REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
24-08-2018	Final Report	14-Aug-2017 - 13-May-2018

4. TITLE AND SUBTITLE
Final Report: Toward an Algorithm for Fast Matrix Multiplication

5a. CONTRACT NUMBER
W911NF-17-1-0380

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Davis
Sponsored Programs
1850 Research Park Drive, Suite 300
Davis, CA 95618 -6153

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
71167-MA-II.3

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas Strohmer
UU	UU	UU	UU		19b. TELEPHONE NUMBER +15-307-5210

RPPR Final Report
as of 24-Aug-2018

Agency Code:

Proposal Number: 71167MAII **Agreement Number: W911NF-17-1-0380**
INVESTIGATOR(S):

Name: Ph.D. Thomas Strohmer
Email: strohmer@math.ucdavis.edu
Phone Number: +15307521071
Principal: Y

Organization: **University of California - Davis**
Address: Sponsored Programs, Davis, CA 956186153
Country: USA
DUNS Number: 047120084 EIN: 946036494
Report Date: 13-Aug-2018 Date Received: 24-Aug-2018
Final Report for Period Beginning 14-Aug-2017 and Ending 13-May-2018
Title: Toward an Algorithm for Fast Matrix Multiplication
Begin Performance Period: 14-Aug-2017 **End Performance Period:** 13-May-2018
Report Term: 0-Other
Submitted By: Ph.D. Thomas Strohmer Email: strohmer@math.ucdavis.edu
Phone: (+15) 307-521071
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 0 **STEM Participants:** 0

Major Goals: Please see uploaded report. Since the project is very mathematical, it contains formulas that are much easier to typeset and read in the pdf format of the uploaded report than in the ascii format available here.

Accomplishments: Please see uploaded report. Since the project is very mathematical, it contains formulas that are much easier to typeset and read in the pdf format of the uploaded report than in the ascii format available here.

Training Opportunities: Nothing to Report. Due to nature of this grant, no students were involved in this project.

Results Dissemination: Due to the nature of this project, the research is still ongoing and thus no papers have been finalized yet. One paper is in preparation right now.

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

Final Report
Towards an Algorithm for Fast Matrix Multiplication

Background:
The multiplication of two matrices is one of the most fundamental operations in linear algebra and in computational mathematics at large. The costs for the multiplication of two dense $n \times n$ matrices is $\mathcal{O}(n^3)$ operations. As such, it forms a bottle neck in many fields ranging from classical applications in scientific computing to recent areas such as Big Data. It is known that reducing the complexity for matrix multiplication would correspondingly also allow one to reduce the complexity of many other procedures in numerical mathematics, such as Gaussian elimination, LU decomposition, computing the determinant or the inverse of a matrix [1, 2]. Matrix multiplication is also used as a subroutine in a plethora of computational problems that, on the face of it, have nothing to do with matrices.

There are numerous algorithms for the fast multiplication of *structured* matrices (Toeplitz, Hankel, sparse, ...). However, in many areas and especially in applications related to Artificial Intelligence and to massive datasets, we have to deal with dense and essentially unstructured matrices. One breakthrough was Strassen's algorithm which can multiply two $n \times n$ matrices in $\mathcal{O}(n^{\log_2(7)}) = \mathcal{O}(n^{2.8074})$ operations [6]. Several variations and improvements of Strassen's algorithm have been proposed in recent years. The most well known is the Coppersmith-Winograd algorithm, which reduces the cost of matrix multiplication to $\mathcal{O}(n^{2.376})$ operations [5]. Yet, due to the large constant involved, the Coppersmith-Winograd algorithm (or any of its recent modifications) is hardly ever used in practice as the improvement only becomes somewhat noticeable for very large matrices. For $n < 10^6$ the current record is just about $\mathcal{O}(n^{2.78})$.

Goals:
The main goal of this STIR project was to study the validity of an hitherto unexplored idea towards carrying out fast matrix multiplication in $\mathcal{O}(n^2 \log n)$ operations. The significance of having such a fast algorithm for matrix multiplication at our disposal is dramatic. It would revolutionize much of computational mathematics, since it impacts a wide range of algorithms and applications, and is thus particularly relevant for very large scale computations. An $\mathcal{O}(n^2 \log n)$ algorithm for matrix multiplication could have an impact on science and technology of the 21st century that is comparable to the enormous impact of the Fast Fourier Transform (FFT) on the 20th century. The potential payoffs for aiding the Army to accomplish its mission are many, including new information technology capabilities, improved methods for signal- and image processing, as well as better data mining tools for massive, complex data sets.

Key ideas of proposed research:
In this project the PI considered a new approach to fast matrix multiplication based on a careful synthesis of group theoretic techniques proposed by Cohn and Umans [4, 3] with novel methods for specific *structured* matrix structures.

