MA398 Matrix Analysis and Algorithms: Exercise Sheet 8

1. (Jacobi) Consider the Jacobi method for solving

$$Ax = b, \quad \text{with } A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \, b = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

and with start value $x^{(0)} = (0, 0, 0)^T$.

(a) State the iteration matrix $R = -D^{-1}(L + U)$, compute its spectral radius $\rho(R)$ and deduce that the Jacobi method converges.

(b) Recall the estimate

$$k \geq k^\sharp = \frac{\log(\|A\|_2 \|e^{(0)}\|_2 / \|b\|_2) - \log(\varepsilon_r)}{\log(\|R\|_2^{-1})}$$

for the number of steps in order to achieve that $\|r^{(k)}\|_2 \leq \varepsilon_r \|b\|_2$.
For the above specific data, give an upper bound for the number of steps required to get the relative error of the residual below 10^{-6}.

(c) Derive the estimate

$$\frac{\|e^{(k)}\|_2}{\|x\|_2} \leq \kappa_2(A) \frac{\|r^{(k)}\|_2}{\|b\|_2}$$

and give an upper bound for the number of steps required to get the relative error of the solution below 10^{-6}.

(d) State the definition of the graph $G(B)$ of a matrix $B \in \mathbb{C}^{n \times n}$.
Prove that $B \in \mathbb{C}^{n \times n}$ is irreducible if and only if its graph $G(B)$ is connected.

Answer:

(a) The iteration matrix of the Jacobi method for the problem is

$$R = -D^{-1}(L + R) = -\begin{pmatrix} -\frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Its characteristic polynomial is $\rho_R(z) = z^3 - \frac{1}{2}z \Rightarrow$ eigenvalues are $\{1/\sqrt{2}, 0, -1/\sqrt{2}\}$. The spectral radius of R is $\rho(R) = 1/\sqrt{2} < 1$. Therefore the Jacobi iteration converges.

(b) Here, $\varepsilon_r = 10^{-6}$. The characteristic polynomial of A is $\rho_A(z) = (z-2)^3 - 2(z-2)$ with zeros $\{2+\sqrt{2}, 2, 2-\sqrt{2}\}$. Recall that for Hermitian matrices the induced norm $\|\cdot\|_2$ is the spectral radius. Hence $\|A\|_2 = 2+\sqrt{2}$. Further, $\|b\|_2 = \sqrt{2}$, and with the solution $x = (1, 1, 1)^T \in \mathbb{R}^3$ for the linear system we have that $\|e_0\|_2 = \|x - x_0\|_2 = \|x\|_2 = \sqrt{3}$. Inserting these numbers and $\|R\|_2 = \rho(R) = 1/\sqrt{2}$ into the estimate for the number of steps we end up with

$$k \geq \frac{\log((2 + \sqrt{2})\sqrt{3}/(10^{-6}\sqrt{2}))}{\log(\sqrt{2})} (\approx 43.99)$$

as a sufficient number of steps.

(c) We have $\|b\|_2 = \|Ax\|_2 \le \|A\|_2\|x\|_2$ so that $\frac{1}{\|x\|_2} \le \frac{\|A\|_2}{\|b\|_2}$. Furthermore, $Ae^{(k)} = r^{(k)}$ so that $\|e^{(k)}\|_2 = \|A^{-1}r^{(k)}\|_2 \le \|A^{-1}\|_2\|r^{(k)}\|_2$. Together we obtain that

$$\frac{\|e^{(k)}\|_2}{\|x\|_2} \le \|A^{-1}\|_2\|r^{(k)}\|_2\frac{\|A\|_2}{\|b\|_2} = \|A\|_2\|A^{-1}\|_2\frac{\|r^{(k)}\|_2}{\|b\|_2} = \kappa_2(A)\frac{\|r^{(k)}\|_2}{\|b\|_2}.$$

We have $\|e^{(k)}\|_2/\|x\|_2 \le \varepsilon_r$ if $\|r^{(k)}\|_2/\|b\|_2 \le \varepsilon_r/\kappa_2(A)$ so we just have to replace ε_r by $\varepsilon_r/\kappa_2(A)$ and proceed as before. Since the eigenvalues of A^{-1} are the inverse eigenvalues of A we have that $\|A^{-1}\|_2 = \rho(A^{-1}) = 1/(2-\sqrt{2})$, hence

$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{2+\sqrt{2}}{2-\sqrt{2}}.$$

Therefore, performing

$$k \ge \frac{\log((2+\sqrt{2})^2(2-\sqrt{2})\sqrt{3}/(10^{-6}\sqrt{2}))}{\log(\sqrt{2})}(\approx 45.99)$$

Jacobi steps ensures that $\|e^{(k)}\|_2/\|x\|_2 \le \varepsilon_r = 10^{-6}$.

