MATH50003 Problem Sheets and Solutions

03/04/2022

Week 1

1. Binary representation

Problem 1.1 What is the binary representation of 1/5?

SOLUTION
Hence we show that

$$(0.00110011001100...)_2 = (2^{-3} + 2^{-4})(1.00010001000...)_2 = (2^{-3} + 2^{-4}) \sum_{k=0}^{\infty} \frac{1}{16^k}$$
$$= \frac{2^{-3} + 2^{-4}}{1 - \frac{1}{2^4}} = \frac{1}{5}$$

2. Integers

Problem 2.1 With 8-bit signed integers, find the bits for the following: $10, 120, -10$.

SOLUTION

We can find the binary digits by repeatedly subtracting the largest power of 2 less than a number until we reach 0, e.g. $10 - 2^3 - 2 = 0$ implies $10 = (1010)_2$. Thus the bits are: 00001010.

For negative numbers we perform the same trick but adding 2^p to make it positive, e.g.,

$$-10 = 2^8 - 10(\text{mod} 2^8) = 246 = 2^7 + 2^6 + 2^5 + 2^4 + 2^2 + 2 = (11110110)_2$$

Thus the bits are 11110110

3. Floating point numbers

Problem 3.1 What are the single precision F_{32} (`Float32`) floating point representations for the following:

$$2, 31, 32, 23/4, (23/4) \times 2^{100}$$

SOLUTION

Recall that we have $\sigma, Q, S = 127, 8, 23$. Thus we write

$$2 = 2^{128-127} * (1.00000000000000000000000)_2$$

The exponent bits are those of

$$128 = 2^7 = (10000000)_2$$

We write

$$31 = (11111)_2 = 2^{131-127} * (1.1111)_2$$

And note that $131 = (10000011)_2$ Hence we have: 00100000111111000000000000000000

On the other hand,

$$32 = (100000)_2 = 2^{132-127}$$

and $132 = (10000100)_2$ hence: 00100001000000000000000000000000

Note that

$$23/4 = 2^{-2} * (10111)_2 = 2^{129-127} * (1.0111)_2$$

and $129 = (10000001)_2$ hence we get: 00100000010111000000000000000000

Finally,

$$23/4 * 2^{100} = 2^{229-127} * (1.0111)_2$$

and $229 = (11100101)_2$ giving us: 00111001010111000000000000000000

Problem 3.2 Let $m(y) = \min\{x \in F_{32} : x > y\}$ be the smallest single precision number greater than y. What is $m(2) - 2$ and $m(1024) - 1024$?

SOLUTION

The next float after 2 is $2 * (1 + 2^{-23})$ hence we get $m(2) - 2 = 2^{-22}$, similarly, for $1024 = 2^{10}$ we find that the difference $m(1024) - 1024$ is $2^{10-23} = 2^{-13}$

4. Arithmetic

Problem 4.1 Suppose $x = 1.25$ and consider 16-bit floating point arithmetic (`Float16`). What is the error in approximating x by the nearest float point number $\mathrm{fl}(x)$? What is the error in approximating $2x$, $x/2$, $x + 2$ and $x - 2$ by $2 \otimes x$, $x \oslash 2$, $x \oplus 2$ and $x \ominus 2$?

SOLUTION

None of these computations have errors since they are all exactly representable as floating point numbers.

Problem 4.2 For what floating point numbers is $x \oslash 2 \neq x/2$ and $x \oplus 2 \neq x + 2$?

SOLUTION

Consider a normal $x = 2^{q-\sigma}(1.b_1 \ldots b_S)_2$. Provided $q > 1$ we have

$$x \oslash 2 = x/2 = 2^{q-\sigma-1}(1.b_1 \ldots b_S)_2$$

However, if $q = 1$ we lose a bit as we shift:

$$x \oslash 2 = 2^{1-\sigma}(0.b_1 \ldots b_{S-1})_2$$

and the property will be satisfy if $b_S = 1$.

