SF2524 Matrix Computations for Large-scale Systems
Exam - Solutions

Aids: None Time: Four hours

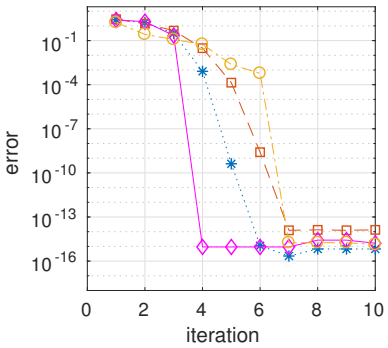
Grades: E: 17 points, D: 19 points, C: 22 points, B: 26 points, A: 29 points (out of the possible 33 points, including bonus points from homeworks).

Problem 1 (5p)

(a) What should ?? be to obtain the dotted curve with stars. Relate to theory for Rayleigh quotient. (Note: Answer without relation to theory will not give full points).

(b) Which one of curves corresponds to Arnoldi's method for eigenvalue problems applied to A with `b=eye(n,1)` and ?? selected as -5. Relate to theory for Arnoldi's method.

```
>> A=[1 3 3 ??; 3 -1 3 2;
  3 3 4 -1; 4 2 -1 -1];
>> x=ones(4,1); s=0;
>> for k=1:10
>>    x=(A-s*eye(4))\x;
>>    x=x/norm(x);
>>    s=x'*A*x;
>> end
```



Solution:

(a) We note that the dotted line is essentially cubic convergence, since from iteration 4 to 5 the error goes from 10^{-3} to 10^{-9}. The algorithm in the notes is the Rayleigh quotient iteration, which has cubic convergence if the matrix is symmetric. The matrix is symmetric if we select ??$= 4$.

(b) The correct solution is the pink line with diamonds. The reason is that we observe an essentially exact solution after 4 iterations. Any Arnoldi method (GMRES, eigenvalues, or matrix functions) leads gives zero error when the number of steps is equal to the problem size. This can be seen from the fact that error is bounded by a factor

$$\min_{p \in P_m} \max_{\lambda \in \lambda(A) \setminus \lambda_i} |p(\lambda)|$$

This factor is zero if the number of steps is equal to the size of the matrix (which is 4).

Problem 2 (6p)

(a) State the minimization definition of the CG-iterates.

(b) How are the iterates of CGN (sometimes called CGNE) defined?

(c) Prove that the CGN iterates are minimizers of a residual with respect to a norm over a space X. Which norm and what is X?

Solution:

(a) The CG-iterates are defined as the minimizers over a Krylov subspace with respect to a particular norm

$$\|Ax_m - b\| = \min_{x \in \mathcal{K}_m(A,b)} \|Ax - b\|_{A^{-1}}$$

(b) The CGN-method is defined as the CG-method applied to

$$A^T A x = A^T b$$

which means that the CGN-iterates satisfy

$$\|Ax_m - b\| = \min_{x \in \mathcal{K}_m(A^T A, A^T b)} \|A^T Ax - A^T b\|_{(A^T A)^{-1}}$$

(c) This follows from properties of norms (the proof was done in the lecture and available in the lecture notes)

$$\begin{aligned} \|A^T Ax - A^T b\|^2_{(A^T A)^{-1}} &= \|A^T(Ax - b)\|^2_{(A^T A)^{-1}} \\ &= (Ax - b)^T A (A^T A)^{-1} A^T (Ax - b) \\ &= (Ax - b)^T A A^{-1} A^{-T} A^T (Ax - b) \\ &= (Ax - b)^T (Ax - b) = \|Ax - b\|^2. \end{aligned}$$

Problem 3 (5p) In this question you shall apply a result for the movement (perturbation) of eigenvalues known as the Bauer-Fike theorem:

> The eigenvalues of $A = B + C$ are contained in discs centered at the eigenvalues of B with radius $K\|C\|$.

where K is the eigenvalue condition number $K = \|V\|\|V^{-1}\|$. We apply GMRES to the matrix $A = \alpha I + C$, where A is a normal matrix such that $\|V\|\|V^{-1}\| = 1$. Provide a convergence factor bound in terms of α and $\|C\|$. You may invoke any theorem/lemma we have used in the course.

