Theoretical Problems

Numerical analysis 2022

Author: Wenchong Huang (EbolaEmperor)

Institute: School of Mathematical Science, Zhejiang University

Date: September 20th, 2022

ElegantLaTeX Program

Elegantly learning.

Chapter 1 Solving Nonlinear Equations

Problem 1.1 Consider the bisection method starting with the initial interval $[1.5, 3.5]$. In the following questions "the interval" refers to the bisection interval whose width changes across different loops.

- What is the width of the interval at the nth step?
- What is the maximum possible distance between the root r and the midpoint of the interval?

Solution *Note that the interval's width is multipled by $\frac{1}{2}$ at each step, and the initial width is 2, hence the width **after** the nth step is $\frac{1}{2^{n-1}}$.*

The maximum distance is not grater than 1 obviously.

Since the loop terminated when $|f(c)| < \varepsilon$, we could construct an increasing function f whose root is $1.5 + \delta$, and $|f(x)| < \varepsilon$ everywhere, hence the bisection loop will terminate at first step, the distance between midpoint and root is $1 - \delta$. Let $\delta \to 0^+$, we know the distance could be infynitely close to 1.

Problem 1.2 In using the bisection algorithm with its initial interval as $[a_0, b_0]$ with $a_0 > 0$, we want to determine the root with its relative error no grater than ε. Prove that this goal of accuracy is guaranteed by the following choice of the number of steps,

$$n \geq \frac{\log(b_0 - a_0) - \log \varepsilon - \log a_0}{\log 2} - 1$$

Solution *Suppose the root is $r \geq a_0$. The relative error **after** the nth step is*

$$\frac{|r - c_n|}{|r|} \tag{1.1}$$

The following inequations hold

$$\frac{|r - c_n|}{|r|} \leq \frac{\frac{1}{2}(b_n - a_n)}{r} \leq \frac{\frac{1}{2}(b_n - a_n)}{a_0} = \frac{b_0 - a_0}{a_0 2^{n+1}} \tag{1.2}$$

Hence when (1.1) holds, we have

$$\begin{aligned} &(n+1)\log 2 \geq \log(b_0 - a_0) - \log \varepsilon - \log a_0 \\ \implies &\log 2^{n+1} \geq \log\left(\frac{b_0 - a_0}{\varepsilon a_0}\right) \\ \implies &2^{n+1} \geq \frac{b_0 - a_0}{\varepsilon a_0} \implies \frac{b_0 - a_0}{a_0 2^{n+1}} \leq \varepsilon \end{aligned}$$

Hence the conclution is proved by (1.2).

Problem 1.3 Perform four iterations of Newton's method for the polynomial equation $p(x) = 4x^3 - 2x^2 + 3 = 0$ with the starting point $x_0 = -1$. Use a hand calculator and organize results of the iterations in a table.

Solution *Firstly we derivate $p(x)$*

$$p'(x) = 12x^2 - 4x$$

The results are shown as the following table.

n	x_n	$p(x_n)$	$p'(x_n)$	$x_n - \frac{f(x_n)}{f'(x_n)}$
0	-1	-3	16	-0.8125
1	-0.8125	-0.46582	11.1719	-0.770804
2	-0.770804	-0.0201359	10.2129	-0.768832
3	-0.768832	-3.98011e-05	10.1686	-0.768828
4	-0.768828			

Problem 1.4 Consider a variation of Newton's method in which only the derivative at x_0 is used,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \tag{1.3}$$

Find C and s such that

$$e_{n+1} = Ce_n^s$$

where e_n is the error of Newton's method at step n, s is a constant, and C may depend on x_n, the given function f and its derivatives.

Solution *Assume the root is r, then $e_n = x_n - r$. Let $g(x) = f(r+x)$. By (1.3), we derive*

$$e_{n+1} = e_n - \frac{g(e_n)}{g'(e_0)} = \left(1 - \frac{g(e_n)}{e_n g'(e_0)}\right) e_n$$

Let $C(n) = 1 - \frac{g(e_n)}{e_n g'(e_0)}$ and $s = 1$, we got $e_{n+1} = C(n)e_n$, and

$$\lim_{n\to\infty} C(n) = 1 - \frac{g'(0)}{g'(e_0)} = 1 - \frac{f'(r)}{f'(x_0)}$$

Problem 1.5 Within $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, will the iteration $x_{n+1} = \tan^{-1} x_n$ converge?

Solution *As we all know that $0 < \tan^{-1} x < x$ $(x > 0)$, so if $x_0 > 0$, we derive*

$$0 < x_{n+1} = \tan^{-1} x_n < x_n$$

And sequence $\{x_n\}$ has lower bound 0, so $\{x_n\}$ is convergent by monotinic sequence theorem.

For $x_0 < 0$, $\{-x_n\}$ is convergent by the discussion above, hence $\{x_n\}$ is convergent.

For $x_0 = 0$, clearly $x_n = 0$ $(\forall n)$.

Problem 1.6 Let $p > 1$. What is the value of the following continued fraction?

$$x = \frac{1}{p + \frac{1}{p + \frac{1}{p + \cdots}}}$$

Prove that the sequence of values converges.

Solution *We construct a sequence by $x_1 = \frac{1}{p}$ and $x_{n+1} = \frac{1}{p+x_n}$ $(n \geq 1)$, then $x = \lim\limits_{n\to\infty} x_n$ if it exists.*

Consider function $g(x) = \frac{1}{p+x}$, clearly $g(x) \in [0,1]$ for all $x \in [0,1]$. And

$$\lambda = \max_{x\in[0,1]} |g'(x)| = \max_{x\in[0,1]} -\frac{1}{(x+p)^2} = \frac{1}{p^2} < 1$$

Hence g is a contraction in $[0,1]$, and consider equation

$$x = g(x) = \frac{1}{p+x}$$

the roots are $\frac{-p\pm\sqrt{p^2+4}}{2}$, hence g has unique fixed-point $\alpha = \frac{-p+\sqrt{p^2+4}}{2}$ in $[0,1]$.

Recall that $x_1 = \frac{1}{p} \in [0,1]$, and $x_{n+1} = g(x_n)$. By Theorem 1.38, $\{x_n\}$ converges and $x = \lim\limits_{n\to\infty} x_n = \alpha$.

