2.5. Accuracy Estimation

• Reading: Trefethen and Bau (1997), Lecture 22

• Accuracy:
  – With roundoff, how accurate is a solution obtained by Gaussian elimination?
  – How sensitive is a solution to perturbations introduced by round off error (stability)?
  – How can we appraise the accuracy of a computed solution?
  – Can the accuracy of a computed solution be improved?

• Let \( \hat{x} \) be a computed solution of

\[
Ax = b
\]

(1)

– Compute the residual

\[
r = b - A\hat{x}
\]

(2)

– Is \( r \) a good measure of the accuracy of \( \hat{x} \)?

• Example 1. Solve (1) with

\[
A = \begin{bmatrix}
1.2969 & 0.8648 \\
0.2161 & 0.1441
\end{bmatrix}, \quad b = \begin{bmatrix}
0.8642 \\
0.1440
\end{bmatrix}
\]

– An approximate solution is \( \hat{x} = [0.9911, -0.4870]^T \)
– The residual is \( r = [-10^{-8}, 10^{-8}]^T \)
– Exact solution:
  – \( \det(A) = 10^{-8} \)
Sensitivity

- *Example 2.* Consider

\[ A = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix} \]

- \( A \) is symmetric and positive definite with \( \det(A) = 1 \)
- Exact solutions

<table>
<thead>
<tr>
<th>( b^T )</th>
<th>( x^T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[23, 32, 33, 31]</td>
<td>[1, 1, 1, 1]</td>
</tr>
<tr>
<td>[22.9, 32.1, 32.9, 31.1]</td>
<td>[−7.2, 6, 2.9, 0.1]</td>
</tr>
<tr>
<td>[22.99, 32.01, 32.99, 31.01]</td>
<td>[0.18, 1.5, 1.19, 0.89]</td>
</tr>
</tbody>
</table>
Perturbations

• **Definition 1:** A matrix is *ill-conditioned* if small changes in $A$ or $b$ produce large changes in the solution.

  - Suppose that $b$ is perturbed to $b + \delta b$
  - The solution $x + \delta x$ of the perturbed system would satisfy

$$A(x + \delta x) = b + \delta b$$

  - Estimate the magnitude of $\delta x$

$$\delta x = A^{-1} \delta b$$

  - Taking a norm

$$||\delta x|| = ||A^{-1} \delta b|| \leq ||A^{-1}|| ||\delta b||$$

  - In addition

$$||b|| = ||Ax|| \leq ||A|| ||x||$$

  - Multiplying the two inequalities

$$||\delta x|| ||b|| \leq ||A|| ||A^{-1}|| ||\delta b|| ||x||$$

or

$$\frac{||\delta x||}{||x||} \leq \kappa(A) \frac{||\delta b||}{||b||} \quad (3a)$$

where

$$\kappa = ||A|| ||A^{-1}|| \quad (3b)$$

* $\kappa(A)$ is called the *condition number*
The Condition Number

• Note:
  i. The condition number relates the relative uncertainty of the data ($||\delta b||/||b||$) to the relative uncertainty of the solution ($||\delta x||/||x||$)
  ii. If $\kappa(A)$ is large, small changes in $b$ may result in large changes in $x$
  iii. If $\kappa(A)$ is large, $A$ is ill-conditioned
  iv. The bound (3a) is sharp: there are $b$s for which equality holds
  v. The det($A$) has nothing to do with conditioning. The matrix

$$
\begin{bmatrix}
10^{-30} & 0 \\
0 & 10^{-30}
\end{bmatrix}
$$

is well-conditioned

• Repeat the analysis for a perturbation in $A$

$$(A + \delta A)(x + \delta x) = b$$

or

$$Ax + A\delta x + \delta A(x + \delta x) = b$$

or

$$\delta x = -A^{-1}\delta A(x + \delta x)$$

or

$$||\delta x|| \leq ||A^{-1}||||\delta A|| ||x + \delta x||$$

or

$$\frac{||\delta x||}{||x + \delta x||} \leq \kappa(A) \frac{||\delta A||}{||A||}$$  \hspace{1cm} (3c)
Condition Numbers

