18.335 Problem Set 4 Solutions

Problem 1: Q's 'R us (10+15 points)

(a) 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.

(b) Trefethen, problem 28.2:

(i) In general, r_{ij} is nonzero (for $i < j$) if column i is non-orthogonal to column j. For a tridiagonal matrix A, only columns within two columns of one another are non-orthogonal (overlapping in the nonzero entries), so R should only be nonzero (in general) for the diagonals and for two entries above each diagonal; i.e. r_{ij} is nonzero only for $i = j$, $i = j-1$, and $i = j-2$.

Each column of the Q matrix involves a linear combination of all the previous columns, by induction (i.e. $q_2$uses q_1, $q_3$uses q_2 and q_1, q_4 uses q_3 and q_2, q_5 uses q_4 and q_3, and so on). This means that an entry (i,j) of Q is zero (in general) only if $a_{i,1:j} = 0$ (i.e., that entire row of A is zero up to the j-th column). For the case of tridiagonal A, this means that Q will have upper-Hessenberg form.

(ii) **Note:** In the problem, you are told that A is symmetric and tridiagonal. You must also assume that A is real, or alternatively that A is Hermitian and tridiagonal. (This is what must be meant in the problem, since tridiagonal matrices only arise in the QR method if the starting point is Hermitian.) In contrast, if A is complex tridiagonal with $A^T = A$, the stated result is not true (RQ is not in general tridiagonal, as can easily be verified using a random tridiagonal complex A in Matlab).

It is sufficient to show that RQ is upper Hessenberg: since $RQ = Q^*AQ$ and A is Hermitian, then RQ is Hermitian and upper-Hessenberg implies tridiagonal. To show that RQ is upper-Hessenberg, all we need is the fact that R is upper-triangular and Q is upper-Hessenberg.

Consider the (i,j) entry of RQ, which is given by $\sum_k r_{i,k}q_{k,j}$. $r_{i,k} = 0$ if $i > k$ since R is upper triangular, and $q_{k,j} = 0$ if $k > j+1$ since Q is upper-Hessenberg, and hence $r_{i,k}q_{k,j} \neq 0$ only

[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.

when $i \le k \le j+1$, which is only true if $i \le j+1$. Thus the (i,j) entry of RQ is zero if $i > j+1$ and thus RQ is upper-Hessenberg.

(iii) Obviously, if A is tridiagonal (or even just upper-Hessenberg), most of each column is already zero—we only need to introduce one zero into each column below the diagonal. Hence, for each column k we only need to do one 2×2 Givens rotation or 2×2 Householder reflection of the k-th and $(k+1)$-st rows, rotating $\begin{pmatrix} \cdot \\ \cdot \end{pmatrix} \to \begin{pmatrix} \bullet \\ 0 \end{pmatrix}$. Each 2×2 rotation/reflection requires 6 flops (multiping a 2-component vector by a 2×2 matrix), and we need to do it for all columns starting from the k-th. However, actually we only need to do it for 3 columns for each k, since from above the conversion from A to R only introduces one additional zero above each diagonal, so most of the rotations in a given row are zero. That is, the process looks like

$$\begin{pmatrix} \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{pmatrix} \to \begin{pmatrix} \bullet & \bullet & \bullet & & & \\ 0 & \bullet & \bullet & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{pmatrix} \to \begin{pmatrix} \cdot & \cdot & \cdot & & & \\ 0 & \bullet & \bullet & \bullet & & \\ & 0 & \bullet & \bullet & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{pmatrix} \to \begin{pmatrix} \cdot & \cdot & \cdot & & & \\ 0 & \cdot & \cdot & \cdot & & \\ & 0 & \bullet & \bullet & \bullet & \\ & & 0 & \bullet & \bullet & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{pmatrix},$$

where $\bullet$ indicates the entries that change on each step. Notice that it gradually converts A to R, with the two nonzero entries above each diagonal as explained above, and that each Givens rotation need only operate on three columns. Hence, only $O(m)$ flops are required, compared to $O(m^3)$ for ordinary QR! [Getting the exact number requires more care that I won't bother with, since we can no longer sweep under the rug the $O(m)$ operations required to construct the 2×2 Givens or Householder matrix, etc.]

Problem 2: (5+10 points)

Suppose A is a diagonalizable matrix with eigenvectors $\mathbf{v}_k$ and eigenvalues λ_k, in decreasing order $|\lambda_1| \ge |\lambda_2| \ge \cdots$. 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 $A_2 = Q^*AQ$. 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 $X^*XS = X^*AX$ via $RS = Q^*AQR = A_2R$. Thus, $S = R^{-1}A_2R$: 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 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 (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 $AQ_n = Q_nH_n$. 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, $Ax = AQ_ny = Q_nH_ny = Q_n(H_ny) \in \mathscr{K}_n$. Q.E.D.

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

(d) If $H_ny = \lambda y$, then $AQ_ny = Q_nH_ny = \lambda Q_ny$, and hence λ is an eigenvalue of A with eigenvector Q_ny.

(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: $x = A^{-1}b = A^{-1}Q_ne_1\|b\| = A^{-1}Q_nH_nH_n^{-1}e_1\|b\| = A^{-1}AQ_nH_n^{-1}e_1\|b\| = Q_nH_n^{-1}e_1\|b\| \in \mathscr{K}_n$. Q.E.D.