(d) The graph $G(B)$ of B is an oriented graph with vertices $1,\dots,n$ and edges $i \to j$ if $a_{i,j} \neq 0$. We first show "$\Rightarrow$" by a contradiction argument. Assume that $G(B)$ is not connected. There is a vertex k to which not all vertices are connected by a chain of edges. Let $S \subsetneq \{1,\dots,n\}$ denote the set of vertices connected to k. Pick any $j \in S$ and any $i \in \{1,\dots,n\}\backslash S$. Then

$$b_{ij} = 0 \tag{$\star$}$$

since otherwise i would be connected to $j \in S$, but since j is connected to k then also i would be connected to k in contradiction to $i \notin S$. After a suitable permutation ($B = P\tilde{B}P^T$) we may assume that $S = \{1,\dots,p\}$ with $p < n$ and let $q = n-p$. By ($\star$) the lower left block of size $q \times p$ in $\tilde{B}$ vanishes, hence B is not irreducible.
Now, we show "$\Leftarrow$". Assume that B is not irreducible. Up to renumbering of the vertices, the graphs of B and $\tilde{B}$ are the same. Therefore, it is sufficient to show that $G(\tilde{B})$ is not connected. Let $i > p$ and $j \le p$ be two vertices of $G(\tilde{B})$ and suppose that there is a chain of edges

$$i = i_0 \to i_1 \to \cdots \to i_k = j$$

connecting them. Necessarily, there is an edge $i_l \to i_{l+1}$ with $i_l > p$ and $i_{l+1} \le p$. But since $\tilde{a}_{i_l,i_{l+1}} = 0$ such an edge cannot exist. Hence, i cannot be connected to j so that $G(\tilde{B})$ is not connected.

2. (SSOR) The <u>symmetric successive over relaxation</u> consists in performing the following iteration:

$$i = 1,\dots,n: \qquad a_{ii}x_i^{(k+\frac{1}{2})} = \omega\Big(-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+\frac{1}{2})} - \sum_{j=i+1}^{n}a_{ij}x_j^{(k)} + b_i\Big) - (\omega-1)a_{ii}x_i^{(k)},$$

$$i = n,\dots,1: \qquad a_{ii}x_i^{(k+1)} = \omega\Big(-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+\frac{1}{2})} - \sum_{j=i+1}^{n}a_{ij}x_j^{(k+1)} + b_i\Big) - (\omega-1)a_{ii}x_i^{(k+\frac{1}{2})}.$$

Here, $x^{(k)}$ stands for the k^{th} iterate, and $x^{(k+\frac{1}{2})}$ is an intermediate value.
Show that SSOR is a linear iterative method with

$$M_{\text{SSOR}}^{-1} = \omega(2-\omega)(D+\omega U)^{-1}D(D+\omega L)^{-1}.$$

Remark: Recalling that SOR uses $M_{\mathrm{SOR}} = \frac{1}{\omega}D + L$ we see that SSOR essentially consists in performing an SOR step followed by a reverse SOR step with $\frac{1}{\omega}D + U$, which explains its name. A couple of SSOR steps sometimes are applied as a preconditioner in CG.

Answer: From the first part of the step we have that

$$a_{ii}x_i^{(k+\frac{1}{2})} + \sum_{j<i}\omega a_{ij}x_j^{(k+\frac{1}{2})} = \omega b_i - \sum_{j=i+1}^{n}\omega a_{ij}x_j^{(k)} - (\omega-1)a_{ii}x_i^{(k)}$$

so that, after dividing by ω,

$$\big(\tfrac{1}{\omega}D + L\big)x^{(k+\frac{1}{2})} = b - \big(U + (1-\tfrac{1}{\omega})D\big)x^{(k)}.$$

Similarly, the second part of the step gives

$$\big(\tfrac{1}{\omega}D + U\big)x^{(k+1)} = b - \big(L + (1-\tfrac{1}{\omega})D\big)x^{(k+\frac{1}{2})}.$$