Similarly, if we are sub-normal, $x = 2^{1-\sigma}(0.b_1 \ldots b_S)_2$ and we have

$$x \oslash 2 = 2^{1-\sigma}(0.0b_1 \ldots b_{S-1})_2$$

and the property will be satisfy if $b_S = 1$. (Or `NaN`.)

Problem 4.3 Explain why for `x = 10.0^100`, we have $x = x + 1$. What is the largest floating point number y such that $y + 1 \neq y$?

SOLUTION

Writing $10 = 2^3(1.01)_2$ we have

$$\mathrm{fl}(10^{100}) = \mathrm{fl}(2^{300}(1 + 2^{-4})^{100}) = 2^{300}(1.\mathrm{b}_1 \ldots \mathrm{b}_{52})_2$$

where the bits b_k are not relevant. We then have:

$$\mathrm{fl}(10^{100}) \oplus 1 = \mathrm{fl}(2^{300}[(1.\mathrm{b}_1 \ldots \mathrm{b}_{52})_2 + 2^{-300}]) = \mathrm{fl}(10^{100})$$

since 2^{-300} is below the necessary precision.

The largest floating point number satisfying the condition is $y = 2^{53} - 1$

Problem 4.4 What are the exact bits for $1/5$, $1/5 + 1$ computed using half-precision arithmetic (`Float16`) (using default rounding)?

SOLUTION

We saw above that

$$1/5 = 2^{-3} * (1.10011001100\ldots)_2 \approx 2^{-3} * (1.1001100110)_2$$

where the $\approx$ is rounded to the nearest 10 bits (in this case rounded down). We write $-3 = 12 - 15$ hence we have $q = 12 = (01100)_2$.

Adding `1` we get:

$$1 + 2^{-3} * (1.1001100110)_2 = (1.001100110011)_2 \approx (1.0011001101)_2$$

Here we write the exponent as $0 = 15 - 15$ where $q = 15 = (01111)_2$. Thus we get: 0011110011001101.

Problem 4.5 Explain why $F_{16}(0.1)/(F_{16}(1.1) - 1)$ does not return 1. Can you compute the bits explicitly?

SOLUTION

For the last problem, note that

$$\frac{1}{10} = \frac{1}{2}\frac{1}{5} = 2^{-4} * (1.10011001100...)_2$$

hence we have

$$\mathrm{fl}(\frac{1}{10}) = 2^{-4} * (1.1001100110)_2$$

and

$$\mathrm{fl}(1 + \frac{1}{10}) = \mathrm{fl}(1.0001100110011\ldots) = (1.0001100110)_2$$

Thus

$$\mathrm{fl}(1.1) \ominus 1 = (0.0001100110)_2 = 2^{-4}(1.1001100000)_2$$

and hence we get

$$\mathrm{fl}(0.1) \oslash (\mathrm{fl}(1.1) \ominus 1) = \mathrm{fl}(\frac{(1.1001100110)_2}{(1.1001100000)_2}) \neq 1$$

To compute the bits explicitly, write $y = (1.10011)_2$ and divide through to get:

$$\frac{(1.1001100110)_2}{(1.10011)_2} = 1 + \frac{2^{-8}}{y} + \frac{2^{-9}}{y}$$

We then have

$$y^{-1} = \frac{32}{51} = 0.627\ldots = (0.101\ldots)_2$$

Hence

$$1 + \frac{2^{-8}}{y} + \frac{2^{-9}}{y} = 1 + (2^{-9} + 2^{-11} + \cdots) + (2^{-10} + \cdots) = (1.00000000111...)_2$$

Therefore we round up (the ... is not exactly zero but if it was it would be a tie and we would round up anyways to get a zero last bit).

Problem 4.6 Find a bound on the *absolute error* in terms of a constant times ϵ_{m} for the following computations

$$(1.1 * 1.2) + 1.3$$
$$(1.1 - 1)/0.1$$

implemented using floating point arithmetic (with any precision).

