MATH 387 ASSIGNMENT 2

SAMPLE SOLUTIONS BY IBRAHIM AL BALUSHI

Problem 4

A matrix $A = [a_{ik}] \in \mathbb{R}^{n\times n}$ is called *symmetric* if $a_{ik} = a_{ki}$ for all i, k, and is called *positive definite* if $x^T Ax \geqslant 0$ for all $x \in \mathbb{R}^n$, with $x^T Ax = 0$ only when $x = 0$. Suppose that $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite.

(a) Show that $a_{ii} > 0$ for all i.
(b) Show that $\max_i a_{ii} = \max_{i,k} |a_{ik}|$.
(c) Let $A_k = [a^{(k)}_{ij}]$ be the matrix that enters in the k-th step of the Gaussian elimination process (with $A_1 = A$). Show that for each $k = 1, \dots, n$, the submatrix $[a^{(k)}_{ij}]_{k\leqslant i,j\leqslant n}$ is symmetric and positive definite. Conclude that Gaussian elimination does not break down (hence in particular, that A is invertible).
(d) Show that $a^{(k)}_{ii} \leqslant a^{(k-1)}_{ii}$ for $k \leqslant i \leqslant n$ and for all $k = 2, 3, \dots, n$. Conclude that for Gaussian elimination in exact arithmetics, the growth factor is 1. Note that in exact arithmetics, the growth factor would be defined by

$$g(A) = \frac{\max_{i,j,k} |a^{(k)}_{ij}|}{\max_{i,j} |a_{ij}|}.$$

Solution

(a) Let $\boldsymbol{e}_j \in \mathbb{R}^n$ be jth canonical basis vector for $\mathbb{R}^n$.

$$\boldsymbol{e}_j^T A\boldsymbol{e}_j = a_{jj} > 0 \quad \forall j = 1, ..., n.$$

(b) Let $x = \boldsymbol{e}_i - \alpha\boldsymbol{e}_j$ for some $\alpha \in \mathbb{R}$.

$$x^T Ax = \boldsymbol{e}_i^T A(\boldsymbol{e}_i - \alpha\boldsymbol{e}_j) - \alpha\boldsymbol{e}_j^T A(\boldsymbol{e}_i - \alpha\boldsymbol{e}_j) = a_{ii} - 2\alpha a_{ij} + \alpha^2 a_{jj}.$$

Suppose that some $i \neq j$ the quantity $|a_{ij}|$ is maximal. The entry a_{ij} cannot be zero; otherwise it will contradict the assumption.

$$x^T Ax = a_{ii} - \alpha a_{ij} + \alpha(\alpha a_{jj} - a_{ji})$$

If a_{ij} is positive then pick $\alpha = 1$ and obtain

$$x^T Ax = (a_{ii} - a_{ij}) + (a_{jj} - a_{ji}) < 0$$

Date: Winter 2016.

whereas if a_{ji} is negative pick $\alpha = -1$ and obtain

$$x^T A x = a_{ii} + a_{ij} + (a_{jj} + a_{ji}) < 0$$

(because diagonal entries are always positive).

(c) We may write $L_j = I - \ell_j \boldsymbol{e}_j^T$ where $\ell_j = \left(\mathbf{0}, \frac{x_{j+1,j}}{x_{jj}}, ..., \frac{x_{nj}}{x_{jj}}\right)^T \in \mathbb{R}^n$.

$$L_j L_{j+1} = (I - \ell_j \boldsymbol{e}_j^T)(I - \ell_{j+1}\boldsymbol{e}_{j+1}^T) = I - \ell_j \boldsymbol{e}_j^T - \ell_{j+1}\boldsymbol{e}_{j+1}^T$$

because $\boldsymbol{e}_j^T \ell_{j+1} = 0$. Therefore at any $k-1$th step,

$$L_1 \cdots L_{k-1} = I - \ell_j \boldsymbol{e}_j^T - \cdots - \ell_{k-1}\boldsymbol{e}_{k-1}^T$$

and $[L_1 \cdots L_{k-1}]_{k \leqslant i,j \leqslant n} = \boldsymbol{O} \in \mathbb{R}^{k \times k}$ zero matrix, so the symmetry of kth step submatrix $[a_{ij}^{(k)}]_{k \leqslant i,j \leqslant n}$ remains unchanged. As for positive definiteness, it follows from the fact

$$x^T A x \geqslant 0 \;\forall x \in \{x \in \mathbb{R}^n : x_j = 0 \;\forall 1 \leqslant j < k\}.$$

It follows that at each step $a_{ii}^{(k)} > 0$ for every $k \leqslant i \leqslant n$ and therefore the resulting matrix $A_n = LA$, with $L = L_1 \cdots L_n$, is upper triangular with non-zero diagonal entries; $\det(A_n)$ is nonzero and $A^{-1} = A_n^{-1} L^{-1}$.

