NEW LOWER BOUNDS FOR THE RANK OF MATRIX MULTIPLICATION

J.M. LANDSBERG

ABSTRACT. The rank of the matrix multiplication operator for \( n \times n \) matrices is one of the most studied quantities in algebraic complexity theory. I prove that the rank is at least \( 3n^2 - o(n^2) \). More precisely, for any integer \( p \leq n-1 \) the rank is at least \( (3 - \frac{1}{p+1})n^2 - (1 + 2p(\frac{2p}{p+1}))n \). The previous lower bound, due to Bläser \([2]\), was \( \frac{3}{2}n^2 - 3n \) (the case \( p = 1 \)). The new bounds improve Bläser’s bound for all \( n > 84 \). I also prove lower bounds for rectangular matrices significantly better than the the previous bound.

1. INTRODUCTION

Let \( X = (x_{ij}^p) \), \( Y = (y_{ij}^p) \) be \( n \times n \)-matrices with indeterminant entries. The rank of matrix multiplication, denoted \( \text{R}(M_{n,n,n}) \), is the smallest number \( r \) of products \( p_\rho = u_\rho(X)v_\rho(Y), 1 \leq \rho \leq r \), where \( u_\rho, v_\rho \) are linear forms, such that the entries of the matrix product \( XY^t \) are contained in the linear span of the \( p_\rho \). This quantity is also called the bilinear complexity of \( n \times n \) matrix multiplication. More generally, one may define the rank \( \text{R}(b) \) of any bilinear map \( b \), see \( \S 2 \).

From the point of view of geometry, rank is badly behaved as it is not semi-continuous. Geometers usually prefer to work with the border rank of a bilinear map \( b \), denoted \( \text{R}^b(b) \), which fixes the semi-continuity problem by fiat: the border rank of a bilinear map \( b \), denoted \( \text{R}^b(b) \), is the smallest \( r \) such that \( b \) can be approximated to arbitrary precision by bilinear maps of rank \( r \). By definition, one has \( \text{R}(b) \geq \text{R}^b(b) \). A more formal definition is given in \( \S 2 \).

Let \( M_{n,m,1} \) denote the multiplication of an \( n \times m \) matrix by an \( m \times 1 \) matrix. In \([5]\) G. Ottaviani and I gave new lower bounds for the border rank of matrix multiplication, namely, for all \( p \leq n-1 \), \( \text{R}^b(M_{n,n,m}) \geq \frac{2p+1}{p+1}nm \). Taking \( p = n-1 \) gives the bound \( \text{R}(M_{n,n,m}) \geq 2nm - m \). In this article it will be advantageous to work with a smaller value of \( p \). The results of \([5]\) are used here to prove:

Theorem 1.1. Let \( p \leq n-1 \) be a natural number. Then

\[
\text{R}(M_{n,n,m}) \geq \frac{2p+1}{p+1}nm + n^2 - (1 + 2p(\frac{2p}{p+1}))n.
\]

The previous bound, due to Bläser \([2]\), was \( \text{R}(M_{n,n,m}) \geq 2nm - m + 2n - 2 \). For square matrices Theorem 1.1 specializes to:

Theorem 1.2. Let \( p \leq n-1 \) be a natural number. Then

\[
\text{R}(M_{n,n,n}) \geq (3 - \frac{1}{p+1})n^2 - (1 + 2p(\frac{2p}{p+1}))n.
\]

In particular, \( \text{R}(M_{n,n,n}) \geq 3n^2 - o(n^2) \).

The “in particular” follows by setting e.g., \( p = \lfloor \sqrt{\log(n)} \rfloor \).

Key words and phrases. rank, matrix multiplication, MSC 68Q17.
supported by NSF grant DMS-1006353.
This improves Bläser’s bound [1] of $\frac{5}{8}n^2 - 3n$ (the case $p = 1$) for all $n > 84$. Working under my direction, Alex Massarenti and Emanuele Raviolo [8, 7] improved the error term in Theorem 1.2. In a preprint of this article I made a mistake in computing the error term. Unfortunately this mistake was not noticed before Massarenti and Raviolo’s paper [8] was published, repeating the error, although their contribution is completely correct and their correct bound will appear in [7].

