DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

ANSWERS OF THE TEST SCIENTIFIC COMPUTING (wi4201)
Friday February 1 2019, 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.
Note that $\|A\|_2 = \max\limits_{1\leq i\leq n} |\lambda_i|$ where λ_i is an eigenvalue of A. Counter example

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

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

(b) Yes
From SPD it follows that $\mathbf{x}^T A\mathbf{x} > 0$. Take the following vectors: $\mathbf{x} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the i^{th} column of the identity matrix I. This implies that

$$\mathbf{e}_i^T A\mathbf{e}_i = a_{ii} > 0$$

which proves the claim.

(c) No
From $\mathbf{r} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2$ it follows that $A^k\mathbf{r} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2$. So all powers of A multiplied with $\mathbf{r}$ are element of the span of $\mathbf{v}_1, \mathbf{v}_2$. This implies that the dimension of $K^{10}(A, \mathbf{r})$ is at most equal to 2.

(d) Yes
Note that the following inequality is valid:

$$\|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^{n} (u_i)^2} \leq \sqrt{n \max_{1\leq i\leq n} |u_i|^2} = \sqrt{n}\|\mathbf{u}\|_\infty.$$

(e) No
$\|A\|_\infty$ is the maximal absolute row sum, and $\|A\|_1$ is the maximal absolute column sum. In general these are not equal. Counter example

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

$\|A\|_\infty = 2$, $\|A\|_1 = 1$.

2. (a) For the discretization we use the Finite Difference Method on the given grid. Using nearest neighbours we have for the internal nodes that

$$\frac{\partial^2 u}{\partial x^2}(x_i, y_j) = \frac{u^h_{i-1,j} - 2u^h_{i,j} + u^h_{i+1,j}}{h^2}\mathcal{O}(h^2) \text{ for } 2 \leq i, j \leq m \tag{1}$$

(and similar for the y-derivative). Using Taylor polynomials the claim that the error is $\mathcal{O}(h^2)$ should be shown. The approximation to the partial differential equation discretized on internal points of the grid can be written as

$$\frac{-u^h_{i,j-1} - u^h_{i-1,j} + 4u^h_{i,j} - u^h_{i+1,j} - u^h_{i,j+1}}{h^2} + u^h_{i,j} = f^h_{i,j} \text{ for } 2 \leq i, j \leq m\,. \tag{2}$$

(b) The stencil on the internal nodes is given by:

$$\frac{1}{h^2}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4+h^2 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{3}$$

The stencil in the lower left corner is:

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

(c) The matrix A has the following structure:

$$\frac{1}{h^2}\begin{pmatrix} T^h & -I & 0 & \dots & \dots & 0 \\ -I & T^h & -I & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & -I & T^h & -I \\ 0 & \dots & \dots & 0 & -I & T^h \end{pmatrix},$$

where T^h is given by

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

Finally, the bandwidth is equal to m.

(d) That the matrix is symmetric is easy to explain. This implies that all eigenvalues are real valued. Using Gershgorin's theorem for the matrix, implies that all eigenvalues are larger than 1. This implies that the matrix is SPD, see the line below Gershgorin's theorem in the lecture notes. Gershgorin's theorem is given

by: **(Gershgorin)** 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

$$\rho_i = \sum_{j=1, j\neq i}^{n} |a_{ij}| \tag{5}$$

i.e.,

$$\lambda \in \sigma(A) \Rightarrow \exists i \text{ such that } |a_{ii} - \lambda| \leq \rho_i. \tag{6}$$

(e) The best Krylov subspace method for SPD matrices is: Conjugate Gradients (CG). Motivation for this choice: CG has short recurrences, its approximation is an element of the Krylov subspace and is has a minimization property that the error is minimal measured in the A norm.

