CME 302: NUMERICAL LINEAR ALGEBRA
FALL 2005/06
LECTURE 9

GENE H. GOLUB

1. Error Analysis of Gaussian Elimination

In this section, we will consider the case of Gaussian elimination and perform a detailed error analysis, 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 $A+E = \bar{L}\bar{U}$. We assume that partial pivoting is used.
(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

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

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

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

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

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}, \quad m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}. \tag{1.1}$$

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

$$b_{ij}^{(k+1)} = a_{ij}^{(k)} - s_{ik}b_{kj}^{(k)} + \epsilon_{ij}^{(k+1)}, \quad s_{ik} = fl\left(\frac{b_{ik}^{(k)}}{b_{kk}^{(k)}}\right). \tag{1.2}$$

Date: November 25, 2005, version 1.1.
Notes originally due to James Lambers. Edited by Lek-Heng Lim.

For $j \geq i$, we have

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

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

$$b_{ij}^{(1)} = b_{ij}^{(i)} + \sum_{k=1}^{i-1} s_{ik} b_{kj}^{(k)} + e_{ij}, \quad j \geq i, \tag{1.3}$$

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

For $i > j$,

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

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

$$\begin{aligned} 0 &= b_{ij}^{(j)} - s_{ij} b_{jj}^{(j)} + b_{jj}^{(j)} \eta_{ij} \\ &= b_{ij}^{(j)} - s_{ij} b_{jj}^{(j)} + \epsilon_{ij}^{(j+1)} \\ &= b_{ij}^{(1)} - \sum_{k=1}^{j} s_{ik} b_{kj}^{(k)} + e_{ij} \end{aligned} \tag{1.4}$$

From (1.3) and (1.4), 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

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

Then,

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

and so,

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

After some manipulations, we obtain

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

With partial pivoting, $|s_{ik}| \leq 1$, provided that $|fl(a/b)| \leq 1$ whenever $|a| \leq |b|$. In most modern implementations of floating-point arithmetic, this is in fact the case. It follows that

$$|\epsilon_{ij}^{(k+1)}| \leq |b_{ij}^{(k+1)}| \frac{\mathsf{u}}{1-\mathsf{u}} + 1 \cdot |b_{ij}^{(k)}| \mathsf{u}.$$

How large can the elements of $B^{(k)}$ be? Returning to exact arithmetic, we assume that $|a_{ij}| \leq a$ and from (1.1), we obtain

$$\begin{aligned} |a_{ij}^{(2)}| &\leq |a_{ij}^{(1)}| + |a_{kj}^{(1)}| \leq 2a \\ |a_{ij}^{(3)}| &\leq 4a \\ &\vdots \\ |a_{ij}^{(n)}| &= |a_{nn}^{(n)}| \leq 2^{n-1} a. \end{aligned}$$

We can show that a similar result holds in floating-point arithmetic:

$$|b_{ij}^{(k)}| \leq 2^{k-1} a + O(\mathsf{u}).$$

This upper bound is achievable (by Hadamard matrices), but in practice it rarely occurs.

2. Error in the LU Factorization

Recall from last time that we were analyzing the error in solving $A\mathbf{x} = \mathbf{b}$ using backward error analysis, in which we assume that our computed solution $\bar{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ is the exact solution to the perturbed problem

$$(A + \delta A)\bar{\mathbf{x}} = \mathbf{b}$$

where δA is a perturbation that has the form

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

and the following relationships hold:

(1) $A + E = \bar{L}\bar{U}$
(2) $(\bar{L} + \delta\bar{L})(\mathbf{y} + \delta\mathbf{y}) = \mathbf{b}$
(3) $(\bar{U} + \delta\bar{U})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{y} + \delta\mathbf{y}$

We concluded that when partial pivoting is used, the entries of $\bar{U}$ were bounded:

$$|b_{ij}^{(k)}| \leq 2^{k-1} a + O(\mathsf{u})$$

where k is the number of steps of Gaussian elimination that effect the ij element and a is an upper bound on the elements of A.