Problem 1.7 What happens in problem 1.2 if $a_0 < 0 < b_0$? Derive an inequality of the number of steps similar to that in problem 1.2. In this case, is the relative error still an appropriate measure?

Solution *In this problem we let the absolutely error $|r - c_n| < \delta$, we derive*

$$|r - c_n| \leq \frac{1}{2}(b_n - a_n) = \frac{b_0 - a_0}{2^{n+1}} \tag{1.4}$$

It is sufficient to let $\frac{b_0-a_0}{2^{n+1}} < \delta$, hence $n \geq \frac{\log(b_0-a_0)-\log\delta}{\log 2} - 1$.

We can't use relative error since r might be zero.

Chapter 2 Polynomial Interpolation

Problem 2.1 For $f \in \mathcal{C}^2[x_0, x_1]$ and $x \in (x_0, x_1)$, linear interpolation of f at x_0 and x_1 yields

$$f(x) - p_1(f; x) = \frac{f''(\xi(x))}{2}(x - x_0)(x - x_1) \tag{2.1}$$

Consider the case $f(x) = \frac{1}{x},\ x_0 = 1,\ x_1 = 2$.

- Determine $\xi(x)$ explicity.
- Extend the domain of ξ continuously from (x_0, x_1) to $[x_0, x_1]$. Find $\max \xi(x)$, $\min \xi(x)$ and $\max f''(\xi(x))$.

Solution

1. *The Lagrange's formula yields*

$$p_1(f; x) = \frac{(x-2)}{(1-2)} + \frac{1}{2} \times \frac{(x-1)}{(2-1)} = -\frac{1}{2}x + \frac{3}{2}$$

Substitute it to (2.1), with $f''(x) = 2x^{-3}$, yield

$$\frac{1}{x} + \frac{1}{2}x - \frac{3}{2} = (x-1)(x-2)\xi^{-3}(x)$$

The result follows from it:

$$\xi(x) = \sqrt[3]{2x}$$

2. *$\xi(x)$ is increasing in $[1, 2]$, hence*

$$\max \xi(x) = \xi(2) = \sqrt[3]{4}, \qquad \min \xi(x) = \xi(1) = \sqrt[3]{2}$$

Also

$$f''(\xi(x)) = 2\left(\sqrt[3]{2x}\right)^{-3} = \frac{1}{x}$$

is decreasing in $[1, 2]$, hence

$$\max f''(\xi(x)) = f''(\xi(1)) = 1$$

Problem 2.2 Let $\mathbb{P}_m^+$ be the set of all polynomials of degree $\leq m$ that are non-negative on the real line,

$$\mathbb{P}_m^+ = \{p : p \in \mathbb{P}_m,\ \forall x \in \mathbb{R},\ p(x) \geq 0\}$$

Find $p \in \mathbb{P}_{2n}^+$ such that $p(x_i) = f_i$ for $i = 0, 1, ..., n$ where $f_i \geq 0$ and x_i are distinct points on $\mathbb{R}$.

Solution *Let $q(x) \in \mathbb{P}_n$ be the unique interpolation polynomial satisfies*

$$q(x_i) = \sqrt{f_i}, \qquad i = 0, 1, ..., n$$

Let $p(x) = q^2(x)$, then $p(x) \in \mathbb{P}_{2n}^+$ and

$$p(x_i) = q^2(x_i) = f_i, \qquad i = 0, 1, ..., n$$

Hence $p(x)$ is what we need. The Lagrange's interpolation formula yields:

$$p(x) = \left(\sum_{i=0}^{n} \sqrt{f_i} \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}\right)^2$$

Problem 2.3 Cnosider $f(x) = e^x$.

- Prove by induction that

$$\forall t \in \mathbb{R}, \qquad f[t, t+1, ..., t+n] = \frac{(e-1)^n}{n!}e^t \tag{2.2}$$

- From Corollary 2.22 we know

$$\exists \xi \in (0, n) \text{ s.t. } f[0, 1, ..., n] = \frac{1}{n!} f^{(n)}(\xi) \tag{2.3}$$

Determine ξ from the above two equations. Is ξ located to the left or to the right of the midpoint $n/2$.

Solution

1. *The Lagrange's formuula yields*

$$p(f; x) = \sum_{k=0}^{n} e^{t+k} \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} = e^t \sum_{k=0}^{n} e^k \frac{(-1)^{n-k} \prod_{j=0,j\neq k}^{n} x - x_j}{k!(n-k)!}$$

Hence

$$f[t, t+1, ..., t+n] = e^t \sum_{k=0}^{n} \frac{(-1)^{n-k} e^k}{k!(n-k)!} = \frac{e^t}{n!} \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} e^k = \frac{(e-1)^n}{n!} e^t$$

2. *Let $t = 0$ in (2.2) and yield*

$$f[0, 1, ..., n] = \frac{(e-1)^n}{n!}$$

Substitute it to (2.3), with $f^{(n)}(x) = e^x$, yield

$$\frac{(e-1)^n}{n!} = \frac{e^\xi}{n!}$$

The result follows from it:

$$\xi = n \ln(e-1) > \frac{n}{2}$$

Hence ξ is located to the right of the midpoint.

Problem 2.4 Consider $f(0) = 5,\ f(1) = 3,\ f(3) = 5,\ f(4) = 12$.

- Use the Newton's formula to obtain $p_3(f; x)$;
- The data suggests that f has a minimum in $x \in (1, 3)$. Find an approximate value for the location x_{min} of the minimum.

Solution

1. *The result follows from Newton's interpolation formula:*

$$p_3(f; x) = 5 - 2x + x(x-1) + \frac{1}{4} x(x-1)(x-3)$$

Transform it into the canonical form:

$$p_3(f; x) = \frac{1}{4} x^3 - \frac{9}{4} x + 5$$

2. *Firstly, calculate the derivative of $p_3(f; x)$:*

$$p_3'(f; x) = \frac{3}{4} x^2 - \frac{9}{4}$$

The first-order necessary condition $p_3'(f; x) = 0$ yields that

$$x_{extreame} = \pm\sqrt{3}$$

In $x \in (1, 3)$, the extreame point might be $x^* = \sqrt{3}$. The second-order condition shows that

$$p_3''(f; x^*) = \frac{3}{2} x^* = \frac{3\sqrt{3}}{2} > 0$$

Hence x^* is the minimum, and $x_{min} = \sqrt{3} \approx 1.73205$.

