WI4201 Written Exam
February 3rd, 2017

This examination consists of the following **5 questions**. *Read the questions* **carefully** *before answering. With each question ten credit points can be gained. This exam is an closed book exam. You are allowed a sufficiently simple calculator. You are* **not** *allowed to use any book or notes.*

Question 1 (10 pnts. - 2 pnts. per subquestion)
Answer the following questions

1. assume A to be the 2-by-2 matrix
$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} .$$
Give the two Gershgorin disks that contain the eigenvalues of the matrix A;

   D_1 has center 1 and radius 2. D_2 has center 4 and radius 3.

2. give an example of a 2-by-2 non-triangular matrix A that is an M-matrix;

   one can take for instance four real positive numbers $a, b, c, d \in \mathbb{R}^+$ such that $a\,d - b\,c \neq 0$, form
$$B = \begin{pmatrix} a & b \\ c & d \end{pmatrix} ,$$
   and set
$$A = B^{-1} = \frac{1}{a\,d - b\,c} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} .$$
   Then $A^{-1} = B \geq 0$ elementwise, and is therefore an M-matrix.

3. assume given an n-by-n real-valued matrix $A \in \mathbb{R}^{n \times n}$ and assume $\mathbf{u}$ to be an eigenvector of A with eigenvalue λ. Assume k to be a positive integer. Then the Krylov subspace $K^k(A, \mathbf{u})$ is a subspace in $\mathbb{R}^n$. Give the number of distinct basis requires to span this space, i.e., give the dimension of this subspace;

   The Krylov subspace $K^k(A, \mathbf{u})$ is the subspace defined as
$$K^k(A, \mathbf{u}) = span\{\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, \ldots, A^{k-1}\mathbf{u}\} .$$
   Given that $A\mathbf{u} = \lambda\mathbf{u}$, this definition reduces to
$$K^k(A, \mathbf{u}) = span\{\mathbf{u}, \lambda\mathbf{u}, \lambda^2\mathbf{u}, \ldots, \lambda^{k-1}\mathbf{u}\} = span\{\mathbf{u}\} .$$
   This Krylov subspace $K^k(A, \mathbf{u})$ is thus one-dimensional.

4. give a distinct advantage of the CGS method over the BiCG method;

   In solving the linear system $A\,x = b$, the BiCG method requires matrix-vectors multiplications with both A and A^T. The CGS method requires matrix-vector multiplications with A only. This can be a big advantage in real live situations in which an efficient matrix-vector multiplication with both A and A^T is hard to realize.

5. show that the smallest eigenvalue of A and the largest eigenvalues of A^{-1} are equal. This question intended to ask the following: show that if A is SPD and if $\lambda_{min}(A)$ is the smallest eigenvalue of A, then $1/\lambda_{min}(A)$ is the largest eigenvalue of A^{-1};

   if A is SPD and has eigenvalues $\{\lambda_1 = \lambda_{min}, \lambda_2, \ldots, \lambda_{n-1}, \lambda_n = \lambda_{max}\}$, then A^{-1} has eigenvalues $\{1/\lambda_n, 1/\lambda_{n-1}, \ldots, 1/\lambda_2, 1/\lambda_1\}$. Indeed, if $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$, then $A^{-1}\mathbf{u}_i = (1/\lambda_i)\mathbf{u}_i$.

Question 2 (10 pnts. - 2 pnt. per subquestion)

In this assignment we consider the finite difference discretization of a diffusion equation with a spatially varying diffusion coefficient on a one-dimensional domain. More specifically, given the domain $x \in \Omega = (0,1)$, given the source function $f(x)$ and given the strictly positive diffusion coefficient $c(x) > 0 \; \forall x$, we consider finding the unknown function $u(x)$ such that $u(x)$ is a solution of the differential equation

$$-\frac{d}{dx}\left[c(x)\frac{du}{dx}\right] = f(x) \text{ for } 0 < x < 1 \text{ with } c(x) > 0 \tag{1}$$

supplied with homogeneous Dirichlet boundary conditions on both end points, i.e., we impose

$$u(0) = 0 \text{ and } u(1) = 1\,. \tag{2}$$

As an aid in answering the questions that follow, we rewrite (1) in the following equivalent form

$$-\frac{d\,\Gamma(x)}{dx} = f(x) \text{ for } 0 < x < 1 \text{ where } \Gamma(x) = c(x)\frac{du}{dx}\,. \tag{3}$$

For the finite difference discretization we consider on Ω a uniform mesh with N elements and a meshwidth $h = 1/N$. We will denote the mesh nodes as $x_i = (i-1)h$ for $1 \le i \le N+1$. In this way the nodes $x_1 = 0$ and $x_{N+1} = 1$ coincide with the left and right end point of Ω, respectively. We also consider the midpoints $x_{i+1/2} = [x_i + x_{i+1}]/2$ for $1 \le i \le N$. The finite difference discretization can be performed in two steps.

