Unit 1.3 Error Bounds

Numerical Analysis

EE/NTHU

Mar. 6, 2017

Error Bounds

- The linear system can be solved accurately using direct methods

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

if $\mathbf{A}$ and $\mathbf{b}$ are exact.

- In the real world, the right hand side $\mathbf{b}$ may not be exact. In this case, the solution is not exact either. We have solved

$$\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}, \tag{1.3.1}$$

then

$$\delta\mathbf{x} = \mathbf{A}^{-1}\delta\mathbf{b} \tag{1.3.2}$$

and

$$\|\delta\mathbf{x}\| \le \|\mathbf{A}^{-1}\|\|\delta\mathbf{b}\| \tag{1.3.3}$$

- For relative error, $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$, we have

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{\|\mathbf{A}^{-1}\|\|\delta\mathbf{b}\|}{\|\mathbf{x}\|} = \|\mathbf{A}^{-1}\|\|\delta\mathbf{b}\|\left\|\frac{\mathbf{A}}{\mathbf{b}}\right\| \le \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \tag{1.3.4}$$

- $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ is the condition number

Condition Number Properties

- Note that the condition number of a matrix $\mathbf{A}$ can be defined with any matrix norm $\|\mathbf{A}\|$. Some popular norms are 1-norm, 2-norm and ∞-norm:
  - $\kappa_1(\mathbf{A}) = \|\mathbf{A}\|_1 \|\mathbf{A}^{-1}\|_1$,
  - $\kappa_2(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$,
  - $\kappa_\infty(\mathbf{A}) = \|\mathbf{A}\|_\infty \|\mathbf{A}^{-1}\|_\infty$,

Property 1.3.1. Condition Number

1. $\kappa(\mathbf{A}) \geq 1$,
2. $\kappa(\mathbf{A}^{-1}) = \kappa(\mathbf{A})$,
3. For all $\alpha \in \mathbb{C}$ and $\alpha \neq 0$, $\kappa(\alpha\mathbf{A}) = \kappa(\mathbf{A})$,
4. If $\mathbf{A}$ is orthogonal, $\kappa_2(\mathbf{A}) = 1$.
5. The condition number of a singular matrix is set equal to infinity.
6. $\kappa_2(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 = \dfrac{\sigma_1(\mathbf{A})}{\sigma_n(\mathbf{A})}$
   where $\sigma_1(\mathbf{A})$ and $\sigma_n(\mathbf{A})$ are the maximum and minimum singular values of $\mathbf{A}$. If $\mathbf{A}$ is symmetric and positive definite
   $$\kappa_2(\mathbf{A}) = \frac{\lambda_{max}}{\lambda_{min}}. \tag{1.3.5}$$

Condition Number Properties, II

- Note that
$$1 = \|\mathbf{I}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| = \kappa(\mathbf{A}).$$

Theorem 1.3.2.

Define the relative distance of $\mathbf{A}$ from the set of singular matrices with respect to the p-norm by
$$\text{dist}_p(\mathbf{A}) = \min\left\{\frac{\|\delta\mathbf{A}\|_p}{\|\mathbf{A}\|_p} : \mathbf{A} + \delta\mathbf{A} \text{ is singular}\right\}. \tag{1.3.6}$$
Then
$$\text{dist}_p(\mathbf{A}) = \frac{1}{\kappa_p(\mathbf{A})}. \tag{1.3.7}$$

- Thus, if the condition number of a matrix $\mathbf{A}$ is large, then $\mathbf{A}$ is close to being singular.

Condition Number Properties, III

Corollary 1.3.3.

$\mathbf{A} + \delta\mathbf{A}$ is nonsingular if

$$\|\delta\mathbf{A}\|_p < \frac{1}{\|\mathbf{A}^{-1}\|_p}. \tag{1.3.8}$$

Proof: Note $\mathbf{A} + \delta\mathbf{A}$ is nonsingular if

$$\frac{\|\delta\mathbf{A}\|_p}{\|\mathbf{A}\|_p} < \text{dist}_p(\mathbf{A}) = \frac{1}{\kappa_p(\mathbf{A})} = \frac{1}{\|\mathbf{A}\|_p\|\mathbf{A}^{-1}\|_p} \tag{1.3.9}$$

and

$$\|\delta\mathbf{A}\|_p < \frac{1}{\|\mathbf{A}^{-1}\|_p}.$$

- Note that this holds for all matrix norm $\|\mathbf{A}\|_p$.
- Equation (1.3.8) can also be written as

$$\|\mathbf{A}^{-1}\|_p\|\delta\mathbf{A}\|_p < 1. \tag{1.3.10}$$

Error bounds, II

- Thus, we have

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \kappa(\mathbf{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \tag{1.3.11}$$