Problem 2.5 Consider $f(x) = x^7$.

- Compute $f[0, 1, 1, 1, 2, 2]$.

- We konw that this devided difference is expressible in terms of the 5th derivative of f evaluated at some $\xi \in (0,2)$. Determine ξ.

Solution

1. *Solve the Hermite's interpolation with a difference table. The result of Newton's form follows:*
$$p(x) = x + 6x(x-1) + 15x(x-1)^2 + 42x(x-1)^3 + 30x(x-1)^3(x-2)$$
Hence
$$f[0,1,1,1,2,2] = 30$$
2. *The 5th derivate of f is*
$$f^{(5)}(x) = 2520x^2$$
Then $\frac{f^{(5)}(x)}{5!} = f[0,1,1,1,2,2]$ *yields*
$$\frac{2520}{5!}\xi^2 = 30 \qquad \Longrightarrow \qquad \xi = \sqrt{\frac{10}{7}} \approx 1.42857 \in (0,2)$$

Problem 2.6 f is a function on $[0,3]$ for which one knows that
$$f(0) = 1, \quad f(1) = 2, \quad f'(1) = -1, \quad f(3) = f'(3) = 0$$

- Estimate $f(2)$ using Hermite's interpolation.
- Estimate the maximum possible error of the above answer if one konws, in addition, that $f \in \mathcal{C}^5[0,3]$ and $|f^{(5)}(x)| \leq M$ on $[0,3]$. Express the answer in terms of M.

Solution

1. *The Hermite's interpolation gives that*
$$p(x) = 1 + x - 2x(x-1) + \frac{2}{3}x(x-1)^2 - \frac{5}{36}x(x-1)^2(x-3)$$
Hence, estimate $f(2)$ as
$$f(2) \approx p(2) = \frac{11}{18} \approx 0.611111$$
2. *Theorem 2.35 gives that*
$$f(x) - p(x) = \frac{f^{(5)}(\xi)}{120}x(x-1)^2(x-3)^2$$
The result follows directly:
$$|f(2) - p(2)| = \left|\frac{f^{(5)}(\xi)}{60}\right| \leq \frac{M}{60}$$

Problem 2.7 Define foward difference by
$$\Delta f(x) = f(x+h) - f(x), \qquad \Delta^{k+1} f(x) = \Delta\Delta^k f(x) = \Delta^k f(x+h) - \Delta^k f(x)$$
and backward difference by
$$\nabla f(x) = f(x) - f(x-h), \qquad \nabla^{k+1} f(x) = \nabla\nabla^k f(x) = \nabla^k f(x) - \nabla^k f(x-h)$$
Prove
$$\Delta^k f(x) = k!h^k f[x_0, x_1, ..., x_k] \tag{2.4}$$
$$\nabla^k f(x) = k!h^k f[x_0, x_{-1}, ..., x_{-k}] \tag{2.5}$$
where $x_j = x + jh$.

Solution *The Lagrange's interpolation formula yields*

$$f[x_0, x_1, ..., x_k] = \sum_{i=0}^{k} f(x_i)\frac{1}{\prod_{j=1,j\neq i}^{k}(x_i - x_j)} = \sum_{i=0}^{k} \frac{(-1)^{k-i}f(x+ih)}{h^k i!(k-i)!}$$

It yields an equivalent form of (2.4):

$$\Delta^k f(x) = k!h^k f[x_0, x_1, ..., x_k] = \sum_{i=0}^{k} \binom{k}{i}(-1)^{k-i}f(x+ih) \tag{2.6}$$

Now prove (2.6) by an induction. For $k = 1$, it could be verified directly:

$$\binom{1}{0}(-1)^{1-0}f(x) + \binom{1}{1}(-1)^{1-1}f(x+h) = f(x+h) - f(x) = \Delta f(x)$$

Suppose (2.6) holds for some $k \geq 1$, then

$$\begin{aligned}
\Delta^{k+1}f(x) &= \Delta\left(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}f(x+ih)\right)\\
&= \sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}f(x+(i+1)h) - \sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}f(x+ih)\\
&= f(x+(k+1)h) - (-1)^k f(x) + \sum_{i=1}^{k}\left(\binom{k}{i-1}(-1)^{k+1-i}f(x+ih) - \binom{k}{i}(-1)^{k-i}f(x+ih)\right)\\
&= f(x+(k+1)h) + (-1)^{k+1}f(x) + \sum_{i=1}^{k}(-1)^{k+1-i}f(x+ih)\left(\binom{k}{i-1} + \binom{k}{i}\right)\\
&= f(x+(k+1)h) + (-1)^{k+1}f(x) + \sum_{i=1}^{k}\binom{k+1}{i}(-1)^{k+1-i}f(x+ih)\\
&= \sum_{i=0}^{k+1}\binom{k+1}{i}(-1)^{k+1-i}f(x+ih)
\end{aligned}$$

It shows that (2.6) holds for $(k+1)$. Hence (2.4) is proved by induction. Now we prove that

$$\Delta^k f(x) = \nabla^k f(x+kh) \tag{2.7}$$

by an induction. For $k = 1$, it could be verified directly:

$$\Delta f(x) = f(x+h) - f(x) = \nabla f(x+h)$$

Suppose (2.7) holds for some $k \geq 1$, then

$$\begin{aligned}
\Delta^{k+1}f(x) &= \Delta\left(\Delta^k f(x)\right) = \Delta\left(\nabla^k f(x+kh)\right) = \nabla^k f(x+(k+1)h) - \nabla^k f(x+kh)\\
&= \nabla\left(\nabla^k f(x+(k+1)h)\right) = \nabla^{k+1}f(x+(k+1)h)
\end{aligned}$$

Hence (2.7) is proved by induction. Finally, (2.5) follows immediately from (2.4),(2.7) and Corollary 2.15.

