DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

ANSWERS OF THE TEST SCIENTIFIC COMPUTING (wi4201)
Friday February 2 2018, 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) True. Since A is SPD it is known that all the eigenvalues of A are real valued and positive. This implies that the inverse of A exists. Suppose that λ and v are an eigenpair, so $Av = \lambda v$. It easily follows that $A^{-1}v = \frac{1}{\lambda}v$. Since A^{-1} is also SPD we know that $\|A^{-1}\|_2$ is equal to the largest eigenvalue of A^{-1}, which is equal to $\frac{1}{\lambda_{min}}$.

   (b) True. Note that all matrices are nonsingular. The condition number is defined as: $\text{cond}_2(A) = \|A\|_2\|A^{-1}\|_2$. Since the 2-norm is a multiplicative norm we have:
   $$\|AB\|_2 \leq \|A\|_2\|B\|_2$$
   and
   $$\|(AB)^{-1}\|_2 \leq \|A^{-1}\|_2\|B^{-1}\|_2$$
   Combination of these inequalities, together with the definition of $\text{cond}_2(A)$ shows that the given inequality is true.

   (c) False. A counterexample is choose for R_1 a nonzero vector of length n and for R_2 a nonzero vector of length n which is perpendicular to R_1. Then the product $R_1^T R_2$ is zero, so the product is not positive definite.

   (d) True. Note that $\|u\|_\infty = \max_{1\leq i\leq n}|u_i|$. Since $\|u\|_1 = |u_1| + ... + |u_n|$. It follows easily that $\|u\|_1 \leq n\|u\|_\infty$. Using the vector, which components are all 1 it appears that $\|u\|_1 = n\|u\|_\infty$, so the bound is sharp.

   (e) True. It follows easily from Gershgorin's law that the minimal real part of the eigenvalues is larger than $a_{ii} - a_{ii} = 0$.

2. (a) The finite difference stencil is given by
   $$\frac{1}{h^2}[-1 \;\; 2 + 4h^2 \;\; -1]$$
   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

$$\lambda_k = 4 + \frac{4}{h^2} sin^2(\frac{\pi h k}{2})$$

(c) If the answer of the previous part is correct, it is easy to see that all the eigenvalues are larger than or equal to 4. Using the stencil notation and the theorem of Gershgorin it also follows that all the eigenvalues are larger than or equal to 4 independent of h.

(d) The 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 $4 + \frac{4}{h^2}$. This implies that $cond_2(A)$ is bounded by $1 + \frac{1}{h^2}$.

(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) 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. This costs $\frac{2}{3}n^3$ 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 n^2 flops.

(b) If we assume that A is not affected by the perturbation, the perturbed solution $u + \Delta u$ solves the system

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

Due to linearity, the perturbation Δu then solves the system

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

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

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

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

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

(c) We have that

$$(M_{n-1} \cdot \ldots \cdot M_1)^{-1} = M_1^{-1} \ldots M_{n-1}^{-1} = \prod_{k=1}^{n-1} \left(I + \boldsymbol{\alpha}^{(k)} \mathbf{e}_k^T\right) = I + \sum_{k=1}^{n-1} \boldsymbol{\alpha}^{(k)} \mathbf{e}_k^T \,. \quad (5)$$

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

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

(d) If $a_{kk}^{(k-1)}$ is equal to zero we consider all elements $a_{jk}^{(k-1)}$ for all $j = k, ..., n$. Suppose that $\hat{j}$ is such that $a_{\hat{j}k}^{(k-1)}$ is in absolute value the largest element in this column. Then $a_{\hat{j}k}^{(k-1)}$ should be nonzero, otherwise the determinant of $A^{(k-1)}$ is zero which implies that the determinant of A is zero. This contradicts the fact that A is nonsingular. Now change the $\hat{j}$ and the kth row, such that the pivot is nonzero, and continue the process.

(e) 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.