A main limitation of the group-theoretic approach is that it is not able to capture certain structural properties of those matrices arising as elements of the group algebra. To illustrate this issue via an example, assume that an irreducible representation leads to an $m \times m$ matrix M. Since this matrix M would not be further reducible by assumption, the computational cost of multiplying M with another $m \times m$ matrix in general would be of order m^3. However for the sake of the argument,

let us assume for the moment that M is, say, of the form $M = DC$, where D is a diagonal matrix and C is a circulant matrix[1]. Note that this structure is no contradiction to the assumption of irreducibility. Then this special property of M, since irreducible in the group-theoretic sense, would go unnoticed by the general group-theoretical approach and the cost from a purely group-theoretic viewpoint for multiplying two matrices of this form would be $\mathcal{O}(n^3)$. Yet, it is easy to see that one could multiply such an M with an arbitrary $m \times m$ matrix in $\mathcal{O}(n^2 \log n)$ operations. There are many other important examples of matrix structures that give rise to fast algorithms but would go unnoticed by a group-theoretic approach, such as e.g. Vandermonde matrices or Cauchy matrices.

There are several finite groups that could prove useful for our agena. One of the most promising one is the *finite Heisenberg group* $\mathbb{H}_n$, which is defined as follows: For $k, l \in \mathbb{C}^n$ we define the *translation operator* T, the *modulation operator* M and the *phase operator* Z by

$$T_k f(m) = f(m-k), \qquad M_l f(m) = e^{2\pi i m l/n} f(m), \qquad Z_s f(m) = e^{2\pi i s/n} f(m), \tag{1}$$

respectively, where translation is understood in a periodic sense. We have the commutation relations

$$M_l T_k = Z_{kl} T_k M_l, \tag{2}$$

hence T_k and M_l commute if only if $k \cdot l$ is a multiple of n. The Heisenberg group $\mathbb{H}_n$ is generated by the objects T_k, M_l, Z_r.

It is easy to verify that the matrices $\{T_k\}_{k=0}^{n-1}, \{M_l\}_{l=0}^{n-1}$ form an orthonormal basis for all $n \times n$ matrices. Proceeding from the group formed by (T_k, M_l, Z_s) to the set formed by (T_k, M_l) corresponds essentially to factoring out the center of $\mathbb{H}$. Hence we can express any matrix $A \in \mathbb{C}^{n \times n}$ as

$$A = \sum_k \sum_l a(k,l) T_k M_n,$$

for appropriate and uniquely defined coefficients $a(k,l)$. A few calculations reveal that the coefficients $a(k,l)$ can be computed via n FFTs of length n, i.e., with a total complexity of $\mathcal{O}(n^2 \log n)$ operations. We furthermore emphasize that there is a bijective correspondence between A and a.

Consider two $n \times n$ matrices A and B with symbols a and b, respectively. We can express the product of the two matrices AB as

$$\begin{aligned} AB = \sum_{k,l} a(k,l) T_k M_l \sum_{k',l'} b(k',l') T_{k'} M_{l'} &= \sum_{k,l} \sum_{k',l'} a(k,l) b(k',l') e^{2\pi i k l'/n} T_{(k+k')} M_{l+l'} \\ &= \sum_{m,j} \Big(\sum_{k,l} a(k,l) b(m-k, j-l) e^{2\pi i (m-k) l/n} \Big) T_l M_m. \end{aligned} \tag{3}$$

We define the *twisted convolution* $a \natural b$ of two two-dimensional arrays $\{a(k,l)\}, \{b(k',l')\} \in \mathbb{C}^{n \times n}$ by

$$(a \natural b)(m,j) = \sum_k \sum_l a(k,l) b(m-k, j-l) e^{2\pi i (m-k) l/n}, \tag{4}$$

[1]It is worth noting that circulant matrices form a group, but matrices that are products of diagonal matrices and circulant matrices do not.

where the indexing is understood in a periodic manner. Thus, with a change of variables we can rewrite (3) as

$$C := AB = \sum_{l,m} c(k,l) T_l M_m,$$

where $c = a \natural b$ is the symbol of C. Using (4) we can rewrite matrix-matrix multiplication of two *unstructured matrices* as *structured* matrix-vector multiplication of an $n^2 \times n^2$ matrix $\mathcal{B}$ with a vector α of length n^2, given by

$$\mathcal{B}\alpha, \tag{5}$$

where

$$\mathcal{B}_{k,l,k',l'} = b(k-k', l-l') e^{2\pi i (k-k')l'/n}. \tag{6}$$

In other words, computing the multiplication of the two $n \times n$ matrices A, B is equivalent to computing the matrix-vector product $\mathcal{B}\alpha$ of the $n^2 \times n^2$ matrix $\mathcal{B}$ with the length-n^2 vector α.