Problem 2.8 Assume f is differentiable at x_0. Prove

$$\frac{\partial}{\partial x_0}f[x_0, x_1, ..., x_n] = f[x_0, x_0, x_1, ..., x_n] \tag{2.8}$$

What about the partial derivate with respect to one of the other variables?

Solution *Firstly, follows from Definition 2.34, we have*

$$\frac{\partial}{\partial x_0}f[x_0] = f'(x_0) = f[x_0, x_0]$$

Prove (2.8) by an induction on n. For $n = 1$, verify it directly:

$$
\begin{aligned}
\frac{\partial}{\partial x_0} f[x_0,x_1] &= \frac{\partial}{\partial x_0}\left(\frac{f[x_1]-f[x_0]}{x_1-x_0}\right) \\
&= \frac{-(x_1-x_0)\frac{\partial}{\partial x_0} f[x_0]+f[x_1]-f[x_0]}{(x_1-x_0)^2} \\
&= \frac{f[x_0,x_1]-f[x_0,x_0]}{x_1-x_0} \\
&= f[x_0,x_0,x_1]
\end{aligned}
$$

Suppose (2.8) holds for some $n \geq 1$, then

$$
\begin{aligned}
\frac{\partial}{\partial x_0} f[x_0,x_1,...,x_{n+1}] &= \frac{\partial}{\partial x_0}\left(\frac{f[x_1,...,x_{n+1}]-f[x_0,...,x_n]}{x_{n+1}-x_0}\right) \\
&= \frac{-(x_{n+1}-x_0)\frac{\partial}{\partial x_0} f[x_0,x_1,...,x_n]+f[x_1,...,x_{n+1}]-f[x_0,...,x_n]}{(x_{n+1}-x_0)^2} \\
&= \frac{-(x_{n+1}-x_0) f[x_0,x_0,x_1,...,x_n]+f[x_1,...,x_{n+1}]-f[x_0,...,x_n]}{(x_{n+1}-x_0)^2} \\
&= \frac{-f[x_0,x_0,x_1,...,x_n]+f[x_0,x_1,...,x_{n+1}]}{x_{n+1}-x_0} \\
&= f[x_0,x_0,x_1,...,x_{n+1}]
\end{aligned}
$$

It shows that (2.8) holds for $(n+1)$, hence proved. Morever, the order of $x_0,...,x_n$ is not important, hence

$$
\frac{\partial}{\partial x_j} f[x_0,x_1,...,x_n] = f[x_0,...,x_{j-1},x_j,x_j,x_{j+1},...,x_n], \qquad \forall j=0,...,n
$$

Problem 2.9 (A min-max problem) For $n \in \mathbb{N}^+$, determine

$$
\min \max_{x\in[a,b]} |a_0x^n + a_1x^{n-1} + ... + a_n| \tag{2.9}
$$

where $a_0 \neq 0$ is fixed and the minimum is taken over all $a_i \in \mathbb{R},\ i=1,2,...,n$.

Solution *The map*

$$
p(x) \mapsto q(x) = \frac{1}{a_0} p\left(a + \frac{b-a}{2}(x+1)\right)
$$

yields a bisection relation between polynomials of degree n defines in $[a,b]$ with leading coefficient a_0 and polynomials of degree n defines in $[0,1]$ with leading coefficient 1. Chebyshev's Theorem gives that

$$
\forall q \in \tilde{\mathbb{P}}_n, \qquad \max_{x\in[-1,1]} \left|\frac{T_n(x)}{2^{n-1}}\right| \leq \max_{x\in[-1,1]} |q(x)|
$$

where T_n is the Chebysheve's polynomial of oeder n. Hence the solution of the min-max problem $p_{min}(x)$ satisfies

$$
\frac{1}{a_0} p_{min}\left(a + \frac{b-a}{2}(x+1)\right) = \frac{T_n(x)}{2^{n-1}}
$$

The result follows immediately:

$$
p_{min}(x) = \frac{a_0}{2^{n-1}} T_n\left(\frac{2}{b-a}(x-a)-1\right)
$$

The min value in (2.8) is $\frac{|a_0|}{2^{n-1}}$.

Problem 2.10 (Imitate the proof of Chebyshev's Theorem) Express the Chebyshev polynomial of degree $n \in \mathbb{N}$ as a polynomial T_n and change its domain from $[-1,1]$ to $\mathbb{R}$. For a fixed $a > 1$, define $\mathbb{P}_n^a := \{p \in \mathbb{P}_n : p(a) = 1\}$ and a polynomial $\hat{p}_n(x) \in \mathbb{P}_n^a$,

$$
\hat{p}_n(x) := \frac{T_n(x)}{T_n(a)}
$$

Prove

$$\forall p \in \mathbb{P}_n^a, \qquad ||\hat{p}_n||_\infty \leq ||p||_\infty$$

where the max-norm of a function $f : \mathbb{R} \to \mathbb{R}$ is defined as $||f||_\infty = \max_{x\in[-1,1]} |f(x)|$.

Solution *First we know that* $||\hat{p}_n||_\infty = \frac{1}{|T_n(a)|}$. *And by the property of* T_n *we have*

$$\hat{p}_n(x'_k) = \frac{(-1)^k}{T_n(a)} \quad \textit{for} \quad x'_k = \cos\frac{k}{n}\pi, \; k = 0, 1, ..., n$$

Now we prove the conclution by using reduction to absurdity. Suppose that:

$$\exists p \in \mathbb{P}_n^a, \quad \textit{s.t.} \quad ||p||_\infty < \frac{1}{|T_n(a)|}$$

Let $q(x) = p(x) - \hat{p}_n(x) \in \mathbb{P}_n$, *then* $q(a) = 0$. *And*

$$q(x'_k) = p(x'_k) - \frac{(-1)^k}{T_n(a)}, \quad k = 0, 1, ..., n$$

We have $sgn(q(x'_k)) \neq sgn(q(x'_{k-1}))$ *for* $k = 1, ..., n$ *since* $||p||_\infty < \frac{1}{|T_n(a)|}$. *By the continuity of* q,

$$\exists -1 = x_n < \xi_n < x_{n-1} < ... < x_1 < \xi_1 < x_0 = 1, \quad \textit{s.t.} \quad q(\xi_1) = \cdots = q(\xi_n) = 0$$

However, $q(a) = 0$ *and* $a > 1$ *shows that* q *has at least* $n + 1$ *zero points, that contradict to* $q \in \mathbb{P}_n$.

