2 | Stationary Iterative Methods

BRIEF THEORY

Fixed point iterations

Let $m \geq 1$ be a natural number, $\mathbf{A} \in \mathbb{R}^{m \times m}$ a nonsingular matrix, and $\mathbf{b} \in \mathbb{R}^m$. Then, the system $\mathbf{Ax} = \mathbf{b}$ has a unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. Here, the focus is on determining an approximation $\mathbf{x}_N$ of $\mathbf{x}$ by applying a mapping $\mathcal{A}: \mathbb{R}^m \to \mathbb{R}^m$. In this context, given an initial guess $\mathbf{x}_0$, a finite sequence $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ is generated by a recursive formula of the form $\mathbf{x}_{n+1} = \mathcal{A}(\mathbf{x}_n)$, that is,

$$(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = \Big(\mathbf{x}_0, \mathcal{A}(\mathbf{x}_0), \mathcal{A}\big(\mathcal{A}(\mathbf{x}_0)\big), \dots, \mathcal{A}^N(\mathbf{x}_0)\Big),$$

where $\mathcal{A}^N = \mathcal{A} \circ \mathcal{A} \circ \cdots \circ \mathcal{A}$ is the N times function composition also called the N-th iterate of $\mathcal{A}$, so that the norm $\|\mathbf{x} - \mathbf{x}_N\|/\|\mathbf{x}\|$ is sufficiently small, which means that the numerically obtained solution $\mathbf{x}_N$ is sufficiently close to the solution $\mathbf{x}$ of the original problem. In other words, the limit of the sequence $(\mathbf{x}_n)$ as $n \to \infty$ needs to exist and to coincide with $\mathbf{x}$. If $\mathcal{A}$ is Lipschitz continuous with Lipschitz constant $L < 1$, also called a contraction, then it can be shown that $\mathbf{x}$ is a so-called unique fixed point of $\mathcal{A}$, that is, it satisfies the equation $\mathcal{A}(\mathbf{x}) = \mathbf{x}$. If $\|\mathbf{x}_n - \mathbf{x}\| \to 0$ as $n \to \infty$, from the contraction definition we obtain $\|\mathcal{A}(\mathbf{x}_n) - \mathcal{A}(\mathbf{x})\| \to 0$ in the limit $n \to \infty$, meaning that each contraction is continuous.

Lipschitz mappings, contractions, and Banach's fixed point theorem

- A function $\mathcal{A}: \mathbb{R}^m \to \mathbb{R}^m$ is said to be Lipschitz continuous, if there exists a positive real constant number $L \geq 0$, such that $\|\mathcal{A}(\mathbf{x}) - \mathcal{A}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$.
- If $\mathcal{A}: \mathbb{R}^m \to \mathbb{R}^m$ is a contraction, then it admits a unique fixed point $\mathbf{x} \in \mathbb{R}^m$ that is obtained as the limit of the sequence with values $\mathbf{x}_{n+1} = \mathcal{A}(\mathbf{x}_n)$.

Manufacturing iterative methods

Here, we consider the m dimensional non-singular square system $\mathbf{Ax} = \mathbf{b}$. Iterative methods for linear systems generate convergent sequences of vectors, $(\mathbf{x}_n) \to \mathbf{x}$ as $n \to \infty$. At each iteration we define the residual

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

which is commonly used to define the termination criteria, meaning that a number N of iterations can be determined so that

$$\frac{\|\mathbf{r}_N\|}{\|\mathbf{r}_0\|} < \epsilon,$$

where $\epsilon > 0$ is a sufficiently small real number, the so-called tolerance.

Remark. In practice, to avoid unnecessary work, when the initial approximation $\mathbf{x}_0$ is far from the limit $\mathbf{x}$, we choose the termination criteria $\|\mathbf{r}_N\|/\|\mathbf{b}\| < \epsilon$. The two given criteria are equivalent for the initial guess $\mathbf{x}_0 = \mathbf{0}$.

