MA 3021: Numerical Analysis I
Direct and Iterative Methods for Solving Linear Systems

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 32001, Taiwan

E-mail: syyang@math.ncu.edu.tw
Website: http://www.math.ncu.edu.tw/~syyang/

A system of linear equations

We are interested in solving systems of linear equations having the form:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n &= b_3, \\ \vdots & \vdots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n. \end{cases}$$

This is a system of n equations in the n unknowns, $x_1, x_2, \cdots, x_n$. The elements a_{ij} and b_i are assumed to be prescribed real numbers.

$Ax = b$

We can rewrite this system of linear equations in a matrix form:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}.$$

We can denote these matrices by A, x, and b, giving the simpler equation:

$$Ax = b.$$

Matrix

A matrix is a rectangular array of numbers such as

$$\begin{bmatrix} 3.0 & 1.1 & -0.12 \\ 6.2 & 0.0 & 0.15 \\ 0.6 & -4.0 & 1.3 \\ 9.3 & 2.1 & 8.2 \end{bmatrix}, \quad \begin{bmatrix} 3 & 6 & \frac{11}{7} & -17 \end{bmatrix}, \quad \begin{bmatrix} 3.2 \\ -4.7 \\ 0.11 \end{bmatrix}.$$

4×3 matrix

1×4 matrix
a row vector

3×1 matrix
a column vector

Matrix properties

- If A is a matrix, the notation a_{ij}, $(A)_{ij}$, or $A(i,j)$ is used to denote the element at the intersection of the ith row and the jth column. For example, let A be the first matrix on the previous slide. Then $a_{32} = (A)_{32} = A(3,2) = -4.0$.
- The transpose of a matrix is denoted by $A^\top$ and is the matrix defined by $(A^\top)_{ij} = a_{ji}$. The transpose of the matrix A is:

$$A^\top = \begin{bmatrix} 3.0 & 6.2 & 0.6 & 9.3 \\ 1.1 & 0.0 & -4.0 & 2.1 \\ -0.12 & 0.15 & 1.3 & 8.2 \end{bmatrix}.$$

- If a matrix A has the property $A = A^\top$, we say that A is symmetric.
- The $n \times n$ matrix

$$I := I_n := I_{n\times n} := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

is called an identity matrix. Note that $IA = A = AI$ for any $n \times n$ matrix A.

Algebraic operations

- *Scalar * Matrix:* If A is a matrix and λ is a scalar, then λA is defined by $(\lambda A)_{ij} = \lambda a_{ij}$.
- *Matrix + Matrix:* If $A = (a_{ij})$ and $B = (b_{ij})$ are $m \times n$ matrices, then $A + B$ is defined by $(A + B)_{ij} = a_{ij} + b_{ij}$.
- *Matrix * Matrix:* If A is an $m \times p$ matrix and B is a $p \times n$ matrix, then AB is an $m \times n$ matrix defined by:

$$(AB)_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}, \qquad 1 \le i \le m,\ 1 \le j \le n.$$

What is the cost of AB?

Answer: mnp multiplications and $mn(p-1)$ additions.

Right inverse and left inverse

If A and B are two matrices such that $AB = I$, then we say that B is a right inverse of A and that A is a left inverse of B. For example:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \alpha & \beta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_{2\times 2}, \qquad \forall \alpha, \beta \in \mathbb{R}.$$

$$\begin{bmatrix} 1 & 0 & \alpha \\ 0 & 1 & \beta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_{2\times 2}, \qquad \forall \alpha, \beta \in \mathbb{R}.$$

Notice that right inverse and left inverse may not unique.

- **Theorem on right inverse:** A square matrix can possess at most one right inverse.

  Proof: Let $AB = I$. Then $\sum_{j=1}^{n} b_{jk} A^{(j)} = I^{(k)}$, $1 \le k \le n$. So, the columns of A form a basis for $\mathbb{R}^n$. Therefore, the coefficients b_{jk} above are uniquely determined.

- **Theorem on matrix inverse:** If A and B are square matrices such that $AB = I$, then $BA = I$.

  Proof: Let $C = BA - I + B$. Then $AC = ABA - AI + AB = A - A + I = I$. Since right inverse for square matrix is at most one, $B = C$. Hence,
  $C = BA - I + B = BA - I + C$, i.e., $BA = I$.

Inverse

- If a square matrix A has a right inverse B, then B is unique and $BA = AB = I$. We then call B the inverse of A and say that A is invertible or nonsingular. We denote $B = A^{-1}$.
- Example:

$$\begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_{2\times 2}.$$

- If A is invertible, then the system of equations $Ax = b$ has the solution $x = A^{-1}b$. If A^{-1} is not available, then in general, A^{-1} should not be computed solely for the purpose of obtaining x.
- How do we get this A^{-1}?

Equivalent systems

- Let two linear systems be given, each consisting of n equations with n unknowns:

$$Ax = b \quad and \quad Bx = d.$$

  If the two systems have precisely the same solutions, we call them equivalent systems.

