Approaches to bounding the exponent of matrix multiplication

Chris Umans
Caltech

Based on joint work with Noga Alon, Henry Cohn, Bobby Kleinberg, Amir Shpilka, Balazs Szegedy

Simons Institute Sept. 17, 2014
Introduction

- Standard method: $O(n^3)$ operations
- Strassen (1969): $O(n^{2.81})$ operations
Introduction

- Standard method: $O(n^3)$ operations
- Strassen (1969): $O(n^{2.81})$ operations

The exponent of matrix multiplication:
smallest number $\omega$ such that for all $\varepsilon > 0$

$O(n^{\omega + \varepsilon})$ operations suffice
History

- Standard algorithm $\omega \leq 3$
- Strassen (1969) $\omega < 2.81$
- Pan (1978) $\omega < 2.79$
- Bini; Bini et al. (1979) $\omega < 2.78$
- Schönhage (1981) $\omega < 2.55$
- Pan; Romani; Coppersmith + Winograd (1981-1982) $\omega < 2.50$
- Strassen (1987) $\omega < 2.48$
- Coppersmith + Winograd (1987) $\omega < 2.375$
- Stothers (2010) $\omega < 2.3737$
- Williams (2011) $\omega < 2.3729$
- Le Gall (2014) $\omega < 2.37286$
Outline

1. main ideas from Strassen 1969 through Le Gall 2014

2. approach via embedding into semi-simple algebra multiplication
   - groups
   - coherent configurations/association schemes

Sept. 17, 2014
The matrix multiplication tensor

\(<n,n,n>\) is a \(n^2 \times n^2 \times n^2\) tensor described by trilinear form \(\sum_{i,j,k} X_{i,j} Y_{j,k} Z_{k,i}\)
The matrix multiplication tensor

\(<n,n,n>\) is a \(n^2 \times n^2 \times n^2\) tensor described by trilinear form \(\sum_{i,j,k} X_{i,j} Y_{j,k} Z_{k,i}\)

\[
\begin{array}{cc}
  a_{11} & a_{12} \\
  a_{21} & a_{22} \\
\end{array}
\times
\begin{array}{cc}
  b_{11} & b_{12} \\
  b_{21} & b_{22} \\
\end{array}
=
\begin{array}{cc}
  c_{11} & c_{12} \\
  c_{21} & c_{22} \\
\end{array}
\]
The matrix multiplication tensor

\[<n,n,n> \text{ is a } n^2 \times n^2 \times n^2 \text{ tensor described by trilinear form } \sum_{i,j,k} X_{i,j} Y_{j,k} Z_{k,i}\]

\[
\begin{array}{cc}
  a_{11} & a_{12} \\
  a_{21} & a_{22} \\
\end{array}
\times
\begin{array}{cc}
  b_{11} & b_{12} \\
  b_{21} & b_{22} \\
\end{array}
=
\begin{array}{cc}
  c_{11} & c_{12} \\
  c_{21} & c_{22} \\
\end{array}
\]
The matrix multiplication tensor

\[ \langle n, n, n \rangle \] is a \( n^2 \times n^2 \times n^2 \) tensor described by trilinear form

\[ \sum_{i,j,k} X_{i,j} Y_{j,k} Z_{k,i} \]
The matrix multiplication tensor

\(<n, n, n>\) is a \(n^2 \times n^2 \times n^2\) tensor described by trilinear form \(\sum_{i,j,k} X_{i,j} Y_{j,k} Z_{k,i}\)

\[
\begin{array}{cc}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{array}
\times
\begin{array}{cc}
b_{11} & b_{12} \\
b_{21} & b_{22}
\end{array}
=
\begin{array}{cc}
c_{11} & c_{12} \\
c_{21} & c_{22}
\end{array}
\]
The matrix multiplication tensor

\(<n,m,p>\) is a \(nm \times mp \times pn\) tensor described by trilinear form

\[\sum_{i,j,k} X_{i,j} Y_{j,k} Z_{k,i}\]

Each of \(np\) slices of \(<n,m,p>\):

\[\ldots\]

Sept. 17, 2014
Strategies for upper bounding the rank of the matrix multiplication tensor
Upper bounds on rank

• Observation: \(<n,n,n>^i = <n^i, n^i, n^i>
  \) \(R(<n^i, n^i, n^i>) \cdot R(<n,n,n>)^i\)

• **Strategy I**: bound rank for small \(n\) by hand
  – \(R(<2,2,2>) = 7\) ! < 2.81
  – \(R(<3,3,3>) = 2 [19..23]\) (worse bound)

