UNSYMMETRIC POSITIVE DEFINITE LINEAR SYSTEMS

by

Gene H. Golub
Charles Van Loan

TR78-352

Department of Computer Science
Cornell University
Ithaca, New York 14853

UNSYMMETRIC POSITIVE DEFINITE LINEAR SYSTEMS

Gene H. Golub and Charles Van Loan

Abstract:

Is it necessary to pivot when solving an unsymmetric positive definite linear system $Ax = b$? Define $T = (A + A^T)/2$ and $S = (A + A^T)/2$. It is shown that pivoting is unnecessary if the quantity $\| S T^{-1} S \|_2 / \| A \|_2$ is suitably small with respect to the working machine precision.

Authors' Addresses:

Professor Gene H. Golub
Department of Computer Science
Stanford University
Stanford, California 94305

Professor Charles Van Loan
Department of Computer Science
Cornell University
Ithaca, New York 14853

1. Introduction

An nxn real matrix A is positive definite if $x^T Ax > 0$ for all nonzero x in R^n. Setting $A = T + S$ where

$$(1.1) \qquad T = (A + A^T)/2 \qquad \text{"the symmetric part of A"}$$

and

$$(1.2) \qquad S = (A - A^T)/2 \qquad \text{"the skew-symmetric part of A"}$$

we see that A is positive definite if and only if T is positive definite. Such matrices have a number of important properties [4]. In particular, if A is positive definite, then each of its principle submatrices is nonsingular and therefore the factorization

$$(1.3) \qquad A = L D M^T \qquad \begin{array}{l} L,M \text{ unit lower triangular} \\ D = \mathrm{diag}(d_1,\ldots,d_n) \end{array}$$

exists.

The LDM^T factorization is a slight rearrangement of the more familiar LU factorization ($U = DM^T$) and can be computed in $n^3/3$ operations using either Gaussian elimination or Crout reduction techniques [6,p.131ff]. Denote computed quantities with the "hat" notation "$\wedge$" and suppose $\hat{L}$, $\hat{D}$, and $\hat{M}$ are used to solve $Ax = b$.

Using the general roundoff error analysis of deBoor and Pinkus [3], it can be shown that the computed solution $\hat{x}$ satisfies

(1.4) $$(A + E)\hat{x} = b$$

where

(1.5) $$|E| \leq u\, c_n\, |\hat{L}|\,|\hat{D}|\,|\hat{M}^T| \quad ,$$

Here, u is the machine precision (i.e. β^{1-t} where t digit, base β arithmetic is used) and c_n is a constant which depends linearly on n and whose exact value depends upon the details of the algorithm The matrix absolute value $|\cdot|$ is defined by $|Z| = (\,|z_{ij}|\,)$ and the relation $|Z| \leq |Y|$ implies $|z_{ij}| \leq |y_{ij}|$ for all i and j .

For general matrices it is customary to pivot and thereby determine the LDM^T factorization of a row permuted version of A. The resulting matrix $|\hat{L}|\,|\hat{D}|\,|\hat{M}^T|$ can then be suitably bounded implying from (1.4) and (1.5) that $\hat{x}$ satisfies a "nearby" linear system.

For several important classes of matrices, however, pivoting is not a numerical necessity. This is always a welcome state of affairs from the standpoint of program simplicity and efficiency. For example, if A is <u>both</u> positive definite <u>and</u> symmetric, then (a) $M = L$ in (1.3), (b) computational requirements are reduced to $n^3/6$, and (c) the matrix E in (1.4) can be shown to satisfy

(1.6) $$\| E \|_\infty \leq u\, q_n\, \| A \|_\infty \quad .$$

where q_n is a constant quadratic in n and $\| A \|_\infty = \max_i \sum_j |a_{ij}|$.

See [3] for details. Since rounding errors of order $u \| A \|_\infty$ are usually present in A to begin with, it can be concluded from (1.4) and (1.6) that pivoting is unnecessary when solving <u>symmetric</u> positive definite linear systems.

Is pivoting necessary when A is positive definite but unsymmetric? Buckley [1,2] considered this question for positive definite systems of the form

$$(1.7) \qquad (I + S)x = b \qquad S^T = -S$$

These systems arise from the discretization of certain differential operators. It turns out that if

$$(I + S) = LU$$

is the LU factorization of $(I + S)$, then the pivots u_{kk} satisfy

$$(1.8) \qquad 1 \leq u_{kk} \leq 1 + \| S \|_2^2$$

where the 2-norm of a matrix Z is defined by

$$\| Z \|_2^2 = \text{the largest eigenvalue of } Z^T Z$$

Since the pivots are nicely bounded away from zero, Buckley concluded that it is safe not to pivot when applying Gaussian elimination to (1.7).

