CS205 Homework #1 Solutions

Problem 1

Arithmetic operations are subject to roundoff error when performed on a finite precision computer. In order to perform an operation x *op* y on the *real* numbers x and y we deviate from the analytic result when discretizing those values to machine precision as well as when we store the resulting value.

Let $\bar{x}$ denote the discretized, floating point version of x that is stored on the computer. You may assume that

$$\bar{x} = (1+\epsilon)x$$

where ϵ is bounded as $0 \le |\epsilon| < \epsilon_{\max}$ where $\epsilon_{\max} \ll 1$ is the machine roundoff precision.

Assume that the result of the arithmetic operation between two floating point numbers $\bar{x}$ and $\bar{y}$ is computed exactly, but when stored on the computer it is once again subject to roundoff error as

$$\overline{\bar{x} \; op \; \bar{y}} = (1+\epsilon')(\bar{x} \; op \; \bar{y})$$

where the roundoff error obeys the same bounds $0 \le |\epsilon'| < \epsilon_{\max}$.

The relative error of a computation is defined as

$$E = \left| \frac{\text{Computed\_Result} - \text{Analytic\_Result}}{\text{Analytic\_Result}} \right|$$

Provide a bound (in terms of $\epsilon_{\max}$) for the relative error induced by the following arithmetic operations, or prove that the relative error is unbounded.

1. Subtraction, Multiplication and Division of two real numbers (for an example on addition see Heath, section 1.3.8)

2. Computing the sum $s_n = \underbrace{x + x + \cdots + x}_{n \text{ terms}}$ using the recurrence

   $$s_1 = x$$
   $$s_k = s_{k-1} + x$$

   [Answer: $\approx n\epsilon_{\max}/2$]

3. Computing the sum $s_n = s_{2^k} = q_k = \underbrace{x + x + \cdots + x}_{n=2^k \text{ terms}}$ where $n = 2^k$ using the recurrence

   $$q_0 = x$$
   $$q_k = q_{k-1} + q_{k-1}$$

For (2) and (3) you may assume for simplicity that $n \ll 1/\epsilon_{\max}$.

Solution

For the following derivation we use the lemma: If $0 \leq |\epsilon_1|, |\epsilon_2|, \ldots, |\epsilon_k| < \epsilon_{\max}$ then there exists an $\epsilon \in [0, \epsilon_{\max})$ such that $(1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_k) = (1+\epsilon)^k$, which holds by virtue of the intermediate value theorem.

For every variable ϵ_i used in the following derivations we will implicity assume it lies within the range $0 \leq |\epsilon_i| < \epsilon_{\max}$.

1. We have $\bar{x} = (1+\epsilon_1)x$ and $\bar{y} = (1+\epsilon_2)y$.

Subtraction We will show that there is no bound on the relative error. Consider $\bar{x} - \bar{y} = (1+\epsilon_1)x - (1+\epsilon_2)y$. Let $x = a + \theta$ and $y = a$ so $x - y = \theta$. Then

$$\bar{x} - \bar{y} = (a+\theta)(1+\epsilon_1) - a(1+\epsilon_2) = \theta + a(\epsilon_1 - \epsilon_2) + \theta\epsilon_1$$

$$\overline{\bar{x} - \bar{y}} = \theta(1+\epsilon_3) + a(\epsilon_1 - \epsilon_2)(1+\epsilon_3) + \theta\epsilon_1(1+\epsilon_3)$$

Then the relative error is given by

$$\begin{aligned} E &= \left|\frac{\overline{\bar{x} - \bar{y}} - (x-y)}{x-y}\right| \\ &= \left|\frac{\theta(1+\epsilon_3) + a(\epsilon_1 - \epsilon_2)(1+\epsilon_3) + \theta\epsilon_1(1+\epsilon_3) - \theta}{\theta}\right| \\ &= \left|\epsilon_3 + \epsilon_1(1+\epsilon_3) + \frac{a}{\theta}(\epsilon_1 - \epsilon_2)(1+\epsilon_3)\right| \end{aligned}$$

which becomes unbounded as $\theta \to 0$.

