MATH 8610 (SPRING 2021) FINAL EXAM

Assigned 04/28/2021 at 8am, due at 11:30am. Unless there is significant urgent accident with proper documented proof, no submission will be accepted after 11:30am.

The final is closed-book. No references to textbooks, notes, online resources, previous homework are allowed. No communications in *any* form with any person other than myself during the exam. You can use a calculator, pen/pencil and paper. No computers or any electronic devices are allowed. Violations of academic integrity would be reported and handled following the University's Graduate Student Handbook.

1. **[Q1]** Let $A \in \mathbb{R}^{n\times n}$ be real symmetric, indefinite and nonsingular. Consider a *signed* Cholesky factorization $A = LDL^T$, where L is lower triangular, and D is a diagonal matrix with ± 1 diagonal elements. Consider a collection of such matrices, for which $\kappa_2(L) \le C_n$ for some moderate constant $C_n > 0$ (assume n is fixed).
   (a) Show that for these matrices, the Cholesky factor L satisfies $\|L\|_2 \le \sqrt{C_n\|A\|_2}$.
   (b) Suppose a signed Cholesky factorization applied to these matrices gives $\widehat{L}$, and diagonal $\widehat{D}$ with ± 1 entries, such that $A + \Delta A = \widehat{L}\widehat{D}\widehat{L}^T$, with $\kappa_2(\widehat{L}) \le C_n$, and $\frac{\|\Delta A\|_2}{\|\widehat{L}\|_2\|\widehat{L}^T\|_2} = \mathcal{O}(\epsilon_{mach})$. Show this algorithm is backward stable for such matrices.
   (Hint: left and right multiply $A = LDL^T$ by L^{-1} and L, respectively, note that D is orthogonal, and find an upper bound on $\|L^TL\|_2$; also need $\|L\|_2\|L^T\|_2 = \|L^TL\|_2$)

   [A1] From $A = LDL^T$, we have $L^{-1}AL = DL^TL$, and $\|DL^TL\|_2 = \|L^{-1}AL\|_2$. Since D is an orthogonal matrix, it follows that $\|L^TL\|_2 = \|DL^TL\|_2 \le \|L^{-1}\|_2\|A\|_2\|L\|_2 = \kappa_2(L)\|A\|_2 \le C_n\|A\|_2$. Meanwhile, let $L = U\Sigma V^T$ be an SVD of L, and we have $\|L^TL\|_2 = \|U\Sigma V^TV\Sigma U^T\|_2 = \|\Sigma^2\|_2 = \max\{\sigma_i^2\} = \|L\|_2^2 = \|L\|_2\|L^T\|_2$. It follows that $\|L\|_2^2 \le C_n\|A\|_2$, or $\|L\|_2 \le \sqrt{C_n\|A\|_2}$.

   Given the actual *computed* signed Cholesky factorization $A+\Delta A = \widehat{L}\widehat{D}\widehat{L}^T$, the above derivation shows that $\|\widehat{L}\|_2\|\widehat{L}^T\|_2 \le C_n\|A+\Delta A\|_2 \le C_n(\|A\|_2 + \|\Delta A\|_2)$. Therefore, from the known relation $\frac{\|\Delta A\|_2}{\|\widehat{L}\|_2\|\widehat{L}^T\|_2} = \mathcal{O}(\epsilon_{mach})$, we have $\frac{\|\Delta A\|_2}{C_n(\|A\|_2+\|\Delta A\|_2)} = \mathcal{O}(\epsilon_{mach})$, or $\frac{\|\Delta A\|_2}{\|A\|_2+\|\Delta A\|_2} \le \widetilde{C_n}\epsilon_{mach}$. Taking the reciprocal of both sides, subtract 1, then taking the reciprocal again, we have $\frac{\|\Delta A\|_2}{\|A\|_2} \le \frac{\widetilde{C_n}\epsilon_{mach}}{1-\widetilde{C_n}\epsilon_{mach}} = \mathcal{O}(\epsilon_{mach})$ for sufficiently small ϵ_{mach}. This established the backward stability of signed Cholesky factorization algorithm for all symmetric matrices whose Cholesky factors satisfy $\kappa_2(L) \le C_n$.