(d) By direct computation: Let $k \leqslant i \leqslant n$. $[\ell_{k-1}\boldsymbol{e}_{k-1}^T]_{ii} = a_{i,k-1}^{(k-1)} / a_{k-1,k-1}^{(k-1)}$

$$\begin{aligned}
a_{ii}^{(k)} &= [L_{k-1}A_{k-1}]_{ii} \\
&= a_{ii}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{i,k-1}^{(k-1)} \\
&= a_{ii}^{(k-1)} - \frac{\left(a_{i,k-1}^{(k-1)}\right)^2}{a_{k-1,k-1}^{(k-1)}} \\
&\leqslant a_{ii}^{(k-1)}.
\end{aligned}$$

It follows that

$$g(A) = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|} \leqslant \frac{\max_{i,j} |a_{ij}^{(1)}|}{\max_{i,j} |a_{ij}|} = 1.$$

Problem 6

(a) Let U be an upper triangular matrix with no zeroes on its diagonal. Let $\tilde{x} \in \mathbb{R}^n$ be the result of back-substitution applied to the system $Ux = b$ in floating point arithmetic (with the "machine epsilon" $\varepsilon > 0$). Show that there exists an upper triangular matrix $\tilde{U}$, such that $\tilde{U}\tilde{x} = b$ in exact arithmetics and that the entries of $\tilde{U} - U$ can be bounded in absolute value by an expression depending only on ε, n, and U. Argue that back-substitution is backward stable.

(b) Recall that Gaussian elimination in floating point arithmetics produces matrices $\tilde{L}$ and $\tilde{U}$, where $\tilde{L}$ is lower triangular with unit diagonal and $\tilde{U}$ is upper triangular, satisfying

$$\|\tilde{L}\tilde{U} - A\|_\infty \leqslant \frac{3ng\varepsilon}{(1-\varepsilon)^2}\|A\|_\infty.$$

Turn this into the following bound

$$\|\tilde{L}\tilde{U} - A\| \leqslant C_n g\varepsilon\|A\|, \qquad \text{for all small } \varepsilon,$$

where $\|\cdot\|$ is the matrix norm induced by the Euclidean norm in $\mathbb{R}^n$. In particular, try get a near-optimal value for the constant C_n.

(c) By combining the preceding two results, perform a backward error analysis of the Gaussian elimination process for solving the equation $Ax = b$. That is, complete the analysis we did in class by taking into account the round-off errors of the forward elimination (solution of $\tilde{L}y = b$) and back substitution (solution of $\tilde{U}x = y$).

Solution

(a) Entries of matrix $U \in \mathbb{R}^{n\times n}$ are given by u_{ij}. Backward substitution algorithm is given by

$$x_j = \left(b_j - \sum_{k=j+1}^{n} x_k u_{jk}\right)\Big/ u_{jj}, \quad j = n, n-1, ..., 1.$$

Let $y_j = b_j - \sum_{k=j+1}^{n} x_k u_{jk}$. Per iteration we do $\tilde{x}_j = y_j \oplus u_{jj}$. Axiom on floating point operations: for some $|\varepsilon_j| \leqslant \varepsilon_{\text{mac}}$ we have $\tilde{x}_j = \frac{y_j}{u_{jj}}(1+\varepsilon_j)$. We want to quantify a perturbation matrix δU of U. Note that

$$1 + \varepsilon_j = \frac{1}{1+\varepsilon_j'} \iff \varepsilon_j' = \frac{-\varepsilon_j}{1+\varepsilon_j}$$

then $\varepsilon_j' = -\varepsilon_j\left(\frac{1}{1-(-\varepsilon_j)}\right) = -\varepsilon_j + \mathcal{O}(\varepsilon_j^2)$ which implies $|\varepsilon_j'| \leqslant \varepsilon_{\text{mac}} + \mathcal{O}(\varepsilon_{\text{mac}}^2)$. So if $y = \frac{1}{x}(1+\varepsilon)$ then $y = \frac{1}{x(1+\varepsilon')}$ for $\varepsilon' = -\varepsilon + \mathcal{O}(\varepsilon^2)$. The computation goes as follows:

$$\tilde{x}_n = b_n \oplus u_{nn} = \frac{b_n}{u_{nn}}(1+\varepsilon_1)$$

where by the previous remark may be written $\tilde{x}_n$ as $\frac{b_n}{u_{nn}(1+\varepsilon_1')}$ for some $|\varepsilon_1'| \leqslant \varepsilon_{\text{mac}} + \mathcal{O}(\varepsilon_{\text{mac}}^2)$.

