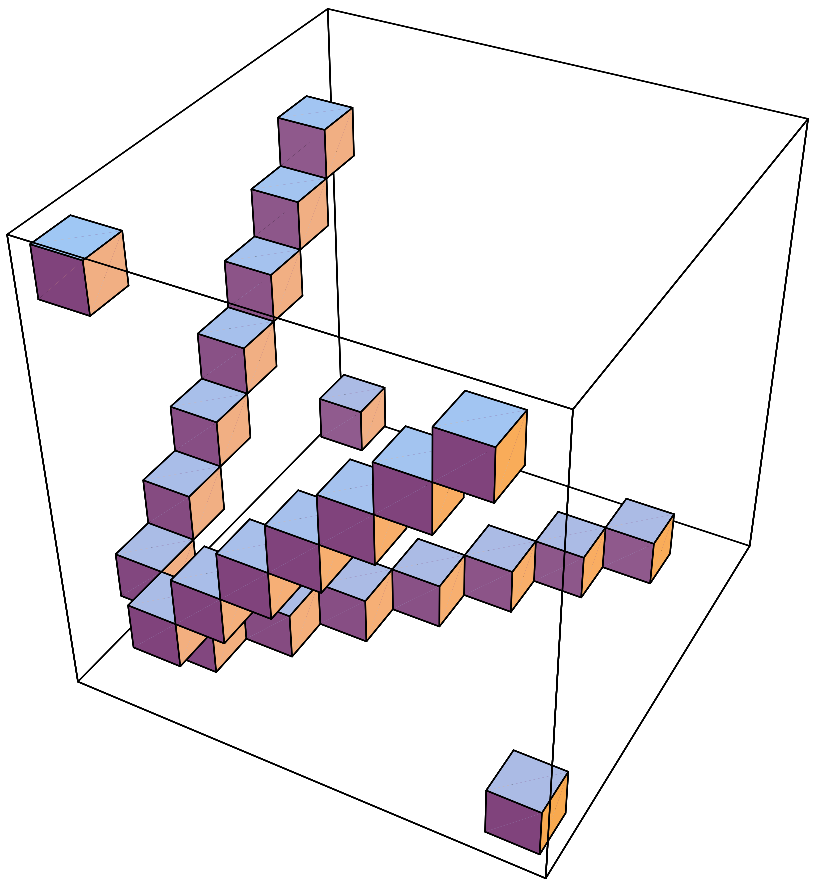
How Fast can We Multiply Matrices?



Amin Shokrollahi

Bell Labs

2×2-Matrix Multiplication

$$\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \cdot \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix}.$$

where

$$\forall\, 1 \le i, j \le 2\text{:} \qquad Z_{ij} = \sum_{\ell=1}^{2} X_{i\ell} Y_{\ell j}.$$

8 multiplications and 4 additions over any ring!

$O(n^3)$ arithemtic operations for $n \times n$-matrices over any ring.

Can we do better?

Strassen's Algorithm

$$
\begin{aligned}
P_1 &:= (-X_{12} + trX_{21} + X_{22})(Y_{11} + Y_{12} - Y_{21}) \\
P_2 &:= (X_{11} + X_{12} - X_{21} - X_{22})(Y_{11} + Y_{12}) \\
P_3 &:= -X_{21}(Y_{11} - Y_{21}) \\
P_4 &:= X_{12}Y_{22} \\
P_5 &:= (-X_{12} + X_{22})(-Y_{11} - Y_{12} + Y_{21} + Y_{22}) \\
P_6 &:= -X_{11}Y_{12} \\
P_7 &:= (X_{21} + X_{22})Y_{21}
\end{aligned}
$$

$$
\begin{aligned}
Z_{11} &= P_1 + P_2 + P_6 + P_7 \\
Z_{12} &= P_4 - P_6 \\
Z_{21} &= -P_3 + P_7 \\
Z_{22} &= P_1 + P_3 + P_4 + P_5.
\end{aligned}
$$

7 multiplications and 18 additions over any ring!

