SF2524 Matrix Computations for Large-scale Systems
Exam

Aids: None Time: Four hours

Grades: E: 16 points, D: 19 points, C: 22 points, B: 25 points, A: 28 points (out of the possible 35 points, including bonus points from homeworks).

Problem 1 (3p) Several iterative methods for linear systems of equations in this course generate iterates $x_1, x_2, \ldots \in \mathbb{R}^m$ which satisfy

$$\min_{x \in S_n} \|Ax - b\|_Z = \|Ax_n - b\|_Z$$

for some norm $\|\cdot\|_Z$ and space S_n. What is S_n and $\|\cdot\|_Z$ when the iterates $x_1, x_2, \ldots$ are generated by (a) GMRES, (b) CG, (c) CGN?

Solution:

(a) GMRES: $\|z\|_2 = \sqrt{z^T z}$ and $S_n = \mathcal{K}_n(A, b)$

(b) CG: $\|z\|_{A^{-1}} = \sqrt{z^T A^{-1} z}$ and $S_n = \mathcal{K}_n(A, b)$.

(c) CGN: $\|z\|_2 = \sqrt{z^T z}$ and $S_n = \mathcal{K}_n(A^T A, A^T b)$. This can be derived from the fact that CGN is CG applied to $A^T A x = A^T b$ and if we set $\tilde{A} = A^T A$,

$$\|\tilde{A}x - A^T b\|^2_{\tilde{A}^{-1}} = (\tilde{A}x - A^T b)^T (A^T A)^{-1} (\tilde{A}x - A^T b) =$$
$$(\tilde{A}x - A^T b)^T A^{-1} (A^T)^{-1} (\tilde{A}x - A^T b) = ((A^T)^{-1}\tilde{A}x - (A^T)^{-1}A^T b)^T ((A^T)^{-1}\tilde{A}x - (A^T)^{-1}A^T b) =$$
$$(Ax - b)^T (Ax - b) = \|Ax - b\|_2^2$$

Note that GMRES and CGN minimize the residual with respect to the same norm, but over different Krylov subspaces.

Problem 2 (2p)

(a) Prove that the result of one step of the shifted QR-method for a symmetric matrix is a symmetric matrix if the shift is real.

(b) What is the result of one step of the basic QR-method for the matrix $A = \begin{bmatrix} 4 & 0 \\ 3 & 0 \end{bmatrix}$?
Hint: You may want to show that the Q-matrix the QR-factorization of a two-by-two matrix is a Givens rotator $Q = \frac{1}{\sqrt{c^2+s^2}} \begin{bmatrix} c & -s \\ s & c \end{bmatrix}$

(c) Let $\tilde{A}$ be the result of one step of the QR-method applied to A. Derive a closed formula for the QR-method applied to the matrix $B = \begin{bmatrix} A & C \\ 0 & R \end{bmatrix}$ where the matrix R is upper triangular.

Solution:

(a) Shifted QR-method: $\bar{A} = RQ + \mu I$ where $QR = A - \mu I$. Hence, $\bar{A} = (Q^T - \mu I)A + \mu I = Q^T AQ$, and

$$\bar{A}^T = (Q^T AQ)^T = QAQ^T = A.$$

(b) We can select a Givens rotator such that

$$Q^T A = Q^T \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

is upper triangular if we select

$$c = a_{11} \quad s = a_{21}.$$

Since

$$Q^T A = \frac{1}{\sqrt{c^2+s^2}} \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \frac{1}{\sqrt{c^2+s^2}} \begin{bmatrix} \times & \times \\ -a_{11}s + a_{21}c & \times \end{bmatrix} = \begin{bmatrix} \times & \times \\ 0 & \times \end{bmatrix}$$

and Q is the Q-matrix in the QR-factorization of A. For the specific matrix in this case

$$Q = \frac{1}{5} \begin{bmatrix} 4 & -3 \\ 3 & 4 \end{bmatrix}$$

The first step of the basic QR-method is

$$RQ = Q^T AQ = \frac{1}{5^2} \begin{bmatrix} 4 & 3 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 3 & 0 \end{bmatrix} \begin{bmatrix} 4 & -3 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 4 & -3 \\ 0 & 0 \end{bmatrix}.$$

This is an upper triangular matrix, so the basic QR-method converged in one step.

