Numerical Analysis MATH50003 (2023–24) Revision Sheet

Problem 1(a) State which real number is represented by an IEEE 16-bit floating point number (with $\sigma = 15, Q = 5$, and $S = 10$) with bits

```
1 01000 0000000001
```

SOLUTION The sign bit is 1 so the answer is negative. The exponent bits correspond to

$$q = 2^3 = 8$$

The significand is

$$(1.0000000001)_2 = 1 + 2^{-10}$$

So this represents

$$-2^{8-\sigma}(1 + 2^{-10}) = -2^{-7}(1 + 2^{-10})$$

END

Problem 1(b) How are the following real numbers rounded to the nearest F_{16}?

$$1/2, 1/2 + 2^{-12}, 3 + 2^{-9} + 2^{-10}, 3 + 2^{-10} + 2^{-11}.$$

SOLUTION 1/2 is already a float. We have

$$1/2 + 2^{-12} = (0.100000000001)_2 = 2^{-1}(1.00000000001)_2$$

This is exactly at the midpoint so is rounded down so the last bit is 0, that is, it is rounded to 1/2. Next we have

$$3 + 2^{-9} + 2^{-10} = (11.0000000011)_2 = 2(1.10000000011)_2.$$

This time we are are exactly at the midpoint but we round up so the last bit is 0 giving us

$$2(1.100000001)_2 = 3 + 2^{-8}.$$

Finally,

$$3 + 2^{-10} + 2^{-11} = 2(1.100000000011)_2$$

This we round up since we are above the midpoint giving us

$$2(1.1000000001)_2 = 3 + 2^{-9}.$$

END

Problem 2(a) Consider a Lower triangular matrix with floating point entries:

$$L = \begin{bmatrix} \ell_{11} & & & \\ \ell_{21} & \ell_{22} & & \\ \vdots & \ddots & \ddots & \\ \ell_{n1} & \cdots & \ell_{n,n-1} & \ell_{nn} \end{bmatrix} \in F_{\sigma,Q,S}^{n\times n}$$

and a vector $\boldsymbol{x} \in F_{\sigma,Q,S}^n$, where $F_{\sigma,Q,S}$ is a set of floating-point numbers. Denoting matrix-vector multiplication implemented using floating point arithmetic as

$$\boldsymbol{b} := \texttt{lowermul}(L, \boldsymbol{x})$$

express the entries $b_k := \mathbf{e}_k^\top \boldsymbol{b}$ in terms of ℓ_{kj} and $x_k := \mathbf{e}_k^\top \boldsymbol{x}$, using rounded floating-point operations $\oplus$ and $\otimes$.

SOLUTION

$$b_k = \bigoplus_{j=1}^{k} (\ell_{kj} \otimes x_j)$$

END

Problem 2(b) Assuming all operations involve normal floating numbers, show that your approximation has the form

$$L\boldsymbol{x} = \texttt{lowermul}(L, \boldsymbol{x}) + \boldsymbol{\epsilon}$$

where, for ϵ_{m} denoting machine epsilon and $E_{n,\epsilon} := \frac{n\epsilon}{1-n\epsilon}$ and assuming $n\epsilon_{\mathrm{m}} < 2$,

$$\|\boldsymbol{\epsilon}\|_1 \leq 2E_{n,\epsilon_{\mathrm{m}}/2} \|L\|_1 \|\boldsymbol{x}\|_1.$$

Here we use the matrix norm $\|A\|_1 := \max_j \sum_{k=1}^n |a_{kj}|$ and the vector norm $\|\boldsymbol{x}\|_1 := \sum_{k=1}^n |x_k|$. You may use the fact that

$$x_1 \oplus \cdots \oplus x_n = x_1 + \cdots + x_n + \sigma_n$$

where

$$|\sigma_n| \leq \|\boldsymbol{x}\|_1 E_{n-1,\epsilon_{\mathrm{m}}/2}.$$

SOLUTION

We have

$$b_k = (\bigoplus_{j=1}^{k} \ell_{kj} \otimes x_j) = (\bigoplus_{j=1}^{k} \ell_{kj} x_j (1+\delta_j)) = (\sum_{j=1}^{k} \ell_{kj} x_j (1+\delta_j)) + \sigma_k$$

where

$$|\sigma_k| \leq M_k E_{k-1,\epsilon_{\mathrm{m}}/2}$$

for

$$M_k := \sum_{j=1}^{k} |\ell_{kj}||x_j||1+\delta_j| \leq 2 \sum_{j=1}^{k} |\ell_{kj}||x_j|.$$

