DELFT UNIVERSITY OF TECHNOLOGY
Faculty of Electrical Engineering, Mathematics and Computer Science

ANSWERS OF THE TEST SCIENTIFIC COMPUTING (wi4201)
Wednesday January 25 2023, 13:30-16:30

This are short answers, which indicate how the exercises can be answered. In most of the cases more details are needed to give a sufficiently clear answer.

1. (a) The matrix norm $\|A\|_{max}$ is defined as $\|A\|_{max} = \max_{1\leq i,j \leq n} |a_{i,j}|$. This norm does not have the multiplicative property. The multiplicative property holds if for any $R_1 \in \mathbb{R}^{m\times q}$ and $R_2 \in \mathbb{R}^{q \times n}$ $\|R_1 R_2\|_p \leq \|R_1\|_p \|R_2\|_p$ A counterexample is $R_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $R_2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$.

   (b) This is not true. $\|R\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |r_{ij}|$ (maximum absolute column sum). A counterexample is: $R = \begin{pmatrix} 2 & 1 \\ 2 & 5 \end{pmatrix}$. Note that to find $\|R\|_1$ we are looking for vector $(x_1, x_2)^T$ such that $|x_1| + |x_2| = 1$ and $|2x_1 + x_2| + |2x_1 + 5x_2|$ is maximal. This leads to $x_1 = 0$ and $x_2 = 1$, so $\|R\|_1 = 6$, which is the maximum absolute column sum of R.

   (c) We know that all p-norms have the multiplicative property. This implies that the following inequalities are valid:

$$1 = \|I\|_p = \|AA^{-1}\|_p \leq \|A\|_p \|A^{-1}\|_p = \kappa_p(A)$$

   (d) This is not true. Since the matrix is symmetric, all eigenvalues are real valued. From Gershgorin's theorem it is easy to see that all eigenvalues are less than or equal to 4.

   (e) If $\rho(A) < 1$, then $I - A$ has eigenvalues bounded away from zero, and is therefore non-singular. We furthermore have the equality

$$(I - A^{k+1}) = (I - A)(I + A + A^2 + \ldots + A^k)\,, \tag{1}$$

   or equivalently

$$(I - A)^{-1}(I - A^{k+1}) = (I + A + A^2 + \ldots + A^k)\,. \tag{2}$$

   Taking the limit as $k \to \infty$ and taking into account that since $\rho(A) < 1$ it follows that $\lim_{k\to\infty} \|A^k\|_2 = 0$ yields the desired result.

2. (a) The finite difference stencil is given by

$$\frac{1}{h^2}[-1 - x_i \quad 2(1 + x_i) + 4h^2 \quad -1 - x_i]$$

In order to show that the method is second order accurate, a Taylor expansion of u in the points x_{i-1} and x_{i+1} should be given around the point x_i where the remainder term is $O(h^4)$. It then follows that

$$\frac{-(1+x_i)u_{i-1} + (2(1+x_i)+4h^2)u_i - (1+x_i)u_{i+1}}{h^2} + 4u_i - x_i^2 =$$

$$-(1+x_i)\frac{d^2u(x_i)}{dx^2} + O(h^2) + 4u(x_i) - x_i^2 = O(h^2)$$

The final equality holds because u is the solution of the given differential equation.

(b) Note that for $N = 4$, we have $h = 0.25$ and there are 3 internal grid point, so the lenght of vector $\mathbf{u}$ is equal to 3. The general formula in an internal grid point is

$$\frac{-(1+x_i)u_{i-1} + (2(1+x_i)+4h^2)u_i - (1+x_i)u_{i+1}}{h^2} + 4u_i = x_i^2$$

Using the values for N and h the following equations should hold:

$$(\frac{2}{0.25^2}(1+0.25)+4)u_1 - \frac{1}{0.25^2}(1+0.25)u_2 = 0.25^2$$

$$-\frac{1}{0.25^2}(1+0.5)u_1(\frac{2}{0.25^2}(1+0.5)+4)u_2 - \frac{1}{0.25^2}(1+0.5)u_3 = 0.5^2$$

$$-\frac{1}{0.25^2}(1+0.75)u_2 + (\frac{2}{0.25^2}(1+0.75)+4)u_3 = 0.75^2$$

This leads to $A = \begin{pmatrix} 44 & -20 & 0 \\ -24 & 52 & -24 \\ 0 & -28 & 60 \end{pmatrix}$. and $\mathbf{f} = \begin{pmatrix} 0.0625 \\ 0.25 \\ 0.5625 \end{pmatrix}$.

