2021 Spring MAS 365
Chapter 7: Iterative Techniques in Matrix Algebra

Donghwan Kim

KAIST

Mar/23,25,30, Apr/1, 2021

Measuring Distances

- The objective of iterative techniques (in Chapter 2) is to find a way to minimize the difference between the approximations and the exact solution.
- We need to determine a way to measure the distance between n-dimensional column vectors.

Vector Norms and Distances

Definition 1

A **vector norm** *on* $\mathbb{R}^n$ *is a function,* $||\cdot||$*, from* $\mathbb{R}^n$ *to* $\mathbb{R}$ *with the following properties:*

1. $||\boldsymbol{x}|| \geq 0$ *for all* $\boldsymbol{x} \in \mathbb{R}^n$*,*
2. $||\boldsymbol{x}|| = 0$ *if and only if* $\boldsymbol{x} = \mathbf{0}$*,*
3. $||\alpha \boldsymbol{x}|| = |\alpha| \, ||\boldsymbol{x}||$ *for all* $\alpha \in \mathbb{R}$ *and* $\boldsymbol{x} \in \mathbb{R}^n$*,*
4. $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$*.*

Definition 2

The l_2 *and* l_∞ *norms for the vectors* $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^t$ *are defined by*

$$||\boldsymbol{x}||_2 = \left\{ \sum_{i=1}^{n} x_i^2 \right\}^{1/2} \quad \text{and} \quad ||\boldsymbol{x}||_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Vector Norms and Distances (cont'd)

Definition 3

If $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^t$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^t$ are vectors in $\mathbb{R}^n$, the l_2 and l_∞ distances between $\boldsymbol{x}$ and $\boldsymbol{y}$ are defined by

$$||\boldsymbol{x} - \boldsymbol{y}||_2 = \left\{ \sum_{i=1}^{n} (x_i - y_i)^2 \right\}^{1/2} \quad \text{and} \quad ||\boldsymbol{x} - \boldsymbol{y}||_\infty = \max_{1 \le i \le n} |x_i - y_i|.$$

Vector Norms and Distances (cont'd)

Ex. The linear system

$$\begin{aligned} 3.3330x_1 + 15920x_2 - 10.333x_3 &= 15913, \\ 2.2220x_1 + 16.710x_2 + 9.6120x_3 &= 28.544, \\ 1.5611x_1 + 5.1791x_2 + 1.6852x_3 &= 8.4254 \end{aligned}$$

has the exact solution $\boldsymbol{x} = (1, 1, 1)^t$, and Gaussian elimination performed using five-digit rounding arithmetic and partial pivoting produces the approximate solution

$$\tilde{\boldsymbol{x}} = (1.2001, 0.9991, 0.92538)^t.$$

Determine the l_2 and l_∞ distances between the exact and approximate solutions.

Sol. Measurements of $\boldsymbol{x} - \tilde{\boldsymbol{x}}$ are given by

$$\begin{aligned} ||\boldsymbol{x} - \tilde{\boldsymbol{x}}||_2 &= [(0.2001)^2 + (0.00009)^2 + (0.07462)^2]^{1/2} = 0.21356, \\ ||\boldsymbol{x} - \tilde{\boldsymbol{x}}||_\infty &= \max\{0.2001, 0.00009, 0.07462\} = 0.2001. \end{aligned}$$

Vector Norms and Distances (cont'd)

Definition 4

A sequence $\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty}$ *of vectors in* $\mathbb{R}^n$ *is said to* **converge** *to* $\boldsymbol{x}$ *with respect to the norm* $||\cdot||$ *if, given any* $\epsilon > 0$, *there exists an integer* $N(\epsilon)$ *such that*

$$||\boldsymbol{x}^{(k)} - \boldsymbol{x}|| < \epsilon, \quad \text{for all } k \geq N(\epsilon).$$

Theorem 1

The sequence of vectors $\{\boldsymbol{x}^{(k)}\}$ *converges to* $\boldsymbol{x}$ *in* $\mathbb{R}^n$ *with respect to the* l_∞ *norm if and only if* $\lim_{k\to\infty} x_i^{(k)} = x_i$, *for each* $i = 1, 2, \ldots, n$.

Theorem 2

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

Vector Norms and Distances (cont'd)

Ex. Show that

$$\boldsymbol{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t = \left(1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k}\sin k\right)^t$$

converges to $\boldsymbol{x} = (1, 2, 0, 0)^t$ with respect to the l_∞ and l_2 norm.

Vector Norms and Distances (cont'd)

Sol. Given any $\epsilon > 0$, there exists an integer $N(\epsilon/2)$ with the property that

$$||\boldsymbol{x}^{(k)} - \boldsymbol{x}||_\infty < \frac{\epsilon}{2},$$

whenever $k \geq N(\epsilon/2)$. By Theorem 2, this implies that

$$||\boldsymbol{x}^{(k)} - \boldsymbol{x}||_2 \leq \sqrt{4}||\boldsymbol{x}^{(k)} - \boldsymbol{x}||_\infty \leq 2(\epsilon/2) = \epsilon,$$

when $k \geq N(\epsilon/2)$. So $\{\boldsymbol{x}^{(k)}\}$ also converges to $\boldsymbol{x}$ with respect to the l_2 norm. $\square$

Vector Norms and Distances (cont'd)

- It can be shown that all norms on $\mathbb{R}^n$ are equivalent with respect to convergence; that is, if $||\cdot||$ and $||\cdot||'$ are any two norms on $\mathbb{R}^n$ and $\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty}$ has the limit $\boldsymbol{x}$ with respect to $||\cdot||$, then $\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty}$ also has the limit $\boldsymbol{x}$ with respect to $||\cdot||'$.

Matrix Norms and Distances

- We also need methods for determining the distance between $n \times n$ matrices.

Definition 5

A **matrix norm** *on the set of all $n \times n$ matrices is a real-valued function, $||\cdot||$, defined on this set, satisfying for all $n \times n$ matrices A and B and all real numbers α:*

1. $||A|| \geq 0$,
2. $||A| = 0$, *if and only if A is the zero matrix,*
3. $||\alpha A|| = |\alpha| \, ||A||$,
4. $||A + B|| \leq ||A|| + ||B||$,
5. $||AB|| \leq ||A|| \, ||B||$.

Matrix Norms and Distances (cont'd)

Theorem 3

If $||\cdot||$ is a vector on $\mathbb{R}^n$, then

$$||A|| = \max_{||\boldsymbol{x}||=1} ||A\boldsymbol{x}|| = \max_{\boldsymbol{z}\neq\boldsymbol{0}} \frac{||A\boldsymbol{z}||}{||\boldsymbol{z}||}$$

is a (natural or induced) matrix norm.

Corollary 1

For any vector $\boldsymbol{z} \neq \boldsymbol{0}$, matrix A and any natural norm $||\cdot||$, we have

$$||A\boldsymbol{z}|| \leq ||A|| \cdot ||\boldsymbol{z}||.$$

Matrix Norms and Distances (cont'd)

- The measure given to a matrix under a natural norm describes how the matrix stretches unit vectors relative to that norm. The maximum stretch is the norm of the matrix.

- The matrix norms have the form

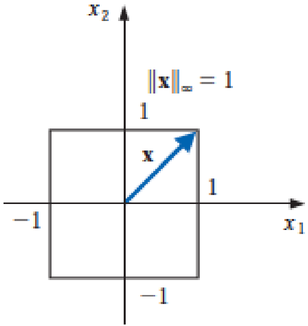
$$\text{the } l_\infty \text{ norm:} \quad ||A||_\infty = \max_{||\boldsymbol{x}||_\infty = 1} ||A\boldsymbol{x}||_\infty,$$

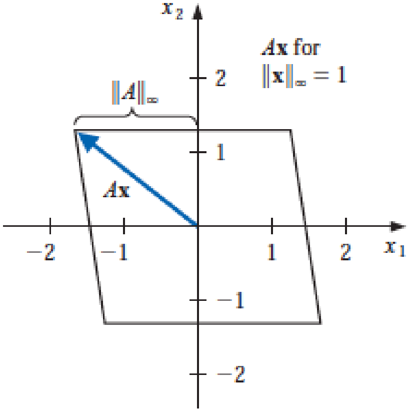
$$\text{the } l_2 \text{ norm:} \quad ||A||_2 = \max_{||\boldsymbol{x}||_2 = 1} ||A\boldsymbol{x}||_2.$$

Matrix Norms and Distances (cont'd)

- An illustration of the norms when $n = 2$ is shown below for the matrix

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

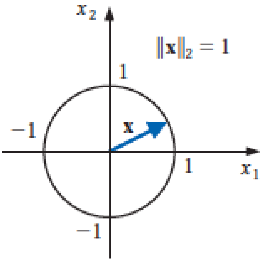


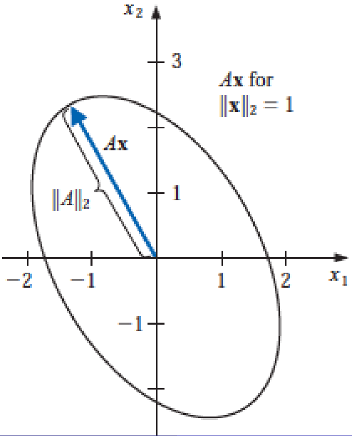


Matrix Norms and Distances (cont'd)

- An illustration of the norms when $n = 2$ is shown below for the matrix

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





Matrix Norms and Distances (cont'd)

Theorem 4

If $A = (a_{ij})$ is an $n \times n$ matrix, then

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

Ex. Determine $||A||_\infty$ for the matrix

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & -1 \\ 5 & -1 & 1 \end{bmatrix}.$$

Sol. Since $\sum_{j=1}^{3} |a_{1j}| = 4$, $\sum_{j=1}^{3} |a_{2j}| = 4$ and $\sum_{j=1}^{3} |a_{3j}| = 7$, we have $||A||_\infty = 7$.