Problem 2.11 Prove Lemma 2.48:

$$\forall k = 0, 1, ..., n, \forall t \in (0, 1), \quad b_{n,k}(t) > 0 \tag{2.10}$$

$$\sum_{k=0}^{n} b_{n,k}(t) = 1 \tag{2.11}$$

$$\sum_{k=0}^{n} k b_{n,k}(t) = nt \tag{2.12}$$

$$\sum_{k=0}^{n} (k - nt)^2 b_{n,k}(t) = nt(1 - t) \tag{2.13}$$

where

$$b_{n,k}(t) = \binom{n}{k} t^k (1 - t)^{n-k}$$

Solution *(2.10) is clearly since* $t \in (0, 1)$.

By the Binomial Theorem we have:

$$1 = (t + (1 - t))^n = \sum_{k=0}^{n} \binom{n}{k} t^k (1 - t)^{n-k} = \sum_{k=0}^{n} b_{n,k}(t)$$

Hence (2.11) is proved.

Again, by the Binomial Theorem we have:

$$(p + q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$$

Partial derivate with respect to p *to both sides yields:*

$$n(p + q)^{n-1} = \sum_{k=0}^{n} \binom{n}{k} k p^{k-1} q^{n-k}$$

Multiple a p *to both sides, yield*

$$np(p + q)^{n-1} = \sum_{k=0}^{n} \binom{n}{k} k p^k q^{n-k} \tag{2.14}$$

Now take $p = t$ and $q = 1 - t$, yield

$$nt = \sum_{k=0}^{n} \binom{n}{k} k t^k (1-t)^{n-k} = \sum_{k=0}^{n} k b_{n,k}(t)$$

Hence (2.12) is proved.

Follows from (2.14), partial derivate again with respect to p to both sides yields:

$$n(p+q)^{n-1} + n(n-1)p(p+q)^{n-2} = \sum_{k=0}^{n} \binom{n}{k} k^2 p^{k-1} q^{n-k}$$

Multiple a p to both sides, yield

$$np(p+q)^{n-1} + n(n-1)p^2(p+q)^{n-2} = \sum_{k=0}^{n} \binom{n}{k} k^2 p^{k} q^{n-k}$$

Now take $p = t$ and $q = 1 - t$, yield

$$nt + n(n-1)t^2 = \sum_{k=0}^{n} k^2 b_{n,k}(t)$$

By (2.11),(2,12) and the result abouve, we got:

$$\begin{aligned}\sum_{k=0}^{n} (k-nt)^2 b_{n,k}(t) &= \sum_{k=0}^{n} k^2 b_{n,k}(t) - 2nt \sum_{k=0}^{n} k b_{n,k}(t) + (nt)^2 \sum_{k=0}^{n} b_{n,k}(t) \\ &= nt + n(n-1)t^2 - 2(nt)^2 + (nt)^2 = nt - nt^2 = nt(1-t)\end{aligned}$$

Hence (2.13) is proved.

Chapter 3 Splines

Problem 3.1 Consider $s \in \mathbb{S}_3^2$ on $[0, 2]$:

$$s(x) = \begin{cases} p(x) & \text{if} x \in [0,1], \\ (2-x)^3 & \text{if} x \in [1,2]. \end{cases}$$

Determine $p \in \mathbb{P}_3$ such that $s(0) = 0$. Is $s(x)$ a natural cubic spline?

Solution *$p(x)$ should satisfy the following condition:*

$$p(0) = 0, \quad p(1) = 1, \quad p'(1) = -3, \quad p''(1) = 6.$$

Use Hermite interpolation, we got

$$p(x) = 7x^3 - 18x^2 + 12x.$$

$s(x)$ is not a natural cubic spline since $s''(0) = -36 \neq 0$.

Problem 3.2 Given $f_i = f(x_i)$ of some scalar function at points $a = x_1 < x_2 < \cdots < x_n = b$, we consider interpolating f on $[a, b]$ with a quadratic spline $s \in \mathbb{S}_2^1$.

(a) Why is an additional condition needed to determine s uniquely?

(b) Define $m_i = s'(x_i)$ and $p_i = s|_{[x_i, x_{i+1}]}$. Determine p_i in terms of f_i, f_{i+1} and m_i for $i = 1, 2, ..., n-1$.

(c) Suppose $m_1 = f'(a)$ is given. Show how $m_2, m_3, ..., m_{n-1}$ can be computed.

Solution *(a) Denote $p_i = s|_{[x_i, x_{i+1}]} \in \mathbb{P}_2$, then there're $3(n-1)$ unknown coefficients in $p_1, ..., p_{n-1}$. First,*

$$p_i(x_i) = f_i, \quad p_i(x_{i+1}) = f_{i+1}, \quad i = 1, ..., n-1$$

gives $2(n-1)$ equations. And

$$p_i'(x_{i+1}) = p_{i+1}'(x_{i+1}), \quad i = 1, ..., n-2$$

gives $n-2$ equations. Now there're $3(n-1)$ unknowns and $3(n-1)-1$ equations. Hence an additional condition is needed.

(b) Suppose that $p_i(x) = a_i x^2 + b_i x + c_i$. The conditions give that:

$$\begin{cases} x_i^2 a_i + x_i b_i + c_i = f_i \\ x_{i+1}^2 a_i + x_{i+1} b_i + c_i = f_{i+1} \\ 2x_i a_i + b_i = m_i \end{cases}$$

Solve the linear equation of a_i, b_i and c_i, we got

$$\begin{aligned} a_i &= \frac{f_{i+1} - f_i}{(x_{i+1} - x_i)^2} - \frac{m_i}{x_{i+1} - x_i} \\ b_i &= \frac{m_i(x_{i+1} + x_i)}{x_{i+1} - x_i} - \frac{2x_i(f_{i+1} - f_i)}{(x_{i+1} - x_i)^2} \\ c_i &= f_i + \frac{x_i^2(f_{i+1} - f_i)}{(x_{i+1} - x_i)^2} - \frac{m_i x_i x_{i+1}}{x_{i+1} - x_i} \end{aligned}$$

Hence p_i is determined.