- Note that A and B can be very different.
- Thus, to solve a linear system of equations, we can instead solve any equivalent system. This simple idea is at the heart of our numerical procedures.

Elementary operations

- Let $\mathcal{E}_i$ denote the i-th equation in the system $Ax = b$. The following are the elementary operations which can be performed:
  - Interchanging two equations in the system: $\mathcal{E}_i \leftrightarrow \mathcal{E}_j$;
  - Multiplying an equation by a nonzero number: $\lambda\mathcal{E}_i \to \mathcal{E}_i$;
  - Adding to an equation a multiple of some other equation: $\mathcal{E}_i + \lambda\mathcal{E}_j \to \mathcal{E}_i$.
- **Theorem on equivalent systems:** If one system of equations is obtained from another by a finite sequence of elementary operations, then the two systems are equivalent.

Elementary operations (continued)

- An elementary matrix is defined to be an $n \times n$ matrix that arises when an elementary operation is applied to the $n \times n$ identity matrix.
- Let A_i be the i-th row of matrix A. The elementary operations expressed in terms of the rows of matrix A are:
  - The interchange of two rows in A: $A_i \leftrightarrow A_j$;
  - Multiplying one row by a nonzero constant: $\lambda A_i \to A_i$;
  - Adding to one row a multiple of another: $A_i + \lambda A_j \to A_i$.
- Each elementary row operation on A can be accomplished by multiplying A on the left by an elementary matrix.

Examples

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \lambda a_{21} & \lambda a_{22} & \lambda a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \lambda a_{21} + a_{31} & \lambda a_{22} + a_{32} & \lambda a_{23} + a_{33} \end{bmatrix}.$$

Invertible matrix

- If matrix A is invertible, then there exists a sequence of elementary row operations can be applied to A, reducing it to I,

$$E_m E_{m-1} \cdots E_2 E_1 A = I.$$

- This gives us an equation for computing the inverse of a matrix:

$$A^{-1} = E_m E_{m-1} \cdots E_2 E_1 = E_m E_{m-1} \cdots E_2 E_1 I.$$

Remark: This is not a practical method to compute A^{-1}.

Eigenvalue and eigenvector

Let $A \in \mathbb{C}^{n \times n}$ be a square matrix. If there exists a nonzero vector $x \in \mathbb{C}^n$ and a scalar $\lambda \in \mathbb{C}$ such that

$$Ax = \lambda x,$$

then λ is called an eigenvalue of A and x is called the corresponding eigenvector of A.

Remark: Computing λ and x is a major task in numerical linear algebra.

Theorem on nonsingular matrix properties

For an $n \times n$ real matrix A, the following properties are equivalent:

- The inverse of A exists; that is, A is nonsingular.
- The determinant of A is nonzero.
- The rows of A form a basis for $\mathbb{R}^n$.
- The columns of A form a basis for $\mathbb{R}^n$.
- As a map from $\mathbb{R}^n$ to $\mathbb{R}^n$, A is injective (one to one).
- As a map from $\mathbb{R}^n$ to $\mathbb{R}^n$, A is surjective (onto).
- The equation $Ax = 0$ implies $x = 0$.
- For each $b \in \mathbb{R}^n$, there is exactly one $x \in \mathbb{R}^n$ such that $Ax = b$.
- A is a product of elementary matrices.
- 0 is not an eigenvalue of A.

Some easy-to-solve systems:

1. Diagonal Structure

$$\begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}.$$

The solution is: (provided $a_{ii} \neq 0$ for all $i = 1, 2, \cdots, n$)

$$x = \left(\frac{b_1}{a_{11}}, \frac{b_2}{a_{22}}, \frac{b_3}{a_{33}}, \cdots, \frac{b_n}{a_{nn}} \right)^\top.$$

- If $a_{ii} = 0$ for some index i, and if $b_i = 0$ also, then x_i can be any real number. The number of solutions is infinity.
- If $a_{ii} = 0$ and $b_i \neq 0$, no solution of the system exists.
- What is the complexity of the method? n divisions.

2. Lower Triangular Systems

$$\begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}.$$

Some simple observations:

- If $a_{11} \neq 0$, then we have $x_1 = b_1/a_{11}$.
- Once we have x_1, we can simplify the second equation, $x_2 = (b_2 - a_{21}x_1)/a_{22}$, provided that $a_{22} \neq 0$.
- Similarly, $x_3 = (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33}$, provided that $a_{33} \neq 0$.
- In general, to find the solution to this system, we use forward substitution (assume that $a_{ii} \neq 0$ for all i):

input n, (a_{ij}), $b = (b_1, b_2, \cdots, b_n)^\top$
for $i = 1$ **to** n **do**

$$x_i \leftarrow \Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j\Big)/a_{ii}$$

end do
output $x = (x_1, x_2, \cdots, x_n)^\top$

2. Lower Triangular Systems (continued)

- Complexity of forward substitution:
  - n divisions.
  - the number of multiplications: 0 for x_1, 1 for x_2, 2 for x_3, $\cdots$
    total $= 0 + 1 + 2 + \cdots + (n-1) \approx 1 + 2 + \cdots + n = (n+1)n/2$,
    $\therefore$ total $= O(n^2)$.
  - the number of subtractions: same as the number of multiplications $= O(n^2)$.
- Forward substitution is an $O(n^2)$ algorithm.
- Remark: forward substitution is a sequential algorithm (not parallel at all).

