Numerical Methods 2023 — Midterm 1

Solutions

Problem 1 (20 pts). In our course, we have already learned double and single precision floating-point formats. Nevertheless, there exists another useful floating-point format that uses only 2 Bytes on the storage:

sign	exponent	significand
1 bit	5 bits	10 bits

which is called half precision floating-point, and this format has the rules:

$$\begin{cases} \pm 0, & \text{exponent} = 00000_2 \text{ and significand } = 0 \\ \pm 2^{-14} \times 0.\text{significand}_2, & \text{exponent} = 00000_2 \text{ and significand } \neq 0 \\ \pm 2^{\text{exponent}-15} \times 1.\text{significand}_2, & 11111_2 > \text{exponent} > 00000_2 \\ \pm \infty, & \text{exponent} = 11111_2 \text{ and significand } = 0 \\ \text{NaN}, & \text{exponent} = 11111_2 \text{ and significand } \neq 0 \end{cases} \tag{1}$$

We further consider rounding even in this problem.

(a) (5 pts) What is the binary representation of the value

$$528.625$$

in half precision floating-point? Please show your the calculation in the answer.

(b) (5 pts) What is the rounding error of encoding

$$528.625$$

in half precision floating-point?

(c) (10 pts) What is the largest relative error in rounding a half precision floating-point number that belongs to the interval $[-10000, 10000]$? Note that the de-normalized number should be considered, so you need to consider the second and the third cases in (1).

Solution.

(a)

$$\begin{aligned} & 528.625 \\ =& 512 + 16 + 0.5 + 0.125 \\ =& 2^9 + 2^4 + 2^{-1} + 2^{-3} \\ =& (10\ 0001\ 0000.101)_2 \\ =& 2^9 \times (1.0000\ 1000\ 0101)_2 \end{aligned} \tag{2}$$

Therefore, the significand part is rounding to

$$0000\ 1000\ 01 \tag{3}$$

and the exponent part is

$$9 + 15 = 24 = 11000_2.$$

The binary representation of 528.625 in half precision floating-point is

$$0\ 11000\ 0000\ 1000\ 01.$$

(b) Let v be the value of encoding 528.625 in half precision floating-point. By (2) and (3), we can calculate the error

$$\begin{aligned} & 528.625 - v \\ =& 2^9 \times ((1.0000\ 1000\ 0101)_2 - (1.0000\ 1000\ 01)_2) \\ =& 2^9 \times (0.0000\ 0000\ 0001)_2 \\ =& 2^{-3} \times (1.0)_2 \end{aligned}$$

Hence, the rounding error is

$$2^{-3} = 0.125.$$

(c) By the formula of relative error

$$\left| \frac{\text{real value} - \text{rounding value}}{\text{real value}} \right|,$$

let us discuss the relative error in two cases:

Case 1: $11111_2 > \text{exponent} > 00000_2$. We have relative error as

$$\begin{aligned} & \left| \frac{2^k \times (1.\text{significand}_2 + \text{rounding error}) - 2^k \times 1.\text{significand}_2}{2^k \times (1.\text{significand}_2 + \text{rounding error})} \right| \\ =& \left| \frac{\text{rounding error}}{(1.\text{significand}_2 + \text{rounding error})} \right|. \end{aligned}$$

Note that

$$\text{rounding error} < 2^{-10}$$

because we have 10 digits for

$$\text{significand}_2.$$

To get the largest relative error in this case, the significand must be as small as possible, which is zero. Therefore, we can use the following optimization problem

$$\max_{x<2^{-10}} \frac{x}{1+x} \equiv \max_{x<2^{-10}} 1 - \frac{1}{1+x}$$

to get the largest relative error, and the solution x^* is

$$2^{-10} \times (0.1111\ldots)_2,$$

which implies the largest relative error is bounded by

$$1 - \frac{1}{1+2^{-10}}$$

in this case.

Case 2: exponent $= 00000_2$. Similarly, we have relative error as

$$\left|\frac{2^{-14} \times (0.\text{significand}_2 + \text{rounding error}) - 2^{-14} \times 0.\text{significand}_2}{2^{-14} \times (0.\text{significand}_2 + \text{rounding error})}\right| = \left|\frac{\text{rounding error}}{(0.\text{significand}_2 + \text{rounding error})}\right|.$$

When significand is equal to zero, the largest relative error is

$$\frac{\text{rounding error}}{(0.0 + \text{rounding error})} = 1$$

in this case.

