SF2524 Matrix Computations for Large-scale Systems
Exam - solution

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).
Notation in exam / course: $P_m := \{$ polynomials of degree $m\}$, $P_m^0 := \{p \in P_m : p(0) = 1\}$.

Problem 1 (4p) We consider the special case of Rayleigh quotient for real matrices and real eigenpairs. Let $r(v) := v^T Av/v^T v$ be the Rayleigh quotient where $A \in \mathbb{R}^{n\times n}$. Suppose $v_* \in \mathbb{R}^n$ is a normalized real eigenvector corresponding to a real eigenvalue $\lambda_* \in \mathbb{R}$.

(a) State the Rayleigh quotient iteration.
(b) Derive formula for $w \in \mathbb{R}^n$ such that $r(v) = \lambda_* + w^T(v - v_*) + \mathcal{O}(\|v - v_*\|^2)$.
Hint: For sufficiently small x, we have $\frac{1}{1+x} = 1 - x + \mathcal{O}(x^2)$
(c) In what sense is the Rayleigh quotient better for symmetric matrices?

Solution:

(a) The Rayleigh quotient iteration is given by

$$\begin{aligned} \mu_k &= r(u_k) \\ x_{k+1} &= (A - \mu_k I)^{-1} u_k \\ u_{k+1} &= \frac{x_{k+1}}{\|x_{k+1}\|}. \end{aligned}$$

(b) Define $\Delta = v - v_\star$. Then

$$\begin{aligned} r(v) = r(v_\star + \Delta) &= \frac{(v_\star + \Delta)^T A(v_\star + \Delta)}{(v_\star + \Delta)^T (v_\star + \Delta)} \\ (v_\star + \Delta)^T A(v_\star + \Delta) &= \lambda \underbrace{v_\star^T v_\star}_{=1} + \lambda \Delta^T v_\star + v_\star^T A\Delta + \Delta^T A\Delta \\ (v_\star + \Delta)^T (v_\star + \Delta) &= \underbrace{v_\star^T v_\star}_{=1} + 2v_\star^T \Delta + \Delta^T \Delta. \end{aligned}$$

Therefore, from the hint with $x := 2v_\star^T \Delta + \Delta^T \Delta$, we have

$$\begin{aligned} r(v) &= \frac{(v_\star + \Delta)^T A(v_\star + \Delta)}{1 + 2v_\star^T \Delta + \Delta^T \Delta} \\ &= \left(\lambda_\star + \left(\lambda v_\star^T + v_\star^T A\right)\Delta + \Delta^T A\Delta\right)\left(1 - 2v_\star^T \Delta - \Delta^T \Delta + O(\|\Delta\|^2)\right) \\ &= \lambda_\star + (\lambda v_\star^T + v_\star^T A - \lambda_\star 2v_\star^T)\Delta + O(\|\Delta\|^2). \end{aligned}$$

Hence, the answer to the problem is

$$w^T = v_\star^T A - \lambda_\star v_\star^T.$$

(c) RQ for symmetric matrices has cubic convergence. RQ for non-symmetric matrices has in general only quadratic convergence.
Additional note not required for full points: In (b), this advantage can be seen from the fact that $v_\star^T A - \lambda_\star v_\star = 0$ for symmetric problems, since the left and the right eigenvectors are equal for symmetric matrices.

Problem 2 (3p) Erik the engineer discretizes a partial differential equation and finds out that he needs to solve $Ax = b$ where $A \in \mathbb{R}^{n\times n}$ is a large sparse symmetric positive definite matrix with condition number $\kappa(A) := \|A\|\|A^{-1}\| \gg 1$. He needs to decide if he should use an implementation of CG (Conjugate Gradients) or CGN (Conjugate Gradients Normal equations) for his problem.

(a) What is the relationship between CG and CGN?

(b) According to a theorem in the course, the error of CG is bounded by

$$\frac{\|e_n\|_A}{\|e_0\|_A} \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^m. \tag{1}$$

Derive a bound for the error in CGN in terms of $\kappa(A)$ based on (1). Which method has in general faster convergence in Erik's case?