3. Upper Triangular Systems

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}.$$

The formal algorithm to solve for x is called backward substitution. It is also an $O(n^2)$ algorithm. Assume that $a_{ii} \neq 0$ for all i:

input n, (a_{ij}), $b = (b_1, b_2, \cdots, b_n)^\top$
for $i = n : -1 : 1$ **do**

$$x_i \leftarrow \left(b_i - \sum_{j=i+1}^{n} a_{ij} x_j\right) / a_{ii}$$

end do
output $x = (x_1, x_2, \cdots, x_n)^\top$

LU decomposition (factorization)

- Suppose that A can be factored into the product of a lower triangular matrix L and an upper triangular matrix U:

$$A = LU.$$

- Then, $Ax = LUx = L(Ux)$. Thus, to solve the system of equations $Ax = b$, it is enough to solve this problem in two stages:

$$\begin{aligned} Lz &= b \quad solve\ for\ z, \\ Ux &= z \quad solve\ for\ x. \end{aligned}$$

Basic Gaussian elimination

Let $A^{(1)} = (a_{ij}^{(1)}) = A = (a_{ij})$ and $b^{(1)} = b$. Consider the following linear system $Ax = b$:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix}.$$

pivot row = row1.
pivot element: $a_{11}^{(1)} = 6$.
row2 - (12/6)*row1 $\to$ row2.
row3 - (3/6)*row1 $\to$ row3.
row4 - (-6/6)*row1 $\to$ row4.

$$\Longrightarrow \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix}.$$

multipliers: $12/6, 3/6, -6/6$.

Basic Gaussian elimination (continued)

We have the following equivalent system $A^{(2)}x = b^{(2)}$:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix}.$$

pivot row = row2.
pivot element $a_{22}^{(2)} = -4$.
row3 - (-12/-4)*row2 $\rightarrow$ row3.
row4 - (2/-4)*row2 $\rightarrow$ row4.

$$\Longrightarrow \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix}.$$

multiplier: $-12/-4, 2/-4$.

Basic Gaussian elimination (continued)

We have the following equivalent system $A^{(3)}x = b^{(3)}$:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix}.$$

pivot row = row3.
pivot element $a_{33}^{(3)} = 2$.
row4 - (4/2)*row3 $\rightarrow$ row4.

$$\Longrightarrow \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -3 \end{bmatrix}.$$

multiplier: 4/2.

Basic Gaussian elimination (continued)

Finally, we have the following equivalent upper triangular system $A^{(4)}x = b^{(4)}$:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -3 \end{bmatrix}.$$

Using the backward substitution, we have

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \\ -2 \\ 1 \end{bmatrix}.$$

The LU decomposition

Display the multipliers in an unit lower triangular matrix $L = (\ell_{ij})$:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{bmatrix}.$$

Let $U = (u_{ij})$ be the final upper triangular matrix $A^{(4)}$. Then we have

$$U = \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix}$$

and one can check that $A = LU$ (the Doolittle Decomposition).

Some remarks

- The entire elimination process will break down if any of the pivot elements are 0.
- The total number of arithmetic operations:

$$M/D = \frac{n^3}{3} + n^2 - \frac{n}{3};$$

$$A/S = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6}.$$

$\therefore$ The GE is an $O(n^3)$ algorithm.

Vector norm

A vector norm on $\mathbb{R}^n$ is a real-valued function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ with the properties:

- $\|x\| \geq 0, \forall\ x \in V$, and $\|x\| = 0$ if and only if $x = 0$;
- $\|\alpha x\| = |\alpha|\|x\|, \forall\ x \in V$ and $\alpha \in \mathbb{R}$;
- $\|x + y\| \leq \|x\| + \|y\|, \forall\ x, y \in V$ (triangle inequality).

Note: $\|x\|$ is called the norm of x, the length or magnitude of x.

Some vector norms on $\mathbb{R}^n$ and distance

- Let $x = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n$:
  - The 2-norm (Euclidean norm, or ℓ^2 norm): $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.
  - The infinity norm (ℓ^∞-norm): $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$.
  - The 1-norm (ℓ^1-norm): $\|x\|_1 = \sum_{i=1}^{n} |x_i|$.
- Let $x = (x_1, x_2, \cdots, x_n)^\top$, $y = (y_1, y_2, \cdots, y_n)^\top \in \mathbb{R}^n$. Then
  - $\|x - y\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$.
  - $\|x - y\|_\infty = \max_{1 \le i \le n} |x_i - y_i|$.
  - $\|x - y\|_1 = \sum_{i=1}^{n} |x_i - y_i|$.

