Set 6: Conditioning, Stability and Factorizations

Kyle A. Gallivan

Department of Mathematics

Florida State University

Foundations of Computational Math 1

Fall 2012

Overview

- Example of stability analysis: sum of n scalars
- Conditioning of solving a linear system
- Results of backward error analysis for LU factorization and its use to solve a system
  - componentwise bounds
  - norm-based bounds
  - growth factor and stability
- Illustrative examples of factorization revisited.

Error Analysis

Consider the sum of n FP numbers,i.e., we are assuming no initial representation error.

$$\sigma_n = \xi_1 + \cdots + \xi_n$$

No assumption is made about ordering of the scalars. We consider the simple recursion:

$$\begin{aligned} \sigma_0 &= 0 \\ \sigma_i &= \sigma_{i-1} + \xi_i \quad 1 \leq i \leq n \end{aligned}$$

Error Analysis

Assume that $n = 4$ and use the standard model of floating point arithmetic of IEEE with the extra bits needed so

$$fl(x\ op\ y) = x\ \boxed{op}\ y.$$

The first iteration yields:

$$\sigma_2 = fl(\xi_1 + \xi_2) = (\xi_1 + \xi_2)(1 + \epsilon_1) = \xi_1(1 + \epsilon_1) + \xi_2(1 + \epsilon_1)$$

where $|\epsilon_1| \leq u$ and u is unit roundoff.

Error Analysis

The second iteration yields:

$$\sigma_3 = fl(\sigma_2 + \xi_3) = (\sigma_2 + \xi_3)(1+\epsilon_2) = \sigma_2(1+\epsilon_2) + \xi_3(1+\epsilon_2)$$
$$= \xi_1(1+\epsilon_1)(1+\epsilon_2) + \xi_2(1+\epsilon_1)(1+\epsilon_2) + \xi_3(1+\epsilon_2)$$

where $|\epsilon_2| \leq u$.

Note that ξ_1 and ξ_2 have been "affected" by an additional ϵ_i.

Error Analysis

The third iteration yields:

$$\begin{aligned}\sigma_4 =& fl(\sigma_3 + \xi_4) = (\sigma_3 + \xi_4)(1+\epsilon_3) = \sigma_3(1+\epsilon_3) + \xi_4(1+\epsilon_3) \\ =& \xi_1(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) + \xi_2(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) \\ &+ \xi_3(1+\epsilon_2)(1+\epsilon_3) + \xi_4(1+\epsilon_3)\end{aligned}$$

$$\begin{aligned}\sigma_4 =& \xi_1(1+\eta_1) + \xi_2(1+\eta_2) + \xi_3(1+\eta_3) + \xi_4(1+\eta_4) \qquad (1) \\ =& (\xi_1 + \xi_2 + \xi_3 + \xi_4) + (\xi_1\eta_1 + \xi_2\eta_2 + \xi_3\eta_3 + \xi_4\eta_4) \qquad (2)\end{aligned}$$

Error Analysis

- (1) is backward error analysis
- (2) is forward error analysis
- the floating point sum is the exact sum of real numbers that are modifications of the original floating point numbers.
- errors have been "cast back" to input data
- this is not always possible – it depends on number of errors and structure of input

Error Analysis

$$
\begin{aligned}
1+\eta_1 &= 1+\epsilon_1+\epsilon_2+\epsilon_3+\epsilon_1\epsilon_2+\epsilon_2\epsilon_3+\epsilon_1\epsilon_2\epsilon_3 \\
&= 1+\epsilon_1+\epsilon_2+\epsilon_3+O(u^2) \\
|\eta_1| &\lesssim 3u \\
1+\eta_2 &= 1+\epsilon_1+\epsilon_2+\epsilon_3+\epsilon_1\epsilon_2+\epsilon_2\epsilon_3+\epsilon_1\epsilon_2\epsilon_3 \\
&= 1+\epsilon_1+\epsilon_2+\epsilon_3+O(u^2) \\
|\eta_2| &\lesssim 3u \\
1+\eta_3 &= 1+\epsilon_2+\epsilon_3+\epsilon_2\epsilon_3 = 1+\epsilon_2+\epsilon_3+O(u^2) \\
|\eta_3| &\lesssim 2u \\
1+\eta_4 &= 1+\epsilon_3 \\
|\eta_4| &\lesssim u
\end{aligned}
$$