$$\tilde{x}_{n-1} = \left[b_{n-1} \ominus (\tilde{x}_n \otimes u_{n-1,n})\right] \oplus u_{n-1,n-1}$$

We may write $b_{n-1} \ominus \tilde{\Sigma}_{n-1} = (b_{n-1} - \tilde{\Sigma}_{n-1})(1+\epsilon_1)$ and $\tilde{\Sigma}_{n-1} = \tilde{x}_n \otimes u_{n-1,n} = \tilde{x}_n u_{n-1,n}(1+\eta_1)$ so

$$\tilde{x}_{n-1} = \frac{b_{n-1} - \tilde{x}_n u_{n-1,n}(1+\eta_1)}{u_{n-1,n-1}(1+\varepsilon_2')(1+\epsilon_1')}.$$

The algorithm terminates for $j = 1$.

$$\tilde{x}_1 = \left[b_1 \ominus \left(\bigoplus_{k=2}^{n} \tilde{x}_k \otimes u_{1k} \right) \right] \oslash u_{11} = \left[b_1 \ominus \left(\bigoplus_{k=2}^{n} \tilde{x}_k \otimes u_{1k} \right) \right] \Big/ u_{11}(1 + \varepsilon_1')$$

It is important to recognize the nesting nature of carrying a sequence of floating point operations when we deal with $\bigoplus_{k=2}^{n}$. Observe that

$$\begin{aligned} a \oplus b \oplus c &= (a \oplus b) \oplus c \\ &= [(a + b)(1 + \epsilon_1)] \oplus c \\ &= ([(a + b)(1 + \epsilon_1)] + c)(1 + \epsilon_2). \end{aligned}$$

Rewrite into

$$b_1 \ominus \left[\bigoplus_{k=2}^{n} \tilde{x}_k \otimes u_{1k} \right] = \left[b_1 \ominus (\tilde{x}_2 \otimes u_{12}) \right] \bigominus_{k=3}^{n} \tilde{x}_k \otimes u_{1k}$$

and $b_1 \ominus (\tilde{x}_2 \otimes u_{12}) = (b_1 - \tilde{x}_2 \otimes u_{12})(1 + \epsilon_2)$ so we rewrite into

$$= \left[b_1 - \tilde{x}_2 \otimes u_{12} \right] \bigominus_{k=3}^{n} \tilde{x}_k \otimes u_{1k}(1 + \epsilon_2) \Big/ (1 + \epsilon_2') \quad \text{for } |\epsilon_2'| \leqslant \varepsilon_{\text{mac}} + \mathcal{O}(\varepsilon_{\text{mac}}^2).$$

We arrive at

$$\tilde{x}_1 = \left\{ \left[b_1 - \tilde{x}_2 \otimes u_{12} \right] \bigominus_{k=3}^{n} \tilde{x}_k \otimes u_{1k}(1 + \epsilon_2) \right] \right\} \Big/ u_{11}(1 + \varepsilon_1')(1 + \epsilon_2').$$

Again,

$$\tilde{x}_1 = \left\{ \left[b_1 - \tilde{x}_2 \otimes u_{12} - \tilde{x}_3 \otimes u_{13}(1 + \epsilon_2) \right] \bigominus_{k=4}^{n} \tilde{x}_k \otimes u_{1k}(1 + \epsilon_2)(1 + \epsilon_3) \right] \right\}$$

$$\Big/ u_{11}(1 + \varepsilon_1')(1 + \epsilon_2')(1 + \epsilon_3').$$

We arrive at (we have included the contribution from $\otimes$):

$$\widehat{x}_1 = \left\{ b_1 - \sum_{k=2}^{n} u_{1k} \tilde{x}_k (1 + \eta_k) \prod_{j=2}^{k-1} (1 + \epsilon_j) \right\} \Big/ u_{11}(1 + \varepsilon_1') \prod_{k=2}^{n} (1 + \epsilon_k').$$

which we can rewrite to

$$u_{11}(1 + \varepsilon_1') \prod_{k=2}^{n} (1 + \epsilon_k') \tilde{x}_1 + \sum_{k=2}^{n} u_{1k} \tilde{x}_k (1 + \eta_k) \prod_{j=1}^{k-1} (1 + \epsilon_j) = b_1.$$

