From *SIAM News*, Volume 38, Number 9, November 2005

Toward an Optimal Algorithm for Matrix Multiplication

By Sara Robinson

Ever since the dawn of the computer age, researchers have been trying to find an optimal way of multiplying matrices, a fundamental operation that is a bottleneck for many important algorithms. Faster matrix multiplication would give more efficient algorithms for many standard linear algebra problems, such as inverting matrices, solving systems of linear equations, and finding determinants. Even some basic graph algorithms run only as fast as matrix multiplication.

The standard method for multiplying $n \times n$ matrices requires $O(n^{3)}$ multiplications. The fastest known algorithm, devised in 1987 by Don Coppersmith and Shmuel Winograd, runs in $O(n^{2.38})$ time. Most researchers believe that an optimal algorithm will run in essentially $O(n^2)$ time, yet until recently no further progress was made in finding one.

Two years ago, a theoretical computer scientist and a mathematician came up with a promising new group theoretic approach to fast matrix multiplication. The researchers—Chris Umans of the California Institute of Technology and Henry Cohn of Microsoft Research—showed that if there are groups that simultaneously satisfy two conditions, then the group theoretic approach will yield nontrivial (< 3) upper bounds on the exponent for matrix multiplication.

Now, Umans and Cohn—together with Robert Kleinberg of the University of California at Berkeley and Balazs Szegedy of the Institute for Advanced Study in Princeton—have found a group that satisfies the two conditions. In a recent paper, the four collaborators describe a series of new matrix multiplication algorithms, the most efficient of which matches the time of the fastest known algorithm. The researchers also came up with two conjectures; a positive resolution to either would imply that the exponent for matrix multiplication is 2. "Our approach makes evident a connection between fast matrix multiplication and group theory, which potentially gives us more tools to work with," Kleinberg says. The researchers presented their results in October at the IEEE Symposium on the Foundations of Computer Science, in Pittsburgh.

From Strassen's Famous Algorithm to Group Theory

The exponent for matrix multiplication has proved so elusive that it has been given a name: ω. Formally, ω is the smallest number such that $O(n^{\omega + \varepsilon})$ multiplications suffice for all $\varepsilon > 0$. Because all n^2 entries must be part of any computation, ω is clearly at least 2.

The road to the current best upper bound on ω was built on clever ideas and increasingly complex combinatorial techniques. Many researchers long believed that the standard, $O(n^3)$ algorithm was the best possible; indeed, one research paper even presented a proof that the standard algorithm is optimal under seemingly reasonable as-sumptions. Then, in 1969, Volker Strassen stunned the research world with his $O(n^{2.81})$ algorithm for multiplying matrices. Still sometimes used in practice, Strassen's algorithm is reminiscent of a shortcut, first observed by Gauss, for multiplying complex numbers. Though finding the product $(a + bi)(c + di) = ac - bd + (bc + ad)i$ seems to require four multiplications, Gauss ob-served that it can actually be done with three—ac, bd, and $(a + b)(c + d)$—because $bc + ad = (a + b)(c + d) - ac - bd$.

In a similar way, Strassen's algorithm reduces the number of multiplications required for computing the product of two 2×2 matrices A and B from eight to seven, by expressing the entries of AB as linear combinations of products of linear combinations of the entries of A and B. This trick, applied to four blocks of $2^n \times 2^n$ matrices, reduces the problem to seven multiplications of $2^{n-1} \times 2^{n-1}$ matrices. Recursive application gives an algorithm that runs in $O(n^{\log_2 7}) = O(n^{2.81})$ time.

Following Strassen's work, a series of papers published through the 1970s and 80s gave increasingly better upper bounds on the exponent, culminating in Coppersmith and Winograd's $O(n^{2.38})$ algorithm. The approaches, which require increasingly sophisticated mathematics, are daunting reading. Yet, viewed at a high level, the methods all rely on the same framework: For some k, they provide a way to multiply $k \times k$ matrices using $m <<< k^3$ multiplications, and apply the technique recursively to show that $\omega < \log_k m$. The main challenge in devising these methods is in the design of the recursion for large k.

The new group theoretic approach follows this blueprint, with a basic theorem from group representation theory used to devise the recursion.

Finite Groups and the Discrete Fourier Transform

One way to understand the group theoretic approach is to view it as a generalization of a method for multiplying polynomials quickly using the fast Fourier transform.

