SIAM J. COMPUT.
Vol. 11, No. 3, August 1982

0097-5397/82/1103-0006 $01.00/0

RAPID MULTIPLICATION OF RECTANGULAR MATRICES*

D. COPPERSMITH†

Abstract. The number of essential multiplications required to multiply matrices of size $N \times N$ and $N \times N^{0.172}$ is bounded by $CN^2 \log^2 N$.

Key words. matrix multiplication, tensor rank, algebraic complexity

Introduction. Let Rank$\langle K, M, N \rangle$ denote the number of essential multiplications required to multiply a $K \times M$ matrix by an $M \times N$ matrix, i.e., the rank of the 3-dimensional tensor defining this matrix multiplication. We show here, by two different routes, the existence of a positive number α such that Rank $\langle N, N, N^\alpha \rangle \leqq CN^2 \log^2 N$.

THEOREM. *There is a positive constant* $\alpha = 2 \log 2/5 \log 5 = 0.17227$ *such that*

$$\text{Rank } (\langle N, N, N^\alpha \rangle) = O(N^2 (\log N)^2).$$

Remark. This agrees well with the trivial lower bound

$$\text{Rank } (\langle N, N, N^\alpha \rangle) \geqq N^2.$$

Proof. The proof may be done in two ways. Each relies on existing basic constructions (each due to Schönhage), and minor modifications to existing techniques for combining basic constructions (i.e., the exponential direct sum theorem, partial matrix multiplication and approximate algorithms). The modifications involve (1) selecting a binomial coefficient to maximize an "area" rather than a "volume" (which allows the agreement between upper and lower bounds) and (2) doing two arguments at once (e.g., the exponential direct sum theorem and approximate algorithms) which serves only to improve the "error bound" from N^ε to $C(\log N)^2$.

Proof version 1 (*partial matrix multiplication*). Begin with the following construction, due to Schönhage [3]:

$$\begin{aligned}
&(a_{11}+x^2a_{12})(b_{21}+x^2b_{11})(c_{11})+(a_{11}+x^2a_{13})(b_{31})(c_{11}-xc_{21})\\
&\quad+(a_{11}+x^2a_{22})(b_{21}-xb_{12})(c_{12})\\
&\quad+(a_{11}+x^2a_{23})(b_{31}+xb_{12})(c_{12}+xc_{21})-(a_{11})(b_{21}+b_{31})(c_{11}+c_{12})\\
&=x^2(a_{11}b_{11}c_{11}+a_{11}b_{12}c_{21}+a_{12}b_{21}c_{11}+a_{13}b_{31}c_{11}+a_{22}b_{21}c_{12}+a_{23}b_{31}c_{12})\\
&\quad+x^3P(a,b,c,x).
\end{aligned}$$

This construction performs an approximate evaluation of the partial matrix product

$$\text{trace}\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \end{pmatrix}\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & 0 \\ b_{31} & 0 \end{pmatrix}\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

with five multiplications. Notice that five is also the trivial lower bound for this matrix product, obtained by counting independent elements of the matrix A. It is this equality

* Received by the editors July 17, 1980, and in final form October 16, 1981.

† Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

of lower and upper bounds which will allow the tight agreement between lower and upper bounds in the theorem, namely $N^2 \leqq \text{Rank}\,\langle N, N, N^\alpha\rangle \leqq CN^2 \log^2 N$.

The matrix product $D = AB$ can be thought of as the sum, as j goes through 1 to 3, of the outer products $a_{*,j}b_{j,*}$; then the indicated trace is the sum, over i and k, of $d_{i,k}c_{k,i}$. Among these three indicated outer products, we have two of dimension (2, 1) and one of dimension (1, 2).

Now iterate (tensorize) this construction M times. The formation of the product of the new, large matrices A and B, now involves 3^M outer products, of which (for each m between 0 and M) we have $\binom{M}{m}2^m$ outer products of dimension $(2^m, 2^{M-m})$.