By combining Case 1 and Case 2, we have the largest relative error as

$$\max\left(1 - \frac{1}{1 + 2^{-10}}, 1\right) = 1.$$

Problem 2 (10 pts). Consider the following matrix

$$A = \begin{bmatrix} 9 & 3 & 18 & 27 \\ 3 & 17 & 10 & 21 \\ 18 & 10 & 62 & 72 \\ 27 & 21 & 72 & 100 \end{bmatrix}.$$

Show every step of performing Cholesky factorization on A with the outer product form. Hint: the resulting L contains only integer values.

Solution.

Step 1: Calculate $\sqrt{\alpha} = \sqrt{9} = 3$, so that

$$L^{(1)} = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 6 & 0 & 1 & 0 \\ 9 & 0 & 0 & 1 \end{bmatrix}$$

in this step. Also, we calculate

$$B - \frac{\boldsymbol{v} \cdot \boldsymbol{v}^T}{\sqrt{\alpha}} = \begin{bmatrix} 17 & 10 & 21 \\ 10 & 62 & 72 \\ 21 & 72 & 100 \end{bmatrix} - \begin{bmatrix} 1 & 6 & 9 \\ 6 & 36 & 54 \\ 9 & 54 & 81 \end{bmatrix} = \begin{bmatrix} 16 & 4 & 12 \\ 4 & 26 & 18 \\ 12 & 18 & 19 \end{bmatrix}.$$

Thus, we have

$$A^{(1)} = \begin{bmatrix} 3 & 1 & 6 & 9 \\ 1 & 16 & 4 & 12 \\ 6 & 4 & 26 & 18 \\ 9 & 12 & 18 & 19 \end{bmatrix}.$$

Step 2: Now we re-do the calculation on $A^{(1)}$, we have $\sqrt{\alpha^{(1)}} = \sqrt{16} = 4$, and

$$L^{(2)} = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 6 & 1 & 1 & 0 \\ 9 & 3 & 0 & 1 \end{bmatrix}$$

in this step. Then, calculate

$$B^{(1)} - \frac{\boldsymbol{v}^{(1)} \cdot \left(\boldsymbol{v}^{(1)}\right)^T}{\sqrt{\alpha^{(1)}}} = \begin{bmatrix} 26 & 18 \\ 18 & 19 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 3 & 9 \end{bmatrix} = \begin{bmatrix} 25 & 15 \\ 15 & 10 \end{bmatrix}.$$

Therefore,

$$A^{(2)} = \begin{bmatrix} 3 & 1 & 6 & 9 \\ 1 & 4 & 1 & 3 \\ 6 & 1 & 25 & 15 \\ 9 & 3 & 15 & 10 \end{bmatrix}.$$

Step 3: Similarly, we have $\sqrt{\alpha^{(2)}} = \sqrt{25} = 5$, and

$$L^{(3)} = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 6 & 1 & 5 & 0 \\ 9 & 3 & 3 & 1 \end{bmatrix}$$

in this step. Furthermore,

$$B^{(2)} - \frac{\boldsymbol{v}^{(2)} \cdot \left(\boldsymbol{v}^{(2)}\right)^T}{\sqrt{\alpha^{(2)}}} = \begin{bmatrix} 10 \end{bmatrix} - \begin{bmatrix} 9 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix}.$$

Therefore,

$$A^{(3)} = \begin{bmatrix} 3 & 1 & 6 & 9 \\ 1 & 4 & 1 & 3 \\ 6 & 1 & 5 & 3 \\ 9 & 3 & 3 & 1 \end{bmatrix}.$$

Step 4: In the final, $\sqrt{\alpha^{(3)}} = \sqrt{1} = 1$, so that

$$L^{(4)} = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 6 & 1 & 5 & 0 \\ 9 & 3 & 3 & 1 \end{bmatrix}$$

is the same as $L^{(3)}$. Moreover, $A^{(4)}$ is also equal to $A^{(3)}$.

Problem 3 (30 pts). Consider the following matrix:

$$A = \begin{bmatrix} 15 & 18 & 60 \\ 5 & 42 & 95 \\ -35 & -21 & -119 \end{bmatrix}$$

(a) (5 pts) In pivoted LU factorization, please give

$$P_1, M_1, P_2, M_2, U$$

such that

$$M_2 P_2 M_1 P_1 A = U.$$

Note that you should choose the pivot which has the largest absolute value. Hint: U has only integer values.

(b) (5 pts) Following (a), what are the P and L such that $PA = LU$?