3. (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

$$\begin{aligned} \mathbf{u}^{k+1} &= M^{-1}N\mathbf{u}^k + M^{-1}\mathbf{f} \\ &= M^{-1}(M-A)\mathbf{u}^k + M^{-1}\mathbf{f} \\ &= \mathbf{u}^k + M^{-1}(\mathbf{f} - A\mathbf{u}^k) \\ &= \mathbf{u}^k + M^{-1}\mathbf{r}^k \end{aligned} \tag{7}$$

The recursion for the residual vector is given by

$$\begin{aligned} \mathbf{r}^{k+1} &= \mathbf{f} - A\mathbf{u}^{k+1} \\ &= \mathbf{f} - A\mathbf{u}^k - AM^{-1}\mathbf{r}^k \\ &= \mathbf{r}^k - AM^{-1}\,\mathbf{r}^k \\ &= (I - AM^{-1})\mathbf{r}^k \end{aligned} \tag{8}$$

(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) Note that the iteration matrix is given by: $B = I - M^{-1}A$. This implies that $B_{RIC} = I - \tau A$. The resulting stencil is:

$$[B_{RIC}] = \begin{bmatrix} \frac{\tau}{h^2} & 1 - 2\frac{\tau}{h^2} & \frac{\tau}{h^2} \end{bmatrix}$$

The stencil for M_{GS} is easy to find:

$$[M_{GS}] = \begin{bmatrix} -\frac{1}{h^2} & \frac{2}{h^2} & 0 \end{bmatrix}$$

However the stencil for B_{GS} is not easy to find.

(d) Using the definition of $B_{RIC} = I - \tau A$ it easily follows that the eigenvalues of B_{RIC} are given by the following expression: $\lambda_{RIC,i} = 1 - \tau\lambda_i$ where λ_i are the eigenvalues of A. For the maximal absolute eigenvalue of B_{RIC} (the spectral radius) we note that $\max_{1\leq i\leq n} |\lambda_{RIC,i}| = max\,[(1-\tau\lambda_1), (\tau\lambda_n - 1)]$. One obtains the best speed up if these values are equal, so $1 - \tau\lambda_1 = \tau\lambda_n - 1$. Solving this leads to the optimal value: $\tau = \frac{2}{\lambda_1+\lambda_n}$

(e) Determine the spectral radius of the Richardson iteration matrix. Note that $1 - \tau\lambda_1 = \frac{19}{20}$ and $1 - \tau\lambda_n = \frac{1}{2}$, so the spectral radius is $\frac{19}{20}$. So if $(\frac{19}{20})^k \leq 10^{-4}$ where k is the number of iterations we have satisfied the stopping criterion. So the answer is: $k = \frac{-4}{log(\frac{19}{20})} = 180$

4. (a) We take $\mathbf{u}^1 = \alpha_0\mathbf{r}^0$ where α_0 is a constant which has to be chosen such that $\|\mathbf{f} - A\mathbf{u}^1\|_A$ is minimal. This leads to

$$\|\mathbf{f}-A\mathbf{u}^1\|_A^2 = (\mathbf{f}-\alpha_0 A\mathbf{r}^0)^T A(\mathbf{f}-\alpha_0 A\mathbf{r}^0) = \mathbf{f}^T A\mathbf{f} - 2\alpha_0(A\mathbf{r}^0)^T A\mathbf{f} + \alpha_0^2(A\mathbf{r}^0)^T AA\mathbf{r}_0\ .$$

The norm given above is minimized if $\alpha_0 = \frac{(A\mathbf{r}^0)^T A\mathbf{f}}{(A\mathbf{r}^0)^T A^2\mathbf{r}^0}$.

(b) The optimality property of CG implies that the approximation $\mathbf{u}^k$ coming from CG satisfies:

$$\|\mathbf{u} - \mathbf{u}^k\|_A = \min_{\mathbf{y}\in K^k(A;\mathbf{r}^0)} \|\mathbf{u} - \mathbf{y}\|_A$$

If the method terminates before we reach $k = n$ we know that we have a 'lucky' breakdown so $\mathbf{u}^k = u$. If not we know that the dimension of K^n is equal to n, thus $K^n = \mathbb{R}^n$ and thus $\mathbf{u}^n = u$.

(c) The convergence of CG depends on the condition number. For SPD matrices the condition number is defined as

$$K_2(A) = \frac{\lambda_n}{\lambda_1}.$$

For a smaller condition number the convergence of CG is faster. Since $K_2(A_1) = 10$ and $K_2(A_2) = 200$, it is clear that we expect that the convergence for A_1 is much faster than for A_2.

