A NEGATIVE RESULT ON SPARSE MATRIX SPLITTING IN THE CONTEXT OF GAUSSIAN ELIMINATION

by

Alan George

Research Report CS-75-10

Department of Computer Science

University of Waterloo
Waterloo, Ontario, Canada

April 1975

This work was supported in part under NASA Grant NGR 47-102-001 while the author was a visiting scientist at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, and in part by Canadian National Research Council Grant A8111. A preliminary version of this paper appeared as an ICASE report.

A NEGATIVE RESULT ON SPARSE MATRIX SPLITTING IN THE CONTEXT OF GAUSSIAN ELIMINATION

by

Alan George

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

This work was supported in part under NASA Grant NGR 47-102-001 while the author was a visiting scientist at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, and in part by Canadian National Research Council Grant A8111. A preliminary version of this paper appeared as an ICASE report.

Abstract

Let $A = B + C$ be a given sparse nonsingular matrix with $B_{ij}C_{ij} = 0$, and let L_AU_A, L_BU_B and $L_{\bar{C}}U_{\bar{C}}$ be triangular factors of A, B and $\bar{C} = I + L_B^{-1}CU_B^{-1}$, computed using Gaussian elimination. Given the problem of solving $Ax = b$, we can factor A directly into L_AU_A and then compute x by solving $L_Ay = b$ and $U_Ax = y$. Alternatively, we can factor B and then compute and factor $\bar{C}$, obtaining the factorization $A = L_BL_{\bar{C}}U_{\bar{C}}U_B$ which can then be used to compute x. In this paper we show that over all splittings of A, this latter strategy cannot be beneficial in terms of reducing arithmetic or storage requirements.

1. Introduction

Consider using Gaussian elimination to solve the sparse nonsingular n by n linear algebraic system

(1.1) $Ax = b,$

where we assume A has a triangular factorization $L_A U_A$. Given L_A and U_A, where L_A is unit lower triangular and U_A is upper triangular, we obtain x by solving the triangular systems $L_A y = b$ and $U_A x = y$.

Alternatively, we might consider an additive splitting B + C of A and solve (1.1) using Gaussian elimination as follows: (a) Factor B into $L_B U_B$ (assuming the factorization exists), (b) Calculate and factor $\bar{C} = I + L_B^{-1} C U_B^{-1}$ into $L_{\bar{C}} U_{\bar{C}}$ (assuming it exists), and (c) solve the four triangular systems $L_B y_1 = b$, $L_{\bar{C}} y_2 = y_1$, $U_{\bar{C}} y_3 = y_2$ and $U_B x = y_3$. Here and elsewhere, it is to be understood that inverses are not actually computed; instead the appropriate triangular systems are solved, using the standard back-substitution procedure.

Why contemplate such a procedure? Except in rather special circumstances, when Gaussian elimination is applied to a sparse matrix A, it suffers "fill"; that is, its triangular factors L_A and U_A typically have nonzeros in positions which are zero in A. Sometimes the judicious removal of a few strategic nonzeros in A can substantially reduce the amount of fill that occurs. That is, in the above notation, a judicious choice of C, involving only a very few components, can make $L_B + U_B$ very much sparser than $L_A + U_A$. However, the crucial question is whether the factors of $\bar{C}$ and B together can have fewer nonzero components than the factors of A.

The purpose of this article is to show that the answer to this question is no. We also show that this splitting technique cannot be beneficial in terms of reducing the number of multiplications required to solve (1.1).

2. Notation, Assumptions, and Preliminary Results

Our introduction implies that the motivation for splitting in this paper is to induce, in a judicious way, some sparsity in B. Thus, throughout the article, when we refer to a splitting B + C of A, we implicitly assume that $B_{ij}C_{ij} = 0$.

In subsequent sections, we require various items defining the structure of vectors and matrices, such as the number of nonzero components in a vector, or the positions of the nonzeros in a vector. In this connection, we denote the number of nonzeros in Δ by $p(\Delta)$, where Δ may be a vector or matrix, and we denote the i-th row and column of a given matrix M by M_{i*} and M_{*i} respectively. In order to discuss the positions of nonzero components, we define the sets

$$\Omega(v) = \{k | v_k \neq 0\},$$

where v is a vector, and

$$\Omega(M) = \{(i,j) | M_{ij} \neq 0\},$$

where M is a matrix.

