DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

ANSWERS OF THE TEST SCIENTIFIC COMPUTING (wi4201)
Wednesday January 19 2022, 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) Yes. Since A is an orthogonal matrix we know that $A^TA = AA^T = I$. Since $\|A\|_2 = \sqrt{\lambda_{max}(A^TA)} = 1$ and $A^{-1} = A^T$ also $\|A^{-1}\|_2 = 1$. This implies that $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = 1$.

   (b) Yes, the Bi-CGSTAB method is a short-recurrency method. So in every iteration the same number of vectorupdates, inner products and matrix vector products are used, so the amount of work per iteration remains constant for all iterations.

   (c) Yes. It is not difficult to see from the definition of a permuation matrix that $P^TP = I$. The reason is as follows. Every row of P has only one non zero element. In the matrix multiplication we compute the inner product of the i-th row with the j-th row. This inner product is only non zero if $i = j$, and then the product is 1. This means that $P^{-1} = P^T$ so P^TAP is a similarity transformation of A and thus the spectra of A and P^TAP are the same.

   (d) No. It is easy to see that the matrix A is symmetric. So all eigenvalues are real valued. From the Gershgorin theorem it follows that $2 \leq \lambda \leq 6$, where λ is een eigenvalue of A. This implies that $\|A\|_2 = \lambda_{max}(A) \leq 6$.

   (e) No. There are various ways to show this. One is to compute the determinant of matrix A which is equal to zero. Another counter example is to multiply matrix A with vector $\mathbf{v}$, which components are all equal to 1. It follows that $A\mathbf{v} = \mathbf{0}$. This implies that matrix A has an eigenvalue equal to zero so matrix A is not invertible.

2. (a)

$$-\frac{\partial^2 u^{[kl]}(x,y)}{\partial x^2} - \frac{\partial^2 u^{[kl]}(x,y)}{\partial y^2} + cu^{[kl]}(x,y) = ((k\pi)^2 + (l\pi)^2 + c) * u^{[kl]}(x,y).$$

   Therefore $u^{[kl]}(x,y)$ is the eigenfunction of the operator. The eigenvalues are

$$((k\pi)^2 + (l\pi)^2 + c) \text{ for } k, l \in \mathbb{N}, k \neq 0 \text{ and } l \neq 0.$$

   (b) Denote $x_i = ih$, $y = jh$, and $()_{i,j}$ stands for the function value at (x_i, y_j). At the internal grid point (x_i, y_j), we have the problem satisfies:

$$-\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} - \left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} + cu_{ij} = f_{i,j} \tag{1}$$

We use the discretization $\frac{(4+ch^2)u_{i,j}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}}{h^2}$ to approximation the left-hand term in (1).

By using the Taylor expansion, we have

$$u_{i-1,j} = u_{i,j} + \left(\frac{\partial u}{\partial x}\right)_{i,j} * (-h) + \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} * \frac{(-h)^2}{2!} + \left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j} * \frac{(-h)^3}{3!} + O(h^4)$$
$$u_{i+1,j} = u_{i,j} + \left(\frac{\partial u}{\partial x}\right)_{i,j} * (h) + \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} * \frac{(h)^2}{2!} + \left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j} * \frac{(h)^3}{3!} + O(h^4)$$
$$u_{i,j-1} = u_{i,j} + \left(\frac{\partial u}{\partial y}\right)_{i,j} * (-h) + \left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} * \frac{(-h)^2}{2!} + \left(\frac{\partial^3 u}{\partial y^3}\right)_{i,j} * \frac{(-h)^3}{3!} + O(h^4)$$
$$u_{i,j+1} = u_{i,j} + \left(\frac{\partial u}{\partial y}\right)_{i,j} * (h) + \left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} * \frac{(h)^2}{2!} + \left(\frac{\partial^3 u}{\partial y^3}\right)_{i,j} * \frac{(h)^3}{3!} + O(h^4)$$

After summing the four equations above and $(4 + ch^2)u_{i,j}$, we have

$$\frac{(4 + ch^2)u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} = -\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} - \left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} + cu_{ij} + O(h^2)$$

The numerical method has a local truncation error of $O(h^2)$.

