MATH50003 Problem Sheets and Solutions

19 April, 2022

Week 1

1. Binary representation

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

SOLUTION
Hence we show that

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

END

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.,

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

Thus the bits are 11110110

END

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

END

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 $2^{10-23} = 2^{-13}$

END

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 $\text{fl}(x)$? 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.

END

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`.)

END

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 $y = 2^{53} - 1$

END

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

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

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: 0001111001100110 1.

END

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

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

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

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

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).

END

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

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

For the second part, we do:

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

Write

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

where

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

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

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

where

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

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

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

Then we get

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

and the error is bounded by:

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

This is quite pessimistic but still captures that we are on the order of ϵ_{m}.

END

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

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

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

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

$$\begin{aligned}&\left|\frac{f(x+h) - f(x-h)}{2h} - f'(x)\right| = \\ &= \left|\frac{1 + (x+h) + (x+h)^2 - 1 - (x-h) - (x-h)^2}{2h} - (1+2x)\right| = \\ &= \left|\frac{2h + 4hx}{2h} - 1 - 2x\right| = 0.\end{aligned}$$

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 $|\delta_x|\le c\epsilon_{\mathrm{m}}$.

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

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

where again

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

In floating point arithmetic, the error is

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

$$\le\left|\frac{(f^{FP}(x+h)-f(x+h))-2(f^{FP}(x)-f(x))+(f^{FP}(x-h)-f(x-h))}{h^2}\right|+\left|\frac{f'''(z_2)-f'''(z_1)}{6}h\right|$$

$$\le\left|\frac{\delta_{x+h}-2\delta_x+\delta_{x-h}}{h^2}\right|+2Ch\le\frac{4c\epsilon_m}{h^2}+2Ch.$$

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 $\tilde{Q}$ is unitary.

Problem 5.6 (A) Consider a Householder reflection with $\mathbf{x} = [1, h]$ with $h = 2^{-n}$. What is the floating point error in computing $\mathbf{y} = \mp\|\mathbf{x}\|\mathbf{e}_1 + \mathbf{x}$ for each choice of sign.

SOLUTION

Since $\|\mathbf{x}\| = \sqrt{1+h^2}$, we have $\mathbf{y} = [1 \mp \sqrt{1+h^2}, h]$. We note first that h^{fp} and $(h^2)^{fp}$ are exact due to the choice of h, so we only need to discuss the floating error in computing $1 \mp \sqrt{1+h^2}$.

Numerically, let the length of the significand be S, then

$$1 \oplus h^2 = \begin{cases} 1+h^2 & n \le S/2 \\ 1 & n > S/2 \end{cases} = 1 + h^2 + \delta_1$$

where $|\delta_1| \le \frac{\epsilon_m}{2}$.

+ PLUS +

Since $\sqrt{1 \oplus h^2}^{fp} > 0$, we know that

$$\begin{aligned} 1 \oplus \sqrt{1 \oplus h^2}^{fp} =& (1+\delta_2)(1 + \sqrt{1 \oplus h^2}^{fp}) \\ =& (1+\delta_2)(1 + \sqrt{1+h^2+\delta_1}(1+\delta_3)) \end{aligned}$$

where $|\delta_2|, |\delta_3| \le \frac{\epsilon_m}{2}$. Then

$$\begin{aligned} \frac{1 \oplus \sqrt{1 \oplus h^2}^{fp}}{1+\sqrt{1+h^2}} =& (1+\delta_2)\left(1 + \frac{\sqrt{1+h^2+\delta_1}(1+\delta_3) - \sqrt{1+h^2}}{1+\sqrt{1+h^2}}\right) \\ =& (1+\delta_2)\left(1 + \frac{(1+\delta_3)(\sqrt{1+h^2+\delta_1} - \sqrt{1+h^2}) + \delta_3\sqrt{1+h^2}}{1+\sqrt{1+h^2}}\right) \\ \approx& (1+\delta_2)\left(1 + \frac{\delta_1}{2(1+\sqrt{1+h^2})\sqrt{1+h^2}} + \delta_3\frac{\sqrt{1+h^2}}{1+\sqrt{1+h^2}}\right) \end{aligned}$$

and we can bound the relative error by

$$|\delta_2| + |\delta_1|\frac{1}{2(1+\sqrt{1+h^2})\sqrt{1+h^2}} + |\delta_3|\frac{\sqrt{1+h^2}}{1+\sqrt{1+h^2}} \le |\delta_2| + \frac{|\delta_1|}{4} + \frac{3|\delta_3|}{4} \le \epsilon_m.$$

In conclusion, it's very accurate to compute $1 + \sqrt{1+h^2}$.

− MINUS −

If $n > S/2$, then $1 \ominus \sqrt{1 \oplus h^2}^{fp} = 1 \ominus \sqrt{1}^{fp} = 1 \ominus 1 = 0$ so the relative error is 100%.

If $n \le S/2$ but not too small, $1 \oplus h^2$ is exactly $1 + h^2$ but $\sqrt{1+h^2}^{fp}$ can have rounding error. Expand $\sqrt{1+h^2}$ into Taylor series:

$$\sqrt{1+h^2} = 1 + \frac{1}{2}h^2 - \frac{1}{8}h^4 + \frac{1}{16}h^6 - O(h^8) = 1 + 2^{-2n-1} - 2^{-4n-3} + 2^{-6n-4} - O(2^{-8n})$$

so

$$\sqrt{1+h^2}^{fp} = \begin{cases} 1 & n = S/2 \\ 1 + \frac{1}{2}h^2 & \frac{S-3}{4} \le n < S/2 \\ 1 + \frac{1}{2}h^2 - \frac{1}{8}h^4 & \frac{S-4}{6} \le n < \frac{S-3}{4} \\ \vdots & \vdots \end{cases}$$

where we can conclude that the absolute error is approximately $\frac{1}{2}h^2, \frac{1}{8}h^4, \frac{1}{16}h^6, ...$ for each stage when h is small. Keeping in mind that $1 - \sqrt{1+h^2} \approx -\frac{1}{2}h^2$ when h is small, the relative error is approximately $1, \frac{1}{4}h^2, \frac{1}{8}h^4, ...$ for each stage. Special note: the relative error is exactly 1 in the first stage when $n = S/2$.

If n is so small that $\sqrt{1+h^2}$ is noticably larger than 1, the absolute error can be bounded by $\frac{\epsilon_m}{2}$ so the relative error is bounded by $\frac{\epsilon_m}{2(\sqrt{1+h^2}-1)} \approx \frac{\epsilon_m}{h^2}$.

Week 4

1. Least Squares and QR

Problem 1.2 (B) Show that every matrix has a QR decomposition such that the diagonal of R is non-negative. Make sure to include the case of more columns than rows.

SOLUTION

Beginning with the square case, a square matrix $A = QR$ with square orthogonal Q and square upper triangular R can always be rewritten in the form $A = QD^{-1}DR$ where D is a diagonal matrix with $sign(R[j,j])$ on the diagonal. As a result, DR is an upper triangular matrix with positive diagonal. It remains to check that $QD^{-1} = -QD$ is still orthogonal - this is easy to check since

$$-QD(-QD)^T = QDDQ^T = QQ^T = I.$$

Note we have made use of the fact that the inverse of a diagonal matrix is diagonal, that any diagonal matrix satisfies $D^T = D$ and that $DD = I$ since $\text{sign}(R[j,j])^2 = 1$.

The same argument works for the non-square cases as long as we take care to consider the appropriate dimensions and pad with the identity matrix. Note that Q is always a square matrix in the QR decomposition. Assume the R factor has more columns m than rows n. Then a square $n \times n$ diagonal matrix with its n diagonal entries being $\text{sign}(R[j,j]), j = 1..n$ works with the same argument as above.

Finally, assume that R has less colums m than rows n. In this case the square $n \times n$ diagonal matrix with its first m diagonal entries being $\text{sign}(R[j,j]), j = 1..m$ and any remaining diagonal entries being 1 works with the same argument as above.

Since the matrix D is always square diagonal and orthogonal by construction, everything else for both of these cases is exactly as in the square case.

Problem 1.3 (B) Show that the QR decomposition of a square invertible matrix is unique, provided that the diagonal of R is positive.

SOLUTION

Assume there is a second decomposition also with positive diagonal

$$A = QR = \tilde{Q}\tilde{R}$$

Then we know

$$Q^\top \tilde{Q} = R\tilde{R}^{-1}$$

Note $Q^\top \tilde{Q}$ is orthogonal, and $R\tilde{R}^{-1}$ has positive eigenvalues (the diagonal), hence all m eigenvalues of $Q^\top \tilde{Q}$ are 1. This means that $Q^\top \tilde{Q} = I$ and hence $\tilde{Q} = Q$, which then implies $\tilde{R} = R$.

2. Gram–Schmidt

Problem 2.1 (B) The modified Gram–Schmidt algorithm is a slight variation of Gram–Schmidt where instead of computing

$$\mathbf{v}_j := \mathbf{a}_j - \sum_{k=1}^{j-1} \underbrace{\mathbf{q}_k^\top \mathbf{a}_j}_{r_{kj}} \mathbf{q}_k$$

we compute it step-by-step:

$$\mathbf{v}_j^1 := \mathbf{a}_j$$
$$\mathbf{v}_j^{k+1} := \mathbf{v}_j^k - \mathbf{q}_k^\top \mathbf{v}_j^k \mathbf{q}_k$$

Show that $\mathbf{v}_j^j = \mathbf{v}_j$.

SOLUTION

Recall that the column vectors of Q are orthonormal i.e. $(\mathbf{q}_i, \mathbf{q}_j) = \delta_{ij}$. Next observe that substituting from $\mathbf{v}_j^1$ for $\mathbf{a}_j$, for each $\mathbf{v}_j^i$ we get i terms (note that the second term in the definition of $\mathbf{v}_j^{k+1}$ contributes only 1 term by orthonormality of Q, retrieving the contribution from $\mathbf{v}_j^{i-1}$ plus one.)

$$\mathbf{v}_j^i := \mathbf{a}_j - \sum_{k=1}^{i-2} q_k^\top \mathbf{a}_j q_k - q_{i-1}^\top \mathbf{a}_j q_{i-1}$$

4. Banded QR with Given's rotations

Problem 4.1 **(A)** Describe an algorithm for computing the QR decomposition of a tridiagonal matrix using rotations instead of reflections to upper-triangularise column-by-column.

SOLUTION

Let A be a tridiagonal matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & \vdots \\ 0 & a_{32} & a_{33} & \ddots & 0 \\ & 0 & \ddots & \ddots & a_{n-1,n} \\ & & 0 & a_{n,n-1} & a_{nn} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix},$$

where each $\mathbf{a}_j \in \mathbb{R}^n$ and $[\mathbf{a}_j]_k = 0$ for $|j-k| > 1$.

Recall that,

$$\frac{1}{\sqrt{a^2+b^2}} \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sqrt{a^2+b^2} \\ 0 \end{bmatrix},$$

and that,

$$\frac{1}{\sqrt{a^2+b^2}} \begin{bmatrix} a & b \\ -b & a \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

is a rotation matrix where $\theta = -\arctan(b/a)$. With this in mind, consider multiplying A from the left by,

$$Q_1 = \begin{bmatrix} \frac{a_{11}}{r_{11}} & \frac{a_{21}}{r_{11}} & & & \\ -\frac{a_{21}}{r_{11}} & \frac{a_{11}}{r_{11}} & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix},$$

where $r_{11} = \sqrt{a_{11}^2 + a_{21}^2}$. This rotates dimensions 1 and 2 through angle $\theta = -\arctan(a_{21}/a_{11})$. We have,

$$Q_1\mathbf{a}_1 = \begin{bmatrix} r_{11} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Q_1\mathbf{a}_2 = \begin{bmatrix} r_{12} := \frac{1}{r_{11}}(a_{11}a_{12} + a_{21}a_{22}) \\ t_1 := \frac{1}{r_{11}}(a_{11}a_{22} - a_{21}a_{12}) \\ a_{32} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Q_1\mathbf{a}_3 = \begin{bmatrix} r_{13} := \frac{1}{r_{11}}a_{21}a_{23} \\ s_1 := \frac{1}{r_{11}}a_{11}a_{23} \\ a_{33} \\ a_{43} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Q_1\mathbf{a}_k = \mathbf{a}_k \text{ for } k > 3.$$

Then we take,

$$Q_2 = \begin{bmatrix} 1 & & & & & \\ & \frac{t_1}{r_{22}} & \frac{a_{32}}{r_{22}} & & & \\ & -\frac{a_{32}}{r_{22}} & \frac{t_1}{r_{22}} & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix},$$

where $r_{22} = \sqrt{t_1^2 + a_{32}^2}$, a matrix which rotates dimensions 2 and 3 through angle $\theta_2 = -\arctan(a_{32}/t_1)$. Then,

$$Q_2Q_1\mathbf{a}_1 = Q_1\mathbf{a}_1 = \begin{bmatrix} r_{11} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Q_2Q_1\mathbf{a}_2 = Q_2 \begin{bmatrix} r_{12} \\ t_1 \\ a_{32} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} r_{12} \\ r_{22} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Q_2Q_1\mathbf{a}_3 = Q_2 \begin{bmatrix} r_{13} \\ s_1 \\ a_{33} \\ a_{43} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} r_{13} \\ r_{23} := \frac{1}{r_{22}}(t_1 s_1 + a_{32}a_{33}) \\ t_2 := \frac{1}{r_{22}}(t_1 a_{33} - a_{32}s_1) \\ a_{43} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$Q_2Q_1\mathbf{a}_4 = Q_2 \begin{bmatrix} 0 \\ 0 \\ a_{34} \\ a_{44} \\ a_{54} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ r_{24} := \frac{1}{r_{22}}a_{32}a_{34} \\ s_2 := \frac{1}{r_{22}}t_1 a_{34} \\ a_{44} \\ a_{54} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Q_2Q_1\mathbf{a}_k = \mathbf{a}_k \text{ for } k > 4.$$

Now, for $j = 3 \to n-1$ we take,

$$Q_j := \begin{bmatrix} \mathbf{I}_{j-1} & & & \\ & \frac{t_{j-1}}{r_{jj}} & \frac{a_{j+1,j}}{r_{jj}} & \\ & \frac{a_{j+1,j}}{r_{jj}} & \frac{t_{j-1}}{r_{jj}} & \\ & & & \mathbf{I}_{n-j-1} \end{bmatrix},$$

where $r_{jj} := \sqrt{t_{j-1}^2 + a_{j+1,j}^2}$.

This gives,

$$Q_j \dots Q_1\mathbf{a}_k = \begin{bmatrix} \mathbf{0}_{k-3} \\ r_{k-2,k} \\ r_{k-1,k} \\ r_{k,k} \\ \mathbf{0}_{n-k} \end{bmatrix},$$

for $k \leq j$, and,

$$Q_j \dots Q_1\mathbf{a}_{j+1} = \begin{bmatrix} \mathbf{0} \\ r_{j-1,j+1} \\ r_{j,j+1} := \frac{1}{r_{jj}}(t_{j-1}s_{j-1} + a_{j+1,j}a_{j+1,j+1}) \\ t_j := \frac{1}{r_{jj}}(t_{j-1}a_{j+1,j+1} - s_{j-1}a_{j+1,j}) \\ a_{j+2,j+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Q_j \dots Q_1\mathbf{a}_{j+2} = \begin{bmatrix} \mathbf{0} \\ r_{j,j+2} := \frac{1}{r_{jj}}a_{j+1,j}a_{j+1,j+2} \\ s_j := \frac{1}{r_{jj}}t_{j-1}a_{j+1,j+2} \\ a_{j+2,j+2} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Finally we define, $r_{nn} = \frac{1}{r_{n-1,n-1}}(t_{n-2}a_{n,n} - a_{n,n-1}s_{n-2})$, to obtain,

$$Q_{n-1} \dots Q_1 A = \begin{bmatrix} r_{11} & r_{12} & r_{13} & 0 & \dots & 0 \\ 0 & r_{22} & r_{23} & r_{24} & & 0 \\ & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & 0 & r_{n-2,n-2} & r_{n-2,n-1} & r_{n-2,n} \\ & & & 0 & r_{n-1,n-1} & r_{n-1,n} \\ & & & & 0 & r_{n,n} \end{bmatrix} =: R$$

so that $A = QR$, for $Q = Q_1^{-1} \dots Q_{n-1}^{-1}$, where each matrix Q_j rotates the coordinates $(j, j+1)$ through the angle $\theta_j = -\arctan(a_{j+1,j}/t_{j-1})$, and thus each matrix Q_j^{-1} rotates the coordinates $(j, j+1)$ through the angle $\arctan(a_{j+1,j}/t_{j-1})$.