The standard method for multiplying two degree-n polynomials requires $O(n^2)$ multiplications. The number of multiplications can be reduced by thinking of the coefficient vectors of the polynomials as elements of the group algebra $C[G]$ of a finite cyclic group G, i.e., all complex-valued vectors indexed by elements of G. If the group is large enough (of order at least $2n$), convolution of two vectors in the group algebra corresponds to the polynomial product.

Convolution in the group algebra can be computed quickly using the fast Fourier transform, an invertible linear map from $C[G]$ to $C^{|G|}$ that transforms convolution into pointwise multiplication. This transformation, and its inverse, can be computed in $O(n \log n)$ time for a cyclic group of size $2n$.

What Cohn and Umans realized is that under certain conditions, matrix multiplication, too, can be embedded into the group algebra of a finite group G, although the embedding is more complex and the group must be non-abelian to yield a nontrivial bound. Nonetheless, as long as the embedding satisfies a certain structural property, when the entries of the matrices A and B are represented by vectors a and b in the group algebra, the entries of the product matrix AB can be read off from the convolution vector $a * b$.

As with the polynomial method, the Fourier transform provides an efficient way to compute the convolution product. For a non-abelian group G, a fundamental theorem of Wedderburn says that the group algebra is isomorphic, via a Fourier transform, to an algebra of block diagonal matrices hav-ing block dimensions $d_1 \ldots d_k$, with $\Sigma d_i^2 = |G|$. Convolution in $C[G]$ is thus transformed into block diagonal matrix multiplication.

Using this framework, Cohn and Umans concocted a new paradigm for matrix multiplication: Two input matrices A and B are first represented as elements a,b of a group algebra $C[G]$. The Fourier transform is used to convert the elements a and b into block diagonal matrix form. Then, via the inverse Fourier transform, the product can be mapped back to the convolution $a * b$. From there, it is possible to read off the entries of AB. This process, which reduces a $k \times k$ matrix product to several smaller matrix products, forms the basis for a recursive algorithm.

As with polynomial multiplication, the approach works only if the group G admits an embedding of matrix multiplication into its group algebra so that the coefficients of the convolution product correspond to the entries of the product matrix. In their 2003 paper, the researchers showed that such an embedding is possible whenever the group has three subgroups, H_1, H_2, H_3, with the property that whenever $h_1 \in H_1$, $h_2 \in H_2$, and $h_3 \in H_3$ with $h_1h_2h_3 = 1$, then $h_1 = h_2 = h_3 = 1$. (The condition can be generalized to subsets of G rather than subgroups.) The researchers call this property of subgroups or subsets the "triple product property."

For the bound on the exponent for matrix multiplication given by this approach to be less than 3, the group must also satisfy conditions on the size of the subgroups or subsets relative to the dimensions of the blocks in the group's matrix algebra representation. In particular, it must be the case that $|H_1||H_2||H_3| > \Sigma d_i^3$—a goal the researchers began to call "beating the sum of the cubes." Cohn and Umans described this property in the 2003 paper, but they were not able to exhibit a group, together with three subsets or subgroups, that satisfied it.

Beating the Sum of the Cubes

For many months, Cohn and Umans tried to find a group with subgroups or subsets satisfying the triple product property that also beat the sum of the cubes. They even wrote a computer program that checked all groups up to order 127 and their subgroups, but found nothing. Faced with a seemingly insurmountable task, Cohn and Umans eventually enlisted reinforcements. Kleinberg, then an intern at Microsoft, had studied pure mathematics before shifting to theoretical com-puter science and seemed to be a natural fit for the problem. Szegedy, then a postdoc at Microsoft working at the intersection of com-binatorics and algebra, also joined the group.

In September 2003, on the eve of Rosh Hashanah, Umans and Cohn wrote in an e-mail message to Kleinberg that they had succeeded in showing that a group had the desired properties. Kleinberg, who does not work during the Jewish New Year, recalls "the agony of spending a two-day holiday knowing they had made this discovery but not being able to work through it."

The elusive group G turned out to be a wreath product of an abelian group of order 17^3 and the symmetric group of order 2. It is this group, and generalizations of it, that are the focus of the paper presented in October at FOCS 2005. Using the wreath product group in a straightforward way, the researchers achieved their first nontrivial bound (2.91) on the exponent for matrix multiplication. From there, they moved to other wreath products of abelian groups with the symmetric group S_n, and devised more sophisticated techniques for choosing the three subsets, giving a series of better bounds. The estimates of these bounds rely on an elementary fact of group representation theory: that the index of the largest abelian subgroup of a group is an upper bound on the size of the maximum block of the block diagonal matrix representation given by Wedderburn's theorem.