Error Analysis

- roundoff errors may cancel — bound may be very loose
- relative perturbation worst case bound $\approx nu$ for ξ_i early in list.
- generically this is a good reason to avoid large magnitudes first
- algorithm is backward stable for reasonable sizes of n
- ill-conditioned sums can still have large relative error
- large n relative to u needs more thought
- error analysis can be adapted to give tighter forward error bounds when more complicated algorithms that include sorting are used.

Condition Number for Inversion

Definition 6.1. If $A \in \mathbb{R}$ is nonsingular then its condition number with respect to inversion, i.e., solving a linear system $Ax = b$, is

$$\kappa(A) = \|A\|\|A^{-1}\|$$

where $\|*\|$ is an induced ($\because$ consistent) matrix norm.

There are various ways to motivate this definition.

A Simple Motivation

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

$$Ax = b$$

$$A(\bar{x}-x) + E\bar{x} = 0$$

$$(\bar{x}-x) = -A^{-1}E\bar{x}$$

$$\|(\bar{x}-x)\| \leq \|A^{-1}\|\|E\|\|\bar{x}\|$$

$$\frac{\|(\bar{x}-x)\|}{\|\bar{x}\|} \leq \|A^{-1}\|\|E\| = \kappa(A)\frac{\|E\|}{\|A\|}$$

Norm-based Bound and Condition Number

We are interested in a neighborhood of nonsingular matrices around A and relative error in x:

Theorem 6.1. *If* $\|A^{-1}E\| < 1$ *then* $(A+E)^{-1}$ *exists and*

$$\frac{\|\bar{x}-x\|}{\|x\|} \leq \frac{\|A^{-1}E\|}{1-\|A^{-1}E\|}$$

and if $\mu = \|A^{-1}\|\|E\| < 1$ *then*

$$\frac{\|\bar{x}-x\|}{\|x\|} \leq \frac{1}{1-\mu}\kappa(A)\frac{\|E\|}{\|A\|} = \frac{1}{1-\kappa(A)\frac{\|E\|}{\|A\|}}\kappa(A)\frac{\|E\|}{\|A\|}$$

The Algorithm

We have a three stage process:

1. Factor $A = LU$
2. Solve $Ly = b$
3. Solve $Ux = y$

- With partial or complete pivoting L is well-conditioned and elements are bounded by 1 in magnitude so (2) is solved accurately.
- Ill-conditioning can show up in U but it tends to be artificial so solving (3) does not magnify error.

Factorization Error

The main source of error is the $O(n^3)$ computation of the factorization.

Theorem 6.2. *(Higham 02) The computed factorization yields* $LU = A + E$ *where componentwise*

$$|E| \quad \leq \quad \omega_n |L||U|$$

where, $\tilde{u} = u/0.9$ *and* $\omega_n = \frac{n\tilde{u}}{1-n\tilde{u}}$.

- The n is pessimistic and tends not to show up except in the elements in the lower right of the matrix.
- It can be replaced by a small constant for matrices with certain structure, e.g., banded.

Norm-based Bound on Factorization Error

Theorem 6.3. *(Higham 02) The computed factorization yields* $LU = A + E$ *where*

$$\frac{\|E\|}{\|A\|} \leq \omega_n \gamma_\epsilon$$

where, $\tilde{u}_M = u/0.9$, $\omega_n = \frac{n\tilde{u}}{1-n\tilde{u}}$ *and* γ_ϵ *is a growth factor.*

Error for the Solution of the System

Theorem 6.4. *(Higham 02) The computed solution $\bar{x}$ using the three stage algorithm with $LU = A + E$ satisfies*

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

where

$$\begin{aligned} |H| &\leq (3\omega_n + \omega_n^2)|L||U| \leq \omega_{3n}|L||U| \\ \frac{\|H\|}{\|A\|} &\leq (3\omega_n + \omega_n^2)\gamma_\epsilon \leq \omega_{3n}\gamma_\epsilon \end{aligned}$$

Growth Factor

- Ideally, since each element of A is involved in $O(n)$ operations the componentwise relative error should be proportional to $n\tilde{u}$, i.e.,

$$|E| \leq \phi_n \tilde{u} |A|$$

  where $\phi_n = O(n)$.