5. PLU decomposition

Problem 5.1 (C) Compute the PLU decompositions for the following matrices:

$$\begin{bmatrix} 0 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{bmatrix}, \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 4 & -2 & 1 \\ -3 & -5 & 6 & 1 \\ -1 & 2 & 8 & -2 \end{bmatrix}$$

SOLUTION Part 1 Compute the PLU decomposition of,

$$\begin{bmatrix} 0 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{bmatrix}$$

We begin by permuting the first and second row as $|2| > |0|$, hence,

$$P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad P_1 A = \begin{bmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 4 \end{bmatrix}$$

We then choose,

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix}, \qquad L_1 P_1 A = \begin{bmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & -2 & \frac{7}{2} \end{bmatrix}$$

There is no need to permute at this stage, so $P_2 = I_3$, the 3−dimensional identity. Then we can choose,

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \qquad L_2 P_2 L_1 P_1 A = \begin{bmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 9/2 \end{bmatrix} =: U$$

Since $P_2 = I_3$, this reduces to $L_2 L_1 P_1 A = U \Rightarrow A = P_1^{-1} L_1^{-1} L_2^{-1} U$. Since P_1 simply permutes two rows, it is its own inverse, and $L_1^{-1} L_2^{-1}$ is simply,

$$L := L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & -1 & 1 \end{bmatrix}$$

Hence, we have $A = PLU$, where,

$$P = P_1^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & -1 & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 9/2 \end{bmatrix}$$

Part 2

Find the PLU decomposition of,

$$A = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 4 & -2 & 1 \\ -3 & -5 & 6 & 1 \\ -1 & 2 & 8 & -2 \end{bmatrix}$$

We see that we must start with the permutation,

$$P_1 := \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad P_1 A = \begin{bmatrix} -3 & -5 & 6 & 1 \\ 2 & 4 & -2 & 1 \\ 1 & 2 & -1 & 0 \\ -1 & 2 & 8 & -2 \end{bmatrix}$$

We then choose L_1,

$$L_1 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{2}{3} & 1 & 0 & 0 \\ \frac{1}{3} & 0 & 1 & 0 \\ -\frac{1}{3} & 0 & 0 & 1 \end{bmatrix}, \qquad L_1 P_1 A = \begin{bmatrix} -3 & -5 & 6 & 1 \\ 0 & \frac{2}{3} & 2 & \frac{5}{3} \\ 0 & \frac{1}{3} & 1 & \frac{1}{3} \\ 0 & \frac{11}{3} & 6 & -\frac{7}{3} \end{bmatrix}$$

We then choose P_2,

$$P_2 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad P_2 L_1 P_1 A = \begin{bmatrix} -3 & -5 & 6 & 1 \\ 0 & \frac{11}{3} & 6 & -\frac{7}{3} \\ 0 & \frac{1}{3} & 1 & \frac{1}{3} \\ 0 & \frac{2}{3} & 2 & \frac{5}{3} \end{bmatrix}$$

We then choose L_2,

$$L_2 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\frac{1}{11} & 1 & 0 \\ 0 & -\frac{2}{11} & 0 & 1 \end{bmatrix}, \qquad L_2 P_2 L_1 P_1 A = \begin{bmatrix} -3 & -5 & 6 & 1 \\ 0 & \frac{11}{3} & 6 & -\frac{7}{3} \\ 0 & 0 & \frac{5}{11} & \frac{6}{11} \\ 0 & 0 & \frac{10}{11} & \frac{23}{11} \end{bmatrix}$$

We then choose P_3 to swap the third and fourth rows,

$$P_3 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad P_3 L_2 P_2 L_1 P_1 A = \begin{bmatrix} -3 & -5 & 6 & 1 \\ 0 & \frac{11}{3} & 6 & -\frac{7}{3} \\ 0 & 0 & \frac{10}{11} & \frac{23}{11} \\ 0 & 0 & \frac{5}{11} & \frac{6}{11} \end{bmatrix}$$

We complete the triangularisation by choosing L_3,

$$L_3 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{2} & 1 \end{bmatrix}, \qquad L_3 P_3 L_2 P_2 L_1 P_1 A = \begin{bmatrix} -3 & -5 & 6 & 1 \\ 0 & \frac{11}{3} & 6 & -\frac{7}{3} \\ 0 & 0 & \frac{10}{11} & \frac{23}{11} \\ 0 & 0 & 0 & -\frac{1}{2} \end{bmatrix}$$

We now consider,

$$L_3 P_3 L_2 P_2 L_1 P_1 = L_3 \tilde{L_2} \tilde{L_1} P_3 P_2 P_1,$$

where $\tilde{L_1}$ and $\tilde{L_1}$ satisfy,

$$P_3 P_2 L_1 = \tilde{L_1} P_3 P_2$$
$$P_3 L_2 = \tilde{L_2} P_3$$

These can be computed via,

$$\tilde{L_1} = P_3 P_2 L_1 P_2^{-1} P_3^{-1}$$
$$\tilde{L_2} = P_3 L_2 P_3^{-1}$$

to obtain,

$$\tilde{L_1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & 1 & 0 & 0 \\ \frac{2}{3} & 0 & 1 & 0 \\ \frac{1}{3} & 0 & 0 & 1 \end{bmatrix}$$

$$\tilde{L_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\frac{2}{11} & 1 & 0 \\ 0 & -\frac{1}{11} & 0 & 1 \end{bmatrix}$$

This gives us,

$$L^{-1} = L_3\tilde{L}_2\tilde{L}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & 1 & 0 & 0 \\ \frac{2}{3} & -\frac{2}{11} & 1 & 0 \\ \frac{1}{3} & -\frac{1}{11} & -\frac{1}{2} & 1 \end{bmatrix},$$

from which it is clear that,

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & 1 & 0 & 0 \\ -\frac{2}{3} & \frac{2}{11} & 1 & 0 \\ -\frac{1}{3} & \frac{1}{11} & \frac{1}{2} & 1 \end{bmatrix}$$

Finally, we have that,

$$P = P_1^{-1}P_2^{-1}P_3^{-1} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

so that $A = PLU$, with,

$$P = P_1^{-1}P_2^{-1}P_3^{-1} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & 1 & 0 & 0 \\ -\frac{2}{3} & \frac{2}{11} & 1 & 0 \\ -\frac{1}{3} & \frac{1}{11} & \frac{1}{2} & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} -3 & -5 & 6 & 1 \\ 0 & \frac{11}{3} & 6 & -\frac{7}{3} \\ 0 & 0 & \frac{10}{11} & \frac{23}{11} \\ 0 & 0 & 0 & -\frac{1}{2} \end{bmatrix}$$

Week 5

1. Positive definite matrices and Cholesky decompositions

Problem 1.1 (C) Use the Cholesky decomposition to determine which of the following matrices are symmetric positive definite:

$$\begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}, \begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 5 \end{bmatrix}, \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 1 & 2 & 2 & 4 \end{bmatrix}$$

SOLUTION

A matrix is symmetric positive definite (SPD) if and only if it has a Cholesky decomposition, so the task here is really just to compute Cholesky decompositions (by hand). Since our goal is to tell if the Cholesky decompositions exist, we do not have to compute L_k's. We only need to see if the decomposition process can keep to the end.

Matrix 1

$$A_0 = \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}$$

$A_1 = 3 - \frac{(-1)\times(-1)}{1} > 0$, so Matrix 1 is SPD.

Matrix 2

$$A_0 = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix} \begin{bmatrix} 2 & 2 \end{bmatrix} = \begin{bmatrix} -3 & -2 \\ -2 & -3 \end{bmatrix}$$

$A_1[1,1] < 0$, so Matrix 2 is not SPD.

Matrix 3

$$A_0 = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 5 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 8 & 4 \\ 4 & 13 \end{bmatrix}$$

$3A_2 = 13 - \frac{4\times 4}{8} > 0$, so Matrix 3 is SPD.

Matrix 4

$$A_0 = \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 1 & 2 & 2 & 4 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix} - \frac{1}{4} \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 2 & 1 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 12 & 4 & 6 \\ 4 & 12 & 6 \\ 6 & 6 & 15 \end{bmatrix}$$

$$4A_2 = \begin{bmatrix} 12 & 6 \\ 6 & 15 \end{bmatrix} - \frac{1}{12} \begin{bmatrix} 4 \\ 6 \end{bmatrix} \begin{bmatrix} 4 & 6 \end{bmatrix} = \frac{4}{3} \begin{bmatrix} 8 & 3 \\ 3 & 9 \end{bmatrix}$$

$3A_3 = 9 - \frac{3\times 3}{8} > 0$, so Matrix 4 is SPD.

Problem 1.2 (B) Recall that an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ on $\mathbb{R}^n$ over the reals $\mathbb{R}$ satisfies, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}$ and $a, b \in \mathbb{R}$: 1. Symmetry: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ 2. Linearity: $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle$ 3. Posive-definite: $\langle \mathbf{x}, \mathbf{x} \rangle > 0, x \neq 0$

Prove that $\langle \mathbf{x}, \mathbf{y} \rangle$ is an inner product if and only if

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top K \mathbf{y}$$

where K is a symmetric positive definite matrix.

SOLUTION

We begin by showing that $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top K \mathbf{y}$ with K spd defines an inner product. To do this we simply verify the three properties: For symmetry, we find

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top K \mathbf{y} = \mathbf{x} \cdot (K\mathbf{y}) = (K\mathbf{y}) \cdot \mathbf{x}$$

$$= (K\mathbf{y})^\top \mathbf{x} = \mathbf{y}^\top K^\top \mathbf{x} = \mathbf{y}^\top K \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle.$$

For linearity:

$$\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = (a\mathbf{x} + b\mathbf{y})^\top K \mathbf{z} = (a\mathbf{x}^\top + b\mathbf{y}^\top) K \mathbf{z}$$

$$= a\mathbf{x}^\top K \mathbf{z} + b\mathbf{y}^\top K \mathbf{z} = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle.$$

Positive-definiteness of the matrix K immediately yields $\langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^\top K \mathbf{x} > 0$. Now we turn to the converse result, i.e. that there exists a symmetric positive definite matrix K for any inner product , such that it

can be written as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top K \mathbf{y}$. Define the entries of K by $K_{ij} = \langle e_i, e_j \rangle$ where e_j is the j-th standard basis vector. Note that by linearity of the inner product any inner product on $\mathbb{R}^n$ can be written as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=0}^n \sum_{l=0}^n x_k y_l \langle e_k, e_l \rangle$ by linearity. But with the elements of K defined as above this is precisely

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=0}^{n} \sum_{l=0}^{n} x_k K_{kl} y_l = \mathbf{x}^\top K \mathbf{y}.$$

What remains is to show that this K is symmetric positive definite. Symmetry is an immediate conseqence of the symmetry of its elements, i.e. $K_{ij} = \langle e_i, e_j \rangle = \langle e_j, e_i \rangle = K_{ji}$. Finally, positive definiteness follows from the positive definiteness of the inner product $\langle \mathbf{x}, \mathbf{x} \rangle > 0$ with $\langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^\top K \mathbf{x}$.

END

Problem 1.3 (A) Show that a matrix is symmetric positive definite if and only if it has a Cholesky decomposition of the form

$$A = UU^\top$$

where U is upper triangular with positive entries on the diagonal.

SOLUTION

We didn't discuss this but note that because a symmetric positive definite matrix has stricly positive eigenvalues: for a normalised eigenvector we have

$$\lambda = \lambda \mathbf{v}^\top \mathbf{v} = \mathbf{v}^\top K \mathbf{v} > 0.$$

Thus they are always invertible. Then note that any such matrix has a Cholesky decomposition of standard form $A = LL^\top$ where L is lower triangular. The inverse of this standard form Cholesky decomposition is then $A^{-1} = L^{-T} L^{-1}$, which is of the desired form since L is lower triangular and L^{-T} is upper triangular. The positive entries on the diagonal follow directly because this is the case for the Cholesky decomposition factors of the original matrix. Thus, since all symmetric positive definite matrices can be written as the inverses of a symmetric positive definite matrix, this shows that they all have a decomposition $A = UU^\top$ (using the Cholesky factors of its inverse).

Alternatively, we can replicate the procedure of computing the Cholesky decomposition beginning in the bottom right instead of the top left. Write:

$$A = \begin{bmatrix} K & \mathbf{v} \\ \mathbf{v}^\top & \alpha \end{bmatrix} = \underbrace{\begin{bmatrix} I & \frac{\mathbf{v}}{\sqrt{\alpha}} \\ & \sqrt{\alpha} \end{bmatrix}}_{U_1} \begin{bmatrix} K - \frac{\mathbf{v}\mathbf{v}^\top}{\alpha} & \\ & 1 \end{bmatrix} \underbrace{\begin{bmatrix} I & \\ \frac{\mathbf{v}^\top}{\sqrt{\alpha}} & \sqrt{\alpha} \end{bmatrix}}_{U_1^\top}$$

The induction proceeds as in the lower triangular case.

END

Problem 1.4 (A) Prove that the following $n \times n$ matrix is symmetric positive definite for any n:

$$\Delta_n := \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}$$

Deduce its two Cholesky decompositions: $\Delta_n = L_n L_n^\top = U_n U_n^\top$ where L_n is lower triangular and U_n is upper triangular.

SOLUTION

We first prove that L_n is lower bidiagonal by induction. Let A be a tridiagonal SPD matrix:

$$A = \left[\begin{array}{c|ccc} \alpha & \beta & 0 & \cdots \\ \hline \beta & & & \\ 0 & & K & \\ \vdots & & & \end{array}\right]$$

where K is again tridiagonal. Denote the Cholesky decomposition of A by $A = LL^\top$. Recalling the proof of the *Theorem (Cholesky & SPD)* from the lecture, we can write

$$L = \left[\begin{array}{c|c} \sqrt{\alpha} & 0 \\ \hline \frac{\beta}{\sqrt{\alpha}} & \\ 0 & \tilde{L} \\ \vdots & \end{array}\right]$$

where $\tilde{L}$ satisfies $\tilde{L}\tilde{L}^\top = K - \begin{bmatrix} \beta^2/\alpha & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ which is a tridiagonal matrix smaller than A. Since L is lower bidiagonal when A is 1×1, we know by induction that L is lower bidiagonal for A of any size.

Once we know that L_n is lower bidiagonal, we can write it as

$$L_n = \begin{bmatrix} a_1 & & & \\ b_1 & \ddots & & \\ & \ddots & \ddots & \\ & & b_{n-1} & a_n \end{bmatrix}$$

which we substitute into $\Delta_n = L_n L_n^\top$ to get

$$\begin{cases} a_1 = \sqrt{2} \\ a_k b_k = -1 \\ b_k^2 + a_{k+1}^2 = 2 \end{cases}$$

for $k = 1, \dots, n-1$.

Now we solve the recurrence. Substituting the second equation into the third one:

$$\frac{1}{a_k^2} + a_{k+1}^2 = 2.$$

Let $c_k = a_k^2$:

$$\frac{1}{c_k} + c_{k+1} = 2.$$

Consider the fixing point of this recurrence: $\frac{1}{x} + x = 2 \Longrightarrow (x-1)^2 = 0$ has the double root $x = 1$ which hints us to consider $\frac{1}{c_k - 1}$. In fact, the recurrence is equivalent to

$$\frac{1}{c_{k+1} - 1} = \frac{1}{c_k - 1} + 1.$$

Recalling that $c_1 = 2$, we know that $\frac{1}{c_k - 1} = k$. As a result, $c_k = (k+1)/k$, $a_k = \sqrt{(k+1)/k}$ and $b_k = -\sqrt{k/(k+1)}$, hence we know L_n.

We can apply the same process to U_n, but this is a special case since flipping Δ_n horizontally and vertically gives itself: $P\Delta_n P^\top = \Delta_n$ where

$$P = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}$$

is the permutation that reverses a vector. So we can also flip L_n to get U_n:

$$U_n = PL_nP$$

so that $U_nU_n^\top = PL_nPPL_n^\top P = P\Delta_nP = \Delta_n$.

Alternatively one can use the procedure from Problem 1.3. That is, write:

$$\Delta_n = \begin{bmatrix} \Delta_{n-1} & -\mathbf{e}_n \\ -\mathbf{e}_n^\top & 2 \end{bmatrix} = \underbrace{\begin{bmatrix} I & \frac{-\mathbf{e}_n}{\sqrt{2}} \\ & \sqrt{2} \end{bmatrix}}_{U_1} \begin{bmatrix} \Delta_{n-1} - \frac{\mathbf{e}_n\mathbf{e}_n^\top}{2} & \\ & 1 \end{bmatrix} \underbrace{\begin{bmatrix} I & \\ \frac{\mathbf{v}^\top}{\sqrt{\alpha}} & \sqrt{\alpha} \end{bmatrix}}_{U_1^\top}$$

2. Matrix norms

Problem 2.1 (B) Prove the following:

$$\|A\|_\infty = \max_k \|A[k,:]\|_1$$
$$\|A\|_{1\to\infty} = \|\mathrm{vec}(A)\|_\infty = \max_{kj} |a_{kj}|$$

SOLUTION

Step 1. upper bounds

$$\|A\mathbf{x}\|_\infty = \max_k \left|\sum_j a_{kj}x_j\right| \le \max_k \sum_j |a_{kj}x_j| \le \begin{cases} \max_j |x_j| \max_k \sum_j |a_{kj}| = \|\mathbf{x}\|_\infty \max_k \|A[k,:]\|_1 \\ \max_{kj} |a_{kj}| \sum_j |x_j| = \|\mathbf{x}\|_1 \|\mathrm{vec}(A)\|_\infty \end{cases}$$

Step 2.1. meeting the upper bound ($\|A\|_{1\to\infty}$)

Let a_{lm} be the entry of A with maximum absolute value. Let $\mathbf{x} = \mathbf{e}_m$, then

$$\|A\mathbf{x}\|_\infty = \max_k \left|\sum_j a_{kj}x_j\right| = \max_k |a_{km}| = |a_{lm}|$$

and

$$\|\mathbf{x}\|_1 \|\mathrm{vec}(A)\|_\infty = 1 \cdot |a_{lm}|.$$

Step 2.2. meeting the upper bound ($\|A\|_\infty$)

Let $A[n,:]$ be the row of A with maximum 1-norm. Let $\mathbf{x} = (\mathrm{sign}.(A[n,:]))^\top$, then $\left|\sum_j a_{kj}x_j\right| \begin{cases} = \sum_j |a_{kj}| = \|A[k,:]\|_1 & k = n \\ \le \sum_j |a_{kj}| = \|A[k,:]\|_1 & k \ne n \end{cases}$,

so

$$\|A\mathbf{x}\|_\infty = \max_k \left|\sum_j a_{kj}x_j\right| = \max_k \|A[k,:]\|_1$$

while

$$\|\mathbf{x}\|_\infty \max_k \|A[k,:]\|_1 = 1 \cdot \max_k \|A[k,:]\|_1.$$

Conclusion

In both cases, equality can hold, so the upper bounds are actually maxima.

END

Problem 2.2 (B) For a rank-1 matrix $A = \mathbf{x}\mathbf{y}^\top$ prove that

$$\|A\|_2 = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Hint: use the Cauchy–Schwartz inequality.

SOLUTION

$$\|A\mathbf{z}\|_2 = \|\mathbf{x}\mathbf{y}^\top \mathbf{z}\|_2 = |\mathbf{y}^\top \mathbf{z}| \|\mathbf{x}\|_2,$$

so it remains to prove that $\|\mathbf{y}\|_2 = \sup_{\mathbf{z}} \frac{|\mathbf{y}^\top \mathbf{z}|}{\|\mathbf{z}\|_2}$.

By Cauchy-Schwartz inequality,

$$|\mathbf{y}^\top \mathbf{z}| = |(\mathbf{y}, \mathbf{z})| \leq \|\mathbf{y}\|_2 \|\mathbf{z}\|_2$$

with the two sides being equal when $\mathbf{y}$ and $\mathbf{z}$ are linearly dependent, in which case the bound is tight.