(c) (10 pts) In pivoted LU factorization, we have a general method that considers the pivot in both rows and columns

$$PAQ = LU,$$

where Q is also a permutation matrix that can interchange the columns. To simplify the calculations, the number of row and column interchange is limited to **ONE** when a pivot is determined, i.e., you cannot pick both row interchange and column interchange. For example, the pivot of the $(1, 1)$ position is the largest absolute value element of the first row and the first column. If it is from the first column, then

$$Q_1 = I,$$

otherwise

$$P_1 = I.$$

Follow the similar idea of (a), please give

$$P_1, Q_1, M_1, P_2, Q_2, M_2$$

and U so that

$$M_2 P_2 M_1 P_1 \cdot A \cdot Q_1 Q_2 = U.$$

Note that you should choose the pivot which has the largest absolute value. If there are more than two candidates of pivot, we choose the one that is the closest to the diagonal part. Note that U has fraction numbers, so you want to be careful in doing the calculation.

(d) (5 pts) Following (c), please calculate P, Q and L such that

$$PAQ = LU.$$

(e) (5 pts) Following (d), solving the linear system

$$A\boldsymbol{x} = \boldsymbol{b}$$

is equivalent to solve

$$P^{-1} L U Q^{-1} \boldsymbol{x} = \boldsymbol{b}.$$

Therefore, we can get the solution $\boldsymbol{x}^*$ by the following steps.

(i) Solve $L\boldsymbol{y} = P\boldsymbol{b}$ to get $\boldsymbol{y}^*$, where $\boldsymbol{y}^* = UQ^{-1}\boldsymbol{x}$.

(ii) Solve $U\boldsymbol{z} = \boldsymbol{y}^*$ to get $\boldsymbol{z}^*$, where $\boldsymbol{z}^* = Q^{-1}\boldsymbol{x}$.

(iii) Solve $Q^{-1}\boldsymbol{x} = \boldsymbol{z}^*$ to get $\boldsymbol{x}^*$.

Now, we have

$$\boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Please solve

$$A\boldsymbol{x} = \boldsymbol{b}$$

by the aforementioned steps.

Solution.

(a) Step 1: The pivot is -35, so we have

$$P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

such that

$$P_1 A = \begin{bmatrix} -35 & -21 & -119 \\ 5 & 42 & 95 \\ 15 & 18 & 60 \end{bmatrix}.$$

Then, we can calculate

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1/7 & 1 & 0 \\ 3/7 & 0 & 1 \end{bmatrix},$$

so that

$$M_1 P_1 A = \begin{bmatrix} -35 & -21 & -119 \\ 0 & 39 & 78 \\ 0 & 9 & 9 \end{bmatrix}.$$

Step 2: The pivot is 39, so

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and $P_2 M_1 P_1 A$ is still

$$\begin{bmatrix} -35 & -21 & -119 \\ 0 & 39 & 78 \\ 0 & 9 & 9 \end{bmatrix}.$$

Thus, M_2 is calculated by

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3/13 & 1 \end{bmatrix},$$

and

$$M_2 P_2 M_1 P_1 A = \begin{bmatrix} -35 & -21 & -119 \\ 0 & 39 & 78 \\ 0 & 0 & -9 \end{bmatrix}.$$

(b) We have

$$P = P_2 P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

and

$$\begin{aligned} L =& M_1^{-1} M_2^{-1} \\ =& \begin{bmatrix} 1 & 0 & 0 \\ -1/7 & 1 & 0 \\ -3/7 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3/13 & 1 \end{bmatrix} \\ =& \begin{bmatrix} 1 & 0 & 0 \\ -1/7 & 1 & 0 \\ -3/7 & 3/13 & 1 \end{bmatrix} \end{aligned}$$

(c) Step 1: The pivot is 60, so

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$Q_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

such that

$$P_1 \cdot A \cdot Q_1 = \begin{bmatrix} 60 & 18 & 15 \\ 95 & 42 & 5 \\ -119 & -21 & -35 \end{bmatrix}.$$

Thereby, we can calculate

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ -95/60 & 1 & 0 \\ 119/60 & 0 & 1 \end{bmatrix}$$

and

$$M_1 P_1 \cdot A \cdot Q_1 = \begin{bmatrix} 60 & 18 & 15 \\ 0 & 27/2 & -75/4 \\ 0 & 147/10 & -21/4 \end{bmatrix}.$$

Step 2: Now the pivot is $-75/4$, so

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$Q_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

such that

$$P_2 M_1 P_1 \cdot A \cdot Q_1 Q_2 = \begin{bmatrix} 60 & 15 & 18 \\ 0 & -75/4 & 27/2 \\ 0 & -21/4 & 147/10 \end{bmatrix}.$$

