ACM106a - Homework 1 Solutions

prepared by Svitlana Vyetrenko

October 4, 2006

1. **Chapter 1, problem 1.3:**

   Since R is a non-singular $m \times m$ upper triangular matrix, its columns r_i form the basis for the space $\mathbb{C}^m$, thus, we can uniquely express any vector in $\mathbb{C}^m$ as a linear combination of r_i. In particular, the canonical vectors can be represented as:

$$e_j = \sum_{i=1}^{m} z_{ij} r_i, j = 1 \dots m \tag{1}$$

   Let Z be the matrix with entries z_{ij}. Then we have $[e_1|e_2|\dots|e_m] = I = RZ$, where Z is the inverse of R.

   Let's determine the structure of Z. For convenience, let's denote the jth component of vector r_i by $r_i^{(j)}$. Similarly, the the jth component of vector e_i is $e_i^{(j)}$. From (1) $e_j = \sum_{i=1}^{m} z_{ij} r_i = z_{1j} r_1 + z_{2j} r_2 + \dots + z_{jj} r_j + z_{j+1,j} r_{j+1} + \dots + z_{mj} r_m$.

   Let $j < m$ and suppose $z_{mj} \neq 0$. Since R is a non-singular upper triangular matrix, all its diagonal entries are different from zero, in particular, $r_m^{(m)} \neq 0$. Because the only basis vector that has a nonzero mth component is r_m and $z_{mj} \neq 0$, it implies that $e_j^{(m)} \neq 0$, which leads to a contradiction. Thus, $z_{mj} = 0$.

   Now let $j < m - 1$ and suppose $z_{m-1,j} \neq 0$. The diagonal entries of R are different from zero, therefore, $r_{m-1}^{(m-1)} \neq 0$. Since we have shown that $z_{mj} = 0$, the only basis vector that can contribute to the representation of e_j and has a nonzero $m - 1$st component is r_{m-1}. Because $z_{m-1,j} \neq 0$, it implies that $e_j^{(m-1)} \neq 0$, which again leads to a contradiction. Thus, $z_{m-1,j} = 0$.

   Proceeding as before, we can show that if $j < k$, then for $k \leq l \leq m$ all $z_{l,j} = 0$. This holds for $k = j + 1 \dots m$. Now let $k = j$. Since for $j + 1 \leq l \leq m$ all $z_{l,j} = 0$, the only basis vector that can contribute to the representation of e_j and has a nonzero jth component is r_j. Now because $e_j^{(j)}$ is equal to 1, $z_{j,j} \neq 0$.

   Therefore, we have shown that for any $1 \leq j \leq m$ $e_j = z_{1j} r_1 + z_{2j} r_2 + \dots + z_{jj} r_j$, or, in other words, $z_{l,j} = 0$ for all $j + 1 \leq l \leq m$. Hence, $Z = R^{-1}$ is upper triangular.

2. **Chapter 3, problem 3.3:**

   (a) $\|x\|_\infty = \sqrt{\max_{i=1..m} |x_i|^2} \leq \sqrt{|x_1|^2 + |x_2|^2 + \dots + |x_m|^2} = \|x\|_2$.
   Consider a vector $x = e_1$. $\|x\|_\infty = 1$, $\|x\|_2 = 1$ and the equality is achieved.

(b) $\|x\|_2 = \sqrt{\sum_{i=1}^m |x_i|^2} \leq \sqrt{m \times \max_{i=1..m} |x_i|^2} = \sqrt{m} \times \|x\|_\infty$.
Consider an m-vector $x = (1, 1, \ldots 1)$. $\|x\|_2 = \sqrt{m}$, $\|x\|_\infty = 1$ and the equality is achieved.

(c) $\|A\|_\infty = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \sqrt{n} \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{n}\|A\|_2$

Recall that $\|A\|_2 = \sqrt{\text{max eigenvalue of } A^*A} = \sqrt{\text{max eigenvalue of } AA^*}$,
$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |a_{ij}|$.

Take A with the entries of the first row equal to one and all other entries equal to zero. Then $\|A\|_\infty = n$ (maximum row-sum).
AA^* is an $m \times m$ matrix with (1,1)-entry equal to n and all other entries equal to zero. Thus, $\|A\|_2 = \sqrt(n)$ and the equality is achieved.

(d) $\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \sqrt{m} \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \sqrt{m}\|A\|_\infty$

Take A with the entries of the first column equal to one and all other entries equal to zero. Then $\|A\|_\infty = 1$ (maximum row-sum).
A^*A is an $n \times n$ matrix with (1,1)-entry equal to m and all other entries equal to zero. Thus, $\|A\|_2 = \sqrt{m}$ and the equality is achieved.

