Numerical Analysis: Solving Systems of Linear Equations

Mirko Navara
http://cmp.felk.cvut.cz/~navara/
Center for Machine Perception, Department of Cybernetics, FEE, CTU
Karlovo náměstí, building G, office 104a

http://math.feld.cvut.cz/nemecek/nummet.html

November 21, 2014

Task: Solve a system of n linear equations with n unknowns $x_1, x_2, \ldots, x_n$

$$\begin{aligned} a_{1,1}\,x_1 + a_{1,2}\,x_2 + \cdots + a_{1,n}\,x_n &= b_1 \\ a_{2,1}\,x_1 + a_{2,2}\,x_2 + \cdots + a_{2,n}\,x_n &= b_2 \\ &\vdots \\ a_{n,1}\,x_1 + a_{n,2}\,x_2 + \cdots + a_{n,n}\,x_n &= b_n \end{aligned}$$

Matrix form:

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

where $\mathbf{A} = (a_{i,j})_{i,j=1,\ldots,n}$ is a (regular) matrix of the system,
$\mathbf{b} = (b_1, b_2, \ldots, b_n)^\top$ vector of right-hand sides,
$\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ vector of unknowns.

Cramer's rule has a high computational complexity and numerical errors.

Matrix of the system:

- full, not very large,
- sparse, often very large (e.g., for a cubic spline).

$$\mathbf{x} = \mathbf{A}^{-1}\,\mathbf{b}$$

A small change of coefficients of the system or of the right-hand side can cause a large change of the solution.

Back substitution of the (imprecise) solution $\mathbf{x}_c$ gives the **residuum of the solution**:

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\,\mathbf{x}_c\,,$$

If matrix $\mathbf{A}^{-1}$ has large entries, the residuum $\mathbf{r}$ can be small, even if the vector $\mathbf{x}_c$ is much different from the exact solution $\overline{\mathbf{x}}$.

$$\begin{aligned} \mathbf{r} &= \mathbf{A}\,\overline{\mathbf{x}} - \mathbf{A}\,\mathbf{x}_c = \mathbf{A}\,(\overline{\mathbf{x}} - \mathbf{x}_c)\,, \\ \overline{\mathbf{x}} - \mathbf{x}_c &= \mathbf{A}^{-1}\,\mathbf{r}\,. \end{aligned}$$

If the entries of matrix $\mathbf{A}^{-1}$ are large, even a small change of an entry of vector $\mathbf{r}$ may cause a large difference $\overline{\mathbf{x}} - \mathbf{x}_c$.

A small residuum does not guarantee a small error of the solution!
Such a system is called **ill-conditioned**.

Example: System

$$\begin{aligned} 2\,x + \quad 6\,y &= 8 \\ 2\,x + 6.00001\,y &= 8.00001 \end{aligned}$$

has the solution $x = 1$, $y = 1$.

A minimal change of coefficients to the system

$$\begin{aligned} 2\,x + \qquad 6\,y &= 8 \\ 2\,x + 5.99999\,y &= 8.00002 \end{aligned}$$

changes the solution to $x = 10$, $y = -2$.

The inverse matrices of both systems have entries of order 10^5, which shows that they are ill-conditioned. The equations in the systems are "almost linearly dependent".

- imprecision of the coefficients of the system and of the right-hand sides,
- round-off errors,
- errors of the method—an infinite process is replaced by a finite number of iterative steps.

Give a (theoretically) precise solution in finitely many steps.

Using equivalent transformations (which do not change the solution), we transform the system to one with an upper triangular matrix, from which the solution can be computed easily by a backward substitution.

Extended matrix of the system has entries

$$\begin{aligned} a^{(0)}_{i,j} &= a_{i,j}\,, \qquad \text{pro } i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n; \\ a^{(0)}_{i,n+1} &= b_i\,, \qquad \text{pro } i = 1, 2, \ldots, n. \end{aligned}$$

System

$$\begin{aligned} a^{(0)}_{1,1}\,x_1 + a^{(0)}_{1,2}\,x_2 + \cdots + a^{(0)}_{1,n}\,x_n &= a^{(0)}_{1,n+1} \\ a^{(0)}_{2,1}\,x_1 + a^{(0)}_{2,2}\,x_2 + \cdots + a^{(0)}_{2,n}\,x_n &= a^{(0)}_{2,n+1} \\ &\vdots \\ a^{(0)}_{n,1}\,x_1 + a^{(0)}_{n,2}\,x_2 + \cdots + a^{(0)}_{n,n}\,x_n &= a^{(0)}_{n,n+1} \end{aligned}$$

can be transformed to

$$\begin{aligned} a^{(0)}_{1,1}\,x_1 + a^{(0)}_{1,2}\,x_2 + \cdots + a^{(0)}_{1,n}\,x_n &= a^{(0)}_{1,n+1} \\ a^{(1)}_{2,2}\,x_2 + \cdots + a^{(1)}_{2,n}\,x_n &= a^{(1)}_{2,n+1} \\ &\vdots \\ a^{(n-1)}_{n,n}\,x_n &= a^{(n-1)}_{n,n+1}\,, \end{aligned}$$

a backward substitution gives the vector of solutions.

For a zero entry on the diagonal, we exchange rows (or columns, but then we **reorder also the entries of the solution!**).

This is always possible if the original matrix of the system was regular.

