18.335 Problem Set 4 Solutions

Problem 1: (10 points)

This is a Galerkin method, very similar to the Rayleigh–Ritz method, and is known as the "FOM" algorithm ("full orthogonalization method"). There are some FOM references on the 18.335 lecture-summary page: in practice, it converges similarly to GMRES, but with some inconvenient quirks, so it is rarely used in practice.

We want to find $x \in \mathscr{K}_n$ where $b - Ax$ is $\perp \mathscr{K}_n$. $x \in \mathscr{K}_n$ implies that $x = Q_n z$ for some $z \in \mathbb{C}^n$, and $b - Ax \perp \mathscr{K}_n$ means $Q_n^*(b - Ax) = 0 = Q_n^* b - Q_n^* A Q_n z$. From class (and the book), $Q_n^* A Q_n = H_n$. So, we have the $n \times n$ system of equations

$$H_n z = Q_n^* b$$

whose solution gives $x = Q_n z$.

In the common case where our iterative method starts with the initial guess $x_1 = 0$, then the first vector in the Arnoldi iteration is $q_1 = b/\|b\|_2$, and the above equation simplies further: $Q_n^* b = e_1 \|b\|_2$ where e_1 is the coordinate vector $(1,0,0,\ldots)^T$ as usual, but this simplification isn't that vital.

Problem 2: (5+10 points)

Suppose A is a diagonalizable matrix with eigenvectors $\mathbf{v}_k$ and eigenvalues λ_k, in decreasing order $|\lambda_1| \geq |\lambda_2| \geq \cdots$. Recall that the power method starts with a random $\mathbf{x}$ and repeatedly computes $\mathbf{x} \leftarrow A\mathbf{x}/\|A\mathbf{x}\|_2$.

(a) After many iterations of the power method, the λ_1 and λ_2 terms will dominate:

$$\mathbf{x} \approx c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2$$

for some c_1 and c_2. However, this is not an eigenvector. Multiplying this by A gives $\lambda_1 c_1 \mathbf{v}_1 + \lambda_2 c_2 \mathbf{v}_2 = \lambda_1 \left(c_1 \mathbf{v}_1 + \frac{\lambda_2}{\lambda_1} c_2 \mathbf{v_2} \right)$, which is not a multiple of $\mathbf{x}$ and hence will be a different vector after normalizing, meaning that it does not converge to any fixed vector.

(b) The key point is that if we look at the vectors $\mathbf{x} \approx c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2$ and $\mathbf{y} \approx \lambda_1 c_1 \mathbf{v}_1 + \lambda_2 c_2 \mathbf{v}_2$ from **two subsequent** iterations, then after **many iterations** these are *linearly independent* vectors that *span the two desired eigenvectors*. We can then employ e.g. a Rayleigh–Ritz procedure to find $\mathbf{v}_1$ and $\mathbf{v}_2$: use Gram–Schmidt to find an orthonormal basis $\mathbf{q}_1 = \mathbf{x}/\|\mathbf{x}\|_2$ and $\mathbf{q}_2 = (\mathbf{y} - \mathbf{q}_1 \mathbf{q}_1^* \mathbf{y}) / \| \cdots \|_2$, form the matrix $Q = (\mathbf{q}_1, \mathbf{q}_2)$ and find the 2×2 matrix $A_2 = Q^* A Q$. The eigenvalues of A_2 (the Ritz values) will then converge to the eigenvalues λ_1 and λ_2 and we obtain $\mathbf{v}_1$ and $\mathbf{v}_2$ (or some multiple thereof) from the corresponding Ritz vectors. The key point is that AQ is in the span of $\mathbf{q}_1$ and $\mathbf{q}_2$ (in the limit of many iterations so that other eigenvectors disappear), so the Ritz vectors are eigenvectors.

Of course, since we don't know λ_3 then we don't know how many iterations to run, but we can do the obvious convergence tests: every few iterations, find the Ritz values from the last two iterations, and stop when these Ritz values stop changing to our desired accuracy.

Alternatively, if we form the matrix $X = (\mathbf{x}, \mathbf{y})$ from the vectors of two subsequent iterations, then we know that (after many iterations) the columns of AX are in $C(X) = \mathbf{x}, \mathbf{y}$. Therefore, the problem $AX = XS$, where S is a 2×2 matrix, has an exact solution S. If we then diagonalize $S = Z \Lambda Z^{-1}$ and multiply both sizes by Z, we obtain $AXZ = XZ\Lambda$, and hence the columns of XZ are eigenvectors of A and the eigenvalues $\operatorname{diag} \Lambda$ of S are the eigenvalues λ_1 and λ_2 of A. However, this is computationally equivalent to the Rayleigh–Ritz procedure above, since to solve $AX = XS$ for S we would first do a QR factorization $X = QR$, and then solve the normal equations $X^* X S = X^* A X$ via $RS = Q^* A Q R = A_2 R$. Thus, $S = R^{-1} A_2 R$: the S and A_2 eigenproblems are similar; in exact arithmetic, the two approaches will give exactly the same eigenvalues and exactly the same Ritz vectors.

[As yet another equivalent alternative, we could write $AXZ = XZ\Lambda$ as above, and then turn it into

$(X^*AX)Z = (X^*X)Z\Lambda$, which is a 2×2 *generalized* eigenvalue problem, or $(X^*X)^{-1}(X^*AX)Z = Z\Lambda$, which is an ordinary 2×2 eigenproblem.]