(c) The stencil located at $(x, y) = (1 - h, 1 - h)$ is

$$\frac{1}{h^2}\begin{bmatrix} 0 & 0 & 0 \\ -1 & 4 + ch^2 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

and the right-hand side is $f_{N-1,N-1} + \frac{1}{h^2}u_{N-1,N} + \frac{1}{h^2}u_{N,N-1} = f_{N-1,N-1} + \frac{2}{h^2}$.

(d) A_h is the matrix we got from the discretization with elimination of the boundary conditions. We need to verify that A_h is symmetric and all its eigenvalues are real and positive.
1. A_h is symmetric based on the stencil and symmetric matrix has real eigenvalues.
2. For matrix A_h, we have $|a_{i,i}| = \frac{4}{h^2} + c$, $\sum_{j=1,j\neq i}^{n} |a_{i,j}| \leq \frac{4}{h^2}$. Since $c > 0$, A_h is strict diagonal diagonal dominant.
3. By using the Gershgorin theorem, we have $0 < \lambda \leq \sum_{j=1,j\neq i}^{n} |a_{i,j}| + a_{i,i}$.
The positive definiteness of A_h then follows.

3. (a) Let $Au = f$, and $A = M - N$ where M is non singular. Derive a formula for u^k and r^k.

i.

$$\begin{aligned} u^{k+1} &= M^{-1}Nu^k + M^{-1}f \\ &= M^{-1}(M-A)u^k + M^{-1}f \\ &= u^k + M^{-1}(f - Au^k) \\ &= u^k + M^{-1}r^k \end{aligned} \tag{2}$$

ii.

$$\begin{aligned} r^{k+1} &= f - Au^{k+1} \\ &= f - A(u^k + M^{-1}r^k) \\ &= f - Au^k - AM^{-1}r^k \\ &= r^k - AM^{-1}r^k \\ &= (I - AM^{-1})r^k \end{aligned} \tag{3}$$

(b) Give the iteration matrix and a sufficient condition for convergence

i. The iteration matrix is given by $B = I - M^{-1}A$.

ii. There are three possible answers

A. $\rho(B) < 1$

B. $\|B\| < 1$

C. $\lim_{k\to\infty} \|B^k\|_2 = 0$

(c) Assume A is lower triangular, show that Gauss-Seidel converges.
Solution: Note that in this case $M = A$ and therefore

$$B = I - M^{-1}A = I - A^{-1}A = 0_{matrix}. \tag{4}$$

Then, $\|B\| < 1$ and therefore GS converges.

(d) Assume A is lower triangular, show Jacobi converges.
Now we have $M = D$, where D is the matrix containing only the diagonal elements of A.
Then $B = I - D^{-1}A = I - (I + L) = L$ where L is a lower diagonal matrix with only zeros on the diagonal. So B is a lower diagonal matrix with only zeros on the diagonal.
It then follows that $\lim_{k\to\infty} \|B^k\|_2 = 0$ so the Jacobi method converges.

(e) Below follow the 3 different stoping criteria and the properties.

i. $\|r^k\| \leq \epsilon$, this criterion is not scaling invariant.

ii. $\frac{\|r^k\|}{\|r^0\|} \leq \epsilon$, depends on goodness of initial guess.

iii. $\frac{\|r^k\|}{\|f\|} \leq \epsilon$, this is a good stopping criterion.

If the student has all three criteria and their properties correct, they are awarded: **2 pt.**
If the student has two out of three criteria and their properties correct, they are awarded: **1 pt.**.

4. (a) If A is SPD show that $< \mathbf{y}, \mathbf{z} >_A = \mathbf{y}^T A\mathbf{z}$ is an inner product.

i.) $< \mathbf{y}, \mathbf{z} >_A = < \mathbf{z}, \mathbf{y} >_A$

$$\begin{aligned} < \mathbf{y}, \mathbf{z} >_A &= \mathbf{y}^T A\mathbf{z} \\ &= \left(\mathbf{y}^T A\mathbf{z}\right)^T \quad \text{(transposition of a scalar)} \\ &= \mathbf{z}^T A^T \mathbf{y} \\ &= \mathbf{z}^T A\mathbf{y} \quad \text{(symmetry of A)} \\ &= < \mathbf{z}, \mathbf{y} >_A \end{aligned}$$