Algorithm:

For $k = 1, 2, \ldots, n-1$, for $i = k+1, k+2, \ldots, n$ a $j = k+1, k+2, \ldots, n+1$

$$a^{(k)}_{i,j} = a^{(k-1)}_{i,j} - \frac{a^{(k-1)}_{i,k}}{a^{(k-1)}_{k,k}}\,a^{(k-1)}_{k,j}.$$

If the **forward step** gives a diagonal entry $a^{(i-1)}_{i,i} = 0$ (resp. $|a^{(i-1)}_{i,i}| < \varepsilon$), then the system is (resp. may be) singular.

Otherwise, we proceed to the **backword substitution**

$$x_i = \frac{1}{a^{(i-1)}_{i,i}}\left(a^{(i-1)}_{i,n+1} - \sum_{j=i+1}^{n} a^{(i-1)}_{i,j}\,x_j\right), \qquad \text{for } i = n, n-1, \ldots, 1.$$

If (the abolute value of) the diagonal entries is small, its small change causes a large change of result of division and round-off errors.

In each step, we choose a diagonal entry with the largest absolute value = the **pivot**.

GEM with pivoting

- **complete**: we choose from all $(n-k)^2$ of entries of the remaining square submatrix (computationally complex),
- **column**: we choose from the column and exchange rows,
- **row**: we choose from the row and exchange columns (**and the order of unknowns!**).

GEM may continue by elimination of entries above the diagonal.
Diagonal entries can be transformed to units.
Then the column of right-hand sides is the vector of solutions.

For a single use the complexity is higher, but it is effective for many tasks with the same right-hand sides (e.g., computation of an inverse matrix, where we solve a system of linear equations simultaneously for n right-hand sides originating from a unit matrix).

Let us denote

$$\mathbf{L}_1 = \begin{pmatrix} 1 & 0 & & \dots & 0 \\ -\frac{a_{2,1}}{a_{1,1}} & 1 & 0 & \dots & 0 \\ -\frac{a_{3,1}}{a_{1,1}} & 0 & 1 & & \\ \vdots & \vdots & & \ddots & \\ & & & 1 & 0 \\ -\frac{a_{n,1}}{a_{1,1}} & 0 & & \dots\ 0 & 1 \end{pmatrix}$$

and multiply $\mathbf{L}_1 \cdot \mathbf{A}$. We obtain an associated system from GEM with zeros under the diagonal in the first column. We continue:

$$\mathbf{A}_0 = \mathbf{A}\,, \qquad \mathbf{A}_{i+1} = \mathbf{L}_{i+1} \cdot \mathbf{A}_i \qquad \text{for } i = 0, 1, \dots, n-2,$$

where

$$\mathbf{L}_{i+1} = \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ 0 & 1 & 0 & \dots & & 0 \\ & 0 & 1 & & & \vdots \\ & & -\frac{a^{(i)}_{i+2,i+1}}{a^{(i)}_{i+1,i+1}} & \ddots & & \\ \vdots & \vdots & \vdots & & 1 & 0 \\ 0 & 0 & -\frac{a^{(i)}_{n,i+1}}{a^{(i)}_{i+1,i+1}} & \dots & 0 & 1 \end{pmatrix}.$$

After $n-1$ matix multiplications we obtain

$$\mathbf{L}_{n-1} \cdot \mathbf{L}_{n-2} \dots \mathbf{L}_2 \cdot \mathbf{L}_1 \cdot \mathbf{A} = \mathbf{U}\,,$$

where matrix $\mathbf{U}$ is upper triangular (=result of the forward step of GEM) and $\overline{\mathbf{L}} = \mathbf{L}_{n-1} \cdot \mathbf{L}_{n-2} \cdot \ldots \cdot \mathbf{L}_2 \cdot \mathbf{L}_1$ is a lower triangular matrix with units on the diagonal. The inverse matrix $\overline{\mathbf{L}}^{-1} = \mathbf{L}$ exists and it is also lower triangular with units on the diagonal.

$$\begin{aligned} \overline{\mathbf{L}} \cdot \mathbf{A} &= \mathbf{U}\,, \\ \mathbf{A} &= \overline{\mathbf{L}}^{-1} \cdot \mathbf{U} = \mathbf{L} \cdot \mathbf{U}\,. \end{aligned}$$

The original system $\mathbf{A}\,\mathbf{x} = \mathbf{L} \cdot \mathbf{U}\,\mathbf{x} = \mathbf{b}$ is replaced by two systems with triangular matrices

$$\begin{aligned} \mathbf{L}\,\mathbf{y} &= \mathbf{b}\,, \\ \mathbf{U}\,\mathbf{x} &= \mathbf{y}\,, \end{aligned}$$

(because $\mathbf{A}\,\mathbf{x} = \mathbf{L} \cdot \mathbf{U}\,\mathbf{x} = \mathbf{L}\,\mathbf{y} = \mathbf{b}$), which are solvable by backward substitution.

Expanding the product $\mathbf{L} \cdot \mathbf{U}$, we obtain

Algorithm:

For $r = 1, 2, \ldots, n$

$$\begin{aligned}
u_{i,r} &= a_{i,r} - \sum_{s=1}^{i-1} l_{i,s}\, a_{s,r} \qquad \text{for } i = 1, 2, \ldots, r, \\
l_{i,r} &= \frac{1}{u_{r,r}} \left(a_{i,r} - \sum_{s=1}^{r-1} l_{i,s}\, u_{s,r} \right) \qquad \text{for } i = r+1, r+2, \ldots, n, \\
y_i &= b_i - \sum_{s=1}^{i-1} l_{i,s}\, y_s \qquad \text{for } i = 1, 2, \ldots, n, \\
x_i &= \frac{1}{u_{i,i}} \left(y_i - \sum_{s=i+1}^{n} u_{i,s}\, x_s \right) \qquad \text{for } i = n, \ldots, 2, 1.
\end{aligned}$$

We need all entries $u_{r,r} \neq 0$ (during the computation) nonzero $\Rightarrow$ pivoting (partial); this also reduces round-off errors.