This conclusion has to be qualified for general positive definite systems as the example

$$A = \begin{bmatrix} \varepsilon & 1 \\ -1 & \varepsilon \end{bmatrix} = \begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad (1 >> \varepsilon > 0)$$

indicates. This matrix has a perfect condition number in the 2-norm,

$$\| A \|_2 \, \| A^{-1} \|_2 = (\varepsilon^2 + 1)^{1/2} \, (\varepsilon^2 + 1)^{-1/2} = 1$$

and therefore, if Gaussian elimination with partial pivoting is used to solve $Ax = b$, the solution will be correct to nearly full working precision. However, if the solution is obtained via the factorization

$$\begin{bmatrix} \varepsilon & 1 \\ -1 & \varepsilon \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/\varepsilon & 1 \end{bmatrix} \begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon - 1/\varepsilon \end{bmatrix} \begin{bmatrix} 1 & 1/\varepsilon \\ 0 & 1 \end{bmatrix}$$

then rounding errors of order u/ε can be expected to contaminate the result.

The problem, of course, is the presence of large entries in L , D , and M . In the next section we show that this can only occur if $\| S\, T^{-1}\, S \|_2$ is large where the matrices T and S are given by (1.1) and (1.2) respectively. Notice that in the above example, this quantity has the value of $1/\varepsilon$.

2. Theoretical Bounds

In this section we put finite precision arithmetic aside and examine how big the factors L, D, and M can get when $A = LDM^T$ is positive definite. Ignoring the distinction between exact and computed quantities, we conclude from (1.5) that a bound on $|L||D||M^T|$ is what we need. For this purpose we shall use the Frobenius norm:

$$\| Z \|_F^2 = \sum_i \sum_j |z_{ij}|^2$$

Theorem

Let A be an $n \times n$ positive definite matrix having the factorization

$$A = L D M^T$$

where L and M are unit lower triangular and $D = \mathrm{diag}(d_1, \ldots, d_n)$. If $T = (A + A^T)/2$ and $S = (A - A^T)/2$, then

$$(2.1) \qquad \| \, |L||D||M^T| \, \|_F \leq n \, [\, \| T \|_2 + \| S T^{-1} S \|_2 \,]$$

Proof

Let the Cholesky decomposition of the symmetric positive definite matrix T be given by

$$T = G G^T \qquad\qquad G \text{ lower triangular}$$

and partition L^{-1} and M^{-1} into their respective rows as follows:

$$L^{-1} = \begin{bmatrix} y_1^T \\ \vdots \\ y_n^T \end{bmatrix} , \qquad M^{-1} = \begin{bmatrix} z_1^T \\ \vdots \\ z_n^T \end{bmatrix}$$

From the equation $L\,D\,M^T = G\,G^T + S$ we find

$$D\,M^T L^{-T} = (G^T L^{-T})^T (G^T L^{-T}) + L^{-1} S\, L^{-T}$$

where L^{-T} denotes the inverse of the transpose of L . Since $M^T L^{-T}$ is unit upper triangular and since the diagonal of the skew-symmetric matrix $L^{-1} S\, L^{-T}$ is zero, it follows by comparing (k,k) entries that

$$(2.2) \qquad d_k = \| G^T y_k \|_2^2 \qquad k = 1,\dots,n \; .$$

Thus, if $D^{1/2} = \mathrm{diag}(\sqrt{d_1}\,,\dots,\sqrt{d_n}\,)$ we obtain

$$\| G^T L^{-T} D^{-1/2} \|_F^2 = \sum_{k=1}^{n} \left(\frac{\| G^T y_k \|_2}{\sqrt{d_k}} \right)^2 = n$$

Similarly,

$$M^{-1} L\,D = (G^T M^{-T})^T (G^T M^{-T}) + M^{-1} S\, M^{-T}$$

implies

$$\| G^T M^{-T} D^{-1/2} \|_F^2 = n$$

Now since $(L\ D^{1/2})\ (D^{1/2}\ M^T) = G\ G^T + S$, we have the identities

$$L\ D^{1/2} = (G + SG^{-T})\ (G^T\ M^{-T}\ D^{-1/2})$$

and

$$D^{1/2} M^T = (G^T\ L^{-T}\ D^{-1/2})^T\ (G^T + G^{-1}S)$$

Taking norms and using the inequalities $\| YZ \|_F \leq \| Y \|_F\ \| Z \|_2$ and $\|YZ\|_F \leq \|Y\|_2\ \|Z\|_F$ we have