(c) Determine p_1 in terms of f_1, f_2 and m_1. Let $m_2 = p_1'(x_2)$.

Determine p_2 in terms of f_2, f_3 and m_2. Let $m_3 = p_2'(x_3)$.

$\vdots$

Determine p_{n-1} in terms of f_{n-1}, f_n and m_{n-1}.

Problem 3.3 Let $s_1(x) = 1 + c(x+1)^3$ where $x \in [-1, 0]$ and $c \in \mathbb{R}$. Determine $s_2(x)$ on $[0, 1]$ such that

$$s(x) = \begin{cases} s_1(x) & \text{if } x \in [-1, 0] \\ s_2(x) & \text{if } x \in [0, 1] \end{cases}$$

is a natural cubic spline on $[-1, 1]$ with knots $-1, 0, 1$. How must c be chosen if one wants $s(1) = -1$?

Solution *Let $s_2(x) = \alpha x^3 + \beta x^2 + \gamma x + \theta$. The following conditions should be satisfied.*

$$s_2(0) = s_1(0) = 1 + c, \quad s_2'(0) = s_1'(0) = 3c, \quad s_2''(0) = s_1''(0) = 6c, \quad s_2(1) = s(1) = -1, \quad s_2''(1) = 0.$$

And these conditions give that:

$$\begin{cases} \theta = 1 + c \\ \gamma = 3c \\ 2\beta = 6c \\ \alpha + \beta + \gamma + \theta = -1 \\ 6\alpha + 2\beta = 0 \end{cases}.$$

Solve the linear system, and we got that $c = -\frac{1}{3}$.

Problem 3.4 Consider $f(x) = \cos\left(\frac{\pi}{2}x\right)$ with $x \in [-1, 1]$.

(a) Determine the natural cubic spline interpolant to f on knots $-1, 0, 1$.

(b) As discussed in the class, natural cubic splines have the minimal total bending energy. Verify this by tanking $g(x)$ be (i) the quadratic polynomial that interpolates f at $-1, 0, 1$, and (ii) $f(x)$.

Solution *(a) The natural cubic spline interpolant to f on knots $-1, 0, 1$ is*

$$s(x) = \begin{cases} -\frac{1}{2}x^3 - \frac{3}{2}x^2 + 1 & \text{if } x \in [-1, 0], \\ \frac{1}{2}x^3 - \frac{3}{2}x^2 + 1 & \text{if } x \in [0, 1]. \end{cases}$$

(b) The bending energy of s is

$$\int_{-1}^{1} [s''(x)]^2 dx = \int_{-1}^{0} (-3x-3)^2 dx + \int_{0}^{1} (3x-3)^2 dx = 6.$$

The quadratic polynomial that interpolates f at $-1, 0, 1$ is

$$p(x) = -x^2 + 1.$$

And its bending energy is

$$\int_{-1}^{1} [p''(x)]^2 dx = \int_{-1}^{1} 4 \, dx = 8 > 6.$$

The bending energy of f is

$$\int_{-1}^{1} [f''(x)]^2 dx = \int_{-1}^{1} \left[-\frac{\pi^2}{4} \cos\left(\frac{\pi}{2}x\right)\right]^2 = \frac{\pi^4}{16} \approx 6.0881 > 6.$$

Problem 3.5 The quadratic B-spline $B_i^2(x)$.

(a) Derive the same explicit expression of $B_i^2(x)$ as that in the notes from the recursive definition of B-splines and the hat function.

(b) Verify that $\frac{\mathrm{d}}{\mathrm{d}x} B_i^2(x)$ is continuous at t_i and t_{i+1}.

(c) Show that only one $x^* \in (t_{i-1}, t_{i+1})$ satisfies $\frac{\mathrm{d}}{\mathrm{d}x} B_i^2(x^*) = 0$. Express x^* in terms of the knots within the interval of support.

(d) Consequently, show $B_i^2(x) \in [0, 1)$.

(e) Plot $B_i^2(x)$ for $t_i = i$.

Solution

(a) *See that*

$$B_i^1(x) = \begin{cases} \frac{x-t_{i-1}}{t_i-t_{i-1}} & x \in (t_{i-1}, t_i], \\ \frac{t_{i+1}-x}{t_{i+1}-t_i} & x \in (t_i, t_{i+1}], \\ 0 & \textit{otherwise}. \end{cases} \qquad B_{i+1}^1(x) = \begin{cases} \frac{x-t_i}{t_{i+1}-t_i} & x \in (t_i, t_{i+1}], \\ \frac{t_{i+2}-x}{t_{i+2}-t_{i+1}} & x \in (t_{i+1}, t_{i+2}], \\ 0 & \textit{otherwise}. \end{cases}$$

And by the recursive definition we have

$$B_i^2(x) = \frac{x - t_{i-1}}{t_{i+1} - t_{i-1}} B_i^1(x) + \frac{t_{i+2} - x}{t_{i+2} - t_i} B_{i+1}^1(x)$$

For $x \in (t_{i-1}, t_i]$,

$$B_i^2(x) = \frac{x - t_{i-1}}{t_{i+1} - t_{i-1}} \cdot \frac{x - t_{i-1}}{t_i - t_{i-1}} + \frac{t_{i+2} - x}{t_{i+2} - t_i} \cdot 0 = \frac{(x - t_{i-1})^2}{(t_{i+1} - t_{i-1})(t_i - t_{i-1})}.$$

For $x \in (t_i, t_{i+1}]$,

$$B_i^2(x) = \frac{(x - t_{i-1})(t_{i+1} - x)}{(t_{i+1} - t_{i-1})(t_{i+1} - t_i)} + \frac{(t_{i+2} - x)(x - t_i)}{(t_{i+2} - t_i)(t_{i+1} - t_i)}.$$