Solution:

(a) CGN is CG applied to the equations

$$\tilde{A}x = A^T b$$

where $\tilde{A} = A^T A$.

(b) We can directly compute a bound from the CG bound

$$\begin{aligned}\frac{\|e_n\|_{\tilde{A}}}{\|e_0\|_{\tilde{A}}} &= 2\left(\frac{\sqrt{\kappa(\tilde{A})}-1}{\sqrt{\kappa(\tilde{A})}+1}\right)^n \\ &= 2\left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^n\end{aligned}$$

since

$$\kappa(\tilde{A}) = \kappa(A^T A) := \underbrace{\|A^T A\|_2}_{\sigma_n(A)^2}\underbrace{\|(A^T A)^{-1}\|_2}_{\sigma_1(A)^2} = \|A\|_2^2\|A^{-1}\|_2^2 = \kappa(A)^2.$$

For symmetric positive definite matrices, CG is in general faster than CGN.
Additional note not required for full points: The advantage of CGN can be seen from the fact that $\kappa(\tilde{A}) = \kappa(A)^2 \gg \kappa(A)$ since $\kappa(A) \gg 1$ and $(\kappa(\tilde{A})-1)/(\kappa(\tilde{A})+1) \gg (\kappa(A)-1)/(\kappa(A)+1)$. Note, however, that the bounds are given in different norms $\|x\|_A$ and $\|x\|_{A^TA}$ and a complete theoretical justification would require relations between the norms.

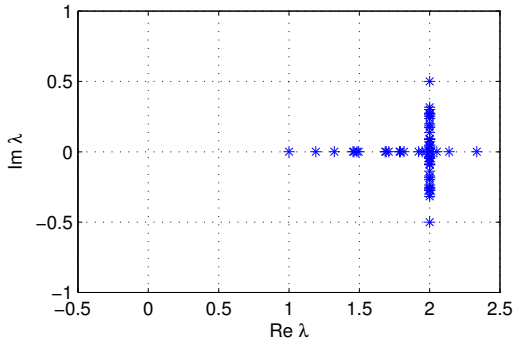
Problem 3 (4p) Suppose the eigenvalues of a matrix A are given as in the figure to the right. Suppose eigenvalues are distinct and $\kappa(V) = 1$.

(a) State a definition of the approximation generated by GMRES.

(b) We apply m steps of GMRES to $Ax = b$ and get approximation x_m. Derive a α and β such that

$$\frac{\|Ax_m - b\|}{\|b\|} \leq \alpha\beta^m.$$

Clearly specify which theorems/results you use and what quantities you observe in the figure. You may use any theorem/result derived in the course.



Solution:

(a) The GMRES approximation is given by

$$x_\star = \underset{x \in \mathscr{K}_n(A,b)}{\operatorname{argmin}} \|Ax - b\|_2.$$

where $\mathscr{K}_n(A,b) = \operatorname{span}(b, Ab, \ldots, A^{n-1}b)$.

(b) Direct application of the min-max bound:

$$\frac{\|r_n\|}{\|b\|} \leq \kappa(V) \max_{p \in P_n^0} \{|p(\lambda_1)|, \cdots, |p(\lambda_n)|\}$$
$$\overset{\text{corollary in lecture notes}}{\leq} \kappa(V) \left|\frac{\rho}{c}\right|^n,$$

where $C(c,\rho)$ is a disc with radius ρ centred at c. We can select $c = 2$ and $\rho = 1 \Longrightarrow$

$$\frac{\|Ax_n - b\|}{\|b\|} \leq \alpha\beta^n,$$

where $\alpha = \kappa(V)$ and $\beta = \left|\frac{\rho}{c}\right| = \frac{1}{2}$.

Problem 4 (4p)
Let

$$A = \begin{bmatrix} 1 & -3 \\ -1 & 1 \end{bmatrix}.$$

(a) Describe the basic QR-method and use the output to the right to compute one step for A. Describe clearly how you use the output.