ii.) $< c\mathbf{y}, \mathbf{z} >_A = c < \mathbf{y}, \mathbf{z} >_A$

$$\begin{aligned} < c\mathbf{y}, \mathbf{z} >_A &= (c\mathbf{y})^T A\mathbf{z} \\ &= c\, \mathbf{y}^T A\mathbf{z} \\ &= c < \mathbf{y}, \mathbf{z} >_A \end{aligned}$$

iii.) $< \mathbf{y} + \mathbf{v}, \mathbf{z} >_A = < \mathbf{y}, \mathbf{z} >_A + < \mathbf{v}, \mathbf{z} >_A$

$$\begin{aligned} < \mathbf{y} + \mathbf{v}, \mathbf{z} >_A &= (\mathbf{y} + \mathbf{v})^T A\mathbf{z} \\ &= (\mathbf{y}^T + \mathbf{v}^T) A\mathbf{z} \\ &= \mathbf{y}^T A\mathbf{z} + \mathbf{v}^T A\mathbf{z} \\ &= < \mathbf{y}, \mathbf{z} >_A + < \mathbf{v}, \mathbf{z} >_A \end{aligned}$$

iv.) $< \mathbf{y}, \mathbf{y} >_A \geq 0$, and $< \mathbf{y}, \mathbf{y} >_A = 0 \Leftrightarrow \mathbf{y} = 0$

$$< \mathbf{y}, \mathbf{y} >_A = \mathbf{y}^T A\mathbf{y} \geq 0 \quad \forall\, \mathbf{y} \in \mathbb{R}^n, \text{ since } A \text{ SPD}$$

(b) We assume that $\mathbf{u}^1 = \alpha_0 \mathbf{r}^0$. Determine α_0 such that $\|\mathbf{u} - \mathbf{u}^1\|_A$ is minimal.

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}^1\|_A^2 &= (\mathbf{u} - \mathbf{u}^1)^T A(\mathbf{u} - \mathbf{u}^1) \\ &= \|\mathbf{u}\|_A - 2\alpha_0 < \mathbf{r}^0, \mathbf{u} >_A + \alpha_0^2 \|\mathbf{r}^0\|_A \end{aligned}$$

$$\frac{d}{d\alpha_0} \|\mathbf{u} - \mathbf{u}^1\|_A^2 = -2 < \mathbf{r}^0, \mathbf{u} >_A + 2\alpha_0 \|\mathbf{r}^0\|_A \tag{5}$$

Then we impose that 5 be equal zero to obtain:

$$\alpha_0 = \frac{< \mathbf{r}^0, \mathbf{u} >_A}{\|\mathbf{r}^0\|_A} = \frac{\mathbf{r}^0 \mathbf{f}}{\|\mathbf{r}^0\|_A}$$

(c) The matrix A corresponds to a shifted discretized Poisson operator. The eigenvalues are given by

$$\lambda_{k,l} = 6 - 2\cos\frac{\pi k}{61} - 2\cos\frac{\pi l}{61}, \quad 1 \leq k, l \leq 60.$$

Determine the linear rate of convergence for the Conjugate Gradient method. The matrix A is SPD, hence we can use:

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

to obtain:

$$\kappa_2(A) = \frac{\lambda_{60,60}(A)}{\lambda_{1,1}(A)} = \frac{9.994}{2.005} = 4.9845$$

So in terms of Theorem 5.5.1 of the lecture notes, the linear rate of convergence is:

$$\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} = 0.3813$$

(d) Based on the first paragraph of section 7.2 of the lecture notes, a preconditioner matrix M should satisfy the following:

- M is SPD,
- the eigenvalues of $M^{-1}A$ are clustered around 1,
- $M^{-1}y$ is obtainable at low cost.

The PCG method is obtained, given a suitable preconditioner $M = PP^T$, by applying the CG method to a preconditioned linear system $\tilde{A}\tilde{\mathbf{u}} = \tilde{\mathbf{y}}$, where $\tilde{A} = P^{-1}AP^{-T}$, $\mathbf{u} = P^{\ H\ RT}\tilde{\mathbf{u}}$ and $\tilde{\mathbf{y}} = P^{\ H\ R1}\mathbf{y}$, and P is a nonsingular matrix. This can also be rewritten such that CG is applied to the system $M^{-1}A\mathbf{u} = \mathbf{f}$.

