DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

**ANSWERS OF THE TEST SCIENTIFIC COMPUTING ( wi4201 )**
**Wednesday January 24 2024, 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) True. Since $Q$ is an orthogonal matrix we know that $QQ^T = Q^TQ = I$. Furthermore $\|A\|_2 = \sqrt{\lambda_{max}(A^TA)}$. This implies:

$$\|Q^TA\|_2 = \sqrt{\lambda_{max}(A^TQQ^TA)} = \sqrt{\lambda_{max}(A^TA)} = \|A\|_2$$

which proves the equality.

(b) 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 $\lambda$ of $A$. So it also holds for the in absolute value largest eigenvalue of $A$ which proves the result.

(c) False. From $\mathbf{r} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2$ it follows that $A^k\mathbf{r} = \lambda_1^k\mathbf{r}$. So all powers of $A$ multiplied with $\mathbf{r}$ are element of the span of $\mathbf{r}$. This implies that the dimension of $K^5(A, \mathbf{r})$ is equal to 1.

(d) True. The symmetry of $A_k$ easily follows from the symmetry of $A$. Since matrix $A$ is SPD we know that $\mathbf{v}^TA\mathbf{v} > 0$ for all $\mathbf{v} \neq \mathbf{0}$. Take $\mathbf{w} \in \mathbb{R}^k$ an arbitrary vector such that $\mathbf{w} \neq \mathbf{0}$. Define $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}(1) = \mathbf{w}(1), ...., \mathbf{v}(k) = \mathbf{w}(k)$ and $\mathbf{v}(k+1) = .... = \mathbf{v}(n) = 0$. It now follows that $\mathbf{w}^TA_k\mathbf{w} = \mathbf{v}^TA\mathbf{v} > 0$ which shows that the statement is true.

(e) True. The condition number is defined as follows: $K_p(A) = \|A\|_p\|A^{-1}\|_p$. The inverse of $I$ is equal to $I$. From the definition of the p-norm it follows that

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

Combination of these expressions shows that $K_p(I) = 1$.

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

$$\frac{1}{h^2}[-1 \;\; 2 - 10h^2 \;\; -1]$$

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

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

This leads to the stencil as given above.

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

$$\lambda_k = -10 + \frac{2}{h^2}[1 - \cos(\pi hk)] = -10 + \frac{4}{h^2} sin^2(\frac{\pi hk}{2})$$

(c) Since the boundary conditions are eliminated, the matrix is symmetric. Note that Gerschgorin's theorem does not imply that all eigenvalues are positive. Using the expression given in (b) with $N = 10$ and $k = 1$ shows that $\lambda_1 = -0.1314$. So eigenvalue $\lambda_1$ is negative with implies that for eigenvector $\mathbf{v}_1$ we have:

$$\mathbf{v}_1^T A \mathbf{v}_1 = \lambda_1 \|\mathbf{v}_1\|_2^2 < 0$$

This implies that the matrix is not SPD.

(d) We have the following ordering:

```
4 9 5
7 3 8
1 6 2
```

This leads to the following non-zero pattern:

```
* 0 0 0 0 * * 0 0
0 * 0 0 0 * 0 * 0
0 0 * 0 0 * * * *
0 0 0 * 0 0 * 0 *
0 0 0 0 * 0 0 * *
* * * 0 0 * 0 0 0
* 0 * * 0 0 * 0 0
0 * * 0 * 0 0 * 0
0 0 * * * 0 0 0 *
```

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

i.

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

ii.

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

(b) 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 upper triangular, show Jacobi converges.
Now we have $M = D$, where $D$ is the matrix containing only the diagonal elements of $A$.
Then $B = I - D^{-1}A = I - (I + U) = U$ where $U$ is an upper triangular matrix with zeros on the diagonal. So $B$ is an upper triangular matrix with only zeros on the diagonal.
It then follows that $B^{n-1} = 0_{matrix}$ so the Jacobi method converges.

(d) Assume $A$ is upper triangular, which Gauss Seidel variant is optimal for this matrix?
*Solution:* There are two Gauss Seidel variants: $M$ is equal to the lower triangular part of $A$ or $M$ is equal to the upper triangular part of $A$. For this matrix the final choice is optimal. The motivation is as follows. Note that in this case $M = A$ and therefore

$$B = I - M^{-1}A = I - A^{-1}A = 0_{matrix}. \tag{3}$$

Then, $\|B\| < 1$ and therefore this variant of GS converges. Furthermore $e^1 = Be^0 = 0_{vector}$, so the method converges after 1 iteration.

