STAT 309: MATHEMATICAL COMPUTATIONS I
FALL 2012
PROBLEM SET 5

1. Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and let $\mathbf{b} \in \mathbb{R}^n$. We shall use $\|\cdot\|$ to denote both the vector and the matrix norm and require that $\|M\mathbf{v}\| \le \|M\|\|\mathbf{v}\|$ for any $M \in \mathbb{R}^{n\times n}$ and $\mathbf{v} \in \mathbb{R}^n$.

(a) Show that given any $\hat{\mathbf{x}} \in \mathbb{R}^n$, we have

$$\frac{1}{\kappa(A)}\frac{\|A\hat{\mathbf{x}} - \mathbf{b}\|}{\|\mathbf{b}\|} \le \frac{\|\hat{\mathbf{x}} - A^{-1}\mathbf{b}\|}{\|A^{-1}\mathbf{b}\|} \le \kappa(A)\frac{\|A\hat{\mathbf{x}} - \mathbf{b}\|}{\|\mathbf{b}\|},$$

where $\kappa(A) = \|A\|\|A^{-1}\|$. Deduce that if $\mathbf{x} = A^{-1}\mathbf{b}$ and $\hat{\mathbf{b}} = A\hat{\mathbf{x}} - \mathbf{b}$, then

$$\frac{1}{\kappa(A)}\frac{\|\hat{\mathbf{b}}\|}{\|\mathbf{b}\|} \le \frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \le \kappa(A)\frac{\|\hat{\mathbf{b}}\|}{\|\mathbf{b}\|}.$$

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

$$\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)}, \tag{1.1}$$

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

2. Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and let $\mathbf{0} \ne \mathbf{b} \in \mathbb{R}^n$. 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.

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

$$\frac{\|\delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \le \kappa(A)\frac{\|\delta A\|}{\|A\|}. \tag{2.2}$$

(b) Suppose $(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ and (1.1) is satisfied. Show that

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

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 (1.1).

Date: December 1, 2012 (Version 1.2); due: December 14, 2012.

(c) Suppose $(A+\delta A)\hat{\mathbf{x}} = \mathbf{b}+\delta\mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b}+\delta\mathbf{b} \neq \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x}+\delta\mathbf{x} \neq \mathbf{0}$. Show that

$$\frac{\|\delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|} + \frac{\|\delta A\|}{\|A\|}\frac{\|\delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|}\right). \tag{2.3}$$

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

$$\frac{\|\delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leq \kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|A\|\|\hat{\mathbf{x}}\|}\right). \tag{2.4}$$

(ii) Show that

$$\frac{1}{\|\hat{\mathbf{x}}\|} \leq \frac{\|A\| + \|\delta A\|}{\|\hat{\mathbf{b}}\|}. \tag{2.5}$$

(iii) Combine (2.4) and (2.5) to get (2.3).

(d) Suppose $(A+\delta A)\hat{\mathbf{x}} = \mathbf{b}+\delta\mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b}+\delta\mathbf{b} \neq \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x}+\delta\mathbf{x} \neq \mathbf{0}$ and (1.1) is satisfied. Use the same ideas in (b) to deduce that

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

3. Let $A \in \mathbb{R}^{m\times n}$ where $m \geq n$ and $\operatorname{rank}(A) = n$. Suppose GECP is performed on A to get

$$\Pi_1 A \Pi_2 = LU$$

where $L \in \mathbb{R}^{m\times n}$ is unit lower triangular, $U \in \mathbb{R}^{n\times n}$ is upper triangular, and $\Pi_1 \in \mathbb{R}^{m\times m}$, $\Pi_2 \in \mathbb{R}^{n\times n}$ are permutation matrices.

(a) Show that U is nonsingular and that L is of the form

$$L = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$$

where $L_1 \in \mathbb{R}^{n\times n}$ is nonsingular.

(b) We will see how the LU factorization may be used to solve the least squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2.$$

(i) Show that the problem may be solved via

$$U\widetilde{\mathbf{x}} = \mathbf{y}, \quad L^\top L\mathbf{y} = L^\top\widetilde{\mathbf{b}},$$

where $\widetilde{\mathbf{b}} = \Pi_1\mathbf{b}$ and $\widetilde{\mathbf{x}} = \Pi_2^\top\mathbf{x}$.

(ii) Describe how you would compute the solution $\mathbf{y}$ in

$$L^\top L\mathbf{y} = L^\top\widetilde{\mathbf{b}}.$$

4. Let $\varepsilon > 0$. Consider the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon \\ 1 & 1-\varepsilon \end{bmatrix}.$$

(a) Why is it a bad idea to solve the normal equation associated with A, i.e.

$$A^\top A\mathbf{x} = A^\top\mathbf{b}$$

when ε is small?

(b) Show that the LU factorization of A is

$$A = LU = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & \varepsilon \end{bmatrix}.$$

(c) Why is it a much better idea to solve the normal equation associated with L, i.e.

$$L^\top L\mathbf{y} = L^\top \widetilde{\mathbf{b}}?$$

This shows that the method in Problem **3** is a more stable method than using the normal equation in (a) directly.