(e) The eigenvalues of the matrix A are:

$$\lambda_1 = 1,\ \lambda_2 = 99,\ \lambda_3 = 101$$

A is symmetric and all its eigenvalues are positive, then A is SPD.

$$\kappa_2(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} = 101$$

$$\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} = 0.819$$

We give an estimate of convergence in terms of the number of iterations k needed to obtain:

$$\frac{\|\mathbf{u} - \mathbf{u}^k\|_A}{\|\mathbf{u} - \mathbf{u}^0\|_A} \leq 2(0.819)^k = 10^{-12} \tag{6}$$

thus k is at most:

$$k = \frac{\log\left(\frac{10^{-12}}{2}\right)}{\log(0.819)} = 141.85 \tag{7}$$

We consider a preconditioner where P is a diagonal matrix whose diagonal elements $p_{i,i} = \sqrt{a_{i,i}}$:

$$P = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

in order to estimate the convergence of the CG method we look at the eigenvalues of the matrix:

$$\tilde{A} = P^{-1}AP^{-T} = \begin{pmatrix} 1 & -\frac{1}{100} & 0 \\ -\frac{1}{100} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{8}$$

these eigenvalues are $\lambda_1 = 1$, $\lambda_2 = \frac{99}{100}$, $\lambda_3 = \frac{101}{100}$

$$\kappa_2(\tilde{A}) = \frac{\lambda_{max}(\tilde{A})}{\lambda_{min}(\tilde{A})} = \frac{101}{99}$$

$$\frac{\sqrt{\kappa_2(\tilde{A})} - 1}{\sqrt{\kappa_2(\tilde{A})} + 1} = 0.005$$

$$\frac{\|\mathbf{u} - \mathbf{u}^k\|_{\tilde{A}}}{\|\mathbf{u} - \mathbf{u}^0\|_{\tilde{A}}} \leq 2(0.005)^k = 10^{-12}$$

thus k is at most:

$$k = \frac{\log\left(\frac{10^{-12}}{2}\right)}{\log(0.005)} = 5.36$$

5. (a) We assume that the eigenvectors are given by $\mathbf{v}_j$ with the property that $\|\mathbf{v}_j\|_2 = 1$. From the assumption we know

$$\mathbf{q}_0 = \sum_{j=1}^{n} a_j \mathbf{v}_j \text{ with } a_1 \neq 0. \tag{9}$$

This implies that

$$A^k \mathbf{q}_0 = a_1 \lambda_1^k (\mathbf{v}_1 + \sum_{j=2}^{n} \frac{a_j}{a_1} (\frac{\lambda_j}{\lambda_1})^k \mathbf{v}_j) \tag{10}$$

Since $|\lambda_2| \geq ... \geq |\lambda_n|$, and $\|\mathbf{v}_j\|_2 = 1$ equation (10) implies:

$$\|\frac{A^k \mathbf{q}_0}{a_1 \lambda_1^k} - \mathbf{v}_1\|_2 \leq \sum_{j=2}^{n} \frac{|a_j|}{|a_1|} (\frac{|\lambda_j|}{|\lambda_1|})^k \|\mathbf{v}_j\|_2 \leq C(\frac{|\lambda_2|}{|\lambda_1|})^k = O(|\frac{\lambda_2}{\lambda_1}|^k) \tag{11}$$

To simplify notation we write (10) as

$$\frac{1}{\|\mathbf{q}_k\|_2}\mathbf{q}_k = \frac{1}{\|A^k\mathbf{q}_0\|_2}A^k\mathbf{q}_0 = \gamma(\mathbf{v}_1 + \mathbf{w}) \tag{12}$$

where $\gamma = \frac{a_1\lambda_1^k}{\|A^k\mathbf{q}_0\|_2}$ and vector $\mathbf{w}$ contains the remaining part and we know that $\|\mathbf{w}\|_2 = O(|\frac{\lambda_2}{\lambda_1}|^k)$. Since $\lambda^{(k)} = \frac{\mathbf{q}_k^T A\mathbf{q}_k}{\|\mathbf{q}_k\|_2^2}$ we obtain

$$\lambda^{(k)} = \gamma(\mathbf{v}_1 + \mathbf{w})^T A\gamma(\mathbf{v}_1 + \mathbf{w}) = (\gamma\mathbf{v}_1 + \gamma\mathbf{w})^T\gamma(\lambda_1\mathbf{v}_1 + A\mathbf{w}) \tag{13}$$