3. **Chapter 3, problem 3.4:**

(a) Note that

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1j} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2j} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{i1} & a_{i2} & \ldots & a_{ij} & \ldots & a_{in} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mj} & \ldots & a_{nn} \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 1 \\ \ldots \\ 0 \end{pmatrix} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \ldots \\ a_{ij} \\ \ldots \\ a_{mj} \end{pmatrix}$$

and

$$\begin{pmatrix} 0 \\ 0 \\ \ldots \\ 1 \\ \ldots \\ 0 \end{pmatrix}^T \times \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1j} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2j} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{i1} & a_{i2} & \ldots & a_{ij} & \ldots & a_{in} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mj} & \ldots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{i1} \\ a_{i2} \\ \ldots \\ a_{ij} \\ \ldots \\ a_{in} \end{pmatrix}^T$$

.

This leads to a conclusion that postmultiplication of A by an $n \times \nu$ matrix columns of which are formed of the canonical vectors e_k if the kth column of A needs to be kept results in an $m \times \nu$ matrix with the selected columns of A (denote the resulting matrix by $\bar{A}$). Similarly, premultiplication of $\bar{A}$ by an $\mu \times m$ matrix rows of which are formed of the canonical vectors e_k if the kth row of A needs to be kept results in an $\mu \times \nu$ matrix B with the selected rows and columns of A.

(b) Denote the postmultiplication matrix from part (a) by C and the premultiplication matrix by R, i.e. $B = R \times A \times C$. Then for any p with $1 \leq p \leq \infty$ we have:
$\|B\|_p = \|R \times A \times C\|_p \leq \|R\|_p \times \|A\|_p \times \|C\|_p$ (the last inequality follows by submultiplicativity of any induced norm).

$\|R\|_p = \sup_{x\neq 0} \frac{\|Rx\|_p}{\|x\|_p} \leq 1$ (because of the structure of R).
Similarly, $\|C\|_p = \sup_{x\neq 0} \frac{\|Cx\|_p}{\|x\|_p} \leq 1$. Therefore, $\|B\|_p \leq \|R\|_p \times \|A\|_p \times \|C\|_p \leq \|A\|_p$.

4. Using the notation of Demmel's book, in IEEE double precision arithmetic $\epsilon \approx 10^{-16}$. Recall that if $\odot$ is one of the four binary operations $+, -, \times, \div$, then $fl(a \odot b) = (a \odot b) \times (1 + \delta)$, where $|\delta| \leq \epsilon$.

   First algorithm:
   $fl\left(\frac{\log(1+x)}{x}\right) = \frac{\log(fl(1+x))}{x} \times (1+\delta_2) = \frac{\log((1+x)(1+\delta_1)))}{x} \times (1+\delta_2) = (\frac{\log(1+x)}{x} + \frac{\log(1+\delta_1)}{x}) \times (1+\delta_2) = \frac{\log(1+x)}{x} + \frac{\log(1+\delta_1)}{x} + \delta_2\frac{\log(1+x)}{x} + \delta_2\frac{\log(1+\delta_1)}{x} \sim \frac{\log(1+x)}{x} + \frac{\delta_1}{x} + \delta_2\frac{\log(1+x)}{x} + \frac{\delta_1\delta_2}{x}$
   First consider the case when $x = 0$, then $\frac{\log((1+x)(1+\delta_1)))}{x} \times (1+\delta_2) \sim \frac{0}{0}$, which results in NaN.
   When $x \neq 0$ the error is given by:
   $fl\left(\frac{\log(1+x)}{x}\right) - \frac{\log(1+x)}{x} = \frac{\delta_1}{x} + \delta_2\frac{\log(1+x)}{x} + \frac{\delta_1\delta_2}{x} \sim \frac{\delta_1}{x} + \delta_2 + \frac{\delta_1\delta_2}{x}$
   Thus, since δ_1 and x can be of the same order of magnitude, the first term in the expression for the error is not negligible, and is, in fact, bounded by 1. Therefore, the first algorithm is unstable near $x = 0$.

   Second algorithm:
   $d = fl(1+x) = (1+x)(1+\delta_1)$
   When $d \neq 1$
   $fl(\frac{\log d}{d-1}) = \frac{\log d}{d-1}(1+\delta_2) = \frac{\log(1+x)(1+\delta_1)}{(1+x)(1+\delta_1)-1}(1+\delta_2) = \frac{\log(1+\delta_1+x+x\delta_1)}{(\delta_1+x+x\delta_1)}(1+\delta_2) \sim \frac{\delta_1+x+x\delta_1}{\delta_1+x+x\delta_1}(1+\delta_2) = 1+\delta_2$.
   Therefore, the error is given by:
   $\frac{\log(1+x)(1+\delta_1)}{(1+x)(1+\delta_1)-1}(1+\delta_2) - \frac{\log(1+x)}{x} \sim 1+\delta_2-1 = \delta_2$ and the algorithm produces accurate answer in floating point arithmetic. Note that the case when $d = 1$ needs a special treatment since for $d = 1$ $\frac{\log d}{d-1} \sim \frac{0}{0}$.