SOLUTION

The first problem is very similar to what we saw in lecture. Write

$$(\mathrm{fl}(1.1) \otimes \mathrm{fl}(1.2)) \oplus \mathrm{fl}(1.3) = [\ 1.1(1+\delta_1) \times 1.2(1+\delta_2)(1+\delta_3) + 1.3(1+\delta_4)\ ] \times (1+\delta_5)$$

We first write

$$1.1(1+\delta_1)1.2(1+\delta_2)(1+\delta_3) = 1.32(1+\delta_6)$$

where

$$|\delta_6| \leq |\delta_1| + |\delta_2| + |\delta_3| + |\delta_1||\delta_2| + |\delta_1||\delta_3| + |\delta_2||\delta_3| + |\delta_1||\delta_2||\delta_3| \leq 4\epsilon_{\mathrm{m}}$$

Then we have

$$1.32(1+\delta_6) + 1.3(1+\delta_4) = 2.62 + \underbrace{1.32\delta_6 + 1.3\delta_4}_{\delta_7}$$

where

$$|\delta_7| \leq 7\epsilon_{\mathrm{m}}$$

Finally,

$$(2.62 + \delta_6)(1 + \delta_5) = 2.62 + \underbrace{\delta_6 + 2.62\delta_5 + \delta_6\delta_5}_{\delta_8}$$

where

$$|\delta_8| \leq 10\epsilon_{\rm m}$$

For the second part, we do:

$$({\rm fl}(1.1) \ominus 1) \oslash {\rm fl}(0.1) = \frac{(1.1(1 + \delta_1) - 1)(1 + \delta_2)}{0.1(1 + \delta_3)}(1 + \delta_4)$$

Write

$$\frac{1}{1 + \delta_3} = 1 + \delta_5$$

where

$$|\delta_5| \leq \left|\frac{\delta_3}{1 + \delta_3}\right| \leq \frac{\epsilon_{\rm m}}{2}\frac{1}{1 - 1/2} \leq \epsilon_{\rm m}$$

using the fact that $|\delta_3| < 1/2$. Further write

$$(1 + \delta_5)(1 + \delta_4) = 1 + \delta_6$$

where

$$|\delta_6| \leq |\delta_5| + |\delta_4| + |\delta_5||\delta_4| \leq 2\epsilon_{\rm m}$$

We also write

$$\frac{(1.1(1 + \delta_1) - 1)(1 + \delta_2)}{0.1} = 1 + \underbrace{11\delta_1 + \delta_2 + 11\delta_1\delta_2}_{\delta_7}$$

where

$$|\delta_7| \leq 17\epsilon_{\rm m}$$

Then we get

$$({\rm fl}(1.1) \ominus 1) \oslash {\rm fl}(0.1) = (1 + \delta_7)(1 + \delta_6) = 1 + \delta_7 + \delta_6 + \delta_6\delta_7$$

and the error is bounded by:

$$(17 + 2 + 34)\epsilon_{\rm m} = 53\epsilon_{\rm m}$$

This is quite pessimistic but still captures that we are on the order of $\epsilon_{\rm m}$.

Week 2

1. Finite-differences

Problem 1.1 Use Taylor's theorem to derive an error bound for central differences

$$f'(x) \approx \frac{f(x + h) - f(x - h)}{2h}.$$

Find an error bound when implemented in floating point arithmetic, assuming that

$$f^{\rm FP}(x) = f(x) + \delta_x$$

where $|\delta_x| \leq c\epsilon_{\rm m}$.

SOLUTION

By Taylor's theorem, the approximation around $x+h$ is

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(z_1)}{6}h^3,$$

for some $z_1 \in (x, x+h)$ and similarly

$$f(x-h) = f(x) + f'(x)(-h) + \frac{f''(x)}{2}h^2 - \frac{f'''(z_2)}{6}h^3,$$

for some $z_2 \in (x-h, x)$.