Thus

$$b_k = \boldsymbol{e}_k^\top L\boldsymbol{x} + \underbrace{\sum_{j=1}^{k} \ell_{kj} x_j \delta_j + \sigma_k}_{\varepsilon_k}.$$

where

$$|\varepsilon_k| \leq \sum_{j=1}^{k} |\ell_{kj}||x_j|(|\delta_j| + 2E_{k-1,\epsilon_{\mathrm{m}}/2}) \leq 2E_{k,\epsilon_{\mathrm{m}}/2} \sum_{j=1}^{k} |\ell_{kj}||x_j|$$

where we use

$$\begin{aligned}(|\delta_j| + 2E_{k-1,\epsilon_{\mathrm{m}}/2}) &\leq \frac{\epsilon_{\mathrm{m}}}{2} + 2\frac{(k-1)\epsilon_{\mathrm{m}}/2}{1-(k-1)\epsilon_{\mathrm{m}}/2} \\ &= \frac{\epsilon_{\mathrm{m}}/2 - (k-1)\epsilon_{\mathrm{m}}^2/4 + 2(k-1)\epsilon_{\mathrm{m}}/2}{1-(k-1)\epsilon_{\mathrm{m}}/2} \\ &\leq \frac{2k\epsilon_{\mathrm{m}}/2}{1-k\epsilon_{\mathrm{m}}/2} = 2E_{k,\epsilon_{\mathrm{m}}/2}.\end{aligned}$$

We then have using $E_{k,\epsilon_{\rm m}/2} \leq E_{n,\epsilon_{\rm m}/2}$,

$$\begin{aligned}
\|\boldsymbol{\epsilon}\|_1 &= \sum_{k=1}^n |\varepsilon_k| \leq 2E_{n,\epsilon_{\rm m}/2} \sum_{k=1}^n \sum_{j=1}^k |\ell_{kj}||x_j| \\
&= 2E_{n,\epsilon_{\rm m}/2} \sum_{j=1}^n |x_j| \sum_{k=1}^{n-j+1} |\ell_{kj}| \leq 2E_{n,\epsilon_{\rm m}/2} \sum_{j=1}^n |x_j| \|L\|_1 \\
&= 2E_{n,\epsilon_{\rm m}/2} \|L\|_1 \|\boldsymbol{x}\|_1.
\end{aligned}$$

END

Problem 3 What is the dual extension of square-roots? I.e. what should $\sqrt{a+b\epsilon}$ equal assuming $a > 0$?

SOLUTION

$$\sqrt{a+b\epsilon} = \sqrt{a} + \frac{b}{2\sqrt{a}}\epsilon$$

END

Problem 4 Use the Cholesky factorisation to determine whether the following matrix is symmetric positive definite:

$$\begin{bmatrix} 2 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

SOLUTION

Here $\alpha_1 = 2$ and $\boldsymbol{v} = [2, 1]$ giving us

$$\begin{aligned}
A_2 &= \begin{bmatrix} 3 & 2 \\ 2 & 2 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 \\ 1 & 3/2 \end{bmatrix}
\end{aligned}$$

Thus $\alpha_2 = 1$ and $\boldsymbol{v} = [1]$ giving us

$$A_3 = [3/2 - 1] = [1/2]$$

As $\alpha_3 = 1/2 > 0$ we know a Cholesky decomposition exists hence A is SPD. In particular we have computed $A = LL^\top$ where

$$L = \begin{bmatrix} \sqrt{2} & & \\ \sqrt{2} & 1 & \\ 1/\sqrt{2} & 1 & 1/\sqrt{2} \end{bmatrix}$$

END

Problem 5 Use reflections to determine the entries of an orthogonal matrix Q such that

$$Q \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix}.$$

SOLUTION

$$
\begin{aligned}
\boldsymbol{x} &:= [2,1,2], \|\boldsymbol{x}\| = 3\\
\boldsymbol{y} &:= \|\boldsymbol{x}\|\boldsymbol{e}_1 + \boldsymbol{x} = [5,1,2], \|\boldsymbol{y}\| = \sqrt{30}\\
\boldsymbol{w} &:= \boldsymbol{y}/\|\boldsymbol{y}\| = [5,1,2]/\sqrt{30}\\
Q &:= I - 2\boldsymbol{w}\boldsymbol{w}^\top = I - \frac{1}{15}\begin{bmatrix}5\\1\\2\end{bmatrix}[5\ 1\ 2] = I - \frac{1}{15}\begin{bmatrix}25 & 5 & 10\\ 5 & 1 & 2\\ 10 & 2 & 4\end{bmatrix}\\
&= \frac{1}{15}\begin{bmatrix}-10 & -5 & -10\\ -5 & 14 & -2\\ -10 & -2 & 11\end{bmatrix}
\end{aligned}
$$