Q. How about l_2 norm of a matrix?

Eigenvalues and Eigenvectors

Definition 6

If A is a square matrix, the **characteristic polynomial** *of A is defined by*

$$p(\lambda) = \det\{A - \lambda I\}\,.$$

Definition 7

If p is the characteristic polynomial of the matrix A, the zeros of p are **eigenvalues** *of the matrix A. If λ is an eigenvalue of A and $\boldsymbol{x} \neq \mathbf{0}$ satisfies $(A - \lambda I)\boldsymbol{x} = \mathbf{0}$, then $\boldsymbol{x}$ is an* **eigenvector** *of A corresponding to the eigenvalue λ.*

Spectral Radius

Definition 8

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

$$\rho(A) = \max |\lambda|, \quad \text{where } \lambda \text{ is an eigenvalue of } A.$$

(For complex $\lambda = \alpha + \beta i$*, we define* $|\lambda| = (\alpha^2 + \beta^2)^{1/2}$*.)*

Theorem 5

If A *is an* $n \times n$ *matrix, then*

1. $||A||_2 = [\rho(A^t A)]^{1/2}$,
2. $\rho(A) \leq ||A||$, *for any natural norm* $||\cdot||$.

Proof

2. Suppose λ is an eigenvalue of A with eigenvector $\boldsymbol{x}$ and $||\boldsymbol{x}|| = 1$. Then $A\boldsymbol{x} = \lambda\boldsymbol{x}$ and

$$|\lambda| = |\lambda| \cdot ||\boldsymbol{x}|| = ||\lambda\boldsymbol{x}|| = ||A\boldsymbol{x}|| \leq ||A||\,||\boldsymbol{x}|| = ||\boldsymbol{A}||.$$

So, $\rho(A) = \max |\lambda| \leq ||A||$. $\square$

Convergent Matrices

- In studying iterative matrix techniques, it is of particular importance to know when powers of a matrix become small.

Definition 9

We call $n \times n$ matrix A **convergent** *if*

$$\lim_{k\to\infty} (A^k)_{ij} = 0, \quad \text{for each } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, n.$$

Theorem 6

The following statements are equivalent.

1. A *is a convergent matrix.*
2. $\lim_{n\to\infty} ||A^n|| = 0$, *for some natural norm.*
3. $\lim_{n\to\infty} ||A^n|| = 0$, *for all natural norms.*
4. $\rho(A) < 1$.
5. $\lim_{n\to\infty} A^n \boldsymbol{x} = \mathbf{0}$. *for every* $\boldsymbol{x}$.

Iterative Methods

- Iterative techniques are seldom used for solving linear systems of small dimension since the time required for sufficient accuracy exceeds that required for direct techniques.
- For large systems with a high percentage of 0 entries, however, iterative techniques are efficient in terms of both computer storage and computation.
- Systems of this type arise frequently in circuit analysis and in the numerical solution of boundary-value problems and partial-differential equations.

Iterative Methods

- Jacobi's method
- Gauss-Seidel method

- An iterative technique to solve the $n \times n$ linear system $A\boldsymbol{x} = \boldsymbol{b}$ starts with an initial approximation $\boldsymbol{x}^{(0)}$ to the solution $\boldsymbol{x}$ and generates a sequence of vectors $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ that converges to $\boldsymbol{x}$.

Jacobi's Method

- The **Jacobi iterative method** is obtained by solving the ith equation in $A\boldsymbol{x} = \boldsymbol{b}$ for x_i to obtain (provided $a_{ii} \neq 0$)

$$x_i = \frac{1}{a_{ii}} \left[-\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j + b_i \right], \quad \text{for } i = 1, 2, \ldots, n.$$

Jacobi's Method

- For each $k \geq 1$, generate the components $x_i^{(k)}$ of $\boldsymbol{x}^{(k)}$ from the components of $\boldsymbol{x}^{(k-1)}$ by

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[-\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k-1)} + b_i \right], \quad \text{for } i = 1, 2, \ldots, n.$$

Jacobi's Method (cont'd)

Ex. The linear system $A\boldsymbol{x} = \boldsymbol{b}$ given by

$$\begin{aligned} E_1 &: 10x_1 - x_2 + 2x_3 = 6, \\ E_2 &: -x_1 + 11x_2 - x_3 + 3x_4 = 25, \\ E_3 &: 2x_1 - x_2 + 10x_3 - x_4 = -11, \\ E_4 &: 3x_2 - x_3 + 8x_4 = 15 \end{aligned}$$

has the unique solution $\boldsymbol{x} = (1, 2, -1, 1)^t$. Use Jacobi's iterative technique to find approximation $\boldsymbol{x}^{(k)}$ to $\boldsymbol{x}$ starting with $\boldsymbol{x}^{(0)} = (0, 0, 0, 0)^t$ until

$$\frac{||\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}||_\infty}{||\boldsymbol{x}^{(k)}||_\infty} < 10^{-3}.$$

Jacobi's Method (cont'd)

Sol. We first solve equation E_i for x_i, for each $i = 1, 2, 3, 4$, to obtain

$$\begin{aligned}
x_1 &= \frac{1}{10}x_2 - \frac{1}{5}x_3 + \frac{3}{5}, \\
x_2 &= \frac{1}{11}x_1 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11}, \\
x_3 &= -\frac{1}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}x_4 - \frac{11}{10}, \\
x_4 &= -\frac{3}{8}x_2 + \frac{1}{8}x_3 + \frac{15}{8}.
\end{aligned}$$

We then have $\boldsymbol{x}^{(1)}$ given by

$$\begin{aligned}
x_1^{(1)} &= \frac{1}{10}x_2^{(0)} - \frac{1}{5}x_3^{(0)} + \frac{3}{5} = 0.6000, \\
x_2^{(1)} &= \frac{1}{11}x_1^{(0)} + \frac{1}{11}x_3^{(0)} - \frac{3}{11}x_4^{(0)} + \frac{25}{11} = 2.2727, \\
x_3^{(1)} &= -\frac{1}{5}x_1^{(0)} + \frac{1}{10}x_2^{(0)} + \frac{1}{10}x_4^{(0)} - \frac{11}{10} = -1.1000, \\
x_4^{(1)} &= -\frac{3}{8}x_2^{(0)} + \frac{1}{8}x_3^{(0)} + \frac{15}{8} = 1.8750.
\end{aligned}$$

Jacobi's Method (cont'd)

- Additional iterates $\boldsymbol{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t$, are generated in a similar manner and are presented below.

k	0	1	2	3	4	5	6	7	8	9	10
$x_1^{(k)}$	0.0000	0.6000	1.0473	0.9326	1.0152	0.9890	1.0032	0.9981	1.0006	0.9997	1.0001
$x_2^{(k)}$	0.0000	2.2727	1.7159	2.053	1.9537	2.0114	1.9922	2.0023	1.9987	2.0004	1.9998
$x_3^{(k)}$	0.0000	−1.1000	−0.8052	−1.0493	−0.9681	−1.0103	−0.9945	−1.0020	−0.9990	−1.0004	−0.9998
$x_4^{(k)}$	0.0000	1.8750	0.8852	1.1309	0.9739	1.0214	0.9944	1.0036	0.9989	1.0006	0.9998

- Terminated after ten iterations because

$$\frac{||\boldsymbol{x}^{(10)} - \boldsymbol{x}^{(9)}||_\infty}{||\boldsymbol{x}^{(10)}||_\infty} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3}.$$

In fact, $||\boldsymbol{x}^{(10)} - \boldsymbol{x}||_\infty = 0.0002$.

Jacobi's Method (cont'd)

- In general, iterative techniques for solving linear systems involve a process that converts the system $A\boldsymbol{x} = \boldsymbol{b}$ into an equivalent system of the form $\boldsymbol{x} = T\boldsymbol{x} + \boldsymbol{c}$ for some fixed matrix T and vector $\boldsymbol{c}$.

- After the initial vector $\boldsymbol{x}^{(0)}$ is selected, the sequence of approximate solution vectors is generated by computing

$$\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$$

for each $k = 1, 2, 3, \ldots$, reminiscent of the fixed-point iteration.