END

Problem 2.3 (B) Show for any orthogonal matrix $Q \in \mathbb{R}^m$ and matrix $A \in \mathbb{R}^{m \times n}$ that

$$\|QA\|_F = \|A\|_F$$

by first showing that $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$ using the *trace* of an $m \times m$ matrix:

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{mm}.$$

SOLUTION

$$\mathrm{tr}(A^\top A) = \sum_k (A^\top A)[k,k] = \sum_k \sum_j A^\top[k,j] A[j,k] = \sum_k \sum_j A[j,k]^2 = \|A\|_F^2.$$

On the other hand,

$$\mathrm{tr}(A^\top A) = \mathrm{tr}(A^\top Q^\top Q A) = \mathrm{tr}((QA)^\top (QA)) = \|QA\|_F^2,$$

so $\|QA\|_F = \|A\|_F$.

END

3. Singular value decomposition

Problem 3.1 (B) Show that $\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\|A\|_2$ where r is the rank of A.

SOLUTION

From Problem 2.3 use the fact that $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$, where $A \in \mathbb{R}^{m \times n}$.

Hence,

$$\|A\|_F^2 = \mathrm{tr}(A^\top A) = \sigma_1^2 + ... + \sigma_m^2$$

where $\sigma_1 \geq ... \geq \sigma_n \geq 0$ are the singular values of A and σ_i^2 are the eigenvalues of $A^\top A$

Knowing that $\|A\|_2^2 = \sigma_1^2$ we have $\|A\|_2^2 \leq \|A\|_F^2$

Moreover, since if the rank of A is r we have that $\sigma_{r+1} = ... = \sigma_m = 0$ and we also know $\sigma_1 \geq ... \geq \sigma_n \geq 0$, we have that

$\|A\|_F^2 = \sigma_1^2 + ... + \sigma_m^2 = \sigma_1^2 + ... + \sigma_r^2 \leq r\sigma_1^2 = r\|A\|_2^2$

Hence,

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\|A\|_2.$$

END

Problem 3.3 (B) For $A \in \mathbb{R}^{m\times n}$ define the *pseudo-inverse*:

$$A^+ := V\Sigma^{-1}U^\top.$$

Show that it satisfies the *Moore-Penrose conditions*: 1. $AA^+A = A$ 2. $A^+AA^+ = A^+$ 3. $(AA^+)^\top = AA^+$ and $(A^+A)^\top = A^+A$

SOLUTION

Let $A = U\Sigma V^\top$ and $A^+ := V\Sigma^{-1}U^\top$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$. Note that $U^\top U = I_m$ and $V^\top V = I_r$.

1. We have
$$AA^+A = U\Sigma V^\top V\Sigma^{-1}U^\top U\Sigma V^\top = U\Sigma\Sigma^{-1}\Sigma V^\top = U\Sigma V^\top = A$$
2. Moreover,
$$A^+AA^+ = V\Sigma^{-1}U^\top U\Sigma V^\top V\Sigma^{-1}U^\top = V\Sigma^{-1}\Sigma\Sigma^{-1}U^\top = V\Sigma^{-1}U^\top = A^+$$
3.
$$(AA^+)^\top = (A^+)^\top A^\top = U\Sigma^{-1}V^\top V\Sigma U^\top = UU^\top = U\Sigma V^\top V\Sigma^{-1}U^\top = AA^+$$
$$(A^+A)^\top = A^\top(A^+)^\top = V\Sigma U^\top U\Sigma^{-1}V^\top = VV^\top = V\Sigma^{-1}U^\top U\Sigma V^\top = A^+A$$

END

Problem 3.4 (A) Show for $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ and rank that $\mathbf{x} = A^+\mathbf{b}$ is the least squares solution, i.e., minimises $\|A\mathbf{x} - \mathbf{b}\|_2$. Hint: extend U in the SVD to be a square orthogonal matrix.

SOLUTION

The proof mimics that of the QR decomposition. Write $A = U\Sigma V^\top$ and let

$$\tilde{U} = \begin{bmatrix} U & K \end{bmatrix}$$

so that $\tilde{U}$ is orthogonal. We use the fact orthogonal matrices do not change norms:

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|U\Sigma V^\top\mathbf{x} - \mathbf{b}\|_2^2 = \|\tilde{U}^\top U\Sigma V^\top\mathbf{x} - \tilde{U}^\top\mathbf{b}\|_2^2 = \| \underbrace{\begin{bmatrix} I_m \\ O \end{bmatrix}}_{\in\mathbb{R}^{m\times n}} \Sigma V^\top\mathbf{x} - \begin{bmatrix} U^\top \\ K^\top \end{bmatrix}\mathbf{b}\|_2^2$$
$$= \|\Sigma V^\top\mathbf{x} - U^\top\mathbf{b}\|_2^2 + \|K^\top\mathbf{b}\|^2$$

The second term is independent of $\mathbf{x}$. The first term is minimised when zero:

$$\|\Sigma V^\top\mathbf{x} - U^\top\mathbf{b}\|_2 = \|\Sigma V^\top V\Sigma^{-1}U^\top\mathbf{b} - U^\top\mathbf{b}\|_2 = 0$$

END

Problem 3.5 (A) If $A \in \mathbb{R}^{m\times n}$ has a non-empty kernel there are multiple solutions to the least squares problem as we can add any element of the kernel. Show that $\mathbf{x} = A^+\mathbf{b}$ gives the least squares solution such that $\|\mathbf{x}\|_2$ is minimised.

SOLUTION

Let $\mathbf{x} = A^+b$ and let $\mathbf{x} + \mathbf{k}$ to be another solution i.e.

$$\|A\mathbf{x} - b\| = \|A(\mathbf{x} + \mathbf{k}) - b\|$$

Following the previous part we deduce:

$$\Sigma V^\top(\mathbf{x}+\mathbf{k}) = U^\top \mathbf{b} \Rightarrow V^\top \mathbf{k} = 0$$

As $\mathbf{x} = V\mathbf{c}$ lies in the span of the columns of V we have $\mathbf{x}^\top \mathbf{k} = 0$. Thus

$$\|\mathbf{x}+\mathbf{k}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{k}\|^2$$

which is minimised when $\mathbf{k} = 0$.

END

Week 6

1. Condition numbers

Problem 1.1 (B) Prove that, if $|\epsilon_i| \leq \epsilon$ and $n\epsilon < 1$, then

$$\prod_{k=1}^{n}(1+\epsilon_i) = 1+\theta_n$$

for some constant θ_n satisfying $|\theta_n| \leq \frac{n\epsilon}{1-n\epsilon}$.

SOLUTION

$$\prod_{k=1}^{n}(1+\epsilon_i) \leq (1+\epsilon)^n = \sum_{k=0}^{n}\binom{n}{k}\epsilon^k \leq 1+\sum_{k=1}^{n} n^k\epsilon^k \leq 1+\sum_{k=1}^{\infty} n^k\epsilon^k = 1+\frac{n\epsilon}{1-n\epsilon}.$$

$$\prod_{k=1}^{n}(1+\epsilon_i) \geq (1-\epsilon)^n = \sum_{k=0}^{n}\binom{n}{k}(-\epsilon)^k \geq 1-\sum_{k=1}^{n} n^k\epsilon^k \geq 1-\sum_{k=1}^{\infty} n^k\epsilon^k = 1-\frac{n\epsilon}{1-n\epsilon}.$$

Problem 1.2 (B) Let $A, B \in \mathbb{R}^{m\times n}$. Prove that if the columns satisfy $\|\mathbf{a}_j\|_2 \leq \|\mathbf{b}_j\|_2$ then $\|A\|_F \leq \|B\|_F$ and $\|A\|_2 \leq \sqrt{\text{rank}(B)}\|B\|_2$.

SOLUTION

Recalling from *Problem Sheet 5 - Problem 2.3* - SOLUTION*, we know that

$$\|A\|_F = \sqrt{\sum_{k,j} A[k,j]^2} = \sqrt{\sum_j \|\mathbf{a}_j\|_2^2} \qquad \text{and} \qquad \|B\|_F = \sqrt{\sum_j \|\mathbf{b}_j\|_2^2}.$$

Since $\|\mathbf{a}_j\|_2 \leq \|\mathbf{b}_j\|_2$, we have $\|A\|_F \leq \|B\|_F$.

Recalling from _Problem Sheet 5 - Problem 3.1*_, we have

$$\|A\|_2 \leq \|A\|_F \leq \|B\|_F \leq \sqrt{\text{rank}(B)}\|B\|_2.$$

Problem 1.3 (C) Compute the 1-norm, 2-norm, and ∞-norm condition numbers for the following matrices:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \begin{bmatrix} 1/3 & 1/5 \\ 0 & 1/7 \end{bmatrix}, \begin{bmatrix} 1 & & & \\ & 1/2 & & \\ & & \ddots & \\ & & & 1/2^n \end{bmatrix}$$

(Hint: recall that the singular values of a matrix A are the square roots of the eigenvalues of the Gram matrix $A^\top A$.)

SOLUTION

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad A^{-1} = -\frac{1}{2}\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 1/3 & 1/5 \\ 0 & 1/7 \end{bmatrix}, \qquad B^{-1} = 21\begin{bmatrix} 1/7 & -1/5 \\ 0 & 1/3 \end{bmatrix}$$

$\|A\|_1 = 6$, $\|A^{-1}\|_1 = 7/2$, so $\kappa_1(A) = 21$.

$\|A\|_\infty = 7$, $\|A^{-1}\|_\infty = 3$, so $\kappa_\infty(A) = 21$.

$\|B\|_1 = 12/35$, $\|B^{-1}\|_1 = 21 \times 8/15 = 56/5$, so $\kappa_1(B) = 96/25$.

$\|B\|_\infty = 8/15$, $\|B^{-1}\|_\infty = 21 \times 12/35$, so $\kappa_\infty(B) = 96/25$

Finally, for the 2-norms: $\kappa_2(A)$: For $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, we have that the singular values are the $\sigma_1 = \sqrt{\lambda_1}, \sigma_2 = \sqrt{\lambda_2}$, where λ_1 and λ_2 are the eigenvalues of $A^T A$.

$$A^T A = \begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix},$$

so an eigenvalue λ of $A^T A$ must satisfy,

$$\begin{aligned} (10-\lambda)(20-\lambda) - 196 &= 0 \\ \Leftrightarrow \lambda &= 15 \pm \sqrt{221}. \end{aligned}$$

The larger eigenvalue corresponds to σ_1, so $\sigma_1 = \sqrt{15+\sqrt{221}}$, and the smaller corresponds to σ_2, so $\sigma_2 = \sqrt{15-\sqrt{221}}$. Finally, we have $\|A\|_2 = \sigma_1, \|A^{-1}\|_2 = 1/\sigma_2$, and so $\kappa_2(A) = \sqrt{\frac{15+\sqrt{221}}{15-\sqrt{221}}}$.

$\kappa_2(B)$: For

$$B = \begin{bmatrix} 1/3 & 1/5 \\ 0 & 1/7 \end{bmatrix},$$

we have that the singular values are the $\sigma_1 = \sqrt{\lambda_1}, \sigma_2 = \sqrt{\lambda_2}$, where λ_1 and λ_2 are the eigenvalues of $A^T A$.

$$A^T A = \begin{bmatrix} 1/9 & 1/15 \\ 1/15 & \frac{74}{5^2 7^2} \end{bmatrix}.$$

An eigenvalue λ must satisfy:

$$\begin{aligned} (1/9-\lambda)\left(\frac{74}{5^2 7^2} - \lambda\right) - \frac{1}{225} &= 0 \\ \Leftrightarrow \lambda &= \frac{1891 \pm 29\sqrt{2941}}{22050}. \end{aligned}$$

With the same logic as above, we can then deduce that

$$\|B\|_2 = \sqrt{\frac{1891 + 29\sqrt{2941}}{22050}}$$

and

$$\|B^{-1}\|_2 \sqrt{\frac{22050}{1891 - 29\sqrt{2941}}}$$

so that,

$$\kappa_2(B) = \sqrt{\frac{1891 + 29\sqrt{2941}}{1891 - 29\sqrt{2941}}}$$

For,

$$A_n = \begin{bmatrix} 1 & & & \\ & 1/2 & & \\ & & \ddots & \\ & & & 1/2^n \end{bmatrix}, \quad A_n^{-1} = \begin{bmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & 2^n \end{bmatrix}$$

It is clear that

$$\|A_n\|_1 = \|A_n\|_\infty = 1,$$

and

$$\|A_n^{-1}\|_1 = \|A_n^{-1}\|_\infty = 2^n,$$

so $\kappa_1(A_n) = \kappa_\infty(A) = 2^n$. Moreover, we can clearly see the singular values $\sigma_1 = 1, \sigma_2 = 1/2, \ldots, \sigma_{n+1} = 1/2^n$. So $\|A_n\|_2 = 1, \|A_n^{-1}\|_2 = 2^n$, $\kappa_2(A_n) = 2^n$

Problem 1.4 (B) State a bound on the relative error on $A\mathbf{v}$ for $\|\mathbf{v}\|_2 = 1$ for the following matrices:

$$\begin{bmatrix} 1/3 & 1/5 \\ 0 & 1/10^3 \end{bmatrix}, \begin{bmatrix} 1 & & & \\ & 1/2 & & \\ & & \cdots & \\ & & & 1/2^{10} \end{bmatrix}$$

Compute the relative error in computing $A\mathbf{v}$ (using `big` for a high-precision version to compare against) where $\mathbf{v}$ is the last column of V in the SVD $A = U\Sigma V^\top$, computed using the `svd` command with `Float64` inputs. How does the error compare to the predicted error bound?

SOLUTION

The Theorem (relative-error for matrix vector) tells us that,

$$\frac{\|\delta A\mathbf{x}\|}{\|A\mathbf{x}\|} \leq \kappa(A)\epsilon,$$

if the relative pertubation error $\|\delta A\| = \|A\|\epsilon$. For the 2−norm, we have,

$$\|\delta A\|_2 \leq \underbrace{\frac{\sqrt{\min(m,n)}n\epsilon_m}{2 - n\epsilon_m}}_{\epsilon} \|A\|_2.$$

The condition number of the first matrix is 453.33 (see code below to compute that), and ϵ defined above is $\frac{2\sqrt{2}\epsilon_m}{2-2\epsilon_m} = 3.14 \cdot 10^{-16}$, so the bound on the relative error is:

$$1.42 \cdot 10^{-13}.$$

The condition number of the second matrix is 2^{10} by the question above, and ϵ defined above is $\frac{10\sqrt{10}\epsilon_m}{2-10\epsilon_m} = 7.02 \cdot 10^{-16}$, the bound on the relative error in this case is then:

$$7.19 \cdot 10^{-13}$$

3. Euler Methods

Problem 3.2 (B) For an evenly spaced grid $t_1, ..., t_n$, use the approximation

$$\frac{u'(t_{k+1}) + u'(t_k)}{2} \approx \frac{u_{k+1} - u_k}{h}$$

to recast

$$u(0) = c$$
$$u'(t) = a(t)u(t) + f(t)$$

as a lower bidiagonal linear system. Use forward-substitution to extend this to vector linear problems:

$$\mathbf{u}(0) = \mathbf{c}$$
$$\mathbf{u}'(t) = A(t)\mathbf{u}(t) + \mathbf{f}(t)$$

SOLUTION

We have,

$$\frac{u_{k+1} - u_k}{h} \approx \frac{u'(t_{k+1}) + u'(t_k)}{2} = \frac{a(t_{k+1})u_{k+1} + a(t_k)u_k}{2} + \frac{1}{2}(f(t_{k+1}) + f(t_k)),$$

so we can write,

$$\left(\frac{1}{h} - \frac{a(t_{k+1})}{2}\right) u_{k+1} + \left(-\frac{1}{h} - \frac{a(t_k)}{2}\right) u_k = \frac{1}{2}(f(t_{k+1}) + f(t_k)).$$

With the initial condition $u(0) = c$, we can write the whole system as,

$$\begin{bmatrix} 1 & & & \\ -\frac{1}{h} - \frac{a(t_1)}{2} & \ddots & \frac{1}{h} - \frac{a(t_2)}{2} & \\ & \ddots & & \ddots \\ & & -\frac{1}{h} - \frac{a(t_{n-1})}{2} & \frac{1}{h} - \frac{a(t_n)}{2} \end{bmatrix} \mathbf{u} = \begin{bmatrix} c \\ \frac{1}{2}(f(t_1) + f(t_2)) \\ \vdots \\ \frac{1}{2}(f(t_{n-1}) + f(t_n)) \end{bmatrix},$$

which is lower bidiagonal.

Now if we wish to use forward substitution in a vector linear problem, we can derive in much the same way as above:

$$\left(\frac{1}{h}I - \frac{A(t_{k+1})}{2}\right) \mathbf{u}_{k+1} + \left(-\frac{1}{h}I - \frac{A(t_k)}{2}\right) \mathbf{u}_k = \frac{1}{2}(\mathbf{f}(t_{k+1}) + \mathbf{f}(t_k)),$$

to make the update equation,

$$\mathbf{u}_{k+1} = \left(I - \frac{h}{2}A(t_{k+1})\right)^{-1} \left(\left(I + \frac{h}{2}A(t_k)\right) \mathbf{u}_k + \frac{h}{2}(\mathbf{f}(t_{k+1}) + \mathbf{f}(t_k))\right),$$

with initial value,

$$\mathbf{u}_1 = \mathbf{c}.$$

Week 7

1. Two-Point Boundary Value Problem

Problem 1.3 (A) Consider Helmholtz with Neumann conditions:

$$u'(0) = c_0$$
$$u'(1) = c_1$$
$$u_{xx} + k^2 u = f(x)$$

Write down the finite difference approximation approximating $u(x_k) \approx u_k$ on an evenly spaced grid $x_k = (k-1)/(n-1)$ for $k = 1, ..., n$ using the first order derivative approximation conditions:

$$u'(0) \approx (u_2 - u_1)/h = c_0$$
$$u'(1) \approx (u_n - u_{n-1})/h = c_1$$

Use pivoting to reduce the equation to one involving a symmetric tridiagonal matrix.

