MT404/MS993

Métodos computacionais de álgebra linear

Designs combinatórios e algoritmos de multiplicação matricial

1 Mapas multilineares

Sejam V_i, $i = 1, \ldots, n$, e W espaços vetoriais. Um mapa

$$f : V_1 \times \cdots \times V_n \to W \tag{1}$$

é dito multilinear se ele for linear em todas suas entradas, i.e., se

$$f(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n) = \vec{w} \in W, \tag{2}$$

com $\vec{v}_i \in V_i$, $i = 1, \ldots, n$, for linear em todas as entradas $\vec{v}_i$. Um mapa (ou transformação) linear é um caso particular em que f tem apenas um argumento. Bilinear e trilinear são, obviamente, os casos com dois e três argumentos, respectivamente. Mapas multilineares do tipo (1) são chamados também de n-lineares. Na maior parte das aplicações, todos os espaços vetoriais que aparecem em (1) estão definidos sobre um mesmo corpo F. Nesses casos, chamamos de funcional multilinear um mapa do tipo

$$f : V_1 \times \cdots \times V_n \to F, \tag{3}$$

linear em todas suas entradas.

Como exemplo ilustrativo de (1), vamos considerar o caso mais simples de um mapa (transformação) linear do tipo

$$f : \mathbb{R}^n \to \mathbb{R}^m. \tag{4}$$

Vamos supor agora que $\{\hat{e}_i\}$, $i = 1, \ldots, n$, e $\{\hat{d}_j\}$, $j = 1, \ldots, m$, sejam, respectivamente, bases de $\mathbb{R}^n$ e $\mathbb{R}^m$. Vetores $\vec{v} \in \mathbb{R}^n$ e $\vec{w} \in \mathbb{R}^m$ arbitrários podem ser escritos como

$$\vec{v} = \sum_{i=1}^{n} v_i \hat{e}_i \quad \text{e} \quad \vec{w} = \sum_{j=1}^{m} w_j \hat{d}_j. \tag{5}$$

Os coeficientes v_i e w_j são as chamadas componentes dos vetores $\vec{v}$ e $\vec{w}$ nas respectivas bases $\{\hat{e}_i\}$ e $\{\hat{d}_j\}$. Da linearidade de f, temos que a ação de f em $\vec{v}$ será

$$\vec{w} = f(\vec{v}) = \sum_{i=1}^{n} v_i f(\hat{e}_i). \tag{6}$$

Notem que, por construção, $f(\hat{e}_i) \in \mathbb{R}^m$ para todo $i = 1, 2, \ldots, n$ e, portanto, os n vetores $f(\hat{e}_i)$ podem ser expressos em termos da base $\{\hat{d}_j\}$

$$f(\hat{e}_i) = \sum_{j=1}^{m} \mathfrak{f}_{ji} \hat{d}_j. \tag{7}$$

Os coeficientes $\mathfrak{f}_{ji}$ dependem apenas das bases $\{\hat{e}_i\}$ e $\{\hat{d}_j\}$. Conhecendo estes coeficientes, podemos escrever a ação geral de f sobre um $\vec{v}$ arbitrário como

$$\vec{w} = \sum_{j=1}^{m} w_j \hat{d}_j = f(\vec{v}) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \mathfrak{f}_{ji} v_i \right) \hat{d}_j. \tag{8}$$

Como os vetores da base $\{\hat{d}_j\}$ são L.I., temos

$$w_j = \sum_{i=1}^{n} \mathfrak{f}_{ji} v_i, \tag{9}$$

que corresponde a representação matricial usual do mapa (ou transformação) linear $f : \mathbb{R}^n \to \mathbb{R}^m$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = \begin{pmatrix} \mathfrak{f}_{11} & \mathfrak{f}_{12} & \cdots & \mathfrak{f}_{1n} \\ \mathfrak{f}_{21} & \mathfrak{f}_{22} & \cdots & \mathfrak{f}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathfrak{f}_{m1} & \mathfrak{f}_{m2} & \cdots & \mathfrak{f}_{mn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \tag{10}$$

nas bases $\{\hat{e}_i\}$ e $\{\hat{d}_j\}$. Os coeficientes $\mathfrak{f}_{ji}$ são chamadas componentes tensoriais do mapa linear f. Um tensor, neste contexto, é o conjunto completo de componentes tensoriais de um mapa multilinear.