Multiplication

$$\begin{aligned} E_\times &= \left|\frac{\overline{\bar{x} \cdot \bar{y}} - xy}{xy}\right| = \left|\frac{xy(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) - xy}{xy}\right| \\ &= \left|(1+\epsilon_4)^3 - 1\right| = \left|3\epsilon_4 + O(\epsilon_{\max}^2)\right| \end{aligned}$$

Division

$$\begin{aligned} E_\div &= \left|\frac{\overline{\bar{x}/\bar{y}} - x/y}{x/y}\right| = \left|\frac{(x/y)\frac{(1+\epsilon_1)(1+\epsilon_3)}{1+\epsilon_2} - x/y}{x/y}\right| = \left|\frac{(1+\epsilon_1)(1+\epsilon_3)}{1+\epsilon_2} - 1\right| \\ &= \left|(1+\epsilon_1)(1+\epsilon_3)[(1-\epsilon_2) + O(\epsilon_{\max}^2)] - 1\right| \\ &= \left|(1+\epsilon_4)^3 - 1 + O(\epsilon_{\max}^2)\right| = \left|3\epsilon_4 + O(\epsilon_{\max}^2)\right| \leq \left|3\epsilon_{\max} + O(\epsilon_{\max}^2)\right| \end{aligned}$$

2. By straightforward manipulation, we have:

$$\begin{aligned} \bar{s}_k &= \left[\bar{s}_{k-1} + (1+\epsilon_1)\,x\right](1+\epsilon_k) \\ &= \left[\left[\bar{s}_{k-2} + (1+\epsilon_1)\,x\right](1+\epsilon_{k-1}) + (1+\epsilon_1)\,x\right](1+\epsilon_k) \\ &= (1+\epsilon_k)(1+\epsilon_1)\,x + (1+\epsilon_k)(1+\epsilon_{k-1})(1+\epsilon_1)\,x + \ldots + \\ &\quad (1+\epsilon_k)(1+\epsilon_{k-1})\ldots(1+\epsilon_2)(1+\epsilon_1)\,x + (1+\epsilon_k)(1+\epsilon_{k-1})\ldots(1+\epsilon_2)(1+\epsilon_1)\,x \\ &= x\left[(1+\epsilon_*)^n + \sum_{k=2}^{n}\left(1+\epsilon_{*_k}\right)^k\right] \end{aligned}$$

Now, we can apply the first-order binomial approximation $(1+\epsilon)^k = 1 + k\epsilon + O\left(\epsilon^2\right)$:

$$\begin{aligned} x\left[(1+\epsilon_*)^n + x\sum_{k=2}^{n}\left(1+\epsilon_{*_k}\right)^k\right] &\leq x\left[1 + n\epsilon_{max} + O\left(\epsilon_{max}^2\right) + \sum_{k=1}^{n-1}\left(1 + (k+1)\,\epsilon_{max} + O\left(\epsilon_{max}^2\right)\right)\right] \\ &= x\left[n\epsilon_{max} + O\left(\epsilon_{max}^2\right) + n + \frac{(n-1)(n+2)}{2}\epsilon_{max} + (n-1)O\left(\epsilon_{max}^2\right)\right] \\ &= x\left[n\epsilon_{max} + n + \left(\frac{n^2+n}{2} + 1\right)\epsilon_{max} + nO\left(\epsilon_{max}^2\right)\right] \\ &= x\left[\left(\frac{n^2+3n}{2} - 1\right)\epsilon_{max} + n + nO\left(\epsilon_{max}^2\right)\right] \end{aligned}$$

Now, we can compute the relative error as follows:

$$\begin{aligned} E_{nx} &= \left|\frac{x\left[\left(\frac{n^2+3n}{2} - 1\right)\epsilon_{max} + n + nO\left(\epsilon_{max}^2\right)\right] - nx}{nx}\right| \\ &= \left|\left(\frac{n+3}{2} - \frac{1}{n}\right)\epsilon_{max} + O\left(\epsilon_{max}^2\right)\right| \end{aligned}$$