1. In the first step the derivative $d\,\Gamma/dx$ in the node $x = x_i$ can be discretized using values of $\Gamma(x)$ in the nodes $x_{i-1/2}$ and $x_{i+1/2}$;
2. In the second step the derivative du/dx in the node $x = x_{i-1/2}$ ($x_{i+1/2}$) can be discretized using values of $u(x)$ in the nodes x_{i-1} and x_i (x_i and x_{i+1}).

Let $A^h\mathbf{u}^h = \mathbf{f}^h$ denote the resulting linear system. Answer the following questions

1. give the finite difference stencil of the matrix A^h corresponding to a grid point that is neither a boundary point nor a point connected to the left or right boundary point. The element of this stencil are functions of the meshwidth h and the function $c(x)$;

   given that

$$\begin{aligned}
-\frac{d\,\Gamma(x)}{dx}|_{x=x_i} &\approx -\frac{\Gamma(x_{i+1/2}) - \Gamma(x_{i-1/2})}{h} \;\; \textit{(central approx. of } \Gamma(x)\textit{)} \\
&\approx -\frac{c(x_{i+1/2})\,u'(x_{i+1/2}) - c(x_{i-1/2})\,u'(x_{i+1/2})}{h} \;\; \textit{(definition of } \Gamma(x)\textit{)} \\
&= -\frac{c(x_{i+1/2})\frac{u(x_{i+1})-u(x_i)}{h} - c(x_{i-1/2})\frac{u(x_i)-u(x_{i-1})}{h}}{h} \;\; \textit{(central approx. of u'(x)} \\
&= \frac{-c(x_{i-1/2})u(x_{i-1}) + [c(x_{i-1/2}) + c(x_{i+1/2})]u(x_i) - c(x_{i+1/2})u(x_{i+1})}{h^2} \;\; \textit{(rearranging)}
\end{aligned}$$

   the stencil is given by

$$[A] = \frac{1}{h^2}\begin{bmatrix} -c(x_{i-1/2}) & [c(x_{i-1/2}) + c(x_{i+1/2})] & -c(x_{i+1/2}) \end{bmatrix}\,.$$

   This stencil is a short-hand notation for the i-th row of the matrix. Observe that this stencil reduces to the more commonly used stencil $[A] = \frac{1}{h^2}\begin{bmatrix} -1 & 2 & -1 \end{bmatrix}$ *in case that* $c(x) = 1$ $\forall x \in \Omega = (0,1)$.

2. assume that the boundary conditions are eliminated from the linear system and give the matrix A^h for $N = 3$ (and thus $h = 1/3$);

   with $N = 3$, the mesh contains 4 grid points, of which 2 are boundary (left and right) points and 2 are interior points. After elimination of the boundary conditions, the matrix A^h is thus a 2-by-2 matrix with entries given by

$$A^h = \frac{1}{h^2} \begin{pmatrix} c(x_{1/2}) + c(x_{3/2}) & -c(x_{3/2}) \\ -c(x_{3/2}) & c(x_{3/2}) + c(x_{5/2}) \end{pmatrix} .$$

   This matrix is symmetric. Given that $c(x) > 0 \; \forall x \in \Omega$, the diagonal elements are positive. The matrix A is diagonally dominant. By the Gershgorin theorem, the diagonal dominance combined with the positive diagonal entries imply that A is positive definite.

3. use the fact that $c(x) > 0 \; \forall x$ to show that the matrix A^h is symmetric and positive definite;

   The symmetry of A^h follows from its stencil. Indeed, the stencil implies that A is tridiagonal. To show that A is symmetric, it suffices to show that $A(i, i+1) = A(i+1, i)$ for all $1 \leq i \leq n-1$. The value of $A(i, i+1)$ is the weight of the connection that node i has with its right neighbour $i+1$ and is by the stencil equal to $-c(x_{i+1/2})$. The value of $A(i+1, i)$ is the weight of the connection that node $i+1$ has with its left neighbour i. This weight is given by the stencil for the $(i+1)$st row and is given by $-c(x_{i+1/2})$. Thus clearly $A(i, i+1) = A(i+1, i)$.

