Efficient m-step Generalization of Iterative Methods†

Sun-Kyung Kim*

Abstract In order to use parallel computers in specific applications, algorithms need to be developed and mapped onto parallel computer architectures. Main memory access for shared memory system or global communication in message passing system deteriorate the computation speed. In this paper, it is found that the m-step generalization of the block Lanczos method enhances parallel properties by forming m simultaneous search direction vector blocks. QR factorization, which lowers the speed on parallel computers, is not necessary in the m-step block Lanczos method. The m-step method has the minimized synchronization points, which resulted in the minimized global communications and main memory access compared to the standard methods.

Key Words : Block Lanczos method, QR factorization, m-step generalization, parallel computer

1. Introduction

Memory contention on shared memory machines constitutes a severe bottleneck for achieving their maximum performances [1]. The same is true for communication costs on a message passing system [2]. It would be desirable to have methods for specific problems, which have low communication costs compared to the computation costs. This is interpreted as a small number of main memory access for the shared memory systems and a small number of global communications for the message passing systems. It also reduces the need for frequent synchronizations of the processors. Linear algebra algorithms, which are implemented efficiently on parallel computers, have also been studied [3,4,5].

Many important scientific and engineering problems require the computation of a small number of eigenvalues of symmetric large sparse matrices. One of the commonly used algorithm for solving a multiple eigenvalue problem is the block Lanczos algorithm. QR factorization process in the block Lanczos method is bottleneck when this method is implemented on parallel computers because QR factorization process requires many synchronization points [6,7]. In this paper we introduce the *m*-step block Lanczos method that need no *QR* factorization process. In the *m*-step method, *m* consecutive steps of the standard method are performed simultaneously. This means, for example, that the inner products needed during *m* steps of the standard method can be performed simultaneously.

The outline of this paper is as follows. The next section discusses the m-step power method and gives some idea for m-step iterative methods. The third section discusses the standard block Lanczos algorithm and the steps of QR factorization process. Also the third section describes the way how to derive the parallel algorithm for the block Lanczos

† 이 논문은 2005년 대구대학교 과제연구비에 의해 수행되었음.
* 대구대학교 컴퓨터·IT공학부

algorithm. The comparison of computational work and the analysis of communication time are discussed in the fourth section.

2. The m-step power method

The power method is a classical method to find a dominant eigenvalue of a large sparse matrix A on Krylov subspace. The power method produces a sequence of vectors and approximation of eigenvalue λ as follows:

Algorithm 2.1 The power method

Choose q_0 with $\|q_0\| = 1$
For $i = 1, 2,...$
1. $z_i = Aq_{i-1}$
2. compute (z_i, z_i)
3. $q_i = z_i/\|z_i\|_2$
4. $\lambda_i = (q_i, Aq_i)$
5. If $\|q_i - q_{i-1}\| < \varepsilon$ then stop

EndFor

There is nothing special about doing a 2-norm normalization of vectors, i.e. q_i is 2-norm normalized vector of Aq_{i-1}.

In the m-step power method, step 1 of algorithm 2.1 is performed m times and then vectors $\{Aq_{i-1}, A^2q_{i-1}, \ldots, A^mq_{i-1}\}$ are obtained. After m times execution of step 1, the vector A^mq_{i-1} is normalized and the approximation of eigenvalue is computed. The m-step power method is as follows:

Algorithm 2.2 The m-step power method

Choose $\hat{q}_0$ with $\|\hat{q}_0\| \neq 0$
For $k = 1, 2, ...$
1. compute
$A\hat{q}_{k-1}, A^2\hat{q}_{k-1}, \ldots, A^m\hat{q}_{k-1}$
$\hat{z}_k = A^m\hat{q}_{k-1}$
2. compute $(\hat{z}_k, \hat{z}_k)$
3. $\hat{q}_k = \hat{z}_k/\|\hat{z}_k\|_2$
4. $\hat{\lambda}_k = (\hat{q}_k, A\hat{q}_k)$
5. If $\|\hat{q}_k - \hat{q}_{k-1}\| < \varepsilon$ then stop

EndFor

One iteration of the m-step power method corresponds to m iterations of standard power method. This means that the vector $\hat{q}_k$, the approximation of eigenvalue $\hat{\lambda}_k$ in algorithm 2.2 correspond to the vector q_{k*m}, the approximation of eigenvalue λ_{k*m} in algorithm 2.1. In the m-step power method $2(m-1)$ inner products for one iteration are decreased compared to the corresponding m iteration of standard method and m matrix vector products can be performed simultaneously.