$$\| L\ D^{1/2} \|_F \leq \| G + SG^{-T} \|_2\ \| G^T M^{-T}\ D^{-1/2} \|_F \leq \sqrt{n}\ \| G + SG^{-T} \|_2$$

and

$$\| D^{1/2}\ M^T \|_F \leq \| (G^T L^{-T}\ D^{-1/2})^T \|_F\ \| G^T + G^{-1}S \|_2 \leq \sqrt{n}\ \| G^T + G^{-1}S \|_2$$

which imply

$$\begin{aligned} \|\ |L|\,|D|\,|M^T|\ \|_F &= \|\ |L\ D^{1/2}|\ |D^{1/2}\ M^T|\ \|_F \\ &\leq \| L\ D^{1/2} \|_F\ \| D^{1/2}\ M^T \|_F \\ &\leq n\ \| G + SG^{-T} \|_2\ \| G^T + G^{-1}S \|_2 \end{aligned}$$

The theorem now follows by using the identities $\| Z \|_2^2 = \| ZZ^T \|_2 = \| Z^T Z \|_2$ to prove

$$(2.3) \qquad \| G + S G^{-T} \|_2^2 = \| (G + S G^{-T})(G + S G^{-T})^T \|_2$$

$$= \| (G + S G^{-T})(G^T - G^{-1} S) \|_2$$

$$= \| G G^T - S G^{-T} G^{-1} S \|_2$$

$$= \| T - S T^{-1} S \|_2$$

$$\leq \| T \|_2 + \| S T^{-1} S \|_2$$

and similarly

$$(2.4) \qquad \| G^T + G^{-1} S \|_2^2 \leq \| T \|_2 + \| S T^{-1} S \|_2$$

What this theorem implies for positive definite linear system solvers will be discussed in the next section. However, as affirmed by our earlier 2x2 example, it is clear that a large $\| S T^{-1} S \|_2$ may imply unacceptably large $|L||D||M^T|$ and a consequent need for pivoting.

We conclude this section by generalizing Buckley's result (1.8) concerning the pivots d_k (his u_{kk}) .

Corollary

$$(2.5)\qquad \frac{1}{\| T^{-1} \|_2} \leqslant d_k \leqslant \| T \|_2 + \| S T^{-1} S \|_2 \qquad (k = 1,...$$

Proof

Since L^{-T} is a unit upper triangular matrix, its k-th colum y_k, has a 2-norm of at least unity. Hence, from (2.2) we obtain

$$d_k = \| G^T y_k \|_2^2 = y_k^T T y_k \geqslant \frac{\| y_k \|_2^2}{\| T^{-1} \|_2} \geqslant \frac{1}{\| T^{-1} \|_2}$$

To establish the upper bound in (2.5), we once again use the identity

$$L D^{1/2} = (G G^T + S) M^{-T} D^{-1/2} = (G + SG^{-T})(G^T M^{-T} D^{-1/2})$$

Let e_k denote the k-th column of the identity. Using (2.2) and (2.3) we find

$$\begin{aligned} d_k &\leqslant \| (LD^{1/2}) e_k \|_2^2 = \| (G + SG^{-T})(G^T M^{-T} D^{-1/2} e_k) \|_2^2 \\ &\leqslant \| G + SG^{-T} \|_2^2 \, \| (G^T M^{-T}) e_k \|_2^2 / d_k \\ &\leqslant \| T \|_2 + \| S T^{-1} S \|_2 \end{aligned}$$

$\blacksquare$

3. Practical Observations

If we assume that

(3.1) $$\| \, |\hat{L}| \, |\hat{D}| \, |\hat{M}^T| \, \|_F \doteq \| \, |L| \, |D| \, |M^T| \, \|_F$$

then (1.5), (2.1), and the inequality $\| T \|_2 \leq \| A \|_2$ imply that the computed solution $\hat{x}$ satisfies

(3.2) $$(A + E)\hat{x} = b$$

where

(3.3) $$\| E \|_F \lessapprox u \, n \, c_n \, [\, \| A \|_2 + \| S \, T^{-1} \, S \|_2 \,]$$

The symbols " $\doteq$ " and " $\lessapprox$ " remind us that these are heuristic results. By setting $S = 0$ in (3.3) we "essentially" obtain the same bound that Wilkinson [7] derived in his classical roundoff error analysis of the Cholesky factorization $A = GG^T$.