Fix values of M and m. Then, mimicking Schönhage's proof of partial matrix multiplication [3], we may create a constant matrix A' of dimension $(2^m, 2^M)$ and a matrix B' of dimension $(2^M, 2^{M-m})$, such that each maximal minor of A' or B' has nonvanishing determinant. (Here we require that our underlying field is large enough; the rationals will do.) Suppose we want to multiply matrices A'' of dimension $(2^m, \binom{M}{m}2^m)$ and B'' of dimension $(\binom{M}{m}2^m, 2^{M-m})$. We start with a 1-1 mapping of the columns of A'' onto those columns of A'' with exactly 2^m nonzero entries. For each such column of A, compute the inverse of the $(2^m, 2^m)$ minor of A' whose rows correspond to the nonzero entries in this column of A. Multiply this inverse by the appropriate column of A''. Fill in the rest of A with zeros. Then we have $A'' = A'A$. Similarly create B such that $BB' = B''$. These matrices are created without essential multiplications, since A' and B' are constant matrices of scalars, and scalar multiplications don't count in the rank of a matrix multiplication problem. Then finally, $A''B'' = (A'A)(BB') = A'(AB)B'$. The multiplication AB can be done in 5^M multiplications in the ring of polynomials in x, and the left multiplication by A' and the right multiplication by B' are again scalar multiplications which don't enter into the calculation of rank.

Thus we have, so far, that

$$\text{Border Rank}\,(\langle 2^m, \tbinom{M}{m}2^m, 2^{M-m}\rangle) \leqq 5^M,$$

where Border Rank is the number of essential x-polynomial multiplications required to perform a given matrix multiplication.

The new wrinkle in our proof lies in our choice of m. Rather than choose m to maximize the "volume" of the left-hand side (the product of all three dimensions), which would enable us to minimize the exponent β for symmetric matrix multiplication ($\text{rank}\,\langle N, N, N\rangle = O(N^\beta)$), we instead select m to maximize the "area" of the projection onto the first two dimensions. Namely, we choose $m = (4M/5)$ as that value of m which maximizes the product of (2^m) and $(\binom{M}{m}2^m)$. This product is just $\binom{M}{m}4^m$, which is a term in the binomial expansion of $(4+1)^M$; thus it is maximized when $m/M = 4/(4+1)$, and for that value of m we have $\binom{M}{m}4^m = 5^M M^{-1/2}K$ for some constant K, by Stirling's formula. (Naively, the largest term in the binomial expansion of $(4+1)^M$ must be at least $(4+1)^M/(M+1)$, and Stirling's formula just gives a tighter bound.)

This gives us that

$$\text{Border Rank}\left(\left\langle 2^{4M/5}, \binom{M}{4M/5}2^{4M/5}, 2^{M/5}\right\rangle\right) \leqq 5^M,$$

or

$$\text{Border Rank}\,(\langle 2^{4M/5}, K5^M 2^{-4M/5}M^{-1/2}, 2^{M/5}\rangle) \leqq 5^M.$$

Now do the same arguments, with first and second dimensions reversed, to get

$$\text{Border Rank } (\langle K5^M 2^{-4M/5} M^{-1/2}, 2^{4M/5}, 2^{M/5}\rangle) \leqq 5^M.$$

Multiply (tensorize) to get

$$\text{Border Rank } (\langle K5^M M^{-1/2}, K5^M M^{-1/2}, 2^{2M/5}\rangle) \leqq 5^{2M}.$$

Letting $N = K5^M M^{-1/2}$ we get

$$\text{Border Rank } (\langle N, N, N^\alpha\rangle) \leqq K' N^2 (\log N),$$

where $\alpha = 2 \log 2/5 \log 5 = 0.17227$, and K' is some constant.

As usual, each multiplication in the ring of polynomials in x can be done as a convolution, via Fourier transforms, and since the degree of the product polynomials does not exceed $8M$ (each basic construction entails polynomials of degree 4, and the degrees are additive in the $2M$ iterations, yielding a total overall degree of $8M$), each such polynomial multiplication involves only $8M+1$ essential multiplications. Combining these results, we have that

$$\text{Rank } (\langle N, N, N^\alpha\rangle) \leqq K'' N^2 (\log N)^2,$$

as desired.

Remarks. If we were more careful, we would probably get a bound of $K'' N^2 (\log N)^{3/2}$.