Solution: The GMRES method for $A^{-1}b$ has a convergence (residual norm) bounded by

$$\frac{\|Ax_m - b\|}{\|c\|} \leq \|V\|\|V^{-1}\|(\frac{\rho}{|c|})^m$$

if the eigenvalues of A are contained in a disc centered at $c \in \mathbb{C}$ and radius ρ. The value $\rho/|c|$ is a convergence factor bound, so we need to apply Bauer-Fike to this expression for $A = \alpha I + C$. Eigenvalues of the matrix αI are all equal to α. Therefore by the Bauer-Fike theorem we have that all eigenvalues of A are contained in a disk of radius $\|C\|$ centered at α. Hence,

$$\frac{\rho}{|c|} \leq \frac{\|C\|}{|\alpha|}.$$

Problem 4 (5p) We define the scalar product $< u, v >= u^T L^T L v$ for some non-singular matrix L. We say that a matrix $Q = [q_1, \ldots, q_m]$ is orthogonal with respect to this scalar product if

$$< q_i, q_j >= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$$

which is equivalent to the equation $Q^T L^T L Q = I$.

(a) Suppose $Q = [q_1, \ldots, q_m]$ is orthogonal with respect to this scalar product. Construct a method of the type

```
>>    h=??
>>    z=??
>>    beta=??
>>    q_new=z/beta;
```

where all operations are done using matrices. The method should, given a vector $b \notin \text{span}(q_1, \ldots, q_m)$, construct q_{m+1} :=`q_new` such that it satisfies $< q_i, q_{m+1} >= 0$ for $i = 1, \ldots, m$ and $< q_{m+1}, q_{m+1} >= 1$ and $b = h_1 q_1 + \ldots + h_m q_m + \beta q_{m+1}$.

(b) We now apply Arnoldi's method with the orthogonalization procedure stated in (a). Let Q_m and $\underline{H}_m$ be the matrices generated. Suppose the matrix A is not symmetric but satisfies instead $AL^T L = L^T L A^T$ which means $< u, Av >=< Au, v >$ for any u and v. What property/structure of H_m does this imply?

Solution:

(a) We want to orthogonalize (and then normalize) the vector b against Q which is orthogonal with respect to the given scalar product.

First we try to find $h_1, \ldots, h_m$ and z such that

$$b = h_1 q_1 + \cdots + h_m q_m + z \tag{1}$$

where z is orthogonal to $q_1, \ldots, q_m$ with respect to $< \cdot, \cdot >$. By considering $< q_i, b >$ for $i = 1, \ldots, m$ applied to (1) we see from the orthogonality of $q_1, \ldots, q_m$ that

$$h = \begin{bmatrix} < q_1, b > \\ \vdots \\ < q_m, b > \end{bmatrix} = \begin{bmatrix} q_1^T L^T L b \\ \vdots \\ q_m^T L^T L b \end{bmatrix} = Q^T L^T L b.$$

In order to compute z we set

$$z = b - q_1 h_1 + \cdots q_m h_m = b - Qh$$

We let q_{m+1} be the z but normalized,

$$\beta = \sqrt{< z, z >} = \sqrt{z^T L^T L z} = \|Lz\|.$$

and

$$q_{m+1} = z/\beta.$$

The algorithm in MATLAB is

```
>>    h=Q'*(L'*(L*b))
>>    z=b-Q*h
>>    beta=norm(L*z)
>>    q_new=z/beta;
```

(b) We run Arnoldi's method with the specified scalar product. This implies that the Arnoldi factorization is satisfied:

$$AQ_m = Q_{m+1} \underline{H}_m \qquad \text{(AF)}$$

but Q_{m+1} is now orthogonal with respect to $< \cdot, \cdot >$ which means (according to (a)) that

$$Q_m^T L^T L Q_m = I$$

In order combine this with (AF) we multiply (AF) with $Q_m^T L^T L$ such that

$$Q_m^T L^T L A Q_m = Q_m^T L^T L \begin{bmatrix} Q_m & q_{m+1} \end{bmatrix} \underline{H_m} = H_m.$$

From the symmetry condition $AL^T L = L^T L A^T$ we see that

$$Q_m^T L^T L A Q_m = Q_m^T A^T L^T L Q_m$$

We identify this as the transpose of $Q_m^T L^T L A Q_m$, therefore

$$H_m = Q_m^T L^T L A Q_m = (Q_m^T L^T L A Q_m)^T = H_m^T$$

which implies that H_m is symmetric. Moreover, since H_m is a Hessenberg matrix it is also a tridiagonal matrix.