Hence, we have

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -21/75 & 1 \end{bmatrix}$$

and then

$$M_2 P_2 M_1 P_1 \cdot A \cdot Q_1 Q_2 = \begin{bmatrix} 60 & 15 & 18 \\ 0 & -75/4 & 27/2 \\ 0 & 0 & 273/25 \end{bmatrix}.$$

(d) We have

$$P = P_2 P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$Q = Q_1 Q_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

and

$$\begin{aligned} L &= M_1^{-1} M_2^{-1} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 95/60 & 1 & 0 \\ -119/60 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 21/75 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 95/60 & 1 & 0 \\ -119/60 & 21/75 & 1 \end{bmatrix}. \end{aligned}$$

(e) In step (i), we have to solve

$$\begin{aligned} & L\boldsymbol{y} = P\boldsymbol{b} \\ \Rightarrow & \begin{bmatrix} 1 & 0 & 0 \\ 95/60 & 1 & 0 \\ -119/60 & 21/75 & 1 \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{b} \\ \Rightarrow & \begin{bmatrix} 1 & 0 & 0 \\ 95/60 & 1 & 0 \\ -119/60 & 21/75 & 1 \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \end{aligned}$$

and the solution $\boldsymbol{y}^*$ can be calculated as

$$\boldsymbol{y}^* = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

In step (ii), we have to solve

$$\begin{aligned} & U\boldsymbol{z} = \boldsymbol{y}^* \\ \Rightarrow & \begin{bmatrix} 60 & 15 & 18 \\ 0 & -75/4 & 27/2 \\ 0 & 0 & 273/25 \end{bmatrix} \boldsymbol{z} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \end{aligned}$$

The solution $\boldsymbol{z}^*$ can be calculated as

$$\boldsymbol{z}^* = \begin{bmatrix} -12/273 \\ 18/273 \\ 25/273 \end{bmatrix}.$$

In step (iii), we have to solve

$$Q^{-1}\boldsymbol{x} = \boldsymbol{z}^*,$$

which implies

$$\boldsymbol{x} = Q\boldsymbol{z}^* = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \boldsymbol{z}^*.$$

Thus,

$$\boldsymbol{x}^* = \begin{bmatrix} 18/273 \\ 25/273 \\ -12/273 \end{bmatrix}.$$

Problem 4 (25 pts). In our lecture slide "linear_matrixcondition2.pdf", we have defined the condition of a matrix A to be $\|A\|\|A^{-1}\|$. In this problem, we explore the relation between the condition number and the singular value decomposition.

Assume that we have a real, invertible, $n \times n$ matrix A. We say that the matrix A has a singular value decomposition if there exists matrices U, Σ, V such that

$$A = U\Sigma V^T$$

where U, V are $n \times n$ orthogonal matrices satisfying $U^T = U^{-1}, V^T = V^{-1}$,
and Σ is an $n \times n$ diagonal matrix with $\Sigma_{ii} = \sigma_i > 0$.

In this problem, let us use u_i and v_i to denote the ith column in U and V, respectively.

(a) (5 pts) Use the definition of matrix norm given in our lecture slide "linear_matrixcondition1.pdf" to show that for A we have

$$\|A\|_2 = \max_{\|x\|_2=1} \left\|\Sigma V^T x\right\|_2.$$

(b) (10 pts) Following subproblem (a), show that

$$\max_{\|x\|_2=1} \left\|\Sigma V^T x\right\|_2 = \max_i \sigma_i$$

to conclude that

$$\|A\|_2 = \max_i \sigma_i.$$

Hint: One way to do the proof is by first showing that

$$\left\|\Sigma V^T x\right\|_2 \leq \max_i \sigma_i \text{ for all } x \text{ satisfying } \|x\|_2 = 1,$$

and then show that

$$\left\|\Sigma V^T x\right\|_2 = \max_i \sigma_i \text{ for some } x \text{ satisfying } \|x\|_2 = 1.$$

(c) (5 pts) Show that for the inverse of A we also have

$$\left\|A^{-1}\right\|_2 = \max_i \frac{1}{\sigma_i}.$$

Then, conclude that the condition number of A induced by the norm $\|\cdot\|_2$ is equivalent to

$$\frac{\max_i \sigma_i}{\min_i \sigma_i}.$$

(d) (5 pts) Recall from lecture slide "linear_matrixcondition2.pdf" that when solving a linear system

$$Ax = b,$$

we have the inequality

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \|A\|_2 \left\|A^{-1}\right\|_2 \frac{\|\delta b\|_2}{\|b\|_2} \tag{4}$$

to quantify the relative change in the solution x when the input b is slightly purturbed.