2. **[Q2]** Let $A \in \mathbb{R}^{m\times n}$ $(m \ge n)$ be of full rank n, with SVD $A = \sum_{j=1}^n \sigma_j u_j v_j^T$, with singular values $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n > 0$. Choose and fix index k $(1 \le k < n)$, define $A_k = \sum_{j=1}^k \sigma_j u_j v_j^T + \sum_{j=k+1}^n \frac{\sigma_{k+1}}{2} u_j v_j^T$, and consider $S = \{B : B \in \mathbb{R}^{m\times n}, \sigma_j(B) \le \frac{\sigma_{k+1}}{2}, k+1 \le j \le n\}$ (similarly, assuming that $\sigma_1(B) \ge \ldots \ge \sigma_n(B)$). Show that

   $$\|A - A_k\|_2 = \inf_{B\in S} \|A - B\|_2.$$

   (Hint: Note that $\|Aw\| \le \|Bw\| + \|(A-B)w\|$, assume that there exists a minimizer $B \ne A_k$, and let w lie in a subspace spanned by certain right singular vectors)

   [A2] By the definition of $A_k \in S$, we have $\|A - A_k\|_2 = \max_{j=k+1}^n \left|\sigma_j - \frac{\sigma_{k+1}}{2}\right| = \frac{\sigma_{k+1}}{2}$. Assume by contradiction that there exists a matrix $B \in S$, such that $\|A - B\|_2 < \frac{\sigma_{k+1}}{2}$. For this matrix, and any vector $w \in \text{span}\{v_{k+1}, \ldots, v_n\}$ (space spanned by the right singular vectors of A corresponding to the smallest singular values) of unit 2-norm, $\|Aw\|_2 \le \|Bw\|_2 + \|(A-B)w\|_2 \le \frac{\sigma_{k+1}}{2} + \|A-B\|_2 < \sigma_{k+1}$. This is a contradiction if we let $w = v_{k+1}$ since $\|Aw\|_2 = \sigma_{k+1}$. In short, there does not exist such a better approximation $B \in S$ to A; that is, $\|A - A_k\|_2 = \inf_{B\in S} \|A - B\|_2$.

3. **[Q3]** Consider the unshifted QR iteration applied to a real symmetric tridiagonal matrix H, described by $Q^{(k)}R^{(k)} = H^{(k-1)}$ and $H^{(k)} = R^{(k)}Q^{(k)}$, with $H^{(0)} = H$. Define $\underline{Q}^{(k)} = Q^{(1)} \cdots Q^{(k)}$ and $\underline{R}^{(k)} = R^{(k)} \cdots R^{(1)}$.

   (a) Is the arithmetic work of each QR iteration $\mathcal{O}(n)$, $\mathcal{O}(n^2)$, or $\mathcal{O}(n^3)$, and why?

   (b) With $H^k = \underline{Q}^{(k)}\underline{R}^{(k)}$, show that under certain mild assumptions, the first and the last column of $\underline{Q}^{(k)}$ converge to the eigenvector of H associated with the largest and the smallest (modulus) eigenvalues, respectively.

   (c) Now consider the *shifted* QR iteration. Assume that the bottom-right 3×3 block of $H^{(k)}$ is $\begin{bmatrix} \times & \eta a & \\ \eta a & a+b & \delta \\ & \delta & b \end{bmatrix}$, with $|\delta|$ sufficiently small, $|a|$ not very small, and $\eta \neq 0$. Assume that the shift $\mu^{(k+1)} = b$ is used to transform $H^{(k)}$ to $H^{(k+1)}$. Give an upper bound on the $(n, n-1)$ entry of $H^{(k+1)}$ in modulus. What does this imply?

   [A3] (a) If each $H^{(k)}$ is a tridiagonal matrix, then each Givens rotation applied to the left side of $H^{(k)}$ only changes at most 6 entries, and so does each Givens rotation applied to the right side of $R^{(k)}$. Therefore, in each QR iteration, the total arithmetic cost to perform QR and compute RQ by $n-1$ Givens rotations is $\mathcal{O}(n)$.

   (b) We multiply both sides of $H^k = \underline{Q}^{(k)}\underline{R}^{(k)}$ by e_1 on the right, and obtain $H^k e_1 = \underline{Q}^{(k)}\underline{R}^{(k)}e_1 = r_{11}^{(k)}\underline{Q}^{(k)}e_1 = r_{11}^{(k)}\underline{q}_1^{(k)}$ (a scalar multiple of the first column of the accumulated Q factor). Since $H^k e_1$ is the vector obtained in the k-th step of the power method with matrix H and starting vector e_1, it typically converges toward the eigenvector associated with the largest (in modulus) eigenvalue of H, if the eigenvalue of such largest modulus is unique, and e_1 has a nonzero component of this eigenvector; the right-hand side $\underline{q}_1^{(k)}$ is the first column of the accumulated Q factor.