Due to (11) we obtain

$$\lambda^{(k)} = \lambda_1(\gamma\mathbf{v}_1)^T(\gamma\mathbf{v}_1) + O(|\frac{\lambda_2}{\lambda_1}|^k) \tag{14}$$

This leads to

$$\lambda^{(k)} = \lambda_1\|\gamma\mathbf{v}_1\|_2^2 + O(|\frac{\lambda_2}{\lambda_1}|^k) \tag{15}$$

From (12) it follows that

$$\gamma\mathbf{v}_1 = \frac{1}{\|\mathbf{q}_k\|_2}\mathbf{q}_k - \mathbf{w}$$

Thus

$$\|\gamma\mathbf{v}_1\|_2 = \|\frac{1}{\|\mathbf{q}_k\|_2}\mathbf{q}_k - \mathbf{w}\|_2 = \frac{1}{\|\mathbf{q}_k\|_2}\|\mathbf{q}_k\|2 + O(|\frac{\lambda_2}{\lambda_1}|^k) = 1 + O(|\frac{\lambda_2}{\lambda_1}|^k)$$

Combining this with (15) shows that

$$\lambda^{(k)} = \lambda_1 + O(|\frac{\lambda_2}{\lambda_1}|^k)$$

which implies:

$$|\lambda^{(k)} - \lambda_1| = O(|\frac{\lambda_2}{\lambda_1}|^k)$$

(b) Assume that $\mathbf{q}_{k-1} = \gamma\mathbf{v}_1 + \mathbf{w}$ where $\|\mathbf{w}\|_2 = \epsilon << 1$. From $\|\mathbf{q}_{k-1}\|_2 = 1$, $\|\mathbf{v}_1\|_2 = 1$ and $\|\mathbf{w}\|_2 = \epsilon$ it follows that $\gamma = 1 + O(\epsilon)$. Putting this in the algorithm shows that

$$\lambda^{(k)} = \mathbf{q}_{k-1}^T A\mathbf{q}_{k-1} = (\gamma\mathbf{v}_1+\mathbf{w})^T A(\gamma\mathbf{v}_1+\mathbf{w}) == (\gamma\mathbf{v}_1+\mathbf{w})^T(\gamma\lambda_1\mathbf{v}_1+A\mathbf{w}) = \gamma^2\lambda_1+O(\epsilon)$$

using the fact that $\gamma = 1 + O(\epsilon)$ shows that $\lambda^{(k)} = \lambda_1 + O(\epsilon)$.

(c) Two options are possible or based on the linear converging result, or based on the residual. For the first stopping criterion we can use:

$$\text{estimate} \quad r \quad \text{from} \quad \tilde{r} = \frac{|\lambda^{(k+1)} - \lambda^{(k)}|}{|\lambda^{(k)} - \lambda^{(k-1)}|}\,,$$

and stop if $\frac{\tilde{r}}{1-\tilde{r}}\frac{|\lambda^{(k+1)}-\lambda^{(k)}|}{|\lambda^{(k+1)}|} \leq \varepsilon$. Or the residual is small

$$\frac{\|\lambda^{(k)}\mathbf{q}_k - A\mathbf{q}_k\|_2}{|\lambda^{(k)}|} < \varepsilon$$

(d) The shift and invert power method is defined as follows: given a shift σ apply the powermethod to the matrix $(A-\sigma I)^{-1}$. The result converges to the in absolute value largest eigenvalue of λ_{min} of $(A-\sigma I)^{-1}$. The relation between the eigenvalues λ_j of A and the eigenvalues $\hat{\lambda}_j$ of $(A-\sigma I)^{-1}$ is as follows: $\hat{\lambda}_j = \frac{1}{\lambda_j-\sigma}$. The rate of convergence of the shift and invert power method is given by $\frac{2-\sigma}{2.1-\sigma}$. If $\sigma = 2$ the matrix $(A-\sigma I)^{-1}$ is singular so the method breaks down. A value of σ close to 2 will lead to fast convergence and there is no problem to compute the inverse of $(A-\sigma I)$.