BLAS IV: A BLAS for Rk Matrix Algebra

John Shaeffer
Matrix Compression Technologies, LLC
Marietta, Georgia
john@shaeffer.com

***Abstract*—Basic Linear Algebra Subroutines (BLAS) are well-known low-level workhorse subroutines for linear algebra vector-vector, matrix-vector and matrix-matrix operations for full rank matrices. The advent of block low rank (Rk) full wave direct solvers, where most blocks of the system matrix are Rk, an extension to the BLAS III matrix-matrix work horse routine is needed due to the agony of Rk addition. This note outlines the problem of BLAS III for Rk LU and solve operations and then outlines an alternative approach, which we will call BLAS IV. This approach utilizes the thrill of Rk matrix-matrix multiply and uses the Adaptive Cross Approximation (ACA) as a methodology to evaluate sums of Rk terms to circumvent the agony of low rank addition.**

Keywords—direct factor method of moments, low rank matrix algebra and electromagnetic scattering.

I. BACKGROUND

Full wave solvers for Maxwell's integral equations are the much-preferred approach when they can be implemented. And direct factor rather than iterative solutions avoids the well-known failures of the latter. However, the direct factor computational cost for N unknowns is immense: N^3 for matrix LU factorization and N^2 for each RHS solution.

The development of the Adaptive Cross Approximation (ACA) for computing the low rank **UV** approximation to system matrix blocks, based on spatial grouping of unknowns, has spawned whole new approaches to solving MOM system matrices. Included is the author's development in 2006 of the first MOM code (Mercury MOM) to LU factor a problem with one million unknowns on a PC computer [1].

Basic linear algebra computational routines are a set of low-level routines for performing common linear algebra operations for vector-vector, matrix-vector and matrix-matrix operations. These subroutines have been highly developed since the early 1990's and collectively are known as the BLAS. They are typically found in specialized numerical matrix libraries for each type of computer architecture where they have been highly optimized. For example, PC computers running Intel processors, the Intel Fortran and C compilers come with BLAS libraries optimized for their line of processors. BLAS routines come in three varieties: BLAS I for vector-vector with O(n) operations; BLAS II for matrix-vector with $O(n^2)$ operations; and BLAS III for matrix-matrix with $O(n^3)$ operations.

II. RK ALGEBRA

A low rank approximation to a (m x n) matrix **A** is the **U V** product of a column and row matrices, **A = U V**, where **U** is (m x k) and **V** is (k x n). Such an approximation to some tolerance ε is said to have rank k where k is usually << (m,n). Memory storage reduces from mn to k(m+n) with a rank fraction compression metric, defined as the ratio of low rank to full rank memory storage, RF = (k*(m+n)) / (mn). A tutorial of Rk matrix algebra is found in [2].

Multiplication of two Rk matrices has significant redeeming value in that the operations count is often reduced from $O(n^3)$ to $O(n^2)$, hence "the thrill of Rk multiplication" [2].

Sums of Rk matrices, however, have no redeeming value since the rank of the sum is the sum of the ranks of individual terms. There is no memory storage savings since the memory storage for the resulting sum is the same as that for all the individual terms of the sum, hence the term "the agony of Rk summation" [2].

The truncation algorithm found in [3] allows recompression of an Rk sum to SVD rank. This is made possible by repeated application of the truncation algorithm of the **UV** approximation which uses QR and SVD factorizations. However, this approach is not feasible for cases where many Rk sum terms must be recompressed.

III. BLAS IV

BLAS III for non Rk matrix-matrix operations is:

$$\mathbf{C} \leftarrow \alpha\, \mathbf{A}\, \mathbf{B} + \beta\, \mathbf{C}, \tag{1}$$

where **C**, **A** and **B** are full matrices and α and β are scalars. This has $O(n^3)$ operations and is usually the most optimized of the BLAS. The common name for this operation is "gemm".

With spatial grouping for electrically large problems (as characterized by tens of thousands to several million unknowns with group sizes from 500 to 10,000 unknowns) most all blocks in the system matrix, except for diagonal self-blocks, become Rk. This includes not only Z blocks but also its L and U factors. And for scattering problems with many RHS illumination angles, the RHS voltage excitation matrix is Rk as well as the current solution J and/or M.