Remark 1.3. If $T$ is a tensor of border rank $r$, where the approximating curve of rank $r$ tensors limits in such a way that $q$ derivatives of the curve are used, then the rank of $T$ is at most $(2q-1)r$, see [3, Prop. 15.26]. In [6] they give explicit, but very large upper bounds on the order of approximation $h$ needed to write a tensor of border rank $r$ as lying in the $h$-jet of a curve of tensors of rank $r$.

The language of tensors will be used throughout. In §2 the language of tensors is introduced and previous work of Bläser and others is rephrased in a language suitable for generalizations. In §3 I describe the equations of [5] and give a very easy proof of a slightly weaker result than Theorem 1.1. In §4 I express the equations in coordinates and prove Theorem 1.1. I work over the complex numbers throughout.

Acknowledgement. I thank the anonymous referee for useful suggestions and C. Ikenmeyer for help with the exposition.

2. Ranks and border ranks of tensors

Let $A, B, C$ be vector spaces, of dimensions $a, b, c$ and with dual spaces $A^*, B^*, C^*$. That is, $A^*$ is the space of linear maps $A \to \mathbb{C}$. Write $A^* \otimes B$ for the space of linear maps $A \to B$ and $A^* \otimes B^* \otimes C$ for the space of bilinear maps $A \times B \to C$. To avoid extra $\ast$-s, I work with bilinear maps $A^* \times B^* \to C$, i.e., elements of $A \otimes B \otimes C$. Let $T : A^* \times B^* \to C$ be a bilinear map. One may also consider $T$ as a linear map $T : A^* \to B \otimes C$ (and similarly with the roles of $A, B, C$ exchanged), or as a trilinear map $A^* \times B^* \times C^* \to \mathbb{C}$.

The rank of a bilinear map $T : A^* \times B^* \to C$, denoted $\operatorname{R}(T)$, is the smallest $r$ such that there exist $a_1, \cdots, a_r \in A, b_1, \cdots, b_r \in B, c_1, \cdots, c_r \in C$ such that $T(\alpha, \beta) = \sum_{i=1}^r a_i(\alpha)b_i(\beta)c_i$ for all $\alpha \in A^*$ and $\beta \in B^*$. The border rank of $T$, denoted $\overline{\operatorname{R}}(T)$, is the smallest $r$ such that $T$ may be written as a limit of a sequence of rank $r$ tensors. Since the set of tensors of border rank at most $r$ is closed, one can use polynomials to obtain lower bounds on border rank. That is, let $P$ be a polynomial on $A \otimes B \otimes C$ such that $P$ vanishes on all tensors of border rank at most $r$: if $T \in A \otimes B \otimes C$ is such that $P(T) \neq 0$, then $\overline{\operatorname{R}}(T) > r$.

The following proposition is a rephrasing of part of the proof in [1]:

Proposition 2.1. Let $P$ be a polynomial of degree $d$ on $A \otimes B \otimes C$ such that $P(T) \neq 0$ implies $\overline{\operatorname{R}}(T) > r$. Let $T \in A \otimes B \otimes C$ be a tensor such that $P(T) \neq 0$ and $T : A^* \to B \otimes C$ is injective. Then $\overline{\operatorname{R}}(T) \geq r + a - d$.

As stated, the proposition is useless, as the degrees of polynomials vanishing on all tensors of border rank at most $r$ are greater than $r$. (A general tensor of border rank $r$ also has rank $r$.) However the conclusion still holds if one can find, for a given tensor $T$, a polynomial, or collection of polynomials on smaller spaces, such that the nonvanishing of $P$ on $T$ is equivalent to the non-vanishing of the new polynomials. Then one substitutes the smaller degree into the statement to obtain the nontrivial lower bound.