The difference between the above norms

- What is the unit ball $\{x \in \mathbb{R}^2 : \ \|x\| \leq 1\}$ for the three norms above?
  - 2-norm: a circle;
  - ∞-norm: a square;
  - 1-norm: a diamond.
- **Example:** Let $x = (-1, 1, -2)^\top \in \mathbb{R}^3$. Then
  $\|x\|_2 = \sqrt{(-1)^2 + 1^2 + (-2)^2} = \sqrt{6}$,
  $\|x\|_\infty = \max\limits_{1 \leq i \leq 3} |x_i| = \max\{|-1|, |1|, |-2|\} = 2$,
  $\|x\|_1 = \sum\limits_{i=1}^{3} |x_i| = |-1| + |1| + |-2| = 6.$
- **Cauchy-Buniakowsky-Schwarz inequality:** For $x = (x_1, x_2, \cdots, x_n)^\top$, $y = (y_1, y_2, \cdots, y_n)^\top \in \mathbb{R}^n$, we have

$$\sum_{i=1}^{n} |x_i y_i| \leq \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \left(\sum_{i=1}^{n} y_i^2 \right)^{1/2} = \|x\|_2 \|y\|_2.$$

Convergence of sequences in $\mathbb{R}^n$

- **Definition:** Let $x, x^{(k)} \in \mathbb{R}^n$ for $k = 1, 2, \cdots$. Then $\lim_{k\to\infty} x^{(k)} = x$ with respect to the norm $\|\cdot\| \iff$ $\forall\ \varepsilon > 0$, $\exists$ an integer $N(\varepsilon) > 0$ such that if $k \geq N(\varepsilon)$ then $\|x^{(k)} - x\| < \varepsilon$.
- $\lim_{k\to\infty} x^{(k)} = x$ with respect to $\|\cdot\|_\infty \iff \lim_{k\to\infty} x_i^{(k)} = x_i$ for $i = 1, 2, \cdots, n$.
- **Example:** $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^\top = (1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k}\sin(k))^\top$.

  $\because \lim_{k\to\infty} 1 = 1,\ \lim_{k\to\infty} (2 + \frac{1}{k}) = 2,\ \lim_{k\to\infty} \frac{3}{k^2} = 0$, and $\lim_{k\to\infty} e^{-k}\sin(k) = 0$.

  $\therefore \lim_{k\to\infty} x^{(k)} = x = (1, 2, 0, 0)^\top$ with respect to $\|\cdot\|_\infty$ norm.

All vector norms on $\mathbb{R}^n$ are equivalent

- For each $x \in \mathbb{R}^n$, $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty$.

  Proof: Let $|x_j| = \|x\|_\infty$. Then

$$\|x\|_\infty^2 = |x_j|^2 = x_j^2 \le \sum_{i=1}^{n} x_i^2 = \|x\|_2^2 \le \sum_{i=1}^{n} x_j^2 = n x_j^2 = n\|x\|_\infty^2.$$

- In fact, all vector norms on $\mathbb{R}^n$ are equivalent!

Matrix norm

Let A be an $n \times n$ real matrix. If $\|\cdot\|$ is any vector norm on $\mathbb{R}^n$, then

$$\|A\| := \max\{\|Ax\| : x \in \mathbb{R}^n, \|x\| = 1\} \left(\Longleftrightarrow \|A\| := \max\{\frac{\|Ax\|}{\|x\|} : x \in \mathbb{R}^n, x \neq 0\} \right)$$

defines a norm on the vector space of all $n \times n$ real matrices.

(This is called the matrix norm associated with the given vector norm)

Proof:

- $\because \|Ax\| \geq 0 \ \forall\ x \in \mathbb{R}^n, \|x\| = 1.\ \therefore \|A\| \geq 0.$
  Exercise: $\|A\| = 0$ if and only if $A = 0$.
- $\|\lambda A\| = \max\{\|\lambda Ax\| : \|x\| = 1\} = \max\{|\lambda|\|Ax\| : \|x\| = 1\}$
  $= |\lambda| \max\{\|Ax\| : \|x\| = 1\} = |\lambda|\|A\|.$
- $\|A + B\| = \max\{\|(A + B)x\| : \|x\| = 1\} \leq \max\{\|Ax\| + \|Bx\| : \|x\| = 1\}$
  $\leq \max\{\|Ax\| : \|x\| = 1\} + \max\{\|Bx\| : \|x\| = 1\} = \|A\| + \|B\|.$

Some additional properties

- $\|Ax\| \leq \|A\|\|x\|, \forall\, x \in \mathbb{R}^n$.

  Proof:

  Let $x \neq 0$. Then $v = \dfrac{x}{\|x\|}$ is of norm 1. $\therefore\ \|A\| \geq \|Av\| = \dfrac{\|Ax\|}{\|x\|}$.

- $\|I\| = 1$.
- $\|AB\| \leq \|A\|\|B\|$.