  – even computer search infeasible…
Upper bounds on rank

• **Border rank** = rank of sequence of tensors approaching target tensor *entrywise*

  \[
  \begin{array}{ccc}
  1 & 1 & 1 \\
  1 & 2 & \\
  \end{array}
  \]
  \[
  \text{rank} = 3
  \]
  \[
  \begin{array}{ccc}
  2^{-1} & 1 & \\
  1 & 2 & \\
  \end{array}
  \]
  \[
  \text{border rank} = 2:
  \]

• **Strategy II**: bound *border rank* for small \(n\)

• Lemma: \(R(<n,n,n>) < r \) ! < \(\log_n r\)
  
  – \(R(<2,2,3>) \cdot 10 ! < 2.79\)
Upper bounds on rank

• Direct sum of tensors $<n,n,n> \otimes <m,m,m>$
  (multiple matrix multiplications in parallel)

• Strategy III: bound (border) rank of direct sums of small matrix multiplication tensors

\[ R(<n_1,n_1,n_1> \otimes \ldots \otimes <n_k,n_k,n_k>) < r ) \sum_i n_i! < r \]

- $R(<4,1,3> \otimes <1,6,1>) \cdot 13 \quad ! < 2.55$
Upper bounds on rank

- **Strategy IV**: Strassen “laser method”
  - tensor with “coarse structure” of MM and “fine structure” components isomorphic to MM
  (many independent MMs in high tensor powers)

```
coarse structure <1,2,1>
```

```
\text{fine} = \text{scalar} \times \text{row vector} \times \text{col vector} \times \text{scalar}
```
Upper bounds on rank

• **Strategy IV:** Strassen “laser method”
  
  tensor with “coarse structure” of MM and “fine structure” components *isomorphic* to MM
  
  (many independent MMs in high tensor powers)

border rank = $q + 1$;

$q = 5$ yields $! < 2.48$
Upper bounds on rank

- Coppersmith-Winograd and beyond: border rank of this tensor is $q+2$:

$$\sum_{i=1}^{q} X_0 Y_i Z_i + X_i Y_0 Z_i + X_i Y_i Z_0 +$$

$$X_0 Y_0 Z_{q+1} + X_0 Y_{q+1} Z_0 + X_{q+1} Y_0 Z_0$$

- 6 “pieces”: target proportions in high tensor power affect # and size of independent MMs
- $q = 6$ yields $! < 2.388$
Upper bounds on rank

- Coppersmith-Winograd and beyond: analyze tensor powers of this tensor

\[ T_q = \sum_{i=1}^{q} X_0 Y_i Z_i + X_i Y_0 Z_i + X_i Y_i Z_0 + X_0 Y_0 Z_{q+1} + X_0 Y_{q+1} Z_0 + X_{q+1} Y_0 Z_0 \]

<table>
<thead>
<tr>
<th>Tensor power</th>
<th># “pieces”</th>
<th>bound</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>36</td>
<td>2.375</td>
<td>C-W</td>
</tr>
<tr>
<td>4</td>
<td>1296</td>
<td>2.3737</td>
<td>Stothers</td>
</tr>
<tr>
<td>8</td>
<td>1679616</td>
<td>2.3729</td>
<td>Williams</td>
</tr>
<tr>
<td>16</td>
<td>2.82 x 10^{12}</td>
<td>2.3728640</td>
<td>Le Gall</td>
</tr>
<tr>
<td>32</td>
<td>7.95 x 10^{24}</td>
<td>2.3728639</td>
<td>Le Gall</td>
</tr>
</tbody>
</table>
Upper bounds on rank

• Coppersmith-Winograd and beyond

<table>
<thead>
<tr>
<th>Tensor power</th>
<th># pieces</th>
<th>bound</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>36</td>
<td>2.375</td>
<td>C-W</td>
</tr>
<tr>
<td>4</td>
<td>1296</td>
<td>2.3737</td>
<td>Stothers</td>
</tr>
<tr>
<td>8</td>
<td>1679616</td>
<td>2.3729</td>
<td>Williams</td>
</tr>
<tr>
<td>16</td>
<td>2.82 x 10^12</td>
<td>2.3728640</td>
<td>Le Gall</td>
</tr>
<tr>
<td>32</td>
<td>7.95 x 10^24</td>
<td>2.3728639</td>
<td>Le Gall</td>
</tr>
</tbody>
</table>

• Ambainis-Filmus 2014: N-th tensor power cannot beat bound of 2.3078
A different approach

