DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

ANSWERS OF THE TEST SCIENTIFIC COMPUTING (wi4201)
Wednesday January 22 2026, 13:30-16:30

This document contains 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) False. The condition number is defined as follows: $K_2(A) = \|A\|_2\|A^{-1}\|_2$. The inverse of I is equal to I. From the definition of the 2-norm it follows that

$$\|I\|_2 = \max_{\|\mathbf{u}\|_2=1} \|I\mathbf{u}\|_2 = 1$$

Combination of these expressions shows that $K_2(I) = 1$ which is not equal to 2.

(b) The matrix norm $\|A\|_{max}$ is defined as $\|A\|_{max} = \max_{1\le i,j\le n} |a_{i,j}|$. This norm does not have the multiplicative property. The multiplicative property holds if for any $R_1 \in \mathbb{R}^{m\times q}$ and $R_2 \in \mathbb{R}^{q\times n}$ $\|R_1R_2\|_p \le \|R_1\|_p \|R_2\|_p$ A counterexample is $R_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $R_2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$.

(c) This is not true. From Gershgorin's theorem it is easy to see that the spectral radius is less than or equal to 4.

(d) No.
Note that $\|A\|_2 = \max\limits_{1\le i\le n} |\lambda_i|$ where λ_i is an eigenvalue of A. Counter example

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$$

Note that $\max\limits_{1\le i\le n} \lambda_i = -1$, whereas $\|A\|_2 = 2$.

(e) True. Every multiplication with A leads to an extra zero diagonal. After n multiplications the resulting product is equal to the zero matrix.

2. (a) Expand the operator:

$$-\nabla \cdot (\alpha \nabla u) = -\alpha(u_{xx} + u_{yy}).$$

Write the second order part in the standard form

$$a\, u_{xx} + 2b\, u_{xy} + c\, u_{yy} + (\text{lower order terms}) = \cdots$$

Here $a = -\alpha$, $b = 0$, $c = -\alpha$, so

$$b^2 - ac = 0 - (-\alpha)(-\alpha) = -\alpha^2 < 0.$$

Hence the PDE is elliptic.

(b) Let grid points be $x_i = ih$, $y_j = jh$ for $i, j = 0, \ldots, n$. Unknowns are the interior values $u_{i,j} \approx u(x_i, y_j)$ for $i, j = 1, \ldots, n-1$ and boundary values are eliminated using $u = 0$ on $\partial\Omega$.

Use central differences for second derivatives. For each interior node (i, j),

$$-\alpha \left(\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} \right) + \sigma u_{i,j} = f_{i,j},$$

where $f_{i,j} = f(x_i, y_j)$.

Equivalently,

$$\frac{\alpha}{h^2}\Big(4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}\Big) + \sigma u_{i,j} = f_{i,j}.$$

Because the central difference approximations of u_{xx} and u_{yy} have truncation error $O(h^2)$, the full scheme has truncation error $O(h^2)$.

(c) For $n = 3$ we have $h = 1/3$ and interior indices are $i, j \in \{1, 2\}$, so there are $(n-1)^2 = 4$ unknowns. Use lexicographic ordering

$$u = \begin{pmatrix} u_{1,1} \\ u_{2,1} \\ u_{1,2} \\ u_{2,2} \end{pmatrix}, \qquad f = \begin{pmatrix} f_{1,1} \\ f_{2,1} \\ f_{1,2} \\ f_{2,2} \end{pmatrix}.$$

The discrete equations give

$$A = \begin{pmatrix} \frac{4\alpha}{h^2} + \sigma & -\frac{\alpha}{h^2} & -\frac{\alpha}{h^2} & 0 \\ -\frac{\alpha}{h^2} & \frac{4\alpha}{h^2} + \sigma & 0 & -\frac{\alpha}{h^2} \\ -\frac{\alpha}{h^2} & 0 & \frac{4\alpha}{h^2} + \sigma & -\frac{\alpha}{h^2} \\ 0 & -\frac{\alpha}{h^2} & -\frac{\alpha}{h^2} & \frac{4\alpha}{h^2} + \sigma \end{pmatrix}.$$

The matrix A is symmetric because each coupling between neighboring grid points appears with the same coefficient in both directions, so $a_{pq} = a_{qp}$.