  Proof:

  $$\begin{aligned}
  &\|AB\| := \max\{\|(AB)x\| : x \in \mathbb{R}^n, \|x\| = 1\} \\
  &\leq \max\{\|A\|\|Bx\| : x \in \mathbb{R}^n, \|x\| = 1\} \\
  &\leq \max\{\|A\|\|B\|\|x\| : x \in \mathbb{R}^n, \|x\| = 1\} = \|A\|\|B\|.
  \end{aligned}$$

Some matrix norms

Let $A_{n\times n} = (a_{ij})$ be an $n \times n$ real matrix. Then

- The ∞-matrix norm:

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

- The 1-matrix norm:

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|.$$

- The 2-matrix norm (ℓ^2-matrix norm):

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

Example

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix}.$$

- The characteristic polynomial $p(\lambda)$ of A is given by

$$\begin{aligned} p(\lambda) &= \det(A - \lambda I) \\ &= (1-\lambda)\{(1-\lambda)^2 + 1\} + (-1)\{-2(1-\lambda)\} \\ &= (1-\lambda)\{\lambda^2 - 2\lambda + 4\}. \end{aligned}$$

  The eigenvalues of A are $\lambda_1 = 1$, $\lambda_2 = 1 + \sqrt{3}i$ and $\lambda_3 = 1 - \sqrt{3}i$.

- The spectral radius $\rho(A)$ of matrix A is defined by

$$\rho(A) = \max\{|\lambda| : \ \lambda \textit{ is an eigenvalue of } A\}.$$

  For the above matrix A, we have $\rho(A) = \max\{|1|, |1+\sqrt{3}i|, |1-\sqrt{3}i|\} = 2$.

The 2-matrix norm

- $\|A\|_2$ is not easy to compute.
- Since $A^\top A$ is symmetric, $A^\top A$ has n real eigenvalues, $\lambda_1, \lambda_2, \cdots, \lambda_n \in \mathbb{R}$. Moreover, one can prove that they are all nonnegative. Then

$$\rho(A^\top A) := \max_{1 \le i \le n} \{\lambda_i\} \ge 0.$$

  is called the spectral radius of $A^\top A$.
- Then the ℓ^2-matrix norm of A is given by

$$\|A\|_2 = \sqrt{\rho(A^\top A)}.$$

  The ℓ^2-matrix norm is also called the spectral norm.

Properties of matrix norm

Let A be an $n \times n$ real matrix. Then

- Then the ℓ^2-matrix norm of A is given by $\|A\|_2 = \sqrt{\rho(A^\top A)}$. The ℓ^2-matrix norm is also called the spectral norm.
- $\rho(A) \le \|A\|$ for any matrix norm $\|\cdot\|$.

  Proof: Suppose that λ is an eigenvalue of A with eigenvector x and $\|x\| = 1$.

  $\Longrightarrow |\lambda| = |\lambda|\|x\| = \|\lambda x\| = \|Ax\| \le \|A\|\|x\| = \|A\|$.

  $\Longrightarrow \rho(A) = \max|\lambda| \le \|A\|$.
- For any $n \times n$ matrix A and any $\varepsilon > 0$, $\exists$ a matrix norm $\|\cdot\|$ such that $\rho(A) \le \|A\| \le \rho(A) + \varepsilon$.

Example

$$
\begin{aligned}
A &= \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix}. \\
A^\top A &= \begin{bmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{bmatrix} \\
\det(A^\top A - \lambda I) &= \det \begin{bmatrix} 3-\lambda & 2 & -1 \\ 2 & 6-\lambda & 4 \\ -1 & 4 & 5-\lambda \end{bmatrix} = -\lambda(\lambda^2 - 14\lambda + 42).
\end{aligned}
$$

$\Longrightarrow \lambda = 0, 7+\sqrt{7}, 7-\sqrt{7}.$

$\Longrightarrow \|A\|_2 = \sqrt{\rho(A^\top A)} = \sqrt{7+\sqrt{7}} \approx 3.106.$

Convergence

- **Definition:** An $n \times n$ matrix A is said to be convergent if $\lim_{k\to\infty} (A^k)_{ij} = 0$ for $i, j = 1, 2, \cdots, n$.
- **Example:**

$$A = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \Longrightarrow A^2 = \begin{bmatrix} \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} \Longrightarrow A^3 = \begin{bmatrix} \frac{1}{8} & 0 \\ \frac{3}{16} & \frac{1}{8} \end{bmatrix} \Longrightarrow \cdots$$

$$A^k = \begin{bmatrix} (\frac{1}{2})^k & 0 \\ \frac{k}{2^{k+1}} & (\frac{1}{2})^k \end{bmatrix}, \quad \lim_{k\to\infty} (\frac{1}{2})^k = 0, \quad \lim_{k\to\infty} \frac{k}{2^{k+1}} = 0.$$

$\therefore$ A is a convergent matrix.

