J} V_j, \quad W = \bigoplus_{\ell \in L} W_\ell$$

(I, J, L sind endliche Indexmengen). Dann bekommt man eine eindeutige Zerlegung

$$t = \sum_{(i,j,\ell) \in I \times J \times L} t(i,j,\ell)$$

in die sogenannten *D-Komponenten* $t(i,j,\ell) \in U_i \otimes V_j \otimes W_\ell$ von t. Die Teilmenge

$$\mathrm{supp}_D t := \{(i,j,\ell) \in I \times J \times L \mid t(i,j,\ell) \neq 0\}$$

des *kombinatorischen Würfels* $I \times J \times L$ nennt man den *D-Träger* von t. Es ist hilfreich, sich unter D eine *Blockzerlegung* des Tensors t vorzustellen: die D-Komponenten beschreiben, was in den einzelnen Blöcken steht, und der D-Träger gibt die grobe Struktur wieder (cf. Abb. 2). Eine Teilmenge Δ eines kombinatorischen Würfels $I \times J \times L$ heisst *Diagonale*, wenn die Projektionen $\Delta \to I$, $\Delta \to J$, $\Delta \to L$ injektiv sind. Die Bedeutung der Diagonalen rührt von folgender Beziehung her:

$$\sum_{(i,j,\ell) \in \Delta} t(i,j,\ell) \simeq \bigoplus_{(i,j,\ell) \in \Delta} t(i,j,\ell).$$

Wir definieren nun den Begriff der kombinatorischen Degeneration für Teilmengen eines kombinatorischen Würfels.

Definition 4 *Sei $\Phi \subseteq I \times J \times L$. Eine Teilmenge Ψ von Φ heisst* (kombinatorische) Degeneration *von Φ, $\Psi \trianglelefteq \Phi$, falls Funktionen $a: I \to \mathbb{Z}$, $b: J \to \mathbb{Z}$, $c: L \to \mathbb{Z}$ existieren mit der Eigenschaft, dass*

$$\begin{aligned} \forall (i,j,\ell) \in \Psi &\quad : \quad a(i) + b(j) + c(\ell) = 0 \\ \forall (i,j,\ell) \in \Phi \setminus \Psi &\quad : \quad a(i) + b(j) + c(\ell) > 0. \end{aligned}$$

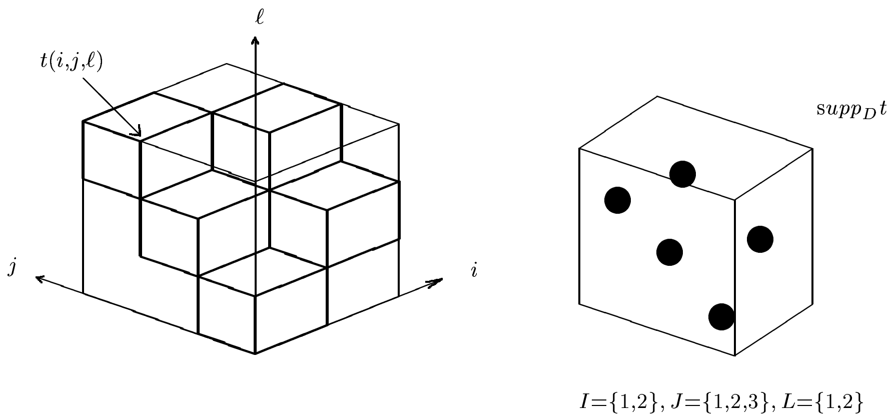


Abbildung 2: Blockzerlegung eines Tensors

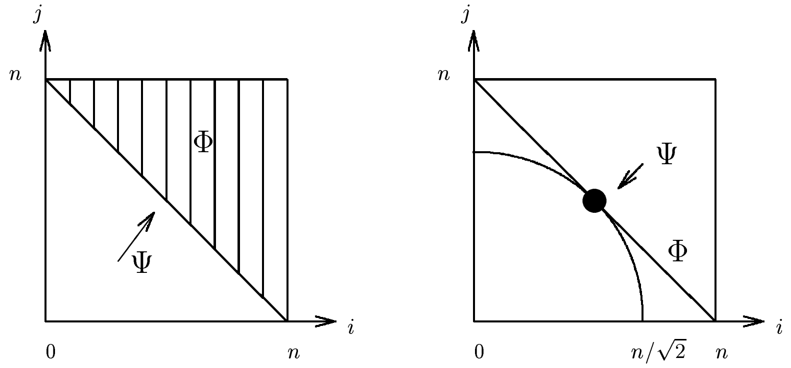
In Abb. 3 wird dieser Begriff an zwei 2-dimensionalen Beispielen illustriert. Im linken Bild wird ein (kombinatorisches) Quadrat durch eine Diagonale in zwei Dreiecke zerlegt. Diese Diagonale ist eine Degeneration beider Dreiecke. Rechts sieht man, dass der Mittelpunkt dieser Diagonale als deren Degeneration erhalten werden kann.