The system $\mathbf{Ax} = \mathbf{b}$ can be written as

$$\mathbf{Ax} + \mathbf{x} = \mathbf{b} + \mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{x} - \mathbf{Ax} + \mathbf{b} \Leftrightarrow \mathbf{x} = (\mathbf{I} - \mathbf{A})\mathbf{x} + \mathbf{b},$$

from which we define the so-called Richardson fixed point iteration formula

$$\mathbf{x}_{n+1} = (\mathbf{I} - \alpha\mathbf{A})\mathbf{x}_n + \alpha\mathbf{b},$$

where $\alpha > 0$. In view of splitting methods, Richardson's recursion formula can be constructed by splitting the coefficient matrix as $\mathbf{A} = \mathbf{M} - \mathbf{K}$, where $\mathbf{M}$ is a non-singular matrix and $\mathbf{K}$ is the remainder; hence,

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{Mx} - \mathbf{Kx} = \mathbf{b} \Leftrightarrow \mathbf{Mx} = \mathbf{Kx} + \mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{M}^{-1}\mathbf{Kx} + \mathbf{M}^{-1}\mathbf{b},$$

from which we define the recursion formula

$$\mathbf{x}_{n+1} = \mathbf{M}^{-1}\mathbf{K}\mathbf{x}_n + \mathbf{M}^{-1}\mathbf{b}.$$

The splitting of $\mathbf{A}$ is chosen so that $\mathbf{M}^{-1}\mathbf{K}$ and $\mathbf{M}^{-1}\mathbf{b}$ are easy to calculate, while avoiding any explicit inverse computation, if possible. Choosing, for instance, $\mathbf{M} = \alpha^{-1}\mathbf{I}$ and $\mathbf{K} = \alpha^{-1}\mathbf{I} - \mathbf{A}$, with $\alpha > 0$, we obtain Richardson's iterations, also written

$$\mathbf{x}_{n+1} = \mathcal{A}(\mathbf{x}_n) = \mathbf{R}\mathbf{x}_n + \alpha\mathbf{b},$$

where $\mathbf{R} = \mathbf{I} - \alpha\mathbf{A}$. Note that

$$\|\mathcal{A}(\mathbf{x}) - \mathcal{A}(\mathbf{y})\| = \|\mathbf{Rx} + \alpha\mathbf{b} - \mathbf{Ry} - \alpha\mathbf{b}\| = \|\mathbf{R}(\mathbf{x} - \mathbf{y})\| \leq \|\mathbf{R}\|\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$. Hence, if $\|\mathbf{R}\| < 1$, then $\mathcal{A}$ is a contraction and as a result, Richardson's iterations converge to the exact solution $\mathbf{x}$. The matrix $\mathbf{R}$ is called the iteration or relaxation matrix of the Richardson method. Generally, methods with recursion formulas of the form $\mathbf{x}_{n+1} = \mathbf{R}\mathbf{x}_n + \mathbf{c}$, where the iteration matrix $\mathbf{R}$ and the vector $\mathbf{c}$ are constant throughout the iterations, are said to be stationary iterative methods. The class of stationary methods contains members such as the Jacobi method, the Gauss-Seidel method, and the (symmetric) successive over-relaxation method ((S)SOR).

Remark. If $\|\mathbf{R}\| < 1$ (sub-multiplicative matrix norm $\|\cdot\|$), then

$$\|\mathbf{S}_n\| = \left\|\sum_{k=0}^{n} \mathbf{R}^k\right\| \leq \sum_{k=0}^{n} \|\mathbf{R}\|^k \leq \frac{1}{1-\|\mathbf{R}\|},$$

meaning that the infinite series $\mathbf{S}_\infty$ is convergent. Further, we have

$$(\mathbf{I}-\mathbf{R})\mathbf{S}_n = \mathbf{S}_n - \mathbf{R}\mathbf{S}_n = \mathbf{I} - \mathbf{R}^{n+1},$$

Hence, letting $n \to \infty$ we obtain $\mathbf{S}_\infty = (\mathbf{I}-\mathbf{R})^{-1}$, since $\|\mathbf{R}\| < 1$.