Recursion

$$\begin{array}{|c|c|}\hline 11 & 12 \\ \hline 21 & 22 \\ \hline \end{array} \; \begin{array}{|c|c|}\hline 11 & 12 \\ \hline 21 & 22 \\ \hline \end{array} = \begin{array}{|c|c|}\hline 11 & 12 \\ \hline 21 & 22 \\ \hline \end{array}$$

$M(n)$:= number of arithmetic operations to multiply two $n \times n$-matrices.

$$M(2^{n+1}) \leq 7 \cdot M(2^n) + 18 \cdot 2^{2n}$$

$$M(2^n) \leq 19 \cdot 7^n$$

$$M(N) \leq 7 \cdot 19 \cdot N^{\log_2 7} = O(N^{2.81})$$

The Exponent

Exponent of matrix multiplication (over field K):

$$\omega = \omega(K) := \inf\{\gamma \mid M(n) = O(n^\gamma)\}$$

Measures asymptotic complexity of matrix multiplication.

- $\omega \geq 2$.
- $\omega \leq \log_2 7$.

Bilinear Maps

U, V, and W finite dimensional K-spaces, $\Phi\colon U \times V \to W$ bilinear map:

$$\Phi(a_1 u_1 + a_2 u_2, b_1 v_1 + b_2 v_2) = \sum_{i,j=1}^{2} a_i b_j \Phi(u_i, v_j).$$

Fixing bases gives tensor $(t_{ij\ell})$:

$$\Phi(u_i, v_j) = \sum_k t_{ijk} w_k.$$

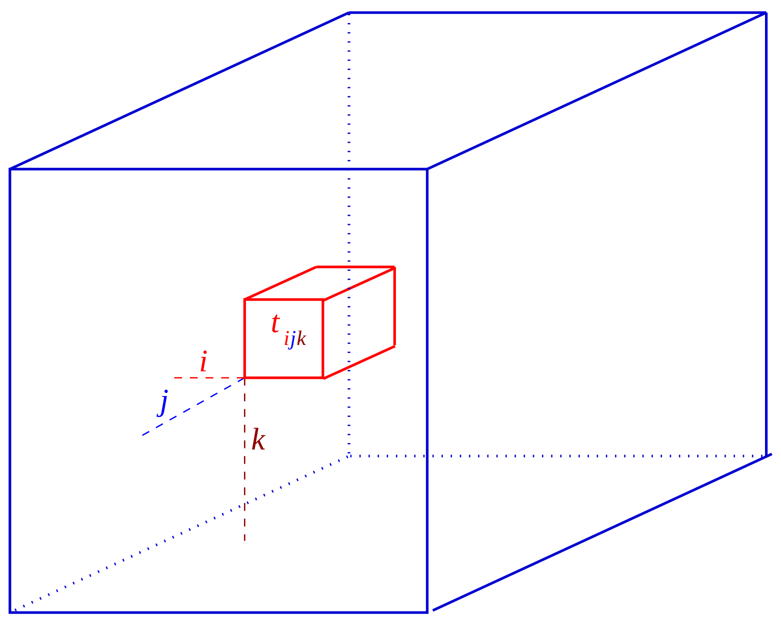
Representations

$$\Phi(u_i, v_j) = \sum_k t_{ijk} w_k.$$

by trilinear form

$$\sum_{ijk} t_{ijk} X_i Y_j Z_k,$$

or tensor-form



Examples

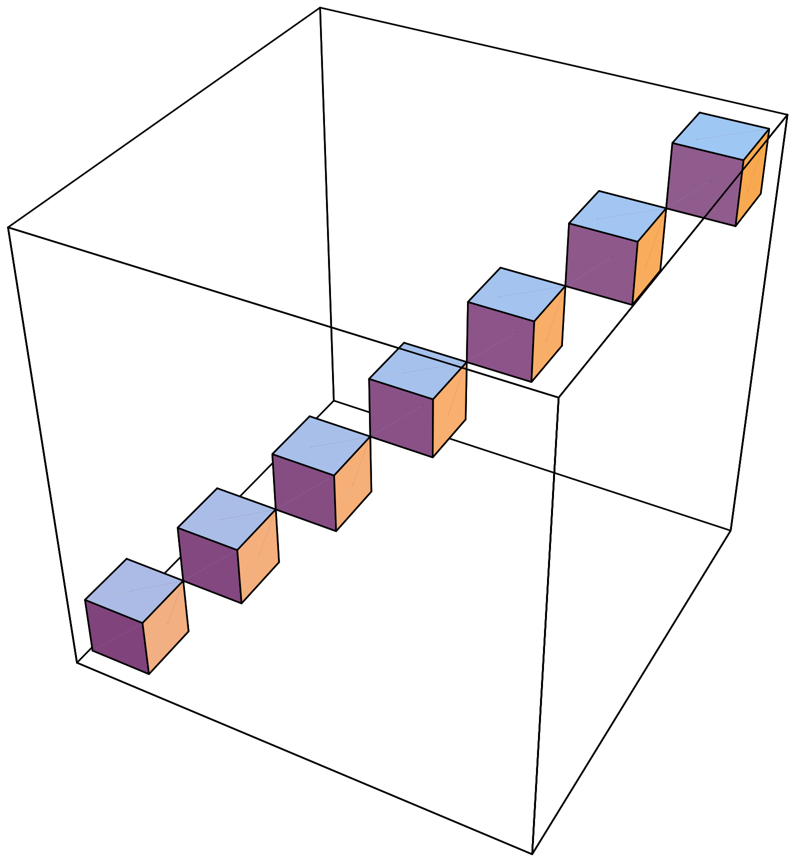
Bilinear map $\langle n \rangle : K^n \to K^n$ defined by

$$((a_1, \ldots, a_n), (b_1, \ldots, b_n)) \mapsto (a_1 b_1, \ldots, a_n b_n)$$

has trilinear form:

$$\sum_{i=1}^{n} X_i Y_i Z_i$$

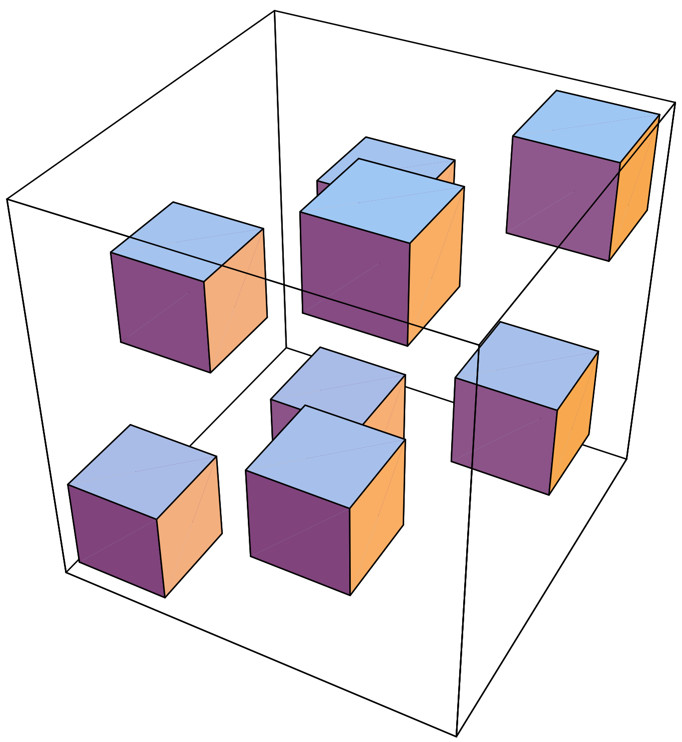
and tensor



Examples

Multiplication of 2×2-Matrices has trilinear form

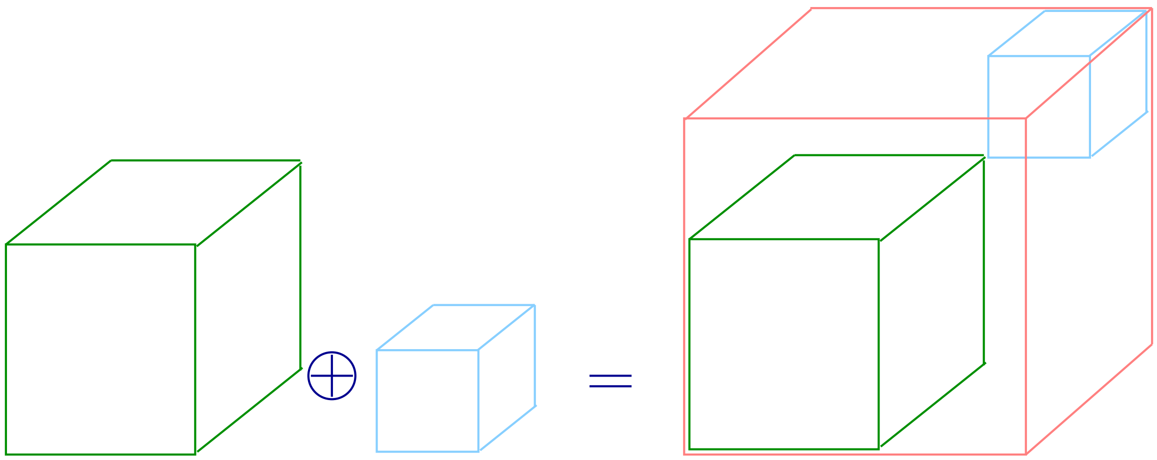
$$\begin{aligned}
&X_{11}Y_{11}Z_{11} + X_{12}Y_{21}Z_{11} + \\
&X_{11}Y_{12}Z_{12} + X_{12}Y_{22}Z_{12} + \\
&X_{21}Y_{11}Z_{21} + X_{22}Y_{21}Z_{21} + \\
&X_{21}Y_{12}Z_{22} + X_{22}Y_{22}Z_{22}
\end{aligned}$$



Direct Sum

$\Phi_i\colon U_i \times V_i \to W_i$, $i = 1, 2$, then

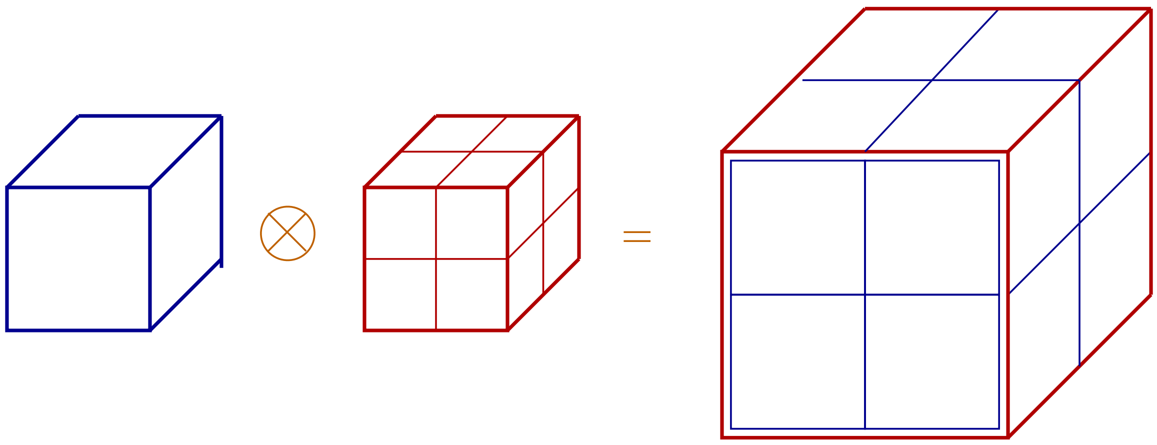
$$\Phi_1 \oplus \Phi_2\colon U_1 \oplus U_2 \times V_1 \oplus V_2 \to W_1 \oplus W_2.$$



Tensor Product

$\Phi_i\colon U_i \times V_i \to W_i$, $i = 1, 2$, then

$$\Phi_1 \otimes \Phi_2\colon U_1 \otimes U_2 \times V_1 \otimes V_2 \to W_1 \otimes W_2.$$



Examples

$\langle e,h,\ell\rangle$: $K^{e\times h}\times K^{h\times\ell}\to K^{e\times\ell}$ Matrix multiplication map.

∘ $\langle n\rangle \simeq \oplus_{i=1}^{n}\langle 1,1,1\rangle$.

∘ $\langle e_1,h_1,\ell_1\rangle\oplus\langle e_2,h_2,\ell_2\rangle$ Multiplication of block diagonal matrices.

∘ $\langle e_1,h_1,\ell_1\rangle\otimes\langle e_2,h_2,\ell_2\rangle\simeq\langle e_1e_2,h_1h_2,\ell_1\ell_2\rangle$.