We denote the unit lower and upper triangular factors of a matrix M, computed using Gaussian elimination, by L_M and U_M. In doing so, we implicitly assume the existence of the factorization $M = L_M U_M$. (We will sometimes drop the subscript when no confusion results.) In a practical

sense we require much more; Gaussian elimination applied to M should be numerically stable. This will be true, for example, if A is diagonally dominant, or symmetric and positive definite [3].

We assume throughout that whenever two nonzero quantities are added or subtracted, the result is nonzero. That is, we ignore any zeros which are created through exact cancellation. Such cancellation rarely occurs, and in general it is difficult to predict, particularly in floating point computation which is subject to rounding error. Thus, such accidental zeros are not normally exploited in implementations. Moreover, there are matrices for which it is possible to guarantee that no such cancellation will occur. With this proviso, our measure of storage requirements is the number of nonzero components in the factorizations.

We measure arithmetic requirements by the number of multiplicative operations (multiplications and divisions). The majority of the arithmetic performed in Gaussian elimination involves sequences of operations occurring in multiplying-add pairs, so the number of additions and multiplications is about the same. Thus, we contend that the multiplication count is a reasonable measure of arithmetic requirements.

We now collect some results on operation counts which we need in Section 3.

Lemma 2.1 (Bunch and Rose [1])

The calculation of the factorization LU of M, using Gaussian elimination and avoiding operating on all zeros, requires $\theta(M)$ multiplications, where

$$\theta(M) = \sum_{i=1}^{n-1} (p(L_{*i})-1)p(U_{i*}).$$

Lemma 2.2 is simply a statement of the following observations. Suppose LU and L'U' are factorizations of M and M', computed using Gaussian elimination, where M' is obtained from M by adding to it a matrix N without cancellation. (Thus, $\Omega(M) \subseteq \Omega(M')$). Since M' is nonzero wherever M is, and we assume numerical cancellation doesn't occur, it follows that $\Omega(L+U) \subseteq \Omega(L'+U')$ and $\theta(M) \leq \theta(M')$.

Now if $N_{ij} \neq 0 \Rightarrow (L+U)_{ij} \neq 0$, (i.e., $\Omega(N) \subseteq \Omega(L+U)$), then the addition of N to M cannot affect the <u>structure</u> of the factors of M, since the positions in L+U corresponding to nonzero components of N were going to be nonzero anyway. Thus, $\Omega(L'+U') = \Omega(L+U)$, and $\theta(M) = \theta(M')$ because θ is a function of the structure of the <u>factors</u> of the matrix rather than its own structure.

On the other hand, if $\Omega(N) \not\subseteq \Omega(L+U)$, then it is clear that $\Omega(L'+U') \neq \Omega(L+U)$, and therefore $\theta(M) < \theta(M')$.

We state these observations as

<u>Lemma 2.2</u>

Let LU and L'U' be triangular factors of M and M' respectively, computed using Gaussian elimination, where M' = M+N and $\Omega(M) \subseteq \Omega(M')$.

Then $\Omega(L+U) \subseteq \Omega(L'+U')$ and $\theta(M) \leq \theta(M')$, with equality occurring if and only if $\Omega(N) \subseteq \Omega(L+U)$. □

It is helpful to be able to denote the number of multiplications required to compute $T^{-1}Y$, where T is a given n by n triangular matrix and Y is an n by r matrix, $r \geq 1$. We denote this number by $\zeta(T,Y)$. The following lemma defines this quantity for r = 1.

Lemma 2.3 (George [2])

Let T be a triangular matrix and let x be the solution to Tx = b, computed using the standard back-substitution procedure for triangular matrices. Then the number of multiplications required to compute x is

$$(2.2) \qquad \zeta(T,b) = \sum_{\ell \in \Omega(x)} (p(T_{*\ell})-\mu),$$

where $\mu = 0$ unless T has a unit diagonal which is exploited, in which case $\mu = 1$.

The corresponding result for $r > 1$ is obvious. The next two lemmas are immediate consequences of Lemma 2.3.