which we can rewrite to

$$u_{11}(1 + \varepsilon_1') \prod_{k=2}^{n} (1 + \epsilon_k') \tilde{x}_1 + \sum_{k=2}^{n} u_{1k} \tilde{x}_k (1 + \eta_k) \prod_{j=2}^{k-1} (1 + \epsilon_j) = b_1.$$

$$\begin{aligned}&(1+\varepsilon_1')(1+\epsilon_2')\cdots(1+\epsilon_n')u_{11}\tilde{x}_1\\&\quad+(1+\eta_2)u_{12}\tilde{x}_2\\&\qquad+(1+\eta_3)(1+\epsilon_2)u_{13}\tilde{x}_3\\&\qquad\quad+\cdots\\&\qquad\qquad+(1+\eta_n)(1+\epsilon_2)\cdots(1+\epsilon_{n-1})u_{1n}\tilde{x}_n=b_1\end{aligned}$$

We note that all $|\varepsilon_i|, |\epsilon_i|, |\eta_i| \leqslant \varepsilon_{\text{mac}}$ and $|\varepsilon_i'| \leqslant \varepsilon_{\text{mac}} + \mathcal{O}(\varepsilon_{\text{mac}}^2)$ and for sufficiently small ε we always have $\prod_{i=1}^m (1+\varepsilon) = 1 + m\varepsilon + \mathcal{O}(\varepsilon^2)$. In light of the expression $\tilde{U}\tilde{x} = (\delta U + U)\tilde{x} = b$, we have

$$\frac{|\delta U|}{|U|} \leqslant \underbrace{\begin{pmatrix} n & 1 & 2 & \cdots & n-2 & n-1 \\ & n-1 & 1 & \cdots & n-3 & n-2 \\ & & & \ddots & \vdots & \vdots \\ & & & & 1 & 2 \\ & & & & 2 & 1 \\ & & & & & 1 \end{pmatrix}}_{\alpha(n)} \varepsilon_{\text{mac}} + \mathcal{O}(\varepsilon_{\text{mac}}^2).$$

where by $|\cdot|$ and / we mean term-wise absolute value and division of entries. $|\tilde{U} - U| = |\delta U|$ so

$$|\tilde{U} - U| = \alpha(n) \cdot |U| \varepsilon_{\text{mac}} + \mathcal{O}(\varepsilon_{\text{mac}}^2) \cdot |U|$$

where the $\cdot$ is also taken as term-wise multiplication.

(b) We first show that $\|M\|_\infty \leqslant \sqrt{n}\|M\| \leqslant n\|M\|_\infty$ for $n \times n$ matrices. Recall that for $x \in \mathbb{R}^n$ we have $\|x\|_\infty \leqslant \|x\| \leqslant \sqrt{n}\|x\|_\infty$

$$\|M\|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^n |M_{ij}|$$

Then

$$\frac{1}{\sqrt{n}}\|\tilde{L}\tilde{U} - A\| \leqslant \|\tilde{L}\tilde{U} - A\|_\infty \leqslant \frac{3ng\varepsilon}{(1-\varepsilon)^2}\|A\|_\infty \leqslant \frac{3ng\varepsilon}{(1-\varepsilon)^2}\sqrt{n}\|A\|$$

and $\frac{\varepsilon}{(1-\varepsilon)^2} = \varepsilon(1 - 2\varepsilon + \mathcal{O}(\varepsilon^2))$ so

$$\|\tilde{L}\tilde{U} - A\| \leqslant 3n^2 g\varepsilon\|A\|.$$

(c) The exact solution x satisfies

$$LUx = b.$$

When we solve this by Gaussian elimination, we perform the following steps:

- Perform the LU decomposition in inexact arithmetics: $\tilde{L}\tilde{U} = A + E$. By (b), the size of E can be estimated as $\|E\| = O(\varepsilon)$.
- Forward elimination: Solve $\tilde{L}y = b$ inexactly, as $(\tilde{L} + \delta L)\tilde{y} = b$. By (a), the size of δL can be estimated as $\|\delta L\| = O(\varepsilon)$.
- Backward substotution: Solve $\tilde{U}z = \tilde{y}$ inexactly, as $(\tilde{U} + \delta U)\tilde{x} = \tilde{y}$. This solution $\tilde{x}$ is the final result. By (a), the size of δU can be estimated as $\|\delta U\| = O(\varepsilon)$.