For the error $\mathbf{e}_n = \mathbf{x} - \mathbf{x}_n$ it holds

$$\mathbf{e}_{n+1} = \mathbf{x} - \mathbf{x}_{n+1} = \mathbf{R}\mathbf{x} + \mathbf{c} - \mathbf{R}\mathbf{x}_n - \mathbf{c} = \mathbf{R}(\mathbf{x} - \mathbf{x}_n) = \cdots = \mathbf{R}^{n+1}\mathbf{e}_0.$$

Assuming that $\|\mathbf{R}\| < 1$, the norm of the error vanishes as $n \to \infty$, since

$$\|\mathbf{e}_{n+1}\| \leq \|\mathbf{R}\|^{n+1}\|\mathbf{e}_0\| \to 0 \text{ as } n \to \infty,$$

from which it follows that $\mathbf{R}$ should be chosen so that $\|\mathbf{R}\|$ is as small as possible.

Jacobi and Gauss-Seidel methods

Here, we are interested in splitting the coefficient matrix $\mathbf{A}$ as

$$\mathbf{A} = \underbrace{\mathbf{D} - \mathbf{U}}_{\mathbf{M}} - \underbrace{\mathbf{L}}_{\mathbf{K}}$$

where $\mathbf{D}$ is the diagonal of $\mathbf{A}$, $-\mathbf{U}$ is the strictly upper part of $\mathbf{A}$, and $-\mathbf{L}$ is its strictly lower part of $\mathbf{A}$.

- If we choose $\mathbf{R} = \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$ and $\mathbf{c} = \mathbf{D}^{-1}\mathbf{b}$, or equivalently $\mathbf{M} = \mathbf{D}$ and $\mathbf{K} = \mathbf{L} + \mathbf{U}$, we obtain the so-called Jacobi method

$$\mathbf{x}_{n+1} = \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x}_n + \mathbf{D}^{-1}\mathbf{b},$$

where the required inverse $\mathbf{D}^{-1}$ is easy to obtain, since $\mathbf{D}$ it is diagonal. If the matrix $\mathbf{A}$ is strictly diagonally dominant, then the Jacobi method converges.

Remark. A matrix $\mathbf{A}$ is said to be (strictly) diagonally dominant if, for each row, the absolute value of the diagonal entry is (strictly) larger than the sum of the absolute values of the other entries.

- If we choose $\mathbf{R} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}$ and $\mathbf{c} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{b}$, or equivalently $\mathbf{M} = \mathbf{D}-\mathbf{L}$ and $\mathbf{K} = \mathbf{U}$, we obtain the so-called Gauss-Seidel method

$$\mathbf{x}_{n+1} = (\mathbf{D}-\mathbf{L})^{-1}(\mathbf{U}\mathbf{x}_n + \mathbf{b}),$$

where the matrix $\mathbf{D}-\mathbf{L}$ is lower triangular, and hence, the effect of the inverse $(\mathbf{D}-\mathbf{L})^{-1}$ can be computed by forward elimination. The Gauss-Seidel method is ensured to converge, if $\mathbf{A}$ is symmetric and positive-definite.

The spectral radius

In the convergence analysis we presented above, we used a generic matrix norm; however, the magnitude of a matrix as measured by one norm can vary significantly, when employing another norm. Note that on a finite-dimensional linear space all norms are equivalent; that is, there exist real numbers $\mu, M > 0$ such that for all $\mathbf{x} \in \mathbb{R}^m$ we have

$$\mu\|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq M\|\mathbf{x}\|_a.$$