Lemma 2.4

Let T be a triangular matrix and let w and x be the computed solutions to Tw = y and Tx = z respectively, where $\Omega(y) \subseteq \Omega(z)$. Then

a) $\Omega(w) \subseteq \Omega(x)$,

and

b) $$\zeta(T,y) = \zeta(T,z) - \sum_{\ell \in \Omega(x) \setminus \Omega(w)} (p(T_{*\ell})-\mu),$$

where μ is as defined in Lemma 2.3 □

Lemma 2.5

Let T be an n by n triangular matrix and let W and X be the computed solutions to TW = Y and TX = Z, where Y and Z are n by r matrices and $\Omega(Y) \subseteq \Omega(Z)$. Then

a) $\Omega(W) \subseteq \Omega(X)$

and

b) $\zeta(T,Y) \leq \zeta(T,Z)$. □

In section 3, we are faced with alternative ways of computing some quantity, say $F = T^{-1}G$, where $G = Y + Z$. One computation may yield F by calculating $W = T^{-1}Y$ and $X = T^{-1}Z$, and then adding W and X, while the alternate computation may add Y and Z yielding G, and then compute $F = T^{-1}G$. Lemma 2.6 below states that this latter computation is always at least as efficient as the first one.

Lemma 2.6

Let T be an n by n triangular matrix and let W and X be the computed solutions to $TW = Y$ and $TX = Z$ respectively. Let F be the computed solution to $TF = G$, where $G = Y + Z$. Then

a) $\Omega(F) = \Omega(X+W)$,

and

b) $\zeta(T,G) \leq \zeta(T,Y) + \zeta(T,Z)$.

Proof Let the lower case letter w denote the i-th column of W, and similarly for X, F, Y, Z and G.

Since $g = y+z$, $\Omega(y) \subseteq \Omega(g) \Rightarrow \Omega(w) \subseteq \Omega(f)$, by Lemma 2.4. Similarly, $\Omega(z) \subseteq \Omega(g) \Rightarrow \Omega(x) \subseteq \Omega(f)$. Thus, since $\Omega(x+w) = \Omega(x) \cup \Omega(w)$, we have

$$\Omega(x+w) \subseteq \Omega(f). \tag{2.3}$$

We now want to show that $\Omega(f) \subseteq \Omega(x+w)$, whence we can conclude part a) of the lemma. For definiteness, we suppose T is lower triangular; the proof for T upper triangular is trivially different.

Let $k \in \Omega(f)$. Then

$$f_k = g_k - \sum_{\ell=1}^{k-1} T_{k\ell} f_\ell. \tag{2.4}$$

Now $f_k \neq 0 \Rightarrow \{$ i) $g_k \neq 0$ or ii) $\exists k_1 < k \ni f_{k_1} \neq 0$ and $T_{kk_1} \neq 0\}$. If i) prevails, then either $y_k \neq 0$ ($\Rightarrow w_k \neq 0$) or $z_k \neq 0$ ($\Rightarrow x_k \neq 0$). Otherwise, we repeat the argument with k replaced by k_1. Ultimately, we must find a k_r for which $f_{k_r} \neq 0$ ($T_{k_{r-1}k_r} \neq 0$) and $g_{k_r} \neq 0 \Rightarrow \{w_{k_r} \neq 0$ or $x_{k_r} \neq 0\}$. Suppose for definiteness that $w_{k_r} \neq 0$.

But we have just established the existence of a sequence of subscripts $(k_{r-1},k_r),(k_{r-2},k_{r-1})\ldots(k,k_1)$ for which T is nonzero; this together with $w_{k_r} \neq 0 \Rightarrow w_k \neq 0$.

Thus, we have shown that $k \in \Omega(f) \Rightarrow \{k \in \Omega(x)$ or $k \in \Omega(w)\}$; i.e.

$$\Omega(f) \subseteq \Omega(x) \cup \Omega(w) = \Omega(x+w). \tag{2.5}$$

Relations (2.3) and (2.5) imply part a) of the lemma. Part b) follows immediately from part a) and lemma 2.4. □

3. Results

Recall that in the introduction, for a splitting B + C of A, we defined the matrix $\bar{C}$ by

$$\bar{C} = I + L_B^{-1} C U_B^{-1}. \tag{3.1}$$

Recall also that the question of whether or not splitting is beneficial depends on the number of nonzeros in the factors of A, compared to the total number of nonzeros in the factors of B and $\bar{C}$. In this section, we show that these latter factors always collectively contain as many nonzeros as the factors of A. We also show that the number of multiplications involved in using the splitting is at least as great as that required to factor A.