(b) Describe the shifted QR-method and compute one step for A.

```
>> F=[1 1; -1 ,1];
>> T=[1,-2;0,-1];
>> F*T
ans =
     1    -3
    -1     1
```

Solution:

(a) Basic QR-method:

$A_1 = A$
Compute $Q_k, R_k : Q_k R_k = A_k,$ where Q_k orthogonal, and R_k upper triangular
Set $A_{k+1} = R_k Q_k$.

By the hint in matlab program

$$A_1 = A = \begin{bmatrix} 1 & -3 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} R,$$

where

$$R = \sqrt{2} \begin{bmatrix} 1 & -2 \\ 0 & -1 \end{bmatrix}.$$

$$A_2 = R_1 Q_1 = \sqrt{2} \begin{bmatrix} 1 & -2 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \frac{1}{\sqrt{2}} = \begin{bmatrix} 3 & -1 \\ 1 & -1 \end{bmatrix}.$$

(b) The shifted QR-method (with Rayleigh quotient shift) is here given by, where $\mu = a_{22}$:

- Let $\tilde{A} = A - \mu I$
- Let Q, R be a QR-factorization of $\tilde{A} = QR$
- Let $A_1 = RQ + \mu I$.

In formulas:

$$\tilde{A} = A - \sigma I \overset{\sigma = a_{22} = 1}{=} \begin{bmatrix} 0 & -3 \\ -1 & 0 \end{bmatrix}.$$

$QR = \tilde{A}$ if

$$Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix}.$$

Hence,

$$RQ + \sigma I = \begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + I = \begin{bmatrix} 0 & -1 \\ -3 & 0 \end{bmatrix} + I = \begin{bmatrix} 1 & -1 \\ -3 & 1 \end{bmatrix}.$$

Problem 5 (4p) Let $T \in \mathbb{R}^{3\times 3}$ be an **upper triangular matrix** with distinct eigenvalues and let

$$f(T) = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{13} \\ f_{31} & f_{32} & f_{33} \end{bmatrix},$$

where the matrix function is defined via the Jordan definition or Taylor definition.

(a) Derive explicit formulas for f_{11}, f_{21}, f_{31}, f_{22}, f_{32}, f_{33} involving only elements of T.
(b) Derive an explicit formula for f_{12} involving only elements of T and the values in (a).

Solution:

(a) Taylor definition

$$f(T) = \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} T^i,$$

where

$$T^i = \begin{bmatrix} t_{11}^i & \times & \cdots & \times \\ 0 & t_{22}^i & \ddots & \vdots \\ \vdots & \ddots & \ddots & \times \\ 0 & \cdots & 0 & t_{nn}^i \end{bmatrix}$$

and $\times$ denotes a non-zero element. Hence,

$$f(T) = \begin{bmatrix} \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} t_{11}^i & \times & \cdots & \times \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \times \\ 0 & \cdots & 0 & \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} t_{nn}^i \end{bmatrix} = \begin{bmatrix} f(t_{11}) & \times & \cdots & \times \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \times \\ 0 & \cdots & 0 & f(t_{nn}) \end{bmatrix}.$$

(b) The proof is essentially the first step in the derivation of the Schur-Parlett method, described in the lectures and Golub and Van Loan. The derivation below is given in a more general form than necessary for this specific problem, in order to illustrate the general principle.

First note that from the Taylor definition we know directly that $f(T)$ and T commute such that $f(T)T = Tf(T)$. From (a) we know that $f(T)$ is upper triangular and

$$f(T)T = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ 0 & f_{22} & f_{13} \\ 0 & 0 & f_{33} \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ 0 & t_{22} & t_{13} \\ 0 & 0 & t_{33} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ 0 & t_{22} & t_{13} \\ 0 & 0 & t_{33} \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ 0 & f_{22} & f_{13} \\ 0 & 0 & f_{33}, \end{bmatrix} = Tf(T) \quad (*)$$

where $f_{ii} = f(t_{ii})$. Multiplication of (*) from the right with

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

and from the left with J^T essentially extracts the top left blocks of the matrices since

$$J^T \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ 0 & f_{22} & f_{13} \\ 0 & 0 & f_{33} \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ 0 & t_{22} & t_{13} \\ 0 & 0 & t_{33} \end{bmatrix} J = \begin{bmatrix} f_{11} & f_{12} \\ 0 & f_{22} \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix}$$

and correspondingly for $Tf(T)$. Hence,

$$\begin{bmatrix} f_{11} & f_{12} \\ 0 & f_{22} \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} \\ 0 & f_{22} \end{bmatrix}$$

In particular the (1,2)-element gives the equation

$$f_{11}t_{12} + f_{12}t_{22} = t_{11}f_{12} + t_{12}f_{22}.$$

We solve this equation explicitly for the only unknown quantity f_{12}:

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

Problem 6 (5p) Let Q_m and $\underline{H}_m$ be an Arnoldi factorization of A.