When the matrix A is symmetric, the Lanczos method is much faster than the power method for finding a dominant eigenvalue [8,9,10]. In the next section we introduce m-step block Lanczos method. The derivation and properties of this method are different from m-step power method except using advanced m vectors from Krylov subspace.

3. The m-step iterative method

3.1. The Block Lanczos method

The block Lanczos algorithm for computing extreme multiple eigenvalues of symmetric matrices is based on the block Lanczos recursion for the block tridiagonalization of a real symmetric matrix. The block Lanczos is as follows:

Algorithm 3.1 The block Lanczos algorithm

$X_1 \in IR^{n\times p}$ given, with $X_1^T X_1 = I_p$.

$M_1 = X_1^T A X_1$

For j = 1 **until** Convergence **Do**

$C_j = AX_j - X_j M_j - X_{j-1}B_{j-1}^T \ (X_0 B_0^T = 0)$

$X_{j+1}B_j = C_j$ (QR factorization process)

$M_{j+1} = X_{j+1}^T A X_{j+1}$

EndFor

At the beginning of the j-th pass through the loop we have

$A[X_1, \ldots, X_j] = [X_1, \ldots, X_j]T_j + C_j[0, \ldots, 0, I_j]$

where $T_j = \begin{bmatrix} M_1 & B_1^T & & \\ B_1 & M_2 & & \\ & \ddots & \ddots & \ddots \\ & & B_{j-1} & M_j \end{bmatrix}$

The eigenvalues of the block tridiagonal matrix T_j are called Rits values of large sparse matrix A. For many matrices and for relatively small j several of the extreme multiple eigenvalues of A, that is several of the algebraically-largest or algebraically-smallest of the eigenvalues of A, are well approximated by eigenvalues of the corresponding matrices T_j. The steps to implement the algorithm 3.1 on parallel computer is as follows:

Select Q_1 of size $n\times p$ where p is at least as large as the number of eigenvalues desired and the columns of Q_1 are orthonormal.

For j = 1 **until** Convergence **Do**

1. compute AQ_j
2. compute (AQ_j, Q_j)
3. $M_j = (AQ_j, Q_j)$: $p \times p$ matrix
4. $D_j = AQ_j - Q_j M_j - Q_{j-1}R_{j-1}^T (Q_0 R_0^T = 0)$
5. Apply a modified Gram-Schmidt orthogonalization to the columns of D_j (that is, $Q_{j+1}R_j = D_j$:QR factorization)

EndFor

Next, compute the p algebraically-largest eigenvalues of T_j. The Ritz vector $Q_j y$ $(= Z)$ obtained from an eigenvector y of a given T_j is an approximation to a corresponding eigenvector of A.

The above block Lanczos procedure with no re-orthogonalization has minimal storage requirements and can therefore be used on very large matrix A, if only approximations to eigenvalues are required. However, the inner products cannot be performed simultaneously because of steps 2, 5 and inner products require global communication among all processors on message passing systems. These force several accesses of vectors Q_j, D_j, AQ_j per one iteration of above procedure. For shared memory systems, main memory accessing may be slow. So in the section 3.3, the m-step block Lanczos method is proposed to perform all inner products needed for m iterations of above procedure.

3.2. Parallel modified Gram-Schmidt method

The modified Gram Schmidt orthogonalization in step 5 makes many synchronization points. The following modified Gram Schmidt algorithm computes orthonormal vector set Q for $D \in IR^{n\times p}$. That is D = QR factorization where R is the upper triangular matrix [10]:

Algorithm 3.2 Modified Gram-Schmidt Algorithm

For j = 1 **to** p

$r_{jj} = \left[\sum_{i=1}^{n} d_{ij}^2\right]^{1/2}$

For i = 1 **to** n

$q_{ij} = d_{ij}/r_{jj}$

EndFor

For $l = j+1$ **to** p

$$r_{jl} = \left[\sum_{i=1}^{n} q_{ij} d_{il}\right]^{1/2}$$

For $i = 1$ **to** n

$q_{il} = d_{il} - q_{ij} r_{jl}$

EndFor

EndFor

EndFor

The steps to implement efficiently the algorithm 3.2 on parallel computer is as follows:

For $j = 1$ **to** p

1. Compute $(\boldsymbol{d}_j, \boldsymbol{d}_j)$
2. $r_{jj} = (\boldsymbol{d}_j, \boldsymbol{d}_j)^{1/2}$
3. $\boldsymbol{q}_j = \boldsymbol{d}_j / r_{jj}$
4. **For** $l = j+1$ **to** p
   Compute $(\boldsymbol{q}_j, \boldsymbol{d}_l)^{1/2}$
   EndFor
5. **For** $l = j+1$ **to** p
   5.1 $r_{jl} = (\boldsymbol{q}_j, \boldsymbol{d}_l)^{1/2}$
   5.2 $\boldsymbol{q}_l = \boldsymbol{d}_l - \boldsymbol{q}_j r_{jl}$
   EndFor

EndFor

In the above procedure $p(p+1)/2$ inner products should be separately repeated $2p-1$ times and so communication time rate is high for this QR factorization process. In the next section we introduce m-step block Lanczos method which has much less synchronization points and need no QR factorization process.

3.3. m-step block Lanczos method

One way that can lead to build a new parallel block Lanczos algorithm is to perform m steps of the standard algorithm simultaneously in parallel using a set of $m \times p$ linearly independent vectors. In this case, the set of $m \times p$ linearly independent vectors $\overline{V_k}$ is spanned by

$$[V_k^1, AV_k^1, \ldots, A^{m-1}V_k^1, \; V_k^1 \in IR^{n \times p}].$$

Remark 1

If $\overline{V_i}$ and $\overline{V_j}$ are orthogonal for $i \neq j$, $V_k = [\overline{V_1}, \overline{V_2}, \ldots, \overline{V_k}]$ can be decomposed into $Q_k * R_k$, where $Q_k = [\overline{Q_1}, \overline{Q_2}, \ldots, \overline{Q_k}]$ and $R_k = diag\,[\overline{R_1}, \overline{R_2}, \ldots, \overline{R_k}]$.

Remark 2

If T_j is a symmetric tridiagonal matrix generated by the standard block Lanczos algorithm and $T_k = R_k^{-1} T_j R_k$ where $j = m*k$, $\overline{T_k}$ becomes a non-symmetric matrix similar to T_j as follows:

$$\overline{T_k} = \begin{bmatrix} G_1 & E_1 & & & \\ F_1 & G_2 & E_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & & E_{k-1} \\ & & & F_{k-1} & G_k \end{bmatrix}$$

where $G_i = [G_i^1, \ldots, G_i^m]$, $G_i^j \in IR^{mp \times p}$ and $E_i = [E_i^1, \ldots, E_i^m]$, $E_i^j \in IR^{mp \times p}$ for $i = 1, \ldots, k$, and $j = 1, \ldots, m$. Here F_i is a specially shaped matrix of which the block in the upper right corner is a triangular matrix with zero elements otherwise. This upper triangular matrix arises from the QR factorization of V_i^j, together with orthonormal vectors. The following is an example of F_i for $p = 3$, and m $= 2$:

$$F_i = \begin{bmatrix} 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Since $\overline{T_k}$ is similar to T_j with $j=k*m$, they

have the same eigenvalues. Note that V_i^l is not orthogonal, but it does not induce any problem in constructing the algorithm. The reason is that, as given in Remark 1, $\overline{V}_i$ and $\overline{V}_j$ are orthogonal for $i \neq j$ and $\overline{V}_i = [V_i^1, V_i^2, \ldots, V_i^m]$ is a set of linearly independent vectors that span $[V_i^1, AV_i^1, \ldots, A^{m-1}V_i^1,\ V_k^i \in R^{n\times p}]$. Based on the above analysis, a new parallel block Lanczos algorithm can be constructed as follows:

Algorithm 3.3 The m-step block Lanczos algorithm

$\overline{V}_0 = 0,\ \overline{V}_1 = [V_1^1, AV_1^1, A^2V_1^1, \ldots, A^{m-1}V_1^1]$

For $k = 1$ **until** Convergence **Do**

Select G_k , E_{k-1} so that $\overline{V}_k$ is orthogonal to $\overline{V}_{k-1}$.