- This can be shown to occur when $|L||U| = |A|$, i.e., no element growth.

- This can be guaranteed for matrices with certain algebraic properties.

Growth Factor

- Size of elements of $|L||U|$ important not the size of the multipliers λ_{ij}.
- Multipliers of $O(1)$ can still have exponential growth factor.
- Small pivots **do not necessarily imply ill-conditioning**.

$$\begin{pmatrix} \epsilon & 1 \\ 1 & 0 \end{pmatrix} \rightarrow \frac{\||L||U|\|}{\|A\|} \sim \epsilon^{-1}$$

Growth Factor and Norm-based Bounds

- $\gamma_n = \frac{|||\tilde{L}||\tilde{U}|||}{||A||} \geq 1$ where $A = \tilde{L}\tilde{U}$ is the exact (or analytical) growth factor
- $\gamma_\epsilon = \frac{|||L_\epsilon||U_\epsilon|||}{||A||} \approx \gamma_n$ where $A + E = L_\epsilon U_\epsilon$ is the computed growth factor
- γ_n used in stability analysis. γ_ϵ is not easily computed during the factorization.
- Wilkinson growth factor: the maximum element in the transformed matrices relative to the maximum element in A.

$$\rho_n = \frac{max|\alpha_{ij}^{(k)}|}{max|\alpha_{ij}^{(1)}|}$$

Growth and Algebraic Properties

We have $\gamma_n = O(1)$ for

- A symmetric positive definite, i.e., $x^T Ax > 0$ for $x \neq 0$, then $\gamma_n = 1$
- A diagonally dominant by columns, i.e., $|\alpha_{ii}| \geq \Sigma_{j=1,i\neq j}^{n} |\alpha_{ji}|$ for $i = 1, \ldots, n$, then $\gamma_n \leq 2$
- A diagonally dominant by rows, i.e., $|\alpha_{ii}| \geq \Sigma_{j=1,i\neq j}^{n} |\alpha_{ij}|$ is also stable.
- A totally positive, i.e., every square submatrix has positive determinant, which also implies L and U both positive then $\gamma_n = 1$.

Growth and Algebraic Properties

- A is a tridiagonal matrix then $\gamma_n \leq 2$
- A is an Hessenberg matrix then $\gamma_n \leq n$
- A is nonsingular then partial pivoting yields $\gamma_n \leq 2^{n-1}$ which is achievable but only rarely.
- A is nonsingular then complete pivoting yields $\gamma_n \leq n^{1/2}(2 \times 3^{1/2} \times \cdots n^{1/(n-1)})^{1/2}$. Not that small but tends to be n or a bit larger in practice.

Factorization

$$A = \begin{pmatrix} \epsilon & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \epsilon^{-1} & 1 \end{pmatrix} \begin{pmatrix} \epsilon & -1 \\ 0 & 1+\epsilon^{-1} \end{pmatrix} = \tilde{L}\tilde{U} \text{ exact}$$

$$LU = \begin{pmatrix} 1 & 0 \\ \epsilon^{-1} & 1 \end{pmatrix} \begin{pmatrix} \epsilon & -1 \\ 0 & \epsilon^{-1} \end{pmatrix} = \begin{pmatrix} \epsilon & -1 \\ 1 & 0 \end{pmatrix} = A + E \text{ computed}$$

$$|\tilde{L}||\tilde{U}| = \begin{pmatrix} \epsilon & 1 \\ 1 & 2\epsilon^{-1}+1 \end{pmatrix} \quad \text{and} \quad |L||U| = \begin{pmatrix} \epsilon & 1 \\ 1 & 2\epsilon^{-1} \end{pmatrix}$$

Growth very large.

