APPENDIX II

SOLUTION OF LINEAR EQUATIONS

This appendix is addressed to those who are uninitiated in numerical analysis. It is included here because of its usefulness in Sections 12 and 18. We consider a linear system

$$(1)\qquad \begin{array}{ccccc} a_{1,1}x(1) & + \dots + & a_{1,n}x(n) & = & y(1) \\ \vdots & & \vdots & & \vdots \\ a_{m,1}x(1) & + \dots + & a_{m,n}x(n) & = & y(m) \end{array}$$

of m equations in n unknowns $x(1),\dots,x(n)$. Let $A = [a_{i,j}]$, $\underset{\sim}{x} = [x(1),\dots,x(n)]^t$ and $\underset{\sim}{y} = [y(1),\dots,y(m)]^t$. Given $\underset{\sim}{y} \in \mathbb{C}^n$, the problem is to find $\underset{\sim}{x} \in \mathbb{C}^n$ such that $A\underset{\sim}{x} = \underset{\sim}{y}$. The $m \times n$ matrix A is called the <u>coefficient matrix</u> of the system (1); it induces a linear map from $\mathbb{C}^n$ to $\mathbb{C}^m$ in a natural way, which we denote also by A. The range of A is the linear span of its columns $\underset{\sim}{a}_1,\dots,\underset{\sim}{a}_n$, where $\underset{\sim}{a}_j(i) = a_{i,j}$. The rank r of A is the maximum number of linearly independent columns of A. Clearly, $r \leq \max\{m,n\}$. When $m = n$, A is invertible if and only if $r = n$ if and only if $\det A \neq 0$; such a matrix is called <u>nonsingular</u>.

Let $\underset{\sim}{e}_k = [0,\dots,0,1,0,\dots,0]^t$, where 1 occurs in the k-th place, and let I_n denote the $n \times n$ identity matrix. An $n \times n$ matrix A is called <u>elementary</u> if $A = I_n + B$, where B has rank 1. Then

$$A = I_n - \underset{\sim}{x}\underset{\sim}{y}^H$$

for some $\underset{\sim}{x}$ and $\underset{\sim}{y}$ in $\mathbb{C}^n$.

We say that an $m \times n$ matrix A is <u>lower</u> (resp., <u>upper</u>) <u>trapezoidal</u> if $a_{i,j} = 0$, whenever $i < j$ (resp., $i > j$); when $m = n$, it is called <u>lower</u> (resp., <u>upper</u>) <u>triangular.</u>

If A is lower trapezoidal, then one can attempt to solve the system (1) by *forward elimination*, and if A is upper trapezoidal then by *back substitution*, in the most natural way. For this reason, trapezoidal matrices are very important in solving linear systems. We shall describe how elementary matrices can be used to reduce a given matrix A to a product of a lower triangular or unitary matrix and an upper trapezoidal matrix.

Gaussian elimination and LR factorization

For $\underset{\sim}{u} = [u(1),\dots,u(m)]^t \in \mathbb{C}^m$ with $u(k) \neq 0$, consider

$$\underset{\sim}{x} = \left[0,\dots,0\ ,\ \frac{u(k+1)}{u(k)}\ ,\ \dots\ ,\ \frac{u(m)}{u(k)}\right]^t .$$

Then the matrix

$$G = I_m - \underset{\sim}{x}\underset{\sim}{e}_k^H \tag{2}$$

is called a <u>Gauss</u> <u>matrix</u>. Note that G is an elementary matrix, and for $\underset{\sim}{y} \in \mathbb{C}^m$, $G\underset{\sim}{y} = \underset{\sim}{y} - y(k)\underset{\sim}{x}$. In particular, $G\underset{\sim}{u} = [u(1),\dots,u(k),\ 0,\dots,0]^t$. Thus, multiplication by a Gauss matrix on the left introduces zeros below a given nonzero entry of a vector. We observe that a Gauss matrix (and hence a product of two Gauss matrices) is a lower triangular matrix with 1's on the diagonal.

<u>Gaussian</u> <u>elimination</u> consists of introducing zeros below the diagonal elements of a matrix A by successively multiplying A on the left by Gauss matrices. In order to achieve this, the k-th diagonal element, called a <u>pivot</u>, must be nonzero at the k-th step. We have the following result.