- Any relative change in the right hand side, the relative change in the solution is amplified by the condition number $\kappa(\mathbf{A})$.
- For symmetric and positive definite matrix $\mathbf{A}$

$$\kappa_2(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\| = \frac{\lambda_{max}}{\lambda_{min}}. \tag{1.3.12}$$

- Thus, the solution accuracy depends on the accuracy of the right hand side vector $\mathbf{b}$ and the condition number, which is a property of the matrix $\mathbf{A}$ alone.
- Note that for a positive definite matrix $\mathbf{A}$, all its eigenvalues are positive.
- If the spread of the eigen values are small, i.e., $\lambda_{max} \approx \lambda_{min}$, then $\kappa(\mathbf{A}) \approx 1$, and the relative solution accuracy tracks the right hand side accuracy.
- If $\lambda_{max} \gg \lambda_{min}$, then any small error in $\mathbf{b}$ will result in a very different solution $\mathbf{x}$.

Error bounds – Example

- Given the linear system

$$\begin{bmatrix} 12 & 0.1 \\ 10 & 0.1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 5.1 \end{bmatrix}$$

  The solution is $\mathbf{x} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$.

- But if the right hand side is slightly different

$$\begin{bmatrix} 12 & 0.1 \\ 10 & 0.1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

  The solution is $\mathbf{x} = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$.

- The solutions are significantly different
- The eigenvalues of the matrix are: $\lambda_1 = 12.0834$, $\lambda_2 = 0.0165516$ and $\lambda_1/\lambda_2 = 730.04$.

Error bounds – Example 2

- Given another linear system

$$\begin{bmatrix} 12 & 0.1 \\ 0.1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 10.05 \end{bmatrix}$$

  The solution is $\mathbf{x} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$

- But if the right hand side is slightly different

$$\begin{bmatrix} 12 & 0.1 \\ 0.1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 9.9 \end{bmatrix}$$

  The solution is $\mathbf{x} = \begin{bmatrix} 0.492 \\ 0.985 \end{bmatrix}$

- The solutions are not too different
- The eigenvalues of the matrix are: $\lambda_1 = 12.005$, $\lambda_2 = 9.99501$ and $\lambda_1/\lambda_2 = 1.201$.

Error bounds, III

- It is also possible that the matrix $\mathbf{A}$ is inexact when we solve the linear system

$$(\mathbf{A}+\delta\mathbf{A})(\mathbf{x}+\delta\mathbf{x}) = \mathbf{b} \tag{1.3.13}$$

and we wish to know $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$.

Lemma 1.3.4.

If $\mathbf{F}$ is an $n \times n$ matrix with $\|\mathbf{F}\| < 1$, then $(\mathbf{I}+\mathbf{F})^{-1}$ exists and satisfies

$$\|(\mathbf{I}+\mathbf{F})^{-1}\| \leq \frac{1}{1-\|\mathbf{F}\|}.$$

Proof:

$$\|(\mathbf{I}+\mathbf{F})\mathbf{x}\| = \|\mathbf{x}+\mathbf{F}\mathbf{x}\| \geq \|\mathbf{x}\| - \|\mathbf{F}\mathbf{x}\| \geq (1-\|\mathbf{F}\|)\|\mathbf{x}\| > 0$$

Thus, $\mathbf{I}+\mathbf{F}$ is nonsingular. Let $\mathbf{C} = (\mathbf{I}+\mathbf{F})^{-1}$ then

$$1 = \|\mathbf{I}\| = \|(\mathbf{I}+\mathbf{F})\mathbf{C}\| = \|\mathbf{C}+\mathbf{F}\mathbf{C}\| \geq \|\mathbf{C}\| - \|\mathbf{C}\|\|\mathbf{F}\| = \|\mathbf{C}\|(1-\|\mathbf{F}\|). \quad \square.$$

Error bounds, IV

Theorem 1.3.5.