Equivalent statements

The following statements are equivalent:

- A is a convergent matrix.
- $\lim_{n\to\infty} \|A^n\| = 0$ for some natural matrix norm.
- $\lim_{n\to\infty} \|A^n\| = 0$ for all natural matrix norms.
- $\rho(A) < 1$.
- $\lim_{n\to\infty} A^n x = 0$ for all x.

Iterative methods

- Basic idea: $Ax = b \Longrightarrow x = Tx + c$ for some fixed matrix T and vector c.
- Given $x^{(0)}$, $x^{(k)} := Tx^{(k-1)} + c$ for $k = 1, 2, \cdots$
- Consider a linear system:

$$\left\{\begin{array}{rcr} 10x_1 - x_2 + 2x_3 + 0 & = & 6, \\ -x_1 + 11x_2 - x_3 + 3x_4 & = & 25, \\ 2x_1 - x_2 + 10x_3 - x_4 & = & -11, \\ 0 + 3x_2 - x_3 + 8x_4 & = & 15. \end{array}\right.$$

Exact unique solution: $x = (1, 2, -1, 1)^\top$.

The Jacobi iterative method

$$
\begin{aligned}
x_1 &= 0 + \frac{1}{10}x_2 - \frac{2}{10}x_3 + 0 + \frac{6}{10}, \\
x_2 &= \frac{1}{11}x_1 + 0 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11}, \\
x_3 &= -\frac{2}{10}x_1 + \frac{1}{10}x_2 + 0 + \frac{1}{10}x_4 - \frac{11}{10}, \\
x_4 &= 0 - \frac{3}{8}x_2 + \frac{1}{8}x_3 + 0 + \frac{15}{8}.
\end{aligned}
$$

$$
x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = Tx + c = \begin{bmatrix} 0 & \frac{1}{10} & -\frac{2}{10} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{2}{10} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{bmatrix} x + \begin{bmatrix} \frac{6}{10} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{bmatrix}.
$$

The Jacobi iterative method (continued)

If $x^{(0)} = (0,0,0,0)^\top$, then

$$x^{(1)} = Tx^{(0)} + c = \begin{bmatrix} \frac{6}{10} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{bmatrix} = \begin{bmatrix} 0.6000 \\ 2.2727 \\ -1.1000 \\ 1.8750 \end{bmatrix}.$$

$$\Longrightarrow x^{(2)} = Tx^{(1)} + c \Longrightarrow \cdots$$

$$\Longrightarrow \frac{\|x^{(10)} - x^{(9)}\|_\infty}{\|x^{(10)}\|_\infty} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3} \quad \textit{stop!} \quad x \approx x^{(10)}.$$

The Jacobi iterative method (continued)

$Ax = b$, $a_{ii} \neq 0$ for all $i = 1, 2, \cdots, n$.

Given $x^{(k-1)}$, $k \geq 1$.

For $i = 1, 2, \cdots, n$,

$$x_i^{(k)} = \frac{-\displaystyle\sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k-1)} + b_i}{a_{ii}}.$$

Theoretical setting

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & & & \\ -a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n1} & \cdots & -a_{nn-1} & 0 \end{bmatrix}$$

$$- \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ & \ddots & \ddots & \vdots \\ & & & -a_{n-1n} \\ & & & 0 \end{bmatrix}$$

$\Longrightarrow A = D - L - U$

D: diagonal matrix; L: lower triangular matrix; U: upper triangular matrix.

Theoretical setting (continued)

$Ax = b$

$\Longrightarrow\ Dx = (L+U)x + b$

$\Longrightarrow\ x = D^{-1}(L+U)x + D^{-1}b$

The Jacobi iterative method: $x^{(k)} = D^{-1}(L+U)x^{(k-1)} + D^{-1}b$, $k = 1, 2, \cdots$

Notation: $x^{(k)} = T_J x^{(k-1)} + c_J$, where $T_J := D^{-1}(L+U)$ and $c_J := D^{-1}b$.

The Gauss-Seidel iterative method

$Ax = b$, $a_{ii} \neq 0$ for all $i = 1, 2, \cdots, n$.

Given $x^{(k-1)}$, $k \geq 1$.

For $i = 1, 2, \cdots, n$,

$$x_i^{(k)} = \frac{-\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + b_i}{a_{ii}}.$$

Example

Letting $x^{(0)} = (0,0,0,0)^\top$, for $k = 1, 2, \cdots$

$$\begin{aligned}
x_1^{(k)} &= 0 + \frac{1}{10}x_2^{(k-1)} - \frac{2}{10}x_3^{(k-1)} + 0 + \frac{6}{10}, \\
x_2^{(k)} &= \frac{1}{11}x_1^{(k)} + 0 + \frac{1}{11}x_3^{(k-1)} - \frac{3}{11}x_4^{(k-1)} + \frac{25}{11}, \\
x_3^{(k)} &= -\frac{2}{10}x_1^{(k)} + \frac{1}{10}x_2^{(k)} + 0 + \frac{1}{10}x_4^{(k-1)} - \frac{11}{10}, \\
x_4^{(k)} &= 0 - \frac{3}{8}x_2^{(k)} + \frac{1}{8}x_3^{(k)} + 0 + \frac{15}{8}.
\end{aligned}$$

$$\Longrightarrow \frac{\|x^{(5)} - x^{(4)}\|_\infty}{\|x^{(5)}\|_\infty} = 4.0 \times 10^{-4} < 10^{-3} \quad \textit{stop!} \quad x \approx x^{(5)}.$$

Theoretical setting

$Ax = b$, $A = D - L - U$.