   Similarly, taking the inverse transpose of $H^k = \underline{Q}^{(k)}\underline{R}^{(k)}$ gives $H^{-k} = \underline{Q}^{(k)}\left(\underline{R}^{(k)}\right)^{-T}$. Multiplying both sides by e_n, we have $H^{-k}e_n = \underline{Q}^{(k)}\left(\underline{R}^{(k)}\right)^{-T}e_n = \frac{1}{r_{nn}^{(k)}}\underline{Q}^{(k)}e_n = \frac{1}{r_{nn}^{(k)}}\underline{q}_n^{(k)}$. Since $H^{-k}e_n$ is the vector obtained in the k-th step of the inverse power method with H and starting vector e_n, it usually converges toward the eigenvector associated with the smallest (modulus) eigenvalue of H, under similar mild assumptions; the right-hand side $\underline{q}_n^{(k)}$ is the last column of the accumulated Q factor.

   (c) To QR factorize the tridiagonal matrix, we first obtain $H^{(k)} - bI$, which has the right bottom 3-by-3 block $\begin{bmatrix} \times & \eta a & \\ \eta a & a & \delta \\ & \delta & 0 \end{bmatrix}$. Then, we apply $n-2$ Givens rotations on the left side of $H^{(k)} - bI$, and note that the $(n-2)$-nd Givens rotation is not identity because the $(n-1, n-2)$ entry is $\eta a \neq 0$. As a result, right before applying the last Givens rotation on the left, we have the temporary matrix $H_{tmp}^{(k)} = \begin{bmatrix} \times & \times & \times \\ \times & \widehat{a} & \widehat{\delta} \\ & \delta & 0 \end{bmatrix}$, where $|\widehat{a}| \leq \sqrt{1+\eta^2}|a|$ and $\widehat{\delta} \leq |\delta|$ as a result of the $(n-2)$-nd Givens rotation applied on the left side (which does not change the 2-norm of the vectors $H_{tmp}^{(k)}(n-2:n-1, n-1)$ and $H_{tmp}^{(k)}(n-2:n-1, n)$). Then, following HW4 Q5(b), the $(n, n-1)$ entry of $H^{(k+1)}$ is $H_{n,n-1}^{(k+1)} = -\frac{\widehat{\delta}\delta^2}{\widehat{a}^2+\delta^2}$. Since we assumed that $|a|$ is not very small and $|\delta|$ is sufficiently small, unless $\widehat{a}$ happens to be very small, we have $|H_{n,n-1}^{(k+1)}| = \mathcal{O}(\delta^3)$, which establishes cubic convergence of the $(n, n-1)$ entry as the QR iteration proceeds.

4. **[Q4]** Consider the Arnoldi relation $AU_k = U_{k+1}\underline{H}_k$, with $U_k^T U_k = I$, $\underline{H}_k \in \mathbb{R}^{(k+1)\times k}$. Let (μ, w) be an eigenpair of H_k (the top k rows of $\underline{H}_k$).

   (a) Show that $(\mu, U_k w)$ satisfies $AU_k w - \mu U_k w \perp \mathcal{K}_k(A, u_1)$, and $\|AU_k w - \mu U_k w\|_2 = |h_{k+1,k} w(k)|$, where $w(k)$ is the last element of w.

   (b) What happens if $\mathrm{col}(U_k)$ is an invariant subspace of A, i.e., $\mathrm{col}(AU_k) \subset \mathrm{col}(U_k)$? Under what condition(s) for u_1 would this scenario happen?

   [A4] (a) The eigenpair (μ, w) satisfies $H_k w = \mu w$. We multiply both sides of the Arnoldi relation on the right by w and get $AU_k w = U_k H_k w + h_{k+1,k} u_{k+1} e_k^T w = \mu U_k w + h_{k+1,k} u_{k+1} e_k^T w$. It follows that the eigenresidual vector $AU_k w - \mu U_k w = h_{k+1,k} u_{k+1} e_k^T w = h_{k+1,k} w(k) u_{k+1}$ is a scalar multiple of vector u_{k+1}. By the construction of Arnoldi's method, $u_{k+1} \perp \mathrm{span}\{u_1, \ldots, u_k\} = \mathcal{K}_k(A, u_1)$. Also, $\|AU_k w - \mu U_k w\|_2 = \|h_{k+1,k} w(k) u_{k+1}\|_2 = |h_{k+1,k} w(k)|$ because $\|u_{k+1}\text{-}2 = 1$.