If $\mathbf{A}$ is an $n \times n$ nonsingular matrix, $\|\delta\mathbf{A}\| < \|\mathbf{A}\|$ and $\mathbf{x}$ and $\delta\mathbf{x}$ satisfy $\mathbf{A}\mathbf{x} = \mathbf{b}$, $(\mathbf{A}+\delta\mathbf{A})(\mathbf{x}+\delta\mathbf{x}) = \mathbf{b}$ then

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\delta\mathbf{A}\|}{1-\|\mathbf{A}^{-1}\delta\mathbf{A}\|}, \tag{1.3.14}$$

also

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1-\kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}. \tag{1.3.15}$$

Proof: From

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

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

Error bounds, V

$$
\begin{aligned}
\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} &\le \|(\mathbf{A}+\delta\mathbf{A})^{-1}[\mathbf{A}-(\mathbf{A}+\delta\mathbf{A})]\| \\
&= \|(\mathbf{A}+\delta\mathbf{A})^{-1}\delta\mathbf{A})\| \\
&= \|(\mathbf{I}+\mathbf{A}^{-1}\delta\mathbf{A})^{-1}\mathbf{A}^{-1}\delta\mathbf{A}\| \\
&\le \|(\mathbf{I}+\mathbf{A}^{-1}\delta\mathbf{A})^{-1}\|\|\mathbf{A}^{-1}\delta\mathbf{A}\| \\
&\le \frac{\|\mathbf{A}^{-1}\delta\mathbf{A}\|}{\|\mathbf{I}-\mathbf{A}^{-1}\delta\mathbf{A}\|} \le \frac{\|\mathbf{A}^{-1}\delta\mathbf{A}\|}{1-\|\mathbf{A}^{-1}\delta\mathbf{A}\|}
\end{aligned}
$$

Then since

$$
\begin{aligned}
\|\mathbf{A}^{-1}\delta\mathbf{A}\| \le \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\| \\
&= \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|/\|\mathbf{A}\| \\
&= \kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}
\end{aligned}
$$

This proves the theorem. □

Error bounds – Example 3

- As the example 1
$$\begin{bmatrix}12 & 0.1\\ 10 & 0.1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}6.1\\ 5.1\end{bmatrix}$$
The solution is $\mathbf{x}=\begin{bmatrix}0.5\\ 1\end{bmatrix}$
- But if the matrix is slightly different
$$\begin{bmatrix}12 & 0\\ 10 & 0.1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}6.1\\ 5.1\end{bmatrix}$$
The solution is $\mathbf{x}=\begin{bmatrix}0.508333\\ 0.166667\end{bmatrix}$
- The solutions are significantly different
- Again, the eigenvalues of the matrix are: $\lambda_1 = 12.0834$, $\lambda_2 = 0.0165516$ and $\lambda_1/\lambda_2 = 730.04$.

Error bounds – Example 4

- Given a different linear system

$$\begin{bmatrix} 12 & 0.1 \\ 0.1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 10.05 \end{bmatrix}$$

  The solution is $\mathbf{x} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$

- But if the matrix is slightly different

$$\begin{bmatrix} 12 & 0 \\ 0.1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 10.05 \end{bmatrix}$$

  The solution is $\mathbf{x} = \begin{bmatrix} 0.508333 \\ 0.999917 \end{bmatrix}$

- The solutions are not very different
- Again, the eigenvalues of the matrix are:
  $\lambda_1 = 12.005$, $\lambda_2 = 9.99501$ and
  $\lambda_1/\lambda_2 = 1.201$.

Error Bounds, VI

- If both $\mathbf{A}$ and $\mathbf{b}$ are inaccurate, we have

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}. \tag{1.3.16}$$

  The following theorem estimates $\delta\mathbf{x}$.

Theorem 1.3.6.

If $\mathbf{A}$ is nonsingular and together with $\delta\mathbf{A}$ satisfy

$$\|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\| < 1. \tag{1.3.17}$$

Then if $\mathbf{x}$ is the solution of $\mathbf{Ax} = \mathbf{b}$, $\mathbf{b} \neq \mathbf{0}$, and $\delta\mathbf{x}$ satisfies equation (1.3.16) for $\delta\mathbf{b}$, then

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\|\delta\mathbf{A}\|/\|\mathbf{A}\|} \left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} \right). \tag{1.3.18}$$