Der Bezug der kombinatorischen Degeneration zum Grenzrang wird durch die folgende Aussage hergestellt.

Proposition 5 *Es sei t ein Tensor mit Blockzerlegung D und $\Psi \trianglelefteq \operatorname{supp}_D t$. Dann gilt $\underline{R}(\sum_{(i,j,\ell)\in\Psi} t(i,j,\ell)) \leq \underline{R}(t)$.*

Zum Beweis erwähnen wir nur, dass $\sum_\Psi t(i,j,\ell)$ aufgrund der Voraussetzung $\Psi \trianglelefteq \operatorname{supp}_D t$ als eine torische Degeneration von t interpretiert werden kann, woraus die Behauptung leicht folgt.

Abbildung 3: Zwei zweidimensionale Beispiele für kombinatorische Degenerationen



$$\Psi \trianglelefteq \Phi$$

$$\underbrace{i}_{a(i)} + \underbrace{(j-n)}_{b(j)} \geq 0 \text{ auf } \Phi$$

Gleichheit auf Ψ

$$\Psi \trianglelefteq \Phi$$

$$\underbrace{i^2}_{a(i)} + \underbrace{(j^2 - n^2/2)}_{b(j)} \geq 0 \text{ auf } \Phi$$

Gleichheit auf Ψ

Wir haben nun die nötigen Hilfsmittel zusammengestellt, um die *Strategie* der sogenannten Lasermethode zu erläutern. Angenommen, wir haben einen Tensor t mit Blockzerlegung D mit folgenden Eigenschaften:

1. alle D-Komponenten sind Matrixtensoren,
2. eine "grosse" Diagonale Δ ist Degeneration von $\mathrm{supp}_D t$,
3. wir kennen eine (gute) obere Schranke r für den Grenzrang von t.

Dann können wir aufgrund von Proposition 5 schliessen, dass

$$\underline{R}\Big(\bigoplus_{(i,j,\ell)\in\Delta} t(i,j,\ell)\Big) \leq \underline{R}(t) \leq r.$$

Jetzt lässt sich die asymptotische Summenungleichung anwenden und wir erhalten eine Abschätzung für den Exponenten.

5 Abschätzung des Exponenten

Das Hauptergebnis dieses Abschnittes ist rein kombinatorischer Natur und liefert eine quantitative Aussage über die Grösse von Diagonalen, die sich aus gewissen Teilmengen von kombinatorischen Würfeln herausdegenerieren lassen. Damit lässt sich die zweite Bedingung der im letzten Abschnitt erläuterten Strategie erfüllen.

Für ein positive natürliche Zahl b setzen wir $I_b := \{-b, -b+1, \ldots, b\} \subseteq \mathbb{Z}$ und $T_b := \{(x, y, z) \in I_b^3 \mid x + y + z = 0\} \subseteq \mathbb{Z}^3$.

Definition 6 *$\Phi \subseteq I \times J \times L$ heisst b-*straff*, falls es ein $r \geq 1$ und injektive Abbildungen $\alpha: I \to \mathbb{Z}^r$, $\beta: J \to \mathbb{Z}^r$, $\gamma: L \to \mathbb{Z}^r$ gibt so, dass*

$$(\alpha, \beta, \gamma)(\Phi) \subseteq (T_b)^r.$$

Eine Teilmenge $\Phi \subseteq I \times J \times L$ ist also b-straff, wenn sie sich unter Erhaltung der Produktstruktur injektiv in ein Produkt von T_b's einbetten lässt.

Der nächste Satz beinhaltet die Kernaussage der Methode, die wir hier beschreiben. Sein Beweis geht auf Coppersmith und Winograd [7, 8] zurück. Die vorliegende geglättete und vereinfachte Variante verdanken wir Strassen [14, 15, 16]. Der Beweis ist ein schönes Beispiel für die Mächtigkeit der probabilistischen Methode in der Kombinatorik.