   The positive definiteness of A^h follows from the fact that A^h has positive diagonal entries and that A^h is diagonally dominant. The diagonal dominance can be seen from the stencil. Gershgorin theorem then constraints the eigenvalues of A^h to lie in the right half of the complex plane. The symmetry of A^h constraints the eigenvalues to be real-valued. The positive definiteness of A^h then follows.

4. show that the method of Jacobi converges for this linear system;

   One of the following three proofs can be given.
   A first proof consists in observing that A^h is an irreducible K-matrix and therefore an M-matrix. Given that $c(x) > 0 \; \forall x \in \Omega$, the method of Jacobi defines a splitting $A^h = M - N$ such that $M^{-1} \geq 0$ (element wise) and $N \geq 0$ (element wise). This splitting is therefore regular. A regular splitting of an M-matrix necessarily converges. The method of Jacobi applied to the A^h thus converges.
   A second proof consists in observing that A^h is diagonally dominant. A^h is actually strictly diagonal dominant due to the handling of the boundary conditions in its first and last row. The Jacobi iteration matrix $B_{JAC} = I - M^{-1}A^h$ is therefore bounded in norm (1-norm or ∞-norm) by one. The strict diagonal dominance leads to a bound strictly lower than one. The method of Jacobi applied to the A^h thus converges.
   A third proof consists in observing that the stencil of the Jacobi iteration matrix $B_{JAC} = I - M^{-1}A^h$ is given by

$$[B_{JAC}] = \left[\frac{c(x_{i-1/2})}{c(x_{i-1/2}) + c(x_{i+1/2})} \quad 0 \quad \frac{c(x_{i+1/2})}{c(x_{i-1/2}) + c(x_{i+1/2})} \right] .$$

   This stencil needs to be adapted in the first and last row to handle the boundary conditions. The Gershgorin theorem then bounds the eigenvalues of the iteration B_{JAC} by one. This argument can be refined by involving the treatment of the boundary conditions to obtain a bound strictly lower than one. The method of Jacobi applied to the A^h thus converges.

5. suppose that the discretization scheme is given to be of order p. Describe a numerical test to verify that the implementation does indeed yields a discretization error of this order;

   The implementation can be ran on a sequences of finer meshes and the computed solution obtained can be compared with an a-priori constructed exact analytical solution. The scaling of this error with the mesh width can then be studied. For a second order discretization scheme for instance, the error should decrease by a factor of four each time the mesh width is decreased by a factor of two.

Question 3 (10 pnts. - 2 pnt. per subquestion)
Answer the following questions

1. given the linear system $A\,\mathbf{u} = \mathbf{f}$ with an n-by-n real-valued coefficient matrix A. Assume a splitting of this coefficient matrix of the form $A = M - N$ where M is non-singular and assume that a basic iterative solution method for the linear system is derived from this splitting. Derive a recursion formula for the iterands $\mathbf{u}^k$. Derive a recursion formula for the residual vector $\mathbf{r}^k$.
   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

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

   The recursion for the residual vector is given by

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

2. given the linear system $A\,\mathbf{u} = \mathbf{f}$ with an n-by-n real-valued coefficient matrix A. Assume the recursion formula for the error vector $\mathbf{e}^{k+1} = \left(I - M^{-1}\,A\right)\mathbf{e}^k$ to be valid. Give a sufficient condition on the matrix n-by-n $B = I - M^{-1}\,A$ for the iterative scheme to converge.
   A sufficient condition for convergence is that

$$\rho(B) = \rho(I - M^{-1}A) < 1\,.$$

3. give the residual equations that give the relation between the error $\mathbf{e}^k$ and the residual vector $\mathbf{r}^k$. Use this relation to derive the defect-correction scheme that use the approximation $\widehat{A}$ to A;
   The error $\mathbf{e}^k$ and the residual vector $\mathbf{r}^k$ are related by $A\,\mathbf{e}^k = \mathbf{r}^k$. We can use this relation to define an iterative scheme by the following sequence of three steps

   - *compute the* residual *or* 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 $\mathbf{u}^{k+1} = \mathbf{u}^k + \widehat{\mathbf{e}}^k$.*