However, since $n \ll 1/\epsilon_{\mathrm{max}}$, we have:

$$\left|\left(\frac{n+3}{2} - \frac{1}{n}\right)\epsilon_{max} + O\left(\epsilon_{max}^2\right)\right| = \left|\left(\frac{n+3}{2}\right)\epsilon_{max} - o\left(\epsilon_{max}\right) + O\left(\epsilon_{max}^2\right)\right|$$

3. To simplify the computation, let ξ_k be the cumulative relative error in q_k, and $\bar{s}_k = q_k(1+\xi_k)$.

$$\begin{aligned} \xi_k &= \left|\frac{\bar{q}_k - q_k}{q_k}\right| = \left|\frac{\overline{\bar{q}_{k-1} + \bar{q}_{k-1}} - q_k}{q_k}\right| \\ &= \left|\frac{\left[(1+\xi_{k-1})2^{k-1}x + (1+\xi_{k-1})2^{k-1}x\right](1+\epsilon_1) - 2^k x}{2^k x}\right| \\ &= \left|(1+\xi_{k-1})(1+\epsilon_1) - 1\right| \\ &= \left|\xi_{k-1} + \epsilon_1 + \xi_{k-1}\epsilon_1\right| \\ &\leq \left|\xi_{k-1} + \epsilon_{\mathrm{max}} + o(\epsilon_{\mathrm{max}})\right| \end{aligned}$$

This last step follows since each application of floating point addition increases the cumulative relative error in its operands by $\epsilon_{\max}$ at most, therefore $\xi_k \leq k\epsilon_{\max}$. Since we assume $k \ll 1/\epsilon_{\max}$ we have $\xi_{k-1}\epsilon_1 = o(\epsilon_{\max})$. Therefore,

$$\xi_k \leq |k\epsilon_{\max} + o(\epsilon_{\max})|$$

If θ_k is the cumulative relative error in the computation of s_k we have $\theta_{2^k} = \xi_k$, therefore

$$\theta_k \leq |\log_2 k \cdot \epsilon_{\max} + o(\epsilon_{\max})|$$

Problem 2

Consider the elimination matrix $\mathbf{M_k} = \mathbf{I} - \mathbf{m_k e_k^T}$ and its inverse $\mathbf{L_k} = \mathbf{I} + \mathbf{m_k e_k^T}$ used in the LU decomposition process, where

$$\mathbf{m_k} = \left(\mathbf{0}, \ldots, \mathbf{0}, \mathbf{m_{k+1}^{(k)}}, \ldots, \mathbf{m_n^{(k)}}\right)$$

and $\mathbf{e_k}$ is the k-th column of the identity matrix. Let $\mathbf{P^{(ij)}}$ be the permutation matrix that results from swapping the i-th and j-th rows (or columns) of the identity matrix.

1. Show that if $i, j > k$ then $\mathbf{L_k P^{(ij)}} = \mathbf{P^{(ij)}}(\mathbf{I} + \mathbf{P^{(ij)} m_k e_k^T})$

2. Recall that the matrix $\mathbf{L}$ resulting from performing Gaussian elimination with partial pivoting is given by

   $$\mathbf{L} = \mathbf{P_1 L_1} \cdots \mathbf{P_{n-1} L_{n-1}}$$

   where the permutation matrix $\mathbf{P_i}$ permutes row i with some row i' where $i < i'$. Show that $\mathbf{L}$ can be rewritten as

   $$\mathbf{L} = \mathbf{P_1} \cdots \mathbf{P_{n-1} L_1^P} \cdots \mathbf{L_{n-1}^P}$$

   where $\mathbf{L_k^P} = \mathbf{I} + (\mathbf{P_{n-1}} \cdots \mathbf{P_{k+1} m_k})\, \mathbf{e_k^T}$.