Schönhage's construction involves two parameters k and n, each of which must be an integer greater than 1. The present theorem goes through exactly for each choice of k and n, and the value of α so obtained is

$$\alpha = \frac{2 \log ((k-1)(n-1)+1)}{(kn+1) \log (kn+1)}.$$

The present theorem selects $k = n = 2$ to maximize α at $2 \log 2/5 \log 5$.

The technique of partial matrix multiplication is due to Schönhage [3], and is valid in more generality than presented here. Here we are specializing his results (particularly by choice of m) to make possible the agreement between upper and lower bounds in our theorem.

Proof version 2 (*via exponential direct sum theorem*). Begin with Schönhage's construction which performs two completely disjoint matrix multiplications, of sizes $\langle k, n, 1\rangle$ and $\langle 1, 1, (k-1)(n-1)\rangle$, in $kn+1$ multiplications over the ring of polynomials in x. The construction is similar to that given above for partial matrix multiplication, and will not be repeated. His specialization to $k = n = 4$ gives the exponent $2.54\cdots$ for symmetric matrix multiplication.

Note again that $kn+1$ is the best possible result for this case, since the number of independent variables in the first two dimensions is kn for the first matrix and 1 for the second, thus $kn+1$ in all. Again this is the fact which will allow the close agreement between upper and lower bounds in the theorem.

Represent this construction as

$$(*) \qquad \text{Border Rank } (\langle k, n, 1\rangle + \langle 1, 1, (k-1)(n-1)\rangle) = kn+1.$$

Again we fix values of k and n which will maximize the eventual value of α', namely $k = n = 3$, and go through the proof for these fixed values, bearing in mind that the proof works for the general values as well.

Then (*) becomes

$$\text{Border Rank } (\langle 3, 3, 1\rangle + \langle 1, 1, 4\rangle) = 10,$$

or in other words,

$$(\langle 3, 3, 1\rangle + \langle 1, 1, 4\rangle) \leftarrow 10\langle 1, 1, 1\rangle,$$

which can be tensorized by $\langle a, b, c\rangle$ to obtain

$$(**) \qquad (\langle 3a, 3b, c\rangle + \langle a, b, 4c\rangle) \leftarrow 10\langle a, b, c\rangle.$$

Here $\leftarrow$ means "homomorphic image by approximating algorithm", and summation of several matrices implies direct sum (the matrices are completely disjoint).

Suppose we are allowed to do M arbitrary multiplications in the ring of polynomials in x. That is, we have at our disposal $M\langle 1, 1, 1\rangle$. We wish to apply (**) as often as possible; thus we divide M into groups of 10, with possibly some left over, and from $M\langle 1, 1, 1\rangle$ we get at least $(M/10-1)\langle 3, 3, 1\rangle + (M/10-1)\langle 1, 1, 4\rangle$. That is, there are at least $(M/10-1)$ disjoint groups of 10 among the M multiplications we are allowed, and each group will yield a $\langle 3, 3, 1\rangle$ and a $\langle 1, 1, 4\rangle$, all disjoint. Apply (**) to the $(M/10-1)\langle 3, 3, 1\rangle$ to get at least $((M/10-1)/10-1)\langle 9, 9, 1\rangle + ((M/10-1)10-1)\langle 3, 3, 4\rangle$. Similarly applying (**) to $(M/10-1)\langle 1, 1, 4\rangle$ we get $((M/10-1)/10-1)\langle 3, 3, 4\rangle + ((M/10-1)/10-1)\langle 1, 1, 16\rangle$. Combining, and doing the implied divisions, we get at least $(M/100-1.1)\langle 9, 9, 1\rangle + (2M/100-2.2)\langle 3, 3, 4\rangle + (M/100-1.1)\langle 1, 1, 16\rangle$. Continue in like fashion. After k iterations we have at least

$$\sum_j \left(\binom{k}{j}\frac{M}{10^k} - 2.5\right)\langle 3^j, 3^j, 4^{k-j}\rangle,$$

as can be proved by induction.

