Jim Lambers
MAT 610
Summer Session 2009-10
Lecture 5 Notes

These notes correspond to Sections 3.3 and 3.4 in the text.

Roundoff Analysis of Gaussian Elimination

In this section, we will perform a detailed error analysis of Gaussian elimination, illustrating the analysis originally carried out by J. H. Wilkinson. The process of solving $A\mathbf{x} = \mathbf{b}$ consists of three stages:

1. Factoring $A = LU$, resulting in an approximate LU decomposition $A + E = \bar{L}\bar{U}$.
2. Solving $L\mathbf{y} = \mathbf{b}$, or, numerically, computing $\mathbf{y}$ such that
$$(\bar{L} + \delta\bar{L})(\mathbf{y} + \delta\mathbf{y}) = \mathbf{b}$$
3. Solving $U\mathbf{x} = \mathbf{y}$, or, numerically, computing $\mathbf{x}$ such that
$$(\bar{U} + \delta\bar{U})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{y} + \delta\mathbf{y}.$$

Combining these stages, we see that

$$\begin{aligned}
\mathbf{b} &= (\bar{L} + \delta\bar{L})(\bar{U} + \delta\bar{U})(\mathbf{x} + \delta\mathbf{x}) \\
&= (\bar{L}\bar{U} + \delta\bar{L}\bar{U} + \bar{L}\delta\bar{U} + \delta\bar{L}\delta\bar{U})(\mathbf{x} + \delta\mathbf{x}) \\
&= (A + E + \delta\bar{L}\bar{U} + \bar{L}\delta\bar{U} + \delta\bar{L}\delta\bar{U})(\mathbf{x} + \delta\mathbf{x}) \\
&= (A + \delta A)(\mathbf{x} + \delta\mathbf{x})
\end{aligned}$$

where $\delta A = E + \delta\bar{L}\bar{U} + \bar{L}\delta\bar{U} + \delta\bar{L}\delta\bar{U}$.

In this analysis, we will view the computed solution $\bar{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ as the exact solution to the perturbed problem $(A + \delta A)\mathbf{x} = \mathbf{b}$. This perspective is the idea behind *backward error analysis*, which we will use to determine the size of the perturbation δA, and, eventually, arrive at a bound for the error in the computed solution $\bar{\mathbf{x}}$.

Error in the LU Factorization

Let $A^{(k)}$ denote the matrix A after $k-1$ steps of Gaussian elimination have been performed, where a step denotes the process of making all elements below the diagonal within a particular column equal to zero. Then, in exact arithmetic, the elements of $A^{(k+1)}$ are given by

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}, \quad m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}. \tag{1}$$

Let $B^{(k)}$ denote the matrix A after $k-1$ steps of Gaussian elimination have been performed using floating-point arithmetic. Then the elements of $B^{(k+1)}$ are

$$b_{ij}^{(k+1)} = b_{ij}^{(k)} - s_{ik}b_{kj}^{(k)} + \epsilon_{ij}^{(k+1)}, \quad s_{ik} = fl\left(\frac{b_{ik}^{(k)}}{b_{kk}^{(k)}}\right). \tag{2}$$

For $j \geq i$, we have

$$\begin{array}{rcl} b_{ij}^{(2)} & = & b_{ij}^{(1)} - s_{i1}b_{1j}^{(1)} + \epsilon_{ij}^{(2)} \\ b_{ij}^{(3)} & = & b_{ij}^{(2)} - s_{i2}b_{2j}^{(2)} + \epsilon_{ij}^{(3)} \\ & \vdots & \\ b_{ij}^{(i)} & = & b_{ij}^{(i-1)} - s_{i,i-1}b_{i-1,j}^{(i-1)} + \epsilon_{ij}^{(i)} \end{array}$$

Combining these equations yields

$$\sum_{k=2}^{i} b_{ij}^{(k)} = \sum_{k=1}^{i-1} b_{ij}^{(k)} - \sum_{k=1}^{i-1} s_{ik}b_{kj}^{(k)} + \sum_{k=2}^{i} \epsilon_{ij}^{(k)}$$

Cancelling terms, we obtain

$$b_{ij}^{(1)} = b_{ij}^{(i)} + \sum_{k=1}^{i-1} s_{ik}b_{kj}^{(k)} - e_{ij}, \quad j \geq i, \tag{3}$$

where $e_{ij} = \sum_{k=2}^{i} \epsilon_{ij}^{(k)}$.