Comment: The whole algorithm can be performed in the same array. The diagonal entries are used for $u_{r,r}$ because the diagonal of $\mathbf{L}$ consists of units which are neither computed nor stored.

Comment: This method is particularly useful for a series of tasks which differ only by the right-hand sides. It can be used also for the computation of the inverse matrix $\mathbf{A}^{-1}$ (with a higher complexity).

$\mathbf{E}$ = unit matrix
$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{E} \Rightarrow \mathbf{A}$ multiplied by the jth column of $\mathbf{A}^{-1}$ equals the jth column of the unit matrix $\mathbf{E}$.
$\mathbf{x}_j = j$th column of $\mathbf{A}^{-1}$
$\mathbf{e}_j = j$th column of $\mathbf{E}$

$$\mathbf{A}\, \mathbf{x}_j = \mathbf{e}_j$$

We obtain a system of equations, where $\mathbf{x}_j$ is the vector of unknowns.

Columns of the inverse matrix $\mathbf{A}^{-1}$ are solutions to the system for various right-hand sides—columns of matrix $\mathbf{E}$.

GEM can be used for one system and several right-hand sides simultaneously; it suffices to "extend" the cycle to a system of type $(n \times 2\,n)$, i.e.,
$j = k+1, k+2, \ldots, 2\,n$ (see Gauss–Jordan reduction).

The use of LU-decomposition: $\mathbf{A} = \mathbf{L} \cdot \mathbf{U}$:

$$\mathbf{A}^{-1} = (\mathbf{L} \cdot \mathbf{U})^{-1} = \mathbf{U}^{-1} \cdot \mathbf{L}^{-1}\,.$$

The computation of $\mathbf{U}^{-1}$ and $\mathbf{L}^{-1}$ is easy; the inverse of a triangular matrix is again triangular.

The definition can be used only for very small orders.

GEM: the product of diagonal entries after the forward step:

$$\det \mathbf{A} = \pm a_{1,1}^{(0)}\, a_{2,2}^{(1)}\, a_{3,3}^{(2)} \ldots a_{n,n}^{(n-1)}.$$

ATTENTION! The exchange of rows or columns (during pivoting) **changes the sign** of the determinant. (It suffices to remember the parity of the number of changes.)

LU-decomposition $\mathbf{A} = \mathbf{L} \cdot \mathbf{U}$:

$$\det \mathbf{A} = \det (\mathbf{L} \cdot \mathbf{U}) = \det \mathbf{L} \cdot \det \mathbf{U} = u_{1,1}\, u_{2,2}\, u_{3,3} \ldots u_{n,n}\,.$$

(These matrices are triangular; moreover $\mathbf{L}$ has units on the diagonal.)
Again **the sign has to be corrected** according to the number of changes of rows or columns.

The exact solution $\overline{\mathbf{x}}$ can be expressed using an imprecise solution $\mathbf{x}_c$:

$$\overline{\mathbf{x}} = \mathbf{x}_c + \delta\,,$$

where $\delta = (\delta_1, \delta_2, \ldots, \delta_n)^T$ is the **vector of corrections**.

$$\begin{aligned}
\mathbf{A}\, \overline{\mathbf{x}} &= \mathbf{b} \\
\mathbf{A}\, (\mathbf{x}_c + \delta) &= \mathbf{b} \\
\mathbf{A}\, \mathbf{x}_c + \mathbf{A}\, \delta &= \mathbf{b} \\
\mathbf{A}\, \delta &= \mathbf{b} - \mathbf{A}\, \mathbf{x}_c = \mathbf{r}
\end{aligned}$$

The vector of corrections is the solution of the same system with the right-hand side replaced by $\mathbf{r}$.

The procedure can be repeated (LU-decomposition may be recommended) and obtain more precise solutions

$$\mathbf{x}_c^{(1)},\ \mathbf{x}_c^{(2)},\ \mathbf{x}_c^{(3)},\ \dots$$

We construct a sequence of vectors convergent to the exact solution.

$\mathbb{R}^n$ ... n-dimensional arithmetical vector space

$\mathbb{R}^{n,n}$ ... space of square matrices of order n

$\mathbf{o} \in \mathbb{R}^n$ ... null vector

$\mathbf{O} \in \mathbb{R}^{n,n}$ ... null matrix

Definition: A **vector norm** is a mapping $\|.\|_v \colon \mathbb{R}^n \to \mathbb{R}$ such that

- $\|\mathbf{x}\|_v \geq 0$, where $\|\mathbf{x}\|_v = 0 \Leftrightarrow \mathbf{x} = \mathbf{o}$, (positive definiteness)
- $\|c\,\mathbf{x}\|_v = |c|\,\|\mathbf{x}\|_v$, (homogeneity)
- $\|\mathbf{x}+\mathbf{y}\|_v \leq \|\mathbf{x}\|_v + \|\mathbf{y}\|_v$. (triangle inequality)

Example:

- $\|\mathbf{x}\|_r = \max\limits_{i=1,\dots,n} |x_i|$ **max-, sup-, Chebyshev**
- $\|\mathbf{x}\|_s = \sum\limits_{i=1}^{n} |x_i|$ **sum-, "Manhattan"**
- $\|\mathbf{x}\|_e = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$ **Euclidean**
- $\|\mathbf{x}\|_q = \left(\sum\limits_{i=1}^{n} |x_i|^q\right)^{\frac{1}{q}}, \quad q \geq 1$ **Hölder** (common generalization of the above)

Change of scale: If $\|.\|_v$ is a vector norm and $r > 0$, then $\|\mathbf{x}\|_u = r\,\|\mathbf{x}\|_v$ is also a vector norm.

