Numerical Methods for Linear Systems

Jan Mandel

October 17, 2023

Contents

1 Jacobi method

Solving system of n linear equations for n unknowns.

Idea: one iteration is computes unknown i from equation i for all i at the same time - using **old** values of x

Example: The system of linear equations

$$
\begin{aligned}
4x_1 - 3x_2 &= -1 \\
2x_1 + 5x_2 &= 19
\end{aligned}
$$

The Jacobi iterative method is: starting from given $x_1^{(0)}$, $x_2^{(0)}$, compute for $k = 0, 1, \ldots$

$$
\begin{aligned}
x_1^{(k+1)} &= \frac{1}{4}(-1 + 3x_2^{(k)}) \\
x_2^{(k+1)} &= \frac{1}{5}\left(19 - 2x_1^{(k)}\right)
\end{aligned}
$$

For a system of n equations:

$$\sum_{j=1}^{n} a_{1j} x_j = b_i, \;\; i = 1, \ldots, n$$

$$
\begin{aligned}
a_{ii}x_i + \sum_{j=1}^{i-1} a_{1j}x_j + \sum_{j=i+1}^{n} a_{1j}x_j &= b_i, \;\; i = 1, \ldots, n \\
a_{ii}x_i^{(k+1)} + \sum_{j=1}^{i-1} a_{1j}x_j^{(k)} + \sum_{j=i+1}^{n} a_{1j}x_j^{(k)} &= b_i, \;\; i = 1, \ldots, n \\
x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{1j}x_j^{(k)} - \sum_{j=i+1}^{n} a_{1j}x_j^{(k)} \right)&, \;\; i = 1, \ldots, n
\end{aligned}
$$

1.1 Matrix form

Write the equations above as

$$\underbrace{\begin{bmatrix} 4 & -3 \\ 2 & 5 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} -1 \\ 19 \end{bmatrix}}_{b}$$

so

$$Ax = b$$

To write the Jacobi iterative method in matrix form, write

$$A = D + L + U$$

where D is diagonal matrix, L is strictly lower triangular, and U is strictly upper triangular. Then $Ax = b$ becomes

$$
\begin{aligned}
(D + L + U)\,x &= b \\
Dx + Lx + Ux &= b
\end{aligned}
$$

and to derive a fixed point form (hence iterations), compute x from the term Dx:

$$
\begin{aligned}
Dx &= b - Lx + Ux \\
x &= D^{-1}\left(b - (L+U)x\right) \\
x^{(k+1)} &= D^{-1}\left(b - (L+U)x^{(k)}\right) \qquad (1)
\end{aligned}
$$

In the example here,

$$
A = \begin{bmatrix} 4 & -3 \\ 2 & 5 \end{bmatrix}, \quad D = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & -3 \\ 0 & 0 \end{bmatrix}
$$

so the iterations (1) become

$$
\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix} \left(\begin{bmatrix} -1 \\ 19 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} \right)
$$

which is again

$$
\begin{aligned}
x_1^{(k+1)} &= \frac{1}{4}(-1 + 3x_2^{(k)}) \\
x_2^{(k+1)} &= \frac{1}{5}\left(19 - 2x_1^{(k)}\right)
\end{aligned}
$$

2 Convergence

Rewrite equation (1)

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

using $A = D + L + U$ as (add and substract $Dx^{(k)}$ iside the bracket)

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

which is, using $D^{-1}Dx^{(k)} = x^{(k)}$, the same as

$$
x^{(k+1)} = x^{(k)} + D^{-1}\left(b - Ax^{(k)}\right)
$$

Now we realize that instead of D we could have used any other invertible matrix M (the same size as A) and get the general iterative method

$$
x^{(k+1)} = x^{(k)} + M^{-1}\left(b - Ax^{(k)}\right) \qquad (2)
$$

When x^* is the exact solution, i.e., $Ax^* = b$, we get

$$
x^* = x^* + M^{-1}\left(b - Ax^*\right) \qquad (3)
$$

since $Ax^* - b = 0$, thus (3) is simply $x^* = x^*$. Now subtract (2) and (3) to get by simple algebra

$$\begin{aligned}
x^{(k+1)} - x^* &= x^{(k)} + M^{-1}\left(b - Ax^{(k)}\right) - x^* - M^{-1}\left(b - Ax^*\right) \\
x^{(k+1)} - x^* &= x^{(k)} + M^{-1}b - M^{-1}Ax^{(k)} - x^* - M^{-1}b + M^{-1}Ax^* \\
x^{(k+1)} - x^* &= \left(x^{(k)} - x^*\right) - M^{-1}A\left(x^{(k)} - x^*\right)
\end{aligned}$$

and, finally, the **error transformation equation**

$$x^{(k+1)} - x^* = \left(I - M^{-1}A\right)\left(x^{(k)} - x^*\right). \tag{4}$$