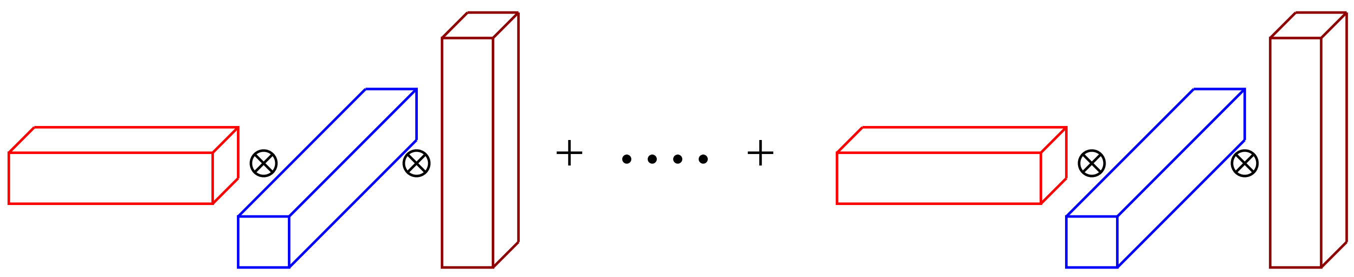
Rank of Bilinear Maps

Φ with trilinear form $t(X, Y, Z) = \sum_{i,j,k} t_{ijk} X_i Y_j Z_k$.

Rank of Φ:

$$R(\Phi) = \min\Big\{ r \mid \exists \text{ linear forms } (f_i, g_i, h_i) \in K[X] \times K[Y] \times K[Z]: \; t(X, Y, Z) = \sum_{i=1}^{r} f_i(X) g_i(Y) h_i(Z) \Big\}.$$

Independent of choice of basis.



r times

Rank and $\oplus/\otimes$

- $R(\Phi_1 \oplus \Phi_2) \leq R(\Phi_1) + R(\Phi_2)$.

- $R(\Phi_1 \otimes \Phi_2) \leq R(\Phi_1) \cdot R(\Phi_2)$.

Rank and Exponent

We have

$$\begin{aligned}\omega &= \inf\{\gamma \mid M(n) = O(n^\gamma)\} \\ &= \inf\{\gamma \mid R(\langle n,n,n\rangle) = O(n^\gamma)\}\end{aligned}$$

over any field.

$$R(\langle n,n,n\rangle) \le r \Longrightarrow \omega \le \log_n r.$$

Symmetrization

$\langle 3, 2, 1 \rangle$:

$$X_{11}Y_1Z_1 + X_{12}Y_2Z_1 + X_{21}Y_1Z_2 + \\ X_{22}Y_2Z_2 + X_{31}Y_1Z_3 + X_{32}Y_2Z_3$$

$\langle 1, 3, 2 \rangle$:

$$X_1Y_{11}Z_1 + X_2Y_{21}Z_1 + X_3Y_{31}Z_1 + \\ X_1Y_{12}Z_2 + X_2Y_{22}Z_2 + X_3Y_{32}Z_2.$$

A change of variables maps one trilinear form into the other. So,

$$R(\langle 3, 2, 1 \rangle) = R(\langle 1, 3, 2 \rangle).$$

In general we have

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

Approximative Algorithms

Bini et al.'s example: $\ulcorner \times \square$ has trilinear form

$$\begin{aligned} F &= X_{11}Y_{11}Z_{11} + X_{11}Y_{12}Z_{12} + X_{12}Y_{21}Z_{11} + \\ &\quad X_{12}Y_{22}Z_{12} + X_{21}Y_{11}Z_{21} + X_{21}X_{12}Z_{22}. \end{aligned}$$

$R(\ulcorner \times \square) = 6$, but...

$$\begin{aligned} F_1(\epsilon) &= (X_{12} + \epsilon X_{11})(Y_{12} + \epsilon Y_{22})Z_{21} + \\ &\quad (X_{21} + \epsilon X_{11})Y_{11}(Z_{11} + \epsilon Z_{12}) - \\ &\quad X_{12}Y_{12}(Z_{11} + Z_{21} + \epsilon Z_{22}) - \\ &\quad X_{21}(Y_{11} + Y_{12} + \epsilon Y_{21})Z_{11} + \\ &\quad (X_{12} + X_{21})(Y_{12} + \epsilon Y_{21})(Z_{11} + \epsilon Z_{22}) \\ \\ &= \epsilon F + \epsilon^2 G(\epsilon) \end{aligned}$$

for some $G(\epsilon) \in K[\epsilon][X, Y, Z]$.