Combining equation (4) with the conclusion in subproblem (c), we can obtain this inequality:

$$\frac{\|\delta x\|_2 \,/\, \|x\|_2}{\|\delta b\|_2 \,/\, \|b\|_2} \leq \frac{\max_i \sigma_i}{\min_i \sigma_i} \tag{5}$$

In this problem, we discuss the situation when the left hand side of (5) becomes equal to the right hand side (i.e., when x becomes the most sensitive to purturbation in b). Let

$$i_{\max} = \arg\max_i \sigma_i \text{ and } i_{\min} = \arg\min_i \sigma_i.$$

You need to show that

$$\text{if } b = c_1 u_{i_{\max}} \text{ and } \delta b = c_2 u_{i_{\min}} \text{ for some } c_1, c_2 \in \mathbb{R}$$
$$\text{then the equality in (5) holds.}$$

(This result means that, when b is parallel to the column vector in U corresponding to the largest singular value and δb is parallel to that of the smallest singular value, the relative change in x is the largest.)

Solution.

(a) By the definition of matrix norm, we have

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \left\|U\Sigma V^T x\right\|_2. \tag{6}$$

Because U is orthogonal, we have

$$\|Uw\|_2 = \sqrt{(Uw)^T(Uw)} = \sqrt{w^T U^T U w} = \sqrt{w^T w} = \|w\|_2 \tag{7}$$

for any $w \in \mathbb{R}^n$. Therefore, (6) can be simplified to

$$\max_{\|x\|_2=1} \left\|\Sigma V^T x\right\|_2.$$

(b) We show that $\left\|\Sigma V^T x\right\|_2$ have an upperbound $\max_i \sigma_i$ and the upperbound is attained by some vector. To derive the upperbound, let

$$y = V^T x = V^{-1} x.$$

Then, we can see that

$$\begin{aligned}
\max_{\|x\|_2=1} \left\|\Sigma V^T x\right\|_2 &= \max_{\|Vy\|_2=1} \|\Sigma y\|_2 \\
&= \max_{\|y\|_2=1} \|\Sigma y\|_2 && \text{(for the same reason in (7))} \\
&= \max_{\|y\|_2=1} \sqrt{\sigma_1^2 y_1^2 + \ldots + \sigma_n^2 y_n^2} \\
&\leq \max_{\|y\|_2=1} \left[(\max_i \sigma_i) \sqrt{y_1^2 + \ldots + y_n^2} \right] \\
&= \max_{\|y\|_2=1} \left[(\max_i \sigma_i) \|y\|_2 \right] \\
&= \max_{\|y\|_2=1} \left[(\max_i \sigma_i) \cdot 1 \right] \\
&= \max_i \sigma_i && (8)
\end{aligned}$$

Next, we need to show that the upperbound is attained. Let

$$i' = \arg\max_i \sigma_i.$$

Then, the upperbound (8) is attained when

$$x = v_{i'}.$$

This is because

$$\left\|\Sigma V^T x\right\|_2 = \|\Sigma e_{i'}\|_2 = \sigma_{i'} = \max_i \sigma_i.$$

In the equation above, $e_{i'}$ is the vector that has 1 in the i'th position and zero at other positions.

(c) The singular value decomposition of A^{-1} can be calculated as

$$A^{-1} = (U\Sigma V^T)^{-1} = (V^T)^{-1}\Sigma^{-1}U^{-1} = V\Sigma^{-1}U^T,$$
$$\text{where } \Sigma_{ii}^{-1} = \frac{1}{\sigma_i}.$$

Then, using the result from subproblem (a) and (b), we can know that

$$\left\|A^{-1}\right\|_2 = \max_i \frac{1}{\sigma_i}.$$

Therefore, the condition number of A is thus

$$\|A\|_2 \left\|A^{-1}\right\|_2 = (\max_i \sigma_i)(\max_i \frac{1}{\sigma_i}) = \frac{\max_i \sigma_i}{\min_i \sigma_i}.$$

(d) We can reorganize (5) into this form:

$$\frac{\|\delta x\|_2}{\|\delta b\|_2} \cdot \frac{\|b\|_2}{\|x\|_2} \leq \frac{\max_i \sigma_i}{\min_i \sigma_i} \tag{9}$$