Proof: From Corollary (1.2.6) and Eq. (1.3.17) we have $\mathbf{A} + \delta\mathbf{A}$ nonsingular, so is $\mathbf{I} + \mathbf{A}^{-1}\delta\mathbf{A}$, and

$$\|(\mathbf{I} + \mathbf{A}^{-1}\delta\mathbf{A})^{-1}\| \leq \frac{1}{1 - \|\mathbf{A}^{-1}\delta\mathbf{A}\|} \leq \frac{1}{1 - \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|} \tag{1.3.19}$$

Error Bounds, VII

And from Eq. (1.3.16) and $\mathbf{Ax} = \mathbf{b}$

$$\begin{aligned}\mathbf{Ax} + \delta\mathbf{Ax} + (\mathbf{A} + \delta\mathbf{A})\delta\mathbf{x} &= \mathbf{b} + \delta\mathbf{b} \\ (\mathbf{A} + \delta\mathbf{A})\delta\mathbf{x} &= \delta\mathbf{b} - \delta\mathbf{Ax}\end{aligned}$$

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

Thus,

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\|\delta\mathbf{A}\|/\|\mathbf{A}\|}(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}). \qquad \square$$

Error Bounds, VIII

- Note that Eq. (1.3.18) can be reduced to Eq. (1.3.11) or Eq. (1.3.15) if $\delta\mathbf{A} = \mathbf{0}$ or $\delta\mathbf{b} = \mathbf{0}$.
- In case of $\delta\mathbf{A} = \mathbf{0}$, we can have stronger bounds.

Theorem 1.3.7.

Assume the conditions of Theorem (1.2.9) holds and $\delta\mathbf{A} = \mathbf{0}$ then

$$\frac{1}{\kappa(\mathbf{A})}\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \le \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \kappa(\mathbf{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{1.3.20}$$

Proof. The second inequality has been proven before.
Since

$$\begin{aligned}\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) &= \mathbf{b} + \delta\mathbf{b} \\ \delta\mathbf{b} &= \mathbf{A}\delta\mathbf{x} \\ \|\delta\mathbf{b}\| &\le \|\mathbf{A}\|\|\delta\mathbf{x}\| \\ \|\mathbf{x}\|\|\delta\mathbf{b}\| &\le \|\mathbf{x}\|\|\mathbf{A}\|\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\mathbf{b}\|\|\mathbf{A}\|\|\delta\mathbf{x}\| \\ &\le \|\mathbf{A}^{-1}\|\|\mathbf{b}\|\|\mathbf{A}\|\|\delta\mathbf{x}\| = \kappa(\mathbf{A})\|\mathbf{b}\|\|\delta\mathbf{x}\|\end{aligned}$$

Thus,

$$\frac{1}{\kappa(\mathbf{A})}\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \le \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Error Bounds, IX

- In engineering problems, one may not get the exact linear system to solve for
$$\mathbf{Ax} = \mathbf{b}.$$
- The right hand side or the matrix itself could have errors
- In these cases, the solution might not be exact
- The condition number $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ plays an important roll in how inaccurate the solutions are
  - The larger the condition number, the larger the solution errors
  - A property of the matrix itself
  - Solution algorithm cannot improve these errors

Improving Solution Accuracy - Scaling

- The linear system below is known to be very sensitive to the error on the right hand side
$$\begin{bmatrix} 12 & 0.1 \\ 10 & 0.1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 5.1 \end{bmatrix} \tag{1.3.21}$$
- If we let $y_1 = 10x_1$, $y_2 = x_2$ then
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{1.3.22}$$
  And the linear system becomes
$$\begin{bmatrix} 12 & 0.1 \\ 10 & 0.1 \end{bmatrix} \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 5.1 \end{bmatrix}$$
$$\begin{bmatrix} 1.2 & 0.1 \\ 1 & 0.1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 5.1 \end{bmatrix} \tag{1.3.23}$$
  The solution is
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

Improving Solution Accuracy - Scaling, II

- If the right hand side is perturbed a little

$$\begin{bmatrix} 12 & 0.1 \\ 10 & 0.1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

- Setting, again, $y_1 = 10x_1$ and $y_2 = x_2$, we obtain

$$\begin{bmatrix} 1.2 & 0.1 \\ 1 & 0.1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

The solution is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}.$$

Though the solution in terms of $\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ is identical as before. The sensitivity of $\mathbf{y}$ appears to be smaller than $\mathbf{x}$. In fact, the matrix of the linear system of $\mathbf{y}$ has the eigenvalues of $\lambda_1 = 1.20834$ and $\lambda_2 = 0.0165516$. And the condition number of $\lambda_1/\lambda_2 = 73.004$.