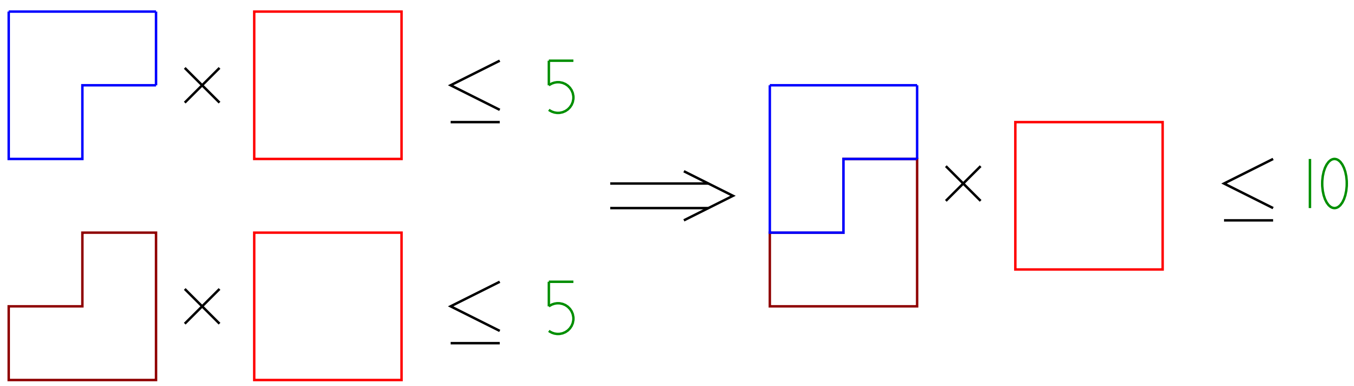
Border Rank

$$\underline{R}(\Phi) = \min\Big\{r \mid \ \exists \text{ linear forms } (f_i, g_i, h_i) \in K[\epsilon][\underline{X}] \times K[\epsilon][\underline{Y}] \times K[\epsilon][\underline{Z}], \ G(\epsilon) \in K[\epsilon][X, Y, Z], q \geq 0: \ \epsilon^q t(X,Y,Z) + \epsilon^{q+1} G(\epsilon) = \sum_{i=1}^{r} f_i(X) g_i(Y) h_i(Z)\Big\}.$$

- $\underline{R}(\Phi_1 \oplus \Phi_2) \leq \underline{R}(\Phi_1) + \underline{R}(\Phi_2)$.

- $\underline{R}(\Phi_1 \otimes \Phi_2) \leq \underline{R}(\Phi_1) \cdot \underline{R}(\Phi_2)$.

- $R(\langle e, h, \ell \rangle) \leq r \Longrightarrow \omega \leq \log_{eh\ell} r$.



Bini-Capovani-Lotti-and Romani et al., 1979:

$$\omega \leq 3 \log_{12} 10 < 2.7797.$$

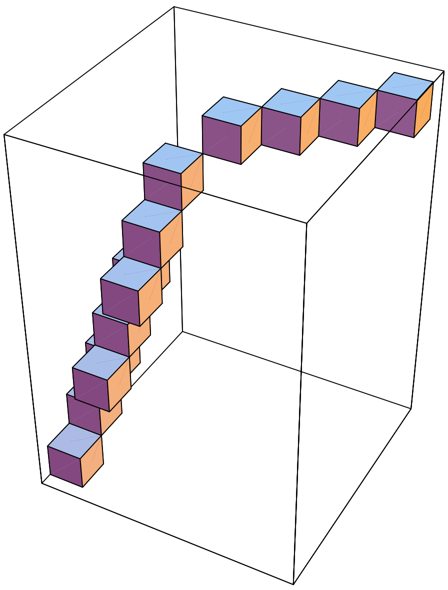
Schönhage's Asypmtotic Sum Inequality
(τ-Theorem)