Q. What are T and $\boldsymbol{c}$ for Jacobi's method?

Jacobi's Method (cont'd)

- D: the diagonal matrix whose diagonal entries are those of A
- $-L$: the strictly lower-triangular part of A
- $-U$: the strictly upper-triangular part of A

$$A = \underbrace{\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix}}_{D} - \underbrace{\begin{bmatrix} 0 & \cdots & \cdots & 0 \\ -a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix}}_{L} - U$$

Jacobi's Method (cont'd)

- The equation $A\boldsymbol{x} = (D - L - U)\boldsymbol{x} = \boldsymbol{b}$ is then transformed into

$$D\boldsymbol{x} = (L + U)\boldsymbol{x} + \boldsymbol{b},$$

and, if D^{-1} exists, that is, if $a_{ii} \neq 0$ for each i, then

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

- The matrix form of the Jacobi iterative technique is

$$\boldsymbol{x}^{(k)} = \underbrace{D^{-1}(L + U)}_{T_j}\boldsymbol{x}^{(k-1)} + \underbrace{D^{-1}\boldsymbol{b}}_{\boldsymbol{c}_j}.$$

Jacobi's Method (cont'd)

- To avoid $a_{ii} = 0$, one should reorder the equations.
- To speed up, one should rearrange so that a_{ii} is as large as possible. Why?
- Jacobi's method: For each $k \geq 1$, generate the components $x_i^{(k)}$ of $\boldsymbol{x}^{(k)}$ from the components of $\boldsymbol{x}^{(k-1)}$ by

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[-\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k-1)} + b_i \right], \quad \text{for } i = 1, 2, \ldots, n.$$

Q. Can we do better?

The Gauss-Seidel Method

- For $i > 1$, the components $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$ of $\boldsymbol{x}^{(k)}$ have already been computed.

- **Gauss-Seidel iterative method** updates as

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

for each $i = 1, 2, \ldots, n$.

The Gauss-Seidel Method (cont'd)

Ex. Use the Gauss-Seidel iterative technique to find approximate solutions to

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

starting with $\boldsymbol{x} = (0, 0, 0, 0)^t$ and iterating until

$$\frac{||\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}||_\infty}{||\boldsymbol{x}^{(k)}||_\infty} < 10^{-3}.$$

The Gauss-Seidel Method (cont'd)

Sol. Gauss-Seidel method updates as

$$
\begin{aligned}
x_1^{(k)} &= \frac{1}{10}x_2^{(k-1)} - \frac{1}{5}x_3^{(k-1)} + \frac{3}{5}, \\
x_2^{(k)} &= \frac{1}{11}x_1^{(k)} + \frac{1}{11}x_3^{(k-1)} - \frac{3}{11}x_4^{(k-1)} + \frac{25}{11}, \\
x_3^{(k)} &= -\frac{1}{5}x_1^{(k)} + \frac{1}{10}x_2^{(k)} + \frac{1}{10}x_4^{(k-1)} - \frac{11}{10}, \\
x_4^{(k)} &= -\frac{3}{8}x_2^{(k)} - \frac{1}{8}x_3^{(k)} + \frac{15}{8}.
\end{aligned}
$$

The Gauss-Seidel Method (cont'd)

- Gauss-Seidel method then generates the values below.

k	0	1	2	3	4	5
$x_1^{(k)}$	0.0000	0.6000	1.030	1.0065	1.0009	1.0001
$x_2^{(k)}$	0.0000	2.3272	2.037	2.0036	2.0003	2.0000
$x_3^{(k)}$	0.0000	−0.9873	−1.014	−1.0025	−1.0003	−1.0000
$x_4^{(k)}$	0.0000	0.8789	0.9844	0.9983	0.9999	1.0000

- Terminated after five iterations because

$$\frac{||\boldsymbol{x}^{(5)} - \boldsymbol{x}^{(4)}||_\infty}{||\boldsymbol{x}^{(5)}||_\infty} = \frac{0.0008}{2.000} = 4 \times 10^{-4} < 10^{-3},$$

which requires twice less iterations than Jacobi's method.

The Gauss-Seidel Method (cont'd)

- **Gauss-Seidel iterative method** updates as

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

  for each $i = 1, 2, \ldots, n$.

- To write the Gauss-Seidel method in a matrix form $\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$, multiply the equation by a_{ii} and rearrange it as

$$a_{i1}x_1^{(k)} + a_{i2}x_2^{(k)} + \cdots + a_{ii}x_i^{(k)} = -a_{i,i+1}x_{i+1}^{(k-1)} - \cdots - a_{in}x_n^{(k-1)} + b_i,$$

  for each $i = 1, 2, \ldots, n$.

The Gauss-Seidel Method (cont'd)

- Writing all equations gives

$$
\begin{aligned}
a_{11}x_1^{(k)} &= -a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1 \\
a_{21}x_1^{(k)} + a_{22}x_2^{(k)} &= -a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2 \\
&\vdots \\
a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \cdots + a_{nn}x_n^{(k)} &= b_n
\end{aligned}
$$

The Gauss-Seidel Method (cont'd)

- With the definitions of D, L, U, Gauss-Seidel method can be represented by

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

and

$$\boldsymbol{x}^{(k)} = \underbrace{(D - L)^{-1}U}_{T_g}\,\boldsymbol{x}^{(k-1)} + \underbrace{(D - L)^{-1}\boldsymbol{b}}_{\boldsymbol{c}_g}$$

for each $k = 1, 2, \ldots$.

- For the lower-triangular matrix $D - L$ to be nonsingular, it is necessary and sufficient that $a_{ii} \neq 0$, for each $i = 1, 2, \ldots, n$.

The Gauss-Seidel Method (cont'd)

- Reordering is also useful in Gauss-Seidel method.

- For some examples, we have seen that Gauss-Seidel method is superior to the Jacobi method.
- This is almost always true, but there are linear systems, where Jacobi method converges, but Gauss-Seidel method does not.

General Iteration Methods

- To study the convergence of general iteration techniques, we need to analyze the formula

$$\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}, \quad \text{for each } k = 1, 2, \ldots,$$

where $\boldsymbol{x}^{(0)}$ is arbitrary.

General Iteration Methods (cont'd)

Lemma 1

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

$$(I - T)^{-1} = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j,$$

Proof We first prove that $(I - T)^{-1}$ exists.

Because $T\boldsymbol{x} = \lambda \boldsymbol{x}$ is true when $(I - T)\boldsymbol{x} = (1 - \lambda)\boldsymbol{x}$, we have λ as an eigenvalue of T when $1 - \lambda$ is an eigenvalue of $I - T$. Since $|\lambda| \leq \rho(T) < 1$, $\lambda = 1$ is not an eigenvalue of T, and thus 0 cannot be an eigenvalue of $I - T$. Hence, $(I - T)^{-1}$ exists.

General Iteration Methods (cont'd)

Proof Let $S_m = I + T + T^2 + \cdots T^m$. Then

$$(I - T)S_m = I - T^{m+1}$$

and, since T is convergent, Theorem 6 implies that

$$\lim_{m\to\infty}(I - T)S_m = \lim_{m\to\infty}(I - T^{m+1}) = I.$$

Thus,

$$(I - T)^{-1} = \lim_{m\to\infty} S_m = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j.$$

□

General Iteration Methods (cont'd)

Theorem 7

For any $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ defined by

$$\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}, \quad \text{for each } k \geq 1,$$

converges to the unique solution of $\boldsymbol{x} = T\boldsymbol{x} + \boldsymbol{c}$ if and only if $\rho(T) < 1$.

Proof First assume that $\rho(T) < 1$. Then,

General Iteration Methods (cont'd)

Proof To prove the converse, we will show that for any $\boldsymbol{z} \in \mathbb{R}^n$, we have $\lim_{k\to\infty} T^k \boldsymbol{z} = \mathbf{0}$, which is equivalent to $\rho(T) < 1$.

Let $\boldsymbol{z}$ be an arbitrary vector, and $\boldsymbol{x}$ be the unique solution to $\boldsymbol{x} = T\boldsymbol{x} + \boldsymbol{c}$. Define $\boldsymbol{x}^{(0)} = \boldsymbol{x} - \boldsymbol{z}$, and, for $k \geq 1$, $\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$. Then, $\{\boldsymbol{x}^{(k)}\}$ converges to $\boldsymbol{x}$. Also,

General Iteration Methods (cont'd)

Corollary 2

If $||T|| < 1$ *for any natural matrix norm and* $\boldsymbol{c}$ *is a given vector, then the sequence* $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ *defined by* $\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$ *converges, for any* $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$*, to a vector* $\boldsymbol{x} \in \mathbb{R}^n$*, with* $\boldsymbol{x} = T\boldsymbol{x} + \boldsymbol{c}$*, and the following error bounds hold:*

1. $||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| \leq ||T||^k ||\boldsymbol{x}^{(0)} - \boldsymbol{x}||$,
2. $||\boldsymbol{x} - \boldsymbol{x}^{(k)}|| \leq \frac{||T||^k}{1-||T||} ||\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}||$.