• So far...
  – bound border rank of small tensor (by hand)
  – asymptotic bound from high tensor powers

• Disadvantages
  – limited universe of “starting” tensors
  – high tensor powers hard to analyze
matrix multiplication
via groups and
coherent configurations / association schemes
The general approach

- Cohn-Umans 2003, 2012:
  - *embed* $n \times n$ matrix multiplication into semi-simple algebra multiplication
  - semi-simple: isomorphic to block-diagonal MM
  - key hope: “nice basis” w/ combinatorial structure
  - reduce $n \times n$ MM to smaller MMs; recurse
The Group Algebra

- given finite group $G$, group algebra $C[G]$ has elements $\sum g \ a_g g$
  with multiplication

$$ (\sum_g a_g g)(\sum_h b_h h) = \sum_f (\sum_{gh = f} a_g b_h) f $$

- structure: $C[G]'(C^{d_1 \times d_1}) \times \ldots \times (C^{d_k \times d_k})$
- group elements are “nice basis”
“Nice basis” embedding:

Subgroups $X, Y, Z$ of $G$ satisfy the triple product property if for all $x \in X$, $y \in Y$, $z \in Z$:

$$xyz = 1 \quad \text{iff} \quad x = y = z = 1.$$
The embedding:

Subsets $X, Y, Z$ of $G$ satisfy the **triple product property** if for all $x \in Q(X), y \in Q(Y), z \in Q(Z)$:

$$xyz = 1 \text{ iff } x = y = z = 1.$$ 

$$A = \sum_{x,y} a_{x,y} (x\ y^{-1}) \quad B = \sum_{y,z} b_{y,z} (y\ z^{-1})$$

**Claim:** $(AB)_{x,z} = \text{coeff. on } (x\ z^{-1}) \text{ in } A^*B.$
The embedding:

Subsets $X, Y, Z$ of $G$ satisfy the triple product property if for all $x \in Q(X), y \in Q(Y), z \in Q(Z)$:

$$xyz = 1 \quad \text{iff} \quad x = y = z = 1.$$ 

$$A = \Sigma a_{x_1,y_1}(x_1y_1^{-1}) \quad B = \Sigma b_{y_2,z_2}(y_2z_2^{-1})$$

Claim: $$(AB)_{x_3,z_3} = \text{coeff. on } (x_3z_3^{-1}) \text{ in } A^*B.$$ 

$$(x_1y_1^{-1})(y_2z_2^{-1}) = x_3z_3^{-1} \quad \Rightarrow \quad x_3^{-1}x_1y_1^{-1}y_2z_2^{-1}z_3 = 1$$
How many multiplications?

Embedding + structure of C[G] yields bound on rank (′ # multiplications):

• we use \( m \leq \Sigma d_i^3 \) mults
• really \( m = \Sigma d_i! \) mults
• at least \( m \geq \Sigma d_i^2 = |G| \) mults

First Challenge: embed \( k \times k \) matrix multiplication in group of size \( \frac{1}{4} k^2 \)
The embedding

First Challenge: embed $k \times k$ matrix multiplication in group of size $\frac{1}{4} k^2$

• simple pigeonhole argument:
  – embedding in an abelian group requires group to have size $k^3$
The triangle construction

**Theorem**: can embed $k \times k$ matrix multiplication in symmetric group of size $k^2 + o(1)$

$n$ objects

• subgroup $X$
• subgroup $Y$
• subgroup $Z$

need $X$, $Y$, $Z$ in $S_n$ all with size $\approx |S_n|^{1/2}$
The triangle construction

- X moves points within rows
- Y moves points within columns
- Z moves points within diagonals
- want: $xyz = 1 \implies x = y = z = 1$
The triangle construction

**Theorem:** can embed $k \times k$ matrix multiplication in symmetric group of size $k^2 + o(1)$

Unfortunately, $d_{\text{max}} > |X| (= |Y| = |Z|)$
What should we be aiming for?

