Iterative Techniques in Matrix Algebra

Simon Fraser University – Surrey Campus
MACM 316 – Spring 2005
Instructor: Ha Le
Overview

- Norms of Vectors and Matrices
- Eigenvalues and Eigenvectors
- Iterative Techniques for Solving Linear Systems
We are interested in solving large linear systems $Ax = b$.

Suppose $A$ is sparse, i.e., it has a high percentage of zeros. We would like to take advantage of this sparse structure to reduce the amount of computational work required.

Gaussian elimination is often unable to take advantage of the sparse structure. For this reason, we consider iterative techniques.
To estimate how well a particular iterate approximates the true solution, we need some measurement of distance. This motivates the notion of a norm.

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

(i) $\| x \| \geq 0$ for all $x \in \mathbb{R}^n$;

(ii) $\| x \| = 0$ if and only if $x = 0$;

(iii) $\| \alpha x \| = |\alpha| \| x \|$ for all $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$;

(iv) $\| x + y \| \leq \| x \| + \| y \|$ for all $x, y \in \mathbb{R}^n$.

**Definition.** A *unit vector* with respect to the norm $\| \cdot \|$ is a vector $x$ that satisfies $\| x \| = 1$. 
Euclidean Norm and Max Norm

Definition. The $l_2$ or Euclidean norm of a vector $x \in \mathbb{R}^n$ is given by

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

Note that this represents the usual notion of distance.

Definition. The infinity or max norm of a vector $x \in \mathbb{R}^n$ is given by

$$
\| x \|_\infty = \max_{1 \leq i \leq n} |x_i|.
$$

Example. For $x = [-1, 1, -2]^T$,

$$
\| x \|_2 = \sqrt{(-1)^2 + (1)^2 + (-2)^2} = \sqrt{6},
\| x \|_\infty = \max \{| -1 |, |1|, |-2| \} = 2.
$$
It is straightforward to check that the max norm satisfies the definition of a norm. Checking that the $l_2$ norm satisfies

$$\| x + y \|_2 \leq \| x \|_2 + \| y \|_2$$

requires

**Cauchy-Schwarz Inequality.** For each $x, y \in \mathbb{R}^n$,

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

$$\| x + y \|_2^2 = \sum_{i=1}^{n} (x_i + y_i)^2$$

$$= \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$$

$$\leq \sum_{i=1}^{n} x_i^2 + 2 \| x \|_2 \| y \|_2 + \sum_{i=1}^{n} y_i^2$$

$$= (\| x \|_2 + \| y \|_2)^2.$$
**Distance between Two Vectors**

**Definition.** For $x, y \in \mathbb{R}^n$,

- the $l_2$ distance between $x$ and $y$ is defined by

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

and

- the $l_\infty$ distance between $x$ and $y$ is defined by

$$\| x - y \|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|.$$

**Example.** For $x_E = [1, 1, 1]^T$, $x_A = [1.2001, 0.99991, 0.92538]^T$, using five-digit rounding arithmetic:

$$\| x_E - x_A \|_\infty = \max \{|1 - 1.2001|, |1 - 0.99991|, |1 - 0.92538|\} = 0.2001,$$

$$\| x_E - x_A \|_2 = \left( (1 - 1.2001)^2 + (1 - 0.99991)^2 + (1 - 0.92538)^2 \right)^{1/2} = 0.21356.$$
Convergence of a Sequence of Vectors

**Definition.** Let \( \{x_n\}_{n=1}^{\infty} \) be an infinite sequence of real or complex numbers. The sequence \( \{x_n\}_{n=1}^{\infty} \) has the limit \( x \) (converges to \( x \)) if, for any \( \epsilon > 0 \), there exists a positive integer \( N(\epsilon) \) such that

\[
|x_n - x| < \epsilon \quad \text{for all } n > N(\epsilon).
\]

The notation \( \lim_{n \to \infty} x_n = x \), or \( x_n \to x \) as \( x \to \infty \), means that the sequence \( \{x_n\}_{n=1}^{\infty} \) converges to \( x \).