4. assume that
$$[A] = \frac{1}{h^2}\begin{bmatrix} -1 & 2-1 \end{bmatrix}$$
to be the stencil of the 1D Laplacian on a uniform mesh. Give stencil for the Jacobi and weighted Jacobi iteration matrix B_{JAC};
The stencil for the Jacobi iteration matrix is given by
$$[B_{JAC}] = \begin{bmatrix} 1/2 & 0 & 1/2 \end{bmatrix} .$$
The stencil for the ω-weighted Jacobi iteration matrix is given by
$$[B_{JAC(\omega)}] = \begin{bmatrix} \omega/2 & 1-\omega & \omega/2 \end{bmatrix} .$$

5. assume A to be SPD and let λ_1 and λ_n denote the smallest and largest eigenvalue of A. Assume $M = \tau^{-1}I$ with τ a real-valued parameter to be the splitting correspond to the Richardson method. Derive optimal value for the parameter τ.
The value of τ should be chosen such that the spectral radius $\rho(I - \tau A)$ is minimized. Then
$$\rho(I - \tau A) = \min\{|1 - \tau\lambda_{min}|, |1 - \tau\lambda_{max}|\}$$
. The optimal value of τ is $\tau = (\lambda_{min} + \lambda_{max})/2$.

Question 4 (10 pnts. - 2 pnt. per subquestion)

In this exercise we consider a linear system $A\mathbf{u} = \mathbf{b}$, where matrix A is real-valued and SPD. As an iterative method we use the (Preconditioned) Conjugate Gradient method.

1. What properties should be satisfied in order that a matrix A is SPD? Give a definition of the matrix A-norm;
A is symmetric iff $A = A^T$. A is positive definite iff $\forall \mathbf{u} \neq 0$ holds that $\mathbf{u}^T A\mathbf{u} > 0$. A is SPD in case that A is both symmetric and positive definite.

2. Assume that our first iterate is given by $\mathbf{u}^1 = \alpha\,\mathbf{b}$. Determine an expression for the parameter α such that the error $\mathbf{u} - \mathbf{u}^1$ has a minimal length in the 2-norm;
$$\|\mathbf{u} - \mathbf{u}^1\|_2^2 = (\mathbf{u} - \alpha\mathbf{b})^T(\mathbf{u} - \alpha\mathbf{b}) = \mathbf{u}^T\mathbf{u} - 2\alpha\mathbf{b}^T\mathbf{u} + \alpha^2\mathbf{b}^T\mathbf{b} . \tag{4}$$
This norm is minimized if $\alpha = \frac{\mathbf{b}^T\mathbf{u}}{\mathbf{b}^T\mathbf{b}}$.

3. We consider two 10-by-10 diagonal matrices A and C. For the first matrix we have $a_{i,i} = i,\ i = 1\ldots10$ and for the second $c_{i,i} = 1,\ i = 1\ldots9$ and $c_{10,10} = 1000$. For the vector $\mathbf{b}$ we have $\mathbf{b}_i = 1,\ i = 1\ldots10$ and we take the zero vector as starting solution. For which of both systems is CG faster to converge (motivate your answer)? What is for both systems the maximum number of iterations before a solution with a sufficient small residual norm, say $\|\mathbf{r}^k\|/\mathbf{b}\| \leq 10^{-15}$, is obtained?
For the matrix A 10 distinct eigenvalues and therefore 10 iterations. For the matrix C 2 distinct eigenvalues and therefore 2 iterations. See Figure 1.

4. We now combine the CG method with a preconditioner M. Give the three properties of M in order to have a good preconditioner;
M should be SPD. A linear system with M as coefficient matrix should be easy to solve. M should ressemble A^{-1} in such a way that $\kappa_2(M^{-1}A) < \kappa_2(A)$.

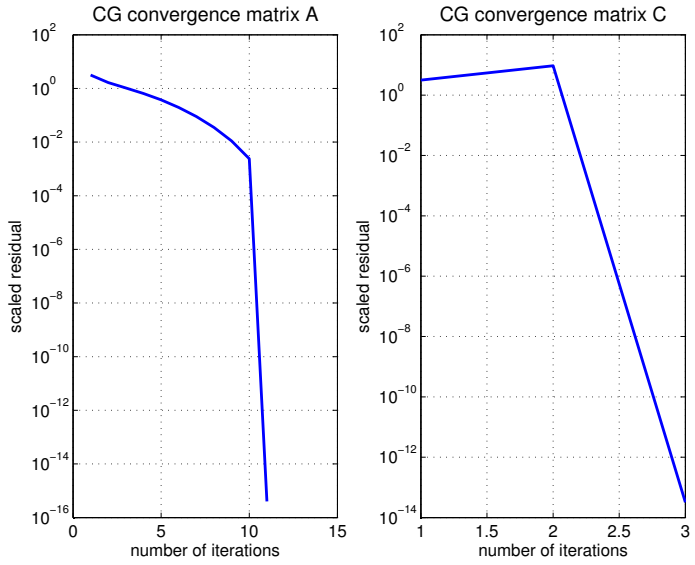

Figure 1: CG convergence history for matrix A and matrix C.