(d) Show that the Moore-Penrose pseudoinverse of A is

$$A^\dagger = \frac{1}{6}\begin{bmatrix} 2 & 2-3\varepsilon^{-1} & 2+3\varepsilon^{-1} \\ 0 & 3\varepsilon^{-1} & -3\varepsilon^{-1} \end{bmatrix}.$$

(e) Describe a method to compute A^+ given L and U. Verify that your method is correct by checking it against the expression in (d).

5. We will now discuss an alternative method to solve the least squares problem in Problem **3** that is more efficient when $m - n < n$.

(a) Show the least squares problem in Problem **3** is equivalent to

$$\min_{\mathbf{z}\in\mathbb{R}^n} \left\| \begin{bmatrix} I_n \\ S \end{bmatrix} \mathbf{z} - \widetilde{\mathbf{b}} \right\|_2$$

where $S = L_2 L_1^{-1}$ and $L_1\mathbf{y} = \mathbf{z}$. Here and below, I_n denotes the $n \times n$ identity matrix.

(b) Write

$$\widetilde{\mathbf{b}} = \begin{bmatrix} \widetilde{\mathbf{b}}_1 \\ \widetilde{\mathbf{b}}_2 \end{bmatrix}$$

where $\widetilde{\mathbf{b}}_1 \in \mathbb{R}^n$ and $\widetilde{\mathbf{b}}_2 \in \mathbb{R}^{m-n}$. Show that the solution $\mathbf{z}$ is given by

$$\mathbf{z} = \widetilde{\mathbf{b}}_1 + S^\top (I_{m-n} + SS^\top)^{-1}(\widetilde{\mathbf{b}}_2 - S\widetilde{\mathbf{b}}_1).$$

(c) Explain why when $m - n < n$, the method in (a) is much more efficient than the method in Problem **3**. For example, what happens when $m = n + 1$?

6. Let $\mathbf{c} \in \mathbb{R}^n$ and consider the linearly constrained least squares problem

$$\min \|\mathbf{w}\|_2 \quad \text{s.t. } A^\top \mathbf{w} = \mathbf{c}.$$

(a) If we write $\widetilde{\mathbf{c}} = \Pi_2^\top \mathbf{c}$ and $\widetilde{\mathbf{w}} = \Pi_1 \mathbf{w}$, show that

$$\widetilde{\mathbf{w}} = L(L^\top L)^{-1} U^{-\top} \widetilde{\mathbf{c}}$$

where $U^{-\top} = (U^{-1})^\top = (U^\top)^{-1}$, a standard notation that we will also use below. (*Hint*: You'd need to use something that you've already determined in an earlier part).

(b) Write

$$\widetilde{\mathbf{w}} = \begin{bmatrix} \widetilde{\mathbf{w}}_1 \\ \widetilde{\mathbf{w}}_2 \end{bmatrix}$$

where $\widetilde{\mathbf{w}}_1 \in \mathbb{R}^n$ and $\widetilde{\mathbf{w}}_2 \in \mathbb{R}^{m-n}$. Show that

$$\widetilde{\mathbf{w}}_1 = L_1^{-\top} U^{-\top} \widetilde{\mathbf{c}} - S^\top \widetilde{\mathbf{w}}_2.$$

(c) Write $\mathbf{d} = L_1^{-\top} U^{-\top} \widetilde{\mathbf{c}}$. Deduce that $\widetilde{\mathbf{w}}_2$ may be obtained either as a solution to

$$\min_{\widetilde{\mathbf{w}}_2 \in \mathbb{R}^{m-n}} \left\| \begin{bmatrix} S^\top \\ I_{m-n} \end{bmatrix} \widetilde{\mathbf{w}}_2 - \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \right\|_2$$

or as

$$\widetilde{\mathbf{w}}_2 = (I_{m-n} + SS^\top)^{-1} S\mathbf{d}.$$

Note that when $m - n < n$, this method is advantageous for the same reason in Problem **5**.

7. So far we have assumed that A has full column rank. Suppose now that $\operatorname{rank}(A) = r < \min\{m, n\}$.

(a) Show that the LU factorization obtained using GECP is of the form

$$\Pi_1 A \Pi_2 = LU = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}$$

where $L_1, U_1 \in \mathbb{R}^{r \times r}$ are triangular and nonsingular.

(b) Show that the above equation may be rewritten in the form

$$\Pi_1 A \Pi_2 = \begin{bmatrix} I_r \\ S_1 \end{bmatrix} L_1 U_1 \begin{bmatrix} I_r & S_2^\top \end{bmatrix}$$

for some lower triangular matrices S_1 and S_2.

(c) Hence show that the Moore-Penrose inverse of A is given by

$$A^\dagger = \Pi_2 \begin{bmatrix} I_r & S_2^\top \end{bmatrix}^\dagger U_1^{-1} L_1^{-1} \begin{bmatrix} I_r \\ S_1 \end{bmatrix}^\dagger \Pi_1.$$

(d) Using the general formula (derived in the lectures) for the Moore-Penrose inverse of a rank-retaining factorization, what do you get for $A^\dagger$?