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

\[
\| x^{(k)} - x \| < \epsilon \quad \text{for all } k \geq N(\epsilon).
\]
Checking convergence in the max norm is facilitated by the following theorem:

**Theorem.** The sequence of vectors \( \{x^{(k)}\}_{k=1}^{\infty} \) converges to \( x \) in \( \mathbb{R}^n \) with respect to \( \| \cdot \|_\infty \) if and only if \( \lim_{k \to \infty} x^{(k)}_i = x_i \) for each \( i \).

**Proof.**

(\(\implies\)) \( \forall \varepsilon > 0, \exists N(\varepsilon) \text{ s.t. } \forall k \geq N(\varepsilon): \)

\[
\max_{1 \leq i \leq n} |x^{(k)}_i - x_i| = \| x^{(k)} - x \|_\infty < \varepsilon
\]

\[\implies |x^{(k)}_i - x_i| < \varepsilon \text{ for each } i\]

\[\implies \lim_{k \to \infty} x^{(k)}_i = x_i \text{ for each } i.\]

(\(\impliedby\)) \( \forall \varepsilon > 0, \exists N_i(\varepsilon) \text{ s.t. } |x^{(k)}_i - x_i| < \varepsilon, \forall k \geq N_i(\varepsilon), 1 \leq i \leq n. \) Let \( N(\varepsilon) = \max_i N_i(\varepsilon) \). If \( k \geq N(\varepsilon) \), then \( |x^{(k)}_i - x_i| < \varepsilon \) for each \( i \) and \( \max_{1 \leq i \leq n} |x^{(k)}_i - x_i| = \| x^{(k)} - x \|_\infty < \varepsilon. \)
Example. Prove that

$$x^{(k)} = \left( \frac{1}{k}, 1 + e^{1-k}, -\frac{2}{k^2} \right)$$

is convergent w.r.t. the infinity norm, and find the limit of the sequence.

$$\lim_{k \to \infty} \frac{1}{k} = 0, \quad \lim_{k \to \infty} 1 + e^{1-k} = 1, \quad \lim_{k \to \infty} -\frac{2}{k^2} = 0.$$ 

Hence, $x^{(k)}$ converges to $[0, 1, 0]^T$ w.r.t. the infinity norm.

- Convergence w.r.t. the $l_2$ norm is complicated to check. Instead, we will use the following theorem:
Theorem. For each \( x \in \mathbb{R}^n \), \( \| x \|_{\infty} \leq \| x \|_2 \leq \sqrt{n} \| x \|_{\infty} \).

Proof. Let \( x_j \) be such that s.t. \( \| x \|_{\infty} = \max_{1 \leq i \leq n} |x_i| = |x_j| \). Then

\[
\| x \|_{\infty}^2 = |x_j|^2 = x_j^2 \leq \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 = nx_j^2 = n \| x \|_{\infty}^2.
\]

Example. Show that \( x^{(k)} = (1/k, 1 + e^{1-k}, -2/k^2) \) converges to \( x = (0, 1, 0)^T \) w.r.t. the \( l_2 \) norm.

From the example on p.11, \( \lim_{k \to \infty} \| x^{(k)} - x \|_{\infty} = 0 \). Hence, \( 0 \leq \| x^{(k)} - x \|_2 \leq \sqrt{3} \| x^{(k)} - x \|_{\infty} = 0 \). This implies \( \{ x^{(k)} \} \) converges to \( x \) w.r.t. the \( l_2 \) norm.

