STAT 309: MATHEMATICAL COMPUTATIONS I
FALL 2013
PROBLEM SET 1

The parts marked "*Bonus*" are optional, i.e., Problems **1**(c), **4**(d), and **4**(e). For Problems **3** and **6**, use any program you like but present your source codes and results in a way that is comprehensible to someone who is unfamiliar with that program (e.g. comment your codes appropriately). Scilab and Octave use Matlab syntax but are open source and freely downloadable.

1. Let $\mathbf{x} \in \mathbb{C}^n$ and $A \in \mathbb{C}^{m\times n}$. We write $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^*\mathbf{x}}$ and $\|A\|_2 = \sup_{\|\mathbf{x}\|_2=1}\|A\mathbf{x}\|_2$ for the vector 2-norm and matrix 2-norm respectively.

(a) Show that there is no ambiguity in the notation, i.e., if $A \in \mathbb{C}^{n\times 1} = \mathbb{C}^n$, then $\|A\|_2$ is the same whether we regard it as the vector or matrix 2-norm. What if $A \in \mathbb{C}^{1\times n}$?

(b) Show that the vector 2-norm is unitarily invariant, i.e.,

$$\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$$

for all unitary matrices $U \in \mathbb{C}^{n\times n}$.

(c) *Bonus*: Show that no other vector p-norm is unitarily invariant, $1 \le p \le \infty$, $p \ne 2$.

(d) Show that the matrix 2-norm is unitarily invariant, i.e.,

$$\|UAV\|_2 = \|A\|_2$$

for all unitary matrices $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{n\times n}$.

(e) Show that the Frobenius norm is unitarily invariant, i.e.,

$$\|UAV\|_F = \|A\|_F$$

for all unitary matrices $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{n\times n}$. (*Hint*: First show that $\|A\|_F^2 = \operatorname{tr}(A^*A) = \operatorname{tr}(AA^*)$).

(f) Let $U \in \mathbb{C}^{n\times n}$. Show that the following are equivalent statements:

(i) $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{C}^n$;

(ii) $(U\mathbf{x})^*U\mathbf{y} = \mathbf{x}^*\mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$;

(iii) U is unitary.

2. Let $A \in \mathbb{C}^{n\times n}$. Let $\|\cdot\|$ be an operator norm of the form

$$\|A\| = \max_{\mathbf{0}\ne\mathbf{v}\in\mathbb{C}^n} \frac{\|A\mathbf{v}\|_\alpha}{\|\mathbf{v}\|_\alpha} \tag{2.1}$$

for some vector norm $\|\cdot\|_\alpha : \mathbb{C}^n \to [0,\infty)$. Show that if $\|A\| < 1$, then $I - A$ is nonsingular and furthermore,

$$\frac{1}{1+\|A\|} \le \|(I-A)^{-1}\| \le \frac{1}{1-\|A\|}.$$

3. We will examine the effect of various parameters on the accuracy of a computed solution to a nonsingular linear system. Relevant commands in Matlab syntax are given in brackets.

(a) Generate $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ as follows:

(i) a_{ij} randomly generated from a standard normal distribution [`randn(n)`];

(ii) a Hilbert matrix, i.e., $a_{ij} = 1/(i+j-1)$ [`hilb(n)`];

(iii) a Pascal matrix, i.e., the entries $a_{ij} = \binom{i+j}{i}$ [`pascal(n)`];

Date: October 10, 2013 (Version 1.4); due: October 22, 2013.

(iv) a magic square, i.e., the entries a_{ij}'s are the integers $1, 2, \ldots, n^2$ arranged in a way that A has equal row, column, and diagonal sums [`magic(n)`].

$$\texttt{hilb(4)} = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{bmatrix}, \quad \texttt{pascal(4)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{bmatrix}, \quad \texttt{magic(4)} = \begin{bmatrix} 16 & 2 & 3 & 13 \\ 5 & 11 & 10 & 8 \\ 9 & 7 & 6 & 12 \\ 4 & 14 & 15 & 1 \end{bmatrix}$$

For simplicity, we will assume that A is stored exactly with no errors even though this is only true for those matrices with integer-valued entries.