Consideremos um segundo exemplo, o mapa bilinear

$$g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n. \tag{11}$$

Não há perda de generalidade em escolher os três espaços com a mesma dimensão n. Vamos admitir que os três são gerados pela mesma base $\{\hat{e}_i\}$, $i = 1, \ldots, n$. Como no exemplo anterior, vamos considerar a ação de g em dois vetores genéricos $\vec{v}, \vec{w} \in \mathbb{R}^n$, dando origem a um vetor $\vec{z} \in \mathbb{R}^n$,

$$\vec{z} = g(\vec{v}, \vec{w}) = g\left(\sum_{i=1}^{n} v_i \hat{e}_i, \sum_{i=j}^{n} w_j \hat{e}_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i w_j g(\hat{e}_i, \hat{e}_j). \tag{12}$$

Como no caso do mapa linear, $g(\hat{e}_i, \hat{e}_j)$ pode também ser completamente descrito em termos da base $\{\hat{e}_i\}$

$$g(\hat{e}_i, \hat{e}_j) = \sum_{k=1}^{n} \mathfrak{g}_{kij} \hat{e}_k \tag{13}$$

De (12), lembrando que $\vec{z} = \sum_{i=1}^{n} z_i \hat{e}_i$, temos que a ação do mapa bilinear g pode ser escrita em termos das componentes tensoriais $\mathfrak{g}_{kij}$ e das componentes dos vetores $\vec{v}$, $\vec{w}$ e $\vec{z}$ como

$$z_k = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathfrak{g}_{kij} v_i w_j, \tag{14}$$

que pode ser encarada como uma generalização (tensorial) da ação matricial (9) e (10) de mapas lineares.

Chamamos de ordem do tensor o número de índices de suas componentes tensoriais. Assim, uma matriz é um tensor de ordem 2. O tensor associado ao mapa bilinear g acima é de ordem 3. Um vetor pode ser visto como um tensor de ordem 1, e um escalar como um de ordem 0. Na maioria das aplicações, o tensor é dado em termos de suas componentes num certo conjunto de bases. As regras de transformações de tensores diante de transformações de base é um capítulo a parte não muito relevante aqui, mas vamos comentar alguns pontos a seguir.

1.1 Covariante, contravariante e a notação de Einstein

Considere um vetor $\vec{v} \in V$, com suas componentes expressas numa certa base $\{\hat{e}_i\}$ de V. A pergunta relevante aqui é: se trocarmos para uma outra base,

por exemplo para $\{\hat{d}_i\}$, como se alterarão as componentes de $\vec{v}$? Vamos primeiro expressar a transformação da mudança de base. Será algo do tipo,

$$\hat{d}_i = \sum_{j=1}^{n} \mathfrak{m}_{ij} \hat{e}_j \tag{15}$$

sendo $\mathfrak{m}_{ij}$ um certo tensor de ordem 2 (a matriz de mudança de base), como introduzido na seção anterior. Porém, para que esta seja uma mudança de base genuína, a transformação deve ser invertível, quer dizer, deverá existir uma transformação inversa $\mathfrak{m}^{-1}_{ij}$ tal que

$$\sum_{k=1}^{n} \mathfrak{m}^{-1}_{ik} \hat{d}_k = \sum_{k=1}^{n} \sum_{j=1}^{n} \mathfrak{m}^{-1}_{ik} \mathfrak{m}_{kj} \hat{e}_j = \hat{e}_i, \tag{16}$$

de onde temos que a transformação inversa deve ser tal que

$$\sum_{k=1}^{n} \mathfrak{m}^{-1}_{ik} \mathfrak{m}_{kj} = \delta_{ij}. \tag{17}$$

Bem, já podemos responder nossa pergunta. Notem que

$$\begin{aligned} \vec{v} = \sum_{i=1}^{n} v_i \hat{e}_i = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} v_i \hat{e}_j &= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \mathfrak{m}^{-1}_{ik} \mathfrak{m}_{kj} v_i \hat{e}_j \\ &= \sum_{k=1}^{n} \underbrace{\left(\sum_{i=1}^{n} \mathfrak{m}^{-1}_{ik} v_i \right)}_{\tilde{v}_k} \underbrace{\left(\sum_{j=1}^{n} \mathfrak{m}_{kj} \hat{e}_j \right)}_{\hat{d}_k} = \sum_{k=1}^{n} \tilde{v}_k \hat{d}_k. \end{aligned} \tag{18}$$

Como vemos, diante da transformação (15), as componentes v_i do vetor $\vec{v}$ se transformam como

$$\tilde{v}_k = \sum_{i=1}^{n} \mathfrak{m}^{-1}_{ik} v_i, \tag{19}$$

quer dizer, se transformam de maneira **contrária** aos vetores de base (15). Componentes deste tipo são ditas contravariantes, e na notação de Einstein são sempre representadas com índices superiores. Assim, com a notação de Einstein, nosso vetor deve ser escrito como $\vec{v} = \sum_{i=1}^{n} v^i \hat{e}_i$. Porém, também faz parte da notação de Einstein **abolir** o somatório sempre que tivermos índices em cima e em baixo iguais. Portanto, basta escrever $\vec{v} = v^i \hat{e}_i$, e o

somatório estará subintendido. Esta notação é muito prática, mas nós não a usaremos aqui, pois não vamos diferenciar entre componentes contravariantes e covariantes, e sem essa diferenciação, comumente a notação de Einstein torna-se ambígua.