Improving Solution Accuracy - Scaling, III

- Comparing the matrices of Eqs. (1.3.21) and (1.3.23), it is apparent that the first column of the matrix in (1.3.23) has been divided by 10 – column scaling.
  - The diagonal elements now have the magnitudes closer to each other.
  - The round off errors during LU decomposition will be smaller as well.
- The solution vector, $\mathbf{x}$, can be found by multiplying $\mathbf{y}$ with the scaling matrix, as shown in Eq. (1.3.22).
  - The sensitivity of $\mathbf{x}$ to $\mathbf{b}$ is, however, unchanged with scaling.
- It is possible to choose the scaling matrix as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.1 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

Then the linear system becomes

$$\begin{bmatrix} 1.2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 6.1 \\ 5.1 \end{bmatrix}$$

and the solution is

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0.1 \end{bmatrix}$$

Improving Solution Accuracy - Scaling, IV

- The perturbed system is
$$\begin{bmatrix} 1.2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$
and the solution is
$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$
- For the scaled variable, $\mathbf{z}$, the sensitivity to the right hand side is small.
- The eigenvalues and the condition number are
$$\lambda_1 = 21.0499, \lambda_2 = 0.0950124, \kappa(\mathbf{A}) = 22.1549.$$
  - The condition number continues to improve.
- The round-off error during LU factorization and forward and backward substitutions can be further reduced.

Improving Solution Accuracy - Scaling, V

Algorithm 1.3.8. Column scaling.

Given a linear system
$$\mathbf{Ax} = \mathbf{b}$$
find a diagonal matrix $\mathbf{C}$ such that the diagonal entries of $\mathbf{F}=\mathbf{AC}$ have similar magnitudes. Solve the linear system
$$\mathbf{Fy} = \mathbf{b}$$
and then the solution $\mathbf{x}$ can be found by
$$\mathbf{x} = \mathbf{Cy}.$$

- Note the matrix condition number improves for the column scaled matrix $\mathbf{F}$.
- For the example,

$$\begin{bmatrix} 12 & 0.1 \\ 10 & 0.1 \end{bmatrix} \qquad \begin{bmatrix} 1.2 & 0.1 \\ 1 & 0.1 \end{bmatrix} \qquad \begin{bmatrix} 1.2 & 1 \\ 1 & 1 \end{bmatrix}$$

$\lambda_1 = 12.0834,$	$\lambda_1 = 1.28443,$	$\lambda_1 = 2.10499,$
$\lambda_2 = 0.0165516,$	$\lambda_2 = 0.0155711,$	$\lambda_2 = 0.0950124,$
$\kappa(\mathbf{A}) = 730.044.$	$\kappa(\mathbf{A}) = 82.488.$	$\kappa(\mathbf{A}) = 22.1549.$

Improving Solution Accuracy - Scaling, VI

Algorithm 1.3.9. Row scaling.

Given a linear system

$$\mathbf{Ax} = \mathbf{b}$$

find a diagonal matrix $\mathbf{R}$ such that the diagonal entries of $\mathbf{G}$=$\mathbf{RA}$ have similar magnitudes. Solve the linear system

$$\mathbf{Gx} = \mathbf{Rb}$$

to find the solution vector $\mathbf{x}$.

- Note the matrix condition number improves for the row scaled matrix also.
- For the example,

$$\begin{bmatrix} 12 & 10 \\ 0.1 & 0.1 \end{bmatrix} \qquad \begin{bmatrix} 1.2 & 1 \\ 0.1 & 0.1 \end{bmatrix} \qquad \begin{bmatrix} 1.2 & 1 \\ 1 & 1 \end{bmatrix}$$

$\lambda_1 = 12.0834,$ $\lambda_2 = 0.0165516,$ $\kappa(\mathbf{A}) = 730.044.$

$\lambda_1 = 1.28443,$ $\lambda_2 = 0.0155711,$ $\kappa(\mathbf{A}) = 82.488.$

$\lambda_1 = 2.10499,$ $\lambda_2 = 0.0950124,$ $\kappa(\mathbf{A}) = 22.1549.$

Improving Solution Accuracy - Scaling, VII

- Scaling techniques can be applied to improve solution accuracy.
  - Round-off error reduced during LU decomposition and forward, backward substitution processes.
  - Sensitivity of the scaled variable to the right hand side can be reduced, but unchanged for the solution vector.
  - Column and row scaling have the same effects.
- Column scaling applies when one of the unknowns in the solution vector has a different magnitude from other unknowns.
  - For example, length measured in μm and voltage in V.
  - The original solution can be found after applying the scaling matrix again.
- Row scaling applied when an equation in the linear system has different magnitudes from other equations.
  - Row scaling should scale the right hand side simultaneously, and the solution can be obtained directly.
- Column scaling and row scaling can be combined.