Factorization

$$PA = \begin{pmatrix} 1 & 1 \\ \epsilon & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \epsilon & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & -1-\epsilon \end{pmatrix} = \tilde{L}\tilde{U} \text{ exact}$$

$$LU = \begin{pmatrix} 1 & 0 \\ \epsilon & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \epsilon & -1+\epsilon \end{pmatrix} = PA + E \text{ computed}$$

$$|\tilde{L}||\tilde{U}| = \begin{pmatrix} 1 & 1 \\ \epsilon & 1+2\epsilon \end{pmatrix} \quad \text{and} \quad |L||U| = \begin{pmatrix} 1 & 1 \\ \epsilon & 1+\epsilon \end{pmatrix}$$

Growth factor is essentially 1.

Perturbation Theorem for Least Squares

- $A \in \mathbb{R}^{m \times k}$ with linearly independent columns.
- $E \in \mathbb{R}^{m \times k}$ is a perturbation matrix.
- $b \in \mathbb{R}^m$
- E_1 projection of E onto $\mathcal{R}(A)$
- b_1 projection of b onto $\mathcal{R}(A)$
- E_2 projection of E onto $\mathcal{R}^{\perp}(A)$
- b_2 projection of b onto $\mathcal{R}^{\perp}(A)$

The vector $x = A^{\dagger}b$ is the minimizer of

$$\min_x \|b - Ax\|_2$$

What is the sensitivity of the problem and the minimizer to E?

Conditioning of Least Squares Problems

Theorem 6.5. *(Stewart 1973) If*

$$\|A^{\dagger}\|_2\|E_1\|_2 < \frac{1}{2}$$

then the columns of $A+E$ are linearly independent and the minimizer of the perturbed problem $\bar{x} = (A+E)^{\dagger}b$ satisfies

$$\frac{\|x-\bar{x}\|_2}{\|x\|_2} \leq 2\kappa\frac{\|E_1\|_2}{\|A\|_2} + 4\kappa^2\frac{\|b_2\|_2}{\|b_1\|_2}\frac{\|E_2\|_2}{\|A\|_2} + 8\kappa^3\frac{\|E_2\|_2^2}{\|A\|_2}$$

where $\kappa = \|A\|_2\|A^{\dagger}\|_2$ and it can be shown that $\kappa^2 = \kappa(A^TA)$. Note $A^{\dagger} = (A^TA)^{-1}A^T$ when A is full rank.

Conditioning Comments

- The third term can usually be ignored due to the square.
- The first term is like our bound on solving systems and says that the part of the relative error that is in $\mathcal{R}(A)$ is expanded by the condition number.
- The second term, the part of the relative error that is in $\mathcal{R}(A)$, has a larger expansion factor, κ^2, but is weighted by the size of the projection of b that is orthogonal to $\mathcal{R}(A)$.

Householder Computations

Householder reflector-based computations tend to be very reliable.

- Construction of H_i is stable.
- Application of a sequence of H_i to a vector is stable.
- Application of a sequence of H_i to a matrix is stable.
- The computed Q factorization yields a stable backward error.
- The computed residual and solution yield a stable backward error.

Least Squares

Theorem 6.6. *(Stewart 98) Let $A \in \mathbb{R}^{n\times k}$, $n > k$, have full rank and suppose the problem $\|b - Ax\|_2$ is solved using Householder reflectors.*

The computed solution satisfies:

$$\frac{\|\hat{x}_{min} - x_{min}\|_2}{\|x_{min}\|_2} \leq 2p(n,k)\kappa_2(A)u + p(n,k)\kappa_2^2(A)\frac{\|r\|_2}{\|A\|_2\|b\|_2}u$$

Least Squares

- $cond_2(A^T) \leq n\kappa_2(A)$
- If $cond_2(A^T)$ not too large then $(1 + n\tilde{\omega}_{nk} cond_2(A^T)) = O(1)$
- Computed residual is therefore close to the true residual plus a constant multiple of the floating point error in evaluating $fl(r)$.
- The relative error in the solution is essentially as good as one would expect given the conditioning of the problem, i.e., how much of b is in $\mathcal{R}(A)$.