5. Explain how a preconditioner can be combined with CG? Give three possible classes of preconditioners. Compare the properties of these preconditioners in the three classes.

   In case that no preconditioner is applied, the CG method is applied to the linear system $A\,\mathbf{u} = \mathbf{b}$. *In case that a preconditioner is applied, the CG method is applied to the linear system* $M^{-1}A\,\mathbf{u} = M^{-1}\mathbf{b}$. *Three possible classes of preconditioners are diagonal scaling, basic iterative methods and incomplete factorisation. Diagonal scaling leads to faster convergence if* $\max(A_{ii}) \gg \min(A_{ii})$. *The combination of a BIM and CG is always faster to converge than the BIM used as a stand alone solver. The incomplete factorization is the most optimal of the three, but is hard to paralize.*

Question 5 (10 pnts. - 2.5 pnt. per subquestion)

In this exercise we consider the Power method to approximate the eigenvalues of an n-by-n real-valued matrix $A \in \mathbb{R}^{n\times n}$.

1. The basic Power method is given by: $\mathbf{q}_k = A\mathbf{q}_{k-1}$. The eigenvalues are ordered such that $|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_n|$. We assume that $\mathbf{q}_0$ can be written as a linear combination of the eigenvectors, with a non-zero component in the eigenvector corresponding to λ_1. Define $\lambda^{(k)} = \frac{\mathbf{q}_k^T A\,\mathbf{q}_k}{\|\mathbf{q}_k\|_2^2}$ and show that

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

See lecture notes Eqns (8.1) and (8.2).

2. Next we consider the advanced Power method:
$\mathbf{q}_0 \in \mathbb{R}^n$ is given
for $k = 1, 2, ...$
$\mathbf{z}_k = A\mathbf{q}_{k-1}$
$\mathbf{q}_k = \mathbf{z}_k / \|\mathbf{z}_k\|_2$
$\lambda^{(k)} = \mathbf{q}_{k-1}^T \mathbf{z}_k$
endfor
Show that if $\mathbf{q}_k$ is close to the eigenvector corresponding to λ_1 then $\lambda^{(k)}$ is a good approximation of λ_1;
See lecture notes Eqns (8.1) and (8.2).

3. Note that from part (1) it follows that the Power method is a linearly converging method. Give a good stopping criterion for the Power method;
the Power method is a linearly converging method. This implies that the following stopping criterion can be used:

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

and stop if $\frac{\tilde{r}}{1-\tilde{r}} \frac{|\lambda^{(k+1)} - \lambda^{(k)}|}{|\lambda^{(k+1)}|} \leq \epsilon$ *. This stopping criterion leads to* $|\lambda_1 - \lambda^{(k+1)}| \leq \epsilon$.

4. Given an n-by-n real-valued matrix $A \in \mathbb{R}^{n\times n}$, where

$$\lambda_1 = 1000 \ , \quad \lambda_{n-1} = 1.1 \ \text{ and } \ \lambda_n = 1 \ .$$

Give a fast converging variant of the Power Method to approximate λ_n.
Various possibilities to approximate λ_n *using the Power Method exist. The Shifted Power Method applies the Power Method to the matrix* $A - cI$ *such that* $|\lambda_n - c| > |\lambda_1 - c|$. *This inequality holds for instance in case that* $c = 501$. *In this the rate of convergence is given by*

$$\frac{|\lambda_2(A - cI)|}{|\lambda_1(A - cI)|} = \frac{|\lambda_2(A) - c|}{|\lambda_1(A) - c|} = \frac{501 - 1.1}{501 - 1} \approx 0.9998 \, .$$

The inverse Power Method consists of applying the Power Method to A^{-1} *without computing this inverse explicitly. The largest and second largest eigenvalue of* A^{-1} *are equal to* $1/\lambda_1$ *and* $1/\lambda_2$, *respectively. In this case the rate of convergence is given by*

$$\frac{|\lambda_2(A^{-1})|}{|\lambda_1(A^{-1})|} = \frac{1/\lambda_2(A)}{1/\lambda_1(A)} = \frac{\lambda_1(A)}{\lambda_2(A)} = \frac{1}{1.1} \approx 0.9091 \, .$$