- Indeed, it can be shown that all norms on \( \mathbb{R}^n \) are equivalent with respect to convergence, i.e.,

If \( \| \cdot \|_a \) and \( \| \cdot \|_b \) are any two norms on \( \mathbb{R}^n \), and \( \{ x^{(k)} \}_{k=1}^{\infty} \) has the limit \( x \) w.r.t. \( \| \cdot \|_a \) then \( \{ x^{(k)} \}_{k=1}^{\infty} \) also has the limit \( x \) w.r.t. \( \| \cdot \|_b \).
Matrix Norm

**Definition.** 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 $\alpha$:

(i) $\| A \| \geq 0$;

(ii) $\| A \| = 0$ if and only if $A = 0$;

(iii) $\| \alpha A \| = |\alpha| \| A \|$;

(iv) $\| A + B \| \leq \| A \| + \| B \|$;

(v) $\| AB \| \leq \| A \| \cdot \| B \|$.
**Definition.** A distance between $n \times n$ matrices $A$ and $B$ w.r.t. a matrix norm $\| \cdot \|$ is $\| A - B \|$.

**Theorem.** If $\| \cdot \|$ is a vector norm on $\mathbb{R}^n$, then $\| A \| = \max_{\| x \| = 1} \| Ax \|$ is a matrix norm.

This is called the *natural* or *induced* matrix norm associated with the vector norm.

The following result gives a bound on the value of $\| Ax \|$:

**Theorem.** For any vector $x \neq 0$, matrix $A$, and any natural norm $\| \cdot \|$, we have $\| Ax \| \leq \| A \| \cdot \| x \|$.

**Proof.** For any vector $z \neq 0$, $x = z / \| z \|$ is a unit vector. Hence,

$$\| A \| = \max_{\| x \| = 1} \| Ax \| = \max_{z \neq 0} \| A \left( \frac{z}{\| z \|} \right) \| = \max_{z \neq 0} \frac{\| Az \|}{\| z \|}.$$
Computing the infinity norm of a matrix is straightforward:

**Theorem.** If $A = (a_{i,j})$ is an $n \times n$ matrix, then

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

**Example.** Find the infinity norm of $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$.

$$\sum_{j=1}^{n} |a_{1,j}| = |2| + | -1 | + |0| = 3,$$

$$\sum_{j=1}^{n} |a_{2,j}| = | -1 | + |2| + | -1 | = 4,$$

$$\sum_{j=1}^{n} |a_{3,j}| = |0| + | -1 | + |2| = 3.$$

Hence, $\| A \|_{\infty} = \max \{3, 4, 3\} = 4.$
Eigenvalues and Eigenvectors

**Definition.** If $A$ is an $n \times n$ matrix, then the polynomial $p$ defined by $p(\lambda) = \det(A - \lambda I)$ is called the *characteristic polynomial* of $A$. It can be shown that $p$ is an $n$-th degree polynomial in $\lambda$.

**Example.**

Let

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

Hence, $p(\lambda) = \det(C) = -(\lambda - 3)^2 (\lambda - 1)$. 
Definition. If $p$ is the characteristic polynomial of an $n \times n$ matrix $A$, then the zeros of $p$ are called eigenvalues, or characteristic values of $A$.

If $\lambda$ is an eigenvalue of $A$ and $x \neq 0$ have the property that $(A - \lambda I)x = 0$, then $x$ is called an eigenvector, or characteristic vector of $A$ corresponding to the eigenvalue $\lambda$.

Example. For the matrix $A$ in the example on p.17, $p(\lambda) = -(\lambda - 3)^2 (\lambda - 1)$. Hence, the eigenvalues are $\lambda_1 = \lambda_2 = 3$, and $\lambda_3 = 1$.