General Iteration Methods (cont'd)

- Jacobi's method: $T_j = D^{-1}(L + U)$
- Gauss-Seidel method: $T_g = (D - L)^{-1}U$

- If $\rho(T_j)$ or $\rho(T_g)$ is less than 1, then the corresponding sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ will converge to the solution $\boldsymbol{x}$ of $A\boldsymbol{x} = \boldsymbol{b}$.

General Iteration Methods (cont'd)

Definition 10

The $n \times n$ matrix A is said to be **diagonally dominant** *when*

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \textit{holds for each } i = 1, 2, \ldots, n.$$

A diagonally dominant matrix is said to be **strictly diagonally dominant** *when*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \textit{holds for each } i = 1, 2, \ldots, n.$$

Theorem 8

A strictly diagonally dominant matrix $\boldsymbol{A}$ is nonsingular. Moreover, in this case, Gaussian elimination can be performed on any linear system of the form $\boldsymbol{Ax} = \boldsymbol{b}$ to obtain its unique solution without row or column interchanges.

General Iteration Methods (cont'd)

Theorem 9

If A is strictly diagonally dominant, then for any choice of $\boldsymbol{x}^{(0)}$, both the Jacobi and Gauss-Seidel methods give sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ that converge to the unique solution of $A\boldsymbol{x} = \boldsymbol{b}$.

General Iteration Methods (cont'd)

- For any matrix T and any $\epsilon > 0$, there exists a natural norm $||\cdot||$ with the property that

$$\rho(T) \leq ||T|| \leq \rho(T) + \epsilon.$$

- By Theorem 9 and above statement, we have

$$||\boldsymbol{x}^{(k)} - \boldsymbol{x}|| \approx \rho(T)^k ||\boldsymbol{x}^{(0)} - \boldsymbol{x}||.$$

- Thus, we would like to select the iterative technique with minimal $\rho(T) < 1$ for a particular system $A\boldsymbol{x} = \boldsymbol{b}$.

General Iteration Methods (cont'd)

- No general results exist to tell which of the two techniques, Jacobi or Gauss-Seidel, will be most successful for an arbitrary linear system.
- However, for some special cases, the answer is known.

Theorem 10

If $a_{ij} \le 0$, for each $i \ne j$ and $a_{ii} > 0$, for each $i = 1, 2, \ldots, n$, then one and only one of the following statements holds:

1. $0 \le \rho(T_g) < \rho(T_j) < 1$
2. $1 < \rho(T_j) < \rho(T_g)$
3. $\rho(T_j) = \rho(T_g) = 0$
4. $\rho(T_j) = \rho(T_g) = 1.$

Residual Vector

- Accelerating convergence can be achieved by choosing a method whose associated matrix has minimal spectral radius.
- For such purpose, we need to first introduce a new means of measuring the amount by which an approximation differs from the true solution to the system.

Definition 11

Suppose $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ is an approximation to the solution of the linear system defined by $A\boldsymbol{x} = \boldsymbol{b}$. The **residual vector** *for $\tilde{\boldsymbol{x}}$ with respect to this system is $\boldsymbol{r} = \boldsymbol{b} - A\tilde{\boldsymbol{x}}$.*

Iterative Method and Residual Vector

- In procedures such as the Jacobi or Gauss-Seidel methods, a residual vector $\boldsymbol{r}$ is associated with each calculation.
- The true objective of those procedures is to generate a sequence of approximations that makes the residual vectors $\boldsymbol{r}$ to rapidly converge to zero.

Iterative Method and Residual Vector (cont'd)

- Let

$$\boldsymbol{r}_i^{(k)} = (r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)})^t$$

be the residual vector for the Gauss-Seidel method, corresponding to $\boldsymbol{x}_i^{(k)}$ defined by

$$\boldsymbol{x}_i^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)})^t.$$

Iterative Method and Residual Vector (cont'd)

- The mth component of $\boldsymbol{r}_i^{(k)}$ is

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^{n} a_{mj} x_j^{(k-1)}$$
$$= b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)}$$

for each $m = 1, 2, \ldots, n$.

- In particular, the ith component of $\boldsymbol{r}_i^{(k)}$ is

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

Iterative Method and Residual Vector (cont'd)

- Recall that in the Gauss-Seidel method, $x_i^{(k)}$ is chosen to be

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

which is equivalent to

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}.$$

Iterative Method and Residual Vector (cont'd)

- Consider $\boldsymbol{r}_{i+1}^{(k)}$, associated with $\boldsymbol{x}_{i+1}^{(k)}$,

$$
\begin{aligned}
r_{i,i+1}^{(k)} &= b_i - \sum_{j=1}^{i} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \\
&= b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k)}
\end{aligned}
$$

  that is $\mathbf{0}$, implying that the Gauss-Seidel chooses each $x_i^{(k)}$ in such a way that the ith component of $\boldsymbol{r}_{i+1}^{(k)}$ is zero.

- However, choosing $x_i^{(k)}$ so that one coordinate of the residual vector is zero is not necessarily the most efficient way to reduce to the norm of the vector $\boldsymbol{r}_{i+1}^{(k)}$. Then how?

Relaxation Methods

- Consider modifying the Gauss-Seidel procedure as

$$x_i^{(k)} = x_i^{(k-1)} + w\frac{r_{ii}^{(k)}}{a_{ii}}$$

for a positive w, which is called **relaxation methods**.

- Methods with $0 < w < 1$ are called **under-relaxation methods**.
- Methods with $1 < w$ are called **over-relaxation methods**, or **SOR** (for successive over-relaxation).

Relaxation Methods (cont'd)

- SOR method can be rewritten as

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

  and

$$a_{ii}x_i^{(k)} + w\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} = (1-w)a_{ii}x_i^{(k-1)} - w\sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} + wb_i,$$

  for each $i = 1, 2, \ldots, n$.

- Considering all n equations, we have the vector form

$$(D - wL)\boldsymbol{x}^{(k)} = [(1-w)D + wU]\boldsymbol{x}^{(k-1)} + w\boldsymbol{b}.$$

Relaxation Methods (cont'd)

- The matrix form of SOR is then

$$\boldsymbol{x}^{(k)} = \underbrace{(D - wL)^{-1}[(1-w)D + wU]}_{T_w}\boldsymbol{x}^{(k-1)} + \underbrace{w(D-wL)^{-1}\boldsymbol{b}}_{\boldsymbol{c}_w}.$$

Relaxation Methods (cont'd)

Ex. The linear system $A\boldsymbol{x} = \boldsymbol{b}$ given by

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

has the solution $(3, 4, -5)^t$. Compare the iterations from the Gauss-Seidel method and the SOR method with $w = 1.25$ using $\boldsymbol{x}^{(0)} = (1, 1, 1)^t$ for both methods.

Sol. For each $k = 1, 2, \ldots,$ the equations for the Gauss-Seidel method are

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

Relaxation Methods (cont'd)

Sol. Then the equations for the SOR method with $w = 1.25$ are

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

Relaxation Methods (cont'd)

Sol. The first seven iterates for each method are listed below.

k	0	1	2	3	4	5	6	7
$x_1^{(k)}$	1	5.250000	3.1406250	3.0878906	3.0549316	3.0343323	3.0214577	3.0134110
$x_2^{(k)}$	1	3.812500	3.8828125	3.9267578	3.9542236	3.9713898	3.9821186	3.9888241
$x_3^{(k)}$	1	−5.046875	−5.0292969	−5.0183105	−5.0114441	−5.0071526	−5.0044703	−5.0027940

k	0
$x_1^{(k)}$	1
$x_2^{(k)}$	1
$x_3^{(k)}$	1

- For the iterates to be accurate to seven decimal places, the Gauss-Seidel requires 34 iterations, while the SOR with $w = 1.25$ requires 14 iterations.

Relaxation Methods (cont'd)

Q. How should we choose w?

Theorem 11

If $a_{ii} \neq 0$, for each $i = 1, 2, \ldots, n$, then $\rho(T_w) \geq |w - 1|$. This implies that the SOR method can converge only if $0 < w < 2$.

Proof Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of T_w. Then

$$\begin{aligned}\prod_{i=1}^{n} \lambda_i &= \det T_w = \det\left((D - wL)^{-1}[(1-w)D + wU]\right) \\ &= \det{}^{-1}\{D - wL\} \det\{(1-w\} D + wU) \\ &= \det D^{-1} \det\{(1-w\} D) \\ &= (a_{11} \cdots a_{nn})^{-1}(1-w)^n(a_{11} \cdots a_{nn}) = (1-w)^n.\end{aligned}$$

Thus,

$$\rho(T_w) = \max_{1 \leq i \leq n} |\lambda_i| \geq |1 - w|,$$

so $0 < w < 2$ (*i.e.*, $|1 - w| < 1$) is required for the SOR method to converge. $\square$

Relaxation Methods (cont'd)

Theorem 12

If A is a positive definite matrix and $0 < w < 2$, then the SOR method converges for any choice of initial approximate vector $\boldsymbol{x}^{(0)}$.

Theorem 13

If A is positive definite and tridiagonal, then $\rho(T_g) = [\rho(T_j)]^2 < 1$, and the optimal choice of w for the SOR method is

$$w = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}}.$$

With this choice of w, we have $\rho(T_w) = w - 1$.