THEOREM 1 Let the leading $k \times k$ principal submatrix of an $m \times n$ matrix A be nonsingular for $k = 1,\ldots,\min\{m-1,n\}$. Then $A = LR$, where L is a $m \times m$ lower triangular matrix with 1's on the diagonal, and R is an $m \times n$ upper trapezoidal matrix; if, in particular, $m = n$ and A is nonsingular, then L and R are unique.

The factors L and R can be computed by employing Algorithm 4.2-1 of [GV]. However, this procedure can fail on simple-looking matrices like $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$, if a leading principal submatrix is singular. To take care of such a situation (and also to achieve stability in case the pivots are nonzero but small), one can interchange the rows of A.

An $m \times m$ matrix P is called a <u>permutation</u> <u>matrix</u> if $P = [\underset{\sim}{e}_{\pi(1)},\ldots,\underset{\sim}{e}_{\pi(m)}]$, where $\pi : \{1,\ldots,m\} \to \{1,\ldots,m\}$ is a one to one map, i.e., a permutation of $\{1,\ldots,m\}$; P is obtained by permuting the columns of I_m according to π. Note that P is unitary, i.e., $P^H P = I = PP^H$, so that $P^{-1} = P^H$ corresponds to the permutation π^{-1}. If A is an $m \times n$ matrix, then the matrix PA is obtained from A by permuting its rows according to the permutation π^{-1}.

In <u>Gaussian</u> <u>elimination</u> <u>with</u> <u>partial</u> <u>pivoting</u>, one performs Gaussian elimination process with the following interchange of rows. Suppose the matrix A $(= A^{(0)})$ is reduced to a matrix

$$(3) \qquad A^{(k-1)} = \begin{bmatrix} A_{1,1}^{(k-1)} & A_{1,2}^{(k-1)} \\ 0 & A_{2,2}^{(k-1)} \end{bmatrix}$$

at the $(k-1)$st step. Then search the first column of $A_{2,2}^{(k-1)}$ for an entry with the largest absolute value, say j-th entry, and then swap the k^{th} and the j^{th} rows of $A^{(k-1)}$. We have the following result.

THEOREM 2 Let the rank of an $m \times n$ matrix A be r. Then $PA = LR$, where P is an $m \times m$ permutation matrix, L is an $m \times m$ lower triangular matrix with 1's on the diagonal and R is upper trapezoidal:

$$(4) \qquad R = \begin{array}{c} \left[\begin{array}{cc} R_{1,1} & R_{1,2} \\ 0 & 0 \end{array}\right] \begin{array}{l} r \\ m-r \end{array} \\ \begin{array}{cc} r & \quad n-r \end{array} \end{array} ,$$

where $R_{1,1}$ is an upper triangular matrix with nonzero diagonal entries.

Because the choice of an entry with the largest absolute value in the first column of $A^{(k-1)}$ is not unique, $k = 1,2,\ldots,$ the factorization in the above result is not unique. It can be computed by using Algorithm 4.4-2 of [GV], which requires $mns - (m+n)s^2/2 + s^3/3$ flops, where $s = \min\{m-1,r\}$, and $O(ms)$ comparisons. A <u>flop</u> is basically one floating-point multiplication and addition of subscripted arguments.

Having found the factorization $PA = LR$, we proceed to solve $A\underset{\sim}{x} = \underset{\sim}{y}$ as follows. Since L is lower triangular and invertible, we can find $\underset{\sim}{z} \in \mathbb{C}^m$ such that $L\underset{\sim}{z} = P\underset{\sim}{y}$ by forward elimination. Next, we consider the equation

$$R\underset{\sim}{u} = \begin{bmatrix} R_{1,1} & R_{1,2} \\ 0 & 0 \end{bmatrix} \underset{\sim}{u} = \underset{\sim}{z} .$$

It has a solution only if the last $(m-r)$ entries of $\underset{\sim}{z}$ are zero. In case $\underset{\sim}{z} = \underset{\sim}{0}$, there are $(n-r)$ linearly independent solutions. The solutions can be obtained by backward substitution since $R_{1,1}$ is upper triangular and invertible.