(b) Generate $\mathbf{x}$ and $\mathbf{b} \in \mathbb{R}^n$ as follows:
  (i) $\mathbf{x} = [1, \ldots, 1]^{\mathsf{T}}$ [`ones(n,1)`];
  (ii) $\mathbf{b} = A\mathbf{x}$ [`b = A*x`].

(c) For each A generated as above, perform the following for $n = 5, 10, 15, \ldots, 500$.
  (i) Solve $A\mathbf{x} = \mathbf{b}$ using your program to get $\widehat{\mathbf{x}}$ [`xhat = A\b`]. Note that in general the result computed by your program will not be exactly the true solution $\mathbf{x} = A^{-1}\mathbf{b}$ because of roundoff errors that occurred during computations.
  (ii) Compute $\delta\mathbf{b} = A\widehat{\mathbf{x}} - \mathbf{b}$ and record the values of $\|\mathbf{x} - \widehat{\mathbf{x}}\|/\|\mathbf{x}\|$, $\kappa(A) = \|A\|\|A^{-1}\|$ and $\kappa(A)\|\delta\mathbf{b}\|/\|\mathbf{b}\|$for $\|\cdot\| = \|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$.
  (iii) Present everything for the $n = 5$ case but only tabulate the relevant trend for general $n > 5$ in a graph.

(d) Discuss and explain the effects of different choices of A, $\mathbf{b}$, $\|\cdot\|$, and n have on the accuracy of the computed solution $\widehat{\mathbf{x}}$.

(e) Instead of solving the linear system directly, compute A^{-1} and then $\widehat{\mathbf{x}} := A^{-1}\mathbf{b}$ [`xhat = inv(A)*b`]. Comment on the accuracy of this approach. Provide numerical evidence to support your conclusion.

(f) Write a program that computes the $(1, 1)$-entry of the matrix A^{-1} that does not involve computing A^{-1}, i.e., if $A^{-1} = [b_{ij}]$, you want the value b_{11} but you are not allowed to compute A^{-1}.

4. Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and let $\mathbf{0} \neq \mathbf{b} \in \mathbb{R}^n$. Let $\mathbf{x} = A^{-1}\mathbf{b} \in \mathbb{R}^n$. In the following, $\delta A \in \mathbb{R}^{n\times n}$ and $\delta\mathbf{b} \in \mathbb{R}^n$ are some arbitrary matrix and vector. We assume that the norm on A satisfies $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$ for all $A \in \mathbb{R}^{n\times n}$ and all $\mathbf{x} \in \mathbb{R}^n$.

(a) Show that if $\delta A \in \mathbb{R}^{n\times n}$ is any matrix satisfying

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)}, \tag{4.2}$$

then $A + \delta A$ must be nonsingular. (*Hint*: If $A + \delta A$ is singular, then there exists nonzero $\mathbf{v}$ such that $(A + \delta A)\mathbf{v} = \mathbf{0}$).

(b) Suppose $(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$. Show that

$$\frac{\|\delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A)\frac{\|\delta A\|}{\|A\|}. \tag{4.3}$$

(c) Suppose $(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ and (4.2) is satisfied. Show that

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)\dfrac{\|\delta A\|}{\|A\|}}{1 - \kappa(A)\dfrac{\|\delta A\|}{\|A\|}}.$$

You may like use the following outline:
  (i) Show that

$$\delta\mathbf{x} = -A^{-1}\delta A\hat{\mathbf{x}}$$

and so

$$\|\delta\mathbf{x}\| \le \kappa(A)\frac{\|\delta A\|}{\|A\|}(\|\mathbf{x}\| + \|\delta\mathbf{x}\|).$$

(ii) Rewrite this inequality as

$$\left(1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}\right)\|\delta\mathbf{x}\| \le \kappa(A)\frac{\|\delta A\|}{\|A\|}\|\mathbf{x}\|$$

and use (4.2).