(e) Below 3 different stopping criteria and their properties are given.

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

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

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

4. (a) We take $\mathbf{u}^1 = \alpha\mathbf{r}^0$ where $\alpha$ is a constant which has to be chosen such that $\|\mathbf{u} - \mathbf{u}^1\|_2$ is minimal. This leads to

$$\|\mathbf{u} - \mathbf{u}^1\|_2^2 = (\mathbf{u} - \alpha\mathbf{r}^0)^T(\mathbf{u} - \alpha\mathbf{r}^0) = \mathbf{u}^T\mathbf{u} - 2\alpha(\mathbf{r}^0)^T\mathbf{u} + \alpha^2(\mathbf{r}^0)^T\mathbf{r}^0 \ .$$

The norm given above is minimized if $\alpha = \frac{(\mathbf{r}^0)^T\mathbf{u}}{(\mathbf{r}^0)^T\mathbf{r}^0}$.

(b) Note that $\mathbf{u}$ is needed in the definition of $\alpha$ as given in part (a). If one uses the A-norm, the following expression should be minimal: $\|\mathbf{u} - \mathbf{u}^1\|_A$. Using the same steps as in part (a) it appears that now $\alpha$ is given by:

$$\alpha = \frac{(\mathbf{r}^0)^T A\mathbf{u}}{(\mathbf{r}^0)^T A\mathbf{r}^0} = \frac{(\mathbf{r}^0)^T\mathbf{f}}{(\mathbf{r}^0)^T A\mathbf{r}^0},$$

which is easy to compute.

(c) 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 $\mathbf{u}^k = \mathbf{u}$. If not we know that the dimension of $K^n$ is equal to $n$, thus $K^n = \mathbb{R}^n$ and thus $\mathbf{u}^n = \mathbf{u}$.

(d) 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 $A_2$.

(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 low cost.

5. (a) 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 $M_k^{-1} = I + \alpha^{(k)}\mathbf{e}_k^T$.

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

$$A(\mathbf{u} + \Delta\mathbf{u}) = \mathbf{f} + \Delta\mathbf{f}\,. \tag{4}$$

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

$$A\Delta\mathbf{u} = \Delta\mathbf{f}\,, \tag{5}$$

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

$$\frac{1}{\|\mathbf{u}\|} \leq \|A\|\,\frac{1}{\|\mathbf{f}\|} \tag{6}$$

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

$$\boxed{\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \|A^{-1}\|\,\|A\|\,\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} = \kappa(A)\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \leq \delta\,\kappa(A)\,,} \tag{7}$$

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

(c) 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. Since the non-zero pattern of $L$ and $U$ is the same as that of $A$, the number of computations per row are: one division to compute the multiplier followed by a multiplication and addition/subtraction to compute the diagonal element of $U$. This leads to $3n$ flops for the decomposition.

In order to find solution $\mathbf{u}$ from $A\mathbf{u} = \mathbf{f}$, we substitute the decomposition into $A\mathbf{u} = \mathbf{f}$, so $LU\mathbf{u} = \mathbf{f}$. If we define $\mathbf{y} = U\mathbf{u}$ we can first solve $L\mathbf{y} = \mathbf{f}$ and then $U\mathbf{u} = \mathbf{y}$. Since these systems are both triangular this is easy to solve. The work for the first system is one multiplication and addition/subtraction to compute one component of $\mathbf{y}$, so in total $2n$ flops for the first system. For the second system one needs a multiplication and addition/subtraction followed by a division to compute one component of $\mathbf{u}$, so in total $3n$ flops for the second system. In total $8n$ flops are needed to solve a system with a tri-diagonal matrix $A$.

(d) After 1 step of the Gaussian elimination process we obtain the following matrix:

$$\begin{pmatrix} 4 & -1 & 0 & 0 & -1 \\ 0 & 3\frac{3}{4} & -1 & 0 & -\frac{1}{4} \\ 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 4 & -1 \\ 0 & -\frac{1}{4} & 0 & -1 & 3\frac{3}{4} \end{pmatrix}$$

Note that the fill in less than $\frac{1}{4}$ in absolute value.