Relaxation Methods (cont'd)

Ex. Find the optimal choice of w for the SOR method for the positive definite and tridiagonal matrix

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}.$$

Error Bounds

- Small $||\boldsymbol{r}||$ where $\boldsymbol{r} = \boldsymbol{b} - A\tilde{\boldsymbol{x}}$ does not necessarily mean $||\boldsymbol{x} - \tilde{\boldsymbol{x}}||$. When?

Ex. The linear system $A\boldsymbol{x} = \boldsymbol{b}$ given by

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

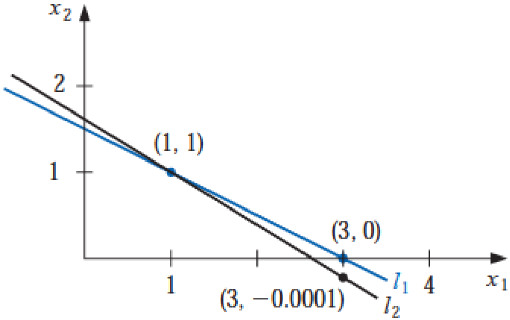
has the unique solution $\boldsymbol{x} = (1, 1)^t$. Determine the residual vector for the poor approximation $\tilde{\boldsymbol{x}} = (3, -0.0001)^t$.

Error Bounds (cont'd)

- The solution is the intersection of the nearly parallel lines

$$l_1 \ : \ x_1 + 2x_2 = 3 \quad \text{and} \quad l_1 \ : \ 1.0001x_1 + 2x_2 = 3.0001$$

- The point $(3, -0.0001)$ lies on l_2, and lies close to l_1, even though it differs significantly from the solution.



- How can we identify such characteristic from the system?

Error Bounds (cont'd)

Theorem 14

Suppose that $\tilde{\boldsymbol{x}}$ is an approximation to the solution of $A\boldsymbol{x} = \boldsymbol{b}$, A is a nonsingular matrix, and $\boldsymbol{r}$ is the residual vector for $\tilde{\boldsymbol{x}}$. Then for any natural norm,

$$||\boldsymbol{x} - \tilde{\boldsymbol{x}}|| \leq ||\boldsymbol{r}|| \cdot ||A^{-1}||$$

and if $\boldsymbol{x} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$,

$$\frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||} \leq ||A|| \cdot ||A^{-1}|| \frac{||\boldsymbol{r}||}{||\boldsymbol{b}||}.$$

Condition Numbers

- The previous theorem implies that $||A^{-1}||$ and $||A|| \cdot ||A^{-1}||$ provide an indication of the connection between the residual vector and the accuracy of the approximation.

- The relative error $||\boldsymbol{x} - \tilde{\boldsymbol{x}}||/||\boldsymbol{x}||$ is of most interest, which is bounded by the product of $||A|| \cdot ||A^{-1}||$ with the relative residual.

Definition 12

The **condition number** *of the nonsingular matrix A relative to a norm $||\cdot||$ is*

$$K(A) = ||A|| \cdot ||A^{-1}||.$$

Condition Numbers (cont'd)

- We then have

$$||\boldsymbol{x} - \tilde{\boldsymbol{x}}|| \leq K(A)\frac{||\boldsymbol{r}||}{||A||} \quad \text{and} \quad \frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||} \leq K(A)\frac{||\boldsymbol{r}||}{||\boldsymbol{b}||}.$$

- For any nonsingular matrix A and natural norm $||\cdot||$,

$$1 = ||I|| = ||A \cdot A^{-1}|| \leq ||A|| \cdot ||A^{-1}|| = K(A).$$

- $K(A) \approx 1$: A is **well-conditioned**,
- $K(A) \gg 1$: A is **ill-conditioned**.

- Conditioning in this context refers to the relative security that a small residual vector implies a correspondingly accurate approximate solution.

Condition Numbers (cont'd)

Ex. Determine the condition number for the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}$$

Sol. Since

$$A^{-1} = \begin{bmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{bmatrix},$$

we have

$$||A||_\infty = \max\{|1| + |2|, |1.001| + |2|\} = 3.0001 \quad \text{and} \quad ||A^{-1}||_\infty = 20000.$$

Thus, for the infinity norm, $K(A) = 20000 \times 3.0001 = 60002$, so making accuracy decisions based on the residual of an approximation should be avoided.

Perturbed Linear System

- In practice, the entries a_{ij} and b_j will be perturbed by an amount δa_{ij} and δb_j, causing the linear system

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

  to be solved instead of $A\boldsymbol{x} = \boldsymbol{b}$.

- Even when exact A and $\boldsymbol{b}$ are used, rounding errors can cause $||\boldsymbol{x} - \tilde{\boldsymbol{x}}||$ to be large.

Perturbed Linear System (cont'd)

Theorem 15

Suppose A is nonsingular and

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

The solution $\tilde{\boldsymbol{x}}$ to $(A+\delta A)\tilde{\boldsymbol{x}} = \boldsymbol{b} + \delta\boldsymbol{b}$ approximates the solution $\boldsymbol{x}$ of $A\boldsymbol{x} = \boldsymbol{b}$ with the error estimate

$$\frac{||\boldsymbol{x}-\tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||} \leq \frac{K(A)||A||}{||A|| - K(A)||\delta A||}\left(\frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||} + \frac{||\delta A||}{||A||}\right)$$

- If A is well-conditioned, then small changes in A and $\boldsymbol{b}$ produce correspondingly small changes in the solution $\boldsymbol{x}$.

Conjugate Gradient Method

- As a Direct Method: in general, Gaussian elimination with pivoting is superior.

- As an Iterative Method: when solving large sparse systems, particularly with nonzero entries occurring in predictable patterns, it is preferred over Gaussian elimination and the previously-discussed iterative methods.

Inner Product

- Assume that A is (symmetric and) positive definite.
- Use the inner product notation

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^t \boldsymbol{y},$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are n-dimensional vectors.

Theorem 16

For any vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ and any real number α, we have

- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$
- $\langle \alpha \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \alpha \boldsymbol{y} \rangle = \alpha \langle \boldsymbol{x}, \boldsymbol{y} \rangle$
- $\langle \boldsymbol{x} + \boldsymbol{z}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{z}, \boldsymbol{y} \rangle$
- $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$
- $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ *if and only if* $\boldsymbol{x} = \mathbf{0}$.

Inner Product (cont'd)

- When A is positive definite,

$$\langle \boldsymbol{x},\, A\boldsymbol{x} \rangle = \boldsymbol{x}^t A \boldsymbol{x} > 0$$

  unless $\boldsymbol{x} = \mathbf{0}$.

- Since A is symmetric, we have $\boldsymbol{x}^t A \boldsymbol{y} = x^t A^t \boldsymbol{y} = (A\boldsymbol{x})^t \boldsymbol{y}$, so we have for each $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\langle A\boldsymbol{x},\, \boldsymbol{y} \rangle = (A\boldsymbol{x})^t \boldsymbol{y} = \boldsymbol{x}^t A^t \boldsymbol{y} = \boldsymbol{x}^t A \boldsymbol{y} = \langle \boldsymbol{x},\, A\boldsymbol{y} \rangle .$$

Linear System and Minimization

Theorem 17

The vector $\boldsymbol{x}^*$ is a solution to the positive definite linear system $A\boldsymbol{x} = \boldsymbol{b}$ if and only if $\boldsymbol{x}^*$ produces the minimal value of

$$g(\boldsymbol{x}) = \langle \boldsymbol{x},\, A\boldsymbol{x} \rangle - 2 \langle \boldsymbol{x},\, \boldsymbol{b} \rangle .$$

Proof Let $\boldsymbol{x}$ and $\boldsymbol{v} \neq \mathbf{0}$ be fixed vectors and t a real number variable. We have

$$\begin{aligned} g(\boldsymbol{x} + t\boldsymbol{v}) &= \langle \boldsymbol{x} + t\boldsymbol{v},\, A\boldsymbol{x} + tA\boldsymbol{v} \rangle - 2 \langle \boldsymbol{x} + t\boldsymbol{v},\, \boldsymbol{b} \rangle \\ &= \langle \boldsymbol{x},\, A\boldsymbol{x} \rangle + t \langle \boldsymbol{v},\, A\boldsymbol{x} \rangle + t \langle \boldsymbol{x},\, A\boldsymbol{v} \rangle + t^2 \langle \boldsymbol{v},\, A\boldsymbol{v} \rangle - 2 \langle \boldsymbol{x},\, \boldsymbol{b} \rangle - 2t \langle \boldsymbol{v},\, \boldsymbol{b} \rangle \\ &= \underbrace{\langle \boldsymbol{x},\, A\boldsymbol{x} \rangle - 2 \langle \boldsymbol{x},\, \boldsymbol{b} \rangle}_{g(x)} - 2t \langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x} \rangle + t^2 \langle \boldsymbol{v},\, A\boldsymbol{v} \rangle \end{aligned}$$