Cholesky factorization and least squares problem

If A is an $n \times n$ positive (definite) matrix, i.e., $\langle A\underset{\sim}{x},\underset{\sim}{x}\rangle = \underset{\sim}{x}^H A\underset{\sim}{x} > 0$ for every $\underset{\sim}{x} \neq 0$, then all the leading principle submatrices of A are nonsingular, and we have

$$A = LR$$

where L is an $n \times n$ lower triangular matrix with 1's on the diagonal and R is an $n \times n$ upper triangular matrix with positive diagonal entries $d_1, \ldots, d_n$. Now,

$$A = L\mathrm{diag}(d_1, \ldots, d_n)\mathrm{diag}(d_1^{-1}, \ldots, d_n^{-1})R ,$$

and since A is self-adjoint,

$$A = A^H = R^H\mathrm{diag}(d_1^{-1}, \ldots, d_n^{-1})\mathrm{diag}(d_1, \ldots, d_n)L^H .$$

But $R^H\mathrm{diag}(d_1^{-1}, \ldots, d_n^{-1})$ is lower triangular with 1's on the diagonal, and $\mathrm{diag}(d_1, \ldots, d_n)L^H$ is upper triangular. Hence by the uniqueness part of Theorem 1, we have

$$R = \mathrm{diag}(d_1, \ldots, d_n)L^H .$$

If we let $G = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})L$, then

$$A = LR = L\mathrm{diag}(d_1, \ldots, d_n)L^H = GG^H .$$

This is called the <u>Cholesky factorization of a positive (definite) matrix</u> A. To implement it, one can use Algorithm 5.2-1 of [GV], which requires $n^3/3$ flops (and no comparisons).

Coming back to a general $m \times n$ matrix A, we note that the $n \times n$ matrix $B = A^H A$ satisfies $\langle Bx, x\rangle = \langle Ax, Ax\rangle \geq 0$ for all $x \in X$, and if A is one to one, then B is, in fact, positive (definite). Now, A is one to one if and only if rank $A = n$. Consider, then, an $m \times n$ matrix A with rank n (in particular, $m \geq n$), and let $\underset{\sim}{y} \in \mathbb{C}^m$. If there is an $\underset{\sim}{x} \in \mathbb{C}^n$ such that $A\underset{\sim}{x} = \underset{\sim}{y}$, then it can be found as follows. Form $B = A^H A$, $\underset{\sim}{z} = A^H \underset{\sim}{y}$, and solve the so-called <u>normal equations</u> $B\underset{\sim}{x} = \underset{\sim}{z}$, yielding

$$\underset{\sim}{x} = B^{-1}\underset{\sim}{z} = (A^H A)^{-1} A^H \underset{\sim}{y} .$$

See Algorithm 6.1-1 of [GV]. This algorithm requires $\frac{n^2}{2}(m + \frac{n}{3})$ flops.

Even when there is no $\underset{\sim}{x} \in \mathbb{C}^n$ such that $A\underset{\sim}{x} = \underset{\sim}{y}$, the vector

$$\underset{\sim}{y}' = A(A^H A)^{-1} A^H \underset{\sim}{y}$$

is the *best approximation* to $\underset{\sim}{y}$ from the range of A . This follows by noting that for every $\underset{\sim}{u} \in \mathbb{C}^n$,

$$\langle \underset{\sim}{y}', A\underset{\sim}{u} \rangle = \langle A^H \underset{\sim}{y}', \underset{\sim}{u} \rangle = \langle A^H \underset{\sim}{y}, \underset{\sim}{u} \rangle = \langle \underset{\sim}{y}, A\underset{\sim}{u} \rangle ,$$

so that $\underset{\sim}{y}' - \underset{\sim}{y}$ is orthogonal to the range of A . (See [L], 23.2.) In other words, the vector

$$\underset{\sim}{x} = (A^H A)^{-1} A^H \underset{\sim}{y} = A^\dagger \underset{\sim}{y} \text{ , say}$$

is the (unique) solution of the <u>least squares problem</u>

(5) ' Find $\underset{\sim}{x} \in \mathbb{C}^n$ such that $\min_{\underset{\sim}{u} \in \mathbb{C}^n} \|A\underset{\sim}{u} - \underset{\sim}{y}\|_2 = \|A\underset{\sim}{x} - \underset{\sim}{y}\|_2$ '.

