Mathematics 372 – Numerical Linear Algebra
Midterm Problem Set Solutions
March 23, 2007

I. (More on Gaussian elimination and LU factorizations).

A) (10) Show that an $n \times n$ matrix A has an LU factorization (equivalently, Gaussian elimination is possible with no row interchanges) if and only if for each $1 \leq k \leq n$, the upper left corner $k \times k$ submatrix

$$A_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}$$

is nonsingular.

Solution: $\Rightarrow$: First assume that Gaussian elimination on A is possible with no row interchanges. We will argue by induction on k. The base case is $k = 1$. If we do not have to interchange rows to find a pivot element in column 1, then $a_{11} = \det(A_1) \neq 0$. So the base case is established. Now assume that we have shown that $\det(A_\ell) \neq 0$ for $\ell < k$. We do the row operations (without row interchanges) to create the zeroes below the main diagonals in columns 1 through ℓ. Since each of these row operations is of type 1 (replacing a row by that row plus a multiple of another row), the determinants of the matrix A and of the principal minors are not changed. In particular, the determinant of the minor $A_{\ell+1}$ will be the same as the determinant of the "eliminated" $(\ell + 1) \times (\ell + 1)$ matrix:

$$\begin{pmatrix} a_{11} & a_{12} & a'_{13} & \cdots & a'_{1\ell} & a_{1,\ell+1} \\ 0 & a'_{22} & a'_{23} & \cdots & a'_{2\ell} & a'_{2,\ell+1} \\ 0 & 0 & a'_{33} & \cdots & a'_{3\ell} & a'_{3,\ell+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a'_{\ell\ell} & a'_{\ell,\ell+1} \\ 0 & 0 & 0 & \cdots & 0 & a'_{\ell+1,\ell+1} \end{pmatrix}$$

The induction hypothesis is that

$$\det(A_\ell) = a_{11}a'_{22} \cdots a'_{\ell\ell} \neq 0.$$

If we can use the diagonal entry as pivot in column $\ell + 1$, then $a_{\ell+1,\ell+1} \neq 0$. Hence

$$\det(A_{\ell+1}) = \det(A_\ell)a'_{\ell+1,\ell+1} \neq 0$$

also.

$\Leftarrow$: For the converse, we essentially just argue in the reverse direction. If all the $\det(A_k) \neq 0$ for $1 \leq k \leq n$, then doing the reductions as above, after completing each column, the diagonal entry in the next column must be nonzero. Hence we never have to interchange rows in the Gaussian elimination.

B) (10) Suppose that A is a matrix that satisfies the condition in part A, and for any k, $1 \leq k \leq n$, partition A as

$$A = \begin{pmatrix} A_k & B \\ C & D \end{pmatrix},$$

where B is $k \times (n-k)$, C is $(n-k) \times k$, and D is $(n-k) \times (n-k)$. Show that there is a unique $(n-k) \times k$ matrix M such that

$$\begin{pmatrix} I_k & 0 \\ -M & I_{n-k} \end{pmatrix} \begin{pmatrix} A_k & B \\ C & D \end{pmatrix} = \begin{pmatrix} A_k & \hat{B} \\ 0 & \hat{D} \end{pmatrix}$$

for some matrices $\hat{B}$ and $\hat{D}$. Also show that $\hat{D} = D - CA_k^{-1}B$.

Solution: Matrices with these block decompositions can be multiplied using the same formulas as for 2×2 scalar matrices. In particular the entry in the "row 2, column 1" of the product on the left side of the equation to be proved is

$$-MA_k + I_{n-k}C = -MA_k + C.$$

Because A_k is assumed to be invertible, there is only one matrix that makes this equation true, namely:

$$(*) \qquad M = CA_k^{-1}.$$

The entry in "row 1, column 2" of the product is $\hat{B} = B$. From (*), $-MB + I_{n-k}D = \hat{D}$, so $\hat{D} = D - MB = D - CA_k^{-1}B$ as desired.

C) (5) Deduce that A has the *block LU decomposition*

$$A = \begin{pmatrix} I_k & 0 \\ M & I_{n-k} \end{pmatrix} \begin{pmatrix} A_k & \hat{B} \\ 0 & \hat{D} \end{pmatrix}.$$

Solution: This follows because

$$\begin{pmatrix} I_k & 0 \\ -M & I_{n-k} \end{pmatrix}^{-1} = \begin{pmatrix} I_k & 0 \\ M & I_{n-k} \end{pmatrix}$$

(check by direct matrix multiplication).

D) (5) By definition, the matrix A_k satisfies the condition in part A. It can be shown that $\hat{D}$ does as well. Hence each of these matrices also has an LU-decomposition:

$A_k = L_1 U_1$, and $\hat{D} = L_2 U_2$. Express the LU-decomposition of A in terms of L_1, L_2, M, U_1, U_2, and B.