(d) *Bonus*: Suppose $(A+\delta A)\hat{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b} \ne \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x} \ne \mathbf{0}$. Show that

$$\frac{\|\delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \le \kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|} + \frac{\|\delta A\|}{\|A\|}\frac{\|\delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|}\right). \tag{4.4}$$

You may like use the following outline:

(i) Show that

$$\delta\mathbf{x} = A^{-1}(\delta\mathbf{b} - \delta A\hat{\mathbf{x}})$$

and so

$$\frac{\|\delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \le \kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|A\|\|\hat{\mathbf{x}}\|}\right). \tag{4.5}$$

(ii) Show that

$$\frac{1}{\|\hat{\mathbf{x}}\|} \le \frac{\|A\| + \|\delta A\|}{\|\hat{\mathbf{b}}\|}. \tag{4.6}$$

(iii) Combine (4.5) and (2.1) to get (4.4).

(e) *Bonus*: Suppose $(A+\delta A)\hat{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b} \ne \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x} \ne \mathbf{0}$ and (4.2) is satisfied. Use the same ideas in (b) to deduce that

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{\kappa(A)\left(\dfrac{\|\delta A\|}{\|A\|} + \dfrac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right)}{1 - \kappa(A)\dfrac{\|\delta A\|}{\|A\|}}.$$

5. Recall that in the lectures, we mentioned that (i) there are matrix norms that are not submultiplicative and an example is the Hölder ∞-norm; (ii) we may always construct a norm that approximates the spectral radius of a given matrix A as closely as we want.

(a) Show that if $\|\cdot\| : \mathbb{C}^{m\times n} \to \mathbb{R}$ is a norm, then there always exists a $c > 0$ such that the constant multiple $\|\cdot\|_c := c\|\cdot\|$ defines a submultiplicative norm, i.e.,

$$\|AB\|_c \le \|A\|_c\|B\|_c$$

for any $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times p}$ (even if $\|\cdot\|$ does not have this property). Find the constant c for the Hölder ∞-norm.

(b) Let $J \in \mathbb{C}^{n\times n}$ be in Jordan form, i.e.,

$$J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{bmatrix}$$

where each block J_r, for $r = 1, \dots, k$, has the form

$$J_r = \begin{bmatrix} \lambda_r & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_r \end{bmatrix}.$$

Let $\varepsilon > 0$ and $D_\varepsilon = \operatorname{diag}(1, \varepsilon, \varepsilon^2, \dots, \varepsilon^{n-1})$. Verify that

$$D_\varepsilon^{-1} J D_\varepsilon = \begin{bmatrix} J_{1,\varepsilon} & & \\ & \ddots & \\ & & J_{k,\varepsilon} \end{bmatrix}$$

where $J_{r,\varepsilon}$ is the matrix you obtain by replacing the 1's on the superdiagonal of J_r by ε's,

$$J_{r,\varepsilon} = \begin{bmatrix} \lambda_r & \varepsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_r \end{bmatrix}$$

(c) Show that
$$\|D_\varepsilon^{-1} J D_\varepsilon\|_\infty \le \rho(J) + \varepsilon.$$

(d) Hence, or otherwise, show that for any given $A \in \mathbb{C}^{n\times n}$ and $\varepsilon > 0$, there exists an operator norm $\|\cdot\|$ of the form (2.1) with the property that
$$\rho(A) \le \|A\| \le \rho(A) + \varepsilon.$$
(*Hint*: Transform A into Jordan form).

6. Let $A = [a_{ij}]$ be an $n \times n$ matrix with entries

$$a_{ij} = \begin{cases} n + 1 - \max(i, j) & i \le j + 1, \\ 0 & i > j + 1. \end{cases}$$

This is an example of an *upper Hessenberg* matrix: it is upper triangular except that the entries on the subdiagonal $a_{i,i+1}$ may also be non-zero. For $n = 12$ and $n = 25$, do the following[1]:

(a) Compute $\|A\|_\infty$ and $\|A\|_1$.
(b) Compute $\rho(A)$ and $\|A\|_2$.
(c) Using Gerschgorin's theorem, describe the domain that contains all of the eigenvalues.
(d) Compute all of the eigenvalues and singular values of A. How many of the eigenvalues are real and how many are complex?

[1]You may use any built-in functions of your program.