To determine eigenvectors associated with the eigenvalue $\lambda = 3$, we solve the homogeneous linear system

\[
\begin{bmatrix}
2 - 3 & 1 & 0 \\
1 & 2 - 3 & 0 \\
0 & 0 & 3 - 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}.
\]
This implies that $x_1 = x_2$ and that $x_3$ is arbitrary. Two linearly independent choices for the eigenvectors associated with the double eigenvalue $\lambda = 3$ are

$$x_1 = [1, 1, 0]^T, \quad x_2 = [1, 1, 1]^T.$$  

The eigenvector associated with the eigenvalue $\lambda = 1$ must satisfy

$$\begin{bmatrix} 2 - 1 & 1 & 0 \\ 1 & 2 - 1 & 0 \\ 0 & 0 & 3 - 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. $$

This implies that we must have $x_1 = -x_2$ and that $x_3 = 0$. One choice for the eigenvector associated with the eigenvalue $\lambda = 1$ is

$$x_3 = [1, -1, 0]^T.$$
Notice that if \( x \) is an eigenvector associated with the eigenvalue \( \lambda \), then \( Ax = \lambda x \). So the matrix \( A \) takes the vector \( x \) into a scalar multiple of itself.

Geometrically, if \( \lambda \) is real, \( A \) has the effect of stretching (or shrinking) \( x \) by a factor of \( \lambda \).

In order to be able to compute the \( l_2 \) norm of a matrix, we need the following

**Definition.** The *spectral radius* \( \rho(A) \) of an \( n \times n \) matrix \( A \) is defined by \( \rho(A) = \max |\lambda| \) where \( \lambda \) is an eigenvalue of \( A \).

**Example.** For the matrix \( A \) in the example on p.17,

\[
\rho(A) = \max\{|3|, |3|, |1|\} = 3.
\]
Theorem. If $A$ is an $n \times n$ matrix then

(i) $\| A \|_2 = (\rho(A^T A))^{1/2}$;

(ii) $\rho(A) \leq \| A \|$ for any natural norm $\| \cdot \|$.

Proof. (ii) Let $\lambda$ be any eigenvalue of $A$ with the corresponding eigenvector $x$. W.l.o.g. (why?) we can assume that $\| x \| = 1$. Since $Ax = \lambda x$,

$$|\lambda| = |\lambda| \| x \| = \| \lambda x \| = \| Ax \| \leq \| A \| \| x \| = \| A \| .$$

Hence, $\rho(A) = \max |\lambda| \leq \| A \|$. 
Example. For the matrix $A$ in the example on p.17,

\[
A^T A = \begin{bmatrix}
2 & 1 & 0 \\
1 & 2 & 0 \\
0 & 0 & 3
\end{bmatrix}
\begin{bmatrix}
2 & 1 & 0 \\
1 & 2 & 0 \\
0 & 0 & 3
\end{bmatrix}
= \begin{bmatrix}
5 & 4 & 0 \\
4 & 5 & 0 \\
0 & 0 & 9
\end{bmatrix}.
\]

The characteristic polynomial $p(\lambda)$ of $A^T A$ is

\[- (\lambda - 1) (\lambda - 9)^2\]

which admits $\lambda = 1$ and $\lambda = 9$ as its roots. Hence.

\[\| A \|_2 = \sqrt{\rho(A^T A)} = \sqrt{\max\{1, 9\}} = 3.\]
When we use iterative matrix technique, we will need to know when powers of a matrix become small.

**Definition.** We call an $n \times n$ matrix $A$ *convergent* if \[ \lim_{k \to \infty} (A^k)_{i,j} = 0 \] for each $i, j$.

**Example.** For $A = \begin{bmatrix} 1/2 & 0 \\ 1/4 & 1/2 \end{bmatrix}$,

\[
A^2 = \begin{bmatrix} 1/4 & 0 \\ 1/4 & 1/4 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 1/8 & 0 \\ 3/16 & 1/8 \end{bmatrix}, \quad A^4 = \begin{bmatrix} 1/16 & 0 \\ 1/8 & 1/16 \end{bmatrix},
\]

and in general, $A^k = \begin{bmatrix} (1/2)^k & 0 \\ \frac{k}{2^{k+1}} (1/2)^k \end{bmatrix}$. Since $\lim_{k \to \infty} (1/2)^k = 0$, and $\lim_{k \to \infty} k/2^{(k+1)} = 0$, $A$ is a convergent matrix.
Note that the convergent matrix $A$ in the last example has $\rho(A) = 1/2 < 1$, since $1/2$ is the only eigenvalue of $A$. This generalizes:

**Theorem.** The following statements are equivalent.

(i) $A$ is a convergent matrix;
(ii) $\rho(A) < 1$;
(iii) $\lim_{n \to \infty} A^n x = 0$ for every $x$;
(iv) $\lim_{n \to \infty} \| A^n \| = 0$ for all natural norms.
Iterative Techniques

- In problems where the matrix $A$ is sparse, iterative techniques are often used to solve the system $Ax = b$ since they preserve the sparse structure of the matrix.
- Iterative techniques convert the system $Ax = b$ into an equivalent system of the form $x = Tx + c$ where $T \in \mathbb{R}^{n \times n}$ is a fixed matrix, and $c \in \mathbb{R}^n$ is a fixed vector.
- An initial vector $x^{(0)}$ is selected, and then a sequence of approximate solution vectors is generated:
  $$x^{(k)} = Tx^{(k-1)} + c.$$ 
- Iterative techniques are rarely used in very small systems. In these cases, iterative methods may be slower since they require several iterations to obtain the desired accuracy.
Iterative Techniques: General Approach

- Split the matrix $A$:

\[
Ax = b \\
(M + (A - M))x = b \\
Mx = b + (M - A)x \\
x = (I - M^{-1}A)x + M^{-1}b.
\]

Iteration becomes

\[
x^{(k+1)} = (I - M^{-1}A)x^{(k)} + M^{-1}b.
\]

Problem. How to choose $M$?
Jacobi Iterative Method

\[ M = D = \text{diag}(A) = \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{pmatrix} \]

To construct the matrix \( T \) and vector \( c \), let

\[ L = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ -a_{2,1} & 0 & \cdots & 0 & 0 \\ -a_{3,1} & -a_{3,2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -a_{n,1} & -a_{n,2} & \cdots & -a_{n,n-1} & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -a_{1,2} & -a_{1,3} & \cdots & -a_{1,n} \\ 0 & 0 & -a_{2,3} & \cdots & -a_{2,n} \\ 0 & 0 & 0 & \cdots & -a_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}. \]
Then \( A = D - L - U \).

\[
Ax = b
\]

\[
(D - L - U)x = b
\]

\[
Dx = (L + U)x + b
\]

\[
x = D^{-1}(L + U)x + D^{-1}b,
\]

which results in the iteration

\[
x^{(k+1)} = D^{-1}(L + U)x^{(k)} + D^{-1}b.
\]
Jacobi Iterative Method: an Example

Solve

\[
\begin{bmatrix}
10 & -1 & 2 & 0 \\
-1 & 11 & -1 & 3 \\
2 & -1 & 10 & -1 \\
0 & 3 & -1 & 8
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{bmatrix}
= 
\begin{bmatrix}
6 \\
25 \\
-11 \\
15
\end{bmatrix}
\]

by Jacobi’s method.

\[
D = 
\begin{bmatrix}
10 & 0 & 0 & 0 \\
0 & 11 & 0 & 0 \\
0 & 0 & 10 & 0 \\
0 & 0 & 0 & 8
\end{bmatrix}
\quad \Rightarrow \quad D^{-1} = 
\begin{bmatrix}
1/10 & 0 & 0 & 0 \\
0 & 1/11 & 0 & 0 \\
0 & 0 & 1/10 & 0 \\
0 & 0 & 0 & 1/8
\end{bmatrix}
\]
\[ L = \begin{bmatrix}
0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
-2 & 1 & 0 & 0 \\
0 & -3 & 1 & 0
\end{bmatrix}, \quad U = \begin{bmatrix}
0 & 1 & -2 & 0 \\
0 & 0 & 1 & -3 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{bmatrix}. \]