If we combine the aforementioned steps, we get

$$(\tilde{L} + \delta L)(\tilde{U} + \delta U)\tilde{x} = b,$$

or

$$(\tilde{L}\tilde{U} + \delta L\tilde{U} + \tilde{L}\delta U + \delta L\delta U)\tilde{x} = LUx.$$

This can be rearranged to yield

$$LU(x - \tilde{x}) = (E + \delta L\tilde{U} + \tilde{L}\delta U + \delta L\delta U)\tilde{x}.$$

Since each of E, δL, δU is of size $O(\varepsilon)$, we conclude that $\|x - \tilde{x}\| = O(\varepsilon)$.

Problem 7

In class, we have shown that if K is a square matrix with $\|K\| < 1$, then $I - K$ is invertible, and

$$I + K + K^2 + \ldots + K^m \to (I - K)^{-1} \qquad \text{as} \quad m \to \infty.$$

We can use this fact to design an iterative method to solve $Ax = b$. The starting point should be to somehow write A in terms of $I - K$, where K has small norm. We can write $A = I - (I - A)$ and set $K = I - A$, but we would need $\|I - A\| < 1$ to ensure convergence. As a simple way to introduce some flexibility, let us multiply $Ax = b$ by some number $\omega \in \mathbb{R}\backslash\{0\}$, to get

$$\omega Ax = \omega b,$$

and then introduce $K = I - \omega A$, yielding

$$(I - K)x = \omega b \qquad \Longleftrightarrow \qquad Ax = b.$$

If $\|K\| = \|I - \omega A\| < 1$, then

$$x_m := (I + K + K^2 + \ldots + K^m)\omega b \to x.$$

The iterates x_m satisfy the recurrent relation

$$\begin{aligned} x_{m+1} &= \omega b + K(I + K + \ldots + K^m)\omega b = \omega b + Kx_m = \omega b + (I - \omega A)x_m \\ &= x_m + \omega(b - Ax_m), \end{aligned}$$

which is convenient for implementation.

(a) Assuming that $\|I - \omega A\| < 1$, derive an estimate on $\|x_m - x\|$ that goes to 0 geometrically as $m \to \infty$.
(b) Assuming that A is diagonalizable, and that all its eigenvalues are positive, estimate $\|I - \omega A\|$ in terms of λ_1, λ_n, and ω. Here λ_1 and λ_n are the smallest and the largest eigenvalues of A, respectively.
(c) In the estimate derived in (b), optimize the choice of the parameter ω.

Solution

(a) We have

$$\begin{aligned} x_m - x &= \omega b + (I - \omega A)x_{m-1} - x \\ &= \omega A x + (I - \omega A)x_{m-1} - x \\ &= (I - \omega A)x_{m-1} - (I - \omega A)x \\ &= (I - \omega A)(x_{m-1} - x). \end{aligned}$$

Then for $0 < \alpha < 1$

$$\|x_m - x\| \leqslant \|I - \omega A\| \|x_{m-1} - x\| \leqslant \alpha \|x_{m-1} - x\|, \quad (m \geqslant 1).$$

Then

$$\|x_m - x\| \leqslant \alpha \|x_{m-1} - x\| \leqslant \alpha^2 \|x_{m-2} - x\| \leqslant \cdots \leqslant \alpha^m \|x_0 - x\|.$$

(b) For invertible matrix Q with unit norm we write $A = QDQ^{-1}$ for some diagonal matrix D. Then

$$I - \omega A = QQ^{-1} - \omega Q \Lambda Q^{-1} = Q(I - \omega \Lambda)Q^{-1}.$$

If Δ is any diagonal matrix with entries Δ_i then $\|\Delta\| = \max_i |\Delta_i|$. Therefore

$$\|I - \omega A\| \leqslant \|Q\| \|I - \omega D\| \|Q^{-1}\| = \max\{|1 - \omega \lambda_1|, |1 - \omega \lambda_n|\}.$$

(c) Look at the function

$$\begin{aligned} f(\omega) &= \max\{|1 - \omega \lambda_1|, |1 - \omega \lambda_n|\} \\ &= \frac{1}{2}\Big(|1 - \omega \lambda_1| + |1 - \omega \lambda_n| + \big||1 - \omega \lambda_1| - |1 - \omega \lambda_n|\big| \Big) \end{aligned}$$

The minimum occurs when $|1 - \omega \lambda_1| = |1 + \omega \lambda_n|$, which corresponds to $\omega = \frac{2}{\lambda_1 + \lambda_2}$.