Suppose that $\|\cdot\|$ denotes a vector norm as well as a compatible matrix norm, so that we have the standard propety $\|Tx\| \le \|T\|\,\|x\|$. Then, from (4), we get

$$\left\|x^{(k+1)} - x^*\right\| \le \left\|I - M^{-1}A\right\| \left\|x^{(k)} - x^*\right\|$$

and, by induction over k and using the property $\|TU\| \le \|T\|\,\|U\|$ of a matrix norm,

$$\left\|x^{(k)} - x^*\right\| \le \left\|\left(I - M^{-1}A\right)^k\right\| \left\|x^{(0)} - x^*\right\| \le \left\|\left(I - M^{-1}A\right)\right\|^k \left\|x^{(0)} - x^*\right\|. \tag{5}$$

Take-home conclusions

1. If $\left\|I - M^{-1}A\right\| < 1$ then the iterations converge
2. If $\left|I - M^{-1}A\right|$ is small, the iterations converge fast
3. If $M \approx A$ (i.e., M is close to A), then $\left\|I - M^{-1}A\right\| = \left\|M^{-1}\left(M - A\right)\right\| \le \left\|M^{-1}\right\| \|M - A\|$
4. In the extreme case when $M = A$, we have $I - M^{-1}A = 0$ and the iterations converge in one step

3 Iterative solvers in practice

3.1 Implementation of an iterative method

The matrix M is called **preconditioner.**

Write (2) in the form

$$x^{(k+1)} = x^{(k)} + M^{-1}r^{(k)}, \quad r^{(k)} = b - Ax^{(k)}$$

Then one iteration can be written as 3 steps.

1. Compute the **residual** $r^{(k)} = b - Ax^{(k)}$
2. Solve the **preconditioning system** $M\delta^{(k)} = r^{(k)}$ for the **increment** $\delta^{(k)}$
3. **Apply** the increment: $x^{(k+1)} = x^{(k)} + \delta^{(k)}$

In application software, multiplication by A and solving the preconditioning system are usually implemented as functions. The matrices A or M are never stored explicitly. That would be just too expensive.

3.2 Design of preconditioners

Solving the preconditioning system $M\delta^{(k)} = r^{(k)}$ should be cheap.

The preconditioning should be close to the actual solution: $M^{-1}A \approx I$

Preconditioner is often constructed from a simplified or approximate version of the same problem.

For example, if A has large diagonal entries compared to the rest of the entries, D can be a good preconditioner. This is the Jacobi method. See "diagonally dominant" in the textbook. And solving a diagonal system is cheap.

4 Infinity norm

For $x \in \mathbb{R}^n$, define the vector norm, called the infinity norm, by

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

What is the induced norm? Consider matrix $A \in \mathbb{R}^{n\times n}$, let $y = Ax$, and estimate:

$$|y_i| = \left|\sum_{j=1}^n a_{ij}x_j\right| \le \sum_{j=1}^n |a_{ij}|\,|x_j| \le \sum_{j=1}^n |a_{ij}| \max_{i=j,\ldots,n} |x_j| = \sum_{j=1}^n |a_{ij}|\,\|x\|_\infty$$

$$\max_{i=1,\ldots,n} |y_i| \le \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}|\,\|x\|_\infty$$

Thus, we have

$$\|Ax\|_\infty \le \|A\|_\infty \|x\|_\infty \tag{6}$$

if we define

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

Exercise: show that the inequality (6) is sharp, that is, the inequality becomes equality for some $x \ne 0$, and we have in fact

$$\|A\|_\infty = \max_{x\ne 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$$