Definition: **Matrix norm** is a mapping $\|.\|_M \colon \mathbb{R}^{n,n} \to \mathbb{R}$ such that

- $\|\mathbf{A}\|_M \geq 0$, where $\|\mathbf{A}\|_M = 0 \Leftrightarrow \mathbf{A} = \mathbf{O}$ (positive definiteness)
- $\|c\,\mathbf{A}\|_M = |c|\,\|\mathbf{A}\|_M$, (homogeneity)
- $\|\mathbf{A}+\mathbf{B}\|_M \leq \|\mathbf{A}\|_M + \|\mathbf{B}\|_M$, (triangle inequality)
- $\|\mathbf{A}\cdot\mathbf{B}\|_M \leq \|\mathbf{A}\|_M\,\|\mathbf{B}\|_M$. (consistency, submultiplicativity, Schwarz's inequality)

Example:

- $\|\mathbf{A}\|_R = \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{i,j}|$ **maximum absolute row sum norm**
- $\|\mathbf{A}\|_S = \max\limits_{j=1,\dots,n} \sum\limits_{i=1}^{n} |a_{i,j}|$ **maximum absolute column sum norm**
- $\|\mathbf{A}\|_E = \sqrt{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} a_{i,j}^2}$ **Euclidean, Frobenius**

What is **not** a matrix norm:

$$\|\mathbf{A}\|_M = \max_{i,j=1,\dots,n} |a_{ij}|$$

satisfies the first 3 conditions, but not the Schwarz's inequality:

$$\left\|\begin{pmatrix}1&1\\1&1\end{pmatrix}\cdot\begin{pmatrix}1&1\\1&1\end{pmatrix}\right\|_M = \left\|\begin{pmatrix}2&2\\2&2\end{pmatrix}\right\|_M = 2 \not\leq \left\|\begin{pmatrix}1&1\\1&1\end{pmatrix}\right\|_M \cdot \left\|\begin{pmatrix}1&1\\1&1\end{pmatrix}\right\|_M = 1\cdot 1 = 1\,.$$

Consequences of the Schwarz's inequality

The scale cannot be arbitrary: Put $\mathbf{B} := \mathbf{E}$ (unit matrix):

$$\|\mathbf{A}\|_M = \|\mathbf{A}\cdot\mathbf{E}\|_M \leq \|\mathbf{A}\|_M\,\|\mathbf{E}\|_M \qquad \Rightarrow \qquad \|\mathbf{E}\|_M \geq 1\,.$$

Convergence of an infinite product: Put $\mathbf{B} := \mathbf{B}^k$, $k \in \mathbb{N}$:

$$\|\mathbf{A}\cdot\mathbf{B}^k\|_M \leq \|\mathbf{A}\|_M\,\|\mathbf{B}\|_M^k \qquad \Rightarrow \qquad \|\mathbf{A}\cdot\mathbf{B}^k\|_M \to 0 \text{ for } \|\mathbf{B}\|_M < 1\,, k \to \infty\,.$$

Definition: Matrix norm $\|.\|_M$ is **consistent** with a vector norm $\|.\|_v$, if

$$\forall \mathbf{A} \in \mathbb{R}^{n,n}\ \forall \mathbf{x} \in \mathbb{R}^n : \|\mathbf{A}\,\mathbf{x}\|_v \leq \|\mathbf{A}\|_M\,\|\mathbf{x}\|_v\,.$$

(A change of the scale of a vector norm has no influence.)

Theorem: For each vector norm $\|.\|_v$, there exists at least one consistent matrix norm $\|.\|_M$, namely

$$\|\mathbf{A}\|_M = \sup_{\mathbf{x}\neq\mathbf{o}} \frac{\|\mathbf{A}\,\mathbf{x}\|_v}{\|\mathbf{x}\|_v} = \sup_{\|\mathbf{x}\|_v=1} \|\mathbf{A}\,\mathbf{x}\|_v\,.$$

The norm $\|.\|_M$ (**induced** by $\|.\|_v$) satisfies $\|\mathbf{E}\|_M = 1$.

Theorem: For each matrix norm $\|.\|_M$, there exists at least one consistent vector norm $\|.\|_v$, namely $\|\mathbf{x}\|_v = \|\mathbf{X}\|_M$, where

$$\mathbf{X} = \begin{pmatrix} x_1 & 0 & 0 & \dots & 0 \\ x_2 & 0 & 0 & \dots & 0 \\ \vdots & & & & \\ x_n & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Proof:

$$\|\mathbf{A}\,\mathbf{x}\|_v = \left\| \begin{pmatrix} \sum_{j=1}^n a_{1,j}\,x_j & 0 & \dots & 0 \\ \sum_{j=1}^n a_{2,j}\,x_j & 0 & \dots & 0 \\ & \vdots & & \\ \sum_{j=1}^n a_{n,j}\,x_j & 0 & \dots & 0 \end{pmatrix} \right\|_M = \|\mathbf{A}\cdot\mathbf{X}\|_M \le \|\mathbf{A}\|_M\,\|\mathbf{X}\|_M = \|\mathbf{A}\|_M\,\|\mathbf{x}\|_v\,.$$

Example: The Euclidean matrix norm is consistent with the Euclidean vector norm:

$$\left\| \begin{pmatrix} x_1 & 0 & 0 & \dots & 0 \\ x_2 & 0 & 0 & \dots & 0 \\ \vdots & & & & \\ x_n & 0 & 0 & \dots & 0 \end{pmatrix} \right\|_E^2 = \sum_{i=1}^n x_i^2 = \|\mathbf{x}\|_e^2\,.$$

Theorem: The following couples of norms (vector, matrix) are consistent:
$\|.\|_r, \|.\|_R$,
$\|.\|_s, \|.\|_S$,
$\|.\|_e, \|.\|_E$.