Satz 7 *Sei $\Phi \subseteq I \times J \times L$ b-straff und etwa $|I| \leq |J| \leq |L|$. Dann existiert eine Diagonale Δ in Φ mit Kardinalität*

$$|\Delta| \geq \frac{1}{K} \min\{|I|, |J|, |L|\},$$

welche Degeneration von Φ ist. Hierbei ist K das 13.5-fache des Maximums von

$$\frac{|I|}{|\Phi|}\max_{i\in I}|p_I^{-1}(i)|,\ \frac{|J|}{|\Phi|}\max_{j\in J}|p_J^{-1}(j)|,\ \frac{|L|}{|\Phi|}\max_{\ell\in L}|p_L^{-1}(\ell)|,\ \frac{4}{9}(2b+1)\frac{|I|}{|\Phi|}$$

und $p_I\colon \Phi\to I$, $p_J\colon \Phi\to J$, $p_L\colon\Phi\to L$ bezeichnet die Projektionen, welche als surjektiv vorausgesetzt werden.

Wir skizzieren hier den Beweis nur in sehr groben Zügen und verweisen für Details auf [4, Chap. 15].

Wir wählen eine Primzahl M, bezeichnen den endlichen Körper $\mathbb{Z}/M\mathbb{Z}$ mit $\mathbb{F}_M$ und setzen

$$\Psi_M := \{(x,y,z)\in\mathbb{F}_M^3 \mid x+y+z=0\}.$$

Ausserdem sei $\varphi\colon \mathbb{F}_M^r\to\mathbb{F}_M$ eine lineare Abbildung. Wir betrachten nun das Diagramm

$$\begin{array}{ccccccc}
I\times J\times L & \overset{(\alpha,\beta,\gamma)}{\hookrightarrow} & \mathbb{Z}^r\times\mathbb{Z}^r\times\mathbb{Z}^r & \to & \mathbb{F}_M^r\times\mathbb{F}_M^r\times\mathbb{F}_M^r & \overset{(\varphi,\varphi,\varphi)}{\to} & \mathbb{F}_M^3\\
\uparrow & & \uparrow & & \uparrow & & \uparrow\\
\Phi & \to & (T_b)^r & \to & (\Psi_M)^r & \to & \Psi_M
\end{array}$$

Die Komposition der unteren drei Abbildungen heisse F_φ. Wenn wir M genügend gross wählen, ist die Komposition von (α,β,γ) mit der Restklassenabbildung $\mathbb{Z}^r\times\mathbb{Z}^r\times\mathbb{Z}^r\to\mathbb{F}_M^r\times\mathbb{F}_M^r\times\mathbb{F}_M^r$ injektiv.

Mit einer direkten Konstruktion kann man leicht eine Diagonale $D\trianglelefteq\Psi_M$ gewinnen mit $|D|\ge M/2$. Durch Zurückziehen sieht man, dass $F_\varphi^{-1}(D)$ Degeneration von Φ ist. Wir schreiben $F_\varphi^{-1}(D)$ als Vereinigung seiner Fasern

$$F_\varphi^{-1}(D) = F_\varphi^{-1}(d_1)\cup F_\varphi^{-1}(d_2)\cup\ldots$$

und degenerieren mit einem weiteren direkten Verfahren jede Faser $F_\varphi^{-1}(d_i)$ zu einer Diagonale Δ_φ^i: $\Delta_\varphi^i\trianglelefteq F_\varphi^{-1}(d_i)$. Weil D diagonal ist, müssen die Fasern $F_\varphi^{-1}(d_i)$ in Würfeln mit paarweise disjunkten Kanten liegen. Daraus erhält man, dass $\Delta_\varphi := \Delta_\varphi^1\cup\Delta_\varphi^2\cup\ldots$ auch diagonal und ausserdem eine Degeneration von Φ ist.

Nun kommt der Clou! Wir wählen die lineare Abbildung $\varphi\colon \mathbb{F}_M^r \to \mathbb{F}_M$ *zufällig*, womit die Kardinalität $|\Delta_\varphi|$ eine Zufallsvariable wird, deren Erwartungswert abgeschätzt werden kann. (Strenggenommen arbeitet man mit einer kleinen Modifikation von α, β, γ, um stochastische Unabhängigkeit zu garantieren.) Man wählt dann die Primzahl M so, dass

$$E(|\Delta_\varphi|) \geq \frac{1}{K} \min\{|I|, |J|, |L|\}.$$

Daraus folgt, dass es mindestens ein φ gibt, für das $|\Delta_\varphi|$ die gewünschte Grösse hat.

