DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

ANSWERS OF THE TEST SCIENTIFIC COMPUTING (wi4201)
Wednesday January 22 2020, 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) No.
   If the matrix is nonsymmetric in general the maximal absolute row sum of A is not equal to maximal absolute column sum of A. Counter example

$$A = \begin{bmatrix} 2 & 1 \\ 2 & 5 \end{bmatrix}$$

   (b) Give the definition of the Gershgorin disk.
   The Gershgorin disk for the first row has center 2 and radius 2.
   The Gershgorin disk for the second row has center 3 and radius 4.
   The Gershgorin disk for the third row has center 4 and radius 3.

   (c) First give the definition of the Krylov subspace. Note that the space is spanned by $\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, ..., A^{k-1}\mathbf{u}$ and since all these vectors are multiplies of $\mathbf{u}$, the dimension of the Krylov subspace is equal to 1.

   (d) True. For the proof we use the definition of the spectral radius: $\rho(A)$ is the in absolute value largest eigenvalue of A. We know that $|\lambda|\|\mathbf{u}\| = \|A\mathbf{u}\|$ for any eigenpair $\lambda, \mathbf{u}$. From the definition of a multiplicative norm $\|.\|$ it follows that $|\lambda|\|\mathbf{u}\| = \|A\mathbf{u}\| \leq \|A\|\|\mathbf{u}\|$. Division by $\|\mathbf{u}\|$ shows that $|\lambda| \leq \|A\|$ for any eigenvalue λ of A. So it also holds for the in absolute value largest eigenvalue of A which proves the result.

   (e) True. Every multiplication with A leads to an extra zero diagonal. After $n-1$ multiplications the resulting product is equal to the zero matrix.

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

$$\frac{1}{h^2}[-1 \;\; 2+\lambda h^2 \;\; -1]$$

   In order to show that the method is second order accurate, a Taylor expansion 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

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

(b) Use the goniometric formula's to show that

$$\lambda_k = \lambda + \frac{2}{h^2}[1 - \cos(\pi hk)] = \lambda + \frac{4}{h^2} sin^2(\frac{\pi hk}{2})$$

(c) Since the boundary conditions are not eliminated, the standard matrix is non symmetric. However after the connections to the boundary nodes are shifted to the right-hand side the matrix is symmetric. Since $\lambda > 0$ it follows from Gerschgorin's theorem, or from the explicit expression from the eigenvalues that all eigenvalues are positive. This is sufficient to conclude that the matrix is SPD.

(d) From

$$A(\mathbf{u} + \Delta\mathbf{u}) = \mathbf{f} + \Delta\mathbf{f}\,. \tag{1}$$

we can conclude that

$$A\Delta\mathbf{u} = \Delta\mathbf{f}\,, \tag{2}$$

from which $\Delta\mathbf{u} = A^{-1}\Delta\mathbf{f}$ and therefore $\|\Delta\mathbf{u}\| \leq \|A^{-1}\|\,\|\Delta\mathbf{f}\|$. From $A\mathbf{u} = \mathbf{f}$ it follows that $\|\mathbf{f}\| \leq \|A\|\,\|\mathbf{u}\|$ and therefore

$$\frac{1}{\|\mathbf{u}\|} \leq \|A\|\,\frac{1}{\|\mathbf{f}\|} \tag{3}$$

Combining these inequalities we arrive at the following bound on the norm of the perturbed solution

$$\boxed{\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \|A^{-1}\|\,\|A\|\,\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} = \kappa(A)\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \leq \delta\,\kappa(A)\,,} \tag{4}$$

where $\kappa(A)$ denotes the condition number of A measured in the norm $\|\cdot\|$.

For an SPD matrix the 2-norm condition number is equal to the ratio of the largest eigenvalue divided by the smallest eigenvalue. Again using Gershgorin or the answer of part (c) we can bound the largest eigenvalue by $\lambda + \frac{4}{h^2}$. This implies that $cond_2(A)$ is bounded by $1 + \frac{4}{\lambda h^2}$.

(e) As direct method the Cholesky decomposition for sparse matrices can be used. This only costs $O(n)$ flops if n is the number of gridpoints. Every iterative method costs at least the same amount of work per iteration. If bad convergence occurs (due to a large condition number) the number of iterations can be very large so an iterative method will cost much more work. So we prefer a direct method.

3. (a) No answer, since this exercise can also be asked as homework exercise.