In our situation, first I will show $P(M_{(n,n,m)}) \neq 0$ if and only if $\tilde{P} (\tilde{M}) \neq 0$ where $\tilde{M}$ is a tensor in a smaller space of tensors and $\tilde{P}$ is a polynomial of lower degree than $P$, see (2). More precisely, note that in the course of the proof, $B \otimes C$ does not play a role, and we will see that the
relevant polynomial, when applied to matrix multiplication $M \in A \otimes B \otimes C = A \otimes \mathbb{C}^n \otimes \mathbb{C}^n$, will not vanish if and only if a polynomial $\tilde{P}$ applied to $\tilde{M} \in A \otimes \mathbb{C}^n \otimes \mathbb{C}^n$ with $\deg(\tilde{P}) = \deg(P)/n$, does not vanish, so the proof below works in this case. Then, in §4, I show that $\tilde{P}(\tilde{M}) \neq 0$ is implied by the non-vanishing of two polynomials of even smaller degrees.

This is why both Bläser’s result and the result of this paper improve the bound of border rank by $\mathbb{C}$, times to see that the pulled back polynomial is not identically zero restricted to $P = \mathbb{C}$.

Lemma 2.3.

Let $\mathbb{C}^a$ be given a basis. Given a polynomial $P$ of degree $d$ on $\mathbb{C}^a$, there exists a set of at most $d$ basis vectors such that $P$ restricted to their span is not identically zero.

The lemma follows by simply choosing a monomial that appears in $P$, as it can involve at most $d$ basis vectors.

Proof of Proposition 2.1.

Let $R(T) = r$ and assume we have written $T$ as a sum of $r$ rank one tensors. Since $T : A^* \otimes B \otimes C$ is injective we may write $T = T' + T''$ with $R(T') = a$, $R(T'') = r - a$ and $T' : A^* \otimes B \otimes C$ injective. Now consider the $a$ elements of $A \otimes B \otimes C$ appearing in $T'$. Since they are linearly independent, by Lemma 2.2 we may choose a subset of $d$ of them such that $P$, evaluated on the sum of terms in $T'$ whose $A$ terms are in the span of these $d$ elements, is not identically zero. Let $T_1$ denote the sum of the terms in $T'$ not involving the (at most) $d$ basis vectors needed for nonvanishing, so $R(T_1) \geq a - d$. Let $T_2 = T - T_1 + T''$. Now $R(T_2) \geq r$ because $P(T_2) \neq 0$. Finally $R(T) = R(T_1) + R(T_2)$.

Let $G(k, V) \subset \mathbb{P} \Lambda^k V$ denote the Grassmannian of $k$-planes through the origin in $V$ in its Plücker embedding. That is, if a $k$-plane is spanned by $v_1, \ldots, v_k$, we write it as $[v_1 \wedge \cdots \wedge v_k]$. One says a function on $G(k, V)$ is a polynomial of degree $d$ if, as a function in the Plücker coordinates, it is a degree $d$ polynomial. The Plücker coordinates $(x^\alpha_\mu)_k$, $k + 1 \leq \mu \leq \dim V = v$, $1 \leq \alpha \leq k$ are obtained by choosing a basis $\epsilon_1, \ldots, \epsilon_v$ of $V$, centering the coordinates at $[\epsilon_1 \wedge \cdots \wedge \epsilon_k]$, and writing a nearby $k$-plane as $[\epsilon_1 + \sum x^i_\mu \epsilon_\mu] \wedge \cdots \wedge [\epsilon_k + \sum x^i_\mu \epsilon_\mu]$. If the polynomial is also homogeneous in the $x^i_\mu$, this is equivalent to it being the restriction of a homogeneous degree $d$ polynomial on $\Lambda^k V$. (The ambiguity of the scale does not matter as we are only concerned with its vanishing.)

Lemma 2.3.

Let $A$ be given a basis. Given a homogeneous polynomial of degree $d$ on the Grassmannian $G(k, A)$, there exists at least $dk$ basis vectors such that, denoting their (at most) $dk$-dimensional span by $A'$, $P$ restricted to $G(k, A')$ is not identically zero.