(d) The superlinear convergence is given in the figure below: The explanation for

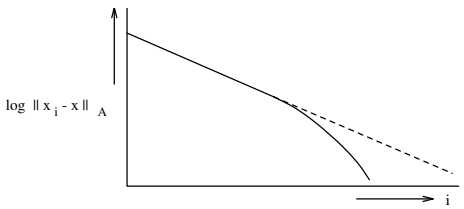

Figure 1: A superlinear convergent behavior

superlinear convergence is that initially the convergence is determined by the condition number $K_2(A) = \frac{\lambda_n}{\lambda_1}$. However after a number of iteration the effect of the smallest eigenvalue component is no longer influencing the convergence, so the convergence is determined by the effective condition number: $\frac{\lambda_n}{\lambda_2}$. So the convergence becomes faster and faster.

(e) The three properties are:

i. The matrix M should be SPD.

ii. the eigenvalues of $M^{-1}A$ should be clustered around 1, or the condition number of $M^{-1}A$ is (much) smaller than the condition number of A.

iii. it should be possible to obtain $M^{-1}\mathbf{y}$ at a low cost.

5. (a) In order to solve the linear system $A\mathbf{u} = \mathbf{f}$ with LU-decomposition without pivotting, we do the following steps:

- Find a lower triangular matrix L and an upper triangular matrix U, such that $LU = A$ and the diagonal elements of L are equal to 1.
- Solve $\mathbf{y}$ from $L\mathbf{y} = \mathbf{f}$.
- Solve $\mathbf{u}$ from $U\mathbf{u} = \mathbf{y}$.

Since L and U are triangular matrices, this solution process is easy to implement. For the derivation of the costs see the lecture notes. The answer for a full matrix is for the decomposition the cost is $\frac{2}{3}n^3$ and for both solution steps together $2n^2$.

(b) If we do the multiplication:

$$(I - \alpha^{(k)}\mathbf{e}_k^T)(I + \alpha^{(k)}\mathbf{e}_k^T)$$

we obtain the following:

$$I - \alpha^{(k)}\mathbf{e}_k^T + \alpha^{(k)}\mathbf{e}_k^T + \alpha^{(k)}\mathbf{e}_k^T\alpha^{(k)}\mathbf{e}_k^T =$$

$$I + \alpha^{(k)}\mathbf{e}_k^T\alpha^{(k)}\mathbf{e}_k^T$$

Due to the zero structure of $\mathbf{e}_k$ and $\alpha^{(k)}$ the product $\mathbf{e}_k^T\alpha^{(k)}$ is equal to zero, so the last term is equal to zero, so

$$(I - \alpha^{(k)}\mathbf{e}_k^T)(I + \alpha^{(k)}\mathbf{e}_k^T) = I$$

which proves the claim that $M_k^{-1} = I + \alpha^{(k)}\mathbf{e}_k^T$.

(c) The perturbed solution $\mathbf{u} + \Delta\mathbf{u}$ solves the system

$$A(\mathbf{u} + \Delta\mathbf{u}) = \mathbf{f} + \Delta\mathbf{f}\,. \tag{9}$$

Due to linearity, the perturbation $\Delta\mathbf{u}$ then solves the system

$$A\Delta\mathbf{u} = \Delta\mathbf{f}\,, \tag{10}$$

from which $\Delta\mathbf{u} = A^{-1}\Delta\mathbf{f}$ and therefore $\|\Delta\mathbf{u}\| \leq \|A^{-1}\|\,\|\Delta\mathbf{f}\|$. It follows form the multiplicative property that $\|\mathbf{f}\| \leq \|A\|\,\|\mathbf{u}\|$ and therefore

$$\frac{1}{\|\mathbf{u}\|} \leq \|A\|\,\frac{1}{\|\mathbf{f}\|} \tag{11}$$

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

$$\boxed{\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \|A^{-1}\|\,\|A\|\,\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} = \kappa(A)\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \leq \delta\,\kappa(A)\,,} \tag{12}$$

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

(d) 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.

(e) The difference is that we now look for a matrix G such that $A = GG^T$, where G is a lower triangular system. Advantages are:

i. Only half of the memory is needed.

ii. Only half of the amount of flops is needed.

iii. The method is stable so no pivotting is needed.