- Define the quadratic function h in t by

$$h(t) = g(\boldsymbol{x} + t\boldsymbol{v}).$$

Linear System and Minimization (cont'd)

Proof h has a minimal value when

$$h'(t) = -2\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x}\rangle + 2t\langle \boldsymbol{v},\, A\boldsymbol{v}\rangle = 0,$$

and thus when

$$\hat{t} = \frac{\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x}\rangle}{\langle \boldsymbol{v},\, A\boldsymbol{v}\rangle}.$$

We then have

$$\begin{aligned} h(\hat{t}) = g(\boldsymbol{x} + \hat{t}\boldsymbol{v}) &= g(\boldsymbol{x}) - 2\hat{t}\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x}\rangle + \hat{t}^2\langle \boldsymbol{v},\, A\boldsymbol{v}\rangle \\ &= g(\boldsymbol{x}) - 2\frac{\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x}\rangle}{\langle \boldsymbol{v},\, A\boldsymbol{v}\rangle}\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x}\rangle + \left(\frac{\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x}\rangle}{\langle \boldsymbol{v},\, A\boldsymbol{v}\rangle}\right)^2\langle \boldsymbol{v},\, A\boldsymbol{v}\rangle \\ &= g(\boldsymbol{x}) - \frac{\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x}\rangle^2}{\langle \boldsymbol{v},\, A\boldsymbol{v}\rangle}. \end{aligned}$$

Linear System and Minimization (cont'd)

Proof So for any vector $\boldsymbol{v} \neq \mathbf{0}$, we have $g(\boldsymbol{x} + \hat{t}\boldsymbol{v}) < g(\boldsymbol{x})$ unless $\langle \boldsymbol{v}, \boldsymbol{b} - A\boldsymbol{x} \rangle = 0$, in which case $g(\boldsymbol{x}) = g(\boldsymbol{x} + \hat{t}\boldsymbol{v}) = h(\hat{t})$.

- "$\Rightarrow$": Suppose $\boldsymbol{x}^*$ satisfies $A\boldsymbol{x}^* = \boldsymbol{b}$. Then $\langle \boldsymbol{v}, \boldsymbol{b} - A\boldsymbol{x}^* \rangle = 0$ for any vector $\boldsymbol{v}$, and $g(\boldsymbol{x})$ cannot be smaller than $g(\boldsymbol{x}^*)$. Thus $\boldsymbol{x}^*$ minimizes g.

- "$\Leftarrow$": Suppose $\boldsymbol{x}^*$ minimizes g. Then for any vector $\boldsymbol{v}$, we have

$$g(\boldsymbol{x}^* + \hat{t}\boldsymbol{v}) \geq g(\boldsymbol{x}^*).$$

Thus $\langle \boldsymbol{v}, \boldsymbol{b} - A\boldsymbol{x}^* \rangle = 0$, implying that $A\boldsymbol{x}^* = \boldsymbol{b}$. $\square$

Search Direction

- Let $\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}$ be the residual vector associated with $\boldsymbol{x}$ and

$$t = \frac{\langle \boldsymbol{v},\, \boldsymbol{b} - A\boldsymbol{x} \rangle}{\langle \boldsymbol{v},\, A\boldsymbol{v} \rangle} = \frac{\langle \boldsymbol{v},\, \boldsymbol{r} \rangle}{\langle \boldsymbol{v},\, A\boldsymbol{v} \rangle}.$$

If $\boldsymbol{r} \neq \mathbf{0}$ and if $\boldsymbol{v}$ and $\boldsymbol{r}$ are not orthogonal, then

$$\boldsymbol{x} + t\boldsymbol{v}$$

gives a smaller value for g than $g(\boldsymbol{x})$ and is presumably closer to $\boldsymbol{x}^*$ than is $\boldsymbol{x}$.

Search Direction (cont'd)

- Let $\boldsymbol{x}^{(0)}$ be an initial approximation to $\boldsymbol{x}^*$, and let $\boldsymbol{v}^{(1)} \neq \mathbf{0}$ be an initial search direction. For $k = 1, 2, 3, \ldots$, we compute

$$t_k = \frac{\langle \boldsymbol{v}^{(k)}, \, \boldsymbol{b} - A\boldsymbol{x}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, \, A\boldsymbol{v}^{(k)} \rangle}$$

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)}$$

  and choose a new search direction $\boldsymbol{v}^{(k+1)}$.

Q. How should we choose the search direction?

Search Direction (cont'd)

- The gradient of g is

$$\nabla g(\boldsymbol{x}) = 2(A\boldsymbol{x} - \boldsymbol{b}) = -2\boldsymbol{r}$$

where $\boldsymbol{r}$ is the residual vector for $\boldsymbol{x}$.

- The direction of greatest decrease in the value of $g(\boldsymbol{x})$ is the direction given by $-\nabla g(\boldsymbol{x})$.

- The method of steepest descent (also known as the gradient descent method) uses

$$\boldsymbol{v}^{(k+1)} = \boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)},$$

which has slow convergence.

A-Orthogonality Condition

- Alternatively, use a set of nonzero direction vectors $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ that satisfy

$$\langle \boldsymbol{v}^{(i)}, A\boldsymbol{v}^{(j)} \rangle = 0, \quad \text{if } i \neq j.$$

Q. Why?

- This is called an A-**orthogonality condition**, and the set of vectors $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ is said to be A-**orthogonal**.

- One can show that a set of A-orthogonal vectors associated with the positive definite matrix A is linearly independent.

A-Orthogonality Condition (cont'd)

Theorem 18

Let $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ be an A-orthogonal set of nonzero vectors associated with the positive definite matrix A, and let $\boldsymbol{x}^{(0)}$ be arbitrary. Define

$$t_k = \frac{\langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - A\boldsymbol{x}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, A\boldsymbol{v}^{(k)} \rangle} \quad \text{and} \quad \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)},$$

for $k = 1, 2, \ldots, n$. Then assuming exact arithmetic, $A\boldsymbol{x}^{(n)} = \boldsymbol{b}$.

Proof Since, for each $k = 1, 2, \ldots, n$, $\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)}$, we have

$$\begin{aligned} A\boldsymbol{x}^{(n)} &= A\boldsymbol{x}^{(n-1)} + t_n A\boldsymbol{v}^{(n)} \\ &= (A\boldsymbol{x}^{(n-2)} + t_{n-1} A\boldsymbol{v}^{(n-1)}) + t_n A\boldsymbol{v}^{(n)} \\ &\vdots \\ &= A\boldsymbol{x}^{(0)} + t_1 A\boldsymbol{v}^{(1)} + t_2 A\boldsymbol{v}^{(2)} + \cdots + t_n A\boldsymbol{v}^{(n)}. \end{aligned}$$

A-Orthogonality Condition (cont'd)

Proof We then have, for each k,

$$\begin{aligned}\langle A\boldsymbol{x}^{(n)} - \boldsymbol{b},\, \boldsymbol{v}^{(k)}\rangle &= \langle A\boldsymbol{x}^{(0)} - \boldsymbol{b},\, \boldsymbol{v}^{(k)}\rangle + t_1 \langle A\boldsymbol{v}^{(1)},\, \boldsymbol{v}^{(k)}\rangle + \cdots + t_n \langle A\boldsymbol{v}^{(n)},\, \boldsymbol{v}^{(k)}\rangle \\ &= \langle A\boldsymbol{x}^{(0)} - \boldsymbol{b},\, \boldsymbol{v}^{(k)}\rangle + t_1 \langle \boldsymbol{v}^{(1)},\, A\boldsymbol{v}^{(k)}\rangle + \cdots + t_n \langle \boldsymbol{v}^{(n)},\, A\boldsymbol{v}^{(k)}\rangle \\ &= \langle A\boldsymbol{x}^{(0)} - \boldsymbol{b},\, \boldsymbol{v}^{(k)}\rangle + t_k \langle \boldsymbol{v}^{(k)},\, A\boldsymbol{v}^{(k)}\rangle,\end{aligned}$$

where the last equality uses the A-orthogonality.

A-Orthogonality Condition (cont'd)

Proof Since $t_k \langle \boldsymbol{v}^{(k)}, A\boldsymbol{v}^{(k)} \rangle = \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - A\boldsymbol{x}^{(k-1)} \rangle$, we have

$$
\begin{aligned}
& t_k \langle \boldsymbol{v}^{(k)}, A\boldsymbol{v}^{(k)} \rangle \\
= & \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - A\boldsymbol{x}^{(0)} + A\boldsymbol{x}^{(0)} - A\boldsymbol{x}^{(1)} + \cdots - A\boldsymbol{x}^{(k-2)} + A\boldsymbol{x}^{(k-2)} - A\boldsymbol{x}^{(k-1)} \rangle \\
= & \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - A\boldsymbol{x}^{(0)} \rangle + \langle \boldsymbol{v}^{(k)}, A\boldsymbol{x}^{(0)} - A\boldsymbol{x}^{(1)} \rangle + \cdots + \langle \boldsymbol{v}^{(k)}, A\boldsymbol{x}^{(k-2)} - A\boldsymbol{x}^{(k-1)} \rangle \\
= & \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - A\boldsymbol{x}^{(0)} \rangle - t_1 \langle \boldsymbol{v}^{(k)}, A\boldsymbol{v}^{(1)} \rangle - \cdots - t_{k-1} \langle \boldsymbol{v}^{(k)}, A\boldsymbol{v}^{(k-1)} \rangle \\
= & \langle \boldsymbol{v}^{(k)}, \boldsymbol{b} - A\boldsymbol{x}^{(0)} \rangle
\end{aligned}
$$

where the third equality uses $A\boldsymbol{x}^{(i)} = A(\boldsymbol{x}^{(i-1)} + t_i \boldsymbol{v}^{(i)})$, and the last equality uses the A-orthogonality.