END

Problem 6 For the function $f(\theta) = \sin 3\theta$, state explicit formulae for its Fourier coefficients

$$\hat{f}_k := \frac{1}{2\pi}\int_0^{2\pi} f(\theta)\mathrm{e}^{-\mathrm{i}k\theta}\mathrm{d}\theta$$

and their discrete approximation:

$$\hat{f}_k^n := \frac{1}{n}\sum_{j=0}^{n-1} f(\theta_j)\mathrm{e}^{-\mathrm{i}k\theta_j}.$$

for *all* integers k, $n = 1, 2, \ldots$, where $\theta_j = 2\pi j/n$.

SOLUTION

We have

$$f(\theta) = \sin 3\theta = \frac{\exp(3i\theta)}{2i} - \frac{\exp(-3i\theta)}{2i}$$

hence $\hat{f}_3 = 1/(2i)$, $\hat{f}_{-3} = -1/(2i)$ and $\hat{f}_k = 0$ otherwise. Thus we have:

$$
\begin{aligned}
&\hat{f}_k^1 = \sum_{k=-\infty}^{\infty} \hat{f}_k = \hat{f}_{-3} + \hat{f}_3 = 0,\\
&\hat{f}_{2k}^2 = 0, \hat{f}_{2k+1}^2 = \hat{f}_{-3} + \hat{f}_3 = 0,\\
&\hat{f}_{3k}^3 = \hat{f}_{-3} + \hat{f}_3 = 0, \hat{f}_{3k+1}^3 = \hat{f}_{3k-1}^3 = 0,\\
&\hat{f}_{4k}^4 = \hat{f}_{4k+2}^4 = 0, \hat{f}_{4k+1}^4 = \hat{f}_{-3} = -1/(2i), \hat{f}_{4k+3}^4 = \hat{f}_3 = 1/(2i)\\
&\hat{f}_{5k}^5 = \hat{f}_{5k+1}^5 = \hat{f}_{5k+4}^5, \hat{f}_{5k+2}^5 = \hat{f}_{-3} = -1/(2i), \hat{f}_{5k+3}^5 = \hat{f}_3 = 1/(2i),\\
&\hat{f}_{6k}^6 = \hat{f}_{6k+1}^6 = \hat{f}_{6k+2}^6 = \hat{f}_{6k+4}^6 = \hat{f}_{6k+5}^6, \hat{f}_{6k+3}^5 = \hat{f}_{-3} + \hat{f}_3 = 0
\end{aligned}
$$

For $n > 6$ we have

$$\hat{f}_{-3+nk}^n = \hat{f}_{-3} = -\frac{1}{2i}, \hat{f}_{3+nk}^n = \hat{f}_3 = \frac{1}{2i}$$

and all other $\hat{f}_k^n = 0$.

END

Problem 7 Consider orthogonal polynomials

$$H_n(x) = 2^n x^n + O(x^{n-1})$$

as $x \to \infty$ and $n = 0, 1, 2, \ldots$, orthogonal with respect to the inner product

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x)g(x)w(x)\mathrm{d}x, \qquad w(x) = \exp(-x^2)$$

Construct $H_0(x)$, $H_1(x)$, $H_2(x)$ and hence show that $H_3(x) = 8x^3 - 12x$. You may use without proof the formulae

$$\int_{-\infty}^{\infty} w(x)\mathrm{d}x = \sqrt{\pi}, \int_{-\infty}^{\infty} x^2 w(x)\mathrm{d}x = \sqrt{\pi}/2, \int_{-\infty}^{\infty} x^4 w(x)\mathrm{d}x = 3\sqrt{\pi}/4.$$

SOLUTION

Because $w(x) = w(-x)$ we know that a_k is zero. We further know that $H_0(x) = 1$ with $\|H_0\|^2 = \sqrt{\pi}$ and $H_1(x) = 2x$ with

$$\|H_1\|^2 = 4\int_{-\infty}^{\infty} x^2 w(x)\mathrm{d}x = 2\sqrt{\pi}.$$

We have

$$xH_1(x) = c_0 H_0(x) + H_2(x)/2$$

where

$$c_0 = \frac{\langle xH_1(x), H_0(x) \rangle}{\|H_0\|^2} = \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1$$