3. Show that $\mathbf{L_1^P} \cdots \mathbf{L_{n-1}^P}$ is lower triangular.

Solution

1. The matrix $\mathbf{m_k e_k^T}$ has nonzero elements only on the k-th column, in the positions corresponding to rows $(k+1)$ through n. Additionally, $\mathbf{m_k e_k^T P^{(ij)}}$ is the result of swapping the i-th and j-th column of $\mathbf{m_k e_k^T}$, which are both zero. Thus $\mathbf{m_k e_k^T P^{(ij)}} = \mathbf{m_k e_k^T}$. Using this result, we have

$$\begin{aligned}
(\mathbf{I} + \mathbf{m_k e_k^T})\mathbf{P^{(ij)}} &= \mathbf{P^{(ij)}} + \mathbf{m_k e_k^T P^{(ij)}} \\
&= \mathbf{P^{(ij)}} + \mathbf{m_k e_k^T} \\
&= \mathbf{P^{(ij)}} + (\mathbf{P^{(ij)}})^2 \mathbf{m_k e_k^T} \qquad \left[(\mathbf{P^{(ij)}})^2 = \mathbf{I}\right] \\
&= \mathbf{P^{(ij)}}(\mathbf{I} + \mathbf{P^{(ij)} m_k e_k^T})
\end{aligned}$$

2. Let $\mathbf{q_k}$ be a vector containing nonzero entries only in the positions $(k+1)$ through n. Then using (1) we have

$$(\mathbf{I} + \mathbf{q_k e_k^T})\mathbf{P_i} = \mathbf{P_i}(\mathbf{I} + \mathbf{P_i q_k e_k^T}) = \mathbf{P_i}(\mathbf{I} + \mathbf{\hat{q}_k e_k^T})$$

   where the vector $\mathbf{\hat{q}_k} = \mathbf{P_i q_k}$ also has nonzero entries in the positions $(k+1)$ through n only.

   Consequently, in the product $\mathbf{P_1 L_1 \cdots P_{n-1} L_{n-1}}$, we can "propagate" each permutation matrix $\mathbf{P_i}$ (in increasing order of the index i) to the left of all matrices $\mathbf{L_k}$ with $k \leq i$ while changing each matrix $\mathbf{L_k}$ according to the equation above (multiplying its second term with $\mathbf{P_i}$ from the left). For example

$$\begin{aligned}
\mathbf{P_1 L_1 P_2 L_2 P_3 L_3} &= \mathbf{P_1(I + m_1 e_1^T)P_2(I + m_2 e_2^T)P_3(I + m_3 e_3^T)} \\
&= \mathbf{P_1 P_2(I + P_2 m_1 e_1^T)(I + m_2 e_2^T)P_3(I + m_3 e_3^T)} \\
&= \mathbf{P_1 P_2(I + P_2 m_1 e_1^T)P_3(I + P_3 m_2 e_2^T)(I + m_3 e_3^T)} \\
&= \mathbf{P_1 P_2 P_3(I + P_3 P_2 m_1 e_1^T)(I + P_3 m_2 e_2^T)(I + m_3 e_3^T)} \\
&= \mathbf{P_1 P_2 P_3 L_1^P L_2^P L_3^P}
\end{aligned}$$

   where $\mathbf{L_k^P} = \mathbf{I} + (\mathbf{P_{n-1} \cdots P_{k+1} m_k})\,\mathbf{e_k^T}$. This argument can be rigorously extended to an arbitrary n via induction.

3. Each matrix $\mathbf{L_k^P}$ can be written as $\mathbf{L_k^P} = \mathbf{I} + \mathbf{\hat{q}_k e_k^T}$ where $\mathbf{\hat{q}_k} = \mathbf{P_{n-1} \cdots P_{k+1} m_k}$, like $\mathbf{m_k}$, only has nonzero entries in the positions $(k+1)$ through n. Furthermore

$$\begin{aligned}
\mathbf{L_1^P L_2^P \cdots L_{n-1}^P} &= \mathbf{(I + \hat{q}_1 e_1^T)(I + \hat{q}_2 e_2^T) \cdots (I + \hat{q}_{n-1} e_{n-1}^T)} \\
&= \mathbf{I + \hat{q}_1 e_1^T + \hat{q}_2 e_2^T + \cdots + \hat{q}_{n-1} e_{n-1}^T}
\end{aligned}$$

   since $\mathbf{e_i^T \hat{q}_j} = \mathbf{0}$ for $i < j$, causing all the cross-terms $(\mathbf{\hat{q}_i e_i^T})(\mathbf{\hat{q}_j e_j^T})$ in the original product to vanish (for $i < j$). Since each term $\mathbf{\hat{q}_i e_i^T}$ contributes nonzero entries only below the diagonal, the entire matrix $\mathbf{I + \hat{q}_1 e_1^T + \hat{q}_2 e_2^T + \cdots + \hat{q}_{n-1} e_{n-1}^T}$ is lower triangular.