For this reason, $\underset{\sim}{x} = A^\dagger \underset{\sim}{y}$ is called the <u>least squares solution</u> of $A\underset{\sim}{u} = \underset{\sim}{y}$; the operator (or the matrix)

$$A^\dagger = (A^H A)^{-1} A^H : \mathbb{C}^m \to \mathbb{C}^n$$

is called the <u>Moore-Penrose inverse</u> of the $m \times n$ matrix A of rank n . Note that it satisfies the four <u>Penrose equations</u>: $AA^\dagger A = A$, $A^\dagger A A^\dagger = A^\dagger$, $(AA^\dagger)^H = AA^\dagger$ and $(A^\dagger A)^H = A^\dagger A$. In particular, if $\underset{\sim}{y}$ belongs to the range of A and $\underset{\sim}{y} = A\underset{\sim}{x}$, then $A^\dagger \underset{\sim}{y} = (A^H A)^{-1} A^H A \underset{\sim}{x} = \underset{\sim}{x}$; i.e., the least squares solution of $A\underset{\sim}{u} = \underset{\sim}{y}$ is, in fact, the solution of $A\underset{\sim}{u} = \underset{\sim}{y}$. In case $m = n$, $A^\dagger = A^{-1}$.

Householder method and QR factorization

There is an alternative, and perhaps better, way of finding the least squares solution. To describe it, we go back to the upper triangularization of an $m \times n$ matrix A .

For $\underset{\sim}{u} \in \mathbb{C}^m$, consider

$$\underset{\sim}{x} = \begin{cases} \underset{\sim}{0} \, , & \text{if } \underset{\sim}{u} = \underset{\sim}{0} \\ 2\underset{\sim}{u}/\underset{\sim}{u}^H\underset{\sim}{u} \, , & \text{if } \underset{\sim}{u} \neq \underset{\sim}{0} \end{cases} .$$

Then the matrix

$$H = I_m - \underset{\sim}{x}\underset{\sim}{u}^H \tag{6}$$

is called a <u>Householder matrix.</u> Note that H is an elementary matrix, and $H\underset{\sim}{u} = -\underset{\sim}{u}$, while $H\underset{\sim}{v} = \underset{\sim}{u}$ if $\underset{\sim}{u}^H\underset{\sim}{v} = 0$; H is called the <u>reflector</u> <u>which reverses</u> $\underset{\sim}{u}$. Observe that H is self-adjoint as well as unitary. Given $\underset{\sim}{a} \in \mathbb{C}^m$, let $\underset{\sim}{u} = \underset{\sim}{a} - \|\underset{\sim}{a}\|_2\underset{\sim}{e}_1$, and H be the reflector which reverses $\underset{\sim}{u}$. Then $H\underset{\sim}{a} = [\|\underset{\sim}{a}\|_2, 0, \ldots, 0]^t$. Thus, like a Gauss matrix, a Householder matrix can be used to introduce zeros in all the entries of a vector except possibly the first.

Let $A = [\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_n]$ be an $m \times n$ matrix. Let $\underset{\sim}{a}_{k_1}$ be the first nonzero column of A . Find a reflector H_1 , as above, such that

$$H_1\underset{\sim}{a}_{k_1} = [\|\underset{\sim}{a}_{k_1}\|_2, 0, \ldots, 0]^t .$$

If

$$H_1A = \begin{bmatrix} 0 \ldots 0 & \|\underset{\sim}{a}_{k_1}\|_2 & * \ldots * \\ . \quad\quad . & 0 & \\ . \quad\quad . & . & A^{(2)} \\ . \quad\quad . & . & \\ 0 \ldots 0 & 0 & \end{bmatrix} ,$$

we repeat the process for the $(m-1) \times (n-k_1)$ matrix $A^{(2)}$ to find an appropriate reflector $\tilde{H}_2$, which is an $(m-1) \times (m-1)$ unitary matrix. Let $H_2 = \mathrm{diag}(I_1, \tilde{H}_2)$, which is also unitary. Continuing this process, we find that

$$(7) \qquad H_r \ldots H_1A = \tilde{R} = \begin{array}{cc} \left[\begin{array}{c} R \\ 0 \end{array}\right] & \begin{array}{c} r \\ m-r \end{array} \\ n & \end{array} ,$$

where r is the rank of A , and $\tilde{R}$ is upper trapezoidal with nonnegative diagonal entries. Letting $U = H_1 \ldots H_r$, we obtain the following result.

THEOREM 3 Let the rank of an $m \times n$ matrix A be r . Then $A = U\tilde{R}$, where U is an $m \times m$ unitary matrix and $\tilde{R}$ is an $m \times n$ upper trapezoidal matrix.