4. (a) Given that a non-singular matrix M exists we can split A as follows: $A = M - N$. The residual is defined by $r^k = f - Au^k$. For the iteration formula we have:

$$Mu^{k+1} = Nu^k + f$$

This can be rewritten as:

$$Mu^{k+1} = Mu^k - Mu^k + Nu^k + f = Mu^k - Au^k + f.$$

Multiplication with M^{-1} and using the definition of the residual gives: $u^{k+1} = u^k + M^{-1} r^k$.

(b) The error is defined as $e^k = u - u^k$. Since $Au = (M - N)u = f$ we have that:

$$Mu = Nu + f$$

For the iteration formula we have:

$$Mu^{k+1} = Nu^k + f$$

combination gives:

$$M(u^{k+1} - u) = N(u^k - u)$$

Since $N = M - A$, this can also be written as: $Me^{k+1} = (M - A)e^k$. Multiplication with M^{-1} shows that

$$e^{k+1} = (I - M^{-1}A)e^k.$$

This holds for all splitting, so also for the Jacobi iteration matrix. In this case $M = D = diag(A)$, so $B_{Jac} = I - D^{-1}A$.

(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 u^k and the full Jacobi step $\bar{u}^{k+1,JAC}$ is computed. We denote the damping parameter by ω, and define the iterant resulting from the damped Jacobi method as

$$u^{k+1} = (1 - \omega)\, u^k + \omega\, \bar{u}^{k+1,JAC}\,. \tag{6}$$

Substituting the expression as give in part (a) with $M = D$ for $\bar{u}^{k+1}$, we obtain that

$$\begin{aligned} u^{k+1} &= (1 - \omega)\, u^k + \omega\, u^k + \omega\, D^{-1} r^k \\ &= u^k + \omega\, D^{-1} r^k \end{aligned} \tag{7}$$

showing that the ω-damped Jacobi method is defined by

$$M_{JAC(\omega)} = \frac{1}{\omega} D \text{ and } B_{JAC(\omega)} = I - \omega D^{-1} A\,. \tag{8}$$

(e) When one starts with the zero vector the first iteration follows from $Mu^{(1)} = f$. So we have to solve:

$$\begin{pmatrix} 2 & 0 & 0 \\ -1 & 2 & 0 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1^{(1)} \\ u_2^{(1)} \\ u_3^{(1)} \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}$$

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

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

5. (a) The iterate u^1 is written as $u^1 = \alpha_0 r^0$ where α_0 is a constant which has to be chosen such that $\|u - u^1\|_2$ is minimal. This leads to $\|u - u^1\|_2^2 = (u - \alpha_0 r^0)^T (u - \alpha_0 r^0) = u^T u - 2\alpha_0 (r^0)^T u + \alpha_0^2 (r^0)^T r_0$. The norm is minimized if $\alpha_0 = \frac{(r^0)^T u}{(r^0)^T r^0}$.

   (b) We check the three properties:

$$< u, v >_A = u^T A v = (u^T A v)^T = v^T A^T u = v^T A u = < v, u >_A$$

   It easily follows that

$$< cu, v >_A = c u^T A v = c < u, v >_A .$$

$$< u + v, w >_A = (u + v)^T A w = u^T A w + v^T A w = < u, w >_A + < v, w >_A$$

   The final property: $< u, u >_A \geq 0$ with equality only for $u = 0$ follows from the fact that A is SPD.

   (c) The rate of convergence of CG is given by $\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}$. The largest eigenvalue is equal to 4 and the smallest eigenvalue is given by $4 - 4\cos\frac{\pi}{31} = 0.0205227$ so $\kappa_2(A) = 195$ so the rate is equal to: 0.8666.

   (d) This is due to the superlinear behaviour of CG. Initially the rate of convergence is determined by the ratio of the largest and the smallest eigenvalue. After some iterations the convergence is determined by the ratio of the largest and the one but smallest eigenvalue. This leads to a faster and faster rate of convergence. Graphically this means that in a double log scale the convergence is faster than a straight line.

   (e) i. CGNR: CG applied to $A^T A u = A^T$, short recurrences, optimal but not based on the Krylov subspace.

   ii. BiCG: short recurrences, not optimal and based on the Krylov subspace.

   iii. GMRES, long recurrences, optimal and based on the Krylov subspace.