Proof. Consider the map $f : A^k \rightarrow G(k, A)$ given by $(a_1, \ldots, a_k) \mapsto [a_1 \wedge \cdots \wedge a_k]$. Then $f$ is surjective. Take the polynomial $P$ and pull it back by $f$. (The pullback $f^*(P)$ is defined by $f^*(P)(a_1, \ldots, a_k) := P(f(a_1, \ldots, a_k))$.) The pullback is of degree $d$ in each copy of $A$. (I.e., fixing $k - 1$ parameters, it becomes a degree $d$ polynomial in the $k$-th.) Now simply apply Lemma 2.2 $k$ times to see that the pulled back polynomial is not identically zero restricted to $A'$, and thus $P$ restricted to $G(k, A')$ is not identically zero.

Remark 2.4. The bound in Lemma 2.3 is sharp, as give $A$ a basis $a_1, \ldots, a_n$ and consider the polynomial on $\Lambda^k A$ with coordinates $x^I = x^{i_1} \cdots x^{i_k}$ corresponding to the vector $\sum_I x^I a_{i_1} \wedge \cdots \wedge a_{i_k}$; $P = x^{i_1} \cdots x^{i_{k+1}} \cdots x^{i_{(d-1)k+1}} \cdots x^{i_{dk}}$. Then $P$ restricted to $G(k, (a_1, \ldots, a_{dk}))$ is non-vanishing but there is no smaller subspace spanned by basis vectors on which it is non-vanishing.
3. Matrix multiplication and its rank

Let \( M_{(m,n,l)} : \text{Mat}_{\mathbb{C}^m} \times \text{Mat}_{\mathbb{C}^n} \rightarrow \text{Mat}_{\mathbb{C}^l} \) denote the matrix multiplication operator. Write \( M = \mathbb{C}^m \), \( N = \mathbb{C}^n \) and \( L = \mathbb{C}^l \). Then

\[
M_{(m,n,l)} : (N \otimes L^*) \times (L \otimes M^*) \rightarrow N \otimes M^*
\]

has the interpretation as \( M_{(m,n,l)} = Id_N \otimes Id_M \otimes Id_L \in (N^* \otimes L^*) \otimes (L^* \otimes M^*) \), where \( Id_X \in N^* \otimes N \) is the identity map. If one thinks of \( M_{(m,n,l)} \) as a trilinear map \((N \otimes L^*) \times (L \otimes M^*) \rightarrow (N \otimes M^*) \rightarrow \mathbb{C} \), in bases it is \((X,Y,Z) \mapsto \text{trace}(XYZ)\). If one thinks of \( M_{(m,n,l)} \) as a linear map \( N \otimes L^* \rightarrow (L^* \otimes M^*) \otimes (N \otimes M^*) \) it is just the identity map tensored with \( Id_M \). In particular, if \( \alpha \in N \otimes L^* \) is of rank \( q \), its image, considered as a linear map \( L \otimes M^* \rightarrow N \otimes M^* \), is of rank \( qm \).

Returning to general tensors \( T \in A \otimes B \otimes C \), from now on assume \( b = c \). When \( T = M_{(m,n,l)} \), one has \( A = N^* \otimes L \), \( B = L^* \otimes M \), \( C = N \otimes M^* \), so \( b = c \) is equivalent to \( l = n \).

The equations of [5] are as follows: given \( T \in A \otimes B \otimes C \), with \( b = c \), take \( A' \subset A \) of dimension \( 2p + 1 \leq a \). Define a linear map

\[
T_{A'}^{A^p} : \Lambda^p A' \otimes B^* \rightarrow \Lambda^{p+1} A' \otimes C
\]

by first considering \( T|_{A' \otimes B \otimes C} : B^* \rightarrow A' \otimes C \) tensored with the identity map on \( \Lambda^p A' \), which is a map \( \Lambda^p A' \otimes B^* \rightarrow \Lambda^p A' \otimes B \otimes C \), and then projecting the image to \( \Lambda^{p+1} A' \otimes C \). Then if the determinant of this linear map is nonzero, the border rank of \( T \) is at least \( \frac{2p+1}{p+1} b \). If there exists an \( A' \) such that the determinant is nonzero, we may think of the determinant as a nontrivial homogeneous polynomial of degree \( \left( \frac{2p+1}{p} \right) b \) on \( G(2p+1, A) \).