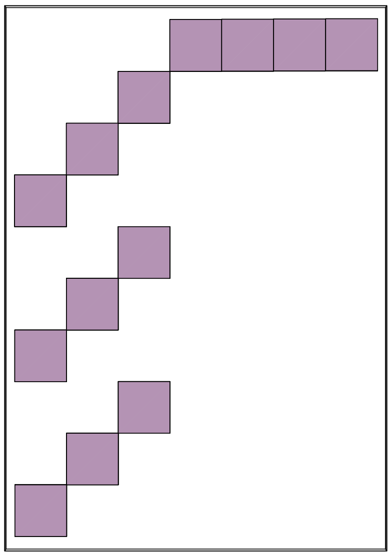
$$\underline{R}\left(\bigoplus_{i=1}^{s}\langle e_i, h_i, \ell_i\rangle\right) \leq r \Longrightarrow \sum_{i=1}^{s}(e_i h_i \ell_i)^{\omega/3} \leq r.$$

Schönhage's example: $\underline{R}\,(\langle e, 1, \ell\rangle \oplus \langle 1, (e-1)(\ell-1), 1\rangle) = e\ell + 1$.

$e = \ell = 4$: $16^{\omega/3} + 9^{\omega/3} \leq 17$, i.e.,

$$\omega < 2.55.$$





Strassen's Laser Method

Strategy up to now: Take a direct sum of matrix tensors for which there exists a good approximative algorithm, and extract asymptotic information using the Asymptotic Sum inequality.

Laser method "recipe":

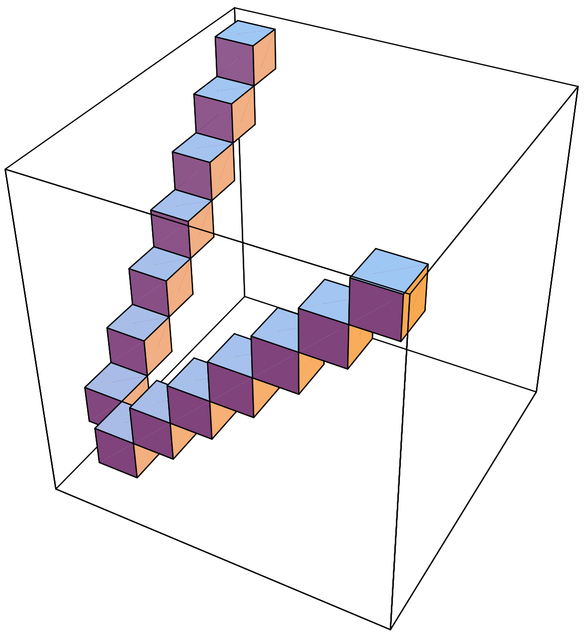
- Start with a tensor with a *block* and a *fine* structure, where each fine structure is isomorphic to a matrix tensor.
- Take tensor powers and identify a large "diagonal" Δ.
- Get rid of the off-diagonal entries using degeneration.
- Apply the Asymptotic Sum Inequality to Δ.

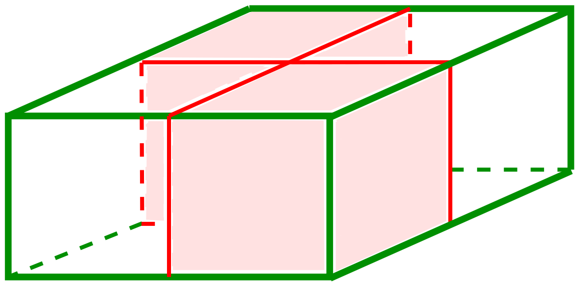
Laser Method

Theorem (Strassen): If t has block structure isomorphic to $\langle e, h, \ell \rangle$ and fine structure isomorphic to $\langle e_i, h_i, \ell_i \rangle$ with $(e_i h_i \ell_i)^3 = q$ for some fixed q, then

$$(eh\ell)^2 q^{\omega} \leq \underline{R}(t)^3.$$

$$F = \sum_{j=1}^{q} (X_0 Y_j Z_j + X_j Y_0 Z_j).$$





Block structure is isomprphic to $\langle 1, 2, 1 \rangle$.

Fast Algorithm

$$\sum_{j=1}^{q}(X_0+\epsilon X_j)(Y_0+\epsilon Y_j)Z_j - X_0Y_0(\sum_{j=1}^{q} Z_j) = \epsilon F + \epsilon^2 G$$

for some $G \in K[\epsilon][X,Y,Z]$.