Having found the factorization $A = U\tilde{R}$, we proceed to determine the solutions of $A\underset{\sim}{u} = \underset{\sim}{y}$ as follows. Since U is unitary, we find $\underset{\sim}{z} \in \mathbb{C}^m$ such that $U\underset{\sim}{z} = \underset{\sim}{y}$ by letting $\underset{\sim}{z} = U^H\underset{\sim}{y}$. Next, we consider the equation

$$\tilde{R}\underset{\sim}{u} = \left[\begin{array}{c} R \\ 0 \end{array}\right]\underset{\sim}{u} = \underset{\sim}{z} .$$

It has a solution only if the last $(m-r)$ entries of $\underset{\sim}{z}$ are zero. In case $\underset{\sim}{z} = \underset{\sim}{0}$, there are $(n-r)$ linearly independent solutions. The solutions can be obtained by backward substitution since R is upper trapezoidal.

Let us consider the case when A has rank n , i.e., $r = n \leq m$. In this case the $n \times n$ upper triangular matrix R is invertible. Let $\underset{\sim}{y} \in \mathbb{C}^m$, and $\underset{\sim}{z} = U^H\underset{\sim}{y} = [c(1),\ldots,c(n),d(1),\ldots, d(m-n)]^t$. Then for every $\underset{\sim}{u} \in \mathbb{C}^n$, we have

$$\|A\underset{\sim}{u} - \underset{\sim}{y}\|_2^2 = \|U^HA\underset{\sim}{u} - U^H\underset{\sim}{y}\|_2^2 = \|R\underset{\sim}{u} - \underset{\sim}{c}\|_2^2 + \|\underset{\sim}{d}\|_2^2 .$$

It is then clear that the quantity $\|A\underset{\sim}{u}-\underset{\sim}{y}\|_2$ is minimized exactly when $R\underset{\sim}{u} = \underset{\sim}{c}$, i.e., $\underset{\sim}{u} = R^{-1}\underset{\sim}{c}$. Thus, the least squares solution of $A\underset{\sim}{u} = \underset{\sim}{y}$ is given by $\underset{\sim}{x} = R^{-1}\underset{\sim}{c}$, where the n-column vector $\underset{\sim}{c}$ consists of the first n entries of $\underset{\sim}{z} = U^H\underset{\sim}{y}$. In other words, $A^\dagger\underset{\sim}{y} = R^{-1}\underset{\sim}{c}$. This is known as the Householder orthogonalization method for finding the least squares solution of an $m \times n$ linear system of rank n . It requires $n^2(m - \frac{n}{3})$ flops.

While on the subject of factoring an $m \times n$ matrix A as a product of a unitary and an upper trapezoidal matrix, we state the following result.

THEOREM 4 (QR factorization) Let A be an $m \times n$ matrix of rank r . Then $A = QR$ where Q is an $m \times r$ matrix satisfying $Q^HQ = I_r$ and R is an $r \times n$ upper trapezoidal matrix with nonnegative diagonal entries; Q and R are unique. If $r = n$, then R is upper triangular with positive diagonal entries.

The proof of the existence is immediate since $A = U\tilde{R} = U\begin{bmatrix} R \\ 0 \end{bmatrix}$ and we can let $Q = UI_{m,r}$, where $I_{m,r}$ consists of the first r columns of I_m . Note that if $A = QR$, then the r columns of Q form an orthonormal set in $\mathbb{C}^m$ since $Q^HQ = I_r$. They are obtained by successively orthonormalizing the linearly independent columns of A in the order $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ by the Gram-Schmidt process. The fact that the diagonal entries of R are nonnegative then gives the uniqueness of this construction.

Perturbation of the solution

We now consider the sensitivity of the solution $\underset{\sim}{x}$ of $A\underset{\sim}{u} = \underset{\sim}{y}$ to the changes in the coefficient matrix A and in the given right hand side $\underset{\sim}{y}$.

First, let $m = n$ and the matrix A be nonsingular. By (9.1) and Problem (9.1) , we obtain the following result.