Now consider the case \( T = M_{(n,n,m)} \), and recall that \( B = L^* \otimes M \), \( C = N \otimes M^* \). The map \((M_{(n,n,m)})_{A'}^{A^p} : \Lambda^p A' \otimes L \otimes M^* \rightarrow \Lambda^{p+1} A' \otimes N \otimes M^* \) is actually a reduced map

\[
\tilde{M}_{A'}^{A^p} : \Lambda^p A' \otimes L \rightarrow \Lambda^{p+1} A' \otimes N
\]

tensored with the identity map \( M^* \rightarrow M^* \), and thus its determinant is non-vanishing if and only if the determinant of \( \tilde{M}_{A'}^{A^p} \) is nonvanishing. But this is a polynomial of degree \( \left( \frac{2p+1}{p} \right) b \ll \left( \frac{2p+1}{p} \right) n^2 \) on \( G(2p+1, n^2) \). Proposition 2.1 with \( d = \left( \frac{2p+1}{p} \right) n \), \( a = n^2 \) and \( r = \frac{2p+1}{p+1} \cdot mnn \), combined with Lemma 2.3 gives the bound

\[
R(M_{(n,n,m)}) \geq \frac{2p+1}{p+1} mn + n^2 - (2p+1) \left( \frac{2p+1}{p} \right) n.
\]

Note that this already gives the \( 3n^2 - o(n^2) \) asymptotic lower bound. The remainder of the paper is dedicated to improving the error term.


Let \( a = 3 \) (so \( p = 1 \)) and \( b = c \), the map (1) expressed in bases is a \( 3b \times 3b \) matrix. If \( a_0, a_1, a_2 \) is a basis of \( A \) and one chooses bases of \( B, C \), then elements of \( B \otimes C \) may be written as matrices, and \( T = a_0 \otimes X_0 + a_1 \otimes X_1 + a_2 \otimes X_2 \), where \( X_j \) are size \( b \) square matrices. Order the basis of \( A \) by \( a_0, a_1, a_2 \) and of \( \Lambda^2 A \) by \( a_1 \wedge a_2, a_0 \wedge a_1, a_0 \wedge a_2 \). We compute

\[
T_{A'}^{(a_0 \otimes \beta)} = \beta(X_0) \otimes a_0 \wedge a_0 + \beta(X_1) \otimes a_1 \wedge a_0 + \beta(X_2) \otimes a_2 \wedge a_0 = -\beta(X_1) \otimes a_0 \wedge a_1 - \beta(X_2) \otimes a_0 \wedge a_2,
\]

\[
T_{A'}^{(a_1 \otimes \beta)} = \beta(X_0) \otimes a_0 \wedge a_1 - \beta(X_2) \otimes a_1 \wedge a_2,
\]

\[
T_{A'}^{(a_2 \otimes \beta)} = \beta(X_0) \otimes a_0 \wedge a_2 + \beta(X_1) \otimes a_1 \wedge a_2,
\]
so the corresponding matrix for $T_A^{\Lambda^1}$ is the block matrix

$$
\text{Mat}(T_A^{\Lambda^1}) = \begin{pmatrix}
0 & -X_2 & X_1 \\
-X_1 & X_0 & 0 \\
-X_2 & 0 & X_0
\end{pmatrix}.
$$

Now assume $X_0$ is invertible and change bases such that it is the identity matrix. Recall the formula for block matrices

$$
\det \begin{pmatrix} X & Y \\ Z & W \end{pmatrix} = \det(W) \det(X - YW^{-1}Z),
$$

assuming $W$ is invertible. Then, using the $(b, 2b) \times (b, 2b)$ blocking (so $X = 0$ in (3))

$$
\det \text{Mat}(T_A^{\Lambda^1}) = \det(X_1X_2 - X_2X_1) = \det([X_1, X_2]).
$$

When $\dim A > 3$, if there exists a three dimensional subspace $A'$ of $A$, such that $\det \text{Mat}(T_A^{\Lambda^1}) = 0$, then $R(T) \geq \frac{3}{2}b$ as this is (1) in the case $p = 1$. These are Strassen’s equations [9].