(b) We have that

$$(M_{n-1} \cdot \ldots \cdot M_1)^{-1} = M_1^{-1} \ldots M_{n-1}^{-1} = \prod_{k=1}^{n-1} \left(I + \boldsymbol{\alpha}^{(k)}\mathbf{e}_k^T\right) = I + \sum_{k=1}^{n-1} \boldsymbol{\alpha}^{(k)}\mathbf{e}_k^T\,.$$

The result follows from the fact that for $1 \leq k \leq n-1$ we have that

$$\left(\boldsymbol{\alpha}^{(k)}\mathbf{e}_k^T\right)\left(\boldsymbol{\alpha}^{(k+1)}\mathbf{e}_{k+1}^T\right) = \boldsymbol{\alpha}^{(k)}\left(\mathbf{e}_k^T\,\boldsymbol{\alpha}^{(k+1)}\right)\mathbf{e}_{k+1}^T = \boldsymbol{\alpha}^{(k)}\left(0\right)\mathbf{e}_{k+1}^T = 0\,.$$

(c) From the construction of L and U it follows that there are only zeroes outside the band with bandwidth m. Within the band, elements which are zero in A become in general non-zero in L and U due to fill in.

(d) The LU decomposition determines an upper triangular matrix U and a lower triangular matrix L, with $l_{ii} = 1$, where $A = LU$. The procedure to obtain this decomposition is using Gauss transformations, such that column k is transformed in a such a way that all element $k+1, ..., n$ of this column become equal to zero. Assume that $n >> m$ for each column one needs $2m^2$ flops, because all elements outside the band are equal to zero. Since there are n columns the total costs is $n2m^2$ flops. In order to find solution u from $Au = f$, we substitute the decomposition into $Au = f$, so $LUu = f$. If we define $y = Uu$ we can first solve $Ly = f$ and then $Uu = y$. Since these systems are both triangular this is easy to solve. The work per solve step is $n2m$ flops.

4. (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 $\|\mathbf{u} - \mathbf{u}_1\|_{A^TA}^2 = \|\mathbf{f} - A\mathbf{u}_1\|_2^2 = (\mathbf{f} - \alpha_0 A\mathbf{f})^T(\mathbf{f} - \alpha_0 A\mathbf{f}) = \mathbf{f}^T\mathbf{f} - 2\alpha_0(A\mathbf{f})^T\mathbf{f} + \alpha_0^2(A\mathbf{f})^TA\mathbf{f}$. The norm is minimized if $\alpha_0 = \frac{(A\mathbf{f})^T\mathbf{f}}{(A\mathbf{f})^TA\mathbf{f}}$.

(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 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. Next to that 5 inner products/norms, and 6 vector updates have to be computed. This is equal to $11 \times 2n$ flops.

(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^TA 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

5. (a) It is easier to assume that $\mathbf{q}_{k-1} = \mathbf{v}_1 + \mathbf{w}$ with $\|\mathbf{w}\|_2 = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$. From the algorithm we know that $\lambda^{(k)} = \bar{\mathbf{q}}_{k-1}^T \mathbf{z}_k$, which is equal to $\lambda^{(k)} = \bar{\mathbf{q}}_{k-1}^T A \mathbf{q}_{k-1}^T = (\mathbf{v}_1 + \mathbf{w})^T(\mathbf{v}_1 + \mathbf{w}) = \lambda_1 \mathbf{v}_1^T \mathbf{v}_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$ In order to prove the result we have to show that $\mathbf{v}_1^T \mathbf{v}_1$ is close to 1. This can be shown as follows: $\mathbf{v}_1^T \mathbf{v}_1 = (\mathbf{q}_{k-1} - \mathbf{w})^T(\mathbf{q}_{k-1} - \mathbf{w}) = 1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$ This proves the result.

(b) For the shifted power method we apply the power method to the matrix $A - cI$. To obtain the original eigenvalue the result of this power method approximation should be shifted back by adding the value c. We know that for the shifted power method the convergence is determined by the ratio $\left|\frac{\lambda_2 - c}{\lambda_1 - c}\right|$ if we assume that $|\lambda_1 - c| > |\lambda_2 - c| \geq |\lambda_n - c|$. We obtain fast convergence if the ratio $\left|\frac{\lambda_2 - c}{\lambda_1 - c}\right|$ is as small as possible. This implies that $|\lambda_2 - c| = |\lambda_n - c|$. This leads to $c = \frac{\lambda_2 + \lambda_n}{2}$.

(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) To approximate the smallest eigenvalue where $\lambda_{n-1} = 1.1$ and $\lambda_n = 1$ the inverse power method is the method of choice. This means that the power method is applied to A^{-1}. If the shifted power method is used the convergence will be very slow $\frac{1001}{1001.1} = 0.9999$, whereas if the inverse power method is used the convergence is given by $\frac{1}{1.1} = 0.9091$. This is much faster.