SOLUTION

We have, with $u(x_k) = u_k$ (and using κ instead of k in the equation $u_{xx} + k^2 u = f(x)$ so as to avoid confusion with the indices):

$$\begin{aligned} \frac{u_2 - u_1}{h} &= c_0, \\ \frac{u_{k-1} - 2u_k + u_{k+1}}{h^2} + \kappa^2 u_k &= f(x_k), \qquad \text{for } k = 2 : n-1 \\ \frac{u_n - u_{n-1}}{h} &= c_1, \end{aligned}$$

which we write in matrix form as:

$$\begin{bmatrix} -\frac{1}{h} & \frac{1}{h} & & & \\ \frac{1}{h^2} & \kappa^2 - \frac{2}{h^2} & \frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{h^2} & \kappa^2 - \frac{2}{h^2} & \frac{1}{h^2} \\ & & & -\frac{1}{h} & \frac{1}{h} \end{bmatrix} \mathbf{u} = \begin{bmatrix} c_0 \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ c_1 \end{bmatrix},$$

which we can make symmetric tridiagonal by multiplying the first row by $1/h$ and the final row by $-1/h$:

$$\begin{bmatrix} -\frac{1}{h^2} & \frac{1}{h^2} & & & \\ \frac{1}{h^2} & \kappa^2 - \frac{2}{h^2} & \frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{h^2} & \kappa^2 - \frac{2}{h^2} & \frac{1}{h^2} \\ & & & \frac{1}{h^2} & -\frac{1}{h^2} \end{bmatrix} \mathbf{u} = \begin{bmatrix} \frac{c_0}{h} \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ -\frac{c_1}{h} \end{bmatrix},$$

2. Convergence

Problem 2.1 (B) For the equation

$$u(0) = c_0$$
$$u' + au = f(x)$$

where $a \in \mathbb{R}$ and $0 \leq x \leq 1$, prove convergence as $n \to \infty$ for the method constructed in PS6 using the approximation where we take the average of the two grid points:

$$\frac{u'(x_{k+1}) + u'(x_k)}{2} \approx \frac{u_{k+1} - u_k}{h}.$$

SOLUTION Using the approximation from PS6 we obtain

$$\frac{f(x_{k+1}) + f(x_k)}{2} = \frac{u'(x_{k+1}) + u'(x_k)}{2} + \frac{a(u(x_{k+1}) + u(x_k))}{2} \approx \frac{(u_{k+1} - u_k)}{h} + \frac{au_{k+1}}{2} + \frac{au_k}{2}$$

So we get

$$\left(\frac{a}{2} - \frac{1}{h}\right) u_k + \left(\frac{a}{2} + \frac{1}{h}\right) u_{k+1} = \frac{f(x_{k+1}) + f(x_k)}{2}$$

We want to prove that $\sup_{k=1,\dots,n-1} |u(x_k) - u_k|$ converges to 0 as $n \to \infty$.

Take $\hat{u} = [u_0, ..., u_{n-1}]^T$ and rewrite the system as

$$\hat{L}\hat{u} = \begin{bmatrix} c_0 \\ \hat{f}^f \end{bmatrix}$$

where $f_k = \frac{f(x_k)+f(x_{k-1})}{2}$, $k = 1, ..., n-1$ and

$$\hat{L} = \begin{bmatrix} 1 & & & & \\ \frac{a}{2} - \frac{1}{h} & \frac{a}{2} + \frac{1}{h} & & & \\ & \frac{a}{2} - \frac{1}{h} & \frac{a}{2} + \frac{1}{h} & & \\ & & \ddots & \ddots & \\ & & & \frac{a}{2} - \frac{1}{h} & \frac{a}{2} + \frac{1}{h} \end{bmatrix}$$

Note that $\hat{L}$ is lower bidiagonal.

Now, similarly to Euler's methods convergence theorem, we study consistency and stability.

Consistency: Our discretisation approximates the true equation.

$$\begin{aligned} Lu &= \begin{bmatrix} c_0 \\ \frac{u(x_1)-u(x_0)}{h} + \frac{a}{2}(u(x_1) + u(x_0)) \\ \vdots \\ \frac{u(x_{n-1})-u(x_{n-2})}{h} + \frac{a}{2}(u(x_{n-1}) + u(x_{n-2})) \end{bmatrix} = \begin{bmatrix} c_0 \\ \frac{1}{2}\left(\frac{u(x_1)-u(x_0)}{h} + \frac{u(x_1)-u(x_0)}{h} + a(u(x_1) + u(x_0))\right) \\ \vdots \\ \frac{1}{2}\left(\frac{u(x_{n-1})-u(x_{n-2})}{h} + \frac{u(x_1)-u(x_0)}{h} + a(u(x_{n-1}) + u(x_{n-2}))\right) \end{bmatrix} \\ &= \begin{bmatrix} c_0 \\ \frac{1}{2}\left(u'(x_0) + au(x_0) + u''(\tau_0)h + u'(x_1) + au(x_1) + u''(\sigma_1)h\right) \\ \vdots \\ \frac{1}{2}\left(u'(x_{n-2}) + au(x_{n-2}) + u''(\tau_{n-2})h + u'(x_{n-1}) + au(x_{n-1}) + u''(\sigma_{n-1})h\right) \end{bmatrix} = \begin{bmatrix} c_0 \\ \frac{f(x_0)+f(x_1)}{2} + \frac{u''(\tau_0)+u''(\sigma_1)}{2}h \\ \vdots \\ \frac{f(x_{n-2})+f(x_{n-1})}{2} + \frac{u''(\tau_{n-2})+u''(\sigma_{n-1})}{2}h \end{bmatrix} \\ &= \begin{bmatrix} c_0 \\ \hat{f}^f \end{bmatrix} + \begin{bmatrix} 0 \\ \delta \end{bmatrix} \end{aligned}$$

where $x_k \leq \tau_k, \sigma_k \leq x_{k+1}$, and uniform boundedness implies that $\|\delta\|_\infty = O(h)$

Stability: The inverse does not blow up the error.

$$\hat{L} = \underbrace{\begin{bmatrix} 1 & & & \\ & \left(\frac{a}{2} + \frac{1}{h}\right) & & \\ & & \ddots & \\ & & & \left(\frac{a}{2} + \frac{1}{h}\right) \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} 1 & & & & \\ \left(\frac{a}{2} + \frac{1}{h}\right)^{-1}\left(\frac{a}{2} - \frac{1}{h}\right) & 1 & & & \\ & \ddots & \ddots & & \\ & & \left(\frac{a}{2} + \frac{1}{h}\right)^{-1}\left(\frac{a}{2} - \frac{1}{h}\right) & 1 \end{bmatrix}}_{L}$$

Thus, we have

$$\|L^{-1}\|_{1\to\infty} \leq \left|\left(\frac{a}{2} + \frac{1}{h}\right)^{-1}\left(\frac{a}{2} - \frac{1}{h}\right)\right|^{n-1} = O(1)$$

as $n \to \infty$, where one can take logarithms and use L'Hopitals rule to show that it actually tends to a limit. Note that

$$\left|\frac{a}{2} + \frac{1}{h}\right|^{-1} = \left|\frac{h}{\frac{ah}{2} + 1}\right| \leq 2h$$

for sufficiently small h (or large n). Combining stability and consistency we have, for sufficently small h,

$$\|\mathbf{u}^f - \mathbf{u}\|_\infty = \|\hat{L}^{-1}(\hat{L}\mathbf{u}^f - \hat{L}\mathbf{u})\|_\infty = \|L^{-1}D^{-1}\begin{bmatrix}0\\ \delta\end{bmatrix}\|_\infty \leq 2h\|L^{-1}\|_{1\to\infty}\|\delta\|_1 = O(h).$$

Problem 2.2 **(A)** Consider the matrices

$$L = \begin{bmatrix} 1 & & & & \\ -a_1 & 1 & & & \\ & -a_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & -a_{n-1} & 1 \end{bmatrix}, \qquad T = \begin{bmatrix} 1 & & & & \\ -a & 1 & & & \\ & -a & 1 & & \\ & & \ddots & \ddots & \\ & & & -a & 1 \end{bmatrix}.$$

By writing down the inverse explicitly prove that if $|a_k| \leq a$ then

$$\|L^{-1}\|_{1\to\infty} \leq \ \|T^{-1}\|_{1\to\infty}.$$

Use this to prove convergence as $n \to \infty$ of forward Euler for

$$\begin{aligned} u(0) &= c_0 \\ u'(x) - a(x)u(x) &= f(x) \end{aligned}$$

SOLUTION

Since

$$L^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & ... & 0 \\ a_1 & 1 & 0 & 0 & 0 & ... & 0 \\ a_1a_2 & a_2 & 1 & 0 & 0 & ... & 0 \\ a_1a_2a_3 & a_2a_3 & a_3 & 1 & 0 & ... & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & 1 & 0 \\ \prod_{i=1}^{n-1} a_i & \prod_{i=2}^{n-1} a_i & ... & ... & a_{n-2}a_{n-1} & a_{n-1} & 1 \end{bmatrix}$$

and

$$T^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & ... & 0 \\ a & 1 & 0 & 0 & 0 & ... & 0 \\ a^2 & a & 1 & 0 & 0 & ... & 0 \\ a^3 & a^2 & a & 1 & 0 & ... & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & 1 & 0 \\ a^{n-1} & a^{n-2} & ... & ... & a^2 & a & 1 \end{bmatrix}$$

Then, $\forall x$

$$\|T^{-1}x\|_\infty = \max_i |(T^{-1}x)_i| = \max_i \left| x_i + \sum_{j=1}^{i-1} a^{i-j}x_j \right| = \begin{cases} 1 & if\ a \in [0,1] \\ a^{n-1} & if\ a \geq 1 \end{cases}$$

since, given $b = \max\{1, a\}$,

$$\max_i \left| x_i + \sum_{j=1}^{i-1} a^{i-j}x_j \right| \leq \max_i \left(|x_i| + \sum_{j=1}^{i-1} |a^{i-j}x_j| \right) \leq b^n \sum_{j=1}^{n} |x_j| = b^n \|x\|_1$$

thus,

$\|T^{-1}\|_{1\to\infty} = \sup_{x\neq 0} \frac{\|T^{-1}x\|_\infty}{\|x\|_1} \leq b^n$ and, in particular,

$$\|T^{-1}\|_{1\to\infty} = b^n$$

since

$$\frac{\|T^{-1}x\|_\infty}{\|x\|_1} = b^n$$

it is obtained using

$$x = \begin{cases} e_1 & b = 1 \\ e_n & b = a \end{cases}$$

Moreover, $|a_j| \leq b$, $\forall j = 1, ..., n$, thus,

$$\|L^{-1}x\|_\infty = \max_i |(L^{-1}x)_i| = \max_i \left| x_i + \sum_{j=1}^{i-1} a_j ... a_{i-1} x_j \right| \leq \max_i |x_i| + \sum_{j=1}^{i-1} |a_j ... a_{i-1} x_j| \leq b^n \|x\|_1$$

Hence,

$$\|L^{-1}\|_{1\to\infty} = \sup_x \frac{\|L^{-1}x\|_\infty}{\|x\|_1} \leq b^n = \|T^{-1}\|_{1\to\infty}$$

Now we prove convergence for the forward Euler method as $n \to \infty$ for

$$\begin{aligned} u(0) &= c_0 \\ u'(x) &= a(x)u(x) + f(x) \end{aligned}$$

Using equidistanced (with step h) points $x_0, ..., x_{n-1}$, we use the approximations $u(x_k) \approx u_k$, where $u_0 = c_0$ and

$$u_{k+1} = u_k + h\left(a(x_k)u_k + f(x_k)\right)$$

In order to study convergence we consider the limit as $n \to \infty$ of

$$\sup_{i=1,...,n-1} |u_i - u(x_i)|$$

Similarly to Euler's methods convergence theorem, we study consistency and stability.

In order to apply the theorem we note that we can define $a_k = a(x_k)$, $k = 1, ...n-1$ and we have that for every k, $|a_k| \leq a := max_{i=1,n-1}|a_i|$.

Consistency: Our discretisation approximates the true equation.

$$\hat{L}u = \begin{bmatrix} c_0 \\ \frac{u(x_1)-u(x_0)}{h} - a_1 u(x_0) \\ \vdots \\ \frac{u(x_{n-1})-u(x_{n-2})}{h} - a_{n-1}u(x_{n-2}) \end{bmatrix} = \begin{bmatrix} c_0 \\ u'(x_0) - a_1 u(x_0) + u''(\tau_0)h \\ \vdots \\ u'(x_{n-2}) - a_{n-1}u(x_{n-2}) + u''(\tau_{n-2})h \end{bmatrix} = \begin{bmatrix} c_0 \\ f(x_0) + u''(\tau_0)h \\ \vdots \\ f(x_{n-2}) + u''(\tau_{n-2})h \end{bmatrix} = \begin{bmatrix} c_0 \\ \mathbf{f}^f \end{bmatrix} + \begin{bmatrix} 0 \\ \delta \end{bmatrix}$$

where $x_k \leq \tau_k \leq x_{k+1}$, and uniform boundedness implies that $\|\delta\|_\infty = O(h)$

Stability: The inverse does not blow up the error. First write, for $l_k = 1 - a_k$

$$\hat{L} = \underbrace{\begin{bmatrix} 1 & & & \\ & h^{-1} & & \\ & & \ddots & \\ & & & h^{-1} \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} 1 & & & \\ -l_1 & 1 & & \\ & \ddots & \ddots & \\ & & -l_{n-1} & 1 \end{bmatrix}}_{L}$$

Thus, we have $\|L^{-1}\|_{1\to\infty} \le \|T^{-1}\|_{1\to\infty} = O(1)$

Combining stability and consistency we have

$$\|\mathbf{u}^f - \mathbf{u}\|_\infty = \|\hat{L}^{-1}(\hat{L}\mathbf{u}^f - \hat{L}\mathbf{u})\|_\infty = \|L^{-1}D^{-1}\begin{bmatrix} 0 \\ \delta \end{bmatrix}\|_\infty \le h\|L^{-1}\|_{1\to\infty}\|\delta\|_1 = O(h)$$

3. Fourier series

Problem 3.1 (C) Give explicit formulae for f_k and f_k^n for the following functions:

$$\cos\theta, \cos 4\theta, \sin^4\theta, \frac{3}{3-\mathrm{e}^{\mathrm{i}}}, \frac{1}{1-2\mathrm{e}^{\mathrm{i}}}$$

Hint: You may wish to try the change of variables $z = \mathrm{e}^{-\mathrm{i}\theta}$.

SOLUTION

1. Just expand in complex exponentials to find that

$$\cos\theta = \frac{\exp(\mathrm{i}\theta) + \exp(-\mathrm{i}\theta)}{2}$$

that is $f_1 = f_{-1} = 1/2$, $f_k = 0$ otherwise. Therefore for $p \in \mathbb{Z}$ we have

$$\begin{aligned} f_k^1 &= f_1 + f_{-1} = 1 \\ f_{2p}^2 &= 0, f_{2p+1}^2 = f_1 + f_{-1} = 1 \\ f_{1+np}^n &= f_{-1+np}^n = 1/2, f_k^n = 0\text{otherwise} \end{aligned}$$

2. Similarly

$$\cos 4\theta = \frac{\exp(4\mathrm{i}\theta) + \exp(-4\mathrm{i}\theta)}{2}$$

that is $f_4 = f_{-4} = 1/2$, $f_k = 0$ otherwise. Therefore for $p \in \mathbb{Z}$ we have

$$\begin{aligned} f_p^1 &= f_4 + f_{-4} = 1 \\ f_{2p}^2 &= f_4 + f_{-4} = 1, f_{2p+1}^2 = 0 \\ f_{3p}^3 &= 0, f_{3p\pm1}^3 = f_{\pm4} = 1/2 \\ f_{4p}^4 &= f_{-4} + f_4 = 1, f_{4p\pm1}^4 = 0, f_{4p+2}^4 = 0 \\ f_{5p}^5 &= 0, f_{5p+1}^5 = f_{-4} = 1/2, f_{5p-1}^5 = f_4 = 1/2, f_{5p\pm2}^5 = 0 \\ f_{6p}^6 &= 0, f_{6p\pm1}^6 = 0, f_{6p+2}^6 = f_{-4} = 1/2, f_{6p-2}^6 = f_4 = 1/2, f_{6p+3}^6 = 0 \\ f_{7p}^7 &= 0, f_{7p\pm1}^7 = 0, f_{7p\pm2}^7 = 0, f_{7p\pm3}^7 = f_{\mp4} = 1/2 \\ f_{8p}^8 &= f_{8p\pm1}^8 = f_{8p\pm2}^8 = f_{8p\pm3}^8 = 0, f_{8p+4}^8 = f_4 + f_{-4} = 1 \\ f_{k+pn}^n &= f_k \text{ for } -4 \le\ k \le\ 4, 0 \text{ otherwise.} \end{aligned}$$

3. Here we have:

$$(\sin\theta)^4 = \left(\frac{\exp(\mathrm{i}\theta)-\exp(-\mathrm{i}\theta)}{2\mathrm{i}}\right)^4 = \left(\frac{\exp(2\mathrm{i}\theta)-2+\exp(-2\mathrm{i}\theta)}{-4}\right)^2 = \frac{\exp(4\mathrm{i}\theta)-4\exp(2\mathrm{i}\theta)+6-4\exp(-2\mathrm{i}\theta)+\exp(-2}{16}$$

that is $f_{-4}=f_4=1/16$, $f_{-2}=f_2=-1/4$, $f_0=3/8$, $f_k=0$ otherwise. Therefore for $p\in\mathbb{Z}$ we have

$$\begin{aligned}
f^1_p &= f_{-4}+f_{-2}+f_0+f_2+f_4=0\\
f^2_k &= 0\\
f^3_{3p} &= f_0=3/8, f^3_{3p+1}=f_{-2}+f_4=-3/16, f^3_{3p-1}=f_2+f_{-4}=-3/16\\
f^4_{4p} &= f_0+f_{-4}+f_4=1/2, f^4_{4p\pm1}=0, f^4_{4p+2}=f_2+f_{-2}=-1/2\\
f^5_{5p} &= f_0=3/8, f^5_{5p+1}=f_{-4}=1/16, f^5_{5p-1}=f_4=1/16, f^5_{5p+2}=f_2=-1/4, f^5_{5p-2}=f_{-2}=-1/4\\
f^6_{6p} &= f_0=3/8, f^6_{6p\pm1}=0, f^6_{6p+2}=f_2+f_{-4}=-3/16, f^6_{6p-2}=f_{-2}+f_4=-3/16, f^6_{6p+3}=0\\
f^7_{7p} &= f_0=3/8, f^7_{7p\pm1}=0, f^7_{7p\pm2}=f_{\pm2}=-1/4, f^7_{7p\pm3}=f_{\mp4}=1/16\\
f^8_{8p} &= f_0=3/8, f^8_{8p\pm1}=0, f^8_{8p\pm2}=f_{\pm2}=-1/4, f^8_{8p\pm3}=0, f^8_{8p+4}=f_4+f_{-4}=1/8\\
f^n_{k+pn} &= f_k \text{ for } -4\le\ k\le\ 4, \text{ 0 otherwise.}
\end{aligned}$$