THEOREM 5 Let A and $\hat{A}$ be $n \times n$ matrices, and A be nonsingular. If $\|A^{-1}(\hat{A}-A)\| < 1$, then $\hat{A}$ is also nonsingular. Also, if $\underset{\sim}{y}$, $\underset{\sim}{x}$, $\hat{\underset{\sim}{y}}$ and $\hat{\underset{\sim}{x}}$ are in $\mathbb{C}^n$ such that $\underset{\sim}{y} \neq \underset{\sim}{0}$ and

$$A\underset{\sim}{x} = \underset{\sim}{y} \quad \text{and} \quad \underset{\sim}{\hat{A}}\hat{\underset{\sim}{x}} = \hat{\underset{\sim}{y}}\,,$$

then

$$\frac{\|\hat{\underset{\sim}{x}} - \underset{\sim}{x}\|}{\|\underset{\sim}{x}\|} \le \|A\|\ \|A^{-1}\|\ \frac{\dfrac{\|\hat{\underset{\sim}{y}} - \underset{\sim}{y}\|}{\|\underset{\sim}{y}\|} + \dfrac{\|\hat{A} - A\|}{\|A\|}}{1-\|A^{-1}(\hat{A}-A)\|}\,. \tag{8}$$

If we let

$$\epsilon = \max\left\{\frac{\|\hat{\underset{\sim}{y}} - \underset{\sim}{y}\|}{\|\underset{\sim}{y}\|}\,,\ \frac{\|\hat{A} - A\|}{\|A\|}\right\},$$

then $\delta \equiv \|A^{-1}(\hat{A}-A)\| \le \|A\|\ \|A^{-1}\|\epsilon$, and hence

$$\begin{aligned}\frac{\|\hat{\underset{\sim}{x}} - \underset{\sim}{x}\|}{\|\underset{\sim}{x}\|} &\le 2\epsilon\|A\|\ \|A^{-1}\|(1 + \delta + \delta^2 + \ldots) \\ &\le 2\|A\|\ \|A^{-1}\|\epsilon + 2\|A\|^2\|A^{-1}\|^2\epsilon^2(1 + \delta + \delta^2 + \ldots)\,.\end{aligned}$$

In other words,

$$\frac{\|\hat{\underset{\sim}{x}}-\underset{\sim}{x}\|}{\|\underset{\sim}{x}\|} = 2\|A\|\ \|A^{-1}\|\epsilon + O(\epsilon^2)\,, \quad \text{as} \quad \epsilon \to 0\,. \tag{9}$$

This shows that the relative change $\|\tilde{\underset{\sim}{x}}-\underset{\sim}{x}\| / \|\underset{\sim}{x}\|$ in the solution $\underset{\sim}{x}$ of $A\underset{\sim}{u} = \underset{\sim}{y}$ is bounded essentially by $2\|A\|\ \|A^{-1}\|$ times the larger of the relative changes in A and in y. For this reason, the quantity

$$k(A) \equiv \|A\|\ \|A^{-1}\| \tag{10}$$

is called the <u>condition</u> <u>number</u> for the linear system having the nonsingular matrix A as its coefficient matrix.

If we use the Euclidean norm $\| \|_2$, then the condition number $k_2(A) = \|A\|_2 \; \|A^{-1}\|_2$ can be given another interpretation. First note that since $\|A^H A\|_2 = \|A\|_2^2 = \|AA^H\|_2^2$, we have

$$k_2(A) = k_2(A^H A)^{1/2} = k_2(AA^H)^{1/2} . \tag{11}$$

Let $\lambda_1(A), \dots, \lambda_n(A)$ denote the eigenvalues of A arranged so that $|\lambda_1(A)| \geq \dots \geq |\lambda_n(A)| \neq 0$. Let B be an $m \times n$ matrix. Then $\sigma_j(B) \equiv \sqrt{\lambda_j(B^H B)}$ is called the j-th <u>singular</u> <u>value</u> of B; it is the positive square root of the j-th largest eigenvalue of $B^H B$. Notice that if A is normal, then $\lambda_j(A^H A) = |\lambda_j(A)|^2$, so that

$$\sigma_j(A) = |\lambda_j(A)| = \sqrt{\lambda_j(A^H A)} \quad (A \text{ normal}) . \tag{12}$$

Now, since $A^H A$ is always normal,

$$\|A\|_2 = \sqrt{\|A^H A\|_2} = \sqrt{\lambda_1(A^H A)} = \sigma_1(A) , \tag{13}$$

and since $\lambda_j(A) = 1/\lambda_{n-j+1}(A^{-1})$ for $j = 1, \dots, n$, we have

$$\|A^{-1}\|_2 = \|(A^H)^{-1}\|_2 = \sqrt{\lambda_1((A^{-1}(A^H)^{-1}}$$
$$= \sqrt{\lambda_1((A^H A)^{-1})} = 1/\sqrt{\lambda_n(A^H A)} = 1/\sigma_n(A) .$$