Because $b = c_1 u_{i_{\max}}$, we have

$$x = A^{-1}b = V\Sigma^{-1}U^T c_1 u_{i_{\max}} = c_1 V\Sigma^{-1} e_{i_{\max}} = \frac{c_1}{\max_i \sigma_i} V e_{i_{\max}} = \frac{c_1}{\max_i \sigma_i} v_{i_{\max}}.$$

Similarly, since $\delta b = c_2 u_{i_{\min}}$, we have

$$\delta x = A^{-1}\delta b = \frac{c_2}{\min_i \sigma_i} v_{i_{\min}}.$$

Plugging all the information we have into (9), we get

$$\frac{\|\delta x\|_2}{\|\delta b\|_2} \cdot \frac{\|b\|_2}{\|x\|_2} = \frac{\frac{c_2}{\min_i \sigma_i}}{c_2} \cdot \frac{c_1}{\frac{c_1}{\max_i \sigma_i}} = \frac{\max_i \sigma_i}{\min_i \sigma_i},$$

so the equality indeed holds.

Problem 5 (15 pts). On page 5 of lecture slide "fp_guarddigit1.pdf", we showed that subtracting near values of the same sign can lead to large relative errors. Then, in slide "fp_guarddigit2.pdf", we showed that with an guard digit, the relative error becomes small. However, the proof only applies to subtracting numbers of the same sign.

In this problem, we discuss the error of adding two floating-point numbers of the same sign when **using p digits with no guard digit**. As in the lecture slides, we will assume that

$$x > y \geq 0$$
$$\text{and } x = x_0.x_1x_2\cdots x_{p-1} \times \beta^0 \text{ with } x_0 \geq 1 \text{ and } \beta \geq 2.$$

We assume that before addition, y is shifted and then **truncated** to p digits. We denote the truncated value of y after shifting as $\overline{y}$. Then, after the addition is done, we assume the value is rounded to the closest number (using **rounding even**) before being stored. The rounding error is denoted δ. To discuss the final error generated in this process, we separate our discussion into two cases. You need to analyze the error in these two cases using $y, \overline{y}$ and δ, similar to the proof in lecture slide "fp_guarddigit2.pdf".

(a) (10 pts) $x + y < \beta$: In this case, we must also have $x + \overline{y} < \beta$ because truncation can only decrease the result. Since there is no carry out to the β^1 bit, no rounding have to performed after addition. The only error arises from the discarded bits when shifting y. Show that the relative error is less than or equal to 2ϵ.

(b) (5 pts) $x + y \geq \beta$: In this case, we must also have $x + \overline{y} \geq \beta$. This is because the addition in the truncated position does not carry over to the untruncated bits, and thus removing it does not affect the most significant bit. Therefore, there is a carry out to the β^1 bit and you also have to consider the rounding error after computing the sum. Show that the relative error is still less than or equal to 2ϵ.

Solution.

(a) Because no rounding have to performed after computation, the error is

$$|(x+y)-(x+\overline{y})| = |y-\overline{y}| = y-\overline{y}.$$

Suppose l bits is shifted out, then the error is at most

$$y-\overline{y} \leq (\beta-1)(\beta^{-p}+\beta^{-p-1}+\cdots+\beta^{-p-l+1}) < \beta^{-p+1}$$

no matter how large l is. Then, the relative error is

$$\frac{y-\overline{y}}{x+y} \leq \frac{\beta^{-p+1}}{1} = 2(\frac{1}{2}\beta^{-p+1}) = 2\epsilon$$

because $x+y \geq 1$.

(b) Because there is a carry out, we have to round off the rightmost bit before storing the result. The rounded result can be written as

$$x+\overline{y}+\delta \text{ where } |\delta| \leq \frac{\beta^{-p+2}}{2}.$$

Therefore, the error is

$$|(x+y)-(x+\overline{y}+\delta)| = |(y-\overline{y})-\delta| \leq |y-\overline{y}|+|\delta| \leq \beta^{-p+1}+\frac{\beta^{-p+2}}{2}.$$

The relative error is then

$$\frac{|(x+y)-(x+\overline{y}+\delta)|}{x+y} \leq \frac{\frac{1}{2}\beta^{-p+2}(1+\frac{2}{\beta})}{\beta} = (1+\frac{2}{\beta})\frac{1}{2}\beta^{-p+1} \leq 2\epsilon.$$

The first inequality is due to $x+y \geq \beta$ and the last inequality is due to $\beta \geq 2$.