Componentes que se transformam como (15) são ditas covariantes e sempre são representadas com índices inferiores. Cabe a pergunta: que componentes seriam covariantes? Para respondermos esta pergunta, devemos lembrar que, dado um espaço vetorial V, temos sempre naturalmente um outro espaço associado, o espaço dual V^*. O espaço dual surge quando consideramos funcionais lineares do tipo

$$l : V \to \mathbb{R}, \tag{20}$$

sendo que estamos considerando, sem perda de generalidade para os nossos propósitos, que V é um espaço vetorial sobre os reais. Como l é linear, sua ação sobre um vetor[1] $\vec{v} = v^i \hat{e}_i$ será

$$l\left(v^i \hat{e}_i\right) = v^i \underbrace{l\left(\hat{e}_i\right)}_{\omega_i} = v^i \omega_i. \tag{21}$$

Dado um um funcional l, sua ação sobre um vetor $\vec{v}$ arbitrário será completamente caracterizada por suas n "componentes" associadas a ω_i. Em outras palavras, o espaço de todos os funcionais tem uma estrutura vetorial. Este é o espaço V^*, que tem a mesma dimensão que V. É natural exigirmos que o valor de $l(\vec{v})$, que é um número real, não dependa das bases que utilizemos em V e V^*. Como sabemos que as componentes v^i são contravariantes, a única maneira de termos (21) invariante por (15) é que as componentes ω_i sejam covariantes. Tensores mistos, aqueles que tem componentes contravariantes e covariantes, sempre podem ser vistos como mapas multilineares do tipo

$$f : V \times V^* \cdots \to V \times V^* \cdots, \tag{22}$$

com as respectivas componentes covariantes e contravariantes associadas aos espaços V^* e V correspondentes. Por exemplo, um tensor do tipo $\mathfrak{t}^i{}_k{}^j$ pode ser, por exemplo, o tensor associado ao mapa multilinear

$$t : V \times V^* \to V. \tag{23}$$

[1] Este será o nosso único uso da notação de Einstein!

2 Estrutura tensorial da multiplicação matricial

Vamos relembrar o algoritmo de Strassen. Seu ponto fundamental é a observação que o produto de duas matrizes 2×2

$$AB = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = C, \tag{24}$$

pode ser calculado com apenas 7 produtos. São eles

$$\begin{aligned} m_1 &= (a_{11} + a_{22})(b_{11} + b_{22}), & m_2 &= (a_{21} + a_{22})b_{11}, \\ m_3 &= a_{11}(b_{12} - b_{22}), & m_4 &= a_{22}(b_{21} - b_{11}), \\ m_5 &= (a_{11} + a_{12})b_{22}, & m_6 &= (a_{21} - a_{11})(b_{11} + b_{12}), \\ m_7 &= (a_{12} - a_{22})(b_{21} + b_{22}), \end{aligned} \tag{25}$$

de onde temos

$$\begin{aligned} c_{11} &= m_1 + m_4 - m_5 + m_7, & c_{12} &= m_3 + m_5, \\ c_{21} &= m_2 + m_4, & c_{22} &= m_1 - m_2 + m_3 + m_6. \end{aligned} \tag{26}$$

Contando-se cuidadosamente, vemos que são 18 adições[2]. Vemos que as entradas c_{ij} da matriz C são combinações de 7 produtos entre entradas de A e de B. Os produtos (25 são do tipo

$$m_K = \left(\sum_{ij} \alpha_{ijK} a_{ij} \right) \left(\sum_{mn} \beta_{mnK} b_{mn} \right), \tag{27}$$

sendo que os índices $i, j, m, n = 1, 2$ e $K = 1, 2, \ldots, 7$. Todos os índices minúsculos aqui percorrem os índices das matrizes em questão, enquanto os maiúsculos percorrem os produtos (25). No caso do algoritmo de Strassen, os coeficientes (componentes tensoriais) α_{ijK} e β_{mnK} assumem apenas os valores $-1, 0$, ou 1. As entradas de C, por sua vez, podem ser escritas como

$$c_{pq} = \sum_{k=1}^{T} \gamma_{qpK} m_K = \sum_{k=1}^{T} \gamma_{qpK} \left(\sum_{ij} \alpha_{ijK} a_{ij} \right) \left(\sum_{mn} \beta_{mnK} b_{mn} \right), \tag{28}$$

com $T = 7$. A troca da ordem dos índices q e p é comum na literatura, ela torna as expressões finais mais simétricas. Os $3 \times 4 \times 7 = 84$ coeficientes

[2]Vejam o EP3 para uma variante (Winograd) com 15 adições.

α_{ijK}, β_{mnK} e γ_{qpK} definem o esquema da multiplicação matricial de Strassen. Convém examinar alguns explicitamente. Por exemplo, da expressão de c_{11}, temos que

$$\gamma_{111} = \gamma_{114} = -\gamma_{115} = \gamma_{117} = 1 \quad \text{e} \quad \gamma_{112} = \gamma_{113} = \gamma_{116} = 0. \tag{29}$$

Da expressão de m_1, temos

$$\alpha_{111} = \alpha_{221} = 1, \quad \alpha_{121} = \alpha_{211} = 0, \quad \beta_{111} = \beta_{221} = 1, \quad \beta_{121} = \beta_{211} = 0. \tag{30}$$

Desta forma, podemos "ler" todos os coeficientes da multiplicação de Strassen. A Tabela 1 apresenta as componentes tensoriais associadas às multiplicações usual e de Strasen de matrizes 2×2.