(a) How are the eigenvalue approximations computed from the Arnoldi factorization in *Arnoldi's method for eigenvalue problems*?

(b) State the Krylov approximation f_m of $f(A)b$.

(c) Under certain conditions on A and f, the error of the Krylov approximation is bounded by

$$\|f(A)b - f_m\| \leq 2\|b\| \min_{p \in P_{m-1}} \max_{z \in \Omega} |f(z) - p(z)|$$

where Ω is a compact set containing all eigenvalues. Suppose the eigenvalues are real and in the interval $I = (0.5, 1.5)$. Determine α and β such that

$$\|f(A)b - f_m\| \leq \alpha \frac{\beta^m}{m!},$$

for all m and for any function satisfying $|f^{(k)}(x)| \leq C$ for all $x \in I$, $k \in \mathbb{Z}$.
Hint: The remainder of the truncated Taylor series satisfies $f(x) - \sum_{k=0}^{m-1} (x-\mu)^k \frac{f^{(k)}(\mu)}{k!} = (x-\mu)^m \frac{f^{(m)}(\xi)}{m!}$, *for some value* $\xi \in [x, \mu]$.

Solution:

(a) If $\underline{H}_m \in \mathbb{R}^{(m+1)\times m}$, the approximate eigenvalues of A are eigenvalues of $H_m \in \mathbb{R}^{m\times m}$, which are also known as Ritz values.

(b) The Arnoldi approximation for functions of matrices is given by $f_m = Q_m f(H_m) e_1 \cdot \|b\|$

(c) We use the corollary and take the truncated Taylor series as $p = q$

$$f(x) = p(x) + \frac{f^{(m)}(\xi)}{m!} \underbrace{(x-\mu)^m}_{\|\cdot\| \leq |0.5|^m} \Longrightarrow$$

$$|f(z) - p(z)| \leq 2\|b\| C \frac{0.5^m}{m!} \Longrightarrow \alpha = 2\|b\|C, \beta = 0.5^m.$$

Problem 7 (5p) Suppose $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, such that $A = V\Lambda V^T$ where $V^T V = I$ and the columns of V are eigenvectors. We start the Arnoldi method with a vector b such that it is orthogonal to the first eigenvector: $x_1^T b = 0$. In this case, the error indicator for Arnoldi's method for eigenvalue problems (for the second eigenvalue λ_2) is bounded by

$$\|(I - Q_m Q_m^T) x_2\| \leq \xi_2 \tilde{\varepsilon}_2^{(m)} \tag{2}$$

for some constant ξ_2, where

$$\tilde{\varepsilon}_2^{(m)} := \min_{\substack{p \in P_{m-1} \\ p(\lambda_2)=1}} \max(|p(\lambda_3)|, \ldots, |p(\lambda_n)|). \tag{3}$$

(a) Suppose $\lambda_k = 1 + \sin\left(\frac{(k-1)\pi}{2(n-1)}\right)$, $k = 1, \ldots n$. Use (2) to derive β such that $\|(I - QQ_m^T)x_2\| \leq \alpha\beta^{m-1}$ for some α.

(b) Prove (2) and (3) and derive a formula for the constant ξ_2.