(c) Let $QU = A$ be the QR-factorization of A. We can explicitly construct a QR-factorization of B as follows:

$$\begin{bmatrix} Q & 0 \\ 0 & Q_2 \end{bmatrix} \begin{bmatrix} U & Z \\ 0 & R \end{bmatrix} = \begin{bmatrix} A & C \\ 0 & R \end{bmatrix}.$$

The Q and Z are determined by considering corresponding blocks in the equation. The $(2,1)$-block is the equation

$$QZ = C$$

such that we should select $Z = Q^T C$. The $(2,2)$-block is

$$Q_2 R = R.$$

That is, $Q_2 = I$. One step of the QR-method for B:

$$\begin{bmatrix} U & Q^T C \\ 0 & R \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} UQ & Q^T C \\ 0 & R \end{bmatrix} = \begin{bmatrix} \tilde{A} & Q^T C \\ 0 & R \end{bmatrix}$$

Problem 3 (a) Given an Arnoldi factorization $AQ_n = Q_{n+1}\underline{H}_n$, how are the Krylov approximations for matrix functions generated? (No derivation required.)

(b) Breakdown of the Arnoldi method corresponds to the case that $h_{n+1,n} = 0$. Prove that the Krylov method for matrix functions generates an exact result if this occurs (which means we have no approximation error). You may assume f is an entire function.

Solution:

(a) The approximation is given by

$$f(A)b \approx Q_m f(H_m)e_1\|b\|.$$

(b) We assume $h_{m+1,m} = 0$ and therefore

$$AQ_m = Q_{m+1}\underline{H}_m = Q_m H_m.$$

Hence, for any i we $Q_m H_m^i = AQ_m H_m^{i-1} = \cdots = A^i Q_m$. The Taylor definition of f gives

$$Q_m f(H_m) = \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} Q_m H_m^i = \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} A_m^i Q_m = f(A)Q_m.$$

Therefore,

$$Q_m f(H_m)e_1\|b\| = f(A)Q_m e_1\|b\| = f(A)q_1\|b\| = f(A)b$$

since the starting vector is $q_1 = b/\|b\|$.

Common mistake in exam: Note that $Q_m \in \mathbb{R}^{n\times m}$ where $n > m$ is a rectangular orthogonal matrix which means that $Q_m^T Q_m = I$ but $Q_m Q_m^T \neq I$. See background pdf on course web page. In particular $A_m \neq Q_m H_m Q_m^T$.

Problem 4 (a) Suppose an Arnoldi factorization $AQ_n = Q_{n+1}\underline{H}_n$ is given, where $q_1 = b/\|b\|$. How is the GMRES approximation for $Ax = b$ computed from the Arnoldi relation?

(b) Derive a closed formula for the approximation generated by one step ($n = 1$) of GMRES, only involving b and A.

Solution:

(a) We first compute $z_n \in \mathbb{R}^n$ which is the solution to the linear least squares problem

$$\min_{z\in\mathbb{R}^n} \|\underline{H}_n z - e_1\|b\|\|_2 = \|\underline{H}_n z_n - e_1\|b\|\|_2.$$

The GMRES-approximation is subsequently given by $x_n = Q_n z_n$.

(b) Let $\gamma := 1/\|b\|$. We have $q_1 = b\gamma$ and

$$w = Aq_1 = \gamma Ab, \quad \text{where we defined } c = \begin{bmatrix} 1 \\ \vdots \\ m \end{bmatrix}.$$

In the orthogonalization process we have

$$h_{1,1} = q_1^T w = \gamma^2 b^T A b$$

and

$$w_\perp = w - q_1 h_{1,1} = \gamma(Ab - \gamma^2 b^T A b b)$$

Therefore

$$h_{2,1} = \gamma \| Ab - \gamma^2 b^T A b b \|$$

and $q_2 = w_\perp / h_{2,1}$. The first step is now a least squares solution to

$$\min_z \| \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix} z - e_1 \|b\| \| \|$$

which can be solved directly with the normal equations $z = h_{11}\|b\|/(h_{11}^2 + h_{21}^2)$. Hence,

$$\tilde{x} = q_1 z = \frac{\gamma h_{11}\|b\|}{h_{11}^2 + h_{21}^2} b = \frac{h_{11}}{h_{11}^2 + h_{21}^2} b = \frac{b^T A b}{(b^T A b)^2 + \|(I - \frac{1}{\|b\|^2} b b^T) A b\|^2} b$$

Common mistake in exam: Note that in GMRES we need to solve an overdetermined linear system here $\|\underline{H}_1 z - e_1 \|b\| \|$, which can be done with the matlab command backslash, or for small problems the normal equations. However, it is not a linear system of equations and $z = H_1^{-1} e_1 \|b\|$ is not the GMRES approximation (it is actually the FOM approximation).

Problem 5 Roxanne the rocket scientist needs to determine the trajectory of her space craft by solving a linear system of equations $Ax = b$, where A is a huge symmetric positive definite matrix. She knows that GMRES that CG have the same convergence factor $\rho = 0.1$ for her particular problem. She also knows that a matrix vector product takes 1 hour, a scalar product of two vectors takes 20 minutes, and the adding a linear combination of two vectors takes 10 minutes. What is the computation time for GMRES and CG to achieve full precision (10^{-16})? Which method should she select? Provide clear justifications of your reasoning and simplifications.