4. Under the change of variables $z=\mathrm{e}^{\mathrm{i}\theta}$ we can use Geoemtric series to determine

$$\frac{3}{3-z}=\frac{1}{1-z/3}=\sum_{k=0}^{\infty}\frac{z^k}{3^k}$$

That is $f_k=1/3^k$ for $k\ge0$, and $f_k=0$ otherwise. We then have for $0\le\ k\le\ n-1$

$$f^n_{k+pn}=\sum_{\ell=0}^{\infty}\frac{1}{3^{k+\ell n}}=\frac{1}{3^k}\frac{1}{1-1/3^n}=\frac{3^n}{3^{n+k}-3^k}$$

5. Now make the change of variables $z=\mathrm{e}^{-\mathrm{i}\theta}$ to get:

$$\frac{1}{1-2/z}=\frac{1}{-2/z}\frac{1}{1-z/2}=\frac{1}{-2/z}\sum_{k=0}^{\infty}\frac{z^k}{2^k}=-\sum_{k=1}^{\infty}\frac{\mathrm{e}^{-\mathrm{i}k\theta}}{2^k}$$

That is $f_k=-1/2^{-k}$ for $k\le\ -1$ and 0 otherwise. We then have for $-n\le k\le\ -1$

$$f^n_{k+pn}=-\sum_{\ell=0}^{\infty}\frac{1}{2^{-k+\ell n}}=-\frac{1}{2^{-k}}\frac{1}{1-1/2^n}=-\frac{2^{n+k}}{2^n-1}$$

Problem 3.2 (B) Prove that if the first $\lambda-1$ derivatives $f(\theta), f'(\theta), ..., f^{(\lambda-1)}(\theta)$ are 2 -periodic and $f^{(\lambda)}$ is uniformly bounded that

$$|f_k|=O(|k|^{-\lambda}) \qquad \text{as } |k|\to\infty$$

Use this to show for the Taylor case ($0=f_{-1}=f_{-2}=\cdots$) that

$$|f(\theta)-\sum_{k=0}^{n-1}f_k\mathrm{e}^{\mathrm{i}k\theta}|=O(n^{1-\lambda})$$

SOLUTION A straightforward application of integration by parts yields the result

$$f_k=\frac{1}{2\pi}\int_0^{2\pi}f(\theta)\mathrm{e}^{-ik\theta}d\theta=\frac{(-i)^\lambda}{2\pi k^\lambda}\int_0^{2\pi}f^{(\lambda)}(\theta)\mathrm{e}^{-ik\theta}d\theta$$

given that $f^{(\lambda)}$ is uniformly bounded, the second part follows directly from this result

$$|\sum_{k=n}^{\infty} f_k \mathrm{e}^{\mathrm{i}k\theta}| \leq \sum_{k=n}^{\infty} |f_k| \leq C \sum_{k=n}^{\infty} k^{-\lambda}$$

for some constant C.

Problem 3.3 (C) If f is a trigonometric polynomial ($f_k = 0$ for $|k| > m$) show for $n \geq 2m+1$ we can exactly recover f:

$$f(\theta) = \sum_{k=-m}^{m} f_k^n \mathrm{e}^{\mathrm{i}k\theta}$$

SOLUTION This proof is nearly identical to the proof of "Theorem (Taylor series converges)" in the lecture notes. Only now one has to also subtract the negative coefficients from the negative approximate coefficients in the chain of arguments.

Problem 3.4 (B) For the general (non-Taylor) case and $n = 2m+1$, prove convergence for

$$f_{-m:m}(\theta) := \sum_{k=-m}^{m} f_k^n \mathrm{e}^{\mathrm{i}k\theta}$$

to $f(\theta)$ as $n \to \infty$. What is the rate of convergence if the first $\lambda - 1$ derivatives $f(\theta), f'(\theta), ..., f^{(\lambda-1)}(\theta)$ are 2 -periodic and $f^{(\lambda)}$ is uniformly bounded?

SOLUTION

Observe that by aliasing (see corollary in lecture notes) and triangle inequality we have the following

$$|f_k^n - f_k| \leq \sum_{p=1}^{\infty} (|f_{k+pn}| + |f_{k-pn}|)$$

Using the result from Problem 3.2 yields

$$|f_k^n - f_k| \leq \frac{C}{n^\lambda} \sum_{p=1}^{\infty} \frac{1}{\left(p + \frac{k}{n}\right)^\lambda} + \frac{1}{\left(p - \frac{k}{n}\right)^\lambda}$$

now we pick $|q| < \frac{1}{2}$ (such that the estimate below will hold for both summands above) and construct an integral with convex and monotonocally decreasing integrand such that

$$(p+q)^{-\lambda} < \int_{p-\frac{1}{2}}^{p+\frac{1}{2}} (x+q)^{-\lambda} dx$$

more over summing over the left-hand side from 1 to ∞ yields a bound by the integral:

$$\int_{\frac{1}{2}}^{\infty} (x+q)^{-\lambda} dx = \frac{1}{\lambda} (\frac{1}{2} + q)^{-\lambda+1}$$

Finally let $q = \pm \frac{k}{n}$ to achieve the rate of convergence

$$|f_k^n - f_k| \leq \frac{C_\lambda}{n^\lambda} \left(\left(\frac{1}{2} + k/n \right)^{-\lambda+1} + \left(\left(\frac{1}{2} - k/n \right) \right)^{-\lambda+1} \right)$$

where C_λ is a constant depending on λ. Note that it is indeed important to split the n coefficients equally over the negative and positive coefficients as stated in the notes, due to the estatime we used above.

Finally, we have (thanks to Anonymous on ed):

$$\begin{aligned}
|f(\theta) - f_{-m:m}(\theta)| &= |\sum_{k=-m}^{m}(f_k - f_k^n)z^k + \sum_{k=m+1}^{\infty} f_k z^k + \sum_{k=-\infty}^{-m-1} f_k z^k| \\
&\leq \sum_{k=-m}^{m}|f_k - f_k^n| + \sum_{k=m+1}^{\infty}|f_k| + \sum_{k=-\infty}^{-m-1}|f_k| \\
&\leq \sum_{k=-m}^{m}\frac{C_\lambda}{n^\lambda}\left(\left(\frac{1}{2}+k/n\right)^{-\lambda+1} + \left(\left(\frac{1}{2}-k/n\right)\right)^{-\lambda+1}\right) + \sum_{k=m+1}^{\infty}|f_k| + \sum_{k=-\infty}^{-m-1}|f_k| \\
&= \frac{C_\lambda}{n^\lambda}2^\lambda + \sum_{k=m+1}^{\infty}|f_k| + \sum_{k=-\infty}^{-m-1}|f_k| \\
&= O(n^{-\lambda}) + O(n^{1-\lambda}) + O(n^{1-\lambda}) \\
&= O(n^{1-\lambda})
\end{aligned}$$

Week 8

1. DFT

Problem 1.1 (C) Show that the DFT Q_n is symmetric ($Q_n = Q_n^\top$) but not Hermitian ($Q_n \neq Q_n^\star$).

SOLUTION

First we remember the definitions we introduced. The DFT is

$$Q_n := \frac{1}{\sqrt{n}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \mathrm{e}^{-\mathrm{i}\theta_1} & \mathrm{e}^{-\mathrm{i}\theta_2} & \cdots & \mathrm{e}^{-\mathrm{i}\theta_{n-1}} \\ 1 & \mathrm{e}^{-\mathrm{i}2\theta_1} & \mathrm{e}^{-\mathrm{i}2\theta_2} & \cdots & \mathrm{e}^{-\mathrm{i}2\theta_{n-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{e}^{-\mathrm{i}(n-1)\theta_1} & \mathrm{e}^{-\mathrm{i}(n-1)\theta_2} & \cdots & \mathrm{e}^{-\mathrm{i}(n-1)\theta_{n-1}} \end{bmatrix}$$

where $\theta_j = 2\pi j/n$ for $j = 0, 1, ..., n$ and $\omega := \mathrm{e}^{\mathrm{i}\theta_1} = \mathrm{e}^{\frac{2\pi\mathrm{i}}{n}}$ are n th roots of unity in the sense that $\omega^n = 1$. So $\mathrm{e}^{\mathrm{i}\theta_j} = \mathrm{e}^{\frac{2\pi\mathrm{i}j}{n}} = \omega^j$. Note that $\theta_j = 2\pi(j-1)/n + 2\pi/n = \theta_{j-1} + \theta_1$. By completing this recurrence we find that $\theta_j = j\theta_1$, from which the following symmetric version follows immediately

$$Q_n = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega^{-1} & \omega^{-2} & \cdots & \omega^{-(n-1)} \\ 1 & \omega^{-2} & \omega^{-4} & \cdots & \omega^{-2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{-(n-1)} & \omega^{-2(n-1)} & \cdots & \omega^{-(n-1)^2} \end{bmatrix}.$$

Now $Q_n^\star$ is found to be

$$Q_n^\star = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \mathrm{e}^{\mathrm{i}\theta_1} & \mathrm{e}^{\mathrm{i}2\theta_1} & \cdots & \mathrm{e}^{\mathrm{i}(n-1)\theta_1} \\ 1 & \mathrm{e}^{\mathrm{i}\theta_2} & \mathrm{e}^{\mathrm{i}2\theta_2} & \cdots & \mathrm{e}^{\mathrm{i}(n-1)\theta_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{e}^{\mathrm{i}\theta_{n-1}} & \mathrm{e}^{\mathrm{i}2\theta_{n-1}} & \cdots & \mathrm{e}^{\mathrm{i}(n-1)\theta_{n-1}} \end{bmatrix} = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega^1 & \omega^2 & \cdots & \omega^{(n-1)} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{(n-1)} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)^2} \end{bmatrix}$$

using the above arguments. Evidently, $Q_n^\star \neq Q_n$ since $\omega \neq \omega^{-1}$.

END

Problem 1.2 **(A)** Show for $0 \leq\ k, \ell \leq\ n-1$

$$\frac{1}{n}\sum_{j=1}^{n}\cos k\theta_j \cos \ell\theta_j = \begin{cases} 1 & k=\ell=0 \\ 1/2 & k=\ell \\ 0 & \text{otherwise} \end{cases}$$

for $\theta_j = \pi(j-1/2)/n$. Hint: Be careful as the θ_j differ from before, and only cover half the period, $[0,\pi]$. Using symmetry may help. You may also consider replacing cos with complex exponentials:

$$\cos\theta = \frac{\mathrm{e}^{\mathrm{i}\theta}+\mathrm{e}^{-\mathrm{i}\theta}}{2}.$$

SOLUTION We have,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\cos(k\theta_j)\cos(l\theta_j) &= \frac{1}{4n}\sum_{j=1}^{n} e^{i(k+l)\theta_j}+e^{-i(k+l)\theta_j}+e^{i(k-l)\theta_j}+e^{-i(k-l)\theta_j} \\ &= \frac{1}{4n}\sum_{j=1}^{n} e^{ia_{kl}\theta_j}+e^{-ia_{kl}\theta_j}+e^{ib_{kl}\theta_j}+e^{-ib_{kl}\theta_j},\end{aligned}$$

where we have defined $a_{kl}=k+l$ and $b_{kl}=k-l$. Now consider, for $a\in\mathbb{Z}$, a $\neq 2kn$ for some $k\in\mathbb{Z}$,

$$\begin{aligned}\sum_{j=1}^{n} e^{ia\theta_j} &= \sum_{j=1}^{n} e^{ia\pi(j-\frac{1}{2})/n} \\ &= e^{-ia\pi/2n}\sum_{j=1}^{n} e^{iaj\pi/n} \\ &= e^{-ia\pi/2n}\sum_{j=1}^{n}(e^{ia\pi/n})^j \\ &= e^{-ia\pi/2n}e^{ia\pi/n}\frac{(e^{ia\pi/n})^n-1}{e^{ia\pi/n}-1} \\ &= e^{ia\pi/2n}\frac{e^{ia\pi}-1}{e^{ia\pi/n}-1},\end{aligned}$$

where our assumptions on a ensure that we are not dividing by 0. Then we have, for a as above,

$$\begin{aligned}\sum_{j=1}^{n} e^{ia\theta_j}+e^{-ia\theta_j} &= e^{ia\pi/2n}\frac{e^{ia\pi}-1}{e^{ia\pi/n}-1}+e^{-ia\pi/2n}\frac{e^{-ia\pi}-1}{e^{-ia\pi/n}-1} \\ &= e^{ia\pi/2n}\frac{e^{ia\pi}-1}{e^{ia\pi/n}-1}+e^{-ia\pi/2n}\cdot\frac{e^{ia\pi/n}}{e^{ia\pi/n}}\cdot\frac{e^{-ia\pi}-1}{e^{-ia\pi/n}-1} \\ &= e^{ia\pi/2n}\frac{e^{ia\pi}-1}{e^{ia\pi/n}-1}+e^{ia\pi/2n}\frac{e^{-ia\pi}-1}{1-e^{ia\pi/n}} \\ &= e^{ia\pi/2n}\frac{e^{ia\pi}-1}{e^{ia\pi/n}-1}-e^{ia\pi/2n}\frac{e^{-ia\pi}-1}{e^{ia\pi/n}-1} \\ &= \frac{e^{ia\pi/2n}}{e^{ia\pi/n-1}}\left(e^{ia\pi}-1-e^{-ia\pi}+1\right) \\ &= \frac{e^{ia\pi/2n}}{e^{ia\pi/n-1}}\frac{1}{2i}\sin(a\pi),\end{aligned}$$

which is 0 for a an integer.

Now, when $k=l=0$, we have $a_{kl}=b_{kl}=0$, and,

$$\frac{1}{n}\sum_{j=1}^{n}\cos(k\theta_j)\cos(l\theta_j) = \frac{1}{4n}\sum_{j=1}^{n}(1+1+1+1)=1.$$

When $k = l \neq 0$, we have $0 < a_{kl} = 2k < 2n$, and $b_{kl} = 0$. Hence,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\cos(k\theta_j)\cos(l\theta_j) &= \frac{1}{4n}\sum_{j=1}^{n}(e^{ia_{kl}\theta_j} + e^{-ia_{kl}\theta_j} + 1 + 1) \\ &= \frac{1}{4n}\left[\left(\sum_{j=1}^{n} e^{ia_{kl}\theta_j} + e^{-ia_{kl}\theta_j}\right) + 2n\right] \\ &= \frac{1}{2},\end{aligned}$$

since a_{kl} meets the conditions for the sum considered above.

When $k \neq l$, we have, $-2n < a_{kl}, b_{kl} < 2n$ and $a_{kl}, b_{kl} \neq 0$, hence,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\cos(k\theta_j)\cos(l\theta_j) &= \frac{1}{4n}\sum_{j=1}^{n}(e^{ia_{kl}\theta_j} + e^{-ia_{kl}\theta_j} + e^{ib_{kl}\theta_j} + e^{-ib_{kl}\theta_j}) \\ &= \frac{1}{4n}\left[\sum_{j=1}^{n}(e^{ia_{kl}\theta_j} + e^{-ia_{kl}\theta_j}) + \sum_{j=1}^{n}(e^{ib_{kl}\theta_j} + e^{-ib_{kl}\theta_j})\right] \\ &= 0.\end{aligned}$$

Problem 1.3 (B) Consider the Discrete Cosine Transform (DCT)

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

for $\theta_j = \pi(j - 1/2)/n$. Prove that C_n is orthogonal: $C_n^\top C_n = C_n C_n^\top = I$. Hint: $C_n C_n^\top = I$ might be easier to show than $C_n^\top C_n = I$ using the previous problem.

SOLUTION

The components of C without the diagonal matrix, which we may call $\hat{C}$ are

$$\hat{C}_{ij} = \cos((j-1)\theta_{i-1}),$$

where $\theta_j = \pi(j - 1/2)/n$. Recalling that the elements of matrix multiplication are given by

$$(ab)_{ij} := \sum_{k=1}^{n} a_{ik}b_{kj}$$

we find that

$$(\hat{C}_n\hat{C}_n^\top)_{ij} = \sum_{k=1}^{n}\cos((i-1)\theta_{k-1})\cos((k-1)\theta_{j-1}).$$

By using the previous problem and the terms on the diagonal matrices which ensure that the 1/2 terms become 1 we know how to compute all of these entries and find that it is the identity.

2. FFT

Problem 2.1 (B) Show that Q_{2n} can also be reduced to Q_n applied to two vectors.

SOLUTION We saw in the lecture notes how this works for $Q_{2n}^\star$:

$$Q_{2n}^\star = \frac{1}{\sqrt{2n}}\left[\mathbf{1}_{2n}|\vec{\omega}_{2n}|\vec{\omega}_{2n}^2|\cdots|\vec{\omega}_{2n}^{2n-1}\right] = \frac{1}{\sqrt{2n}}P_\sigma^\top \begin{bmatrix} \mathbf{1}_n & \vec{\omega}_n & \vec{\omega}_n^2 & \cdots & \vec{\omega}_n^{n-1} & \vec{\omega}_n^n & \cdots & \vec{\omega}_n^{2n-1} \\ \mathbf{1}_n & \omega_{2n}\vec{\omega}_n & \omega_{2n}^2\vec{\omega}_n^2 & \cdots & \omega_{2n}^{n-1}\vec{\omega}_n^{n-1} & \omega_{2n}^n\vec{\omega}_n^n & \cdots & \omega_{2n}^{2n-1}\vec{\omega}_n^{2n-1} \end{bmatrix}$$
$$= \frac{1}{\sqrt{2}}P_\sigma^\top \begin{bmatrix} Q_n^\star & Q_n^\star \\ Q_n^\star D_n & -Q_n^\star D_n \end{bmatrix} = \frac{1}{\sqrt{2}}P_\sigma^\top \begin{bmatrix} Q_n^\star & \\ & Q_n^\star \end{bmatrix}\begin{bmatrix} I_n & I_n \\ D_n & -D_n \end{bmatrix}$$

A very similar chain of arguments can be made for Q_{2n} but we can also infer it directly from the above, since $(Q_{2n}^\star)^\star = Q_{2n}$, we find that

$$Q_{2n} = \left(\frac{1}{\sqrt{2}}P_\sigma^\top \begin{bmatrix} Q_n^\star & \\ & Q_n^\star \end{bmatrix}\begin{bmatrix} I_n & I_n \\ D_n & -D_n \end{bmatrix}\right)^\star = \frac{1}{\sqrt{2}}\begin{bmatrix} I_n & D_n \\ I_n & -D_n \end{bmatrix}\begin{bmatrix} Q_n & \\ & Q_n \end{bmatrix}P_\sigma$$

3. Orthogonal polynomials

Problem 3.1 **(B)** Construct $p_0(x), p_1(x), p_2(x), p_3(x)$, monic OPs for the weight $\sqrt{1-x^2}$ on $[-1,1]$. Hint: first compute $\int_{-1}^1 x^k\sqrt{1-x^2}\mathrm{d}x$ for $0 \le\ k \le\ 2$ using a change-of-variables.