This gives

$$V_{k+1}^1 = AV_k^m - \overline{V}_{k-1}E_{k-1}^m - \overline{V}_kG_k^m$$

Select C_k^j such that $\overline{V}_{k-1}$ is orthogonal to $[V_k^1, AV_k^1, \ldots, A^{m-1}V_k^1]$ which gives $V_{k+1}^j = A^{j-1}V_{k+1}^1 - \overline{V}_kC_k^j$ for $j = 2, \ldots, m$.

EndFor

We will present the reduction matrix generated by m-step block Lanczos method for the special case of m = 2 and $p = 2$:

$$\overline{T}_k = \left[\begin{array}{cccc|cccc|cc} * & * & * & * & * & * & * & * & & \\ * & * & * & * & * & * & * & * & & \\ * & * & * & * & * & * & * & * & & \\ & * & * & * & * & * & * & * & & \\ \hline & & 1 & 0 & * & * & * & * & \cdot & \\ & & & 1 & * & * & * & * & \cdot & \\ & & & & * & * & * & * & \cdot & \\ & & & & & * & * & * & \cdot & \\ & & & & & & & & & \\ & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right]$$

Next, we demonstrate how to determine the parameters G_k, E_{k-1}, C_k in the above process:

Scalar 1: $\overline{V}_k{}^T A\,\overline{V}_k = \overline{V}_k{}^T\overline{V}_kG_k$,

$$\overline{V}_{k-1}{}^T A\,\overline{V}_k = \overline{V}_{k-1}{}^T\overline{V}_{k-1}E_{k-1}$$

Scalar 2: $0 = \overline{V}_k{}^TA^{j-1}V_k^1 - \overline{V}_k{}^T\overline{V}_kC_k^j$

The inner products $(\overline{V}_k, \overline{V}_k)$, $(\overline{V}_k, A\overline{V}_k)$, $(\overline{V}_{k-1}, A\overline{V}_k)$, $(\overline{V}_k, A^{j-1}V_k^1)$ can be reduced to $(V_k^1, V_k^1), (V_k^1, AV_k^1), \ldots, (V_k^1, A^{2m-1}V_k^1)$ in a similar way to the s-step Lanczos method[7]. We now formulate the m-step block Lanczos algorithm.

Select V_1^1

For $k = 1$ **until** Convergence **Do**

1. Compute $AV_k^1, A^2V_k^1, \ldots, A^mV_k^1$
2. Compute $2p^2m$ inner products
3. Compute Scalars

   C_{k-1}^j for $j = 2, \ldots, m$

   E_{k-1}^j, G_k^j for $i = 1, \ldots, m$
4. Compute

   $V_k^j = A^{j-1}V_k^1 - \overline{V}_{k-1}C_{k-1}^j$

   (with $\overline{V}_0 = 0$, $C_0^j = 0$)
5. Compute AV_k^m
6. Compute

   $V_{k+1}^1 = AV_k^m - \overline{V}_{k-1}E_{k-1}^m - \overline{V}_kG_k^m$

   (with $E_0^m = 0$)

EndFor

In the above algorithm all the inner products are performed at once in the step 2. Matrix-vector operations are performed in the step 1 and 5, and the other steps are for vector update operation.

4. Analysis of m-step iterative method

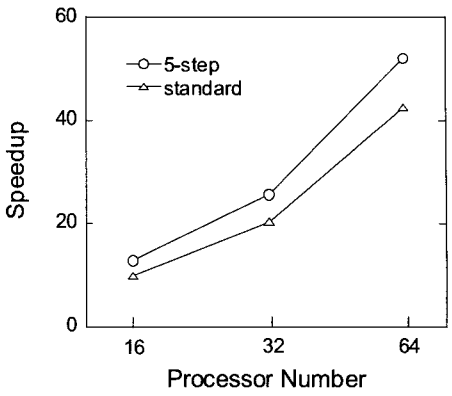
During the procedure of the standard block Lanczos algorithm in the Section 3.1, $p^2+p(p+1)/2$ inner products should be separately repeated at $2p$ times. Thus on message passing system, lots of times are needed for global communication, which in turn reduces the efficiency of the parallel system.

However, the new parallel algorithm introduced in Section 3.3 can perform all the needed inner products at once during m iterations of the corresponding standard method, the time necessary for inner products is reduced by a factor of $1/(2pm)$compared to the standard method. The main cause for the increased memory latency time for the shared memory system is too many synchronizing point. In a new parallel algorithm 3.3 such a bottleneck phenomena can be reduced. Table 1 shows the number of vector operations and the data communications during a single iteration of the m-step block Lanczos algorithm and m iterations of the corresponding standard method.