A multiplicação usual também pode ser descrita na forma (28). Porém, neste caso, como são 8 produtos, teremos necessariamente $T = 8$. A multiplicação de duas matrizes quaisquer $N \times N$ pode sempre ser descrita no forma (28), e teremos um esquema subcúbico sempre que tivermos $T < N^3$. Notem que os coeficientes α_{ijK}, β_{mnK} e γ_{qpK} podem ser vistos como as componentes tensoriais de transformações lineares entre o espaço das componentes matriciais $\mathbb{R}^N \times \mathbb{R}^N$ e o espaço dos produtos $\mathbb{R}^T$. Vejam, por exemplo, α_{ijK}, cujo papel é determinar o primeiro fator $\sum_{ij} \alpha_{ijK} a_{ij}$ do produto m_k. Essa é a expressão de uma transformação linear que atua no espaço das componentes da matriz A, e tem valores em $\mathbb{R}^T$, *i.e.*, uma transformação linear do tipo $\mathfrak{A} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^T$, cujas componentes tensoriais são os coeficientes α_{ijK}. Obviamente, o mesmo se aplica para os outros coeficientes, com a curiosidade que a interpretação mais direta do tensor de componentes γ_{qpK} seja a de uma transformação linear $\mathfrak{A} : \mathbb{R}^T \to \mathbb{R}^N \times \mathbb{R}^N$.

Analisemos, agora, o seguinte problema. Quando um conjunto α_{ijK}, β_{mnK} e γ_{qpK} de $3N^2T$ componentes tensoriais definem um esquema de multiplicação matricial de matrizes $N \times N$? Bem, de (28), eles devem ser tais que

$$\sum_{k=1}^{T} \gamma_{qpK} \left(\sum_{i,j=1}^{N} \alpha_{ijK} a_{ij} \right) \left(\sum_{m,n=1}^{N} \beta_{mnK} b_{mn} \right) = \sum_{\ell=1}^{N} a_{p\ell} b_{\ell q}, \tag{31}$$

para todas as matrizes A e B ou, de forma equivalente, para todos os conjuntos de componentes a_{ij} e b_{mn}. Podemos determinar a expressão tensorial satisfeita pelo conjunto α_{ijK}, β_{mnK} e γ_{qpK} considerando **todas** as possíveis componentes a_{ij} e b_{mn}, mas isto é pouco prático no caso de matrizes quadradas de dimensão arbitrária N. Uma maneira mais sistemática é explorar

K	α_{ijK}	β_{ijK}	γ_{jiK}	K	α_{ijK}	β_{ijK}	γ_{jiK}
1	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	2	$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$
3	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	4	$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$
5	$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	6	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$
7	$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	8	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$

K	α_{ijK}	β_{ijK}	γ_{jiK}	K	α_{ijK}	β_{ijK}	γ_{jiK}
1	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	2	$\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$
3	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$	4	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} -1 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$
5	$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}$	6	$\begin{pmatrix} -1 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$
7	$\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$				

Tabela 1: Componentes tensoriais associadas à multiplicação de matrizes 2×2. Acima: multiplicação usual. Abaixo: Multiplicação de Strassen. Fixado K, as componentes tensoriais de α_{ijK}, β_{ijK} e γ_{ijK} devem ser lidas como componentes de matrizes 2×2. Notem que os índices do tensor γ_{ijK} estão invertidos.

que a expressão (31) é uma identidade válida para todos os valores das componentes matriciais a_{ij} e b_{mn} e que, portanto, podemos derivar em relação a a_{ij} e b_{mn} e obter novas identidades válidas. Derivando-se ambos os lados de (31) em relação a a_{rs} e a b_{uv}, e usando-se que

$$\frac{\partial a_{ij}}{\partial a_{rs}} = \delta_{ir}\delta_{js}, \quad \frac{\partial b_{mn}}{\partial b_{uv}} = \delta_{mu}\delta_{nv}, \quad \frac{\partial b_{mn}}{\partial a_{rs}} = \frac{\partial a_{ij}}{\partial b_{uv}} = 0, \tag{32}$$

teremos

$$\begin{aligned} \sum_{k=1}^{T} \gamma_{qpK} \left(\sum_{ij} \alpha_{ijK}\delta_{ir}\delta_{js} \right) \left(\sum_{mn} \beta_{mnK}\delta_{mu}\delta_{nv} \right) &= \sum_{\ell=1}^{N} \delta_{pr}\delta_{\ell s}\delta_{\ell u}\delta_{qv} \\ \sum_{k=1}^{T} \alpha_{rsK}\beta_{uvK}\gamma_{qpK} &= \delta_{pr}\delta_{su}\delta_{vq}, \end{aligned} \tag{33}$$

que é a chamada equação de Brent (ver, por exemplo, [1, 2]), a equação satisfeita pelos tensores que definem uma multiplicação matricial. Notem que são N^6 equações e $3N^2T$ incógnitas. Trata-se de um sistema cúbico sobre-determinado e pessimamente comportado. No entanto, sabemos que para $T = N^3$ ele sempre tem solução, pois sabemos que existe o esquema usual de multiplicação. A questão é se existe ou não solução com $T < N^3$, como existe no caso 2×2.