Now from the usual sensitivity analysis of linear systems [6,p.194ff.] we know that if $Ax = b$ and $(A + F)y = b$, then roughly speaking

$$\frac{\| x - y \|_2}{\| x \|_2} \lessapprox \frac{\| F \|_2}{\| A \|_2} \kappa_2(A)$$

where

$$\kappa_2(A) = \| A \|_2 \, \| A^{-1} \|_2$$

is the condition of A. Hence, from (3.2) and (3.3) we obtain the key result

$$(3.4) \qquad \frac{\| x - \hat{x} \|_2}{\| x \|_2} \lesssim u\, n\, c_n \left[1 + \frac{\| S T^{-1} S \|_2}{\| A \|_2} \right] \kappa_2(A)$$

On the other hand, the corresponding result for $\hat{x}_{piv}$, the computed solution obtained via Gaussian elimination with pivoting, is roughly of the form

$$(3.5) \qquad \frac{\| x - \hat{x}_{piv} \|_2}{\| x \|_2} \lesssim u\, \rho\, n^3\, \kappa_2(A)$$

where ρ is the "growth factor". Assuming $\rho \doteq 1$ and ignoring the unimportant factors nc_n and n^3, we see that the main distinction between (3.4) and (3.5) is the factor

$$\Omega = \frac{\| S T^{-1} S \|_2}{\| A \|_2}$$

Based on dozens of "randomly generated" examples, our experience has been that it is safe not to pivot in problems having a small Ω. In other words, heuristic (3.4) can be trusted in practice. However, if Ω is large, then it is more difficult to draw conclusions from (3.4). A large Ω may certainly imply a need for pivoting as the 2x2 example of Section 1 indicates. But to show that an accurate x can be obtained despite a large Ω, we consider the linear system

$$(3.6) \qquad Ax = (GG^T + S)x = b \quad .$$

where

$$G = (g_{ij}) \qquad g_{ij} = \begin{cases} 0 & i < j \\ 1 & i = j \\ -1 & i > j \end{cases} ,$$

$$S = (s_{ij}) \qquad s_{ij} = \begin{cases} -10 \times (-1)^{j + i(i-1)/2} & i < j \\ 0 & i = j \\ 10 \times (-1)^{j + i(i-1)/2} & i > j \end{cases} ,$$

d b is chosen such that the exact solution is given by $= (1,1,\ldots,1)^T$. The system has the property that

$$\Omega \doteq O(4^n)$$

d so by increasing the dimension n we can arbitrarily increase e bound in (3.4). Using IBM/370 short precision arithmetic ($u = 10^{-6}$) solved (3.6) for $n = 5,10,15,20$, and 25. The following table mmarizes our findings:

(b) $\frac{\lVert x - \hat{x} \rVert_2}{\lVert x \rVert_2}$	(c) $u\,\Omega\,\kappa_2(A)$	(d) $u\,\kappa_2(A)$	(e) $\lVert \lvert\hat{L}\rvert \lvert\hat{D}\rvert \lvert\hat{M}^T\rvert \rVert_F$	(f) Range of $\hat{d}_k$
4×10^{-6}	1×10^{-2}	2×10^{-5}	8×10^{2}	$1 - 10^{2}$
1×10^{-5}	2×10^{1}	3×10^{-5}	2×10^{5}	$1 - 10^{2}$
2×10^{-5}	3×10^{4}	5×10^{-5}	3×10^{5}	$1 - 10^{2}$
2×10^{-5}	4×10^{7}	7×10^{-5}	5×10^{5}	$1 - 10^{2}$
4×10^{-5}	6×10^{10}	1×10^{-4}	8×10^{5}	$1 - 10^{2}$

We tabulated $u\Omega\,\kappa_2(A)$ and $u\,\kappa_2(A)$ since these are the key factors in the upper bounds of (3.4) and (3.5) respectively. It is clear from columns (b) and (c) in the table that the upper bound in (3.4) is extremely pessimistic. For all the values of n selected, the matrix A is fairly well conditioned as evidenced by column (d) which indicates the error which can be expected when Gaussian elimination with partial pivoting is applied to (3.6). There is now way we know how to explain why such good solutions were obtained without pivoting in view of the largeness of Ω and the numbers in column (e). Perhaps a more refined analysis would reveal that the size of the pivots has a critical bearing on the relative error, for as we see in these examples, the d_k are rather nicely behaved.

As a final example, we demonstrate how our analysis can be applied to the linear system which arises from the discretization of the boundary value problem

$$y''(x) = p(x)y'(x) + q(x)y(x) + r(x) \quad (-1 \leq x \leq 1).$$

$$y(-1) = \alpha \quad , \quad y(1) = \beta \; , \; q(x) \geq 0$$

Although this is a simple problem, it illustrates how an estimate of Ω can be obtained from information relating to the underlying differential equation.