For $i > j$,

$$\begin{array}{rcl} b_{ij}^{(2)} & = & b_{ij}^{(1)} - s_{i1}b_{1j}^{(1)} + \epsilon_{ij}^{(2)} \\ & \vdots & \\ b_{ij}^{(j)} & = & b_{ij}^{(j-1)} - s_{i,j-1}b_{j-1,j}^{(j-1)} + \epsilon_{ij}^{(j)} \end{array}$$

where $s_{ij} = fl(b_{ij}^{(j)}/b_{jj}^{(j)}) = b_{ij}^{(j)}/b_{jj}^{(j)}(1+\eta_{ij})$, and therefore

$$\begin{array}{rcl} 0 & = & b_{ij}^{(j)} - s_{ij}b_{jj}^{(j)} + b_{ij}^{(j)}\eta_{ij} \\ & = & b_{ij}^{(j)} - s_{ij}b_{jj}^{(j)} + \epsilon_{ij}^{(j+1)} \\ & = & b_{ij}^{(1)} - \displaystyle\sum_{k=1}^{j} s_{ik}b_{kj}^{(k)} + e_{ij} \end{array} \tag{4}$$

From (3) and (4), we obtain

$$\bar{L}\bar{U} = \begin{bmatrix} 1 & & & \\ s_{21} & 1 & & \\ \vdots & & \ddots & \\ s_{n1} & \cdots & \cdots & 1 \end{bmatrix} \begin{bmatrix} b_{11}^{(1)} & b_{12}^{(1)} & \cdots & b_{1n}^{(1)} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & b_{nn}^{(n)} \end{bmatrix} = A + E.$$

where

$$s_{ik} = fl(b_{ik}^{(k)}/b_{kk}^{(k)}) = \frac{b_{ik}^{(k)}}{b_{kk}^{(k)}}(1+\eta_{ik}), \quad |\eta_{ik}| \leq \mathbf{u}$$

Then,

$$fl(s_{ik}b_{kj}^{(k)}) = s_{ik}b_{kj}^{(k)}(1+\theta_{ij}^{(k)}), \quad |\theta_{ij}^{(k)}| \leq \mathbf{u}$$

and so,

$$\begin{aligned} b_{ij}^{(k+1)} &= fl(b_{ij}^{(k)} - s_{ik}b_{kj}^{(k)}(1+\theta_{ij}^{(k)})) \\ &= (b_{ij}^{(k)} - s_{ik}b_{kj}^{(k)}(1+\theta_{ij}^{(k)}))(1+\varphi_{ij}^{(k)}), \quad |\varphi_{ij}^{(k)}| \leq \mathbf{u}. \end{aligned}$$

After applying (2), and performing some manipulations, we obtain

$$\epsilon_{ij}^{(k+1)} = b_{ij}^{(k+1)}\left(\frac{\varphi_{ij}^{(k)}}{1+\varphi_{ij}^{(k)}}\right) - s_{ik}b_{kj}^{(k)}\theta_{ij}^{(k)}.$$

We then have the following bound for the entries of E:

$$|E| \leq (1+\ell)Ga\mathbf{u} \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & \cdots & \cdots & \cdots & \cdots & 1 \\ 1 & 2 & \cdots & \cdots & \cdots & 2 \\ \vdots & \vdots & 3 & \cdots & \cdots & 3 \\ & & & \ddots & \cdots & \vdots \\ 1 & 2 & 3 & \cdots & n-1 & n-1 \end{bmatrix} + O(\mathbf{u}^2),$$

where $\ell = \max_{i,j}|s_{ij}|$ and G is the *growth factor* defined by

$$G = \frac{\max_{i,j,k}|\bar{b}_{ij}^{(k)}|}{\max_{i,j}|a_{ij}|}.$$

The entries of the above matrix indicate how many of the $\epsilon_{ij}^{(k)}$ are included in each entry e_{ij} of E.

Bounding the perturbation in A

From a roundoff analysis of forward and back substitution, which we do not reproduce here, we have the bounds

$$\begin{aligned}\max_{i,j} |\delta \bar{L}_{ij}| &\leq n\mathbf{u}\ell + O(\mathbf{u}^2),\\ \max_{i,j} |\delta \bar{U}_{ij}| &\leq n\mathbf{u}Ga + O(\mathbf{u}^2)\end{aligned}$$

where $a = \max_{i,j} |a_{ij}|$, $\ell = \max_{i,j} |\bar{L}_{ij}|$, and G is the growth factor. Putting our bounds together, we have

$$\begin{aligned}\max_{i,j} |\delta A_{ij}| &\leq \max_{i,j} |e_{ij}| + \max_{i,j} |\bar{L}\delta\bar{U}_{ij}| + \max_{i,j} |\bar{U}\delta\bar{L}_{ij}| + \max_{i,j} |\delta\bar{L}\delta\bar{U}_{ij}|\\ &\leq n(1+\ell)Ga\mathbf{u} + n^2\ell Ga\mathbf{u} + n^2\ell Ga\mathbf{u} + O(\mathbf{u}^2)\end{aligned}$$

from which it follows that

$$\|\delta A\|_\infty \leq n^2(2n\ell + \ell + 1)Ga\mathbf{u} + O(\mathbf{u}^2).$$