Solution: We have, by regrouping terms using associativity of matrix multiplication,

$$\begin{aligned} A &= \begin{pmatrix} I_k & 0 \\ M & I_{n-k} \end{pmatrix} \begin{pmatrix} A_k & B \\ 0 & \hat{D} \end{pmatrix} \\ &= \begin{pmatrix} I_k & 0 \\ M & I_{n-k} \end{pmatrix} \begin{pmatrix} L_1 U_1 & B \\ 0 & L_2 U_2 \end{pmatrix} \\ &= \begin{pmatrix} I_k & 0 \\ M & I_{n-k} \end{pmatrix} \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix} \begin{pmatrix} U_1 & L_1^{-1} B \\ 0 & U_2 \end{pmatrix} \\ &= \begin{pmatrix} L_1 & 0 \\ M L_1 & L_2 \end{pmatrix} \begin{pmatrix} U_1 & L_1^{-1} B \\ 0 & U_2 \end{pmatrix}. \end{aligned}$$

This is an ordinary LU factorization for A.

II. (The matrix 2-norm, revisited.) By the general theory of the matrix operator norms associated to vector norms on $\mathbf{R}^n$, the matrix 2-norm of $A \in M_{n \times n}(\mathbf{R})$ is given by

$$\|A\|_2 = \max_{\|u\|_2 = 1} \|Au\|_2. \tag{1}$$

Your goal in this problem is to prove the formula

$$\|A\|_2 = \sqrt{\max\{|\lambda| : \lambda \text{ is an eigenvalue of } A^t A\}} \tag{2}$$

that we mentioned in class. The equation (1) shows that we need to solve a *constrained optimization problem* to determine $\|A\|_2$.

A) (5) What standard method from multivariable calculus applies to solve constrained optimization problems of the following form:

$$\begin{aligned} \text{maximize} &: f(x_1, \ldots, x_n) \\ \text{subject to} &: g(x_1, \ldots, x_n) = 0? \end{aligned}$$

Describe how the method works in general.

Solution: The method is the *Lagrange multiplier* method for constrained optimization. One way to say the process is that you consider the system of $n + 1$ equations in $n + 1$ variables $x_1, \ldots, x_n, \lambda$ obtained from the components of the vector equation $\nabla f = \lambda \nabla g$, together with the gradient equation. The maximum on the constraint set will be among the solutions of this system.

B) (15) For our purposes in (1), it will be most convenient to square both the function $\|Au\|_2$ we seek to maximize and the constraint equation $\|u\|_2 = 1$. Writing $u =$

$(u_1, \ldots, u_n)^t$, apply the method with $f(u_1, \ldots, u_n) = \|Au\|_2^2$ and $g(u_1, \ldots, u_n) = u_1^2 + \cdots + u_n^2 - 1$. Show that the equations for the constrained maximum imply that the vector u where f achieves its maximum must be an eigenvector of A^tA, and that the multiplier λ in the method must be the corresponding eigenvalue.

Solution: With

$$f = \|Au\|_2^2 = \langle Au, Au \rangle = \langle u, A^tAu \rangle$$

and $g = \|u\|_2^2 = \langle u, u \rangle$, in the Lagrange equations from $\nabla f = \lambda \nabla g$ we have the equations

$$\frac{\partial f}{\partial u_i} = \lambda \frac{\partial g}{\partial u_i}$$

for $i = 1, \ldots, n$. From the form of g, the right side is easy: $2\lambda u_i$. The left side is somewhat messy, but simplest if you use the form $f = \langle u, A^tAu \rangle$. When we take this expression and compute the derivative with respect to u_i, we get 2 times the ith component of A^tAu. Hence the equation $\nabla f = \lambda \nabla g$ is just $2A^tAu = 2\lambda u$, or $A^tAu = \lambda u$. Since u is a unit vector from the constraint equation, it is nonzero. Hence u is an eigenvector of A^tA with eigenvalue λ.

C) (5) Deduce that (2) holds, giving the formula for $\|A\|_2$.