Problem 5 (4p) Let

$$P = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}, \qquad f(P) = F = \begin{bmatrix} F_A & F_B \\ 0 & F_C \end{bmatrix}$$

where $C = \alpha I$ and $A, B, C, F_A, F_B, F_C \in \mathbb{C}^{n\times n}$. Derive a formula for F_B [1] only involving $A, B, \alpha, f(A)$, and $f(C)$.

Solution: The proof is a generalization of the derivation of the Schur-Parlett method. We use that P and F commute such that

$$PF = FP$$

in other words

$$\begin{bmatrix} A & B \\ 0 & C \end{bmatrix} \begin{bmatrix} F_A & F_B \\ 0 & F_C \end{bmatrix} = \begin{bmatrix} F_A & F_B \\ 0 & F_C \end{bmatrix} \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

Analogous to the derivation of Schur-Parlett, we consider the two-one block of that matrix equation which leads to a different matrix equation:

$$AF_B + BF_C = F_A B + F_B C \qquad (\star)$$

Since P is block triangular, $F_A = f(A)$ and $f_C = f(C)$. In the statement of the problem $C = \alpha I$ and therefore $f(\alpha I) = f(\alpha) I$ and $(\star)$ becomes

$$AF_B + f(\alpha)B = f(A)B + \alpha F_B$$

Rearranging the terms $f(\alpha)B - f(A)B = \alpha F_B - AF_B = (\alpha I - A)F_B$ which in turn implies that

$$F_B = (\alpha I - A)^{-1}(f(\alpha)B - f(A)B).$$

Problem 6 (5p)

(a) Prove shift invariance of Arnoldi factorization by showing a relation of the form

$$(A - \mu I)Q_m = Q_{m+1}???$$

(b) In a particular application we discretize a parameter dependent PDE which leads to a parameter dependent linear system of equations $g(\mu) = (A - \mu I)^{-1} b$ where $A \in \mathbb{R}^{n\times n}$ is large and sparse. We want to evaluate $g(\mu)$ for $\mu = \mu_1, \ldots, \mu_p$ for a large p-value. Derive and explain (for instance in the form of a program) a method based on GMRES, which only requires the computation of $N = 100$ matrix-vector products with matrix A, to compute *all* vectors $g(\mu_1), \ldots, g(\mu_p)$. (That is, the number of matrix-vector products with A is independent of p.) You may assume that GMRES converges in N steps.

[1] A typo in the problem formulation has been corrected here. The exam said F_C when it should have been F_B.

Solution:

a) By assumption $AQ_m = Q_{m+1}\underline{H}_m$ which implies that

$$(A - \mu I)Q_m = AQ_m - \mu Q_m = Q_{m+1}\underline{H}_m - \mu Q_m.$$

We now define the extended identity matrix

$$\underline{I} = \begin{bmatrix} I \\ 0 \end{bmatrix} \in \mathbb{C}^{(m+1)\times m}.$$

Since the first m columns of Q_{m+1} is Q_m we can write $Q_m = Q_{m+1}\underline{I}$ and

$$(A - \mu I)Q_m = Q_{m+1}\underline{H}_m - \mu Q_{m+1}\underline{I} = Q_{m+1}(\underline{H}_m - \underline{I}).$$

b) We use Arnoldi's method to compute $AQ_N = Q_N\underline{H}_N$ which only requires N matrix vector products. We wish to use this to numerically solve many linear systems $g(\mu) = (A-\mu I)^{-}b$. The GMRES-approximation is the minimizer of

$$\min_{x\in\mathcal{K}_N(A-\mu I,b)} \|(A - \mu I)x - b\|_2 = \min_{z\in\mathbb{C}^N} \|(A - \mu I)Q_N z - b\|$$

we can simplify this expression analogous to the derivation of GMRES by using the shift invariance of the Arnoldi factorization (proven in (a)):

$$\|(A - \mu I)Q_N z - b\| = \|Q_{N+1}(\underline{H}_N - \mu\underline{I})z - Q_{N+1}e_1\|b\|\| = \\ \|(\underline{H}_N - \mu\underline{I})z - e_1\|b\|\|$$

In other words, after computing the Arnoldi factorization we can compute the approximation of $g(\mu)$ as

$$g(\mu) = (A - \mu I)^{-1}b \approx \|b\|Q_m((\underline{H}_N - \mu\underline{I}))\backslash e_1$$

The right-hand side does not involve a matrix-vector product with A, and we can therefore evaluate it for $g(\mu_1), \ldots, g(\mu_p)$ with a computation time which is essentially independent of p.