Bounding the error in the solution

Let $\bar{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ be the computed solution. Since the exact solution to $A\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = A^{-1}\mathbf{b}$, we are also interested in examining $\bar{\mathbf{x}} = (A + \delta A)^{-1}\mathbf{b}$. Can we say something about $\|(A+\delta A)^{-1}\mathbf{b} - A^{-1}\mathbf{b}\|$?

We assume that $\|A^{-1}\delta A\| = r < 1$. We have

$$A + \delta A = A(I + A^{-1}\delta A) = A(I - F), \quad F = -A^{-1}\delta A.$$

From the manipulations

$$\begin{aligned}(A+\delta A)^{-1}\mathbf{b} - A^{-1}\mathbf{b} &= (I + A^{-1}\delta A)^{-1}A^{-1}\mathbf{b} - A^{-1}\mathbf{b}\\ &= (I + A^{-1}\delta A)^{-1}(A^{-1} - (I + A^{-1}\delta A)A^{-1})\mathbf{b}\\ &= (I + A^{-1}\delta A)^{-1}(-A^{-1}(\delta A)A^{-1})\mathbf{b}\end{aligned}$$

and this result from Lecture 2,

$$\|(I-F)^{-1}\| \leq \frac{1}{1-r},$$

we obtain

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|(\mathbf{x} + \delta\mathbf{x}) - \mathbf{x}\|}{\|\mathbf{x}\|}$$

$$
\begin{aligned}
&= \frac{\|(A+\delta A)^{-1}\mathbf{b} - A^{-1}\mathbf{b}\|}{\|A^{-1}\mathbf{b}\|} \\
&\leq \frac{1}{1-r}\|A^{-1}\|\|\delta A\| \\
&\leq \frac{1}{1-\|A^{-1}\delta A\|}\kappa(A)\frac{\|\delta A\|}{\|A\|} \\
&\leq \frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}.
\end{aligned}
$$

Note that a similar conclusion is reached if we assume that the computed solution $\bar{\mathbf{x}}$ solves a nearby problem in which *both* A and $\mathbf{b}$ are perturbed, rather than just A.

We see that the important factors in the accuracy of the computed solution are

- The growth factor G
- The size of the multipliers m_{ik}, bounded by ℓ
- The condition number $\kappa(A)$
- The precision $\mathbf{u}$

In particular, $\kappa(A)$ must be large with respect to the accuracy in order to be troublesome. For example, consider the scenario where $\kappa(A) = 10^2$ and $\mathbf{u} = 10^{-3}$, as opposed to the case where $\kappa(A) = 10^2$ and $\mathbf{u} = 10^{-50}$. However, it is important to note that even if A is well-conditioned, the error in the solution can still be very large, if G and ℓ are large.

Pivoting

During Gaussian elimination, it is necessary to interchange rows of the augmented matrix whenever the diagonal element of the column currently being processed, known as the *pivot element*, is equal to zero.

However, if we examine the main step in Gaussian elimination,

$$a_{ik}^{(j+1)} = a_{ik}^{(j)} - m_{ij}a_{jk}^{(j)},$$

we can see that any roundoff error in the computation of $a_{jk}^{(j)}$ is amplified by m_{ij}. Because the multipliers can be arbitrarily small, it follows from the previous analysis that the error in the computed solution can be arbitrarily large. Therefore, Gaussian elimination is *numerically unstable*.

Therefore, it is helpful if it can be ensured that the multipliers are small. This can be accomplished by performing row interchanges, or *pivoting*, even when it is not absolutely necessary to do so for elimination to proceed.

Partial Pivoting

One approach is called *partial pivoting.* When eliminating elements in column j, we seek the largest element in column j, on or below the main diagonal, and then interchanging that element's row with row j. That is, we find an integer p, $j \leq p \leq n$, such that

$$|a_{pj}| = \max_{j \leq i \leq n} |a_{ij}|.$$

Then, we interchange rows p and j.