Solution: Since the convergence factor is 0.1, the error for both methods behave as

$$\text{error} \sim 0.1^k,$$

so full precision is achieved in 16 iterations (under the assumption that the error behaves exactly as the convergence factor).

CG and GMRES both require one matrix vector product per iteration, and both methods require 16 matrix vector products.

Other operations (rough estimates which can depend on which version of CG is used):

- CG: requires 2 scalar products per iterations and forming linear combination of 3 vectors per iteration. Hence, in total 32 scalar products and 48 linear combinations of vectors.

- GMRES: At step k we need to orthogonalize against $k-1$ vectors. Orthogonalizing against $k-1$ vectors requires (with single GS) $k-1$ scalar products and $k-1$ linear combinations. Normalization requires 1 scalar product. Hence, in total we need

$$\sum_{k=1}^{16} k = 136$$

scalar products and (approximately) as many linear combinations.

Taking the computation time of the operations specified in the question into account:

- CG: $16 * 60 + 32 * 20 + 48 * 10 = 2080$ minutes
- GMRES: $16 * 60 + 136 * 10 + 136 * 20 = 5040$ minutes

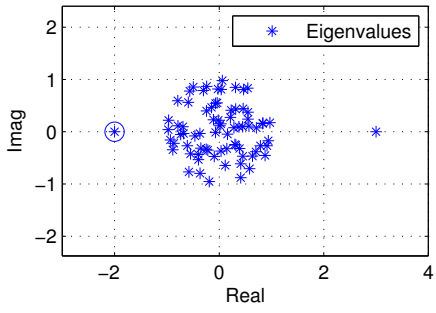
Clearly CG is faster since it is a low-term reccurrence. This is substantial since the computation cost for forming linear combinations and scalar products is substantial in comparison to the matrix-vector product computation.

Common mistake in exam: The matrix H_m is very small (17×16) and the computational effort to solve the overdetermined linear system $\min \|\underline{H}_m z - e_1 \|b\|\|$ is negligable. This is in general the case for GMRES.

Problem 6 A theorem in this course states that the error indicator in Arnoldi's method for eigenvalue problems can (under appropriate conditions) be bounded as

$$\|(I - Q_n Q_n^T)x_j\| \leq \alpha \min_{\substack{p \in P_{n-1} \\ p(\lambda_j)=1}} \max_{i \neq j} |p(\lambda_i)|. \qquad (\star)$$

The eigenvalues of the matrix $A \in \mathbb{R}^{n \times n}$ is given to the right.



(a) The eigenvalue λ_1 is marked with a circle in the figure. Use $(\star)$ to determine a convergence factor γ such that

$$\|(I - Q_n Q_n^T)x_1\| \leq \alpha \gamma^k$$

where $\rho < 1$. Describe clearly what you identify in the figure.

(b) Consider the more generalization for a matrix with eigenvalues as in the figure. Suppose eigenvalues $\lambda_3, \ldots, \lambda_n \in D(\rho, 0)$ and suppose $|\lambda_1| > |\lambda_2| > \rho$. Derive a formula for $\tilde{\gamma}$ such that

$$\|(I - QQ_n^T)x_1\| \leq \beta \tilde{\gamma}^k$$

such that we always have $\tilde{\gamma} < 1$.

Solution:

(a) If we set a disk at $c = 1$ and $\rho = 2$ the figure appears to include all eigenvalues except $\lambda_1 = -2$. We have

$$\gamma = \frac{\rho}{|c - \lambda_1|} = \frac{2}{3}.$$

(b) This question has several correct answers. One alternative: We use the min-max result with polynomial First note that the specific polynomial $p(z) = \frac{z-\lambda_2}{\lambda_1-\lambda_2} q(z)$ is $p \in P_{n-1}$ if $q \in P_{n-2}$. Moreover, since $p(\lambda_2) = 0$, we have

$$\max_{i \neq 1} |p(\lambda)| = \max_{i \neq 1, i \neq 2} |p(\lambda)| \leq \max_{i \neq 2} \frac{|\lambda_i - \lambda_2|}{|\lambda_2 - \lambda_1|} \max_{i \neq 1, i \neq 2} q(\lambda)$$

Therefore,

$$\min \max |p(\lambda)| \leq \max_{i \neq 2} \frac{|\lambda_i - \lambda_2|}{|\lambda_2 - \lambda_1|} \min_{q \in P_{n-2}, q(\lambda_1)=1} \max_{i \neq 1,2} |q(\lambda_i)| \leq \left(\max_{i \neq 2} \frac{|\lambda_i - \lambda_2|}{|\lambda_2 - \lambda_1|} \right) \frac{\rho^{k-2}}{|c - \lambda_1|^{k-2}}$$

So the convergence factor is

$$\tilde{\gamma} = \frac{\rho}{|c - \lambda_1|}$$

Beyond the scope of the question: The coefficient $\tilde{\beta}$ is $\tilde{\beta} = \alpha \left(\max_{i \neq 2} \frac{|\lambda_i - \lambda_2|}{|\lambda_2 - \lambda_1|} \right) \rho / |c - \lambda_1|$ and the specific setting in (a) gives $\tilde{\gamma} = 1/2$ which improves the bound in a.