Hence,

\[ T = D^{-1}(L+U) = \begin{bmatrix}
0 & 1/10 & -1/5 & 0 \\
1/11 & 0 & 1/11 & -3/11 \\
-1/5 & 1/10 & 0 & 1/10 \\
0 & -3/8 & 1/8 & 0
\end{bmatrix}, \quad c = D^{-1}b = \begin{bmatrix}
3/5 \\
25/11 \\
-11/10 \\
15/8
\end{bmatrix}. \]
Take \( x^{(0)} = [0, 0, 0, 0]^T \). Then

\[
\begin{align*}
  x^{(1)} &= Tx^{(0)} + c = c = [0.6000, 2.2727, -1.1000, 1.8750]^T, \\
  x^{(2)} &= Tx^{(1)} + c = [1.0473, 1.7159, -0.8052, 0.8852]^T, \\
  & \vdots \quad \vdots \quad \vdots \\
  x^{(9)} &= Tx^{(8)} + c = [0.9997, 2.0004, -1.0004, 1.0006]^T, \\
  x^{(10)} &= Tx^{(9)} + c = [1.1001, 1.9998, -0.9998, 0.9998]^T.
\end{align*}
\]

The decision to stop after ten iterations was based on the criterion

\[
\frac{\| x^{(10)} - x^{(9)} \|_{\infty}}{\| x^{(10)} \|_{\infty}} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3}.
\]
Comments on Jacobi’s Method

\[ x^{(k+1)} = D^{-1}(L + U)x^{(k)} + D^{-1}b. \]

1. The algorithm requires that \( a_{i,i} \neq 0 \) for each \( i \). If one of the \( a_{i,i} = 0 \), and the system is nonsingular, then a reordering of the equations can be performed so that no \( a_{i,i} = 0 \);

2. To accelerate convergence, the equations should be arranged so that \( a_{i,i} \) is as large as possible;

3. A possible stopping criterion is to iterate until

\[ \frac{\| x^{(k)} - x^{(k-1)} \|}{\| x^{(k)} \|} < \epsilon. \]
Gauss-Seidel Iterative Method

- Write out Jacobi’s method \( x^{(k+1)} = D^{-1}(L + U) x^{(k)} + D^{-1}b \), we find that

\[
x_i^{(k+1)} = \frac{\sum_{j=1, j \neq i}^{n} (-a_{i,j} x_j^{(k)}) + b_i}{a_{i,i}} \quad \text{for } 1 \leq i \leq n.
\]

Notice that to compute \( x_i^{(k+1)} \), the components \( x_i^{(k)} \) are used. However, for \( i > 1 \), \( x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{i-1}^{(k+1)} \) have already been computed, and are likely better approximations to the actual solutions than \( x_1^{(k)}, x_2^{(k)}, x_{i-1}^{(k)} \). Hence, it seems reasonable to compute with these most recently computed values, i.e.,

\[
x_i^{(k+1)} = \frac{-\sum_{j=1}^{i-1} (a_{i,j} x_j^{(k+1)}) - \sum_{j=i+1}^{n} (a_{i,j} x_j^{(k)}) + b_i}{a_{i,i}}.
\]
Matrix Formulation. Set \( M = D - L \).

\[
Ax = b \\
(D - L - U)x = b \\
(D - L)x = Ux + b \\
x = (D - L)^{-1}Ux + (D - L)^{-1}b.
\]

Hence, iteration becomes

\[
x^{(k+1)} = (D - L)^{-1}U x^{(k)} + (D - L)^{-1}b.
\]

Notice that \((D - L)\) is lower triangular. It is invertible if and only if \(a_{i,i} \neq 0\).
Gauss-Seidel Method: an Example