For $x \in (t_{i+1}, t_{i+2}]$,

$$B_i^2(x) = \frac{x - t_{i-1}}{t_{i+1} - t_{i-1}} \cdot 0 + \frac{t_{i+2} - x}{t_{i+2} - t_i} \cdot \frac{t_{i+2} - x}{t_{i+2} - t_{i+1}} = \frac{(t_{i+2} - x)^2}{(t_{i+2} - t_i)(t_{i+2} - t_{i+1})}.$$

Hence we derived

$$B_i^2(x) = \begin{cases} \frac{(x-t_{i-1})^2}{(t_{i+1}-t_{i-1})(t_i-t_{i-1})} & x \in (t_{i-1}, t_i], \\ \frac{(x-t_{i-1})(t_{i+1}-x)}{(t_{i+1}-t_{i-1})(t_{i+1}-t_i)} + \frac{(t_{i+2}-x)(x-t_i)}{(t_{i+2}-t_i)(t_{i+1}-t_i)} & x \in (t_i, t_{i+1}], \\ \frac{(t_{i+2}-x)^2}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})} & x \in (t_{i+1}, t_{i+2}], \\ 0 & \textit{otherwise}. \end{cases} \tag{3.1}$$

(b) *Follows from (3.1), we derived*

$$\frac{d}{dx} B_i^2(x) = \begin{cases} p_1(x) = \frac{2(x-t_{i-1})}{(t_{i+1}-t_{i-1})(t_i-t_{i-1})} & x \in (t_{i-1}, t_i], \\ p_2(x) = \frac{t_{i+1}+t_{i-1}-2x}{(t_{i+1}-t_{i-1})(t_{i+1}-t_i)} + \frac{t_{i+2}+t_i-2x}{(t_{i+2}-t_i)(t_{i+1}-t_i)} & x \in (t_i, t_{i+1}], \\ p_3(x) = \frac{2(x-t_{i+2})}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})} & x \in (t_{i+1}, t_{i+2}], \\ 0 & \textit{otherwise}. \end{cases} \tag{3.2}$$

We have

$$\begin{aligned} p_1(t_i) &= \frac{2(t_i - t_{i-1})}{(t_{i+1} - t_{i-1})(t_i - t_{i-1})} = \frac{2}{t_{i+1} - t_{i-1}} \\ p_2(t_i) &= \frac{t_{i+1} + t_{i-1} - 2t_i}{(t_{i+1} - t_{i-1})(t_{i+1} - t_i)} + \frac{t_{i+2} + t_i - 2t_i}{(t_{i+2} - t_i)(t_{i+1} - t_i)} \\ &= \frac{t_{i-1} - t_i}{(t_{i+1} - t_{i-1})(t_{i+1} - t_i)} + \frac{1}{t_{i+1} - t_{i-1}} + \frac{1}{t_{i+1} - t_i} \\ &= \frac{t_{i-1} - t_i + t_{i+1} - t_{i-1}}{(t_{i+1} - t_{i-1})(t_{i+1} - t_i)} + \frac{1}{t_{i+1} - t_{i-1}} \\ &= \frac{2}{t_{i+1} - t_{i-1}} = p_1(t_i) \end{aligned}$$

Hence $\frac{d}{dx} B_i^2(x)$ *is continuous at* t_i. *Similarly,*

$$\begin{aligned} p_3(t_{i+1}) &= \frac{2(t_{i+1} - t_{i+2})}{(t_{i+2} - t_i)(t_{i+2} - t_{i+1})} = -\frac{2}{t_{i+2} - t_i} \\ p_2(t_{i+1}) &= \frac{t_{i+1} + t_{i-1} - 2t_{i+1}}{(t_{i+1} - t_{i-1})(t_{i+1} - t_i)} + \frac{t_{i+2} + t_i - 2t_{i+1}}{(t_{i+2} - t_i)(t_{i+1} - t_i)} = -\frac{2}{t_{i+2} - t_i} = p_3(t_{i+1}) \end{aligned}$$

Hence $\frac{d}{dx} B_i^2(x)$ *is continuous at* t_{i+1}.

(c) *We konw $\frac{d}{dx}B_i^2(x)$ is continuous, and is a linear function at each interval $(t_{i-1}, t_i]$, $(t_i, t_{i+1}]$ and $(t_{i+1}, t_{i+2}]$. And we have that*

$$\frac{d}{dx}B_i^2(t_{i-1}) = 0, \qquad \frac{d}{dx}B_i^2(t_i) = \frac{2}{t_{i+1} - t_{i-1}} > 0.$$

So by the property of linear function,

$$\frac{d}{dx}B_i^2(x) > 0, \quad x \in (t_{i-1}, t_i]$$

Morever,

$$\frac{d}{dx}B_i^2(t_{i+1}) = -\frac{2}{t_{i+2} - t_i} < 0$$

Hence by the property of linear function, there is unique $x^* \in (t_i, t_{i+1})$ such that $\frac{d}{dx}B_i^2(x^*) = 0$. Follows from (3.2) we have the following equation.

$$\frac{t_{i+1} + t_{i-1} - 2x^*}{t_{i+1} - t_{i-1}} + \frac{t_{i+2} + t_i - 2x^*}{t_{i+2} - t_i} = 0$$

Solve it and we got

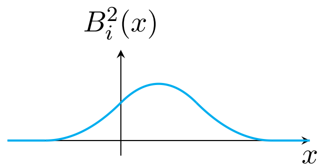
$$x^* = \frac{t_{i+2}t_{i+1} - t_i t_{i-1}}{(t_{i+2} + t_{i+1}) - (t_i + t_{i-1})}.$$

(d) *By (c) we know that:*

$$\frac{d}{dx}B_i^2(x) > 0, \quad x \in (t_{i-1}, x^*)$$
$$\frac{d}{dx}B_i^2(x) < 0, \quad x \in (x^*, t_{i+2})$$

Also $B_i^2(t_{i-1}) = B_i^2(t_{i+2}) = 0$. And $B(x^*) < 1$ could be verified by a trivial computation. Hence $B_i^2(x) \in [0, 1)$.