I now phrase the equations of [5] in coordinates. Let $\dim A = 2p + 1$. Write $T = a_0 \otimes X_0 + \cdots + a_{2p} \otimes X_{2p}$. The expression of (1) in bases is as follows: write $a_I := a_{i_1} \land \cdots \land a_{i_p}$ for $\Lambda^p A$, require that the first $(2^p_{p+1})$ basis vectors have $i_1 = 0$, that the second $(2^p_p)$ do not, and call these multi-indices $0J$ and $K$. Order the bases of $\Lambda^{p+1}A$ such that the first $(2^p_{p+1})$ multi-indices do not have $0$, and the second $(2^p_p)$ do, and furthermore that the second set of indices is ordered the same way as $K$, only we write $0K$ since a zero index is included. Then the resulting matrix is of the form

$$
\begin{pmatrix} 0 & Q \\ \breve{Q} & R \end{pmatrix}
$$

where this matrix is blocked $((\binom{2p}{p+1})b, (\binom{2p}{p})b) \times ((\binom{2p}{p+1})b, (\binom{2p}{p})b)$,

$$
R = \begin{pmatrix} X_0 & \cdots \\ \vdots & \ddots & \vdots \\ X_0 & & X_0
\end{pmatrix},
$$

and $Q, \breve{Q}$ have entries in blocks consisting of $X_1, \ldots, X_{2p}$ and zero. Thus if $X_0$ is the identity matrix, so is $R$ and the determinant equals the determinant of $QQ$. If $X_0$ is the identity matrix, when $p = 1$ we have $QQ = [X_1, X_2]$ and when $p = 2$

$$
QQ = \begin{pmatrix}
0 & [X_1, X_2] & [X_1, X_3] & [X_1, X_4] \\
[X_2, X_1] & 0 & [X_2, X_3] & [X_2, X_4] \\
[X_3, X_1] & [X_3, X_2] & 0 & [X_3, X_4] \\
[X_4, X_1] & [X_4, X_2] & [X_4, X_3] & 0
\end{pmatrix}.
$$

In general, when $X_0$ is the identity matrix, $QQ$ is a block $((\binom{2p}{p+1})b \times (\binom{2p}{p})b)$ matrix whose block entries are either zero or commutators $[X_i, X_j]$.

To prove Theorem 1.1 we work with $\hat{M}_i^{\Lambda^p}$ of (2), so $b = n$. First apply Lemma 2.2 to choose $n$ basis vectors such that restricted to them $\det(X_0)$ is non-vanishing, and then we consider our polynomial $\det(QQ)$ as defined on $G(2p, (2p + 1)n^2 - 1)$, and apply Lemma 2.3, using $2p(\binom{2p}{p})n$ basis vectors to insure it is non-vanishing. Our error term is thus $n + 2p(\binom{2p}{p-1})n$, and the theorem follows.
Remark 4.1. In [8, 7], they show the matrix $Q \tilde{Q}$ can be made to have a nonzero determinant by a subtle combination of factoring and splitting it into a sum of two matrices that carries a lower cost than just taking its determinant.

References


7. Alex Massarenti and Emanuele Raviolo, *Erratum to: The rank of $n \times n$ matrix multiplication is at least $3n^2 - 2\sqrt{2}n^3/2 - 3n$*, to appear.

8. , *The rank of $n \times n$ matrix multiplication is at least $3n^2 - 2\sqrt{2}n^3/2 - 3n$*, Linear Algebra Appl. 438 (2013), no. 11, 4500–4509. MR 3034546


E-mail address: jml@math.tamu.edu