In order to prove these results, it is helpful to be able to rule out some splittings by showing that they can never be any better than certain other splittings. This is the object of Theorem 3.1, which states that there is no point in considering splittings for which $C_{ij} \neq 0$ and $(L_B+U_B)_{ij} \neq 0$. Recall from section 2 that for our purposes, reference to a splitting B + C of A implies that $B_{ij}C_{ij} = 0$. In the theorem below, $\bar{F}$ is defined by (3.1) with B and C replaced by E and F.

Theorem 3.1

Let B + C and E + F be two splittings of A, where $\Omega(C) \subseteq \Omega(F)$ and $\Omega(F) \setminus \Omega(C) \subseteq \Omega(L_E + U_E)$. Then

a) $\Omega(\bar{C}) \subseteq \Omega(\bar{F})$,

and

b) The factorization of E and calculation of $\bar{F}$ requires at least as many multiplications as the factorization of B and calculation of $\bar{C}$.

Proof Notice first that since $B_{ij}C_{ij} = 0$, $E_{ij}F_{ij} = 0$, and $\Omega(C) \subseteq \Omega(F)$, we have B = E+(F-C) and $\Omega(E) \subseteq \Omega(B)$. Setting M' = B, M = E, and N = F-C in Lemma 2.2 and using the fact that $\Omega(F-C) \subseteq \Omega(L_E+U_E)$, we have

$$\Omega(L_E+U_E) = \Omega(L_B+U_B) \tag{3.2}$$

and

$$\theta(E) = \theta(B). \tag{3.3}$$

Thus, (3.2) implies that the calculation of $I+L_B^{-1}CU_B^{-1}$ requires the same number of multiplications as that required to compute $I+L_E^{-1}CU_E^{-1}$, which (under our no cancellation assumption) cannot be greater than the number required to compute $I+L_E^{-1}FU_E^{-1}$ since $\Omega(C) \subseteq \Omega(F)$. It follows from Lemma 2.4 that $\Omega(\bar{C}) \subseteq \Omega(\bar{F})$. □

The above result allows us to simplify the proof of the theorem which follows, by allowing us to make some assumptions about C.

Theorem 3.2

Let B + C be a splitting of a given n by n sparse matrix A, where A, B, and $\bar{C} = I+L_B^{-1}CU_B^{-1}$ have triangular factorizations L_AU_A, L_BU_B and $L_{\bar{C}}U_{\bar{C}}$ respectively, computed using Gaussian elimination. Then

a) $p(L_A+U_A) \leq p(L_B+U_B) + p(L_{\bar{C}}+U_{\bar{C}})$,

and

b) $\theta(A) \leq \theta(B) + \theta(\bar{C}) + \sigma(\bar{C})$,

where $\sigma(\bar{C})$ is the number of multiplications required to compute $\bar{C}$, given L_B, U_B and C.

Proof We assume $\bar{C}$ is computed as $I+(L_B^{-1}C)U_B^{-1}$; the modification of the proof if $\bar{C}$ is computed as $I+L_B^{-1}(CU_B^{-1})$ is straightforward.

The proof is by induction on the order of the matrix. The result obviously holds for 1 by 1 and 2 by 2 matrices, and we suppose it holds for n by n matrices. Consider the splitting B' + C' of the (n+1) by (n+1) matrix A' shown below. Since $(L_{B'}+U_{B'})_{11} \neq 0$, by Theorem 3.1 we can with no loss of generality set $C'_{11} = 0$.

$$(3.4) \qquad A' = \begin{pmatrix} d & u_2^T+v_2^T \\ u_1+v_1 & \bar{B}+C \end{pmatrix} = \begin{pmatrix} d & u_2^T \\ u_1 & \bar{B} \end{pmatrix} + \begin{pmatrix} 0 & v_2^T \\ v_1 & C \end{pmatrix} = B' + C'$$