Solution:

(a)

$$\begin{aligned}
\lambda_1 &= 1 \\
\lambda_2 &= 1+\sin(\frac{\pi}{2(n-1)}) \\
\lambda_3 &= 1+\sin(\frac{2\pi}{2(n-1)}) = 1+\sin(\frac{\pi}{n-1}) \\
&\vdots \\
\lambda_n &= 1+\sin(\frac{(n-1)\pi}{2(n-1)}) = 2.
\end{aligned}$$

Note that the eigenvalues $\lambda_3,\ldots,\lambda_n$ are contained in the interval $I=[1+\sin(\frac{\pi}{n-1}),2]$. We use the polynomial defined via a disk that we have used in several proofs in the course. We can select

$$c = \frac{1+\sin(\frac{\pi}{n-1})+2}{2} = \frac{3}{2}+\frac{1}{2}\sin(\frac{\pi}{n-1}),$$

and

$$\rho = 2-c = 0.5-\frac{1}{2}\sin(\frac{\pi}{n-1}),$$

such that $\lambda_3,\ldots,\lambda_n$ are contained in a disk of radius ρ centered at c. By using the polynomial

$$p(z) = \left(\frac{z-c}{\lambda_2-c}\right)^{m-1},$$

we obtain for any $z=\lambda_i$, $i=3,\ldots,n$, that

$$\begin{aligned}
|p(z)| &\leq \left|\frac{\rho}{\lambda_2-c}\right|^{m-1} \\
&= \left(\frac{1-\sin(\frac{\pi}{n-1})}{1-2\sin(\frac{\pi}{2(n-1)})+\sin(\frac{\pi}{n-1})}\right)^{m-1} = \beta^{m-1}.
\end{aligned}$$

(b) The starting vector can be expressed as a linear combination of eigenvectors (since the eigenvectors span $\mathbb{R}^n$),

$$b = \alpha_1 x_1 + \cdots + \alpha_n x_n,$$

where the eigenvectors are normalized. Since, $x_1^T b = 0$ from the assumption in the question and the eigenvectors are orthogonal we have

$$0 = x_1^T b = \alpha_1 x_1^T x_1 + \cdots + \alpha_n x_1^T x_n = \alpha_1 x_1^T x_1.$$

Due to the fact that the eigenvectors are normalied, we have

$$\alpha_1 = 0.$$

The proof follows the same line of reasoning as the proof in the lecture notes, except "step 3". The following sequence of equalities is the same as in the lecture notes except equality (**), where we use that $\alpha_1 = 0$.

$$
\begin{aligned}
\|(I - QQ^T)\alpha_2 x_2\| &= \min_{p \in P_{m-1}} \|\alpha_2 x_2 - p(A) \sum_{j=1}^{n} \alpha_j x_j\| \\
&= \min_{p \in P_{m-1}} \left\| \alpha_2 x_2 - \sum_{j=1}^{n} \alpha_j p(\lambda_j) x_j \right\| \\
&= \min_{p \in P_{m-1}} \left\| \alpha_2 x_2 - \sum_{j=1, j\neq 1}^{n} \alpha_j p(\lambda_j) x_j \right\| \qquad (**) \\
&\leq \min_{\substack{p \in P_{m-1} \\ p(\lambda_2)=1}} \left\| \alpha_2 x_2 - \sum_{j=1, j\neq 1}^{n} \alpha_j p(\lambda_j) x_j \right\| \\
&= \min_{\substack{p \in P_{m-1} \\ p(\lambda_2)=1}} \|\alpha_2 x_2 - \alpha_2 x_2 - \sum_{\substack{j=1, j\neq 1 \\ j\neq 2}}^{n} \alpha_j p(\lambda_j) x_j\| \\
&= \min_{\substack{p \in P_{m-1} \\ p(\lambda_2)=1}} \| \sum_{\substack{j=1 \\ j\neq 2, j\neq 1}}^{n} \alpha_j p(\lambda_j) x_j\| \\
&\leq (\sum_{\substack{j=1 \\ j\neq 2}}^{n} |\alpha_j|) \cdot \min_{\substack{p \in P_{m-1} \\ p(\lambda_2)=1}} \max_{j\neq 2, j\neq 1} (|p(\lambda_j)|) \\
&= (\sum_{\substack{j=1 \\ j\neq 2}}^{n} |\alpha_j|) \cdot \tilde{\varepsilon}_2^{(m)}
\end{aligned}
$$

The conclusion is established by dividing the equation by $|\alpha_2|$.