Um primeiro passo comum na análise de existência de esquemas subcúbicos de multiplicação é permitir que os valores de α_{ijK}, β_{mnK} e γ_{qpK} sejam reais arbitrários, e não apenas 0, -1 e 1 como no caso de Strassen. Notem que se provarmos que para um certo T a equação de Brent não tem solução real, então certamente estaremos excluindo também os casos com valores inteiros. Uma maneira de se atacar este problema é considerar a minimização da função não negativa

$$f(\alpha, \beta, \gamma) = \sum_{rsuvqp} \left(\sum_{k=1}^{T} \alpha_{rsK}\beta_{uvK}\gamma_{qpK} - \delta_{pr}\delta_{su}\delta_{vq} \right)^2, \tag{34}$$

cujos mínimos serão zero se e somente se a equação de Brent tiver solução para algum T. Há interesse atual em se empregar técnicas de machine learning para estudar os mínimos desse tipo de função, ver, por exemplo, [3].

Antes de partirmos para a construção de esquemas subcúbicos usando-se os designs (esféricos) combinatórios, há uma questão que merece menção explícita. O fato de permitir que os tensores α_{ijK}, β_{mnK} e γ_{qpK} tenham

componentes reais não altera a complexidade do esquema de multiplicação. Considerem, por exemplo, os produtos do tipo

$$\begin{aligned} m_1 &= (a_{11}+a_{22})(b_{11}+b_{22}), & m_2 &= (a_{21}+a_{22})b_{11}, \\ m_3 &= \tfrac{1}{2}a_{11}(2b_{12}-2b_{22}), & m_4 &= a_{22}(b_{21}-b_{11}), \\ m_5 &= (a_{11}+a_{12})b_{22}, & m_6 &= (a_{21}-a_{11})(b_{11}+b_{12}), \\ m_7 &= (a_{12}-a_{22})(b_{21}+b_{22}). \end{aligned} \tag{35}$$

Obviamente, trata-se de um esquema subcúbico do tipo Strassen, porém "pouco otimizado", pois há claramente três produtos desnecessários no cálculo de m_3. O ponto chave é que estes são produtos por "escalar", não produtos entre as componentes da matriz. Suponha que você implemente o produto de matrizes dessa forma mesmo, com as multiplicações por escalar desnecessárias. Qual a complexidade de um esquema de multiplicação de matrizes $2^k \times 2^k$ baseado nesses produtos? Bem, do EP3, sabemos que o número de produtos P_k e de somas S_k necessários obedecerão às equações[3]

$$P_k = 7P_{k-1} + 3\cdot 4^{k-1}, \quad S_k = 7S_{k-1} + 18\cdot 4^{k-1}, \tag{36}$$

de onde temos que a complexidade é inalterada: $O\left(N^{\log_2 7}\right)$. Podemos, inclusive, considerar o "pior dos casos", quando nenhuma componente dos tensores α_{ijK}, β_{mnK} e γ_{qpK} se anula e são todas diferentes de 1 e -1. O cálculo de cada produto m_k nesse caso requer 8 produtos de matrizes $2^{k-1}\times 2^{k-1}$ por um escalar. No total, os 7 produtos m_k necessitarão de $56\cdot 4^{k-1}$ produtos elementares. O cálculo de cada entrada da matriz produto c_{pq} irá requerer, no pior dos casos, mais 7 produtos produtos de matrizes $2^{k-1}\times 2^{k-1}$ por um escalar. O número total de produtos será

$$P_k = 7P_{k-1} + 7\cdot 2^{2k-1} + 56\cdot 4^{k-1}, \tag{37}$$

e vemos que o número de operações novamente[4] crescerá como $N^{\log_2 7}$.

3 Designs esféricos e multiplicação matricial

Vamos agora discutir uma construção explícita de algoritmos subcúbicos de multiplicação matricial. De fato, é a única construção explícita que conheço

[3] Resolva-as!