Solution: From the previous part, the solutions of the Lagrange equations will consist of (u, λ) where u is a unit eigenvector of A^tA, and λ is the corresponding eigenvector. The maximum is obtained for the one of these that gives the largest value for f. But notice,

$$f(u) = \langle Au, Au \rangle = \langle u, A^tAu \rangle = \langle u, \lambda u \rangle = \lambda \langle u, u \rangle = \lambda.$$

In other words, *the value of f at the eigenvector u is just the corresponding eigenvalue of A^tA.* Hence the maximum value of f will be

$$\max\{\lambda : \lambda \text{ is an eigenvalue of } A^tA\}$$

(note the absolute value is actually unnecessary because A^tA is symmetric and positive semidefinite, so all the eigenvalues are reals with $\lambda \geq 0$). The desired equation (2) comes from taking square roots:

$$\|A\|_2 = \sqrt{\max\{\lambda : \lambda \text{ is an eigenvalue of } A^tA\}}.$$

D) (10) Recall the 60×60 adjacency matrix B of the "bucky ball" from Lab/Problem Set 3. Use (2) above and MATLAB to compute $\|B\|_2$. Some technical notes: (a) The MATLAB command `eig` returns a list of all the eigenvalues of a square matrix; (b) the built-in matrix `bucky` is a sparse matrix. If you want to check your answer using the `norm` command, you will need to convert `bucky` to "full" form.

Solution: See MATLAB solutions.

E) (10) If $A = (a_{ij})$, the matrix Frobenius norm is defined by

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2}$$

(note: the absolute values here are unnecessary for real matrices, but necessary if some of the entries in A are complex numbers with nonzero imaginary parts). Show that for all matrices A

$$\|A\|_2 \le \|A\|_F \le \sqrt{n}\|A\|_2 \tag{3}$$

Solution: For the first inequality, subdivide A into its rows and call these vectors $A_1, \ldots, A_n$. Then for some unit vector u,

$$\begin{aligned}
\|A\|_2^2 &= \|Au\|_2^2 \\
&= \langle A_1, u\rangle^2 + \cdots + \langle A_n, u\rangle^2 \\
&\le \|A_1\|_2^2\|u\|_2^2 + \cdots + \|A_n\|_2^2\|u\|_2^2 \quad \text{(Cauchy} - \text{Schwarz)} \\
&= \|A_1\|_2^2 + \cdots + \|A_n\|_2^2 \quad (\text{since } \|u\|_2 = 1)
\end{aligned}$$

This last line is the sum of the squares of all the entries in the matrix, which is $\|A\|_F^2$. Hence $\|A\|_2 \le \|A\|_F$ for all $n \times n$ matrices A.
For the second inequality, split A into its columns, $A^{(1)}, \ldots, A^{(n)}$. Then the Frobenius norm squared is the sum of the squares of the 2-norms of the columns. Using the fact that $A^{(j)} = Ae_j$ and standard properties of matrix norms,

$$\begin{aligned}
\|A\|_F^2 &= \|A^{(1)}\|_2^2 + \ldots + \|A^{(n)}\|_2^2 \\
&= \|Ae_1\|_2^2 + \ldots + \|Ae_n\|_2^2 \\
&\le \|A\|_2\|e_1\|_2^2 + \ldots + \|A\|_2\|e_1\|_2^2 \\
&\le n\|A\|_2
\end{aligned}$$

(since $\|e_i\|_2^2 = 1$ for all of the standard basis vectors e_i). Hence taking square roots, $\|A\|_F \le \sqrt{n}\|A\|_2$.

F) (5) Compute the Frobenius norm of the bucky ball matrix B using MATLAB and verify that (3) holds in this case.

Solution: See MATLAB solutions.

III. (Roundoff considerations and least squares.) In many situations, the way a problem is set up can have a significant effect on the susceptibility to roundoff errors when we apply standard methods. But fortunately, there are often ways to minimize those effects. For instance, suppose that we want to compute the best-fit line for the following data points

x_i	1.001	1.002	1.003	1.004	1.005	1.006	1.007
y_i	3.664	3.789	3.891	4.022	4.233	5.200	5.329

by solving the normal equations $X^tX \begin{pmatrix} m \\ b \end{pmatrix} = X^tY$ for the least squares problem.

A) (5) Using MATLAB, find the ∞-norm condition number of the coefficient matrix X^tX in the system of normal equations. What does your answer indicate about this system of equations?

   Solution: See MATLAB solutions.

B) (10) Our standard way of setting up the normal equations corresponds to using the basis $\{x, 1\}$ for the vector space of linear polynomials. But we could use any other basis we like too. Suppose we use $\{300(x - 1.004), 1\}$ instead (that is, translate and rescale the x_i values). What is the ∞-norm condition number of the coefficient matrix in the normal equations now? Show that the best-fit line can still be computed if we set the problem up this way.

   Solution: See MATLAB solutions.

C) (5) Explain your results in A and B geometrically.

   Solution: The idea is that the matrix X^tX is nearly singular (this is what the large condition numbers mean. The lines in $\mathbf{R}^2$ coming from the two equations in the system $X^tX \begin{pmatrix} m \\ b \end{pmatrix} = X^tY$ are nearly parallel. By rescaling (multiplying the x-coordinates by 300), the slopes are made much *smaller*, so perturbing the lines shifts the position of the intersection less in a relative sense. This is indicated by the smaller condition number of the second matrix.