Again nothing is new; this is just a proof of the exponential direct sum theorem [3]. But now we choose our j to maximize, again, the "area" of the resulting expression, i.e., its projection to the first two dimensions, rather than its volume. Indeed, choosing j to maximize the product of the binomial coefficient with the first two dimensions, that is, roughly,

$$\left(\binom{k}{j}\frac{M}{10^k}\right)(9^j),$$

we get the value of $j = 9k/10$. Fixing k we may choose M so that the factor $(\binom{k}{j}M/10^k)$ is just greater than 2.5; this will insure us that we will have at least one piece of size $\langle 3^j, 3^j, 4^{k-j}\rangle$. Thus we choose:

$$j = \frac{9k}{10}, \qquad M = 1 + \frac{2.5(10^k)}{\binom{k}{j}}.$$

With these choices, we get that

$$\text{Border Rank } (\langle 3^j, 3^j, 4^{k-j}\rangle) \leqq M,$$

or, in terms of k,

$$\text{Border Rank } (\langle 3^{9k/10}, 3^{9k/10}, 4^{k/10}\rangle) \leqq \frac{2.5(10^k)}{\binom{k}{9k/10}}.$$

Again we use Stirling's approximation to get that

$$\binom{k}{9k/10} \geqq C' 10^k 9^{-9k/10} k^{-1/2}.$$

Letting $N = 3^{9k/10}$ and substituting, we get

$$\text{Border Rank } (\langle N, N, N^{\alpha'} \rangle) < C'' N^2 (\log N)^{1/2}.$$

Here α' is $(2 \log 4)/(9 \log 9) = .1402$, or in general, $(2 \log ((k-1)(n-1)))/(kn \log (kn))$.

Again, we may replace Border Rank by Rank (i.e., eliminate the x-polynomials) at the price of a factor of N^{ε}. (The error bound is not as good as before, since we have no nice bound on the degrees of the x-polynomials.) Thus we get

$$\text{Rank } (\langle N, N, N^{\alpha'} \rangle) = O(N^{2+\varepsilon}).$$

Conclusion. We present, by two different routes, means by which matrix multiplication problems of size $\langle N, N, N^{\alpha} \rangle$ can be done with $N^{2+\varepsilon}$ operations, for numbers α strictly bounded away from 0. We do not see directly how this can be used to accelerate the symmetric matrix multiplication problem, but we present the result as interesting in its own right, and also with the hope that progress may be made in this new and different direction towards the solution of the symmetric matrix multiplication problem.

Perhaps one can find a way of arguing that if this theorem holds for some α between 0 and 1, then it must hold for a larger α, and that the sequence of α's so obtained would converge to 1. But I see no path which such a construction would take.

Another result which would be of interest, and which we cannot seem to obtain, would be the existence of a number $\alpha > 1$ such that Rank $(\langle N, N, N^{\alpha} \rangle) = O(N^{1+\alpha+\varepsilon})$. We can almost get this result, namely, by similar techniques, for each $\beta > 0$ there is an $\alpha > 1$ such that

$$\text{Rank } (\langle N, N, N^{\alpha} \rangle) = O(N^{\alpha+1+\beta} (\log N)^{3/2}).$$

A literature search shows that in 1976 Brockett and Dobkin obtained the related result

$$\text{Rank } (\langle N, N, \log N \rangle) = N^2 + o(N^2).$$

The present result is incomparable, in the sense that we do a larger problem and require a larger number of multiplications.

Note. More recent results by Coppersmith and Winograd [4] (this issue, pp. 472–492), combined with these techniques, all yield a better estimate of α: Rank $(\langle N, N, N^{\alpha} \rangle) = O(N^{2+\varepsilon})$ for $\alpha = 0.197$.

Acknowledgment. The present paper was inspired by a conversation with Victor Pan.

REFERENCES

[1] R. W. BROCKETT AND D. DOBKIN, *On the number of multiplications required for matrix multiplication*, this Journal, 5 (1976), pp. 624–628.

[2] V. YA. PAN, *New combinations of methods for the acceleration of matrix multiplication*, Comput. Math. with Appl., 7 (1981), pp. 73–125.

[3] A. SCHÖNHAGE, *Partial and total matrix multiplication*, this Journal, 10 (1981), pp. 434–455.

[4] D. COPPERSMITH AND S. WINOGRAD, *On the asymptotic complexity of matrix multiplication*, this Journal, this issue, pp. 472–492.