[4] Mostre!

válida para qualquer N. Da seção anterior, vimos que a multiplicação matricial pode ser expressa tensorialmente como

$$c_{pq} = \sum_{r,s,u,v} \delta_{pr}\delta_{su}\delta_{vq}a_{rs}b_{uv} = \sum_{s,u,v} \delta_{su}\delta_{vq}a_{ps}b_{uv} = \sum_{s,u} \delta_{su}a_{ps}b_{uq} = \sum_{s} a_{ps}b_{sq}, \quad (38)$$

e que a existência de algoritmos subcúbicos de multiplicação dependem da existência de uma decomposição do tipo (33), a saber

$$\sum_{k=1}^{T} \alpha_{rsK}\beta_{uvK}\gamma_{qpK} = \delta_{pr}\delta_{su}\delta_{vq}, \quad (39)$$

com $T < N^3$. Notem a estrutura dos índices dos dois lados da equação, há uma permutação cíclica

$$((r,s),(u,v),(q,p)) \Rightarrow ((p,r),(s,u),(v,q)). \quad (40)$$

Vamos olhar com um pouco mais de cuidado essa decomposição. O tensor do lado direito de (39) tem N^6 componentes independentes, sendo muitas delas nulas. Porém, a soma do lado esquerdo tem $3TN^2$ componentes independentes. Esta é uma situação muito parecida ao problema da decomposição em produtos de Kronecker do EP2! De fato, trata-se de um problema de decomposição tensorial em somas de produtos de Kronecker, mas de tensores de ordem 3, enquanto no EP2 discutimos o caso de ordem 2. O problema é de ordem 3 porque podemos fazer o *flattening* nos índices (r,s), (u,v) e (q,p) do lado esquerdo, e vemos que, de fato, estamos somando T produtos de 3 vetores (tensores de ordem 1). O problema da decomposição tensorial em produtos de Kronecker para tensores de ordem 3 é bastante complicado, com problemas ainda em aberto! Dado um tensor, o número mínimo de produtos de Kronecker necessário para decompô-lo como em (39) recebe o nome de *border rank* do tensor. Sabe-se que para $N = 2$, o border rank do tensor da multiplicação matricial é 7, o que implica que não existe nada mais eficiente para a multiplicação de matrizes 2×2 que o algoritmo de Strassen. Para $N = 3$, basicamente não se sabe nada! Sabe-se apenas que o border rank é menor ou igual a 23, pois são conhecidos algoritmos do tipo Strassen para $N = 3$ com $T = 23$. Como $\log_3 23 > \log_2 7$, os esquemas de multiplicação de matrizes 3×3 com 23 produtos são piores que o esquema de Strassen para matrizes 2×2.

Uma maneira bastante interessante, mas (infelizmente!) não muito eficiente, de se obter algoritmos do tipo Strassen para N arbitrário, ou em outras

palavras, decomposições do tipo (39) com $T < N^3$, envolve uma construção combinatória-geométrica bastante nova, os chamados desings esféricos, que exploraremos a partir de agora. O Apêndice destas notas é dedicado ao problema clássico dos designs combinatórios, pois é uma assunto que vale a pena ser visto, mesmo que superficialmente.

Comecemos como uma definição e em seguida vamos explorar suas aplicações. Um conjunto de s vetores $\vec{W}^k \in \mathbb{R}^d$, $k = 1, \ldots, s$ é dito um 2-design esférico se

$$\sum_{k=1}^{s} \vec{W}^k = 0, \tag{41}$$

e

$$\sum_{k=1}^{s} w_i^k w_j^k = \frac{s}{N}\delta_{ij}. \tag{42}$$

A primeira condição é simples de ser interpretada. Supondo-se que os vetores $\vec{W}^k$ correspondem a posições de pontos em $\mathbb{R}^N$, a primeira condição diz que o baricentro dos pontos é a origem, quer dizer, os pontos estão dispostos de maneira simétrica em torno da origem. A segunda condição pode ser escrita matricialmente como

$$\sum_k V^k (V^k)^t = \frac{s}{N} 1_N, \tag{43}$$

sendo 1_N a matriz identidade $N \times N$. Como a matriz identidade tem obviamente posto completo, temos que $s \geq N$ necessariamente, vejam o EP2. É sempre possível construir 2-designs para qualquer N arbitrário com $s = N+1$ vetores. Um exemplo explícito para $N = 2$ são os pontos sobre um triângulo equilátero inscrito na circunferência unitária:

$$W^1 = (0,1)^t, \quad W^2 = (-\tfrac{\sqrt{3}}{2}, -\tfrac{1}{2}), \quad W^3 = (\tfrac{\sqrt{3}}{2}, -\tfrac{1}{2}). \tag{44}$$

É fácil ver que (41) é verificada. Para (42), notem que

$$W^1(W^1)^t = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad W^2(W^2)^t = \tfrac{1}{4}\begin{pmatrix} 3 & \sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix},$$
$$W^3(W^3)^t = \tfrac{1}{4}\begin{pmatrix} 3 & -\sqrt{3} \\ -\sqrt{3} & 1 \end{pmatrix}, \tag{45}$$

e portanto

$$\sum_k W^k (W^k)^t = \frac{3}{2} 1_2, \tag{46}$$

como esperado. Para $N = 3$, os vértices de um tetraedro com baricentro na origem definem também um 2-design, e de maneira análoga para N arbitrário com os vértices de um N-simplex com baricentro na origem. O importante é que 2-designs esféricos sempre existem para qualquer dimensão N.