<Table 1> Comparison of the numbers of the vector operations and the data communications for block Lanczos method

	standard algorithm	m-step algorithm
inner product	$[p^2 + p(p+1)/2]m$	$2p^2m$
vector update	$(2p+1)pm$	$pm(m+1)$
matrix-vector multiplication	pm	$p(m+1)$
Global communication	$2pm$	1
Local communication	m	2

As shown in Table 1, the communication cost can be greatly decreased by introducing more effective algorithm in parallel process.



<Fig. 1> Perfomance of the standard and the m-step block Lanczos algorithm in Cray T3E.

Fig. 1 shows the efficiency of the 5-step block Lanczos algorithm implemented on MP parallel computer Cray T3E. All the inner products needed for m iterations of the standard method are performed at the same time in the m-step method, so that only one global communication is required and the communication cost is decreased on a message passing system like Cray T3E. But the m-step block Lanczos method needs $p(m+1)$ times of matrix-vector operations compared to pm times in the standard methods. But in the case of vector updates, the number is slightly smaller in the m-step method. The m-step algorithm is also effective in the matrix-vector operations since the communication time between the adjacent processors of Cray T3E can be reduced. While the efficiency of parallelism increases with increasing m and p, but a loss of accuracy for eigenvalues has been observed for large $m > 5$.

5. Conclusions

Parallel processing systems equipped with from a few to many thousand processors are

now being used in many areas. As the number of the processors involved in the parallel system is increased, the relative importance of the communication cost grows. In this paper, we proposed new m-step iterative method suitable to reduce the communication cost. The m-step block Lanczos algorithm utilizes a reduced matrix similar to that in the standard block Lanczos method with the same eigenvalues, but m-step method is more effective in the parallel system because a large amount of the inner products can be done at once. This process can reduce the data communication time in a message passing system. The m-step method also help to reduce the memory latency time in shared memory systems by reducing the synchronization point showing a better data locality. The new m-step method has the better performance compared to the standard method in MP parallel computer Cray T3E.

References

[1] Paul E. Saylor, "Leapfrog Variants of Iterative Methods for Linear Algebraic Equations", *Journal of Computational and Applied Mathematics*24 (1988) 169-193

[2] Sanjay Ranka, Youngju Won, and Sartaj Sahni, "Programming the NCUBE Hypercube", *Tech. Rep. Csci* No 88-13, Univ. of Minnesota, (1988)

[3] Ameet K. Dave and Iain S. Duff, "Sparse Matrix Calculations on the CRAY-2", *Parallel Computing* 5 (1987) 55-64

[4] Inge Gutheil and Werner Krotz-Vogel, "Performance of a Parallel Matrix Multiplication Routine on Intel iPSC/860", *Parallel Computing* 20 (1994) 953-974

[5] K.K. Mathur and S. Lennart Johnsson, "Multiplication of Matrices of Arbitrary Shape on a Data Parallel Computers", *Parallel Computing* 20 (1994) 919-951

[6] Claus Bendtsen, Per Christian Hansen, Kaj Madsen, Hans Bruun Nielsen, Mustafa Pinar, "Implementation of QR up- and downdating on a massively parallel computer", *Parallel Computing* 21 (1995) 49-61

[7] Pontus Matstoms, "Parallel sparse QR factorization on shared memory architectures", *Parallel Computing* 21 (1995) 473-486

[8] Jane K. Cullum and Raph A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalues Computation*, Birkhauser Boston, Inc. (1985)

[9] James W. Demmel, *Applied Numerical Linear Algebra*, the Society for Industrial and Applied Mathematics press. (1997)

[10] G. H. Golub, C. F. Van Loan, *MATRIX Computations*, Johns Hopkins University Press. (1996)

김 선 경 (Sun-Kyung Kim)

- 1979년 이화여자대학교 수학과 학사
- 1982년 한국과학기술원 전산학과 석사
- 1991년 미국 Minnesota대학원 전산학과 박사
- 1992년~현재 대구대학교 컴퓨터·IT공학부 교수
- 관심분야 : 과학계산, 병렬처리, 알고리즘, 가상강의