Iterative Refinement

- In Unit 1.1, we have shown that if the computer's number system has few significant digits, then even LU decomposition method can result in significant errors.
- Example, using 4-digit decimal system to solve

$$\begin{bmatrix} 0.001 & 2.42 \\ 1 & 1.58 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5.2 \\ 4.57 \end{bmatrix}$$

We get the solution

$$\mathbf{x}^{(0)} = \begin{bmatrix} 2 \\ 2.148 \end{bmatrix}$$

Note that

$$\mathbf{A}\mathbf{x}^{(0)} = \begin{bmatrix} 5.2 \\ 5.394 \end{bmatrix}$$

And

$$\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ -0.824 \end{bmatrix}$$

Significant errors are obtained.

Iterative Refinement, II

- Let $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$, and solve for $\mathbf{y}$ in $\mathbf{A}\mathbf{y} = \mathbf{r}$,

$$\begin{bmatrix} 0.001 & 2.42 \\ 1 & 1.58 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.824 \end{bmatrix}$$

Since the LU factors have been obtained, $\mathbf{y}$ can be found quickly.

$$\mathbf{y} = \begin{bmatrix} -0.8247 \\ 0.0003408 \end{bmatrix}$$

Let $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{y}$,

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1.175 \\ 2.148 \end{bmatrix}$$

Then,

$$\mathbf{A}\mathbf{x}^{(1)} = \begin{bmatrix} 5.199 \\ 4.569 \end{bmatrix}, \qquad \mathbf{r}^{(1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(1)} = \begin{bmatrix} 0.001 \\ 0.001 \end{bmatrix}.$$

We have a good approximation to the linear system solution.

Iterative Refinement, III

Algorithm 1.3.10. Iterative Refinement.

Given a linear system $\mathbf{Ax} = \mathbf{b}$, and small number ε.
Let $\mathbf{x}^{(0)} = \mathbf{0}$, $\mathbf{r}^{(0)} = \mathbf{b}$, $err = 1 + \varepsilon$, $k = 1$.
While $(err >= \varepsilon)$ {

Solve

$$\mathbf{Ay}^{(k)} = \mathbf{r}^{(k-1)}$$

by LU decomposition to get $\mathbf{y}^{(k)}$,

Let

$$\begin{aligned}
&\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{y}^{(k)}, \\
&\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{Ax}^{(k)}, \\
&err = |\mathbf{r}^{(k)}|, \\
&k = k + 1.
\end{aligned}$$

}

- The number of iterations is not known a priori and thus the CPU time needed cannot be predicted.
  - Iterative method.

Summary

- Error bounds
- Condition number
- Error examples
- Errors due to RHS and matrix
- Improving solution accuracy
  - Scaling
  - Iterative refinement

Vector Norms

Definition 1.3.11.

A vector norm on a vector space $\mathbb{X}$ is a real-valued function on $\mathbb{X}$, which satisfies the following three conditions:

1. $\|\mathbf{x}\| \geq 0, \quad \forall \mathbf{x} \in \mathbb{X}$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
2. $\|\alpha \mathbf{x}\| = |\alpha| \, \|\mathbf{x}\|$, $\forall \mathbf{x} \in \mathbb{X}$, $\forall \alpha \in \mathbb{C}$.
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{X}$.

Definition 1.3.12.

For the case $\mathbb{X} = \mathbb{C}^n$, the Euclidean norm of a vector is defined by

$$\|\mathbf{x}\|_2 = (\mathbf{x}^T \mathbf{x})^{1/2}. \tag{1.3.24}$$

Vector Norms, II

- The most commonly used vector norms in numerical linear algebra are special cases of the Hölder norms.