For complete pivoting, Wilkinson gave a bound, denoted G, or *growth factor*. Until 1990, it was conjectured that $G \leq k$. It was shown to be true for $n \leq 5$, but there have been examples constructed for $n > 5$ where $G \geq n$. In any event, we have the following bound for the entries of E:

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

3. Error Analysis of Forward Substitution

We now study the process of forward substitution, to solve

$$\begin{bmatrix} t_{11} & & 0 \\ \vdots & \ddots & \\ t_{n1} & & t_{nn} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}.$$

Using forward substitution, we obtain

$$\begin{aligned} u_1 &= \frac{h_1}{t_{11}} \\ &\vdots \\ u_k &= \frac{h_k - t_{k1}u_1 - \cdots - t_{k,k-1}u_{k-1}}{t_{kk}} \end{aligned}$$

which yields

$$\begin{aligned} fl(u_k) &= \frac{h_k(1+\epsilon_k)(1+\eta_k) - \sum_{i=1}^{k-1} t_{ki}u_i(1+\xi_{ki})(1+\epsilon_k)(1+\eta_k)}{t_{kk}} \\ &= \frac{h_k - \sum_{i=1}^{k-1} t_{ki}u_i(1+\xi_{ki})}{\frac{t_{kk}}{(1+\epsilon_k)(1+\eta_k)}} \end{aligned}$$

or

$$\sum_{i=1}^{k} u_i t_{ki}(1+\lambda_{ki}) = h_k$$

which can be rewritten in matrix notation as

$$T\mathbf{u} + \begin{bmatrix} \lambda_{11}t_{11} & & \\ \lambda_{12}t_{12} & \lambda_{22}t_{22} & \\ \vdots & \vdots & \ddots \end{bmatrix} \mathbf{u} = \mathbf{h}.$$

In other words, the computed solution $\mathbf{u}$ is the exact solution to the perturbed problem $(T+\delta T)\mathbf{u} = \mathbf{h}$, where

$$|\delta T| \leq \mathsf{u} \begin{bmatrix} |t_{11}| & & & \\ |t_{21}| & 2|t_{22}| & & \\ \vdots & & \ddots & \\ (n-1)|t_{n1}| & \cdots & \cdots & 2|t_{nn}| \end{bmatrix} + O(\mathsf{u}^2).$$

Note that the perturbation δT actually depends on $\mathbf{h}$.

4. Bounding the perturbation in A

Recall that our computed solution $\mathbf{x} + \delta\mathbf{x}$ solves

$$(A + \delta A)\bar{\mathbf{x}} = \mathbf{b}$$

where δA is a perturbation that has the form

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

For partial pivoting, $|\bar{l}_{ij}| \leq 1$, and we have the bounds

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

were $a = \max_{i,j} |a_{ij}|$ and G is the growth factor for partial pivoting. Putting our bounds together, we have

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

from which it follows that

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

We conclude that Gaussian elimination is *backward stable*.

5. Bounding the error in the solution

Let $\bar{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ be the computed solution. Then, from $(A + \delta A)\bar{\mathbf{x}} = \mathbf{b}$ we obtain

$$\delta A\bar{\mathbf{x}} = \mathbf{b} - A\bar{\mathbf{x}} = \mathbf{r}$$

where $\mathbf{r}$ is called the *residual vector*. From our previous analysis,

$$\frac{\|\mathbf{r}\|_\infty}{\|\bar{\mathbf{x}}\|_\infty} \leq \|\delta A\|_\infty \leq 2n^2(n+1)Ga\mathsf{u}.$$

Also, recall

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\dfrac{\|\delta A\|}{\|A\|}} \frac{\|\delta A\|}{\|A\|}.$$

We know that $\|A\|_\infty \leq na$, so

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \leq 2n(n+1)G\mathsf{u}.$$

Note that if $\kappa(A)$ is large and G is large, our solution can be very inaccurate. The important factors in the accuracy of the computed solution are:

- The growth factor G
- The condition number κ
- The accuracy u

In particular, κ must be large with respect to the accuracy in order to be troublesome. For example, consider the scenario where $\kappa = 10^2$ and $\mathsf{u} = 10^{-3}$, as opposed to the case where $\kappa = 10^2$ and $\mathsf{u} = 10^{-50}$.