Moreover,
matrix norm $\|.\|_R$ is induced by vector norm $\|.\|_r$,
matrix norm $\|.\|_S$ is induced by vector norm $\|.\|_s$, but
matrix norm $\|.\|_E$ **is not** induced by vector norm $\|.\|_e$.

Definition: A sequence of vectors $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ **converges to a vector** $\mathbf{x} \in \mathbb{R}^n$ if

$$\lim_{k\to\infty} x_i^{(k)} = x_i \qquad \text{for } i = 1, 2, \dots, n.$$

Theorem: A sequence of vectors $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ converges to a vector $\mathbf{x} \in \mathbb{R}^n$ iff

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_v = 0\,,$$

where the norm can be any of the above vector norms.

(The convergence does not depend on the choice of the norm.)

Definition: A sequence of matrices $\mathbf{A}^{(0)} = \left(a_{i,j}^{(0)}\right)_{i,j=1}^n$, $\mathbf{A}^{(1)} = \left(a_{i,j}^{(1)}\right)_{i,j=1}^n$, $\mathbf{A}^{(2)} = \left(a_{i,j}^{(2)}\right)_{i,j=1}^n, \dots$ **converges to a matrix** $\mathbf{A} = \left(a_{i,j}\right)_{i,j=1}^n$ if

$$\lim_{k\to\infty} a_{i,j}^{(k)} = a_{i,j} \qquad \text{for all } i, j = 1, 2, \dots, n.$$

($\mathbf{A}^{(k)}$ ... kth element of the sequence)

Definition: Matrix $\mathbf{B}$ is **convergent** if the sequence of matrices $\mathbf{B}, \mathbf{B}^2, \mathbf{B}^3, \mathbf{B}^4, \ldots$ converges to the zero matrix. Otherwise, matrix $\mathbf{B}$ is called **divergent**.
($\mathbf{B}^k$ ... kth power of matrix $\mathbf{B}$)

Definition: Number $\lambda \in \mathbf{C}$ is an **eigenvalue** of a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ if there exists a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ (**eigenvector**) satisfying

$$\lambda\,\mathbf{x} = \mathbf{A}\,\mathbf{x}\,.$$

Theorem: Number λ is an eigenvalue of a matrix $\mathbf{A}$ iff $\det(\mathbf{A} - \lambda\,\mathbf{E}) = 0$.

Proof: If λ is an eigenvalue of a matrix $\mathbf{A}$, then $\exists \mathbf{x} \neq \mathbf{o} : \lambda\,\mathbf{x} = \mathbf{A}\,\mathbf{x}$, hence $\mathbf{A}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{o}$, $(\mathbf{A} - \lambda\,\mathbf{E})\,\mathbf{x} = \mathbf{o}$. Let $\det(\mathbf{A} - \lambda\,\mathbf{E}) \neq 0$, so matrix $\mathbf{A} - \lambda\,\mathbf{E}$ is regular. Homogenous system of equations with a regular matrix has only the trivial solution $\Rightarrow$ a contradiction.

Conversely, the assumption $\det(\mathbf{A} - \lambda\,\mathbf{E}) = 0$ implies that matrix $\mathbf{A} - \lambda\,\mathbf{E}$ is singular and $\exists \mathbf{x} \neq \mathbf{o}$: $(\mathbf{A} - \lambda\,\mathbf{E})\,\mathbf{x} = \mathbf{o}$. Hence $\mathbf{A}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{o}$, $\lambda\,\mathbf{x} = \mathbf{A}\,\mathbf{x}$.

For matrices of small orders:

- Find $\det(\mathbf{A} - \lambda\,\mathbf{E})$, which is a polynomial in variable λ of order n (**characteristic polynomial of matrix** $\mathbf{A}$).
- Eigenvalues λ_i are roots of the characteristic polynomial of matrix $\mathbf{A}$ (equation $\det(\mathbf{A} - \lambda\,\mathbf{E}) = 0$ is the **characteristic equation**).

Comment: "Eigenvalues" of a real matrix may be complex. (We may speak of complex eigenvalues in a complex vector space, not in a real one, where the multiplication by a complex number is undefined.

Definition: **Spectral radius** of a matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is the number

$$\varrho(\mathbf{A}) = \max_{i=1,2,\ldots,n} |\lambda_i|\,,$$

where $\lambda_i, i = 1, 2, \ldots, n$, are eigenvalues of matrix $\mathbf{A}$.