Thus, we have

$$k_2(A) = \frac{\sigma_1(A)}{\sigma_n(A)} . \tag{14}$$

Let us now consider a more general case when A is an $m \times n$ matrix of rank n. Then the equation $A\underset{\sim}{u} = \underset{\sim}{y}$ has at most one solution for every $\underset{\sim}{y} \in \mathbb{C}^m$. Unless $m = n$, there exists $y \in \mathbb{C}^m$ for which there is no solution. All the same, there is a unique least squares solution for every $y \in \mathbb{C}^m$. In analogy with the square nonsingular matrix case, we define the condition number of an $m \times n$ matrix A of rank n by

$$k(A) = \|A\| \, \|A^{\dagger}\| \, . \tag{15}$$

Note that for the Euclidean norm $\| \ \|_2$,

$$\begin{aligned} \|A\|_2^2 &= \|A^HA\|_2 \, , \\ \|A^{\dagger}\|_2^2 &= \|A^{\dagger}(A^{\dagger})^H\|_2 = \|(A^HA)^{-1}A^HA(A^HA)^{-1}\|_2 = \|(A^HA^{-1})\|_2 \, . \end{aligned} \tag{16}$$

Hence by (12) and (14)

$$k_2(A) = \sqrt{\|A^HA\|_2}\,\sqrt{\|(A^HA)^{-1}\|_2} = \sqrt{k_2(A^HA)} = \frac{\sigma_1(A)}{\sigma_n(A)} \tag{17}$$

as in (10) and (14).

Let $\underset{\sim}{y} \in \mathbb{C}^m$. In the method of normal equations to find the least squares solution of $A\underset{\sim}{u} = \underset{\sim}{y}$, we let $B = A^HA$ and find $\underset{\sim}{x}_N$ such that $B\underset{\sim}{x}_N = A^H\underset{\sim}{y}$. Let $\hat{B}$ be another $n \times n$ matrix such that $\|(\hat{B}-A^HA)(A^HA)^{-1}\| < 1$, so that $\hat{B}$ is nonsingular. If $\hat{B}\hat{\underset{\sim}{x}}_N = \hat{z}$, then since $k_2(A^HA) = k_2(A)$ by (9),

$$\frac{\|\hat{\underset{\sim}{x}}_N - \underset{\sim}{x}\|_2}{\|\underset{\sim}{x}\|_2} \le 2[k_2(A)]^2\epsilon + O(\epsilon^2) \, , \quad \text{as} \quad \epsilon \to 0 \, , \tag{18}$$

where $\epsilon = \max\{\|\hat{z} - A^H\underset{\sim}{y}\|_2 \,/\, \|A^H\underset{\sim}{y}\|_2 \, , \, \|\hat{B}-A^HA\|_2 \,/\, \|A^HA\|_2\}$.

In case $\underset{\sim}{y}$ belongs to the range of A , the relative change in the least squares solution of $A\underset{\sim}{u} = \underset{\sim}{y}$ has a better bound, as the following result shows.

THEOREM 6 Let A and $\hat{A}$ be $m \times n$ matrices and let A have rank n . If $\|A^{\dagger}(A-\hat{A})\| < 1$, then $\hat{A}$ is also of rank n . Let $\underset{\sim}{y}$ and $\hat{\underset{\sim}{y}}$ be in $\mathbb{C}^m$, and $\underset{\sim}{x}$ and $\hat{\underset{\sim}{x}}$ be in $\mathbb{C}^n$ such that

$$A\underset{\sim}{x} = \underset{\sim}{y} \quad \text{and} \quad \|\hat{A}\hat{\underset{\sim}{x}}-\underset{\sim}{y}\|_2 = \min_{\underset{\sim}{u}\in\mathbb{C}^n} \|\hat{A}\underset{\sim}{u}-\hat{\underset{\sim}{y}}\|_2 \ .$$