A-Orthogonality Condition (cont'd)

Proof Thus,

$$\begin{aligned}\langle A\boldsymbol{x}^{(n)} - \boldsymbol{b},\, \boldsymbol{v}^{(k)}\rangle &= \langle A\boldsymbol{x}^{(0)} - \boldsymbol{b},\, \boldsymbol{v}^{(k)}\rangle + t_k \langle \boldsymbol{v}^{(k)},\, A\boldsymbol{v}^{(k)}\rangle \\ &= \langle A\boldsymbol{x}^{(0)} - \boldsymbol{b},\, \boldsymbol{v}^{(k)}\rangle + \langle \boldsymbol{v}^{(k)},\, \boldsymbol{b} - A\boldsymbol{x}^{(0)}\rangle = 0.\end{aligned}$$

Hence the vector $A\boldsymbol{x}^{(n)} - \boldsymbol{b}$ is orthogonal to the A-orthogonal set of vectors $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$. Since the set $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ is linearly independent, there is a collection of constants $a_1, \ldots, a_n$ with

$$A\boldsymbol{x}^{(n)} - \boldsymbol{b} = \sum_{i=1}^{n} a_i \boldsymbol{v}^{(i)}.$$

Then,

$$\langle A\boldsymbol{x}^{(n)} - \boldsymbol{b},\, A\boldsymbol{x}^{(n)} - \boldsymbol{b}\rangle = \sum_{i=1}^{n} a_i \langle A\boldsymbol{x}^{(n)} - \boldsymbol{b},\, \boldsymbol{v}^{(i)}\rangle = 0,$$

which implies that $A\boldsymbol{x}^{(n)} - \boldsymbol{b} = \boldsymbol{0}$. $\square$

A-Orthogonality Condition (cont'd)

Ex. The linear system

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

has the exact solution $\boldsymbol{x}^* = (3, 4, -5)^t$. Show that the method in Theorem 18 with $\boldsymbol{x}^{(0)} = (0, 0, 0)^t$ and an A-orthogonal set of vectors

$$\boldsymbol{v}^{(1)} = (1, 0, 0)^t, \quad \boldsymbol{v}^{(2)} = (-3/4, 1, 0)^t, \quad \boldsymbol{v}^{(3)} = (-3/7, 4/7, 1)^t.$$

produces this exact solution after three iterations.

A-Orthogonality Condition (cont'd)

Sol. We then have

$$\boldsymbol{r}^{(1)} = \boldsymbol{b} - A\boldsymbol{x}^{(1)} = (0, 12, -24)^t, \quad t_1 = \frac{\langle \boldsymbol{v}^{(2)}, \boldsymbol{r}^{(1)} \rangle}{\langle \boldsymbol{v}^{(2)}, A\boldsymbol{v}^{(2)} \rangle} = \frac{48}{7},$$

$$\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + t_1 \boldsymbol{v}^{(2)} = (6, 0, 0)^t + \frac{48}{7}\left(-\frac{3}{4}, 1, 0\right)^t = \left(\frac{6}{7}, \frac{48}{7}, 0\right)^t.$$

Finally,

$$\boldsymbol{r}^{(2)} = \boldsymbol{b} - A\boldsymbol{x}^{(2)} = \left(0, 0, -\frac{120}{7}\right)^t, \quad t_2 = \frac{\langle \boldsymbol{v}^{(3)}, \boldsymbol{r}^{(2)} \rangle}{\langle \boldsymbol{v}^{(3)}, A\boldsymbol{v}^{(3)} \rangle} = -5,$$

$$\boldsymbol{x}^{(3)} = \boldsymbol{x}^{(2)} + t_2 \boldsymbol{v}^{(3)} = \left(\frac{6}{7}, \frac{48}{7}, 0\right)^t + (-5)\left(-\frac{3}{7}, \frac{4}{7}, 1\right)^t = (3, 4, -5)^t.$$

Conjugate Direction Method

- The use of A-orthogonal set $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\}$ of direction vectors gives what is called a **conjugate direction** method.

Theorem 19

The residual vectors $\boldsymbol{r}^{(k)}$, where $k = 1, 2, \ldots, n$, for a conjugate direction method, satisfy the equations

$$\langle \boldsymbol{r}^{(k)}, \boldsymbol{v}^{(j)} \rangle = 0, \quad \textit{for each } j = 1, 2, \ldots, k.$$

- The **conjugate gradient** method chooses the search directions $\{\boldsymbol{v}^{(k)}\}$ during the iterative process so that the residual vectors $\{\boldsymbol{r}^{(k)}\}$ are **mutually orthogonal**.

Q. How can we choose such search directions $\{\boldsymbol{v}^{(k)}\}$?

Conjugate Gradient Method

- Start with $\boldsymbol{x}^{(0)}$ and use the steepest descent direction $\boldsymbol{r}^{(0)} = \boldsymbol{b} - A\boldsymbol{x}^{(0)}$.

- Assume that the conjugate directions $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(k-1)}$ and the approximations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k-1)}$ have been computed with

$$\boldsymbol{x}^{(k-1)} = \boldsymbol{x}^{(k-2)} + t_{k-1}\boldsymbol{v}^{(k-1)},$$

  where

$$\langle \boldsymbol{v}^{(i)}, A\boldsymbol{v}^{(j)} \rangle = 0 \quad \text{and} \quad \langle \boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)} \rangle = 0, \quad \text{for } i \neq j.$$

- If $\boldsymbol{x}^{(k-1)}$ is the solution $A\boldsymbol{x} = \boldsymbol{b}$, we are done.
  Otherwise, $\boldsymbol{r}^{(k-1)} = \boldsymbol{b} - A\boldsymbol{x}^{(k-1)} \neq \boldsymbol{0}$ and Theorem 19 implies that

$$\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{v}^{(i)} \rangle = 0$$

  for each $i = 1, 2, \ldots, k-1$.

Conjugate Gradient Method: How to Choose $\boldsymbol{v}^{(k)}$

- We generate $\boldsymbol{v}^{(k)}$ using $\boldsymbol{r}^{(k-1)}$ as

$$\boldsymbol{v}^{(k)} = \boldsymbol{r}^{(k-1)} + s_{k-1}\boldsymbol{v}^{(k-1)},$$

  and choosing s_{k-1} so that

$$\langle \boldsymbol{v}^{(k-1)}, A\boldsymbol{v}^{(k)} \rangle = 0.$$

Q. Why?

Conjugate Gradient Method: How to Choose $\boldsymbol{v}^{(k)}$ (cont'd)

- Since $A\boldsymbol{v}^{(k)} = A\boldsymbol{r}^{(k-1)} + s_{k-1}A\boldsymbol{v}^{(k-1)}$ and

$$\langle \boldsymbol{v}^{(k-1)}, A\boldsymbol{v}^{(k)} \rangle = \langle \boldsymbol{v}^{(k-1)}, A\boldsymbol{r}^{(k-1)} \rangle + s_{k-1} \langle \boldsymbol{v}^{(k-1)}, A\boldsymbol{v}^{(k-1)} \rangle,$$

  we have $\langle \boldsymbol{v}^{(k-1)}, A\boldsymbol{v}^{(k)} \rangle = 0$ when

$$s_{k-1} = -\frac{\langle \boldsymbol{v}^{(k-1)}, A\boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k-1)}, A\boldsymbol{v}^{(k-1)} \rangle}$$

- We can also show that such s_{k-1} guarantees $\langle \boldsymbol{v}^{(k)}, A\boldsymbol{v}^{(i)} \rangle = 0$, for each $i = 1, 2, \ldots, k-2$. Thus $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(k)}\}$ is an A-**orthogonal** set.

- This implies the **mutual orthogonality** of $\{\boldsymbol{r}^{(k)}\}$, *e.g.*,

$$\langle \boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k-1)} \rangle = \langle \boldsymbol{r}^{(k)}, \boldsymbol{v}^{(k)} - s_{k-1}\boldsymbol{v}^{(k-1)} \rangle = 0.$$

Conjugate Gradient Method: How to Choose t_k

- Having chosen $\boldsymbol{v}^{(k)}$, we have

$$
\begin{aligned}
t_k = \frac{\langle \boldsymbol{v}^{(k)},\, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)},\, A\boldsymbol{v}^{(k)} \rangle} &= \frac{\langle \boldsymbol{r}^{(k-1)} + s_{k-1}\boldsymbol{v}^{(k-1)},\, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)},\, A\boldsymbol{v}^{(k)} \rangle} \\
&= \frac{\langle \boldsymbol{r}^{(k-1)},\, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)},\, A\boldsymbol{v}^{(k)} \rangle} + s_{k-1}\frac{\langle \boldsymbol{v}^{(k-1)},\, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)},\, A\boldsymbol{v}^{(k)} \rangle} \\
&= \frac{\langle \boldsymbol{r}^{(k-1)},\, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)},\, A\boldsymbol{v}^{(k)} \rangle},
\end{aligned}
$$

where the last equality uses Theorem 19, $\langle \boldsymbol{v}^{(k-1)},\, \boldsymbol{r}^{(k-1)} \rangle = 0$.