In a numerical context, though the limits with respect to two equivalent norms coincide, the norm of the iteration matrix $\mathbf{R}$ of a stationary iterative scheme does not describe the performance of the corresponding recursion formula accurately. Recall that the set $\sigma(\mathbf{A}) = \{\alpha \in \mathbb{C}: \phi_{\mathbf{A}}(\alpha) = 0\}$ of the eigenvalues of $\mathbf{A}$ is called the spectrum of $\mathbf{A}$. The number $\varrho(\mathbf{A}) = \max\{|\alpha| \in \mathbb{R}: \alpha \in \sigma(\mathbf{A})\}$ is called the spectral radius of $\mathbf{A}$. The spectral radius satisfies the inequality $\varrho(\mathbf{A}) \leq \|\mathbf{A}\|$ for all induced matrix norms, and hence, it is independent of any matrix norm. Further, for any $\epsilon > 0$ there is a norm $\|\cdot\|$ on $\mathbb{R}^m$ such that $\varrho(\mathbf{A}) > \|\mathbf{A}\| - \epsilon$. Thus, if $\varrho(\mathbf{R}) < 1$, then stationary iterations of the form $\mathbf{x}_{n+1} = \mathbf{R}\mathbf{x}_n + \mathbf{c}$ are convergent.

EVALUATION QUIZ

2.1 Characterize the following statements as true (T) or false (F) and justify your characterization briefly. Here, $\mathbf{A} \in \mathbb{R}^{m\times m}$, with $m > 1$, is the coefficient matrix of a well-posed linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$.

(a) If a function is Lipschitz continuous, then it has a unique fixed point.

(b) The function with values $f(x) = \sqrt{x}$ is not Lipschitz continuous.

(c) A diagonally dominant matrix is non-singular.

(d) A strictly diagonally dominant matrix can be singular.

(e) Iterative schemes of the form $\mathbf{x}_{n+1} = \mathbf{R}_n\mathbf{x}_n + \mathbf{b}_n$ are called stationary.

(f) A stationary iterative method with relaxation matrix $\mathbf{R}$ is convergent if $\|\mathbf{R}\| = 1$.

(g) If $\mathbf{A}$ is diagonally dominant, then the Jacobi method generates a convergent sequence.

(h) If $\mathbf{A}$ is positive definite, then the Gauss-Seidel method is convergent.

(i) For $\omega = 1$ the recursive formula of the SOR scheme reduces to Jacobi iterations.

2.2 After writing the Gauss-Seidel method in component form, use the function template `out(x,N)/in(A,b,x0,tol)` to

(a) implement the Richardson, Jacobi and Gauss-Seidel methods.

(b) Make a comparison between Richardson, Jacobi, and Gauss-Seidel methods, in terms of the number of iterations till convergence, for a (strictly diagonally dominant) pseudorandom matrix of size $m = 1000$.

(c) Measure the execution time of Richardson, Jacobi, and Gauss-Seidel methods; compare the execution times of each one of these methods with the time it takes to solve the linear system, when using a direct method.

(d) Use Jacobi and Gauss-Seidel iterations to solve linear systems with coefficient matrices of the form `A=rand(m)+a*eye(m)` for `a=1:100` and constant right-hand side filled with ones. For each one of these matrices and each method, compute the number of iterations and make two plots (one for each method) of the number of iterations versus `a`. Comment on the results.

2.3 For the solution of the linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is an $m \times m$ matrix, we consider the so-called relaxation scheme

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

for $i = 1,2,\ldots m$, where ω is a real number. When $\omega = 1$, this scheme coincides with the Gauss-Seidel method, while when $\omega \in (1,2)$, it defines the so-called successive over-relaxation (SOR) method.

(a) Find the explicit form of the corresponding iteration matrix.

(b) Show that the condition $\omega \in (0,2)$ is necessary for this scheme to generate convergent successive approximations.

(c) Implement this relaxation method and let ω be an input parameter.

(d) Perform numerical experiments to find a sufficiently good ω value for the problem solved in the mini project of the introductory lecture. Make a plot of the number of iterations versus ω.