Norwegian University of Science
and Technology
Department of Mathematical
Sciences

TMA4205 Numerical Linear Algebra Fall 2017

Semester project – part 1

1 Consider the following one-dimensional boundary value problem (BVP) for the advection–diffusion equation

$$\begin{aligned} -U_{xx} + aU_x &= f \qquad \text{in } \Omega = (0,1), \\ U(0) &= 1, \qquad U(1) = -1, \end{aligned} \tag{1}$$

where $a = a(x)$ and $f = f(x)$ are given functions. For the discretisation of this equation, we first subdivide the interval Ω into n subintervals of length $h = 1/n$ with end-points $x_j = jh$, $j = 0, \ldots, n$. Then we try to find a numerical approximation $u = (u_j)_{j=0,\ldots,n}$ of the solution U of (1) at the grid points x_j by solving a linear system obtained from (1) by replacing all differential operators by suitable finite difference approximations. For the diffusive term $-U_{xx}$ we choose a (standard) 3-point central scheme

$$U_{xx}(x_j) \approx \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2}. \tag{2}$$

For the advective term U_x, we may choose either of the discretisations

$$U_x(x_j) \approx \frac{u_{j+1} - u_{j-1}}{2h}, \quad U_x(x_j) \approx \frac{u_{j+1} - u_j}{h}, \quad \text{or} \quad U_x(x_j) \approx \frac{u_j - u_{j-1}}{h}. \tag{3}$$

These choices correspond to central, forward, and backward finite differences, respectively, and it is possible to choose a different discretisation at each of the points x_j. Inserting the approximations (2) and (3) into (1) evaluated (collocated) at $x = x_j$, we arrive at a sparse linear system of equations $Au = b$ for the vector of unknowns $u = (u_1, \ldots, u_{n-1})^T$. Note that $u_0 = U(0) = 1$ and $u_n = U(1) = -1$ are just the boundary values and therefore known a-priori. We can now write the linear system in the form

$$A = \begin{bmatrix} \alpha_1 & \delta_1 & & & \\ \gamma_2 & \alpha_2 & \delta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{n-2} & \alpha_{n-2} & \delta_{n-2} \\ & & & \gamma_{n-1} & \alpha_{n-1} \end{bmatrix}, \qquad b = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n-2} \\ \beta_{n-1} \end{bmatrix},$$

where the values β_j have the form $\beta_j = h^2 f(x_j) + \tau_j$ with τ_j accounting for the boundary contributions; all of these values depend on the choice of the discretisation of (3).

a) Abbreviating $a_j := a(x_j)$, compute the explicit forms of α_j, γ_j, δ_j, and τ_j for each of the possible discretisations of the advection term in (3).

b) Formulate a strategy for selecting the discretisation of the advection term in (3) in such a way (depending on a) that the resulting matrix A is guaranteed to

be irreducibly row diagonally dominant for all grid sizes $h > 0$.[1] (This will guarantee that both the Jacobi and the Gauß–Seidel method for this system converge for all starting values $u^{(0)}$.) Note that it might be necessary to choose different discretisations for different points x_j.

From now on, we assume that a is the constant function $a(x) = -2$ and we choose forward differences for all the discretisations of the advection term (that is, the second possibility in (3)), which implies that $\alpha = \alpha_j$, $\delta = \delta_j$, and $\gamma = \gamma_j$ are all independent of j.

c) Provide an explicit formula for the eigenvalues of A.

Hint: Use the note "Eigenvalues of tridiagonal Toeplitz matrices", which can be found on the home page. No derivations of the formulas there are required.

We will now study the behaviour of simple matrix-splitting methods for the solution of our linear system. Write $A = D - E - F$, where D, $-E$, and $-F$ denote the diagonal, strict lower triangular, and strict upper triangular parts of A, respectively (see Section 4.1 in YS).

d) Consider Jacobi iteration, $u^{(k+1)} = G_J u^{(k)} + D^{-1}b$, with $G_J = D^{-1}(E + F)$. Using the same methods as for **c)**, find the eigenvalues of the iteration matrix G_J and determine the spectral radius of G_J.

e) How would you expect the error $e^{(k)} = u - u^{(k)}$, that is, the difference between the k-th iterate $u^{(k)}$ obtained by means of Jacobi iteration and the actual solution u of $Au = b$, to behave as a function of k and n in this particular problem? Specifically, assume you double n, how should you change k in order to obtain an error $e^{(k)}$ of approximately the same size? Does the choice of the norm in which you measure the error influence your estimates?

Hint: Taylor series expansions with respect to $h = n^{-1}$ for small parameter h can be helpful for obtaining useful asymptotic estimates for large n.

f) Consider again the problem (1) with exact solution given by $U(x) = \cos(\pi x)$. What is the corresponding right-hand side f? Let $n = 20$ and use Jacobi iteration to solve the corresponding discrete system with this choice of f. Define u_* to be the vector with entries $U(x_i)$, $i = 1, \ldots, n-1$, that is, the continuous solution evaluated at the interior grid points. Define also $e_*^{(k)} = u_* - u^{(k)}$ and plot $\log(\|e_*^{(k)}\|_\infty)$ as a function of k. Iterate until the error $e_*^{(k)}$ no longer changes. Next, increase n to 40, and repeat the solution process. Finally, do the same with $n = 80$. Compare the convergence behaviour for all three cases (e.g. in one single plot). Are the results as expected? Can you explain your observations?

Hint: $u_* - u^{(k)} = (u_* - u) + (u - u^{(k)})$.

g) In the notation of **f)**, set $n = 40$ and compare the behaviour of the Jacobi

[1] Replace all of the non-zero entries of A by 1, and interpret the resulting matrix as the adjacency matrix of a directed graph. The original matrix A is irreducible, if the resulting directed graph is strongly connected.

method and both the forward and the backward Gauß–Seidel methods,[2]

$$\begin{aligned} u_J^{(k+1)} &= D^{-1}(E+F)u_J^{(k)} + D^{-1}b, \\ u_f^{(k+1)} &= (D-E)^{-1}Fu_f^{(k)} + (D-E)^{-1}b, \\ u_b^{(k+1)} &= (D-F)^{-1}Eu_b^{(k)} + (D-F)^{-1}b, \end{aligned}$$

for instance by plotting $\log(\|e_*^{(k)}\|_\infty)$ for the different methods as a function of k.

Try to find a (heuristic) explanation as to why one version of the Gauß–Seidel method outperforms the other one for this particular problem.

[2] The following definitions, especially of the forward and backward Gauß–Seidel methods, should be seen as purely formal. In an actual implementation the updates of the different components of u all have the form

$$u_i \leftarrow \frac{1}{a_{ii}}\big(b_i - \sum_{j\neq i} a_{ij}u_j\big),$$

and the difference between the methods is only the order in which these operations are executed:

- simultaneously in the Jacobi method—this can also be implemented using a matrix–vector multiplication;
- in the order $i = 1, 2, \ldots, n$ in the forward Gauß–Seidel method;
- in the order $i = n, n-1, \ldots, 1$ in the backward Gauß–Seidel method.