Problem 3

Two vector norms $\|\mathbf{x}\|_a$ and $\|\mathbf{x}\|_b$ are called equivalent if there exist $c, d > 0$ such that $c\|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq d\|\mathbf{x}\|_a$.

1. Prove that $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ are equivalent.

2. Prove that equivalence of two vector norms implies that their induced matrix norms are also equivalent. (The definition for equivalence of matrix norms is analogous to that of vector norms, i.e there must exist $c, d > 0$ s.t. $c\|\mathbf{A}\|_a \leq \|\mathbf{A}\|_b \leq d\|\mathbf{A}\|_a$)

Solution

1. $\|x\|_1$ equivalent to $\|x\|_\infty$: $\|x\|_1 = \sum_{i=1}^n |x_i| \le n \max_i |x_i| = n\|x\|_\infty$ and $\|x\|_\infty = \max_i |x_i| \le \sum_{i=1}^n |x_i| = \|x\|_1$. $\|x\|_\infty$ is equivalent to $\|x\|_2$: $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{1/2} \le (n \max_i(x_i^2))^{1/2} \le \sqrt{n}\sqrt{(\max_i |x_i|)^2} = \sqrt{n}\|x\|_\infty$ and $\|x\|_\infty = \max_i |x_i| = \sqrt{(\max_i |x_i|)^2} \le \sqrt{\sum_{i=1}^n x_i^2} = \|x\|_2$. $\|x\|_\infty \le \|x\|_1 = \sum_{i=1}^n |x_i| = [(\sum_{i=1}^n |x_i|)^2]^{1/2} = \|x\|_2$. $\|x\|_1$

   equivalent to $\|x\|_2$: We have $\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty$ and $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty$ by the previous results. From these we can write

$$\frac{1}{\sqrt{n}}\|x\|_2 \le \frac{\sqrt{n}}{\sqrt{n}}\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty \le n\|x\|_2.$$

2. Suppose two equivalent vector norms $\|\cdot\|_a$ and $\|\cdot\|_b$. Then we have appropriate positive constants c_a, d_a, c_b, d_b such that $c_a\|x\|_a \le \|x\|_b \le d_a\|x\|_a$ and $c_b\|x\|_b \le \|x\|_a \le d_b\|x\|_b$ hold $\forall x$. Then we can make the bound:

$$\frac{\|Ax\|_a}{\|x\|_a} \le \frac{d_b\|Ax\|_b}{\|x\|_a} \le \frac{d_b\|Ax\|_b}{c_b\|x\|_b}$$

   Thus,

$$\|A\|_a = \max_{x\neq 0} \frac{\|Ax\|_a}{\|x\|_a} \le \max_{x\neq 0} \frac{d_b}{c_b}\left(\frac{\|Ax\|_b}{\|x\|_b}\right) = \frac{d_b}{c_b} \max_{x\neq 0} \frac{\|Ax\|_b}{\|x\|_b}$$

   Similarly for the other side:

$$\frac{\|Ax\|_a}{\|x\|_a} \ge \frac{c_b\|Ax\|_b}{\|x\|_a} \ge \frac{c_b\|Ax\|_b}{d_b\|x\|_b}$$

   So,

$$\|A\|_a = \max_{x\neq 0} \frac{\|Ax\|_a}{\|x\|_a} \ge \max_{x\neq 0} \frac{c_b}{d_b}\left(\frac{\|Ax\|_b}{\|x\|_b}\right) = \frac{c_b}{d_b} \max_{x\neq 0} \frac{\|Ax\|_b}{\|x\|_b}$$