**Theorem:** in group $G$ supporting $k \times k$ matrix multiplication with character degrees $d_1, d_2, d_3, \ldots$, we obtain:

$$k^\omega \cdot \sum_i d_i^\omega$$

- If $X, Y, Z \in G$ satisfy T.P.P. and
  - $(|X|\cdot|Y|\cdot|Z|)^{1/3} = k \cdot |G|^{1/2} - o(1)$
  - $d_{\text{max}} \cdot |G|^{1/2} - 2$

then $! = 2$
Constructions in linear groups

• Good candidate family: \( SL(n, q) \) for fixed dimension \( n \)

• In \( SL(n, \mathbb{R}) \) these three subgroups satisfy the triple product property:
  – upper-triangular with ones on the diagonal
  – lower-triangular with ones on the diagonal
  – the special orthogonal group \( SO(n, \mathbb{R}) \)

and dim. of each is \( \frac{1}{2} \) dim. of \( G \) as \( n \to 1 \)
Group algebra approach

• [CKSU 2005] wreath product groups yield:
  – \(1 < 2.48, 1 < 2.41\)
  – key part of construction is combinatorial
  – two conjectures implying \(1 = 2\)

• Main disadvantage:
  – non-trivial results require non-abelian groups
  – most ideas foiled by too-large char. degrees
General semi-simple algebras

- (finite dimensional, complex) algebra specified by
  - “nice basis” \( e_1, e_2, \ldots, e_r \)
  - structure constants \( \delta_{i,j,k} \) satisfying
    \[
    e_i e_j = \sum_k \delta_{i,j,k} e_k
    \]
  “realizes” MM if contains*: MM tensor \( <n,n,n> \)

*structural tensor of algebra mult.
Weighted vs. unweighted MM

• Technical problem:
  – MM tensor \( <n,n,n> \) given by \( \sum_{i,j,k} X_{i,j}Y_{j,k}Z_{k,i} \)
  – embedding into algebra bounds rank of tensor given by
    \[ \sum_{i,j,k} \delta_{i,j,k} X_{i,j}Y_{j,k}Z_{k,l} \]
      (with \( \delta_{i,j,k} \neq 0 \))
  – group algebra: \( \delta_{i,j,k} \) always 0 or 1
Weighted vs. unweighted MM

s-rank of tensor $T$: minimum rank of tensor with same support as $T$

Does upper bound on s-rank of MM tensor imply upper bound on ordinary rank?

Example:

\[
\begin{array}{cc}
  a_{11} & a_{12} \\
  a_{21} & a_{22} \\
\end{array}
\times
\begin{array}{cc}
  b_{11} & b_{12} \\
  b_{21} & b_{22} \\
\end{array}
= 
\begin{array}{cc}
  a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\
  a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\
\end{array}
\]
Weighted vs. unweighted MM

**s-rank** of tensor $T$: minimum rank of tensor with same **support** as $T$

Does upper bound on s-rank of MM tensor imply upper bound on ordinary rank?

**Example:**

\[
\begin{array}{cc}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{array}
\times
\begin{array}{cc}
b_{11} & b_{12} \\
b_{21} & b_{22}
\end{array}
\]

\[
\begin{array}{c}
a_{11}b_{11} + a_{12}b_{21} \\
a_{11}b_{12} + a_{12}b_{22}
\end{array}
\]

\[
\begin{array}{c}
a_{21}b_{11} + a_{22}b_{21} \\
a_{21}b_{12} + 2\sigma a_{22}b_{22}
\end{array}
\]

does it help if can compute this in 6 multiplications?
Weighted vs. unweighted MM

• s-rank can be much smaller than rank:

\[
\begin{array}{cccc}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 \\
\end{array}
\]

same support:

\[
\begin{array}{cccc}
0 & 1 & 2 & 3 \\
3 & 0 & 1 & 2 \\
2 & 3 & 0 & 1 \\
1 & 2 & 3 & 0 \\
\end{array}
\]

\(\mathbb{R} = \text{n-th root of unity}\)

\[
\begin{array}{cccc}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
\end{array}
\]

maybe it’s easy to show s-rank of \(n \leq n\) matrix multiplication is \(n^2\) (!!)
Weighted vs. unweighted MM

\[ ! = \inf \{ \omega : \text{rank}(\langle n,n,n \rangle) \cdot O(n^{\omega}) \} \]
\[ !_s = \inf \{ \omega : s\text{-rank}(\langle n,n,n \rangle) \cdot O(n^{\omega}) \} \]

**Theorem:** \[ ! \cdot \frac{(3!_s - 2)}{2} \]

in particular, \[ !_s \cdot 2 + \frac{3}{2} \]

\[ ! \cdot 2 + \frac{(3/2)^2} \]

- **Proof idea:**
  - find \( \frac{1}{4} n^2 \) copies of \( \langle n,n,n \rangle \) in 3\(^{rd}\) tensor power
  - when broken up this way, can rescale
A promising family of semisimple algebras
Coherent configurations

“group theory without groups”