5 Convergence of Jacobi method for strictly diagonally dominant matrices

From (5), we have for the Jacobi method,

$$\left\|x^{(k)} - x^*\right\| \le \left\|\left(I - D^{-1}A\right)\right\|^k \left\|x^{(0)} - x^*\right\|$$

where D is the diagonal of A. Matrix $A = [a_{ij}]$ is called strictly diagonally dominant if for all i,

$$\sum_{\substack{j=1\\ j\ne i}}^n |a_{ij}| < |a_{ii}| \,. \tag{7}$$

So, suppose that $A \in \mathbb{R}^{n\times n}$ is strictly diagonally dominant, and compute $I - D^{-1}A$

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

Thus

$$\left\|I - D^{-1}A\right\| = \max_{i=1,\ldots,n} \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \max_{i=1,\ldots,n} \frac{\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|}{|a_{ii}|} < 1$$

because A is strictly diagonally dominant (7).

6 Gauss Seidel method

Solving system of n linear equations for n unknowns.

Idea: one iteration is computes unknown i from equation i for all i at the same time - using **new** values of x as soon as they are available

Example: The system of linear equations

$$\begin{aligned} 4x_1 - 3x_2 &= -1 \\ 2x_1 + 5x_2 &= 19 \end{aligned}$$

The Gauss-Seidel iterative method is: starting from given $x_1^{(0)}$, $x_2^{(0)}$, compute for $k = 0, 1, \ldots$

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{4}(-1 + 3x_2^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{5}\left(19 - 2x_1^{(k+1)}\right) \end{aligned}$$

For a system of n equations:

$$\sum_{j=1}^{n} a_{1j}x_j = b_i,\ i = 1, \ldots, n$$

$$a_{ii}x_i + \sum_{j=1}^{i-1} a_{1j}x_j + \sum_{j=i+1}^{n} a_{1j}x_j = b_i,\ i = 1, \ldots, n$$

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

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

Coding is simple: for $k = 1, 2, \ldots$

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

7 SOR

Write Gauss-Seidel in correction form - move $x_i^{(k)}$ out of the bracket

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

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

Now instead of the correction, **make ω times the correction** in every step

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

Equivalent writing - move $x_i^{(k)}$ back inside

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

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

Coding is simple: for $k = 1, 2, \ldots$

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

A significant improvement for $\omega > 1$. Can be proved to converge for $0 < \omega < 2$ (under assumptions).

SOR was a **major** improvement for matrices that come from discretizing differential equations, such as

$$\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$$

SOR made solving such equations numerically practical!

For more on such iterative methods, see the easy exposition [Var62]. For an exhaustive study of SOR and extensions, see [You03].

8 Choleski Decomposition

$$A = \begin{bmatrix} a_{11} & b^T \\ b & C \end{bmatrix}, \quad R = \begin{bmatrix} \alpha & \beta^T \\ 0 & \Gamma \end{bmatrix}$$

$$R^T R = \begin{bmatrix} \alpha & 0 \\ \beta & \Gamma^T \end{bmatrix} \begin{bmatrix} \alpha & \beta^T \\ 0 & \Gamma \end{bmatrix} = \begin{bmatrix} \alpha^2 & \alpha\beta \\ \beta\alpha & \Gamma^T\Gamma + \beta\beta^T \end{bmatrix} = \begin{bmatrix} a_{11} & b^T \\ b & C \end{bmatrix}$$

$$\alpha^2 = a_{11} \Rightarrow \alpha = \sqrt{a_{11}}$$
$$\beta\alpha = b \Rightarrow \beta = b/\alpha$$
$$\Gamma^T\Gamma + \beta\beta^T = C \Rightarrow \Gamma^T\Gamma = C - \beta\beta^T$$

Update lower right block to $C - \beta\beta^T$, use same method i dimension 1 less to find Γ such that $\Gamma^T\Gamma = C - \beta\beta^T$

$\beta\beta^T$ is called “outer product” hence this is called outer product cholesky algorithm

References

[Var62] Richard S. Varga. *Matrix iterative analysis.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1962.

[You03] David M. Young. *Iterative solution of large linear systems.* Dover Publications, Inc., Mineola, NY, 2003. Unabridged republication of the 1971 edition [Academic Press, New York-London, MR 305568].