Subtracting the second expression from the first we obtain

$$f(x+h) - f(x-h) = f'(x)(2h) + \frac{f'''(z_1) + f'''(z_2)}{6}h^3.$$

Hence,

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \underbrace{\frac{f'''(z_1) + f'''(z_2)}{12}h^2}_{\delta_{\text{Taylor}}}.$$

Thus, the error can be bounded by

$$\left|\delta_{\text{Taylor}}\right| \leq \frac{M}{6}h^2,$$

where

$$M = \max_{y \in [x-h, x+h]} \left|f'''(y)\right|.$$

In floating point we have

$$\begin{aligned}(f^{\text{FP}}(x+2h) \ominus f^{\text{FP}}(x-2h)) \oslash (2h) &= \frac{f(x+h) + \delta_{x+h} - f(x-h) - \delta_{x-h}}{2h}(1+\delta_1) \\ &= \frac{f(x+h) - f(x-h)}{2h}(1+\delta_1) + \frac{\delta_{x+h} - \delta_{x-h}}{2h}(1+\delta_1)\end{aligned}$$

Applying Taylor's theorem we get

$$(f^{\text{FP}}(x+h) \ominus f^{\text{FP}}(x-h)) \oslash (2h) = f'(x) + \underbrace{f'(x)\delta_1 + \delta_{\text{Taylor}}(1+\delta_1) + \frac{\delta_{x+h} - \delta_{x-h}}{2h}(1+\delta_1)}_{\delta^{\text{CD}}_{x,h}}$$

where

$$|\delta^{\text{CD}}_{x,h}| \leq \frac{|f'(x)|}{2}\epsilon_{\text{m}} + \frac{M}{3}h^2 + \frac{2c\epsilon_{\text{m}}}{h}$$

To compute the errors of the central difference approximation of $f'(x)$ we compute

$$\left|\frac{f(x+h) - f(x-h)}{2h} - f'(x)\right| =$$

$$= \left|\frac{1 + (x+h) + (x+h)^2 - 1 - (x-h) - (x-h)^2}{2h} - (1+2x)\right| =$$

$$= \left|\frac{2h + 4hx}{2h} - 1 - 2x\right| = 0.$$

As we can see, in this case the central difference approximation is exact. The errors we start observing for small step sizes are thus numerical in nature. The values of the function at $f(x+h)$ and $f(x-h)$ eventually become numerically indistinguishable and thus this finite difference approximation to the derivative incorrectly results in 0.

To compute the errors of the central difference approximation of $g'(x)$ we compute

$$\left|\frac{g(x+h)-g(x-h)}{2h} - g'(x)\right| =$$

$$= \left|\frac{1+\frac{(x+h)}{3}+(x+h)^2-1-\frac{(x-h)}{3}-(x-h)^2}{2h} - \left(\frac{1}{3}+2x\right)\right| =$$

$$= \left|\frac{1}{3}+2x-\frac{1}{3}-2x\right| = 0.$$

Problem 1.3 (A) Use Taylor's theorem to derive an error bound on the second-order derivative approximation

$$f''(x) \approx \frac{f(x+h)-2f(x)+f(x-h)}{h^2}$$

Find an error bound when implemented in floating point arithmetic, assuming that

$$f^{\mathrm{FP}}(x) = f(x) + \delta_x$$

where $|\delta_x| \leq c\epsilon_{\mathrm{m}}$.

SOLUTION

Using the same two formulas as in 1.1 we have

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(z_1)}{6}h^3,$$

for some $z_1 \in (x, x+h)$ and

$$f(x-h) = f(x) + f'(x)(-h) + \frac{f''(x)}{2}h^2 - \frac{f'''(z_2)}{6}h^3,$$

for some $z_2 \in (x-h, x)$.