(c) Since matrix A is non-symmetric it is possible that the eigenvalues are complex valued. To bound eigenvalues we can use Gershgorin's theorem:
If $\lambda \in \sigma(A)$, then λ is located in one of the n closed disks in the complex plane that has center a_{ii} and radius

$$\rho_i = \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

For $i = 1$ we have

$$a_{1,1} = \frac{2(1+x_1)+4h^2)}{h^2} \text{ and } \rho_1 = \frac{1+x_1}{h^2},$$

for $i = 2, ..., N-1$ we have

$$a_{i,i} = \frac{2(1+x_i)+4h^2)}{h^2} \text{ and } \rho_i = \frac{2(1+x_i)}{h^2},$$

and finally for $i = N$ we have

$$a_{N,N} = \frac{2(1+x_N)+4h^2)}{h^2} \text{ and } \rho_N = \frac{1+x_N}{h^2}.$$

These expressions combined with Gershgorin's theorem shows that the real part of all eigenvalues of A is larger than or equal to 4.

3. (a) Given that a non-singular matrix M exists we can split A as follows: $A = M - N$. The residual is defined by $\mathbf{r}^k = \mathbf{f} - A\mathbf{u}^k$. For the iteration formula we have:

$$M\mathbf{u}^{k+1} = N\mathbf{u}^k + \mathbf{f}$$

This can be rewritten as:

$$M\mathbf{u}^{k+1} = M\mathbf{u}^k - M\mathbf{u}^k + N\mathbf{u}^k + \mathbf{f} = M\mathbf{u}^k - A\mathbf{u}^k + \mathbf{f}.$$

Multiplication with M^{-1} and using the definition of the residual gives: $\mathbf{u}^{k+1} = \mathbf{u}^k + M^{-1}\mathbf{r}^k$.

(b) The error is defined as $\mathbf{e}^k = \mathbf{u} - \mathbf{u}^k$. Since $A\mathbf{u} = (M - N)\mathbf{u} = \mathbf{f}$ we have that:

$$M\mathbf{u} = N\mathbf{u} + \mathbf{f}$$

For the iteration formula we have:

$$M\mathbf{u}^{k+1} = N\mathbf{u}^k + \mathbf{f}$$

combination gives:

$$M(\mathbf{u}^{k+1} - \mathbf{u}) = N(\mathbf{u}^k - \mathbf{u})$$

Since $N = M - A$, this can also be written as: $M\mathbf{e}^{k+1} = (M - A)\mathbf{e}^k$. Multiplication with M^{-1} shows that

$$\mathbf{e}^{k+1} = (I - M^{-1}A)\mathbf{e}^k.$$

This holds for all splitting, so also for the Jacobi iteration matrix. In this case $M = D = diag(A)$, so $B_{Jac} = I - D^{-1}A$.

(c) For the 2D Poisson equation the stencil is

$$A : \frac{1}{h^2}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \text{ and } D : \frac{1}{h^2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

So the Jacobi iteration matrix: $B_{Jac} = I - D^{-1}A$ has the following stencil

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 1 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{bmatrix}$$

(d) In the damped Jacobi method a weighted average of the current iterant $\mathbf{u}^k$ and the full Jacobi step $\bar{\mathbf{u}}^{k+1,JAC}$ is computed. We denote the damping parameter by ω, and define the iterant resulting from the damped Jacobi method as

$$\mathbf{u}^{k+1} = (1-\omega)\,\mathbf{u}^k + \omega\,\bar{\mathbf{u}}^{k+1,JAC}\,. \tag{3}$$

Substituting the expression as give in part (a) with $M = D$ for $\bar{\mathbf{u}}^{k+1}$, we obtain that

$$\begin{aligned} \mathbf{u}^{k+1} &= (1-\omega)\,\mathbf{u}^k + \omega\,\mathbf{u}^k + \omega\,D^{-1}\mathbf{r}^k \qquad (4)\\ &= \mathbf{u}^k + \omega\,D^{-1}\mathbf{r}^k \end{aligned}$$

showing that the ω-damped Jacobi method is defined by

$$M_{JAC(\omega)} = \frac{1}{\omega}D \text{ and } B_{JAC(\omega)} = I - \omega D^{-1}A\,. \tag{5}$$

(e) When one starts with the zero vector the first iteration follows from $M\mathbf{u}^{(1)} = \mathbf{f}$. So we have to solve:

$$\begin{pmatrix} 2 & -1 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} u_1^{(1)} \\ u_2^{(1)} \\ u_3^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix}$$

After solving the lower triangular system the first iteration is equal to:

$$\begin{pmatrix} u_1^{(1)} \\ u_2^{(1)} \\ u_3^{(1)} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \\ 2 \end{pmatrix}$$

4. (a) The iterate $\mathbf{u}_1$ is written as $\mathbf{u}_1 = \alpha\mathbf{f}$ where α is a constant which has to be chosen such that $\|\mathbf{u} - \mathbf{u}_1\|_{A^TA}$ is minimal. This leads to $\|\mathbf{u} - \mathbf{u}_1\|_{A^TA}^2 = \|\mathbf{f} - A\mathbf{u}_1\|_2^2 = (\mathbf{f} - \alpha A\mathbf{f})^T(\mathbf{f} - \alpha A\mathbf{f}) = \mathbf{f}^T\mathbf{f} - 2\alpha(A\mathbf{f})^T\mathbf{f} + \alpha^2(A\mathbf{f})^TA\mathbf{f}$. The norm is minimized if $\alpha = \frac{(A\mathbf{f})^T\mathbf{f}}{(A\mathbf{f})^TA\mathbf{f}}$.