$\Longrightarrow \ (D-L)x^{(k)} = Ux^{(k-1)} + b$

That is,

$$
\begin{array}{rcl}
a_{11}x_1^{(k)} & = & -a_{12}x_2^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1, \\
a_{21}x_1^{(k)} + a_{22}x_2^{(k)} & = & -a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2, \\
\vdots & = & \vdots \\
a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \cdots + a_{nn}x_n^{(k)} & = & b_n.
\end{array}
$$

$\Longrightarrow \ x^{(k)} = (D-L)^{-1}Ux^{(k-1)} + (D-L)^{-1}b$ for $k = 1, 2, \cdots$

The Gauss-Seidel iterative method: $x^{(k)} = T_S x^{(k-1)} + c_S$,

where $T_S := (D-L)^{-1}U$ and $c_S := (D-L)^{-1}b$.

Note: $a_{ii} \neq 0$, $i = 1, 2, \cdots, n \Longleftrightarrow D - L$ is nonsingular!

Theorem on convergence

- If $\rho(T) < 1$, then $(I - T)^{-1}$ exists and

$$(I - T)^{-1} = I + T + T^2 + \cdots (:= \sum_{n=0}^{\infty} T^n).$$

- For any $x^{(0)} \in \mathbb{R}^n$, the sequence $\{x^{(k)}\}$ defined by $x^{(k)} := Tx^{(k-1)} + c$, $k \geq 1$, converges to the unique solution of $x = Tx + c \iff \rho(T) < 1$.

Corollaries

- $x^{(0)} \in \mathbb{R}^n$, $x^{(k)} := Tx^{(k-1)} + c$, $k \geq 1$. If $\|T\| < 1$ for any natural matrix norm then $\{x^{(k)}\}$ converges to the unique solution of $x = Tx + c$ and
  - $\|x - x^{(k)}\| \leq \|T\|^k \|x - x^{(0)}\|$.
  - $\|x - x^{(k)}\| \leq \dfrac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\|$.
- If A is strictly diagonally dominant, then for any $x^{(0)} \in \mathbb{R}^n$, both the Jacobi and Gauss-Seidel methods give sequences $\{x^{(k)}\}$ that converge to the unique solution of $Ax = b$ $(x = Tx + c)$.

Successive Over-Relaxation (SOR)

- **The Gauss-Seidel method:**

$$x_i^{(k)} = \frac{1}{a_{ii}} \left\{ -\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + b_i \right\}.$$

- **Successive over-relaxation:**

$$x_i^{(k)} = (1-\omega) x_i^{(k-1)} + \frac{\omega}{a_{ii}} \left\{ -\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + b_i \right\},$$

where $\omega > 0$. In general,

- $\omega = 1$: the Gauss-Seidel method;
- $0 < \omega < 1$: when G-S diverges;
- $\omega > 1$: when G-S converges!

SOR (continued)

$$a_{ii}x_i^{(k)} + \omega \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} = (1-\omega)a_{ii}x_i^{(k-1)} - \omega \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} + \omega b_i.$$

$$\Longrightarrow \quad (D - \omega L)x^{(k)} = \Big((1-\omega)D + \omega U\Big)x^{(k-1)} + \omega b.$$

$$\Longrightarrow \quad x^{(k)} = (D - \omega L)^{-1}\Big((1-\omega)D + \omega U\Big)x^{(k-1)} + \omega (D - \omega L)^{-1} b.$$

$$\Longrightarrow \quad x^{(k)} = T_\omega x^{(k-1)} + c_\omega.$$

Example

- Consider a linear system:

$$\begin{cases} 4x_1 + 3x_2 + 0 &= 24, \\ 3x_1 + 4x_2 - x_3 &= 30, \\ 0 - x_2 + 4x_3 &= -24. \end{cases}$$

  Exact unique solution: $x = (3, 4, -5)^\top$.