In view of the definition of the multiplier, $m_{ij} = a_{ij}^{(j)}/a_{jj}^{(j)}$, it follows that $|m_{ij}| \leq 1$ for $j = 1, \ldots, n-1$ and $i = j+1, \ldots, n$. Furthermore, while pivoting in this manner requires $O(n^2)$ comparisons to determine the appropriate row interchanges, that extra expense is negligible compared to the overall cost of Gaussian elimination, and therefore is outweighed by the potential reduction in roundoff error. We note that when partial pivoting is used, the growth factor G is 2^{n-1}, where A is $n \times n$.

Complete Pivoting

While partial pivoting helps to control the propagation of roundoff error, loss of significant digits can still result if, in the abovementioned main step of Gaussian elimination, $m_{ij}a_{jk}^{(j)}$ is much larger in magnitude than $a_{ij}^{(j)}$. Even though m_{ij} is not large, this can still occur if $a_{jk}^{(j)}$ is particularly large.

Complete pivoting entails finding integers p and q such that

$$|a_{pq}| = \max_{j \leq i \leq n, j \leq q \leq n} |a_{ij}|,$$

and then using both row *and column* interchanges to move a_{pq} into the pivot position in row j and column j. It has been proven that this is an effective strategy for ensuring that Gaussian elimination is *backward stable*, meaning it does not cause the entries of the matrix to grow exponentially as they are updated by elementary row operations, which is undesirable because it can cause undue amplification of roundoff error.

The LU Decomposition with Pivoting

Suppose that pivoting is performed during Gaussian elimination. Then, if row j is interchanged with row p, for $p > j$, before entries in column j are eliminated, the matrix $A^{(j)}$ is effectively multiplied by a *permutation matrix* $P^{(j)}$. A permutation matrix is a matrix obtained by permuting the rows (or columns) of the identity matrix I. In $P^{(j)}$, rows j and p of I are interchanged, so that multiplying $A^{(j)}$ on the left by $P^{(j)}$ interchanges these rows of $A^{(j)}$. It follows that the process of Gaussian elimination with pivoting can be described in terms of the matrix multiplications

$$M^{(n-1)}P^{(n-1)}M^{(n-2)}P^{(n-2)} \cdots M^{(1)}P^{(1)}A = U,$$

where $P^{(k)} = I$ if no interchange is performed before eliminating entries in column k.

However, because each permutation matrix $P^{(k)}$ at most interchanges row k with row p, where $p > k$, there is no difference between applying all of the row interchanges "up front", instead of applying $P^{(k)}$ immediately before applying $M^{(k)}$ for each k. It follows that

$$[M^{(n-1)}M^{(n-2)}\cdots M^{(1)}][P^{(n-1)}P^{(n-2)}\cdots P^{(1)}]A = U,$$

and because a product of permutation matrices is a permutation matrix, we have

$$PA = LU,$$

where L is defined as before, and $P = P^{(n-1)}P^{(n-2)}\cdots P^{(1)}$. This decomposition exists for *any* nonsingular matrix A.

Once the LU decomposition $PA = LU$ has been computed, we can solve the system $A\mathbf{x} = \mathbf{b}$ by first noting that if $\mathbf{x}$ is the solution, then

$$PA\mathbf{x} = LU\mathbf{x} = P\mathbf{b}.$$

Therefore, we can obtain $\mathbf{x}$ by first solving the system $L\mathbf{y} = P\mathbf{b}$, and then solving $U\mathbf{x} = \mathbf{y}$. Then, if $\mathbf{b}$ should change, then only these last two systems need to be solved in order to obtain the new solution; as in the case of Gaussian elimination without pivoting, the LU decomposition does not need to be recomputed.

Practical Computation of Determinants, Revisited

When Gaussian elimination is used without pivoting to obtain the factorization $A = LU$, we have $\det(A) = \det(U)$, because $\det(L) = 1$ due to L being unit lower-triangular. When pivoting is used, we have the factorization $PA = LU$, where P is a permutation matrix. Because a permutation matrix is orthogonal; that is, $P^TP = I$, and $\det(A) = \det(A^T)$ for any square matrix, it follows that $\det(P)^2 = 1$, or $\det(P) = \pm 1$. Therefore, $\det(A) = \pm\det(U)$, where the sign depends on the sign of $\det(P)$.

To determine this sign, we note that when two rows (or columns) of A are interchanged, the sign of the determinant changes. Therefore, $\det(P) = (-1)^p$, where p is the number of row interchanges that are performed during Gaussian elimination. The number p is known as the *sign* of the permutation represented by P that determines the final ordering of the rows. We conclude that $\det(A) = (-1)^p\det(U)$.