Problem 7 (a) Compute $f(A)$ with the (simplified) Schur-Parlett method when $a < b < 10$ when

$$A = \begin{bmatrix} a & 1 & 0 \\ & b & 1 \\ & & 10 \end{bmatrix}$$

(b) What is $f(B)$ when $B = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}$ for some constant a?

(c) Suppose $B \in \mathbb{R}^{2\times 2}$, $c \in \mathbb{R}^2$ and $d \in \mathbb{R}$. Derive a formula for $f_c \in \mathbb{R}^2$ such

$$f(A) = \begin{bmatrix} f(B) & f_c \\ 0 & f(d) \end{bmatrix} \text{ when } A = \begin{bmatrix} B & c \\ 0 & d \end{bmatrix}.$$

The formula should be a linear system expressed in terms of $f(B)$, $f(d)$, c, d.

(d) The (simplified) Schur-Parlett method will fail for the matrix in (a) if $b = a$. Use (b)-(c) to derive formula for $f(A)$ when $b = a$. If you encounter a 2×2 linear system of equations, you do not need to explicitly solve it.

Solution:

(a) The problem can be solved directly if the algorithm is memorized. The values are such that the necessary quantities can be derived by hand. From commutativity of A and $F = f(A)$ we have

$$AF = FA, \qquad (\star\star)$$

and since A is upper triangular, $F = f(A)$ is upper triangular

$$F = \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ & f_{22} & f_{23} \\ & & f_{33} \end{pmatrix}.$$

We consider the first row and the second column of ($\star\star$):

$$af_{12} + 1f_{22} = 1f_{11} + f_{12}b.$$

This equation can be solved for f_{12}, such that

$$f_{12} = \frac{f_{22} - f_{11}}{b - a} = \frac{f(b) - f(a)}{b - a}.$$

Similarly, we can consider the second row and third column of ($\star\star$)

$$bf_{23} + 1f_{33} = 1f_{22} + f_{23}10.$$

such that

$$f_{23} = \frac{f(10) - f(b)}{10 - b}.$$

Finally, we consider the first row and third column of ($\star\star$):

$$f_{33} = \frac{1}{f(10) - f(a)} \left(\frac{f(10) - f(b)}{10 - b} - \frac{f(b) - f(a)}{b - a} \right)$$

(b) This is the definition of the matrix function of a Jordan block:

$$f(B) = \begin{pmatrix} f(a) & \frac{1}{1!}f'(a) \\ 0 & f(a) \end{pmatrix}$$

(c) We consider the commutator

$$\begin{bmatrix} f(B) & f_c \\ 0 & f(d) \end{bmatrix} \begin{bmatrix} B & c \\ 0 & d \end{bmatrix} - \begin{bmatrix} B & c \\ 0 & d \end{bmatrix} \begin{bmatrix} f(B) & f_c \\ 0 & f(d) \end{bmatrix} = 0$$

The (1,2)-block corresponds to the equation

$$f(B)c + f_c d - Bf_c + cd = 0$$

and

$$(B - dI)f_c = f(B)c - cf(d)$$

which can be solved explicitly if B, d, c and $f(B)$ is available.

(d) In order to apply (c) we first identify that

$$f(A) = \begin{pmatrix} f(B) & f_c \\ 0 & f(d) \end{pmatrix}$$

where $d = 10$,

$$f(B) = \begin{pmatrix} f(a) & \frac{1}{1!}f'(a) \\ 0 & f(a) \end{pmatrix}$$

and

$$f_c = (B - dI)^{-1}(f(B)c - cf(d)) =$$
$$\begin{pmatrix} a-d & 1 \\ & a-d \end{pmatrix}^{-1} \left[\begin{pmatrix} f(a) & \frac{1}{1!}f'(a) \\ 0 & f(a) \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} f(d) \right] =$$
$$\begin{pmatrix} a-d & 1 \\ & a-d \end{pmatrix}^{-1} \begin{pmatrix} f'(a) \\ f(a) - f(d) \end{pmatrix} = \frac{1}{a-d} \begin{pmatrix} f'(a) - \frac{f(a)-f(d)}{a-d} \\ f(a) - f(d) \end{pmatrix}$$