Nun können wir diesen Satz mit unserer Strategie kombinieren. Wir wählen einen Tensor t mit einer Blockzerlegung D so, dass alle D-Komponenten Matrixtensoren sind und mit der Eigenschaft, dass der D-Träger straff ist. Dann erhalten wir aus der asymptotischen Summenungleichung eine Abschätzung für den Exponenten. Wir können diese noch verbessern, indem wir das Verfahren auf eine hohe Tensorpotenz $t^{\otimes N}$ mit der zugehörigen Blockzerlegung anwenden. Auf diese Art kann man folgendes zeigen:

Korollar 8 (Coppersmith and Winograd [7]) $\omega < 2.38$.

Obwohl seit 1987 keine besseren Abschätzungen für den Exponenten gefunden wurden, ist hier das letzte Wort sicherlich noch nicht gesprochen.

Die vergebliche Suche nach nichttrivialen unteren Schranken für den Exponenten, sowie die massiven Verbesserungen bei den oberen Schranken lassen die Vermutung zu, dass der Exponent ω *zwei* sein könnte. Diese Vermutung wird gestützt durch das folgende Resultat von Coppersmith [6]

$$R(\langle h, h, \lfloor h^{0.17} \rfloor \rangle) = O(h^2 \log^2 h).$$

Demnach lassen sich eine quadratische h-reihige und eine rechteckige $h \times h^{0.17}$-Matrix mit einem Aufwand multiplizieren, der beinahe linear in der Inputgrösse h^2 ist.

Ich möchte den Vortrag mit der Frage schliessen, ob tatsächlich $\omega = 2$ ist. Dies ist sicher eine der zentralen offenen Probleme der algebraischen Komplexitätstheorie.

Literatur

[1] W. Baur and V. Strassen. The complexity of partial derivatives. *Theoret. Comp. Sc.*, 22:317–330, 1983.

[2] D. Bini. Relation between exact and approximate bilinear algorithms. Applications. *Calcolo*, 17:87–97, 1980.

[3] D. Bini, M. Capovani, G. Lotti, and F. Romani. $O(n^{2.7799})$ complexity for matrix multiplication. *Inf. Proc. Letters*, 8:234–235, 1979.

[4] P. Bürgisser, M. Clausen, and M.A. Shokrollahi. *Algebraic Complexity Theory*, *Grundlehren der mathematischen Wissenschaften*, Bd. 315. Springer Verlag, 1996.

[5] M. Clausen. Beiträge zum Entwurf schneller Spektraltransformationen. Habilitationsschrift, Universität Karlsruhe, 1988.

[6] D. Coppersmith. Rapid multiplication of rectangular matrices. *SIAM J. Comp.*, 11:467–471, 1982.

[7] D. Coppersmith and S. Winograd. Matrix multiplications via arithmetic progression. In *Proc. 19th ACM STOC*, pages 1–6, 1987.

[8] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. Symb. Comp.*, 9:251–280, 1990.

[9] V.Ya. Pan. Field extension and trilinear aggregating, uniting and cancelling for the acceleration of matrix multiplication. In *Proc. of the 20th Ann. IEEE Symp. on Foundations of Comp. Sc.*, pages 28–38, 1979.

[10] A. Schönhage. Partial and total matrix multiplication. *SIAM J. Comp.*, 10:434–455, 1981.

[11] V. Strassen. Gaussian elimination is not optimal. *Num. Math.*, 13:354–356, 1969.

[12] V. Strassen. Vermeidung von Divisionen. *Crelles J. Reine Angew. Math.*, 264:184–202, 1973.

[13] V. Strassen. Relative bilinear complexity and matrix multiplication. *Crelles J. Reine Angew. Math.*, 375/376:406–443, 1987.

[14] V. Strassen. The asymptotic spectrum of tensors. *Crelles J. Reine Angew. Math.*, 384:102–152, 1988.

[15] V. Strassen. Degeneration and complexity of bilinear maps: some asymptotic spectra. *Crelles J. Reine Angew. Math.*, 413:127–180, 1991.

[16] V. Strassen. Private Mitteilung, 1995.