For the linear system on p.28,

\[ T_g = \begin{bmatrix} 0 & 1/10 & -1/5 & 0 \\ 0 & \frac{1}{110} & \frac{4}{55} & -3/11 \\ 0 & -\frac{21}{1100} & \frac{13}{275} & \frac{4}{55} \\ 0 & -\frac{51}{8800} & -\frac{47}{2200} & \frac{49}{440} \end{bmatrix}, \quad c_g = \begin{bmatrix} 3/5 \\ \frac{128}{55} \\ -\frac{543}{550} \\ \frac{3867}{4400} \end{bmatrix}. \]

Take \( x^{(0)} = [0, 0, 0, 0]^T \). Then

\[
x^{(1)} = T_g x^{(0)} + c_g = c_g = [0.6000, 2.3272, -0.9873, 0.8789]^T,
\]

\[
\ldots \ldots \ldots \ldots \ldots
\]

\[
x^{(4)} = T_g x^{(3)} + c_g = [1.0009, 2.0003, -1.0003, 0.9999]^T,
\]

\[
x^{(5)} = T_g x^{(4)} + c_g = [1.1001, 2.0000, -1.0000, 1.0000]^T.
\]

Since

\[
\frac{\| x^{(5)} - x^{(4)} \|_\infty}{\| x^{(5)} \|_\infty} = \frac{0.0008}{2.0000} = 4 \times 10^{-4},
\]

\( x^{(5)} \) is accepted as a reasonable approximation to the solution.
Convergence of General Iteration Techniques

\[ x^{(k)} = T x^{(k-1)} + c \]

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

\[ (I - T)^{-1} = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j. \]

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

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

converges to the unique solution \( x = T x + c \) if and only if \( \rho(T) < 1 \).
Proof.

\((\iff)\) Assume that \(\rho(T) < 1\). Then

\[
x^{(k)} = T x^{(k-1)} + c
\]

\[
= T (T x^{(k-2)} + c) + c = T^2 x^{(k-2)} + (T + I)c
\]

\[
\ldots
\]

\[
= T^k x^{(0)} + (T^{k-1} + \cdots + T + I)c.
\]

Since \(\rho(T) < 1\), the matrix \(T\) is convergent, and by the theorem (iv) on p.23, \(\lim_{k \to \infty} T^k x^{(0)} = 0\). The Lemma on p.35 implies that

\[
\lim_{k \to \infty} x^{(k)} = \lim_{k \to \infty} T^k x^{(0)} + \left(\sum_{j=0}^{\infty} T^j\right) c = (I - T)^{-1} c.
\]

Hence, the sequence \(\{x^{(k)}\}_{k=0}^{\infty}\) converges to the vector \(x = (I - T)^{-1} c\) and \(x = Tx + c\).
(⇒) We show that for any \( z \in \mathbb{R}^n \), \( \lim_{k \to \infty} T^k z = 0 \). By the theorem on p.23, this is equivalent to \( \rho(T) < 1 \).

Let \( z \) be an arbitrary vector, and \( x \) be the unique solution to \( x = Tx + c \). Define

\[
x^{(k)} = \begin{cases} 
  x - z & \text{if } k = 0, \\
  Tx^{(k-1)} + c & \text{if } k \geq 1.
\end{cases}
\]

Then \( \{x^{(k)}\}_{k=0}^\infty \) converges to \( x \). Also,

\[
x - x^{(k)} = (Tx + c) - (Tx^{(k-1)} + c) = T(x - x^{(k-1)}) \]
\[
= T^2(x - x^{(k-2)}) = \cdots = T^k(x - x^{(0)}) = T^k z.
\]

Hence, \( \lim_{k \to \infty} T^k z = 0 \). Since \( z \in \mathbb{R}^n \) is arbitrary, \( T \) is a convergent matrix (p.23 (i)), and that \( \rho(T) < 1 \) (p.23 (ii)).
This allows us to derive some related results on the rates of convergence.