6. Iterative Refinement

The process of *iterative refinement* proceeds as follows to find a solution to $A\mathbf{x} = \mathbf{b}$:

$$\begin{aligned}\mathbf{x}^{(0)} &= \mathbf{0} \\ \mathbf{r}^{(i)} &= \mathbf{b} - A\mathbf{x}^{(i)} \\ A\boldsymbol{\delta}^{(i)} &= \mathbf{r}^{(i)} \\ \mathbf{x}^{(i+1)} &= \mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}\end{aligned}$$

Numerically, this translates to

$$\begin{aligned}(A + \delta A^{(i)})\boldsymbol{\delta}^{(i)} &= (I + E^{(i)})\mathbf{r}^{(i)} \\ \mathbf{x}^{(i+1)} &= (I + F^{(i)})(\mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)})\end{aligned}$$

where the matrices $E^{(i)}$ and $F^{(i)}$ denote roundoff error. Let $\mathbf{z}^{(i)} = \mathbf{x} - \mathbf{x}^{(i)}$. Then

$$\begin{aligned}
\mathbf{x}^{(i+1)} - \mathbf{x} &= (I + F^{(i)})(\mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}) - \mathbf{x} \\
&= (I + F^{(i)})(\mathbf{x}^{(i)} - \mathbf{x}) + F^{(i)}\mathbf{x} + (I + F^{(i)})\boldsymbol{\delta}^{(i)} \\
&= (I + F^{(i)})[-\mathbf{z}^{(i)} + (I + A^{-1}\delta A^{(i)})^{-1}\mathbf{z}^{(i)} \\
&\qquad + (I + A^{-1}\delta A^{(i)})^{-1}(A^{-1}E^{(i)}A)\mathbf{z}^{(i)}] + F^{(i)}\mathbf{x} \\
&= (I + F^{(i)})(I + A^{-1}\delta A^{(i)})^{-1}(A^{-1}\delta A^{(i)}\mathbf{z}^{(i)} + A^{-1}E^{(i)}A\mathbf{z}^{(i)}) + F^{(i)}\mathbf{x}
\end{aligned}$$

which we rewrite as

$$\mathbf{z}^{(i+1)} = K^{(i)}\mathbf{z}^{(i)} + \mathbf{c}^{(i)}$$

Taking norms yields

$$\|\mathbf{z}^{(i+1)}\| \leq \|K^{(i)}\|\|\mathbf{z}^{(i)}\| + \|\mathbf{c}^{(i)}\|.$$

Under the assumptions

$$\|K^{(i)}\| \leq \tau, \quad \|\mathbf{c}^{(i)}\| \leq \sigma\|\mathbf{x}\|$$

we obtain

$$\begin{aligned}
\|\mathbf{z}^{(i+1)}\| &\leq \tau\|\mathbf{z}^{(i)}\| + \sigma\|\mathbf{x}\| \\
&\leq \tau^{i+1}\|\mathbf{z}^{(0)}\| + \sigma(1 + \tau + \cdots + \tau^{i})\|\mathbf{x}\| \\
&\leq \tau^{i+1}\|\mathbf{z}^{(0)}\| + \sigma\frac{1 - \tau^{(i+1)}}{1 - \tau}\|\mathbf{x}\|
\end{aligned}$$

Assuming $\|A^{-1}\|\|\delta A^{(i)}\| \leq \alpha$ and $\|E^{(i)}\| \leq \omega$,

$$\tau = \frac{(1 + \epsilon)(\alpha + \kappa(A)\omega)}{1 - \alpha}$$

where $\|F^{(i)}\| \leq \epsilon$. For sufficiently large i, we have

$$\frac{\|\mathbf{z}^{(i)}\|}{\|\mathbf{x}\|} \leq \frac{\epsilon}{1 - \tau} + O(\epsilon^2)$$

From

$$1 - \tau = \frac{(1 - \alpha) - (1 + \epsilon)(\alpha + \kappa(A)\omega)}{1 - \alpha}$$

we obtain

$$\frac{1}{1 - \tau} = \frac{1 - \alpha}{(1 - \alpha) - (1 + \epsilon)(\alpha + \kappa(A)\omega)} \approx \frac{1 - \alpha}{1 - 2\alpha - \kappa(A)\omega}.$$

Therefore, $1/(1 - \tau) \leq 2$ whenever

$$\alpha \leq \frac{1}{3} - \frac{2}{3}\kappa(A)\omega,$$

approximately.

It can be shown that if the vector $\mathbf{r}^{(k)}$ is computed using double or extended precision that $\mathbf{x}^{(k)}$ converges to a solution where almost all digits are correct when $\kappa(A)\mathtt{u} \leq 1$.

Department of Computer Science, Gates Building 2B, Room 280, Stanford, CA 94305-9025
E-mail address: golub@stanford.edu