Set $h = 2/(n+1)$ and $x_i = a + ih$, $i = 0,\ldots,n+1$. Following rtega [5,p.96ff.], if we replace the above derivatives with appro-riate divided differences, then we are led to a linear system $y = d$ in the unknowns $y_i \doteq y(x_i)$ $(i = 1,\ldots,n)$ where

$$A = \begin{bmatrix} a_1 & -c_1 & & & \bigcirc \\ -b_2 & a_2 & -c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -b_{n-1} & a_{n-1} & -c_{n-1} \\ \bigcirc & & & b_n & a_n \end{bmatrix}$$

nd

$$a_i = 2 + h^2 q_i \qquad\qquad q_i = q(x_i).$$

$$b_i = 1 + p_i h/2 \qquad\qquad p_i = p(x_i)$$

$$c_i = 1 - p_i h/2$$

he symmetric matrix $T = (A + A^T)/2$ has the form $T = T_1 + T_2$ here

$$T_1 = \begin{bmatrix} a_1 & -1 & & \bigcirc \\ -1 & a_2 & -1 & \\ & \ddots & \ddots & -1 \\ \bigcirc & & -1 & a_n \end{bmatrix}$$

and

$$T_2 = \begin{bmatrix} 0 & w_2 & & \bigcirc \\ w_2 & 0 & w_3 & \\ & \ddots & \ddots & w_n \\ \bigcirc & & w_n & 0 \end{bmatrix} \qquad w_i = \frac{-h}{4}(p_i - p_{i-1}).$$

while the skew-symmetric matrix $S = (A - A^T)/2$ has the form

$$S = \begin{bmatrix} 0 & s_2 & & & \\ -s_2 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 0 & s_n \\ & & & -s_n & 0 \end{bmatrix}. \qquad s_i = (p_{i-1} + p_i)h/2$$

Now if we assume that $p'(x)$ is continuous on $[-1,1]$ and that

$$\max_{-1 \le x \le 1} |p'(x)| = p'_{max} \le 2 \tag{3.7}$$

then

$$|w_i| = \frac{h^2}{4} \left| \frac{p_i - p_{i-1}}{h} \right| \le \frac{h^2}{4} p'_{max} \le \frac{h^2}{2}$$

and thus

$$\| T_2 \|_2 \le h^2$$

Since the smallest eigenvalue of T_1 is bounded below,

$$\lambda_{min}(T_1) \ge 2 - 2\cos\left(\frac{\pi h}{2}\right) \ge 2h^2 \qquad (h \le 1) ,$$

we obtain the following lower bound for the smallest eigenvalue of T:

$$\lambda_{min}(T) \ge \lambda_{min}(T_1) - \| T_2 \|_2 \ge 2h^2 - h^2 = h^2$$

Thus, $T = (A + A^T)/2$ is positive definite and moreover, (3.8) $$\|T^{-1}\|_2 \leq 1/h^2$$

Setting

$$p_{max} = \max_{-1 \leq x \leq 1} |p(x)|$$

it is easy to verify that

$$\|S\|_2 \leq 2h \; p_{max}$$

and therefore

$$\Omega \leq \frac{\|S\,T^{-1}\,S\|_2}{\|A\|_2} \leq \|S\|_2^2 \; \|T^{-1}\|_2 \leq (2\,p_{max})^2$$

Thus, if the function $p(x)$ is both smooth enough and small enough, the matrix A is positive definite and pivoting unnecessary.

Acknowledgements

The authors are grateful to Professors Jim Ortega and Olof Widlund for sharing their thoughts on the problem discussed in this paper.

REFERENCES

[1] A.Buckley , A Note on Matrices A = I+H, H Skew-Symmetric, Z.Angew. Math.Mech. 54(1974),#2,125-126.

[2] A.Buckley , On the Solution of Certain Skew-Symmetric Linear Systems, SIAM J.Num.Anal. 14(1977),566-570.

[3] C.deBoor and A.Pinkus, Backward Error Analysis for Totally Positive Linear Systems , Numer.Math. 27(1977),485-490.

[4] K.Fan , On Real Matrices with Positive Definite Symmetric Component , Linear and Multilinear Algebra 1(1973),1-4 .

[5] J.Ortega, Numerical Analysis - A Second Course , Academic Press, New York, 1972.

[6] G.W.Stewart, Introduction to Matrix Computations , Academic Press, New York, 1973.

[7] J.H.Wilkinson, A Priori Error Analysis of Algebraic Processes , Proc. Proc.International Congress Math.(Moscow, 1968), pp.629-639, Moscow: Izdat.Mir. 1968.