(d) Gershgorin's theorem says:

> 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
>
> $$\rho_i = \sum_{j=1,\, j \neq i}^{N} |a_{ij}|,$$
>
> that is
>
> $$\lambda \in \sigma(A) \implies |a_{ii} - \lambda| \leq \rho_i.$$

Assume $\alpha = 1$ and $\sigma = 1$. Using the second order central differences on the uniform grid with meshwidth h, we obtain at an interior grid point (x_i, y_j)

$$-\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} - \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h^2} + u_{i,j} = f(x_i, y_j),$$

so that the linear system $Au = f$ has, for the unknown $u_{i,j}$,

$$a_{ii} = \frac{4}{h^2} + 1, \qquad a_{i,i\pm1} = a_{i,i\pm(n-1)} = -\frac{1}{h^2},$$

whenever the corresponding neighbor is an interior unknown. (Neighbors on $\partial\Omega$ are eliminated using $u = 0$.)

Applying Gershgorin with center $a_{ii} = \frac{4}{h^2} + 1$ gives the following disks (real intervals) depending on how many interior neighbors the node has:

- Interior nodes (4 interior neighbors):

$$\rho_i = \frac{4}{h^2}, \qquad \lambda \in \left[\left(\frac{4}{h^2} + 1\right) - \frac{4}{h^2},\ \left(\frac{4}{h^2} + 1\right) + \frac{4}{h^2}\right] = \left[1,\ 1 + \frac{8}{h^2}\right].$$

- Edge nodes excluding corners (3 interior neighbors):

$$\rho_i = \frac{3}{h^2}, \qquad \lambda \in \left[\left(\frac{4}{h^2} + 1\right) - \frac{3}{h^2},\ \left(\frac{4}{h^2} + 1\right) + \frac{3}{h^2}\right] = \left[1 + \frac{1}{h^2},\ 1 + \frac{7}{h^2}\right].$$

- Corner nodes of the interior grid (2 interior neighbors):

$$\rho_i = \frac{2}{h^2}, \qquad \lambda \in \left[\left(\frac{4}{h^2} + 1\right) - \frac{2}{h^2},\ \left(\frac{4}{h^2} + 1\right) + \frac{2}{h^2}\right] = \left[1 + \frac{2}{h^2},\ 1 + \frac{6}{h^2}\right].$$

Hence all eigenvalues satisfy

$$1 \le \lambda \le 1 + \frac{8}{h^2}.$$

(e) Now take $\alpha = 1$ and $\sigma = -1$. The discretization gives

$$a_{ii} = \frac{4}{h^2} - 1, \qquad a_{i,i\pm1} = a_{i,i\pm(n-1)} = -\frac{1}{h^2}$$

for existing interior neighbors.

Applying Gershgorin with center $a_{ii} = \frac{4}{h^2} - 1$ yields

- Interior nodes (4 interior neighbors):

$$\rho_i = \frac{4}{h^2}, \qquad \lambda \in \left[\left(\frac{4}{h^2} - 1\right) - \frac{4}{h^2},\ \left(\frac{4}{h^2} - 1\right) + \frac{4}{h^2}\right] = \left[-1,\ -1 + \frac{8}{h^2}\right].$$

- Edge nodes excluding corners (3 interior neighbors):

$$\rho_i = \frac{3}{h^2}, \qquad \lambda \in \left[\left(\frac{4}{h^2} - 1\right) - \frac{3}{h^2},\ \left(\frac{4}{h^2} - 1\right) + \frac{3}{h^2}\right] = \left[-1 + \frac{1}{h^2},\ -1 + \frac{7}{h^2}\right].$$

- Corner nodes of the interior grid (2 interior neighbors):

$$\rho_i = \frac{2}{h^2}, \qquad \lambda \in \left[\left(\frac{4}{h^2} - 1\right) - \frac{2}{h^2},\ \left(\frac{4}{h^2} - 1\right) + \frac{2}{h^2}\right] = \left[-1 + \frac{2}{h^2},\ -1 + \frac{6}{h^2}\right].$$