Theorem: Each matrix norm $\|\,.\,\|_M$ satisfies

$$\|\mathbf{A}\|_M \geq \varrho(\mathbf{A})\,.$$

Proof: Let $\varrho(\mathbf{A}) = |\lambda|$,
λ = the eigenvalue with the largest absolute value,
$\mathbf{y}$ = an eigenvector corresponding to λ, $\lambda\,\mathbf{y} = \mathbf{A}\,\mathbf{y}$.
Matrix norm $\|\,.\,\|_M$ is consistent with some vector norm $\|\,.\,\|_v$ satisfying

$$\|\mathbf{A}\|_M = \sup_{\mathbf{x}\neq\mathbf{o}} \frac{\|\mathbf{A}\,\mathbf{x}\|_v}{\|\mathbf{x}\|_v} \geq \frac{\|\mathbf{A}\,\mathbf{y}\|_v}{\|\mathbf{y}\|_v} = \frac{|\lambda|\,\|\mathbf{y}\|_v}{\|\mathbf{y}\|_v} = |\lambda| = \varrho(\mathbf{A})\,.$$

Theorem: Matrix $\mathbf{B}$ is convergent $\iff$ $\varrho(\mathbf{B}) < 1$.

Theorem: **Sufficient** condition for a matrix $\mathbf{B}$ to be convergent:

- $\|\mathbf{B}\|_M < 1$ for **some** matrix norm,

  i.e.,

- the linear mapping represented by matrix $\mathbf{B}$ is contractive (w.r.t. some vector norm, consistent with matrix norm $\|\,.\,\|_M$).

We look for a sequence of vectors $\mathbf{x}^{(k)} \in \mathbb{R}^n$ satisfying $\mathbf{x}^{(k)} \to \overline{\mathbf{x}}$, i.e., $\|\mathbf{x}^{(k)} - \overline{\mathbf{x}}\| \to 0$.
We use a recurrent formula

$$\mathbf{x}^{(k+1)} = F_k(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)}, \ldots, \mathbf{x}^{(k-m)})\,.$$

We restrict attention to **linear single-point stationary matrix iterative methods**, i.e., F_k is independent of k, depends only (linearly) on $\mathbf{x}^{(k)}$,

$$\mathbf{x}^{(k+1)} = \mathbf{B}\,\mathbf{x}^{(k)} + \mathbf{c}\,.$$

E.g.,

$$\begin{aligned} \mathbf{b} &= \mathbf{A}\,\mathbf{x}\,, \\ \mathbf{x}+\mathbf{b} &= \mathbf{x}+\mathbf{A}\,\mathbf{x} = (\mathbf{E}+\mathbf{A})\,\mathbf{x}\,, \\ \mathbf{x} &= (\mathbf{E}+\mathbf{A})\,\mathbf{x}-\mathbf{b}\,, \\ \mathbf{x}^{(k+1)} &= \underbrace{(\mathbf{E}+\mathbf{A})}_{\mathbf{B}}\,\mathbf{x}^{(k)}\,\underbrace{-\mathbf{b}}_{\mathbf{c}}\,. \end{aligned}$$

Vector of errors:

$$\begin{aligned} \varepsilon^{(k)} &= \mathbf{x}^{(k)}-\overline{\mathbf{x}} \qquad (\text{we assume } \neq \mathbf{o}) \\ \mathbf{x}^{(k)} &= \mathbf{B}\,\mathbf{x}^{(k-1)}+\mathbf{c} \\ \overline{\mathbf{x}} &= \mathbf{B}\,\overline{\mathbf{x}}+\mathbf{c} \\ \mathbf{x}^{(k)}-\overline{\mathbf{x}} &= \mathbf{B}\,(\mathbf{x}^{(k-1)}-\overline{\mathbf{x}}) = \mathbf{B}^2\,(\mathbf{x}^{(k-2)}-\overline{\mathbf{x}}) = \ldots = \mathbf{B}^k\,(\mathbf{x}^{(0)}-\overline{\mathbf{x}}) \\ \varepsilon^{(k)} &= \mathbf{B}\,\varepsilon^{(k-1)} = \ldots = \mathbf{B}^k\,\varepsilon^{(0)} \end{aligned}$$

$$\lim_{k\to\infty}\varepsilon^{(k)} = \mathbf{o} \qquad \Leftrightarrow \qquad \lim_{k\to\infty}\mathbf{B}^k = \mathbf{O}\,.$$

Theorem: Necessary and sufficient condition for convergence of the iterative method of the form $\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)}+\mathbf{c}$ is $\varrho(\mathbf{B}) < 1$.

Comment: The convergence is independent of the right-hand side and the initial estimate.

Theorem: Sufficient condition for convergence of the iterative method of the form $\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)}+\mathbf{c}$ is $\|\mathbf{B}\|_M < 1$ (for **some** matrix norm).

We express matrix $\mathbf{A}$ in the form

$$\mathbf{A} = \mathbf{D}+\mathbf{L}+\mathbf{U}\,,$$

where $\mathbf{D}$ is diagonal, $\mathbf{L}$ is strictly lower triangular and $\mathbf{U}$ is strictly upper triangular.

$$\begin{aligned} \mathbf{A}\,\mathbf{x} = (\mathbf{D}+\mathbf{L}+\mathbf{U})\,\mathbf{x} = \mathbf{D}\,\mathbf{x}+(\mathbf{L}+\mathbf{U})\,\mathbf{x} &= \mathbf{b}\,, \\ \mathbf{D}\,\mathbf{x} &= -(\mathbf{L}+\mathbf{U})\,\mathbf{x}+\mathbf{b}\,, \\ \mathbf{x} &= \mathbf{D}^{-1}\,(-(\mathbf{L}+\mathbf{U})\,\mathbf{x}+\mathbf{b})\,, \end{aligned}$$

we choose

$$\mathbf{x}^{(k+1)} = \underbrace{-\mathbf{D}^{-1}\,(\mathbf{L}+\mathbf{U})}_{\mathbf{B}_{\mathrm{JIM}}}\,\mathbf{x}^{(k)}+\underbrace{\mathbf{D}^{-1}\,\mathbf{b}}_{\mathbf{c}_{\mathrm{JIM}}}\,.$$

We assume that the main diagonal of matrix $\mathbf{A}$ has no zero entries (this can be achieved by an exchange of rows or columns).