Following the hint, we first calculate $\int_{-1}^1 x^k\sqrt{1-x^2}\mathrm{d}x$. By symmetry, it's zero when k is odd and double the integral on $[0,1]$ when k is even.

$$\underbrace{\int_0^1 x^k\sqrt{1-x^2}\mathrm{d}x}_{I_k} =_{x=\sin t} \underbrace{\int_0^{\pi/2}\sin^k(t)\cos^2(t)\mathrm{d}t}_{I_k} = \underbrace{\int_0^{\pi/2}\sin^k t\mathrm{d}t}_{J_k} - \underbrace{\int_0^{\pi/2}\sin^{k+2}t\mathrm{d}t}_{J_{k+2}}.$$

Meanwhile,

$$J_k = -\int_0^{\pi/2}\sin^{k-1}t\mathrm{d}(\cos t) =_{\text{integral by part}} (k-1)I_{k-2}.$$

Putting the above 2 equations together, we have $I_k = (k-1)I_{k-2} - (k+1)I_k$, so $I_k = \frac{k-1}{k+2}I_{k-2}$. Since $I_0 = \pi/4$, we have $I_k = \frac{(k-1)!!}{(k+2)!!}\frac{\pi}{2}$ for positive even k. (Note that the denominator multiplies to 4, not to 2.) Keep in mind that the integral we want is double this value when k is even.

Remark Check the beta function.

Let $p_0(x) = 1$, then $\|p_0\|^2 = 2I_0 = \pi/2$. We know from the 3-term recurrence that

$$xp_0(x) = a_0p_0(x) + p_1(x)$$

where

$$a_0 = \frac{\langle p_0, xp_0\rangle}{\|p_0\|^2} = 0.$$

Thus $p_1(x) = x$ and $\|p_1\|^2 = 2I_2 = \pi/8$. From

$$xp_1(x) = c_0p_0(x) + a_1p_1(x) + p_2(x)$$

we have

$$c_0 = \frac{\langle p_0, xp_1\rangle}{\|p_0\|^2} = 2I_2/2I_0 = 1/4$$
$$a_1 = \frac{\langle p_1, xp_1\rangle}{\|p_1\|^2} = 0$$
$$p_2(x) = xp_1(x) - c_0 - a_1p_1(x) = x^2 - 1/4$$

$$\|p_2\|^2 = 2I_4 - I_2 + 1/8 I_0 = \pi/32$$

Finally, from

$$xp_2(x) = c_1 p_1(x) + a_2 p_2(x) + p_3(x)$$

we have

$$c_1 = \frac{\langle p_1, xp_2\rangle}{\|p_1\|^2} = (2I_4 - 1/2 I_2)/(\pi/8) = 1/4$$

$$a_2 = \frac{\langle p_2, xp_2\rangle}{\|p_2\|^2} = 0$$

$$p_3(x) = xp_2(x) - c_1 p_1(x) - a_2 p_2(x) = x^3 - 1/2x$$

Problem 3.2 (C, 3-term recurrence, 1st form) Show that if $\{p_n\}$ are OPs then there exist real constants $A_n \neq 0$, B_n, and C_n such that

$$p_1(x) = (A_0 x + B_0) p_0(x)$$
$$p_{n+1}(x) = (A_n x + B_n) p_n(x) - C_n p_{n-1}(x)$$

Write this as a lower triangular linear system, given $p_0(x) = \mu \in \mathbb{R}$:

$$L_x \begin{bmatrix} p_0(x) \\ \vdots \\ p_{n+1}(x) \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

SOLUTION

The 2nd form of 3-term recurrence is

$$xp_0(x) = a_0 p_0(x) + b_0 p_1(x)$$

and

$$xp_n(x) = c_{n-1} p_{n-1}(x) + a_n p_n(x) + b_n p_{n+1}(x)$$

which is equivalent to

$$p_1(x) = \left(\frac{1}{b_0}x - \frac{a_0}{b_0}\right) p_0(x),$$

$$p_{n+1}(x) = \left(\frac{1}{b_n}x - \frac{a_n}{b_n}\right) p_n(x) - \frac{c_{n-1}}{b_n} p_{n-1}(x).$$

So we have $A_n = 1/b_n$, $B_n = -\frac{a_n}{b_n}$ and $C_n = \frac{c_{n-1}}{b_n}$.

Writing down the recurrence for every n, we have a lower tridiagonal linear system

$$\underbrace{\begin{bmatrix} 1 & & & & \\ -A_0 x - B_0 & 1 & & & \\ C_1 & -A_1 x - B_1 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & C_n & -A_n x - B_n & 1 \end{bmatrix}}_{L_x} \underbrace{\begin{bmatrix} p_0(x) \\ \vdots \\ p_{n+1}(x) \end{bmatrix}}_{\mathbf{P}(x)} = \underbrace{\begin{bmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\mu \mathbf{e}_1}$$

Problem 3.3 (B) Show that if $f(x)$ is a degree $n-1$ polynomial

$$f(x) = [p_0(x)|\cdots|p_{n-1}(x)] \underbrace{\begin{bmatrix} c_0 \\ \vdots \\ c_{n-1} \end{bmatrix}}_{\mathbf{c}}$$

then evaluation at a point can be recast as inverting an upper triangular system (*Clenshaw's algorithm*):

$$f(x) = \mu \mathbf{e}_1^\top U_x^{-1} \mathbf{c}.$$

SOLUTION

Using the same notation as the previous part we have:

$$f,(x) = \mathbf{P}(x)^\top \mathbf{c} = (\mu L_x^{-1} \mathbf{e}_1)^\top \mathbf{c} = \mu \mathbf{e}_1^\top (L_x^\top)^{-1} \mathbf{c}$$

that is $U_x = L_x^\top$.

Problem 3.4 (B) If $w(-x) = w(x)$ for a weight supported on $[-b, b]$ show that $a_n = 0$. Hint: first show that the (monic) polynomials $p_{2n}(x)$ are even and $p_{2n+1}(x)$ are odd.

SOLUTION

An integral is zero if its integrand is odd. Moreover an even function times an odd function is odd and an odd function times an odd fuction is even. Note that $p_0(x)$ and $w(x)$ are even and x is odd.

We see that a_0 is zero:

$$\langle p_0, x p_0(x) \rangle = \int_{-b}^{b} x w(x) \mathrm{d}x = 0$$

since $xw(x)$ is odd, which shows that

$$p_1(x) = x p_0(x)$$

is odd. We now proceed by induction. Assume that p_{2n} is even and p_{2n-1} is odd. We have:

$$\langle p_{2n}, x p_{2n}(x) \rangle = \int_{-b}^{b} x w(x) p_{2n}(x)^2 \mathrm{d}x = 0$$

since $xw(x)p_{2n}(x)^2$ is odd, therefore $a_{2n} = 0$. Thus from

$$p_{2n+1}(x) = (x p_{2n}(x) - c_{2n-1} p_{2n-1}(x))/b_{2n}$$

we see that p_{2n+1} is odd. Then

$$\langle p_{2n+1}, x p_{2n+1}(x) \rangle = \int_{-b}^{b} x w(x) p_{2n+1}(x)^2 \mathrm{d}x = 0$$

since $xw(x)p_{2n+1}(x)^2$ is odd, therefore $a_{2n+1} = 0$. and hence

$$p_{2n+2}(x) = (x p_{2n+1}(x) - c_{2n} p_{2n}(x))/b_{2n+1}$$

is even.

END

Week 9

1. Jacobi matrices

Problem 1.1 (C) What are the upper 3x3 sub-block of the Jacobi matrix for the monic and orthonormal polynomials with respect to the following weights on $[-1, 1]$:

$$1 - x, \sqrt{1 - x^2}, 1 - x^2$$

SOLUTION

Monic We know that for monic ($b_n = 1$) orthogonal polynomials we can write the upper 3x3 block in the form

$$X = \begin{bmatrix} a_0 & c_0 & 0 \\ 1 & a_1 & c_1 \\ 0 & 1 & a_2 \end{bmatrix}$$

Note that for non-orthonormal polynomials the Jacobi matrix is technically the tranpose, i.e. X^T.

1. $w(x) = 1 - x$

Take $p_0(x) = k_0 = 1$ (monic), then $p_1(x) = x + k_1^{(0)}$

For the recurrence formula

$$xp_0(x) = a_0 p_0(x) + p_1(x)$$

i.e. $x = a_0 + x + k_1^{(0)}$, that gives $k_1^{(0)} = -a_0$

and for orthogonality, using recurrence formula and multiplying it by p_0 (inner product), we have

$$a_0 = \frac{\langle p_0, xp_0 \rangle}{\|p_0\|^2} = \frac{\int_{-1}^{1} x(1-x)dx}{2} = -\frac{1}{3}$$

and $k_1^{(0)} = \frac{1}{3}$

For $p_2(x) = x^2 + k_2^{(1)}x + k_2^{(0)}$ we obtain

$$xp_1(x) = c_0 p_0(x) + a_1 p_1(x) + p_2(x)$$

i.e. $x^2 + \frac{1}{3}x = c_0 + a_1 x + \frac{1}{3}a_1 + x^2 + k_2^{(1)}x + k_2^{(0)}$, that gives $a_1 + k_2^{(1)} = \frac{1}{3}$ and $c_0 + \frac{1}{3}a_1 + k_2^{(0)} = 0$

As before we then find

$$c_0 = \frac{\langle p_0, xp_1 \rangle}{\|p_0\|^2} = \frac{\int_{-1}^{1} (x + \frac{1}{3})x(1-x)dx}{2} = \frac{2}{9}$$

and

$$a_1 = \frac{\langle p_1, xp_1 \rangle}{\|p_1\|^2} = \frac{\int_{-1}^{1} (x + \frac{1}{3})^2 x(1-x)dx}{\int_{-1}^{1} (x + \frac{1}{3})^2 (1-x)dx} = -\frac{1}{15}$$

Thus, $k_2^{(0)} = -\frac{1}{5}$ and $k_2^{(1)} = \frac{2}{5}$.

Now

$$xp_2(x) = c_1 p_1(x) + a_2 p_2(x) + p_3(x)$$

And once again as before:

$$c_1 = \frac{\langle p_1, xp_2\rangle}{\|p_1\|^2} = \frac{\int_{-1}^{1}(x+\frac{1}{3})(x^2+\frac{1}{9}x-\frac{1}{5})x(1-x)dx}{\int_{-1}^{1}(x+\frac{1}{3})^2(1-x)dx} = \frac{16}{45}$$

and

$$a_2 = \frac{\langle p_2, xp_2\rangle}{\|p_2\|^2} = \frac{\int_{-1}^{1}(x^2+\frac{1}{9}x-\frac{1}{5})^2x(1-x)dx}{\int_{-1}^{1}(x^2+\frac{1}{9}x-\frac{1}{5})^2(1-x)dx} = \frac{2085}{3451}$$

2. $w(x) = \sqrt{1-x^2}$

Take $p_0(x) = k_0 = 1$ (monic), then $p_1(x) = x + k_1^{(0)}$

From the recurrence we have

$$xp_0(x) = a_0p_0(x) + p_1(x)$$

i.e. $x = a_0 + x + k_1^{(0)}$, which gives $k_1^{(0)} = -a_0$.

We then proceed as before to obtain

$$a_0 = \frac{\langle p_0, xp_0\rangle}{\|p_0\|^2} = \frac{\int_{-1}^{1} x\sqrt{1-x^2}dx}{\pi/2} = 0$$

and $k_1^{(0)} = 0$

Likewise for $p_2(x) = x^2 + k_2^{(1)}x + k_2^{(0)}$:

$$xp_1(x) = c_0p_0(x) + a_1p_1(x) + p_2(x)$$

i.e. $x^2 = c_0 + a_1x + x^2 + k_2^{(1)}x + k_2^{(0)}$, which gives $a_1 + k_2^{(1)} = 0$ and $c_0 + k_2^{(0)} = 0$

Proceeding as before once again we obtain:

$$c_0 = \frac{\langle p_0, xp_1\rangle}{\|p_0\|^2}\frac{\int_{-1}^{1} x^2\sqrt{1-x^2}dx}{\pi/2} = \frac{\pi/8}{\pi/2} = \frac{1}{4}$$

and

$$a_1 = \frac{\langle p_1, xp_1\rangle}{\|p_1\|^2} = \frac{\int_{-1}^{1} x^3\sqrt{1-x^2}dx}{\int_{-1}^{1} x^2\sqrt{1-x^2}dx} = 0$$

Thus, $k_2^{(0)} = -\frac{1}{4}$ and $k_2^{(1)} = 0$.

Finally:

$$xp_2(x) = c_1p_1(x) + a_2p_2(x) + p_3(x)$$

and thus

$$c_1 = \frac{\langle p_1, xp_2\rangle}{\|p_1\|^2} = \frac{\int_{-1}^{1}(x^2 - \frac{1}{4})x^2\sqrt{1-x^2}dx}{\int_{-1}^{1} x^2\sqrt{1-x^2}dx} = \frac{\pi/32}{\pi/8} = \frac{1}{4}$$

and

$$a_2 = \frac{\langle p_2, xp_2\rangle}{\|p_2\|^2} = \frac{\int_{-1}^{1}(x^2 - \frac{1}{4})^2 x\sqrt{1-x^2}dx}{\int_{-1}^{1}(x^2 - \frac{1}{4})^2\sqrt{1-x^2}dx} = 0$$

.

3. $w(x) = 1 - x^2$

Take $p_0(x) = k_0 = 1$ (monic), then $p_1(x) = x + k_1^{(0)}$

From recurrence we have

$$xp_0(x) = a_0 p_0(x) + p_1(x)$$

i.e. $x = a_0 + x + k_1^{(0)}$, which gives $k_1^{(0)} = -a_0$

The rest works as before:

$$a_0 = \frac{\langle p_0, xp_0\rangle}{\|p_0\|^2} = \frac{\int_{-1}^{1} x(1-x^2)dx}{4/3} = 0$$

and $k_1^{(0)} = 0$

Then from

$$xp_1(x) = c_0 p_0(x) + a_1 p_1(x) + p_2(x)$$

i.e. $x^2 = c_0 + a_1 x + x^2 + k_2^{(1)} x + k_2^{(0)}$, which gives $a_1 + k_2^{(1)} = 0$ and $c_0 + k_2^{(0)} = 0$

we find

$$c_0 = \frac{\langle p_0, xp_1\rangle}{\|p_0\|^2}\frac{\int_{-1}^{1} x^2(1-x^2)dx}{4/15} = \frac{4/15}{4/3} = \frac{1}{5}$$

and

$$a_1 = \frac{\langle p_1, xp_1\rangle}{\|p_1\|^2} = \frac{\int_{-1}^{1} x^3(1-x^2)dx}{\int_{-1}^{1} x^2(1-x^2)dx} = 0$$

i.e. $k_2^{(0)} = -\frac{1}{5}$ and $k_2^{(1)} = 0$.

Finally,

$$xp_2(x) = c_1 p_1(x) + a_2 p_2(x) + p_3(x)$$

and thus

$$c_1 = \frac{\langle p_1, xp_2\rangle}{\|p_1\|^2} = \frac{\int_{-1}^{1}(x^2 - \frac{1}{5})x^2(1-x^2)dx}{\int_{-1}^{1} x^2(1-x^2)dx} = \frac{32/525}{4/15} = \frac{8}{35}$$

and

$$a_2 = \frac{\langle p_2, xp_2\rangle}{\|p_2\|^2} = \frac{\int_{-1}^{1}(x^2 - \frac{1}{4})^2x(1-x^2)dx}{\int_{-1}^{1}(x^2 - \frac{1}{4})^2(1-x^2)dx} = 0$$

Orthonormal We know that for orthonormal ($b_n = c_n$) polynomials we can write the upper 3x3 block of the Jacobi matrix in the form

$$X = \begin{bmatrix} a_0 & b_0 & 0 \\ b_0 & a_1 & b_1 \\ 0 & b_1 & a_2 \end{bmatrix}$$

We could proceed step by step as above to find the entries of the recurrences but we may also take a short-cut since we already know the monic recurrence. Given any recurrence relationship coefficients a_n, b_n, c_n for orthogonal polynomials for a given weight, the(symmetric) Jacobi matrix corresponding to the orthogonal polynomials for the same weight has entries $\widehat{b}_n = \sqrt{c_n b_n}$ and $\widehat{a}_n = a_n$. Using this the 3×3 blocks of the Jacobi matrices belonging to the orthonormal polynomials are straightforwardly computed from the above.

Problem 1.2 (B) Consider the *truncated Jacobi matrix* associated with orthonormal polynomials $q_n(x)$:

$$X_n := \begin{bmatrix} a_0 & b_0 & & \\ b_0 & \ddots & \ddots & \\ & \ddots & a_{n-2} & b_{n-2} \\ & & b_{n-2} & a_{n-1} \end{bmatrix} \in \mathbb{R}^{n\times n}$$

Show that

$$x[q_0(x)|\cdots|q_{n-1}(x)] = [q_0(x)|\cdots|q_{n-1}(x)]X_n + b_{n-1}q_n(x)\mathbf{e}_n^\top.$$

SOLUTION

We have that in the infinite form of X:

$$x[q_0, ..., q_{n-1}, ...] = [q_0, ..., q_{n-1}, ...]X$$

Thus, selecting just the first n terms, we have

$$x[q_0, ..., q_{n-1}] = [[q_0, ..., q_{n-1}, q_n]X]_{1,...,n} = [[q_0, ..., q_{n-1}, ...]X_{n+1}]_{1,...,n}$$

considering that we don't need the terms after the $n+1$ since the relevant terms in the matrix X would be 0s.