First consider applying Gaussian elimination to A'. After one step, we have the partial factorization.

$$(3.5)\qquad A' = \begin{pmatrix} 1 & 0 \\ d^{-1}(u_1+v_1) & I_n \end{pmatrix} \begin{pmatrix} d & u_2^T+v_2^T \\ 0 & A \end{pmatrix}$$

where

$$(3.6)\qquad A = \bar{B} + C - d^{-1}(u_1+v_1)(u_2+v_2)^T$$

$$= \bar{B} + C - d^{-1}(u_1u_2^T + u_1v_2^T + v_1u_2^T + v_1v_2^T).$$

Thus, the number of nonzeros in the factors of A' is given by

$$(3.7)\qquad \mu = p(L_A+U_A) + p(u_1+v_1) + p(u_2+v_2) + 1.$$

The number of multiplications required to factor A' is given by

$$(3.8)\qquad \nu = \theta(A) + p(u_1+v_1)(p(u_2+v_2) + 1)$$

$$\leq \theta(A) + p(u_1) + p(v_1) + p(u_1)p(u_2) + \omega,$$

where

$$(3.9)\qquad \omega = \min\{p(u_1)p(v_2) + p(v_1)p(u_2+v_2), p(u_1+v_1)p(v_2) + p(v_1)p(u_2)\}.$$

The reason for defining ω will be apparent later in the proof.

Now consider the use of the proposed splitting. Factoring B', we have

$$(3.10)\qquad B' = \begin{pmatrix} 1 & 0 \\ d^{-1}u_1 & L_B \end{pmatrix} \begin{pmatrix} d & u_2^T \\ 0 & U_B \end{pmatrix},$$

where L_BU_B is the factorization of

$$(3.11)\qquad B = \bar{B} - (d^{-1}u_1)u_2^T.$$

Some elementary matrix algebra yields

$$(3.12)\qquad \bar{C}' = \begin{pmatrix} 1 & v_2^T U_B^{-1} \\ d^{-1}(L_B^{-1}v_1) & I+[L_B^{-1}(C-(d^{-1}u_1)v_2^T)-(d^{-1}L_B^{-1}v_2)u_2^T]U_B^{-1} \end{pmatrix}$$

where the parentheses indicate the order in which quantities are computed. After one step of Gaussian elimination is applied to $\bar{C}'$, we obtain the partial factorization

$$(3.13)\qquad \bar{C}' = \begin{pmatrix} 1 & 0 \\ d^{-1}(L_B^{-1}v_1) & I_n \end{pmatrix}\begin{pmatrix} 1 & v_2^T U_B^{-1} \\ 0 & \bar{C} \end{pmatrix}$$

where

$$(3.14)\qquad \bar{C} = I+[L_B^{-1}(C-(d^{-1}u_1)v_2^T)-(d^{-1}L_B^{-1}v_1)u_2^T]U_B^{-1}-(d^{-1}L_B^{-1}v_1)(v_2^T U_B^{-1}).$$

Applying Lemma 2.6 a number of times, we conclude $\bar{C}$ has the same structure as it would have if it was computed as

$$(3.15)\qquad \bar{C} = I+L_B^{-1}[C-d^{-1}(u_1v_2^T+v_1u_2^T+v_1v_2^T)]U_B^{-1}.$$

Now, by the induction hypothesis, the factors of $\bar{C}$ and the factors of $B = L_B U_B$ have in total as many nonzero components as the factors of the matrix

$$\begin{aligned} &B + C - d^{-1}(u_1v_2^T+v_1u_2^T+v_1v_2^T) \\ &\quad = (\bar{B}-d^{-1}u_1u_2^T) + (C-d^{-1}(u_1v_2^T+v_1u_2^T+v_1v_2^T)), \end{aligned}$$

which is just A (equation (3.6)). Thus, we conclude that

(3.16) $\quad p(L_B+U_B) + p(L_{\bar{C}}+U_{\bar{C}}) \geq p(L_A+U_A).$

Using (3.8), (3.10), (3.13), and (3.16), we conclude that the number of non-zeros μ^s in the factors of B' and C' satisfies

$$\begin{aligned}\mu^s &\geq p(L_B+U_B) + p(L_{\bar{C}}+U_{\bar{C}}) + p(u_1) + p(u_2) + p(v_1) + p(v_2) + 1 \\ &\geq p(L_A+U_A) + p(u_1+v_1) + p(v_2+u_2) + 1 \\ &= \mu.\end{aligned}$$

This establishes part a) of the theorem.