   (b) If $\mathrm{col}(AU_k) \subset \mathrm{col}(U_k)$, then $\mathcal{K}_{k+1}(A, u_1) = \mathrm{span}\{u_1\} + A\mathcal{K}_k(A, u_1) = \mathrm{span}\{u_1\} + \mathrm{col}(AU_k) \subset \mathrm{col}(U_k) = \mathcal{K}_k(A, u_1)$. Meanwhile, since $\mathcal{K}_k(A, u_1) \subset \mathcal{K}_{k+1}(A, u_1)$, we have $\mathcal{K}_k(A, u_1) = \mathcal{K}_{k+1}(A, u_1)$; that is, the dimension of $\mathcal{K}_k(A, u_1)$ will no longer increase with k. This would happen if the starting vector u_1 is a linear combination of eigenvectors of A associated with k distinct eigenvalues.

5. **[Q5]** Let $r_0 = b - Ax_0$ be the initial residual vector of the linear system $Ax = b$, and $r_k = r_0 - Az_k$ with $z_k \in \mathcal{K}_k(A, r_0)$.

   (a) Show that $p_k \perp A\mathcal{K}_k(A, r_0)$ for CG, and $r_k \perp A\mathcal{K}_k(A, r_0)$ for GMRES. As a result, show that $(r_j, p_k) = (r_k, p_k)$ for CG, and $(r_j, r_k) = (r_k, r_k)$ for GMRES $(1 \le j < k)$.

   (b) Let $AU_k = U_{k+1}\underline{H}_k$ be the Lanczos/Arnoldi relation for solving $Ax = b$, where $u_1 = \frac{r_0}{\|r_0\|_2}$. Let the k-th iterate of CG or GMRES be $x_k = x_0 + U_k y_k$. Show that $H_k y_k = \|r_0\|_2 e_1$ for CG, whereas $\underline{H}_k^T \underline{H}_k y_k = \|r_0\|_2 \underline{H}_k^T e_1$ for GMRES.

   (Hint: for GMRES, consider the normal equation for the linear least squares)

   [A5] (a) For CG, we have $\mathrm{span}\{r_0, \ldots, r_{k-1}\} = \mathrm{span}\{p_0, \ldots, p_{k-1}\} = \mathcal{K}_k(A, r_0)$, with $p_0 = r_0$. Since $p_i^T A p_j = 0$ for all $i \ne j$, we have $p_k \perp \mathrm{span}\{Ap_0, \ldots, Ap_{k-1}\} = A\mathcal{K}_k(A, r_0)$. Recall that for Krylov subspace methods, the residual at the k-th step is $r_k = r_0 - Az_k$ with $z_k \in \mathcal{K}_k(A, r_0)$. Note that $r_k - r_j = (r_0 - Az_k) - (r_0 - Az_j) = A(z_j - z_k) \in A\mathcal{K}_k(A, r_0)$ because $1 \le j < k$. It follows that $r_k - r_j \perp p_k$, which is equivalent to $(r_j, p_k) = (r_k, p_k)$. For GMRES, since the residual $r_k = r_0 - Az_k = r_0 - AU_k y_k$ is minimized in 2-norm for all $y \in \mathbb{R}^k$, the least squares condition we learned from Chapter 2 states that y_k must satisfy $r_k = r_0 - AU_k u_k \perp \mathrm{col}(AU_k) = A\mathcal{K}_k(A, r_0)$. Replacing p_k we have shown for CG with r_k of GMRES, we obtain $(r_j, r_k) = (r_k, r_k)$.

   For CG, $r_k \perp \mathrm{span}\{r_0, \ldots, r_{k-1}\} = \mathcal{K}_k(A, r_0) = \mathrm{col}(U_k)$. Therefore, $r_k = r_0 - Az_k = U_k e_1 \|r_0\|_2 - AU_k y_k = U_k e_1 \|r_0\|_2 - U_k H_k y_k - h_{k+1,k} u_{k+1} e_k y_k$ satisfies

   $$U_k^T r_k = U_k^T \left(U_k e_1 \|r_0\|_2 - U_k H_k y_k - h_{k+1,k} u_{k+1} e_k y_k\right) = e_1 \|r_0\|_2 - H_k y_k = 0,$$

   which gives $H_k y_k = \|r_0\|_2 e_1$.
   For GMRES, we have $r_k = U_{k+1} e_1 \|r_0\|_2 - U_{k+1} \underline{H}_k y_k = U_{k+1}(e_1 \|r_0\|_2 - \underline{H}_k y_k)$ satisfying $r_k \perp \mathrm{col}(AU_k) = \mathrm{col}(U_{k+1}\underline{H}_k)$. This leads to

   $$(AU_k)^T = \left(U_{k+1}\underline{H}_k\right)^T r_k = \underline{H}_k^T (e_1 \|r_0\|_2 - \underline{H}_k y_k) = 0,$$

   which gives $\underline{H}_k^T \underline{H}_k y_k = \|r_0\|_2 \underline{H}_k^T e_1$.