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

(i) \( \| x - x^{(k)} \| \leq \| T \|^k \| x^{(0)} - x \| \);

(ii) \( \| x - x^{(k)} \| \leq \frac{\| T \|^k}{1-\| T \|} \| x^{(1)} - x^{(0)} \| \).

Recall that \( \rho(A) \leq \| A \| \) for any natural norm (the theorem on p.20). In practice

\[
\| x - x^{(k)} \| \approx \rho(T)^k \| x^{(0)} - x \| .
\]

Hence, it is desirable to have \( \rho(T) \) as small as possible.
Some results for Jacobi and Gauss-Seidel methods.

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

**Remark.** No general results exist to tell which of the two methods will converge more quickly.
The following result applies in a variety of examples.

**Theorem.** (Stein-Rosenberg)

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

a. $0 \leq \rho(T_g) < \rho(T_j) < 1$;

b. $1 \leq \rho(T_j) < \rho(T_g)$;

c. $\rho(T_j) = \rho(T_g) = 0$;

d. $\rho(T_j) = \rho(T_g) = 1$.

**Note.** If one method converges, both do and Gauss-Seidel method converges faster. Otherwise, if one method diverges, both do. The divergence for Gauss-Seidel is more pronounced.

**Warning.** This result only holds when $a_{i,j} \leq 0$ for $i \neq j$, and $a_{i,i} > 0$. 
Successive Over Relaxation (SOR)

- Suppose $\tilde{x}^{(k+1)}$ is the iterate from Gauss-Seidel using $x^{(k)}$. The $(k+1)$-th iterate of SOR is defined by

$$x^{(k+1)} = w \tilde{x}^{(k+1)} + (1 - w)x^{(k)}$$

where $1 < w < 2$.

- Matrix notation.

$$x^{(k)} = T_w x^{(k-1)} + c_w,$$

where

$$T_w = (D - wL)^{-1}((1 - w)D + wU), \quad c_w = w(D - wL)^{-1}b.$$
Example. Solve

\[
\begin{bmatrix}
4 & 3 & 0 \\
3 & 4 & -1 \\
0 & -1 & 4
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
= \begin{bmatrix}
24 \\
30 \\
-24
\end{bmatrix}.
\]

\[T_w = (D - wL)^{-1}((1 - w)D + wU)\]

\[
= \begin{bmatrix}
1 - w & -3/4w & 0 \\
-3/16 w (4 - 4w) & \frac{9}{16} w^2 + 1 - w & 1/4 w \\
-\frac{3}{64} w^2 (4 - 4w) & \frac{9}{64} w^3 + 1/16 w (4 - 4w) & 1 + 1/16 w^2 - w
\end{bmatrix},
\]

\[
c_w = w(D - wL)^{-1}b = \begin{bmatrix}
6 w \\
-9/2 w^2 + 15/2 w \\
-\frac{9}{8} w^3 + \frac{15}{8} w^2 - 6 w
\end{bmatrix}.
\]
Take $x^{(0)} = [1, 1, 1]^T$. Then for $w = 1.25$,
\[
\begin{align*}
    x^{(1)} &= T_w x^{(0)} + c_w = [6.312500, 3.5195313, -6.6501465]^T, \\
    x^{(2)} &= T_w x^{(2)} + c_w = [2.6223145, 3.9585266, -4.6004238]^T, \\
    \vdots & \vdots \vdots \\
    x^{(6)} &= T_w x^{(5)} + c_w = [2.9963276, 4.0029250, -4.9982822]^T, \\
    x^{(7)} &= T_w x^{(6)} + c_w = [3.0000498, 4.0002586, -5.0003486]^T.
\end{align*}
\]
Note that the exact solution is $[3, 4, -5]^T$. 
It can be difficult to select $w$ optimally. Indeed, the answer to this question is not known for general $n \times n$ linear systems. However, we do have the following results:

**Theorem.** (Kahan)

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

**Theorem.** (Ostrowski-Reich)

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

**Theorem.** 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 $\rho(T_w) = w - 1$. 

44