Hence $H_2(x) = 2xH_1(x) - H_0(x) = 4x^2 - 2$, which satisfies

$$\|H_2\|^2 = 16\int_{-\infty}^{\infty} x^4 w(x)\mathrm{d}x - 16\int_{-\infty}^{\infty} x^2 w(x)\mathrm{d}x + 4\int_{-\infty}^{\infty} w(x)\mathrm{d}x = (12 - 8 + 4)\sqrt{\pi} = 8\sqrt{\pi}.$$

We further have

$$\langle xH_2(x), H_1(x) \rangle = \int_{-\infty}^{\infty} (8x^4 - 4x^2)w(x)\mathrm{d}x = (6 - 2)\sqrt{\pi} = 4\sqrt{\pi}$$

Finally we have

$$xH_2(x) = c_1 H_1(x) + H_3(x)/2$$

where

$$c_1 = \frac{\langle xH_2(x), H_1(x) \rangle}{\|H_1\|^2} = \frac{4\sqrt{\pi}}{2\sqrt{\pi}} = 2$$

Hence

$$H_3(x) = 2xH_2(x) - 4H_1(x) = 8x^3 - 12x.$$

END

Problem 8(a) Derive the 3-point Gauss quadrature formula

$$\int_{-\infty}^{\infty} f(x)\exp(-x^2)\mathrm{d}x \approx w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3)$$

with analytic expressions for x_j and w_j.

SOLUTION

We know x_k are the roots of $H_3(x) = 8x^3 - 12x$ hence we have $x_2 = 0$ and the other roots satisfy

$$2x^2 - 3 = 0,$$

i.e., $x_1 = -\sqrt{3/2}$ and $x_2 = \sqrt{3/2}$. To deduce the weights the easiest approach is to use Lagrange interpolation. An alternative is to orthonormalise. Note the Jacobi matrix satisfies

$$x[H_0|H_1|H_2|H_3|\dots] = [H_0|H_1|H_2|H_3|\dots] \underbrace{\begin{bmatrix} 0 & 1 & & & \\ 1/2 & 0 & 2 & & \\ & 1/2 & 0 & \ddots & \\ & & 1/2 & \ddots & \\ & & & \ddots & \end{bmatrix}}_{X}$$

To find $q_k = d_k H_k$, orthonormalised versions of Hermite, we need to choose d_k to symmetrise X, that is for $D = \mathrm{diag}(d_0, d_1, \dots)$ we have

$$x[q_0|q_1|\dots] = x[H_0|H_1|\dots]D = [H_0|H_1|\dots]XD = [q_0|q_1|\dots]D^{-1}XD$$

where

$$D^{-1}XD = \begin{bmatrix} 0 & d_1/d_0 & & & \\ d_0/(2d_1) & 0 & 2d_2/d_1 & & \\ & d_1/(2d_2) & 0 & \ddots & \\ & & d_2/(2d_3) & \ddots & \\ & & & \ddots & \end{bmatrix}$$

Note $d_0 = 1/\sqrt{\int_{-\infty}^{\infty} \exp(-x^2)\mathrm{d}x} = 1/\pi^{1/4}$ then we have

$$d_0^2 = 2d_1^2 \Rightarrow d_1 = 1/(\sqrt{2}\pi^{1/4})$$
$$d_1^2 = 4d_2^2 \Rightarrow d_2 = 1/(2\sqrt{2}\pi^{1/4})$$

We thus have

$$\begin{aligned} w_1 &= \frac{1}{q_0(-\sqrt{3/2})^2 + q_1(-\sqrt{3/2})^2 + q_2(-\sqrt{3/2})^2} = \frac{1}{d_0^2 + 4d_1^2(3/2) + d_2^2(6-2)^2} = \frac{\sqrt{\pi}}{6} \\ w_2 &= \frac{1}{q_0(0)^2 + q_1(0)^2 + q_2(0)^2} = \frac{1}{d_0^2 + d_2^2(2)^2} = \frac{2\sqrt{\pi}}{3} \\ w_3 &= w_1 = \frac{\sqrt{\pi}}{6}. \end{aligned}$$

END

Problem 8(b) Compute the 2-point and 3-point Gaussian quadrature rules associated with $w(x) = 1$ on $[-1, 1]$.