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}. \tag{1.3.25}$$

- The following norms are more important in practice,

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|, \tag{1.3.26}$$

$$\|\mathbf{x}\|_2 = \left(|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2 \right)^{1/2}, \tag{1.3.27}$$

$$\|\mathbf{x}\|_\infty = \max_{i=1}^{n} |x_i|. \tag{1.3.28}$$

- The Cauchy-Schwarz inequality can also be written as

$$|\mathbf{x}^T \mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2. \tag{1.3.29}$$

Matrix Norms

Definition 1.3.13.

Given a matrix $\mathbf{A} \in \mathbb{C}^{n\times m}$, the matrix norm is defined as

$$\|\mathbf{A}\|_{pq} = \max_{\mathbf{x}\in\mathbb{C}^m, \mathbf{x}\neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_q}. \tag{1.3.30}$$

- In this definition, the norm $\|\cdot\|_{pq}$ is induced by two vector norms, $\|\cdot\|_p$ and $\|\cdot\|_q$. And these norms satisfy the usual properties of norms,

$$\|\mathbf{A}\| \geq 0, \quad \forall \mathbf{A} \in \mathbb{C}^{n\times m}, \text{ and } \|\mathbf{A}\| = 0 \text{ if and only if } \mathbf{A} = \mathbf{0}. \tag{1.3.31}$$

$$\|\alpha\mathbf{A}\| = |\alpha|\,\|\mathbf{A}\|, \quad \forall \mathbf{A} \in \mathbb{C}^{n\times m}, \forall \alpha \in \mathbb{C}. \tag{1.3.32}$$

$$\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|, \quad \forall \mathbf{A},\mathbf{B} \in \mathbb{C}^{n\times m}. \tag{1.3.33}$$

- When $p = q$ the matrix norm is simplified as $\|\cdot\|_p$ and is called a matrix p-norm.
- The most important cases are still $p = 1, 2$ and ∞.
- Matrix norms defined above satisfy the following property.

$$\|\mathbf{A}\mathbf{B}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_p. \tag{1.3.34}$$

- Matrix norms satisfy the above equation is also called consistent.

Matrix Norms, II

- A consistent matrix p-norm also satisfies the following:

$$\|\mathbf{A}^k\|_p \leq \|\mathbf{A}\|_p^k. \tag{1.3.35}$$

- Thus, if any of the p-norms of $\mathbf{A}$ is less than 1, then $\mathbf{A}^k \to \mathbf{0}$ as $k \to \infty$.

Definition 1.3.14.

The Frobenius norm of an $n \times m$ matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_F = \left(\sum_{j=1}^{m}\sum_{i=1}^{n} |a_{i,j}|^2\right)^{1/2}. \tag{1.3.36}$$

- Frobenius norm is shown to be consistent.
- Unlike matrix p-norm, $\|\mathbf{I}\|_F \neq 1$; while $\|\mathbf{I}\|_p = 1$.

Matrix Norms, III

- It can be shown that the followings are consequences of the definition beforehand.

$$\|\mathbf{A}\|_1 = \max_{j=1}^{m} \sum_{i=1}^{n} |a_{i,j}|, \tag{1.3.37}$$

$$\|\mathbf{A}\|_\infty = \max_{i=1}^{n} \sum_{j=1}^{m} |a_{i,j}|, \tag{1.3.38}$$

$$\|\mathbf{A}\|_2 = \left(\rho(\mathbf{A}^H\mathbf{A})\right)^{1/2} = \left(\rho(\mathbf{A}\mathbf{A}^H)\right)^{1/2}, \tag{1.3.39}$$

$$\|\mathbf{A}\|_F = \left(\text{tr}(\mathbf{A}^H\mathbf{A})\right)^{1/2} = \left(\text{tr}(\mathbf{A}\mathbf{A}^H)\right)^{1/2}. \tag{1.3.40}$$

- The eigenvalues of $\mathbf{A}^H\mathbf{A}$ are nonnegative, and their square roots are called singular values of $\mathbf{A}$ and are denoted by σ_i, $i = 1, 2, \cdots, m$, ordered from large to small.
- Thus, $\|\mathbf{A}\|_2 = \sigma_1$, the largest singular value of $\mathbf{A}$.
- Note 1. The spectrum of $\mathbf{A}$, $\sigma(\mathbf{A})$, is the set of all eigenvalues of matrix $\mathbf{A}$
- Note 2. The spectral radius of $\mathbf{A}$, $\rho(\mathbf{A})$, is the largest absolute value of the eigenvalues of $\mathbf{A}$.
- Note 3. The trace of $\mathbf{A}$ is $\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.