Now consider the number of multiplications involved in utilizing the splitting. The factorization of B' requires $\theta(B')$ multiplications, where

(3.17) $\quad \theta(B') = p(u_1) + p(u_1)p(u_2) + \theta(B).$

The repeated application of Lemma 2.6 shows us that the number of multiplications required to compute $\bar{C}$ as indicated by (3.14) is at least as great as that required to compute it as

(3.18) $\quad \bar{C} = I + (L_B^{-1}W)U_B^{-1},$

where

(3.19) $\quad W = C - (d^{-1}u_1)v_2^T + (d_1^{-1}v_1)v_2^T + (d^{-1}v_1)u_2^T.$

Now the number of multiplications required to compute W, given $d^{-1}v_1$, $d^{-1}u_1$, u_2,v_2 and C is at least ω, which is defined by (3.9). Counting the divisions

involved in calculating $d^{-1}(L_B^{-1}v_1)$ in $\bar{C}'$, which is at least $p(v_1)$, and letting $\sigma(\bar{C})$ be the cost of computing $\bar{C}$ given L_B, U_B and W, we have that the number of multiplications required to utilize the splitting is

$$\begin{aligned}
\nu^S &\geq \theta(B') + \omega + p(v_1) + \theta(\bar{C}) + \sigma(\bar{C}) \\
&= p(u_1) + p(v_1) + p(u_1)p(u_2) + \theta(B) + \theta(\bar{C}) + \sigma(\bar{C}) + \omega \quad \text{(using 3.17)}, \\
&\geq \nu - \theta(A) + \theta(B) + \theta(\bar{C}) + \sigma(\bar{C}), \qquad \text{(using 3.8)}.
\end{aligned}$$

But by the induction hypothesis, $\theta(A) \leq \theta(B) + \theta(\bar{C}) + \sigma(\bar{C})$, which implies that $\nu^S \geq \nu$, concluding the proof. □

To summarize, Theorem 3.2 tells us two things. First, part a) says that for <u>any</u> splitting B + C of A, the number of nonzeros in the factors of A cannot exceed the number of nonzeros in the factors of B and $\bar{C}$. Since this is our measure of storage requirements for solving (1.1), we conclude that the splitting strategy cannot reduce storage requirements. Second, part b) says that the number of multiplications required to factor A cannot exceed the total number required to factor B, calculate $\bar{C}$, and factor $\bar{C}$. It follows that the splitting scheme cannot reduce the number of multiplications <u>required to solve (1.1)</u>.

4. <u>Concluding Remarks</u>

There are other ways to utilize the splitting A = B + C. For example, we could compute and factor $\tilde{C} = I + B^{-1}C$, obtaining the factorization $A = L_B U_B L_{\tilde{C}} U_{\tilde{C}}$ which can then be used in solving Ax = b. It is not difficult to construct an A for which this strategy can pay handsomely in terms of storage and arithmetic reduction. However, numerous examples

suggest that we can always find a permutation matrix P such that the factorization or a block factorization of PAP^T is as efficient as using the splitting as described above. It appears to be difficult to prove or disprove this conjecture. In this connection, however, it is easy to show that Theorem 3.1 holds if $\bar{C}$ is defined as $I + B^{-1}C$ or $I + CB^{-1}$, and similarly for $\bar{F}$.

5. Acknowledgements

The author is grateful to Drs. J.M. Ortega and W.G. Poole of ICASE for comments and suggestions, and to D.R. McIntyre for a careful reading and criticism of the manuscript.

6. References

[1] J.R. Bunch and D.J. Rose, "Partitioning, tearing, and modification of sparse linear systems", J. Math. Anal. and Appl., to appear.

[2] Alan George, "On block elimination for sparse linear systems", SIAM J. Numer. Anal., 11 (1974), pp.585-603.

[3] J.H. Wilkinson, "Error analysis of direct methods of matrix inversion", J. Assoc. Comput. Mach. 8, (1961), pp.281-330.