Problem 3 (15 points)

Trefethen, problem 27.5. The basic answer here is that their *is* a big roundoff error, but it is in the direction of the eigenvector we want, so we don't care (since overall scale factors are irrelevant to eigenvector computations).

In finite precision, instead of $w = A^{-1}v$, we will get $\tilde{w} = w + \delta w$ where $\delta w = -(A+\delta A)^{-1}\delta A\, w$ (from the formula on page 95), where $\delta A = O(\varepsilon_{\text{machine}})\|A\|$ is the backwards error. [Note that we cannot use $\delta w \approx -A^{-1}\delta A\, w$, which neglects the $\delta A \delta w$ terms, because in this case δw is not small.] The key point, however, is to show that δw is mostly parallel to q_1, the eigenvector corresponding to the smallest-magnitude eigenvalue λ_1 (it is given that all other eigenvalues have magnitude $\geq |\lambda_2| \gg |\lambda_1|$). Since w is also mostly parallel to q_1, this will mean that $\tilde{w}/\|\tilde{w}\|_2 \approx q_1 \approx w/\|w\|_2$.

First, exactly as in our analysis of the power method, note that $w = A^{-1}v = \alpha_1 q_1[1 + O(\lambda_1/\lambda_2)]$, since A^{-1} amplifies the q_1 component of v by $1/|\lambda_1|$ which is much bigger than the inverse of all the other eigenvalues. Thus, $w/\|w\|_2 = q_1[1 + O(\lambda_1/\lambda_2)]$.

Second, if we Taylor-expand $(A+\delta A)^{-1}$ in powers of δA, i.e. in powers of $\varepsilon_{\text{machine}}$, we obtain:[1] $(A+\delta A)^{-1} = A^{-1} - A^{-1}\delta A A^{-1} + O(\varepsilon^2_{\text{machine}})$. Since all of the terms in this expansion are multiplied on the *left* by A^{-1}, when multiplied by *any* vector they will again amplify the q_1 component much more than any other component. In particular, the vector $\delta A\, w$ is a vector in a random direction (since δA comes from roundoff and is essentially random) and hence will have some nonzero q_1 component. Thus, $\delta w = -(A+\delta A)^{-1}\delta A\, w = \beta_1 q_1[1 + O(\lambda_1/\lambda_2)]$ for some constant β_1.

Putting these things together, we see that $\tilde{w} = (\alpha_1 + \beta_1)q_1[1 + O(\lambda_1/\lambda_2)]$, and hence $\tilde{w}/\|\tilde{w}\|_2 = q_1[1 + O(\lambda_1/\lambda_2)] = \frac{w}{\|w\|_2}[1 + O(\lambda_1/\lambda_2)]$. Q.E.D.

Problem 4 (5+5+5+5+5 pts):

Trefethen, problem 33.2:

(a) In this case, the q_{n+1} vector is multiplied by a zero row in $\tilde{H}_n$, and we can simplify 33.13 to $AQ_n = Q_n H_n$. If we consider the full Hessenberg reduction, $H = Q^*AQ$, it must have a "block Schur" form:

$$H = \begin{pmatrix} H_n & B \\ 0 & H' \end{pmatrix},$$

where H' is an $(m-n)\times(m-n)$ upper-Hessenberg matrix and $B \in \mathbb{C}^{n\times(m-n)}$. (It is *not* necessarily the case that $B = 0$; this is only true if A is Hermitian.)

(b) Q_n is a basis for $\mathscr{K}_n$, so any vector $x \in \mathscr{K}_n$ can be written as $x = Q_n y$ for some $y \in \mathbb{C}^n$. Hence, from above, $Ax = AQ_n y = Q_n H_n y = Q_n(H_n y) \in \mathscr{K}_n$. Q.E.D.

(c) The $(n+1)$ basis vector, $A^n b$, is equal to $A(A^{n-1}b)$ where $A^{n-1}b \in \mathscr{K}_n$. Hence, from above, $A^n b \in \mathscr{K}_n$ and thus $\mathscr{K}_{n+1} = \mathscr{K}_n$. By induction, $\mathscr{K}_\ell = \mathscr{K}_n$ for $\ell \geq n$.

(d) If $H_n y = \lambda y$, then $AQ_n y = Q_n H_n y = \lambda Q_n y$, and hence λ is an eigenvalue of A with eigenvector $Q_n y$.

(e) If A is nonsingular, then H_n is nonsingular (if it had a zero eigenvalue, A would too from above). Hence, noting that b is proportional to the first column of Q_n, we have: $x = A^{-1}b = A^{-1}Q_n e_1\|b\| = A^{-1}Q_n H_n H_n^{-1}e_1\|b\| = A^{-1}AQ_n H_n^{-1}e_1\|b\| = Q_n H_n^{-1}e_1\|b\| \in \mathscr{K}_n$. Q.E.D.

[1] Write $(A+\delta A)^{-1} = [A(I + A^{-1}\delta A)]^{-1} = (I + A^{-1}\delta A)^{-1}A^{-1} \approx (I - A^{-1}\delta A)A^{-1} = A^{-1} - A^{-1}\delta A A^{-1}$. Another approach is to let $B = (A+\delta A)^{-1} = B_0 + B_1 + \cdots$ where B_k is the k-th order term in δA, collect terms order-by-order in $I = (B_0 + B_1 + \cdots)(A + \delta A) = B_0 A + (B_0\delta A + B_1 A) + \cdots$, and you immediately find that $B_0 = A^{-1}$, $B_1 = -B_0\delta A A^{-1} = -A^{-1}\delta A A^{-1}$, and so on.