Observe that

$$\begin{aligned}
&\big(L + (1-\tfrac{1}{\omega})D\big)\big(\tfrac{1}{\omega}D + L\big)^{-1}\\
=&\big(L + \tfrac{1}{\omega}D + (1-\tfrac{2}{\omega})D\big)\big(\tfrac{1}{\omega}D + L\big)^{-1}\\
=&I + (1-\tfrac{2}{\omega})D\big(\tfrac{1}{\omega}D + L\big)^{-1}\\
=&I + (\omega-2)D\big(D+\omega L\big)^{-1}.
\end{aligned}\tag{1}$$

Inserting the formula for $x^{(k+\frac{1}{2})}$ into the one for $x^{(k)}$ therefore yields

$$\begin{aligned}
\big(\tfrac{1}{\omega}D + U\big)x^{(k+1)} &= b - \big(L + (1-\tfrac{1}{\omega})D\big)\big(\tfrac{1}{\omega}D + L\big)^{-1}\big(b - \big(U + (1-\tfrac{1}{\omega})D\big)x^{(k)}\big)\\
&= -(\omega-2)D\big(D+\omega L\big)^{-1}b\\
&+ \big(U + (1-\tfrac{1}{\omega})D\big)x^{(k)}\\
&+ (\omega-2)D\big(D+\omega L\big)^{-1}\big(U + (1-\tfrac{1}{\omega})D\big)x^{(k)}.
\end{aligned}$$

Similarly to (1)

$$\big(\tfrac{1}{\omega}D + U\big)^{-1}\big(U + (1-\tfrac{1}{\omega})D\big) = I + \big(D+\omega U\big)^{-1}(\omega-2)D.$$

We conclude that

$$\begin{aligned}
x^{(k+1)} &= \big(\tfrac{1}{\omega}D + U\big)^{-1}(2-\omega)D\big(D+\omega L\big)^{-1}b\\
&+ \big(I + \big(D+\omega U\big)^{-1}(\omega-2)D\big)x^{(k)}\\
&+ \big(\tfrac{1}{\omega}D + U\big)^{-1}(\omega-2)D\big(D+\omega L\big)^{-1}\big(U + (1-\tfrac{1}{\omega})D\big)x^{(k)}\\
&= M_{\mathrm{SSOR}}^{-1}b + x^{(k)}\\
&+ \big(D+\omega U\big)^{-1}(\omega-2)D\omega\big(D+\omega L\big)^{-1}\tfrac{1}{\omega}\big(D+\omega L\big)x^{(k)}\\
&- M_{\mathrm{SSOR}}^{-1}\big(U + (1-\tfrac{1}{\omega})D\big)x^{(k)}\\
&= M_{\mathrm{SSOR}}^{-1}b - M_{\mathrm{SSOR}}^{-1}\big(-M_{\mathrm{SSOR}}\big)x^{(k)}\\
&- M_{\mathrm{SSOR}}^{-1}\big(\underbrace{\tfrac{1}{\omega}D + L + U + (1-\tfrac{1}{\omega})D}_{=D+L+U=A}\big)x^{(k)}\\
&= M_{\mathrm{SSOR}}^{-1}b - M_{\mathrm{SSOR}}^{-1}\big(\underbrace{A - M_{\mathrm{SSOR}}}_{=:N_{\mathrm{SSOR}}}\big)x^{(k)}\\
&= M_{\mathrm{SSOR}}^{-1}\big(b - N_{\mathrm{SSOR}}x^{(k)}\big).
\end{aligned}$$

3. Implement the Gauss-Seidel method, a variant of the Jacobi method, where $M := L + D$ and $N := U$. Write a Python function with the signature def gauss_seidel_method(A, b, x0, max_iter, tol, omega=1.0), where omega is the relaxation parameter.

    (a) Test your function on a system of linear equations with a diagonally dominant matrix A and different choices of omega. Plot the norm of the residual vector as a function of the number of iterations for different choices of omega.

    (b) Use your implementation to investigate the impact of the relaxation parameter on the convergence of the method. For which values of omega does the method converge fastest?

    (c) If possible, compare the performance (in terms of both accuracy and computational cost) of your Gauss-Seidel implementation with a basic Jacobi method implementation.