SOLUTION

For the weights $w(x) = 1$, the orthogonal polynomials of degree ≤ 3 are the Legendre polynomials,

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= \frac{1}{2}(3x^2 - 1), \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x) \end{aligned}$$

which can be found from, e.g, the Rodriguez formula or by direct construction. We can normalise each to get $q_j(x) = P_j(x)/\|P_j\|$, with $\|P_j\|^2 = \int_{-1}^{1} P_j^2 dx$. This gives,

$$\begin{aligned}
q_0(x) &= \frac{1}{\sqrt{2}}, \\
q_1(x) &= \sqrt{\frac{3}{2}}x, \\
q_2(x) &= \sqrt{\frac{5}{8}}(3x^2 - 1), \\
q_3(x) &= \sqrt{\frac{7}{8}}(5x^3 - 3x).
\end{aligned}$$

For the first part we use the roots of $P_2(x)$ which are $\boldsymbol{x} = \left\{\pm\frac{1}{\sqrt{3}}\right\}$. The weights are,

$$w_j = \frac{1}{\alpha_j^2} = \frac{1}{q_0(x_j)^2 + q_1(x_j)^2} = \frac{1}{\frac{1}{2} + \frac{3}{2}x_j^2},$$

where α_j is the same as in III.6 Lemma 2, so that,

$$w_1 = w_2 = 1,$$

and the Gaussian Quadrature rule is,

$$\Sigma_2^w[f] = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

For the second part, we use the roots of $P_3(x)$ which are $\boldsymbol{x} = \left\{0, \pm\sqrt{\frac{3}{5}}\right\}$. The weights are then,

$$w_j = \frac{1}{\alpha_j^2} = \frac{1}{q_0(x_j)^2 + q_1(x_j)^2 + q_2(x_j)^2} = \frac{1}{\frac{9}{8} - \frac{9}{4}x_j^2 + \frac{45}{8}x_j^4}$$

Giving us,

$$\begin{aligned}
w_1 = w_3 = \frac{1}{\frac{9}{8} - \frac{9}{4}\frac{3}{5} + \frac{45}{8}\frac{9}{25}} &= \frac{5}{9} \\
w_2 &= \frac{8}{9}
\end{aligned}$$

Then the Gaussian Quadrature rule is,

$$\Sigma_3^w[f] = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]$$

END

Problem 9 Solve Problem 4(b) from PS8 using **Lemma 12 (discrete orthogonality)** with $w(x) = 1/\sqrt{1-x^2}$ on $[-1, 1]$. That is, use the connection of $T_n(x)$ with $\cos n\theta$ to show that the Discrete Cosine Transform

$$C_n := \begin{bmatrix} \sqrt{1/n} & & & \\ & \sqrt{2/n} & & \\ & & \ddots & \\ & & & \sqrt{2/n} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ \cos\theta_1 & \cdots & \cos\theta_n \\ \vdots & \ddots & \vdots \\ \cos(n-1)\theta_1 & \cdots & \cos(n-1)\theta_n \end{bmatrix}$$

for $\theta_j = \pi(j-1/2)/n$ is an orthogonal matrix.

SOLUTION

Our goal is to show that $C_n C_n^\top = I$. By Lemma 12 (Discrete Orthogonality) and PS10 Q5, we have,

$$\Sigma_n^w[q_l q_m] = \frac{\pi}{n}\sum_{j=1}^n q_l(x_j)q_m(x_j) = \delta_{lm}.$$

where for the weight $w(x) = \frac{1}{\sqrt{1-x^2}}$ we have the orthonormal polynomials $q_0(x_j) = \frac{1}{\sqrt{\pi}}$, $q_k(x_j) = \sqrt{\frac{2}{\pi}}\cos(k\theta_j)$. Thus we have:

$$\begin{aligned}
\boldsymbol{e}_1^\top C_n C_n^\top \boldsymbol{e}_1 &= \sqrt{1/n}[1,1,\ldots,1]\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}\sqrt{1/n} = \frac{1}{n}\sum_{j=1}^n 1 = 1\\
\boldsymbol{e}_k^\top C_n C_n^\top \boldsymbol{e}_1 &= \boldsymbol{e}_1^\top C_n C_n^\top \boldsymbol{e}_k = \sqrt{1/n}[1,1,\ldots,1]\begin{bmatrix}\cos(k-1)\theta_1\\ \vdots\\ \cos(k-1)\theta_n\end{bmatrix}\sqrt{2/n}\\
&= \frac{1}{n}\pi\sum_{\ell=1}^n q_k(x_\ell)q_0(x_\ell) = 0\\
\boldsymbol{e}_k^\top C_n C_n^\top \boldsymbol{e}_j &= \sqrt{2/n}[\cos(k-1)\theta_1,\ldots,\cos(k-1)\theta_n]\begin{bmatrix}\cos(j-1)\theta_1\\ \vdots\\ \cos(j-1)\theta_n\end{bmatrix}\sqrt{2/n}\\
&= \frac{\pi}{n}\sum_{\ell=1}^n q_k(x_\ell)q_j(x_\ell) = \delta_{kj}.
\end{aligned}$$

END