So, $\underline{R}(F) \leq q+1$, and $4q^\omega \leq (q+1)^\omega$. Put $q = 5$ (Strassen, 1987):

$$\omega < 2.48$$

Coppersmith-Winograd's Extension

The block structure need not be a matrix tensor, but only planar.

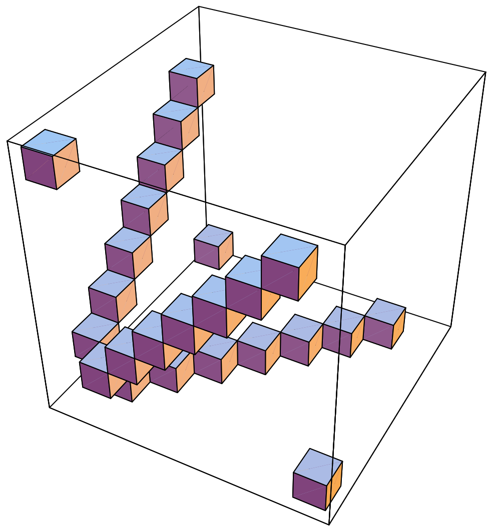
Theorem(Coppersmith-Winograd): If t has a planar block structure B and for each $b \in B$ the corresponding fine structure is isomorphic to some matrix tensor $\langle e_b, h_b, \ell_b \rangle$, then for any probability distribution P on B we have

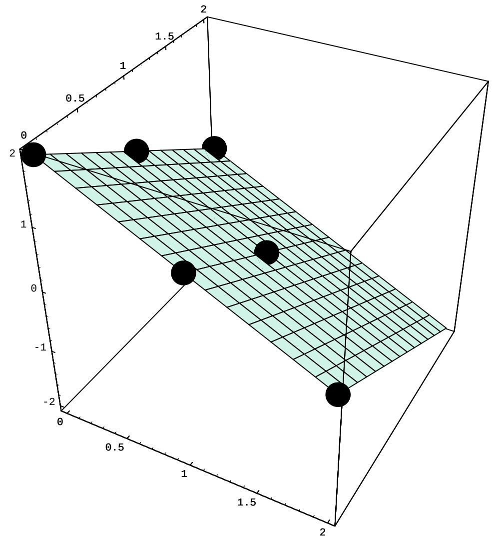
$$\min_{1 \leq m \leq 3} H(P_m) + \frac{\omega}{3} \sum_{b \in B} P(b) \log(e_b h_b \ell_b) \leq \log \underline{R}(t).$$

The main ingredient of the proof is the use of random sampling to find a large diagonal. (*The probabilistic method*.)

Application

$$F = \sum_{j=1}^{q} (X_0 Y_j Z_j + X_j Y_0 Z_j + X_j Y_j Z_0) + X_0 Y_0 Z_{q+1} + X_0 Y_{q+1} Z_0 + X_{q+1} Y_0 Z_0.$$





Block structure is planar and lies on hyperplane $x + y + z = 2$.

Application

$$
\begin{aligned}
&\sum_{i=1}^{q} \epsilon(X_0 + \epsilon X_i)(Y_0 + \epsilon Y_i)(Z_0 + \epsilon Z_i) \\
&\quad - (X_0 + \epsilon^2 \sum_{i=1}^{q} X_i)(Y_0 + \epsilon^2 \sum_{i=1}^{q} Y_i)(Z_0 + \epsilon^2 \sum_{i=1}^{q} Z_i) \\
&\quad + (1 - q\epsilon)(X_0 + \epsilon^3 X_{q+1})(Y_0 + \epsilon^3 Y_{q+1}) \\
&\qquad (Z_0 + \epsilon^3 Z_{q+1}) \\
= \;& \epsilon^3 F + \epsilon^4 G
\end{aligned}
$$

for some $G \in K[\epsilon][X, Y, Z]$.

So, $\underline{R}(F) \leq q + 2$ and (Coppersmith-Winograd, 1987-90)

$$\omega < 2.39$$

$\omega = 2$?

Theorem (Coppersmith): Over an infinite field we have

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

where $\alpha = \log 4/(5 \log 5) \approx 0.17$.

- Classify all (n, n, n)-tensors of border rank n.

- Is there an efficient computational procedure to determine all these tensors for $n \leq 4$ and $K := \mathbf{F}_2$, say?