The structure of $\mathcal{B}$ is a mixture of block-circulant with block-Vandermonde. While fast algorithms are known for each individual structure, it is as of today not clear yet—and has been one of the key tasks of this project—how to develop a fast algorithm for the particular structure underlying $\mathcal{B}$ to enable fast matrix-vector multiplication. It is essential here to note that the size of $\mathcal{B}$ is $n^2 \times n^2$, thus we need an algorithm for matrix-vector multiplication that is almost linear in the matrix dimension, such as e.g., log-linear in n^2. Only then would this translate into an algorithm for multiplying A and B with a cost of $\mathcal{O}(n^2 \log n)$.

We stress again that the initial matrices A and B are arbitrary and not structured at all. The structure we discussed before emerges only by embedding the matrix multiplication into a larger space - albeit at the cost of increasing the dimension. The potential benefits are that we may gain enough structure to not just compensate for the increase in dimensions, but actually reduce the overall computational costs.

Results:
An essential insight of this project was that one can bring to bear recent breakthrough results by Lek-Heng Lim and collaborators [8]. In their seminal work, the authors investigated how the celebrated Cohn-Umans method may be used for bilinear operations such as *structured matrix-vector multiplication*, which is exactly the bilinear operation we are interested in. The authors further relate it to Strassen's tensor rank approach, the traditional framework for investigating bilinear complexity. The key point here is that the authors succeed in relating the bilinear complexity of (structured) matrix-vector multiplication to the border rank of the corresponding structure tensor. More precisely, the strucure tensor is defined as follows:

Definition 1. *Let $\beta : U \times V \to W$ be a bilinear map. Then there exists a unique tensor $\mu_\beta \in U^* \times V^*$ such that for any given $(u,v) \in U \times V$ we have*

$$\beta(u,v) = \mu_\beta(u,v,\cdot) \in W.$$

We call μ_β the structure tensor of the bilinear map β.

We already know, due to Strassen's Algorithm and its improvements, that the complexitiy of matrix-matrix multiplication is $\mathcal{O}(n^{2.8074})$. Hence, the cost of a structured matrix-vector multiplication of the form outlined in (5),(6) must also be at most $\mathcal{O}(n^{2.8074})$ (and hopefully less!). Thus,

a promising road map towards fast algorithms for matrix-matrix multiplication can now be formed by combining our approach in (5),(6) with the border rank techniques of Lim [8].

The next questions we need to address to pursue the main goal of this project are therefore:

1. Can we prove that the border rank of the structure tensor μ_β of matrix-vector multiplication is $\mathcal{O}(n^{2.8074})$ (or even lower)? And which algebra do we need to represent the matrix B in, so that such a proof becomes possible?

2. If we succeed in item (1), can we then further reduce the border rank to $\mathcal{O}(n^2 \log n)$? Which algebra do we need to choose for this purpose?

The main obstacle we face is that we need to find a proper algebra in which to express unstructured matrix-matrix multiplication as structured matrix-vector multiplication and determine its border rank. So far we have not succeeded and the mathematical tools currently available in bilinear algebra and not yet specialized and sophisticated enough for this purpose. Some potential algebras may be Clifford algebras and those related to extraspecial groups. We plan to investigate these in our future work.

A rather surprising insight of this project, that is not directly related to the main goal of this research but likely very useful, is the following: Our investigations of proper, efficient embeddings for fast matrix-matrix multiplication gave us the insight that somewhat related ideas may be useful in specific applications, where computational devices are rather limited in terms of computation power, memory, and battery supply. One such scenario arises when one tries to run deep learning algorithms for image classification on very small devices, such as very lightweight drones or small stand-alone devices as they may arise in the future Internet-of-Things.

By first projecting the measured data on a properly chosen subspace, one can dramatically reduce the size of the data, and thus the memory requirements, computational cost, and battery demands. The key here is that we can approximate the optimal projection provided by principal component analysis via certain structured matrices *and* that these matrices can be implemented in hardware. Thus, the projection can be integrated in the hardware of the sensor, thereby paving the way for marrying two cutting-edge techniques: compressive sensing and deep learning. We are currently investigating this very promising direction. Preliminary results will be reported in [7].

References

[1] Alfred V Aho and John E Hopcroft. *Design & Analysis of Computer Algorithms*. Pearson Education India, 1974.

[2] D.A. Bini. Fast matrix multiplication. In Leslie Hogben, editor, *Handbook of linear algebra*, chapter 47. CRC Press, 2006.

[3] Henry Cohn, Robert Kleinberg, Balazs Szegedy, and Christopher Umans. Group-theoretic algorithms for matrix multiplication. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 379–388. IEEE, 2005.

[4] Henry Cohn and Christopher Umans. A group-theoretic approach to fast matrix multiplication. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 438–449. IEEE, 2003.

[5] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 1–6. ACM, 1987.

[6] Volker Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 13(4):354–356, 1969.

[7] T. Strohmer, Y. Li, and D. Pinkney. Compressive Deep Learning and the Internet-of-Things. *Manuscript, in preparation*, 2018.

[8] Ke Ye and Lek-Heng Lim. Fast structured matrix computations: Tensor rank and Cohn–Umans method. *arXiv preprint arXiv:1601.00292*, 2016.