Moreover, we want the diagonal entries to be "large".

Componentwise:

$$\begin{aligned} x_1 &= -\frac{1}{a_{1,1}}\,(a_{1,2}\,x_2+a_{1,3}\,x_3+\ldots+a_{1,n}\,x_n)+\frac{b_1}{a_{1,1}} \\ x_2 &= -\frac{1}{a_{2,2}}\,(a_{2,1}\,x_1+a_{2,3}\,x_3+\ldots+a_{2,n}\,x_n)+\frac{b_2}{a_{2,2}} \\ &\vdots \\ x_n &= -\frac{1}{a_{n,n}}\,(a_{n,1}\,x_1+a_{n,2}\,x_2+\ldots+a_{n,n-1}\,x_{n-1})+\frac{b_n}{a_{n,n}} \end{aligned}$$

$$x_i^{(k+1)} = -\frac{1}{a_{i,i}}\left(\sum_{j=1}^{i-1}a_{i,j}\,x_j^{(k)}+\sum_{j=i+1}^{n}a_{i,j}\,x_j^{(k)}\right)+\frac{b_i}{a_{i,i}}\,, \qquad i=1,2,\ldots,n.$$

Ending condition: $\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|_v < \varepsilon$.

Each already computed entry of vector $\mathbf{x}^{(k+1)}$ is immediately used in further computations.

$$x_i^{(k+1)} = -\frac{1}{a_{i,i}}\left(\sum_{j=1}^{i-1}a_{i,j}\,x_j^{(k+1)}+\sum_{j=i+1}^{n}a_{i,j}\,x_j^{(k)}\right)+\frac{b_i}{a_{i,i}}\,, \qquad i=1,2,\ldots,n.$$

A single vector $\mathbf{x}$ suffices for saving the results.
Matrix form:

$$\begin{aligned}(\mathbf{D}+\mathbf{L})\,\mathbf{x} &= -\mathbf{U}\,\mathbf{x}+\mathbf{b}\,,\\ \mathbf{x} &= -(\mathbf{D}+\mathbf{L})^{-1}\,(\mathbf{U}\,\mathbf{x}-\mathbf{b})\,,\\ \mathbf{x}^{(k+1)} &= \underbrace{-(\mathbf{D}+\mathbf{L})^{-1}\,\mathbf{U}}_{\mathbf{B}_{\mathrm{GSM}}}\,\mathbf{x}^{(k)}+\underbrace{(\mathbf{D}+\mathbf{L})^{-1}\,\mathbf{b}}_{\mathbf{c}_{\mathrm{GSM}}}\,.\end{aligned}$$

Theorem: JIM (resp. GSM) converges for any initial estimate iff $\varrho(\mathbf{B}_{\mathrm{JIM}})<1$ (resp. $\varrho(\mathbf{B}_{\mathrm{GSM}})<1$).

If $\varrho(\mathbf{B}_{\mathrm{JIM}})=\varrho(\mathbf{B}_{\mathrm{GSM}})=0$, we obtain (theoretically) an exact solution within finitely many steps.
It may happen that only one (or none) of the two methods converges.
Problem: computation of eigenvalues of a matrix.
Theorem: Let $\|\mathbf{B}_{\mathrm{JIM}}\|_M<1$ (resp. $\|\mathbf{B}_{\mathrm{GSM}}\|_M<1$). Then JIM (resp. GSM) converges from any initial estimate and the following error estimate holds:

$$\|\mathbf{x}^{(k)}-\overline{\mathbf{x}}\|_v \le \frac{\|\mathbf{B}\|_M}{1-\|\mathbf{B}\|_M}\,\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|_v\,,$$

where $\mathbf{B}$ is the corresponding iteration matrix (the norms of matrices and vectors must be **consistent**).

Definition: Matrix $\mathbf{A}\in\mathbb{R}^{n,n}$ is **strictly diagonally dominant**, if

$$|a_{i,i}| > \sum_{j=1,j\neq i}^{n}|a_{i,j}| \quad \text{for all } i=1,2,\ldots,n.$$

Comment: For the transposed matrix, we obtain another condition (as useful as this one).

Theorem: Let matrix $\mathbf{A}\in\mathbb{R}^{n,n}$ be strictly diagonally dominant. Then JIM and GSM converge for any initial estimate.

Comment: E.g., the matrix of the system of equations for coefficients of a cubic spline is strictly diagonally dominant. Moreover, it is sparse, so the complexity of one iteration is proportional to n.

Definition: Matrix $\mathbf{A}\in\mathbb{R}^{n,n}$ is **positive definite** if each nonzero vector $\mathbf{x}\in\mathbb{R}^n$ satisfies

$$\mathbf{x}^{\top}\mathbf{A}\,\mathbf{x}>0\,.$$

Theorem: Let matrix $\mathbf{A}\in R^{n,n}$ be symmetric and positive definite. Then GSM converges.

Comment: E.g., the matrix of the system of normal equations in LSM is symmetric and positive definite.