• Example 3. The matrix $A$ of Example 1 and its inverse are

$$A = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}, \quad A^{-1} = 10^8 \begin{bmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{bmatrix}$$

Calculate

$$||A||_\infty = 2.1617, \quad ||A^{-1}||_\infty = 1.5130 \times 10^8$$

$$\kappa_\infty(A) = 3.2707 \times 10^8$$

• Example 4. The matrix $A$ of Example 2 and its inverse are

$$A = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} 68 & -41 & -17 & 10 \\ -41 & 25 & 10 & -6 \\ -17 & 10 & 5 & -3 \\ 10 & -6 & -3 & 2 \end{bmatrix}$$

and

$$||A||_\infty = 33, \quad ||A^{-1}||_\infty = 136$$

$$\kappa_\infty(A) = 4488$$
Stability of Gaussian Elimination

- How do rounding errors propagate in Gaussian elimination?
  - Let’s discuss this without explicit treatment of pivoting
    * Assume that \( A \) has its rows arranged so that pivoting is unnecessary

- **Theorem 1**: Let \( A \) be an \( n \times n \) real matrix of floating-point numbers on a computer with unit round-off error \( u \). Assume that no zero pivots are encountered. Let \( \hat{L} \) and \( \hat{U} \) be the computed triangular factors of \( A \) and let \( \hat{y} \) and \( \hat{x} \) be the computed solutions of

\[
\hat{L}y = b, \quad \hat{U}x = \hat{y}
\]  

(4a)

Then

\[
(A + \delta A)\hat{x} = b
\]  

(4b)

where

\[
|\delta A| \leq 3nu|L||U| + O(n^2u^2)
\]  

(4c)

- **Remark 1**: Recall (Section 2.1) that \( |A| \) is a matrix with elements \( |a_{ij}| \), \( i, j = 1:n \)

- **Remark 2**: The computed solution \( \hat{x} \) is the exact solution of a system \( A + \delta A \)

- **Remark 3**: Additionally, \( \hat{L} \) and \( \hat{U} \) are the exact factors of a perturbed matrix \( A + E \)

- **Proof**: Follow the arguments used in the round-off error analysis of the dot product of Section 2.1 (cf. Demmel, Section 2.4) \( \square \)
Growth Factor

- **Remark:** For the infinity, one and Frobenius norms,
  \[ \|z\|_\infty = \|z\| \]

  - Taking a norm of (4c)
  \[ \|\delta A\| \leq 3nu\|L\|\|U\| + O(n^2u^2) \] (5)

- We would like to estimate \(\|L\|\) and \(\|U\|\)

  - With partial pivoting, the elements of \(L\) are bounded by unity
    * Thus, \(\|L\|_\infty \leq n\)
  - Bounding \(\|U\|\) is much more difficult
    * Define the *growth factor* as
      \[ \rho = \frac{\max_{i,j} |u_{ij}|}{\|A\|_\infty} \] (6)
    * Then \(|U| \leq \rho\|A\|\) and
      \[ \|U\|_\infty \leq \rho n\|A\|_\infty \]
  - Using the bounds for \(\|L\|_\infty\) and \(\|U\|_\infty\) in (5), we find
    \[ \|\delta A\|_\infty \leq n^3u\rho C'\|A\|_\infty \] (7)
    * The constant \(C'\) accounts for our “casual” treatment of the \(O(n^2u^2)\) term
Growth Factors