(b) Due to the definition of CGNR we know that the method computes an approximation $\mathbf{u}_k$ in the Krylov subspace $K^k(A^TA, A^T\mathbf{r_0})$ such that the norm $\|\mathbf{u} - \mathbf{u}_k\|_{A^TA}$ is minimal. It appears that

$$\|\mathbf{u} - \mathbf{u}_k\|_{A^TA} = \|A\mathbf{u} - A\mathbf{u}_k\|_2 = \|\mathbf{f} - A\mathbf{u}_k\|_2 = \|\mathbf{r}_k\|_2$$

Since in every iteration the dimension of the Krylov subspace will increase (except if 'lucky' breakdown occurs) one can conclude that the sequence $\|\mathbf{r}_k\|_2$ is monotone decreasing.

(c) We know that CG converges in one iteration if the 2-norm condition number of the iteration matrix is 1. For CGNR the iteration matrix is A^TA. If we choose the matrix such that $A^TA = I$, we know that the 2-norm condition number of A^TA is equal to 1 (this is called an orthogonal matrix). A 4 × 4 example is:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

(d) The following vectors should be stored in memory: $\mathbf{r}, \mathbf{u}, \tilde{\mathbf{r}}, \mathbf{w}, \mathbf{p}, \hat{\mathbf{v}}, \mathbf{q}, \hat{\mathbf{w}}$. Furthermore the matrices A and M should be stored in memory.

Per iteration two matrix vector products with A and two preconditioning vector products have to be computed. Next to that 2 inner products/norms, and 7 vector updates have to be computed. This is equal to $9 \times 2n$ flops.

(e) Per method at least 3 properties should be mentioned and / or compared. For the two methods the following properties are known:
CGNR: robust (only lucky breakdown), short recurrences, optimisation property, not based on the Krylov subspace $K^k(A, \mathbf{r_0})$, in general slow convergence since the condition number of A^TA is equal to the square of the condition number A.
CGS: not robust, short recurrences, no optimisation property, based on the Krylov subspace $K^k(A, \mathbf{r_0})$, in general fast convergence, rounding errors can become very large.

5. (a) The definition of the Galerkin Coarse Correction matrix is

$$A_H = I_h^H A_h I_H^h$$

Since A_h is symmetric and $I_H^h = (I_h^H)^T$ it easily follows that A_H is symmetric. To check that A_H is positive definite we consider a non-zero vector $\mathbf{x}_H$. Using the definition of A_H it follows that

$$\mathbf{x}_H^T A_H \mathbf{x}_H = \mathbf{x}_H^T I_h^H A_h I_H^h \mathbf{x}_H = \mathbf{y}^T A_h \mathbf{y},$$

where $\mathbf{y} = I_H^h \mathbf{x}_H$ is a non zero vector. Since A_h is SPD it follows that $\mathbf{y}^T A_h \mathbf{y} > 0$ which implies that A_H is an SPD matrix.

(b) Combining the definition of $A_H = I_h^H A_h I_H^h$ with $A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$

and $I_H^h = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$ leads to $A_H = \begin{pmatrix} 72 & 0 \\ 0 & 72 \end{pmatrix}$.

(c) Using the definition of the iteration matrix for a Basic Iterative Method (BIM) it follows that $B_{CGC} = I - I_H^h (A_H)^{-1} I_h^H A_h$.

(d) In the two grid method the convergence is a combination of the reduction of the error due to the smoothing operator and the Coarse Grid Correction operator. The eigenvalues 1 till 16 are reduced in size by the Coarse Grid Correction operator. The eigenvalues 17 till 32 are reduced in size by the smoothing operator. Comparing the eigenvalues in absolute value for the eigenvalues 17 till 32 shows that the choice $\omega = \frac{2}{3}$ gives the largest reduction, therefore $\omega = \frac{2}{3}$ gives the best smoother.

(e) To store solution vector $\mathbf{u}_h$ we need n memory positions, for $\mathbf{u}_{2h}$ we need $\frac{n}{2}$ memory positions, for $\mathbf{u}_{4h}$ we need $\frac{n}{4}$ memory positions, etc. In total we need

$$(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + ...)n = \frac{1}{1 - \frac{1}{2}} n = 2n$$

memory positions.