Summing the two we obtain

$$f(x+h) + f(x-h) = 2f(x) + f''(x)h^2 + \frac{f'''(z_1)}{6}h^3 - \frac{f'''(z_2)}{6}h^3.$$

Thus,

$$f''(x) = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} + \frac{f'''(z_2)-f'''(z_1)}{6}h.$$

Hence, the error is

$$\left|f''(x) - \frac{f(x+h)-2f(x)+f(x-h)}{h^2}\right| = \left|\frac{f'''(z_2)-f'''(z_1)}{6}h\right| \leq 2Ch,$$

where again

$$C = \max_{y\in[x-h,x+h]}\left|\frac{f'''(y)}{6}\right|.$$

In floating point arithmetic, the error is

$$\begin{aligned}
\left|f''^{FP}(x) - f''(x)\right| &= \left|\frac{f^{FP}(x+h) - 2f^{FP}(x) + f^{FP}(x-h)}{h^2} - \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{f'''(z_2) - f'''(z_1)}{6}h\right| \\
&\leq \left|\frac{(f^{FP}(x+h) - f(x+h)) - 2(f^{FP}(x) - f(x)) + (f^{FP}(x-h) - f(x-h))}{h^2}\right| + \left|\frac{f'''(z_2) - f'''(z_1)}{6}h\right| \\
&\leq \left|\frac{\delta_{x+h} - 2\delta_x + \delta_{x-h}}{h^2}\right| + 2Ch \leq \frac{4c\epsilon_m}{h^2} + 2Ch.
\end{aligned}$$

Week 3

1. Banded Matrices

Problem 2.2 (B) Given $\mathbf{x} \in \mathbb{R}^n$, find a lower triangular matrix of the form

$$L = I - 2\mathbf{v}\mathbf{e}_1^\top$$

such that:

$$L\mathbf{x} = x_1\mathbf{e}_1.$$

What does $L\mathbf{y}$ equal if $\mathbf{y} \in \mathbb{R}^n$ satisfies $y_1 = \mathbf{e}_1^\top \mathbf{y} = 0$?

SOLUTION

By straightforward computation we find

$$Lx = x - 2\mathbf{v}\mathbf{e}_1^\top x = x - 2\mathbf{v}x_1$$

and thus we find such a lower triangular L by choosing $v_1 = 0$ and $v_k = \frac{x_k}{2x_1}$ for $k = 2..n$ and $x_1 \neq 0$.

2. Orthogonal Matrices

Problem 5.1 (C) Show that orthogonal matrices preserve the 2-norm of vectors:

$$\|Q\mathbf{v}\| = \|\mathbf{v}\|.$$

SOLUTION

$$\|Q\mathbf{v}\|^2 = (Q\mathbf{v})^\top Q\mathbf{v} = \mathbf{v}^\top Q^\top Q\mathbf{v} = \mathbf{v}^\top \mathbf{v} = \|\mathbf{v}\|^2$$

Problem 5.2 (B) Show that the eigenvalues λ of an orthogonal matrix Q are on the unit circle: $|\lambda| = 1$.

SOLUTION Let $\mathbf{v}$ be a unit eigenvector corresponding to λ: $Q\mathbf{v} = \lambda\mathbf{v}$ with $\|\mathbf{v}\| = 1$. Then

$$1 = \|\mathbf{v}\| = \|Q\mathbf{v}\| = \|\lambda\mathbf{v}\| = |\lambda|.$$

Problem 5.3 (A) Explain why an orthogonal matrix Q must be equal to I if all its eigenvalues are 1.

SOLUTION

Note that Q is normal ($Q^\top Q = I$) and therefore by the spectral theorem for normal matrices we have

$$Q = \tilde{Q}\Lambda\tilde{Q}^\star = \tilde{Q}\tilde{Q}^\star = I$$

since Q is unitary.

Problem 5.4 (B) Complete the implementation of a type representing reflections that supports `Q[k,j]` and such that `*` takes $O(n)$ operations.