• points $X$, partition $R_1, R_2, \ldots, R_r$ of $X^2$
  – diagonal $\{(x,x) : x \in X\}$ is the union of some classes
  – for each $i$, there is $i^*$ such that $R_i^* = \{(y,x) : (x,y) \in R_i\}$
  – exist integers $p_{i,j}^k$ such that for all $(x,y) \in R_k$:
    $\#\{z : (x,z) \in R_i$ and $(z,y) \in R_j\} = p_{i,j}^k$

  if one class: “association scheme”
  $p_{i,j}^k = p_{j,i}^k$ : commutative
Coherent configs: examples

- **Hamming scheme:**
  - points 0/1 vectors
  - classes determined by hamming distance

- **distance-regular graph:**
  - points = vertices
  - classes determined by distance in graph metric
Coherent configs: examples

• scheme based on finite group $G$
  – set $X = \text{finite group } G$
  – classes $R_g = \{(x,xg) : x \in X\}$

$\begin{align*}
p_{f,g}^h &= 1 \text{ if } fg = h, \ 0 \text{ otherwise} 
\end{align*}$

• “Schurian”:
  – group $G$ acts on set $X$
  – classes = orbits of (diagonal) $G$-action on $X^2$
Coherent configs: examples

• “Schurian”:
  – group \( G \) acts on set \( X \)
  – classes = orbits of (diagonal) \( G \)-action on \( X^2 \)

• one Schurian scheme: “group scheme”
  – group \( G \times G \) acts on \( G \) via \((g,h) \cdot x = gxh^{-1}\)
  – orbits all of the form \(\{(x,y): xy^{-1} \in 2 \ C_i\}\) for conjugacy class \( C_i \)
  – always commutative!
Adjacency algebra

CC: points $X$, partition $R_1$, $R_2$, …, $R_r$ of $X^2$

- for each class $R_i$, matrix $A_i$ with
  \[ A_i[x,y] = 1 \text{ iff } (x,y) \in R_i \]

- 3 CC axioms )
  \[ \{A_i\} \text{ generate a semisimple algebra} \]
  - e.g., 3\text{rd} axiom implies $A_i A_j = \sum_k p_{ij}^k A_k$
  - if the CC based on group $G$, algebra is $C[G]$
Nice basis conditions

• group algebra $\mathbb{C}[G]$: “nice basis” yields triple product property

• adjacency algebras of CCs: “nice basis” yields triangle condition:

$$\circ(k,i') \oplus (i,j') \ominus (j,k')$$

can look like

iff $i = i', j = j', k = k'$

Sept. 17, 2014
Nice basis conditions

- Schurian CCs: “nice basis” yields
  - group $G$ acts on set $X$
  - subsets $A, B, C$ of $X$ realize $\langle |A|, |B|, |C| \rangle$ if:

\[
\begin{align*}
&fgh = 1 \implies a = a', b = b', c = c' \\
&\text{Diagram:}
\end{align*}
\]
Coherent configs vs. groups

Generalization for generalization’s sake?

• recall group framework:
  – non-commutative necessary

**Theorem:** in group G realizing n£n matrix multiplication, with character degrees \(d_1, d_2, d_3, \ldots\), we obtain:

\[
R(<n,n,n>) \cdot \sum_i d_i^\omega \cdot d_{\text{max}}^{\omega-2} \leq |G|
\]

goals: \(|G| \leq \frac{1}{4} n^2 and small d_{\text{max}}\)
Coherent configs vs. groups

Generalization for generalization’s sake?

• coherent configuration framework:
  – commutative suffices!
  – combinatorial constructions from old setting yield
    \( !_s < 2.48, !_s < 2.41 \)
  – conjectures from old setting (if true) would imply \( !_s = 2 \)

in commutative Schurian CC’s even group schemes even symmetric

Sept. 17, 2014
Proof idea

we prove a general transformation:

if can realize several independent matrix multiplications in CC…

• can do this in abelian groups
• conjectures: can “pack optimally”

… then high symmetric power of CC realizes single matrix multiplication

– reproves Schönhage’s Asymptotic Sum Inequality
Commutative CCs suffice

Main point

embedding $n \times n$ matrix multiplication into a commutative coherent configuration of rank $\frac{1}{4} n^2$ is a viable route to $! = 2$

(no representation theory needed)
Open problems

• find a construction in new framework that
  – proves non-trivial bound on $s$
  – is not based on constructions from old setting

• is the (border) s-rank of $<2,2,2>$ = 6?

• embed $n \leq n$ MM into commutative coherent configuration of rank $\frac{1}{4} n^2$