Dado um 2-design em dimensão N, podemos, a partir de (42), escrever (39) como

$$\delta_{pr}\delta_{su}\delta_{vq} = \left(\frac{N}{N+1}\right)^3 \sum_{A,B,C} w_p^A w_r^A w_s^B w_u^B w_v^C w_q^C, \tag{47}$$

sendo que as somas em A, B e C são de 1 a $s = N+1$. Há $(N+1)^3$ termos sendo somados do lado esquerdo e, portanto, esta não é uma decomposição tensorial útil para os nossos propósitos. Analisemos, porém, a seguinte expressão

$$\mathcal{T} = \sum_{A,B,C} w_p^A (w_r^A - w_r^B) w_s^B (w_u^B - w_u^C) w_v^C (w_q^C - w_q^A). \tag{48}$$

A primeira observação é que há vários termos que são nulos na série. Em particular, a soma deve ser feita sobre os índices A, B e C diferentes. Serão, portanto, $(N+1)N(N-1) = N^3 - N$ termos apenas. Vamos agora expandir todos os termos dessa expressão:

$$\begin{aligned}
\mathcal{T} \;=\; & \sum_{A,B,C} w_p^A w_r^A w_s^B w_u^B w_v^C w_q^C - \sum_{A,B,C} w_p^A w_r^A w_s^B w_u^B w_v^C w_q^A \\
& - \sum_{A,B,C} w_p^A w_r^A w_s^B w_u^C w_v^C w_q^C + \sum_{A,B,C} w_p^A w_r^A w_s^B w_u^C w_v^C w_q^A \\
& - \sum_{A,B,C} w_p^A w_r^B w_s^B w_u^B w_v^C w_q^C + \sum_{A,B,C} w_p^A w_r^B w_s^B w_u^B w_v^C w_q^A, \\
& + \sum_{A,B,C} w_p^A w_r^B w_s^B w_u^C w_v^C w_q^C - \sum_{A,B,C} w_p^A w_r^B w_s^B w_u^C w_v^C w_q^A.
\end{aligned} \tag{49}$$

São 8 termos. A observação crucial é que apenas o primeiro e o último são não nulos, pois todos os outros possuem termos que correspondem a um somatório de um termo linear no índice A, B ou C, e essa soma será zero por (41). Além disso, o primeiro termo é exatamente o somatório de (47). Ficamos, finalmente, com a seguinte decomposição em $N^3 - N + 1$ produtos do tensor da multiplicação matricial

$$\delta_{pr}\delta_{su}\delta_{vq} = \delta_{rs}\delta_{uv}\delta_{pq} + \left(\frac{N+1}{N}\right)^3 \mathcal{T} = \sum_{K=0}^{N^3-N} \alpha_{rsK}\beta_{utK}\gamma_{pqK}, \tag{50}$$

com

$$\alpha_{rs0} = \delta_{rs}, \quad \beta_{uv0} = \delta_{uv}, \quad \gamma_{pq0} = \delta_{pq}, \tag{51}$$

e as outras componentes correspondem $N^3 - N$ termos não nulos de (48). Vamos escrevê-las explicitamente para o caso $N = 2$, o caso de N genérico é análogo, mas muito mais longo.

$$\begin{aligned}
\left(\frac{N+1}{N}\right)^3 \mathcal{T} &= \left(\frac{N+1}{N}\right)^3 \sum_{A,B,C} w_p^A (w_r^A - w_r^B) w_s^B (w_u^B - w_u^C) w_v^C (w_q^C - w_q^A) \\
&= \underbrace{\frac{3}{2}(w_r^1 - w_r^2)w_s^2}_{\alpha_{rs1}} \underbrace{\frac{3}{2}(w_u^2 - w_u^3)w_v^3}_{\beta_{uv1}} \underbrace{\frac{3}{2}(w_q^3 - w_q^1)w_p^1}_{\gamma_{qp1}} \\
&+ \underbrace{\frac{3}{2}(w_r^1 - w_r^3)w_s^3}_{\alpha_{rs2}} \underbrace{\frac{3}{2}(w_u^3 - w_u^2)w_v^2}_{\beta_{uv2}} \underbrace{\frac{3}{2}(w_q^2 - w_q^1)w_p^1}_{\gamma_{qp2}} \\
&+ \underbrace{\frac{3}{2}(w_r^2 - w_r^1)w_s^1}_{\alpha_{rs3}} \underbrace{\frac{3}{2}(w_u^1 - w_u^3)w_v^3}_{\beta_{uv3}} \underbrace{\frac{3}{2}(w_q^3 - w_q^2)w_p^2}_{\gamma_{qp3}} \\
&+ \underbrace{\frac{3}{2}(w_r^2 - w_r^3)w_s^3}_{\alpha_{rs4}} \underbrace{\frac{3}{2}(w_u^3 - w_u^1)w_v^1}_{\beta_{uv4}} \underbrace{\frac{3}{2}(w_q^1 - w_q^2)w_p^2}_{\gamma_{qp4}} \\
&+ \underbrace{\frac{3}{2}(w_r^3 - w_r^2)w_s^2}_{\alpha_{rs5}} \underbrace{\frac{3}{2}(w_u^2 - w_u^1)w_v^1}_{\beta_{uv5}} \underbrace{\frac{3}{2}(w_q^1 - w_q^3)w_p^3}_{\gamma_{qp5}} \\
&+ \underbrace{\frac{3}{2}(w_r^3 - w_r^1)w_s^1}_{\alpha_{rs6}} \underbrace{\frac{3}{2}(w_u^1 - w_u^2)w_v^2}_{\beta_{uv6}} \underbrace{\frac{3}{2}(w_q^2 - w_q^3)w_p^3}_{\gamma_{qp6}} \\
&= \sum_{k=1}^{T=6} \alpha_{rsK}\beta_{uvK}\gamma_{qpK}. \qquad (52)
\end{aligned}$$