Conjugate Gradient Method (cont'd)

- In summary, we have

$$\boldsymbol{r}^{(0)} = \boldsymbol{b} - A\boldsymbol{x}^{(0)}, \quad \boldsymbol{v}^{(1)} = \boldsymbol{r}^{(0)},$$

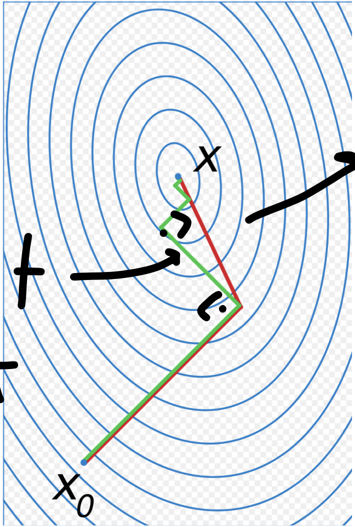
and, for $k = 1, 2, \ldots, n$,

$$t_k = \frac{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{v}^{(k)}, A\boldsymbol{v}^{(k)} \rangle}, \quad \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{v}^{(k)}, \quad \boldsymbol{r}^{(k)} = \boldsymbol{r}^{(k-1)} - t_k A\boldsymbol{v}^{(k)},$$

$$s_k = \frac{\langle \boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)} \rangle}{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}, \quad \boldsymbol{v}^{(k+1)} = \boldsymbol{r}^{(k)} + s_k \boldsymbol{v}^{(k)}.$$

Conjugate Gradient Method (cont'd)

steepest descent



conjugate gradient

Convergence Rate of Steepest Descent Method

- Steepest descent method (where $\nabla g(\boldsymbol{x}^{(k)}) = -2\boldsymbol{r}^{(k)}$)

$$t_k = \frac{\langle \boldsymbol{r}^{(k-1)}, \boldsymbol{r}^{(k-1)} \rangle}{\langle \boldsymbol{r}^{(k-1)}, A\boldsymbol{r}^{(k-1)} \rangle}, \quad \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \boldsymbol{r}^{(k-1)}$$

- Equivalent to the steepest descent method with an exact line search

$$\boldsymbol{x}^{(k)} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \{ g(\boldsymbol{x}) \ : \ \boldsymbol{x} = \boldsymbol{x}^{(k-1)} - t\nabla g(\boldsymbol{x}^{(k-1)}), \ t \in \mathbb{R} \}$$

- This has a rate

≥ 1

$$g(\boldsymbol{x}^{(k)}) - g(\boldsymbol{x}^*) \leq \left(\frac{K(A) - 1}{K(A) + 1} \right)^{2k} (g(\boldsymbol{x}^{(0)}) - g(\boldsymbol{x}^*)).$$

condit. number

Convergence Rate of Conjugate Gradient Method

- Conjugate gradient method is equivalent to

$$
\begin{aligned}
\boldsymbol{x}^{(k)} &= \arg\min_{\boldsymbol{x}\in\mathbb{R}^n}\{g(\boldsymbol{x}) \ : \ \boldsymbol{x} = \boldsymbol{x}^{(k-1)} + t\boldsymbol{v}^{(k)},\ t\in\mathbb{R}\} \\
&= \arg\min_{\boldsymbol{x}\in\mathbb{R}^n}\{g(\boldsymbol{x}) \ : \ \boldsymbol{x} \in \boldsymbol{x}^{(0)} + \operatorname{span}\{\nabla g(\boldsymbol{x}^{(0)}),\ldots,\nabla g(\boldsymbol{x}^{(k-1)})\}\} \\
&= \arg\min_{\boldsymbol{x}\in\mathbb{R}^n}\{g(\boldsymbol{x}) \ : \ \boldsymbol{x} = \boldsymbol{x}^{(k-1)} - \alpha\nabla g(\boldsymbol{x}^{(k-1)}) + \beta\underbrace{(\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k-2)})}_{\text{momentum}}, \alpha,\beta\in\mathbb{R}\}
\end{aligned}
$$

- Conjugate gradient method has a rate

$$
g(\boldsymbol{x}^{(k)}) - g(\boldsymbol{x}^*) \le 4\left(\frac{\sqrt{K(A)}-1}{\sqrt{K(A)}+1}\right)^{2k}(g(\boldsymbol{x}^{(0)}) - g(\boldsymbol{x}^*)).
$$

Of course, after n iterations, conjugate gradient method will find the exact solution.

Preconditioning

- If the matrix A is ill-conditioned, the conjugate gradient method is highly susceptible to rounding errors.
- In such case, one could apply conjugate gradient method to another positive definite matrix with a smaller condition number, rather than A.
- Note that it should be easy to find the solution of the original linear system from the solution of the another linear system.

Preconditioning (cont'd)

- Multiply on each side by a nonsingular matrix C^{-1} as

$$\tilde{A} = C^{-1}A(C^{-1})^t,$$

which preserves the positive definiteness, with a hope that $\tilde{A}$ has a smaller condition number compared to A.

- Then, consider the linear system

$$\tilde{A}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$$

where $\tilde{\boldsymbol{x}} = C^t\boldsymbol{x}$ and $\tilde{\boldsymbol{b}} = C^{-1}\boldsymbol{b}$. We have

$$\tilde{A}\tilde{\boldsymbol{x}} = (C^{-1}A(C^{-1})^t)(C^t\boldsymbol{x}) = C^{-1}A\boldsymbol{x} = C^{-1}\boldsymbol{b}.$$

Preconditioning (cont'd)

Ex. Find the eigenvalues and condition number of the matrix

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix}$$

and compare with those of the preconditioned matrix
$\tilde{A} = D^{-1/2} A D^{-1/2}$, where $D = \text{diag}\{[\ 0.2 \quad 4 \quad 60 \quad 8 \quad 700\ ]\}$.

ones on diagonal

Sol. Eigenvalues of A: $700.031, 60.0284, 0.0570757, 8.33845, 3.74533$

Eigenvalues of $\tilde{A}$: $1.88052, 0.156370, 0.852686, 1.10159, 1.00884$

$K(A) = 13961.7$ and $K(\tilde{A}) = 16.1155$

$\cdot\ \tilde{A} = I$

Preconditioning (cont'd)

Ex. Compare the Jacobi, Gauss-Seidel, SOR, (Preconditioned) Conjugate Gradient method on the linear system $A\boldsymbol{x} = \boldsymbol{b}$ with

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}.$$

Preconditioning (cont'd)

Sol. The solution is $\tilde{A} = $ Wellconditioned

$$x^* = (7.859713071, 0.4229264082, -0.07359223906, -0.5406430164, 0.01062616286)^t$$

$K(\tilde{A}) \approx 1$

Method	Number of Iterations	$\mathbf{x}^{(k)}$	$\lVert \mathbf{x}^* - \mathbf{x}^{(k)} \rVert_\infty$
Jacobi	49	$(7.86277141, 0.42320802, -0.07348669, -0.53975964, 0.01062847)^t$	0.00305834
Gauss-Seidel	15	$(7.83525748, 0.42257868, -0.07319124, -0.53753055, 0.01060903)^t$	0.02445559
SOR ($\omega = 1.25$)	7	$(7.85152706, 0.42277371, -0.07348303, -0.53978369, 0.01062286)^t$	0.00818607
Conjugate Gradient	5	$(7.85341523, 0.42298677, -0.07347963, -0.53987920, 0.008628916)^t$	0.00629785
Conjugate Gradient (Preconditioned)	4	$(7.85968827, 0.42288329, -0.07359878, -0.54063200, 0.01064344)^t$	0.00009312

$A \approx CC^t$

Q. How should we choose C^{-1} other than $C = \text{diag}\{[\ a_{11} \quad \cdots \quad a_{nn}\ ]\}$?

Preconditioning: How to choose C^{-1}

Corollary 3

The matrix A is positive definite if and only if A can be factored in the form LL^t, where L is lower triangular with nonzero diagonal entries.

- Consider a Cholesky factorization $A = LL^t$. Letting $C = L$ yields

$$\tilde{A} = C^{-1}A(C^{-1})^t = L^{-1}LL^t(L^{-1})^t = I.$$

$\hookrightarrow O(n^3)$ - instead Gaus

- In practice, we choose $C \approx L$ (perhaps by ignoring some small entries in A when computing Cholesky factorization) yielding

$$\tilde{A} \approx I.$$