At some point, Szegedy realized that some of the combinatorial structures from the 1987 Coppersmith–Winograd paper could be used to select the three subsets in the wreath product groups in a more sophisticated way, thus improving the bounds. Applying these tools to the wreath product construction, the researchers were able to devise algorithms that match several earlier bounds on the exponent for matrix multiplication, including that of the fastest known algorithm. Given this mixing and matching of techniques, an obvious question is whether the group theoretic approach is really new or just a different way of viewing the earlier approaches. The researchers acknowledge that they do not yet fully understand the interconnections. Nevertheless, Umans says, the two approaches "seem different in crucial ways" and result in different algorithms.

Ultimately, the researchers distilled their insights into two conjectures, one that has an algebraic flavor and one that is purely combinatorial (see sidebar). (The conjectures apply only to the wreath product construction.) A proof of either conjecture would imply that $\omega = 2$. "It's easy to write down impossibly difficult combinatorial conjectures, but I don't think these conjectures fall into that category," Umans says. "I hope we prove $\omega = 2$, but part of putting these conjectures out is that we hope somebody else might, too."

"The new approach seems at least as promising as older approaches, not necessarily because it is more powerful, but because it is easier to understand what is needed to make it work," says Bjorn Poonen, a professor of mathematics at the University of California, Berkeley, who attended a three-hour seminar on the effort. "Experts in finite group theory and design theory should leap at the opportunity to make a contribution."

Even if someone manages to prove one of the conjectures—thereby demonstrating that $\omega = 2$—the wreath product approach is unlikely to be applicable to the large matrix problems that arise in practice. Although the wreath product algorithms perform better than Strassen asymptotically, Umans says, the input matrices must be astronomically large for the difference in time to be apparent. Still, the discovery of other classes of groups that satisfy the two conditions would provide new toolkits for developing fast matrix multiplication algorithms, possibly even practical ones.

"One of the things that has kept us enthusiastic about this method is that many things have worked out in a beautiful way," Umans says. "This makes it seem like the right approach."

Sara Robinson is a freelance writer based in Los Angeles, California.

The Combinatorial Conjecture

X	X	Y	Y	Y	Z
X	Y	X	Y	Z	Y
X	Y	Y	Y	Z	Z
Y	X	X	Z	Y	Y
Y	X	Y	Z	Y	Z
Y	Y	X	Z	Z	Y

		Y	Y	Y	

			Z		Z

A 6 × 6 strong USP, along with 2 of its 18 pieces.

In a 2005 FOCS paper, Henry Cohn, Robert Kleinberg, Balazs Szegedy, and Chris Umans describe two conjectures, one combinatorial and the other algebraic. A proof of either conjecture would imply that ω, the exponent for matrix multiplication, is 2.

The combinatorial conjecture concerns an object called a "uniquely solvable puzzle," or USP, which can be thought of as a jigsaw puzzle with pieces of three colors. As in the assembled puzzle shown here, each row is formed from three puzzle pieces, each made up of cells of a single color. The rule for assembly is that each row of the solution is composed of three overlaid pieces, one of each color. Like a jigsaw puzzle, a USP has the property that it can be put together in only one way without any overlapping of pieces. (Obviously, a USP can have no duplicate pieces.) A theorem from the 1987 paper of Coppersmith and Winograd implies that there exist USPs that are as large as possible, given the defining constraint.

A slightly more restricted version of a USP—known as a strong USP—has the property that every assembly of the puzzle pieces (other than the solution) includes at least one cell in which precisely two pieces overlap. The combinatorial conjecture, in essence, says that there are strong USPs that are as large as possible.

In their paper, the researchers describe how they systematically transform strong USPs into sets that satisfy the triple product property, with larger strong USPs producing better bounds on ω. If the conjecture is true, this method will yield ω = 2.

Details about both conjectures can be found in "Group-theoretic Algorithms for Matrix Multiplication," by Cohn, Kleinberg, Szegedy, and Umans, at http://www.cs.caltech.edu/~umans/papers/CKSU05.pdf.