Hence all eigenvalues satisfy

$$-1 \le \lambda \le -1 + \frac{8}{h^2}.$$

(f) Recall that $\alpha > 0$. Let A_h denote the discrete operator for $-\Delta$ with elimination of boundary values, so that

$$A = \alpha A_h + \sigma I.$$

If μ is an eigenvalue of A_h with eigenvector v, then

$$Av = (\alpha A_h + \sigma I)v = (\alpha\mu + \sigma)v,$$

so every eigenvalue of A has the form $\lambda = \alpha\mu + \sigma$ with $\mu \in \sigma(A_h)$.
For standard Poisson and Dirichlet (A_h), we have

$$0 < \mu \le \frac{8}{h^2} \qquad \text{for all } \mu \in \sigma(A_h).$$

If $\sigma > 0$, then for all $\mu > 0$ we have $\lambda = \alpha\mu + \sigma > \sigma > 0$, hence $0 \notin \sigma(A)$ and A is nonsingular.
If $\sigma < -\frac{8\alpha}{h^2}$, then for all $\mu \le \frac{8}{h^2}$ we have

$$\lambda = \alpha\mu + \sigma \le \alpha\frac{8}{h^2} + \sigma < 0,$$

so again $0 \notin \sigma(A)$ and A is nonsingular.
For $\sigma \in \left[-\frac{8\alpha}{h^2}, 0\right]$, these bounds do not allow us to guarantee nonsingularity, because $\alpha\mu + \sigma$ can take values in an interval that contains 0.

3. (a) Let $Ax = f$ and let $x + \Delta x$ solve $A(x + \Delta x) = f + \Delta f$. Note that A is symmetric positive definite (SPD), which implies that A is nonsingular. Then

$$A\Delta x = \Delta f \quad \Rightarrow \quad \Delta x = A^{-1}\Delta f.$$

For any submultiplicative norm,

$$\|\Delta x\| \le \|A^{-1}\|\,\|\Delta f\| \le \delta\,\|A^{-1}\|\,\|f\|.$$

Since $f = Ax$, we have

$$\|f\| = \|Ax\| \le \|A\|\,\|x\|.$$

Combining gives

$$\frac{\|\Delta x\|}{\|x\|} \le \delta\,\|A^{-1}\|\,\|A\| = \delta\,\kappa(A).$$

(b) Since A is SPD, there exists an orthogonal matrix Q and a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i > 0$ such that

$$A = Q\Lambda Q^T.$$

Let $\lambda_{\min}(A) = \min_i \lambda_i$ and $\lambda_{\max}(A) = \max_i \lambda_i$.
By definition,

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

Write $x = Qy$ so that $\|y\|_2 = \|x\|_2 = 1$. Using orthogonality of Q,

$$\|Ax\|_2 = \|Q\Lambda Q^T x\|_2 = \|\Lambda y\|_2,$$

hence

$$\|A\|_2 = \max_{\|y\|_2=1} \|\Lambda y\|_2.$$

Moreover,

$$\|\Lambda y\|_2^2 = \sum_{i=1}^{n} \lambda_i^2 y_i^2 \le \lambda_{\max}(A)^2 \sum_{i=1}^{n} y_i^2 = \lambda_{\max}(A)^2,$$

so $\|A\|_2 \le \lambda_{\max}(A)$. Taking $y = e_k$ where $\lambda_k = \lambda_{\max}(A)$ gives equality, hence

$$\|A\|_2 = \lambda_{\max}(A).$$

Since $A^{-1} = Q\Lambda^{-1}Q^T$ with $\Lambda^{-1} = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_n)$, the same argument yields

$$\|A^{-1}\|_2 = \lambda_{\max}(A^{-1}) = \max_i \frac{1}{\lambda_i} = \frac{1}{\lambda_{\min}(A)}.$$

Therefore,

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

(c) From $(A + E)\hat{x} = f$ and $Ax = f$ we get $(A + E)\hat{x} = Ax$. Multiply by A^{-1} to obtain

$$(I + A^{-1}E)\hat{x} = x.$$

Let $M = A^{-1}E$. Then $\hat{x} = (I + M)^{-1}x$ and

$$\hat{x} - x = \big((I + M)^{-1} - I\big)x = -(I + M)^{-1}Mx.$$

Taking 2-norms gives

$$\frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \|(I + M)^{-1}\|_2 \, \|M\|_2.$$

Moreover,

$$\|M\|_2 \leq \|A^{-1}\|_2 \, \|E\|_2 \leq \varepsilon \, \|A^{-1}\|_2 \, \|A\|_2 = \varepsilon \, \kappa_2(A).$$

If $\varepsilon\kappa_2(A) < 1$, then $\|M\|_2 < 1$ and the Neumann series bound yields

$$\|(I + M)^{-1}\|_2 \leq \frac{1}{1 - \|M\|_2}.$$

Hence

$$\frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \frac{\|M\|_2}{1 - \|M\|_2} \leq \frac{\varepsilon \, \kappa_2(A)}{1 - \varepsilon \, \kappa_2(A)}.$$