Writing the n^{th} term explicitly we obtain

$$xq_{n-1} = [q_{n-2}, q_{n-1}, q_n]\{X_{n+1}\}_{n,n-1:n+1} = b_{n-2}q_{n-2} + a_{n-1}q_{n-1} + b_{n-1}q_n$$

While

$$[q_{n-2}, q_{n-1}, q_n]\{X_n\}_{n,n-1:n+1} = b_{n-2}q_{n-2} + a_{n-1}q_{n-1}$$

Thus, the only extra term that this form is missing is $b_{n-1}q_n$ in the n^{th} term which can be added using the vector e_n^T, hence

$$x[q_0(x)|\cdots|q_{n-1}(x)] = [q_0(x)|\cdots|q_{n-1}(x)]X_n + b_{n-1}q_n(x)\mathbf{e}_n^\top$$

Problem 1.3 (A) Prove the *Christoffel-Darboux Formula*:

$$\sum_{k=0}^{n} q_k(x)q_k(y) = b_n \frac{q_{n+1}(x)q_n(y) - q_n(x)q_{n+1}(y)}{x-y}$$

Hint: Consider

$$(x-y)[q_0(x)|\cdots|q_n(x)] \begin{bmatrix} q_0(y) \\ \vdots \\ q_n(y) \end{bmatrix}.$$

SOLUTION

We have to prove that

$$(x-y)[q_0(x)|\cdots|q_n(x)] \begin{bmatrix} q_0(y) \\ \vdots \\ q_n(y) \end{bmatrix} = (x-y)\sum_{k=0}^{n} q_k(x)q_k(y) = b_n(q_{n+1}(x)q_n(y) - q_n(x)q_{n+1}(y))$$

Using 1.2 we have

$x[q_0(x)|\cdots|q_n(x)] = [q_0(x)|\cdots|q_n(x)]X_{n+1} + b_n q_{n+1}(x)e_{n+1}^T$

$y[q_0(y)|\cdots|q_n(y)] = [q_0(y)|\cdots|q_n(y)]Y_{n+1} + b_n q_{n+1}(y)e_{n+1}^T$

Right-multiplying these formulas respectively by $[q_0(y)|\cdots|q_n(y)]^T$ and $[q_0(x)|\cdots|q_n(x)]^T$ we obtain

$x[q_0(x)|\cdots|q_n(x)][q_0(y)|\cdots|q_n(y)]^T = [q_0(x)|\cdots|q_n(x)]X_{n+1}[q_0(y)|\cdots|q_n(y)]^T + b_n q_{n+1}(x)q_n(y)$

$y[q_0(y)|\cdots|q_n(y)][q_0(x)|\cdots|q_n(x)]^T = [q_0(y)|\cdots|q_n(y)]Y_{n+1}[q_0(x)|\cdots|q_n(x)]^T + b_n q_{n+1}(y)q_n(x)$

which (transposed of a number) corresponds to

$y[q_0(x)|\cdots|q_n(x)][q_0(y)|\cdots|q_n(y)]^T = [q_0(x)|\cdots|q_n(x)]Y_{n+1}^T[q_0(y)|\cdots|q_n(y)]^T + b_n q_n(x)q_{n+1}(y)$

Subtracting from the first equation the second (transposed) we get

$$(x-y)[q_0(x)|\cdots|q_n(x)] \begin{bmatrix} q_0(y) \\ \vdots \\ q_n(y) \end{bmatrix} = [q_0(x)|\cdots|q_n(x)](X_{n+1} - Y_{n+1}^T) \begin{bmatrix} q_0(y) \\ \vdots \\ q_n(y) \end{bmatrix} + b_n(q_{n+1}(x)q_n(y) - q_n(x)q_{n+1}(y))$$

But since the matrices X_{n+1} and Y_{n+1} do not depend on the values of x and y, but on the polynomials, then they are the same matrix.

In particular, they are symmetric matrices since the q_ks are orthonormal. Thus, $X_{n+1} - Y_{n+1}^T = 0$

Hence, $[q_0(x)|\cdots|q_n(x)](X_{n+1} - Y_{n+1}^T) \begin{bmatrix} q_0(y) \\ \vdots \\ q_n(y) \end{bmatrix} = 0$, thus

$$\sum_{k=0}^{n} q_k(x)q_k(y) = [q_0(x)|\cdots|q_n(x)] \begin{bmatrix} q_0(y) \\ \vdots \\ q_n(y) \end{bmatrix} = \frac{b_n(q_{n+1}(x)q_n(y) - q_n(x)q_{n+1}(y))}{x-y}$$

2. Chebyshev polynomials

Problem 2.1 (C) What is the Jacobi matrix for $T_n(x)$? What scaling $q_n(x) = D_n T_n(x)$ gives us orthonormal polynomials?

SOLUTION

Let $w(x) = 1/\sqrt{1-x^2}$ on $[-1, 1]$ and $q_n(x) = D_n T_n(x)$.

We know $T_0(x) = 1$ and we want $\|q_0\| = 1$

Thus,

$$1 = \int_{-1}^{1} D_0^2 w(x) dx = D_0^2 \pi$$

Hence, $D_0 = \sqrt{\frac{1}{\pi}}$

Now let $\|q_1\| = 1$

$$1 = \int_{-1}^{1} D_1^2 (\cos(acos(x)))^2 w(x) dx = D_1^2 \int_{-1}^{1} \frac{x^2}{\sqrt{1-x^2}} dx = D_1^2 \frac{\pi}{2}$$

Hence, $D_1 = \sqrt{\frac{2}{\pi}}$

In general we want $\|q_n\| = 1$, thus

$$1 = \int_{-1}^{1} D_n^2 (\cos(n \times acos(x)))^2 w(x) dx = D_n^2 \frac{\pi}{2}$$

Hence, $\forall n \geq 1$,

$$D_n = \sqrt{\frac{2}{\pi}}$$

Problem 2.2 (B) Consider the function

$$f(x) = \sum_{k=0}^{\infty} \frac{T_k(x)}{k!}$$

For what coefficients c_k does

$$(x^2 + 1) f(x) = \sum_{k=0}^{\infty} c_k T_k(x)?$$

SOLUTION

We need

$$(x^2 + 1) \sum_{k=0}^{\infty} \frac{T_k(x)}{k!} = \sum_{k=0}^{\infty} c_k T_k(x)$$

We can use the following to rewrite the LHS

$$x T_0(x) = T_1(x)$$

$$x T_n(x) = \frac{T_{n-1}(x) + T_{n+1}(x)}{2}$$

Indeed,

$$(x^2+1)\sum_{k=0}^{\infty}\frac{T_k(x)}{k!}=\sum_{k=0}^{\infty}\frac{x^2T_k(x)}{k!}+\sum_{k=0}^{\infty}\frac{T_k(x)}{k!}=xT_1(x)+\sum_{k=1}^{\infty}x\frac{T_{k-1}(x)+T_{k+1}(x)}{2k!}+\sum_{k=0}^{\infty}\frac{T_k(x)}{k!}=$$

Using

$$x\frac{T_0(x)+T_2(x)}{2*1!}=\frac{T_1(x)}{2}+\frac{T_1(x)+T_3(x)}{4}$$

and

$$\sum_{k=2}^{\infty}x\frac{T_{k-1}(x)+T_{k+1}(x)}{2k!}=\sum_{k=2}^{\infty}\left(\frac{T_{k-2}(x)+T_k(x)}{4k!}+\frac{T_k(x)+T_{k+2}(x)}{4k!}\right)$$

we obtain

$$=\frac{T_0(x)+T_2(x)}{2}+\frac{T_1(x)}{2}+\frac{T_1(x)+T_3(x)}{4}+\sum_{k=2}^{\infty}\left(\frac{T_{k-2}(x)+T_k(x)}{4k!}+\frac{T_k(x)+T_{k+2}(x)}{4k!}\right)+\sum_{k=0}^{\infty}\frac{T_k(x)}{k!}$$

So, we have

$$c_0=\frac{1}{2}+\frac{1}{4*2!}+1=\frac{13}{8}$$

$$c_1=\frac{1}{2}+\frac{1}{4*1!}+\frac{1}{4*3!}+1=\frac{43}{24}$$

$$c_2=\frac{1}{2}+\frac{2}{4*2!}+\frac{1}{4*4!}+1$$

$$c_3=\frac{1}{4}+\frac{2}{4*3!}+\frac{1}{4*5!}+1$$

$$c_4=\frac{1}{4*2!}+\frac{2}{4*4!}+\frac{1}{4*6!}+1$$

For $n\geq 3$ we have

$$c_n=\frac{1}{4*(n-2)!}+\frac{1}{2*n!}+\frac{1}{4*(n+2)!}+1$$

Problem 2.3 (B) Consider orthogonal polynomials with respect to $\sqrt{1-x^2}$ on $[0,1]$ with the normalisation

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

Prove that

$$U_n(\cos\theta)=\frac{\sin(n+1)\theta}{\sin\theta}$$

SOLUTION

We need to verify: 1. graded polynomials 2. orthogonal w.r.t. $\sqrt{1-x^2}$ on $[-1,1]$, and 3. have the leading coefficient 2^n.

(2) follows under a change of variables

$$\int_{-1}^{1} U_n(x)U_m(x)\sqrt{1-x^2}\mathrm{d}x = \int_0^{\pi} U_n(\cos\theta)U_m(\cos\theta)\sin^2\theta\mathrm{d}\theta = \int_0^{\pi}\sin(n+1)\theta\sin(m+1)\theta\mathrm{d}\theta = \frac{\pi}{2}\delta_{mn}$$

where the last step is a result from Fourier theories.

To see that they are graded we use the fact that

$$xU_n(x) = \frac{\cos\theta\sin(n+1)\theta}{\sin\theta} = \frac{\sin(n+2)\theta+\sin n\theta}{2\sin\theta}$$

In other words $2xU_n(x) = U_{n+1}(x) + U_{n-1}(x)$. Since each time we multiply by $2x$ and $U_0(x) = 1$ we have

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

which also proves (3).

Problem 2.4 (B) Show that

$$xU_0(x) = U_1(x)/2$$
$$xU_n(x) = \frac{U_{n-1}(x)}{2} + \frac{U_{n+1}(x)}{2}.$$

SOLUTION

SOLUTION

The first result is trivial.

To get the second result, recall that

$$U_n(\cos\theta) = \cos\theta U_{n-1}(\cos\theta) + \cos n\theta$$

and

$$\cos n\theta = \cos\theta\cos(n-1)\theta - (1-\cos^2\theta)U_{n-2}(\cos\theta).$$

The first equation gives $\cos n\theta$ in terms of U_n and U_{n-1}. Substitute the result into the second equation to get

$$U_n(x) - xU_{n-1}(x) = x(U_{n-1}(x) - xU_{n-2}(x)) - (1-x^2)U_{n-2}(x)$$

which reduces to the desired recurrence.

Problem 2.5 (C) What is the Jacobi matrix for $U_n(x)$? What scaling $q_n(x) = D_n U_n(x)$ gives us orthonormal polynomials?

Solution

Problem 2.4 gives the Jacobi matrix

$$\begin{bmatrix} 0 & 1/2 & \\ 1/2 & \ddots & \ddots \\ & \ddots & \end{bmatrix}$$

Problem 2.3 - Claim 3 gives the norms $\|U_n\| = \sqrt{\pi/2}$, so $D_n = \sqrt{2}\pi$.

3. Hermite polynomials

Problem 3.1 (B) Consider Hermite polynomials orthogonal with respect to the weight $\exp(-x^2)$ on $\mathbb{R}$ with the normalisation

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

Prove the Rodrigues formula

$$H_n(x) = (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp(-x^2)$$

SOLUTION

We need to verify: 1. graded polynomials 2. orthogonal to all lower degree polynomials on $\mathbb{R}$, and 3. have the right leading coefficient 2^n.

Comparing the Rodrigues formula for n and $n-1$, we find that

$$(-1)^n \exp(-x^2) H_n(x) = \frac{\mathrm{d}}{\mathrm{d}x} \left((-1)^{n-1} \exp(-x^2) H_{n-1}(x) \right)$$

which reduces to

$$H_n(x) = 2x H_{n-1}(x) - H'_{n-1}(x).$$

(1) and (3) then follows from induction since $H_0(x) = 1$.

(2) follows by integration by parts. If r_m is a degree $m < n$ polynomial we have:

$$\int_{-\infty}^{\infty} H_n(x) r_m(x) \exp(-x^2) \mathrm{d}x = \int_{-\infty}^{\infty} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp(-x^2) \exp(-x^2) \mathrm{d}x = \cdots \text{integration by parts} \cdots = (-1)^n \int_{-\infty}^{\infty} \exp(-x^2) r_m^{(n)}($$

Problem 3.2 (C) What are $k_n^{(1)}$ and $k_n^{(2)}$ such that

$$H_n(x) = 2^n x^n + k_n^{(1)} x^{n-1} + k_n^{(2)} x^{n-2} + O(x^{n-3})$$

SOLUTION

From Problem 3.1,

$$H_n(x) = 2x H_{n-1}(x) - H'_{n-1}(x).$$

Thus we have

$$k_n^{(1)} = 2k_{n-1}^{(1)}$$
$$k_n^{(2)} = 2k_{n-1}^{(2)} - (n-1)2^{n-1}$$

Since $k_0^{(1)} = 0$, we have $k_n^{(1)} = 0$. For the second recurrence, divide both sides by 2^n:

$$2^{-n} k_n^{(2)} = 2^{-(n-1)} k_{n-1}^{(2)} - \frac{n-1}{2}$$

Since $k_0^{(2)} = 0$, we have $2^{-n} k_n^{(2)} = -\frac{1+\cdots+(n-1)}{2} = -\frac{n(n-1)}{4}$, so $k_n^{(2)} = n(n-1)2^{n-2}$.

Problem 3.3 (B) Deduce the 3-term recurrence relationship for $H_n(x)$.

SOLUTION

Our goal is to find a_n, b_n and c_n such that

$$x H_n(x) = c_{n-1} H_{n-1}(x) + a_n H_n(x) + b_n H_{n+1}(x).$$

Compare the 3 leading coefficients on both sides and use the results from Problem 3.1 and Problem 3.2:

$$2^n = 0 + 0 + b_n 2^{n+1}$$
$$0 = 0 + a_n 2^n + 0$$
$$n(n-1)2^{n-2} = c_{n-1} 2^{n-1} + 0 + b_n (n+1) n 2^{n-1}$$

Thus we have $b_n = 1/2$, $a_n = 0$ and $c_{n-1} = -n$.

4. Interpolation

Problem 4.1 (C) Use Lagrange interpolation to interpolate the function $\cos x$ by a polynomial at the points $[0, 2, 3, 4]$ and evaluate at $x = 1$.

SOLUTION

- $l_0(x) = \frac{(x-2)(x-3)(x-4)}{(0-2)(0-3)(0-4)} = -\frac{1}{24}(x-2)(x-3)(x-4)$
- $l_2(x) = \frac{(x-0)(x-3)(x-4)}{(2-0)(2-3)(2-4)} = \frac{1}{4}x(x-3)(x-4)$
- $l_3(x) = \frac{(x-0)(x-2)(x-4)}{(3-0)(3-2)(3-4)} = -\frac{1}{3}x(x-2)(x-4)$
- $l_4(x) = \frac{(x-0)(x-2)(x-3)}{(4-0)(4-2)(4-3)} = \frac{1}{8}x(x-2)(x-3)$

$p(x) = \cos(0)l_0(x) + \cos(2)l_2(x) + \cos(3)l_3(x) + \cos(4)l_4(x)$

$l_0(1) = 1/4$, $l_2(1) = 3/2$, $l_3(1) = -1$, $l_4(1) = 1/4$, so $p(1) = 1/4\cos(0) + 3/2\cos(2) - \cos(3) + 1/4\cos(4)$.

Problem 4.2 (A) Consider the re-expanding the Lagrange basis in monomials. Use this to construct an explicit formula for the inverse of the Vandermonde matrix.

SOLUTION

We use two methods to interpolate $f(x)$

Vandermonde matrix

$$p(x) = \sum_{k=1}^{n} c_k x^{k-1} = \underbrace{\begin{bmatrix} 1 & x & \cdots & x^{n-1} \end{bmatrix}}_{\mathbf{M}(x)} \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}}_{\mathbf{c}}$$

where

$$V\mathbf{c} = \underbrace{\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}}_{\mathbf{f}}$$

Lagrange interpolation

$$p(x) = \sum_{k=1}^{n} f(x_k) l_k(x) = \underbrace{\begin{bmatrix} l_1(x) & \cdots & l_n(x) \end{bmatrix}}_{\mathbf{L}(x)} \mathbf{f}$$

Comparing the two results, we have

$$\mathbf{M}(x)V^{-1}\mathbf{f} = \mathbf{L}(x)\mathbf{f},$$

so

$$\mathbf{M}(x)V^{-1} = \mathbf{L}(x)$$

which means that the (i, j)-entry of V^{-1} is the coefficient of x^{i-1} in $l_j(x)$.

See here for the complete explicit expression.

Week 10

1. Orthogonal Polynomial Roots

Problem 1.1 (C) Compute the roots of $P_3(x)$, orthogonal with respect to $w(x) = 1$ on $[-1, 1]$, by computing the eigenvalues of a 3×3 truncation of the Jacobi matrix.