Faster convergence can be achieved by any modification which reduces the spectral radius of matrix $\mathbf{B}$.

$$\begin{aligned}\mathbf{D}\,\mathbf{x} &= \mathbf{D}\,\mathbf{x}+\omega\,\overbrace{(-\mathbf{A}\,\mathbf{x}+\mathbf{b})}^{\mathbf{o}} =\\ &= \mathbf{D}\,\mathbf{x}+\omega\,((-\mathbf{L}-\mathbf{D}-\mathbf{U})\,\mathbf{x}+\mathbf{b})\,,\\ \mathbf{D}\,\mathbf{x}+\omega\,\mathbf{L}\,\mathbf{x} &= (1-\omega)\,\mathbf{D}\,\mathbf{x}-\omega\,\mathbf{U}\,\mathbf{x}+\omega\,\mathbf{b}\,, \qquad (\text{GSM: } \omega:=1)\\ (\mathbf{D}+\omega\,\mathbf{L})\,\mathbf{x} &= [(1-\omega)\,\mathbf{D}-\omega\,\mathbf{U}]\,\mathbf{x}+\omega\,\mathbf{b}\,,\\ \mathbf{x} &= (\mathbf{D}+\omega\,\mathbf{L})^{-1}\,([(1-\omega)\,\mathbf{D}-\omega\,\mathbf{U}]\,\mathbf{x}+\omega\,\mathbf{b}) =\\ &= (\mathbf{D}+\omega\,\mathbf{L})^{-1}\,[(1-\omega)\,\mathbf{D}-\omega\,\mathbf{U}]\,\mathbf{x}+\omega\,(\mathbf{D}+\omega\,\mathbf{L})^{-1}\,\mathbf{b}\,,\\ \mathbf{x}^{(k+1)} &= \underbrace{(\mathbf{D}+\omega\,\mathbf{L})^{-1}\,[(1-\omega)\,\mathbf{D}-\omega\,\mathbf{U}]}_{\mathbf{B}_\omega}\,\mathbf{x}^{(k)}+\underbrace{\omega\,(\mathbf{D}+\omega\,\mathbf{L})^{-1}\,\mathbf{b}}_{\mathbf{c}_\omega}\,,\end{aligned}$$

where ω is **relaxation factor**. (For $\omega=1$, we obtain GSM.) Componentwise:

$$x_i^{(k+1)}=\frac{\omega}{a_{i,i}}\Big(-\sum_{j=1}^{i-1}a_{i,j}\,x_i^{(k+1)}-\sum_{j=i+1}^{n}a_{i,j}\,x_i^{(k)}+b_j\Big)+(1-\omega)\,x_i^{(k)}$$

Theorem: SOR converges for any initial estimate iff $\varrho(\mathbf{B}_\omega) < 1$.

Theorem: (Ostrowski) Let $\mathbf{A}$ be a symmetric matrix with positive diagonal entries. Then $\varrho(\mathbf{B}_\omega) < 1$ iff matrix $\mathbf{A}$ is positive definite and $0 < \omega < 2$.

Comment: It is often better to overestimate the relaxation factor.

Comment: The complexity of solving a system of n linear equations with n unknowns and with a single vector of right-hand sides using GEM or other direct methods is proportional to n^3. If the matrix of the system is triangular (in particular, during the backward part of GEM or for a *known* LU-decomposition), the complexity proportional to n^2. Also the multiplication of the vector of right-hand sides by a *known* inverse matrix has a complexity is proportional to n^2. If the matrix of the system is diagonal, the complexity is proportional to n.

For iterative methods in general, the complexity of a single iteration is proportional to n^2. The number of iterations depends on the required precision and the speed of convergence (and this on the spectral radius of the iteration matrix).

Task:

$$\mathbf{A}\,\mathbf{x} = \mathbf{B}$$

Comment: Here matrix $\mathbf{B}$ representes the right-hand sides of the system of equations, not the iteration matrix.

- $\mathbf{A}$ is dense (has many nonzero entries), $\mathbf{B}$ has a few columns: GEM or its modifications
- $\mathbf{A}$ is dense, $\mathbf{B}$ has many columns: computation of $\mathbf{A}^{-1}$ and matrix multiplication or LU-decomposition
- $\mathbf{A}$ is sparse, has a few nonzero entries distributed irregularly: GEM with pivoting
- $\mathbf{A}$ is sparse with nonzero entries distributed regularly: iterative methods JIM, GSM, SOR

References

[Navara, Němeček] Navara, M., Němeček, A.: *Numerical Methods* (in Czech). 2nd edition, CTU, Prague, 2005.

[Knuth] Knuth, D.E.: *Fundamental Algorithms.* Vol. 1 of *The Art of Computer Programming,* 3rd ed., Addison-Wesley, Reading, MA, 1997.

[KJD] Kubíček, M., Janovská, D., Dubcová, M.: *Numerical Methods and Algorithms.* Institute of Chemical Technology, Prague, 2005. `http://old.vscht.cz/mat/NM-Ang/NM-Ang.pdf`

[Num. Recipes] Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: *Numerical Recipes (The Art of Scientific Computing).* 2nd edition, Cambridge University Press, Cambridge, 1992. `http://www.nrbook.com/a/bookcpdf.php`

[Stoer, Bulirsch] Stoer, J., Bulirsch, R.: *Introduction to Numerical Analysis.* Springer Verlag, New York, 2002.

[Handbook Lin. Alg.] Hogben, L. (ed.): *Handbook of Linear Algebra.* Chapman & Hall/CRC, Boca Raton/London/New York, 2007.