- Let $x^{(0)} = (1, 1, 1)^\top$. The G-S method:

$$\begin{cases} x_1^{(k)} &= -0.75x_2^{(k-1)} + 6, \\ x_2^{(k)} &= -0.75x_1^{(k)} + 0.25x_3^{(k-1)} + 7.5, \\ x_3^{(k)} &= 0.25x_2^{(k)} - 6. \end{cases}$$

- Let $x^{(0)} = (1, 1, 1)^\top$. The SOR with $\omega = 1.25$:

$$\begin{cases} x_1^{(k)} &= -0.25x_1^{(k-1)} - 0.9375x_2^{(k-1)} + 7.5, \\ x_2^{(k)} &= -0.9375x_1^{(k)} - 0.25x_2^{(k-1)} + 0.3125x_3^{(k-1)} + 9.375, \\ x_3^{(k)} &= 0.3125x_2^{(k)} - 0.25x_3^{(k-1)} - 7.5. \end{cases}$$

Theorems on convergence

- If $a_{ii} \neq 0$, $i = 1, 2, \cdots, n$, then $\rho(T_\omega) \geq |\omega - 1|$. This implies the SOR method can converge only if $0 < \omega < 2$.
- If A is SPD, $0 < \omega < 2$, then the SOR method converges for any $x^{(0)}$.

Some error analysis

- Suppose that we want to solve the linear system $Ax = b$, but b is somehow perturbed to $\widetilde{b}$ (this may happen when we convert a real b to a floating-point b).
- Then actual solution would satisfy a slightly different linear system

$$A\widetilde{x} = \widetilde{b}.$$

- Question: Is $\widetilde{x}$ very different from the desired solution x of the original system?
- Of course, the answer should depend on how good the matrix A is.
- Let $\|\cdot\|$ be a vector norm, we consider two types of errors:
  - absolute error: $\|x - \widetilde{x}\|$?
  - relative error: $\|x - \widetilde{x}\|/\|x\|$?

The absolute error

For the absolute error, we have

$$\|x - \widetilde{x}\| = \|A^{-1}b - A^{-1}\widetilde{b}\| = \|A^{-1}(b - \widetilde{b})\| \leq \|A^{-1}\|\|b - \widetilde{b}\|.$$

Therefore, the absolute error of x depends on two factors: the absolute error of b and the matrix norm of A^{-1}.

The relative error

For the relative error, we have

$$\begin{aligned}\|x-\widetilde{x}\| &= \|A^{-1}b - A^{-1}\widetilde{b}\| = \|A^{-1}(b-\widetilde{b})\| \\ &\le \|A^{-1}\|\|b-\widetilde{b}\| = \|A^{-1}\|\|Ax\|\frac{\|b-\widetilde{b}\|}{\|b\|} \\ &\le \|A^{-1}\|\|A\|\|x\|\frac{\|b-\widetilde{b}\|}{\|b\|}.\end{aligned}$$

That is

$$\frac{\|x-\widetilde{x}\|}{\|x\|} \le \|A^{-1}\|\|A\| \ \frac{\|b-\widetilde{b}\|}{\|b\|}.$$

Therefore, the relative error of x depends on two factors: the relative error of b and $\|A\|\|A^{-1}\|$.

Condition number

- Therefore, we define a condition number of the matrix A as

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

  $\kappa(A)$ measures how good the matrix A is.
- Example: Let $\varepsilon > 0$ and

$$A = \begin{bmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{bmatrix} \Longrightarrow A^{-1} = \varepsilon^{-2} \begin{bmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{bmatrix}.$$

  Then $\|A\|_\infty = 2+\varepsilon$, $\|A^{-1}\|_\infty = \varepsilon^{-2}(2+\varepsilon)$, and $\kappa(A) = \left(\frac{2+\varepsilon}{\varepsilon}\right)^2 \geq \frac{4}{\varepsilon^2}$.

Condition number (continued)

- For example, if $\varepsilon = 0.01$, then $\kappa(A) \geq 40000$.
- What does this mean?
  It means that the relative error in x can be 40000 times greater than the relative error in b.
- If $\kappa(A)$ is large, we say that A is ill-conditioned, otherwise A is well-conditioned.
- In the ill-conditioned case, the solution is very sensitive to the small changes in the right-hand vector b (higher precision in b may be needed).

Another way to measure the error

Consider the linear system $Ax = b$. Let $\widetilde{x}$ be a computed solution (an approximation to x).

- Residual vector:
$$r = b - A\widetilde{x}.$$
- Error vector:
$$e = x - \widetilde{x}.$$
- They satisfy
$$Ae = r.$$
(Proof: $Ae = Ax - A\widetilde{x} = b - A\widetilde{x} = r$)
- Moreover, we have
$$\frac{1}{\kappa(A)} \frac{\|r\|}{\|b\|} \le \frac{\|e\|}{\|x\|} \le \kappa(A) \frac{\|r\|}{\|b\|}.$$
(Theorem on bounds involving condition number)

Proof of the Theorem

$\because Ae = r.$

$\therefore e = A^{-1}r.$

$\therefore \|e\|\|b\| = \|A^{-1}r\|\|Ax\| \leq \|A^{-1}\|\|r\|\|A\|\|x\|.$

$$\therefore \frac{\|e\|}{\|x\|} \leq \kappa(A)\,\frac{\|r\|}{\|b\|}.$$

On the other hand, we have $\|r\|\|x\| = \|Ae\|\|A^{-1}b\| \leq \|A\|\|e\|\|A^{-1}\|\|b\|.$

$$\therefore \frac{1}{\kappa(A)}\,\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|}.$$