(d) Using part (b) for the SPD matrix A, we have

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

With the assumed bounds,

$$\kappa_2(A) \leq \frac{c_2/h^2}{c_1} = \frac{c_2}{c_1} \, h^{-2} = O(h^{-2}).$$

Combining with (a), a relative perturbation of size δ in the right hand side can lead to a relative solution perturbation bounded by

$$\frac{\|\Delta u\|}{\|u\|} \leq \delta \, \kappa(A) = O(\delta \, h^{-2}).$$

If machine precision is fixed and δ is on the order of machine precision, then the error upper bound grows like h^{-2} as $h \to 0$. **Grading note:The proof in (d) is independent of (a) and follows directly from the statement already given in the question formulation (b) plus the given eigenvalue bounds, only the final interpretation sentence uses (a).**

4. (a) The linear system $A\mathbf{u} = \mathbf{f}$ can then be written as $M\mathbf{u} = N\mathbf{u} + \mathbf{f}$. By multiplying to the left and right by M^{-1} we can define an iterative scheme

$$\begin{aligned} \mathbf{u}^{k+1} &= M^{-1}N\mathbf{u}^k + M^{-1}\mathbf{f} \\ &= M^{-1}(M - A)\mathbf{u}^k + M^{-1}\mathbf{f} \\ &= \mathbf{u}^k + M^{-1}(\mathbf{f} - A\mathbf{u}^k) \\ &= \mathbf{u}^k + M^{-1}\mathbf{r}^k \end{aligned} \tag{1}$$

The recursion for the residual vector is given by

$$\begin{aligned} \mathbf{r}^{k+1} &= \mathbf{f} - A\mathbf{u}^{k+1} \\ &= \mathbf{f} - A\mathbf{u}^{k} - AM^{-1}\mathbf{r}^{k} \\ &= \mathbf{r}^{k} - AM^{-1}\,\mathbf{r}^{k} \\ &= (I - AM^{-1})\mathbf{r}^{k} \end{aligned} \tag{2}$$

(b) The error $\mathbf{e}^k$ and the residual $\mathbf{r}^k$ are related in the following way $\mathbf{r}^k = A\mathbf{e}^k$. We can use this relation to define an iterative scheme by the following sequence of three steps

- compute the *defect*: $\mathbf{r}^k = \mathbf{f} - A\mathbf{u}^k$;
- compute the approximate *correction* by solving the approximate residual equations: $\widehat{A}\widehat{\mathbf{e}}^k = \mathbf{r}^k$;
- add the correction to the previous iterand $\mathbf{u}^{k+1} = \mathbf{u}^k + \widehat{\mathbf{e}}^k$.

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

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

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:

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

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

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

5. (a) The iterate $\mathbf{u}_1$ is written as $\mathbf{u}_1 = \alpha_0 \mathbf{f}$ where α_0 is a constant which has to be chosen such that $\|\mathbf{u} - \mathbf{u}_1\|_{A^TA}$ is minimal. This leads to $\|\mathbf{u} - \mathbf{u}_1\|_{A^TA}^2 = \|\mathbf{f} - A\mathbf{u}_1\|_2^2 = (\mathbf{f} - \alpha_0 A\mathbf{f})^T(\mathbf{f} - \alpha_0 A\mathbf{f}) = \mathbf{f}^T\mathbf{f} - 2\alpha_0(A\mathbf{f})^T\mathbf{f} + \alpha_0^2(A\mathbf{f})^T A\mathbf{f}$. The norm is minimized if $\alpha_0 = \frac{(A\mathbf{f})^T\mathbf{f}}{(A\mathbf{f})^T A\mathbf{f}}$.

(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 3×3 example is:

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

(d) The following vectors should be stored in memory: $\mathbf{r}, \mathbf{p}, \mathbf{v}, \mathbf{s}, \mathbf{u}, \mathbf{z}, \mathbf{t}, \hat{\mathbf{u}}$. 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. The matrix vector product with A costs $9 \times n$ flops, whereas the solution of the lower and upper triangular matrix costs also $9 \times n$ flops in total $36 \times n$ flops. Next to that 5 inner products/norms, and 6 vector updates have to be computed. This is equal to $11 \times 2n$ flops. So per iteration $58 \times n$ flops are needed.

(e) Per iteration 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 $A^T A$ is equal to the square of the condition number A.
Bi-CGSTAB: not robust, short recurrences, no optimisation property, based on the Krylov subspace $K^k(A, \mathbf{r_0})$, in general fast convergence