Assume that

$$\epsilon = \max\left\{\frac{\|\hat{\underset{\sim}{y}} - \underset{\sim}{y}\|_2}{\|\underset{\sim}{y}\|_2} \ , \ \frac{\|\hat{A} - A\|_2}{\|A\|_2}\right\} < \frac{1}{k_2(A)} \ .$$

Then

$$\frac{\|\hat{\underset{\sim}{x}} - \underset{\sim}{x}\|_2}{\|\underset{\sim}{x}\|_2} = 2k_2(A)\epsilon + O(\epsilon^2) \ , \quad \text{as} \quad \epsilon \to 0 \ .$$

The first part of the above theorem is easy to prove: Let $\hat{A}\underset{\sim}{u} = \underset{\sim}{0}$. Since $\|A^{\dagger}(A-\hat{A})\| < 1$ and

$$\|\underset{\sim}{u}\| = \|A^{\dagger}A\underset{\sim}{u}\| = \|A^{\dagger}(A\underset{\sim}{u}-\hat{A}\underset{\sim}{u})\| \leq \|A^{\dagger}(A-\hat{A})\| \ \|\underset{\sim}{u}\| \ ,$$

we see that $\|\underset{\sim}{u}\| = 0$, i.e., $\underset{\sim}{u} = \underset{\sim}{0}$. This shows that $\hat{A}$ is one to one, i.e., rank $\hat{A} = n$. The second part of the theorem is difficult to prove and we refer the reader to pages 141, 143 and 144 of [GV].

Numerical stability

While solving a system of linear equations $A\underset{\sim}{x} = \underset{\sim}{y}$ on a computer, we have to use the floating-point representation of the entries of A and $\underset{\sim}{y}$; thus the entries are only approximately correct. Further, in the process of solving the problem, round-off errors arise due to the floating-point arithmetic of the computer. (See Section 18 for some details.) For many well known methods of solving a linear system, it can be shown that the computed solution $\hat{\underset{\sim}{x}}$, in fact, satisfies a nearby system $\hat{A}\hat{\underset{\sim}{x}} = \hat{\underset{\sim}{y}}$. In these cases, the perturbation analysis of the solution given earlier becomes applicable.

In the case of Gaussian elimination with partial pivoting for an $n \times n$ nonsingular system, the computed solution satisfies a linear system with coefficient matrix $\hat{A}$ such that

$$\|\hat{A}-A\|_\infty = 8n^3\rho\|A\|_\infty\delta + O(\delta^2) , \tag{19}$$

where $\delta = \frac{1}{2}\beta^{1-t}$, with β the machine base and t the machine precision, and ρ is a certain growth factor which measures how large the entries become in the solution process. Empirically ρ is known to be of modest size ([GV], p.67).

If $m \neq n$, then a round-off error analysis for Gaussian elimination with partial pivoting for an $m \times n$ system is not feasible, because the pivots are not uniquely determined. Thus, it is not possible to associate a unique $\underset{\sim}{x} \in \mathbb{C}^n$ as the 'solution' of $A\underset{\sim}{u} = \underset{\sim}{y}$ for an arbitrary $\underset{\sim}{y} \in \mathbb{C}^m$.

In the case of the Householder method for the $m \times n$ least squares problem $A\underset{\sim}{u} = \underset{\sim}{y}$ of rank n , it can be shown that the computed solution $\hat{\underset{\sim}{x}}$ satisfies

$$\|\hat{A}\hat{\underset{\sim}{x}} - \hat{\underset{\sim}{y}}\|_2 = \min_{\underset{\sim}{u}\in\mathbb{C}^n} \|\hat{A}\underset{\sim}{u} - \hat{\underset{\sim}{y}}\|_2 ,$$

where $\hat{A}$ and $\hat{\underset{\sim}{y}}$ satisfy

$$\begin{aligned} \|\hat{A} - A\|_2 &= n\delta(6m-3n+41)\alpha + O(\delta^2) , \\ \|\hat{\underset{\sim}{y}} - \underset{\sim}{y}\|_2 &= n\delta(6m-3n+40)\|\underset{\sim}{y}\|_2 + O(\delta^2) , \end{aligned} \tag{20}$$

with $\alpha^2 = \sum_{i,j=1}^{n} |a_{i,j}|^2$. (See p.149 of [GV].)

While Gaussian elimination with partial pivoting has a smaller flop count, the Householder method has guaranteed stability. There are several other methods for solving linear systems. But the above two are often recommended from the point of economy and numerical stability.