• Note:
  
i. The bound (7) for $\delta A$ is acceptable unless $\rho$ is large
  
  ii. J.H. Wilkinson (1961) \(^1\) shows that if $|a_{ij}| \leq 1$, $i, j = 1: n$, 
  
  $\rho ||A||_{\infty} \leq 2^{n-1}$ with partial pivoting 
  
  $\rho ||A||_{\infty} \leq 1.8 n^{0.25 \ln n}$ with complete pivoting 
  
  - The bound with partial pivoting is sharp! Consider 

  \[
  A = \begin{bmatrix}
  1 & 1 \\
  -1 & 1 & 1 \\
  -1 & -1 & 1 & 1 \\
  -1 & -1 & -1 & 1 & 1 \\
  -1 & -1 & -1 & -1 & 1 \end{bmatrix}
  \]

  which has 

  \[
  L = \begin{bmatrix}
  1 & & & & \\
  -1 & 1 \\
  -1 & -1 & 1 \\
  -1 & -1 & -1 & 1 \\
  -1 & -1 & -1 & -1 & 1 \end{bmatrix}, \quad 
  U = \begin{bmatrix}
  1 & 1 \\
  & 2 \\
  & 4 \\
  & 8 \\
  & 16 \end{bmatrix}
  \]

  Thus, $\rho = 16$. For an $n \times n$ matrix, $\rho = 2^{n-1}$ 

  iii. In most cases, $\rho ||A||_{\infty} \leq 8$ with partial pivoting 

  iv. Our bounds are conservative. In most cases 

  $||\delta A||_{\infty} \leq n u ||A||_{\infty}$

\(^1\)J.H. Wilkinson, “Error analysis of direct methods of matrix inversion,” J. A.C.M., 8, 281-330
Iterative Improvement

- Iterative improvement can enhance the accuracy of a computed solution
  - It works best when double precision arithmetic is available but costs much more than single precision arithmetic
- Let $\hat{x}$ be an approximate solution of (1).
- Calculate the residual (2)
  \[ r = b - A\hat{x} = A(x - \hat{x}) \]
- Solve
  \[ A\delta x = r \]  
  for $\delta x$ and calculate an “improved” solution
  \[ x = \hat{x} + \delta x \]  
- $\hat{x}$ cannot be improved unless:
  - The residual is calculated in double precision
  - The original (not the factored) $A$ is used to calculate $r$
- The procedure is
  1. Calculate $r = b - A\hat{x}$ using double precision
  2. Solve $\hat{L}y = Pr$ by forward substitution
  3. Solve $\hat{U}\delta x = y$ by backward substitution
  4. $x = \hat{x} + \delta x$
Iterative Improvement

• Notes:
  
  i. Round the double precision residual to single precision
  
  ii. The cost of iterative improvement:
      
          - One matrix-by-vector multiplication for the residual
            and one forward and backward substitution
          
          - Each cost $n^2$ multiplications if double precision hard-
            ware is available
          
          - The total cost of $2n^2$ multiplications is small relative
            to the factorization
  
  iii. Continue the iteration for further improvement
  
  iv. Have to save the original $A$ and the factors $L$ and $U$
  
  v. Have to do mixed-precision arithmetic

• Accuracy of the improved solution
  
      - From (8)

        \[ \delta x := x - \hat{x} = A^{-1}(b - A\hat{x}) = A^{-1}r \]

      - Take a norm

        \[ \|\delta x\|_\infty \leq \|A^{-1}\|_\infty \|r\|_\infty \]

      - $\hat{x}$ satisfies (4c), so

        \[ r = (A + \delta A)\hat{x} - A\hat{x} = \delta A\hat{x} \]

      - Take a norm

        \[ \|r\|_\infty \leq \|\delta A\|_\infty \|\hat{x}\|_\infty \]
Still Improving

- Combining results

\[ \| \delta x \|_\infty \leq \| A^{-1} \|_\infty \| \delta A \|_\infty \| \hat{x} \|_\infty \]

- Using (3b)

\[ \frac{\| \delta x \|_\infty}{\| \hat{x} \|_\infty} \leq \kappa(A) \frac{\| \delta A \|_\infty}{\| A \|_\infty} \quad (9a) \]

- We could estimate \( \| \delta A \|_\infty \) using (7)
  * Golub and Van Loan (1996) suggest

\[ \| \delta A \|_\infty = nu \| A \|_\infty \]