A complexidade de um esquema de multiplicação baseado na expansão em 2-designs esféricos será $O(N^\omega)$ com $\omega = \log_N(N^3 - N + 1)$. Será subcúbica sempre. Para $N = 2$, temos a mesma complexidade do algoritmo de Strassen. Para $N = 3$, temos $\omega = \log_3 25$, que é pior que os algoritmos conhecidos com $\omega = \log_3 23$. Para $N = 4$, temos $\log_4 61$, que é obviamente menor que 3, mas bastante pior que a iteração do algoritmo de Strassen, que exigiria apenas $7^2 = 49$ multiplicações. O mais decepcionante desta construção é que $\lim_{N\to\infty} \omega = 3$, *i.e.*, para matrizes muito grandes o produto é cúbico na prática. De qualquer forma, trata-se de uma construção que pode ser explorada em

situações mais gerais e, talvez, possa ser generalizada para outros produtos de interesse.

A Apêndice: Designs combinatórios

Designs são estruturas combinatórias com algum grau de simetria ou regularidade. O "folclore" usual atribui a Fisher e Yates, dois estatísticos importantíssimos do início do século XX, a introdução deste conceito, ver [4] para mais referências sobre o assunto. Vamos introduzir a ideia mais geral com um exemplo clássico. Suponha que você tenha N marcas de café, e queira submetê-las a um teste de qualidade com m "degustadores" independentes, mas equivalentes do ponto de vista de qualidade. Para evitar qualquer tipo de viés e ter um resultado final o mais homogêneo possível, cada marca deve ser provada por exatamente r degustadores diferentes, e cada degustador deve provar exatamente p amostras de cafés diferentes. É possível "planejar" (*design*, em inglês) um experimento com essas propriedades? Bem, obviamente isto depende dos valores de N, m, r e p. Suponha que são 6 marcas de café, $X = \{a, b, c, d, e, f\}$, e 10 degustadores. A partição do conjunto X nos seguintes blocos, sendo que cada um deles corresponde a um degustador,

$$\begin{aligned} B \;=\; & \{\{a,b,c\},\{a,b,d\},\{a,c,e\},\{a,d,f\},\{a,e,f\}, \\ & \{b,c,f\},\{b,d,e\},\{b,e,f\},\{c,d,e\},\{c,d,f\}\}, \end{aligned} \tag{53}$$

nos garante que:

1. cada degustador provará exatamente 3 marcas diferentes de café,
2. cada marca de café será provada por exatamente 5 degustadores diferentes,
3. cada par de marcas de café será provada por exatamente 2 degustadores diferentes.

Quer dizer, para $N = 6$, $m = 10$, $r = 5$, $p = 3$, é possível fazer o experimento de forma "balanceada", sem nenhum tipo de viés para marcas ou degustadores. Não é difícil perceber que nem sempre há soluções desse tipo para N, m, r e p arbitrários. Estamos em condições de apresentar a definição mais rigorosa de um design combinatório.

Definição. *Seja $X \neq \varnothing$ um conjunto com N elementos e B uma coleção de $m > 0$ subconjuntos (blocos) $b_i \subset X$ distintos, cada um com cardinalidade $p > 0$. O par (X, B) é denominado um t-design com parâmetros (N, p, q), $0 < t < p < N$ e $q > 0$, se cada subconjunto $x_j \subset X$ de cardinalidade t estiver contido em exatamente q blocos de B.*

Referências

[1] R.P. Brent, *Algorithms for matrix multiplication*, Technical Report TR-CS-70-157, DCS, Stanford (1970).

[2] R.W. Johnson and A.M. McLoughlin, *Noncommutative Bilinear Algorithms for* 3×3 *Matrix Multiplication*, SIAM Journal on Computing **15**, 595 (1986).

[3] A. Fawzi, M. Balog, A. Huang, *et al.*, *Discovering faster matrix multiplication algorithms with reinforcement learning*, Nature **610** 47 (2022).

[4] C.J. Colbourn, J.H. Dinitz (Eds.), *Handbook of Combinatorial Designs*, Chapman and Hall/CRC (2006).