To see the need for a BLAS IV operation for Rk matrices, we need to examine the block formulas for LU factorization and the forward/backward solve operation, [4]:

Submitted On: September 8, 2020
Accepted On: September 9, 2020

https://doi.org/10.47037/2020.ACES.J.351102

1054-4887 © ACES

$$\mathbf{U}_{iBlk,jBlk} = \mathbf{Z}_{iBlk,jBlk} - \sum_{pBlk=1}^{pBlk=iBlk-1} \mathbf{L}_{iBlk,pBlk}\, \mathbf{D}^{-1}_{pBlk,pBlk}\, \mathbf{U}_{pBlk,jBlk}$$

$$\mathbf{Y}_i = \left[\mathbf{V}_i - \sum_{p=1}^{i-1} \mathbf{L}_{ip}\mathbf{D}^{-1}_{pp}\mathbf{Y}_p\right] \; ; \; \mathbf{J}_i = \mathbf{D}^{-1}_i\left[\mathbf{Y}_i - \sum_{p=n-1}^{i-1} \mathbf{U}_{ip}\mathbf{J}_p\right]. \tag{2}$$

When the block matrices in (2) are of Rk form, we see the need for the BLAS IV operation involving the summation of matrix-matrix products:

$$\mathbf{C} \leftarrow \alpha \sum_{p=1}^{k} \mathbf{A}_p\, \mathbf{B}_p + \beta\, \mathbf{C}\,, \tag{3}$$

where **C, A** and **B** are Rk matrices. This is the matrix-matrix multiply summation form of the BLAS III case. One could argue that this is simply a repeated operation of BLAS III, and indeed it is. The difficulty is the agony of computing the Rk sum.

Matrix blocks, for perspective, are typically 500 x 500 to 10 000 x 10 000 and the number of sum terms may be in the hundreds.

A methodology for evaluating (3), using the thrill of Rk multiplication and the Adaptive Cross Approximation for bypassing the agony of low rank Rk sum evaluation is as follows.

Use Rk multiplication to set $\mathbf{A}_p\mathbf{B}_p = \mathbf{S}_p$, where **S**p terms are also Rk, so that (3) is rewritten as:

$$\mathbf{C} \leftarrow \alpha \sum_{p=1}^{k} \mathbf{S}_p + \beta\, \mathbf{C}\,. \tag{4}$$

The sum term is computed using the ACA where we recall that the ACA needs rows and columns of the matrix being approximated. Writing (4) in full Rk **UV** form we have:

$$\begin{bmatrix} C_u \\ \\ \end{bmatrix}\begin{bmatrix} C_v & & \end{bmatrix} = \alpha \sum_{p=1}^{k-1} \left(\begin{bmatrix} \mathbf{S}_u \\ \\ \end{bmatrix}\begin{bmatrix} \mathbf{S}_v & & \end{bmatrix} \right)_p + \beta \begin{bmatrix} C_u \\ \\ \end{bmatrix}\begin{bmatrix} C_v & & \end{bmatrix}. \tag{5}$$

The ACA algorithm to compute the left-hand side requires the rows / columns of the right-hand side of (5). This is a straight forward vector-matrix and matrix-vector evaluation of a gemv form:

$$\begin{bmatrix} \rightarrow & \rightarrow & \rightarrow \end{bmatrix} = \sum \left(\begin{bmatrix} \rightarrow \end{bmatrix}\begin{bmatrix} \downarrow & \downarrow & \downarrow \end{bmatrix} \right)_p$$

$$\begin{bmatrix} \downarrow \\ \downarrow \\ \downarrow \end{bmatrix} = \sum \left(\begin{bmatrix} \rightarrow \\ \rightarrow \\ \rightarrow \end{bmatrix}\begin{bmatrix} & \downarrow & \end{bmatrix} \right)_p . \tag{6}$$

IV. PC WORKSTATIONS

BLAS IV methodology has allowed the use inexpensive PC workstations, such as found on engineer’s desks, to use a direct LU factorization for full wave electromagnetic solvers. Access to very costly and limited time slot availability of super computer clusters is not needed. PC workstation problem sizes up to five million unknowns (with RWG average edge lengths of 0.1 λ) have been accomplished [5].

V. CONCLUDING REMARKS

The need for a BLAS IV for low rank Rk sums of matrix-matrix multiply was demonstrated. A computational approach using Rk multiplication for the multiply terms and the ACA for computing the sum term was outlined.

REFERENCES

[1] J. Shaeffer, “Direct solve of electrically large integral equations for problem sizes to 1 M unknowns,” IEEE Trans. Antennas Propag., vol. 56, no. 8, pp. 2306-2313, Aug. 2008.

[2] J. Shaeffer, “Low rank Matrix Algebra for the method of moments,” ACES Journal, Oct. 2018.

[3] M. Bebendorf, Hierarchical Matrices, Berlin, Springer-Verlag, 2008.

[4] J. Shaeffer, “Direct solve of electrically large integral equations for problem Sizes to 1M unknowns,” NASA/CR-2008-215353, Sept. 2008.

[5] J. Shaeffer, “Five million unknown MOM LU factorization on a PC workstation,” Antenna Measurement Techniques Association Meeting, Long Beach, CA, Oct. 11-16, 2015.