* Trefethen and Bau (1997) suggest \( \rho = O(n^{1/2}) \)

\[ \frac{\| \delta x \|_\infty}{\| \hat{x} \|_\infty} \leq n u \kappa(A) \quad (9b) \]

- If \( u \propto \beta^{1-t} \) and \( \kappa(A) \propto \beta^{s-1} \)

\[ \frac{\| \delta x \|_\infty}{\| \hat{x} \|_\infty} \leq n c \beta^{s-t} \quad (9c) \]

- \( c \) represents constants arising from \( u \) and \( \kappa \)

- The computed solution \( \hat{x} \) will have about \( t - s \) correct \( \beta \)-digits
  - If \( s \) is small, \( A \) is well conditioned and the error is small
  - If \( s \) is large, \( A \) is ill-conditioned and the error is large
  - If \( s > t \), the error grows relative to \( \hat{x} \)
It Don’t Get Much Better

• Let $\hat{x}^{(0)} = \hat{x}$, $\delta x^{(0)} = \delta x$, and $\hat{x}^{(1)} = \hat{x}^{(0)} + \delta x^{(0)}$ be the solution computed by iterative improvement.

• Assume $\delta x \approx \delta x^{(0)}$ and Calculate an estimate of $\kappa(A)$ from (9b) as

$$\frac{1}{nu} \frac{||\delta x^{(0)}||_\infty}{||\hat{x}^{(0)}||_\infty} \leq \kappa(A)$$

• An empirical result that holds widely is

$$\frac{1}{n} \kappa(A) \leq \frac{1}{nu} \frac{||\delta x^{(0)}||_\infty}{||\hat{x}||_\infty} \leq \kappa(A) \quad (10)$$

• **Theorem 2:** Suppose

$$r^{(k)} = b - A\hat{x}^{(k)}, \quad \hat{x}^{(k+1)} = \hat{x}^{(k)} + \delta x^{(k)} \quad (11)$$

are calculated in double precision and $\delta x^{(k)}$ is the solution of

$$A\delta x^{(k)} = r^{(k)}$$

in single precision. Thus,

$$(A + \delta A^{(k)})\delta x^{(k)} = r^{(k)}$$

If $||A^{-1}\delta A^{(k)}|| \leq 1/2$ then $||x^{(k)} - \hat{x}|| \to 0$ as $k \to \infty$

• **Corollary 2:** If $||\delta A^{(k)}|| \leq nu||A||$ and $nu\kappa(A) \leq 1/2$ then $||x^{(k)} - \hat{x}|| \to 0$ as $k \to \infty$


• **Remark:** Iterative improvement converges rather generally and each iteration gives $t - s$ correct $\beta$ digits
A Hilbert Matrix

- Example 5. An \( n \times n \) Hilbert matrix has entries

\[
h_{ij} = \frac{1}{i + j - 1}
\]

- With \( n = 4 \):

\[
H = \begin{bmatrix}
1 & 1/2 & 1/3 & 1/4 \\
1/2 & 1/3 & 1/4 & 1/5 \\
1/3 & 1/4 & 1/5 & 1/6 \\
1/4 & 1/5 & 1/6 & 1/7 \\
\end{bmatrix}
\]

- It is symmetric and positive definite
- It arises in the polynomial approximation of functions

- Condition numbers of Hilbert matrices

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \kappa_2(H) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.9282e+01</td>
</tr>
<tr>
<td>4</td>
<td>1.5514e+04</td>
</tr>
<tr>
<td>8</td>
<td>1.5258e+10</td>
</tr>
<tr>
<td>16</td>
<td>6.5341e+17</td>
</tr>
</tbody>
</table>

- Solve \( Hx = b \) with \( n = 12 \)
  
  - Select \( b \) the first column of the identity matrix
  - \( x \) is the first column of \( H^{-1} \)

- Solve by Gaussian elimination with one step of iterative improvement

\[
\frac{||x^{(0)} - x||_2}{||x||_2} = 0.0784, \quad \frac{||x^{(1)} - x||_2}{||x||_2} = 0.0086
\]