SOLUTION

We have, $P_0(x) = 1$. Though recall that in order to use Lemma (zeros), the Jacobi matrix must be symmetric and hence the polynomials orthonormal. So Take $Q_0(x) = 1/||P_0(x)|| = \frac{1}{\sqrt{2}}$. Then we have, by the three term recurrence relationship,

$$xQ_0(x) = a_0Q_0(x) + b_0Q_1(x),$$

and taking the inner product of both sides with $Q_0(x)$ we get,

$$a_0 = \langle xQ_0(x), Q_0(x)\rangle = \int_{-1}^{1} x/2dx = 0.$$

Next recall that $P_1(x) = x$ and so $Q_1(x) = x/||P_1(x)|| = \sqrt{\frac{3}{2}}x$. We then have, taking the innner product of the first equation above with $Q_1(x)$,

$$b_0 = \langle xQ_0(x), Q_1(x)\rangle = \int_{-1}^{1} \frac{\sqrt{3}}{2}x^2dx = \frac{1}{\sqrt{3}},$$

and also $b_0 = c_0$ by the Corollary (orthonormal 3-term recurrence). We have,

$$a_1 = \langle xQ_1(x), Q_1(x)\rangle = \int_{-1}^{1} \frac{3}{2}x^3dx = 0.$$

Recall that $P_2(x) = \frac{1}{2}(3x^2 - 1)$, so that $Q_2(x) = P_2(x)/||P_2(x)|| = \sqrt{\frac{5}{8}}(3x^2 - 1)$, and that,

$$xQ_1(x) = c_0Q_0(x) + a_1Q_1(x) + b_1Q_2(x).$$

Taking inner the inner product of both sides with $Q_2(x)$, we see that,

$$c_1 = b_1 = \langle xQ_1(x), Q_2(x)\rangle = \int_{-1}^{1} \sqrt{\frac{5}{8}} \cdot \sqrt{\frac{3}{2}}(3x^2 - 1) \cdot x \cdot xdx = \frac{2}{\sqrt{15}}.$$

Finally,

$$a_2 = \langle Q_2(x), xQ_2(x)\rangle = \frac{5}{8}\int_{-1}^{1} (3x^2 - 1)^2xdx = 0.$$

This gives us the truncated Jacobi matrix,

$$X_3 = \begin{bmatrix} a_0 & b_0 & 0 \\ b_0 & a_1 & b_1 \\ 0 & b_1 & a_2 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{15}} \\ 0 & \frac{2}{\sqrt{15}} & 0 \end{bmatrix},$$

whose eigenvalues are the zeros of $Q_3(x)$, and hence the zeros of $P_3(x)$ since they are the same up to a constant. To work out the eigenvalues, we have,

$$\begin{aligned} |X_3 - \lambda I| = \begin{vmatrix} -\lambda & \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{3}} & -\lambda & \frac{2}{\sqrt{15}} \\ 0 & \frac{2}{\sqrt{15}} & -\lambda \end{vmatrix} &= 0 \\ \Leftrightarrow -\lambda(\lambda^2 - \frac{4}{15}) - \frac{1}{\sqrt{3}} \cdot \frac{-\lambda}{\sqrt{3}} &= 0 \\ \Leftrightarrow -\lambda^3 + \frac{3}{5}\lambda &= 0, \end{aligned}$$

which has solutions $\lambda = 0, \pm\sqrt{\frac{3}{5}}$

END

Problem 1.2 (B) Give an explicit diagonalisation of

$$X_n = \begin{bmatrix} 0 & 1/2 & & \\ 1/2 & 0 & \ddots & \\ & \ddots & \ddots & 1/2 \\ & & 1/2 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

for all n by relating it to the Jacobi matrix for $U_n(x)$.

SOLUTION

Recall the three term recurrence for the Chebyshev Polynomials U_n,

$$\begin{aligned} xU_0(x) &= \frac{1}{2}U_1(x), \\ xU_n(x) &= \frac{U_{n-1}(x)}{2} + \frac{U_{n+1}(x)}{2}, \end{aligned}$$

and hence, we can see that,

$$X_n = \begin{bmatrix} 0 & 1/2 & & \\ 1/2 & 0 & \ddots & \\ & \ddots & \ddots & 1/2 \\ & & 1/2 & 0 \end{bmatrix},$$

is the $n \times n$ truncation of the Jacobi matrix. If $x_1, \dots, x_n$ are the zeros of $U_n(x)$, by Lemma (zeros) we have that,

$$X_n Q_n = Q_n \begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_n \end{bmatrix},$$

for,

$$Q_n = \begin{bmatrix} U_0(x_1) & \cdots & U_0(x_n) \\ \vdots & \ddots & \vdots \\ U_{n-1}(x_1) & \cdots & U_{n-1}(x_n) \end{bmatrix} \tilde{D} = \tilde{Q}_n \tilde{D}$$

where

$$\tilde{D} = \begin{bmatrix} 1/\sqrt{U_0(x_1)^2 + \cdots + U_{n-1}(x_1)^2} & \\ & \ddots \\ & 1/\sqrt{U_0(x_n)^2 + \cdots + U_{n-1}(x_n)^2} \end{bmatrix}$$

guarantess that Q_n is orthogonal. Recall that if $x = \cos\theta$ then $U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}$, so in particular the roots

of $U_n(x)$ are $x_k = \cos\left(\frac{k\pi}{n+1}\right)$ for $k = 1, \dots, n$, (where $\sin\left(\frac{k\pi}{n+1}\right) \neq 0$). Hence, we have,

$$\begin{aligned}
X_n &= Q_n \begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_n \end{bmatrix} Q_n^\top \\
&= \tilde{Q}_n \tilde{D} \begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_n \end{bmatrix} \tilde{D}^{-1} \tilde{Q}_n^{-1} \\
&= \tilde{Q}_n \begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_n \end{bmatrix} \tilde{D}\tilde{D}^{-1} \tilde{Q}_n^{-1} \\
&= \tilde{Q}_n \begin{bmatrix} \cos\left(\frac{\pi}{n+1}\right) & & & \\ & \cos\left(\frac{2\pi}{n+1}\right) & & \\ & & \ddots & \\ & & & \cos\left(\frac{n\pi}{n+1}\right) \end{bmatrix} \tilde{Q}_n^{-1} \\
&= \tilde{Q}_n \Lambda_n \tilde{Q}_n^{-1},
\end{aligned}$$

where,

$$\tilde{Q}_n = \begin{bmatrix} 1 & \cdots & 1 \\ \frac{\sin(2\cdot\frac{2\pi}{n+1})}{\sin(\frac{2\pi}{n+1})} & \cdots & \frac{\sin(2\cdot\frac{n\pi}{n+1})}{\sin(\frac{n\pi}{n+1})} \\ \vdots & & \vdots \\ \frac{\sin(n\cdot\frac{2\pi}{n+1})}{\sin(\frac{2\pi}{n+1})} & \cdots & \frac{\sin(n\cdot\frac{n\pi}{n+1})}{\sin(\frac{n\pi}{n+1})} \end{bmatrix},$$

and,

$$\Lambda_n = \begin{bmatrix} \cos\left(\frac{\pi}{n+1}\right) & & & \\ & \cos\left(\frac{2\pi}{n+1}\right) & & \\ & & \ddots & \\ & & & \cos\left(\frac{n\pi}{n+1}\right) \end{bmatrix}$$

END

Problem 1.3 (A) Give an explicit solution to heat on a graph

$$\begin{aligned}
\mathbf{u}(0) &= \mathbf{u}_0 \in \mathbb{R}^n \\
\mathbf{u}_t &= \Delta \mathbf{u}
\end{aligned}$$

where

$$\Delta := \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & \ddots & & \\ & 1 & \ddots & 1 & \\ & & \ddots & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

(which corresponds to Dirichlet conditions.) Hint: use Problem 1.2 to diagonalise the problem.

SOLUTION

We have,

$$\begin{aligned}
\mathbf{u}_t &= \Delta \mathbf{u} \\
&= 2(X_n - I),
\end{aligned}$$

with X_n defined as above. Observe, for Q_n and Λ_n defined as above

$$\begin{aligned} 2(X_n - I) &= 2(Q_n \Lambda Q_n^\top - I) \\ &= Q_n(2(\Lambda - I))Q_n^\top \end{aligned}$$

We then have

$$\mathbf{v}(t) := Q_n^\top \mathbf{u}(t)$$

satisfies

$$\mathbf{v}'(t) = Q_n^\top \mathbf{u}'(t) = Q_n^\top \Delta \mathbf{u}(t) = Q_n^\top \Delta Q_n \mathbf{v}(t) = \Lambda \mathbf{v}(t)$$

This is a diagonal problem so we know that

$$\mathbf{v}(t) = \exp(\Lambda t)\mathbf{v}(0) = \exp(\Lambda t) Q_n^\top \mathbf{u}_0$$

I.e.

$$\mathbf{u}(t) = Q_n \exp(\Lambda t) Q_n^\top \mathbf{u}_0 = Q_n \begin{bmatrix} e^{2(\cos(\frac{\pi}{n+1})-1)t} & & & \\ & e^{2(\cos(\frac{2\pi}{n+1})-1)t} & & \\ & & \ddots & \\ & & & e^{2(\cos(\frac{n\pi}{n+1})-1)t} \end{bmatrix} Q_n^\top \mathbf{u}_0.$$

END

2. Interpolatory quadrature

Problem 2.1 (C) Compute the interpolatory quadrature rule for $w(x) = \sqrt{1-x^2}$ with the points $[-1, 1/2, 1]$.

SOLUTION

For the points $\mathbf{x} = \{-1, 1/2, 1\}$ we have the Lagrange polynomials:

$$\ell_1(x) = \left(\frac{x - 1/2}{-1 - 1/2}\right) \cdot \left(\frac{x-1}{-1-1}\right) = \frac{1}{3}\left(x^2 - \frac{3}{2}x + \frac{1}{2}\right),$$

and

$$\ell_2(x) = -\frac{4}{3}x^2 + \frac{4}{3}, \ell_3(x) = x^2 + \frac{1}{2}x - \frac{1}{2},$$

similarly. We can then compute the weights,

$$w_j = \int_{-1}^{1} \ell_j(x) w(x) dx,$$

using,

$$\int_{-1}^{1} x^k \sqrt{1-x^2} dx = \begin{cases} \frac{\pi}{2} & k = 0 \\ 0 & k = 1 \\ \frac{\pi}{8} & k = 2 \end{cases}$$

to find,

$$w_j = \begin{cases} \frac{\pi}{8} & j = 1 \\ \frac{\pi}{2} & j = 2 \\ -\frac{\pi}{8} & j = 3, \end{cases}$$

so that the interpolatory quadrature rule is:

$$\Sigma_3^{w,\mathbf{x}}(f) = \frac{\pi}{2}\left(\frac{1}{4}f(-1) + f(1/2) - \frac{1}{4}f(1)\right)$$

END

Problem 2.2 (C) Compute the 2-point interpolatory quadrature rule associated with roots of orthogonal polynomials for the weights $\sqrt{1-x^2}$, 1, and $1-x$ on $[-1,1]$ by integrating the Lagrange bases.

SOLUTION For $w(x) = \sqrt{1-x^2}$ the orthogonal polynomial of degree 2 is $U_2(x) = 4x^2 - 1$, with roots $\mathbf{x} = \{x = \pm\frac{1}{2}\}$. The Lagrange polynomials corresponding to these roots are,

$$\ell_1(x) = \frac{x - 1/2}{-1/2 - 1/2} = \frac{1}{2} - x,$$
$$\ell_2(x) = \frac{x + 1/2}{1/2 + 1/2} = x + \frac{1}{2}$$

We again work out the weights

$$w_j = \int_{-1}^{1} \ell_j(x) w(x) dx,$$

to find,

$$w_1 = w_2 = \frac{\pi}{4},$$

and thus the interpolatory quadrature rule is,

$$\Sigma_2^{w,\mathbf{x}}(f) = \frac{\pi}{4}(f(-1/2) + f(1/2)).$$

For $w(x) = 1$, the orthogonal polynomial of degree 2 is, using Legendre Rodriguez formula:

$$P_2(x) = \frac{1}{(-2)^2 2!} \frac{d^2}{dx^2} \left(1 - x^2\right)^2 = -\frac{1}{2} + \frac{3}{2}x^2.$$

This has roots $\mathbf{x} = \left\{\pm\frac{1}{\sqrt{3}}\right\}$. We then have,

$$\ell_1(x) = -\frac{\sqrt{3}}{2}x + \frac{1}{2}$$
$$\ell_2(x) = \frac{3}{2}x + \frac{1}{2},$$

from which we can compute the weights,

$$w_1 = w_2 = 1,$$

which give the quadrature rule:

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

Finally, with $w(x) = 1 - x$ we use the solution to PS9 Problem 1.1, which states that

$$p_2(x) = x^2 + 2x/5 - 1/5$$

which has roots, $\mathbf{x} = \left\{-\frac{1}{5} \pm \frac{\sqrt{6}}{5}\right\}$. The Lagrange polynomials are then,

$$\begin{aligned}
\ell_1(x) &= \frac{x - (-\frac{1}{5} + \frac{\sqrt{6}}{5})}{-\frac{1}{5} - \frac{\sqrt{6}}{5} - (-\frac{1}{5} + \frac{\sqrt{6}}{5})} \\
&= \frac{x - (-\frac{1}{5} + \frac{\sqrt{6}}{5})}{-\frac{2\sqrt{6}}{5}} \\
&= -\frac{5}{2\sqrt{6}}x - \frac{1}{2\sqrt{6}} + \frac{1}{2} \\
\ell_2(x) &= \frac{x - (-\frac{1}{5} - \frac{\sqrt{6}}{5})}{\frac{2\sqrt{6}}{5}} \\
&= \frac{5}{2\sqrt{6}}x + \frac{1}{2\sqrt{6}} + \frac{1}{2}
\end{aligned}$$

From which we can compute the weights,

$$\begin{aligned}
w_1 &= 1 + \frac{\sqrt{6}}{9}, \\
w_2 &= 1 - \frac{\sqrt{6}}{9},
\end{aligned}$$

giving the quadrature rule,

$$\Sigma_2^{w,\mathbf{x}}(f) = \left[\left(1 + \frac{\sqrt{6}}{9}\right) f\left(-\frac{1}{5} - \frac{\sqrt{6}}{5}\right) + \left(1 - \frac{\sqrt{6}}{9}\right) f\left(-\frac{1}{5} + \frac{\sqrt{6}}{5}\right)\right]$$

END

3. Gaussian quadrature

Problem 3.1 (C) 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,

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

We can normalise each to get $q_j'(x) = q_j(x)/||q_j||$, with $||q_j||^2 = \int_{-1}^{1} q_j^2 dx$. This gives,

$$q_0'(x) = \frac{1}{\sqrt{2}}$$
$$q_1'(x) = \sqrt{\frac{3}{2}}x$$
$$q_2'(x) = \sqrt{\frac{5}{8}}(3x^2 - 1)$$
$$q_3'(x) = \sqrt{\frac{7}{8}}(5x^3 - 3x)$$

For the first part we use the roots of $q_2(x)$ which are $\mathbf{x} = \left\{\pm\frac{1}{\sqrt{3}}\right\}$. The weights are,

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

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 $q_3(x)$ which are $\mathbf{x} = \left\{0, \pm\sqrt{\frac{3}{5}}\right\}$. 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,

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

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 3.2 (A) Show for $w(x) = 1/\sqrt{1-x^2}$ that the Gaussian quadrature rule is

$$\frac{\pi}{n}\sum_{j=1}^{n} f(x_j)$$

where $x_j = (j - 1/2)\pi/n$ for all n.

SOLUTION

For $w(x) = \frac{1}{\sqrt{1-x^2}}$, the orthogonal polynomials are the Chebyshev polynomials $T_n(x) = \cos(n \arccos(x))$. To make them orthonormal, we have,

$$T_0'(x) = \frac{1}{\sqrt{\pi}},$$
$$T_n'(x) = \frac{2}{\pi}\cos(n \arccos(x)), \qquad (n > 0)$$

We have,

$$\begin{aligned} T_n'(x) = 0 &\Leftrightarrow \cos(n \arccos(x)) = 0 \\ &\Leftrightarrow n \arccos(x) = j\pi - \frac{\pi}{2}, \qquad (j \in \mathbb{Z}) \\ &\Leftrightarrow x = \cos\left(\frac{(j-\frac{1}{2})\pi}{n}\right), \qquad (j \in \mathbb{Z}), \end{aligned}$$

which has unique solutions $\{x_j = \cos((j-\frac{1}{2})\pi/n) : j = 1, \dots, n\}$.

We then have,

$$w_j = \frac{1}{\alpha_j^2} = \frac{1}{T_0'(x_j)^2 + T_1'(x_j)^2 + \dots T_{n-1}'(x_j)^2}$$

Consider, writing $\theta_j = (j-\frac{1}{2})\pi/n$

$$\begin{aligned} \alpha_j^2 &= \sum_{k=0}^{n-1} T_k'(x_j)^2 \\ &= \frac{1}{\pi} + \frac{2}{\pi}\sum_{k=1}^{n-1} \cos^2(k\theta_j) \\ &= \frac{1}{\pi} + \frac{1}{2\pi}\sum_{k=1}^{n-1} (e^{ik\theta_j} + e^{-ik\theta_j})^2 \\ &= \frac{1}{\pi} + \frac{1}{2\pi}\sum_{k=1}^{n-1} (e^{2ik\theta_j} + e^{-2ik\theta_j} + 2) \\ &= \frac{n}{\pi} + \frac{1}{2\pi}\sum_{k=1}^{n-1} (e^{2ik\theta_j} + e^{-2ik\theta_j}) \end{aligned}$$

Using a geometric sum (in essentially the same way as the solution of Problem 1.2 in Problem Sheet 8) we can show that the second term is 0 and thus $w_j = \frac{1}{\alpha_j^2} = \frac{\pi}{n}$.

END

Problem 3.3 (B) Solve Problem 1.2 from PS8 using **Lemma (discrete orthogonality)** with $w(x) = 1/\sqrt{1-x^2}$ on $[-1, 1]$.

SOLUTION

By the Lemma (Discrete Orthogonality), we have,

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

By the previous question, for the weight $w(x) = \frac{1}{\sqrt{1-x^2}}$ we have $q_0(x_j) = \frac{1}{\sqrt{\pi}}$, $q_k(x_j) = \sqrt{\frac{2}{\pi}}\cos(k\theta_j)$. For

$l = m = 0$ then we have,

$$\frac{1}{\pi}\sum_{j=1}^{n}\cos(l\theta_j)\cos(m\theta_j) = \sum_{j=1}^{n} q_l(x_j)q_m(x_j) = \frac{n}{\pi}\delta_{lm} = \frac{n}{\pi}$$
$$\Rightarrow \frac{1}{n}\sum_{j=1}^{n}\cos(l\theta_j)\cos(m\theta_j) = 1$$

Now, for $l = m \neq 0$, we have,

$$\frac{2}{\pi}\sum_{j=1}^{n}\cos(l\theta_j)\cos(m\theta_j) = \sum_{j=1}^{n} q_l(x_j)q_m(x_j) = \frac{n}{\pi}\delta_{lm} = \frac{n}{\pi}$$
$$\Rightarrow \frac{1}{n}\sum_{j=1}^{n}\cos(l\theta_j)\cos(m\theta_j) = \frac{1}{2}$$

Finally, for $l \neq m$, we have,

$$C_{lm}\sum_{j=1}^{n}\cos(l\theta_j)\cos(m\theta_j) = \sum_{j=1}^{n} q_l(x_j)q_m(x_j) = \frac{n}{\pi}\delta_{lm} = 0,$$

for some constant $C_{lm} \neq 0$ which is $\frac{1}{\pi}$ if $l = 0$ or $m = 0$ and $\frac{2}{\pi}$ otherwise (it doesn't matter what it is so long as it is not 0). Therefore, for $l \neq m$ we have,

$$\frac{1}{n}\sum_{j=1}^{n}\cos(l\theta_j)\cos(m\theta_j) = 0$$

END