(e) *Clearly the image of $B_i^2(x)$ with different i could be obtained by translation. So we just draw with $i = 0$.*



Problem 3.6 Verify Theorem 3.32 algebraically for the case of $n = 2$, i.e.

$$(t_{i+2} - t_{i-1})[t_{i-1}, t_i, t_{i+1}, t_{i+2}](t - x)_+^2 = B_i^2(x).$$

Solution *For $x \in (t_{i-1}, t_i]$, by Lagrange's formula we have:*

$$\begin{aligned}
[t_{i-1}, t_i, t_{i+1}, t_{i+2}](t - x)_+^2 =& \frac{(t_i - x)^2}{(t_i - t_{i-1})(t_i - t_{i+1})(t_i - t_{i+2})} + \frac{(t_{i+1} - x)^2}{(t_{i+1} - t_{i-1})(t_{i+1} - t_i)(t_{i+1} - t_{i+2})} \\
&+ \frac{(t_{i+2} - x)^2}{(t_{i+2} - t_{i-1})(t_{i+2} - t_i)(t_{i+2} - t_{i+1})} \\
=& \frac{(x - t_{i-1})^2}{(t_{i+2} - t_{i-1})(t_{i+1} - t_{i-1})(t_i - t_{i-1})} = \frac{B_i^2(x)}{t_{i+2} - t_{i-1}}
\end{aligned}$$

For $x \in (t_i, t_{i+1}]$, by Lagrange's formula we have:

$$\begin{aligned}
[t_{i-1}, t_i, t_{i+1}, t_{i+2}](t - x)_+^2 =& \frac{(t_{i+1} - x)^2}{(t_{i+1} - t_{i-1})(t_{i+1} - t_i)(t_{i+1} - t_{i+2})} + \frac{(t_{i+2} - x)^2}{(t_{i+2} - t_{i-1})(t_{i+2} - t_i)(t_{i+2} - t_{i+1})} \\
=& \frac{B_i^2(x)}{t_{i+2} - t_{i-1}}
\end{aligned}$$

For $x \in (t_{i+1}, t_{i+2}]$, by Lagrange's formula we have:

$$\begin{aligned}[t_{i-1}, t_i, t_{i+1}, t_{i+2}](t-x)_+^2 =& \frac{(t_{i+2}-x)^2}{(t_{i+2}-t_{i-1})(t_{i+2}-t_i)(t_{i+2}-t_{i+1})} \\ =& \frac{B_i^2(x)}{t_{i+2}-t_{i-1}}\end{aligned}$$

Hence we verified

$$(t_{i+2}-t_{i-1})[t_{i-1}, t_i, t_{i+1}, t_{i+2}](t-x)_+^2 = B_i^2(x)$$

in the support of $B_i^2(x)$. And clearly, the equation is also right when $B_i^2(x)$ vanishes.

Problem 3.7 Scaled integral of B-splines.

Deduce from the Theorem on deriviates of B-splines that the scaled integral of a B-spline $B_i^n(x)$ over its support is independent of its index i even if the spacing of the knots is not uniform.

Solution *By the Theorem on derivates of B-splines, we have*

$$\frac{d}{dx}B_i^{n+1}(x) = \frac{(n+1)B_i^n(x)}{t_{i+n}-t_{i-1}} - \frac{(n+1)B_{i+1}^n(x)}{t_{i+n+1}-t_i}, \qquad n = 1, 2, ...$$

Integral to both side, we have:

$$\int_{t_{i-1}}^{t_{i+n+1}} \frac{d}{dx}B_i^{n+1}(x)dx = \int_{t_{i-1}}^{t_{i+n+1}} \frac{(n+1)B_i^n(x)}{t_{i+n}-t_{i-1}} - \frac{(n+1)B_{i+1}^n(x)}{t_{i+n+1}-t_i}dx, \qquad n = 1, 2, ...$$

For the left side, we have:

$$LHS = B_i^{n+1}(t_{i+n+1}) - B_i^{n+1}(t_{i-1}) = 0 - 0 = 0.$$

For the right side, we have:

$$RHS = (n+1)\left(\int_{t_{i-1}}^{t_{i+n}} \frac{B_i^n(x)}{t_{i+n}-t_{i-1}}dx - \int_{t_i}^{t_{i+n+1}} \frac{B_{i+1}^n(x)}{t_{i+n+1}-t_i}dx\right)$$

Then we got

$$\int_{t_{i-1}}^{t_{i+n}} \frac{B_i^n(x)}{t_{i+n}-t_{i-1}}dx = \int_{t_i}^{t_{i+n+1}} \frac{B_{i+1}^n(x)}{t_{i+n+1}-t_i}dx$$

Hence the scaled integral of $B_i^n(x)$ over its support is independent of i.

Problem 3.8 Symmetric Polynomials.

We have a theorem on expressing complete symmetric polynomials as divided difference of monomials.

(a) Verify this theorem for $m = 4$ and $n = 2$ by working out the table of divided difference and comparing the result to the definition of complete symmetric polynomials.

(b) Prove this theorem by the lemma on the recursive relation on complete symmetric polynomials.

Solution

(a) By the definition,

$$\tau_2(x_i, x_{i+1}, x_{i+2}) = x_i^2 + x_{i+1}^2 + x_{i+2}^2 + x_i x_{i+1} + x_i x_{i+2} + x_{i+1}x_{i+2}.$$

Make a table of divided difference as following.

$$\begin{array}{c|lll} x_i & x_i^4 & & \\ x_{i+1} & x_{i+1}^4 & (x_{i+1}^2+x_i^2)(x_{i+1}+x_i) & \\ x_{i+2} & x_{i+2}^4 & (x_{i+2}^2+x_{i+1}^2)(x_{i+2}+x_{i+1}) & \frac{(x_{i+2}^2+x_{i+1}^2)(x_{i+2}+x_{i+1})-(x_{i+1}^2+x_i^2)(x_{i+1}+x_i)}{x_{i+2}-x_i} \end{array}$$

Then the result follows from

$$\begin{aligned}
&\frac{(x_{i+2}^2 + x_{i+1}^2)(x_{i+2} + x_{i+1}) - (x_{i+1}^2 + x_i^2)(x_{i+1} + x_i)}{x_{i+2} - x_i} \\
=&\frac{(x_{i+2}^3 - x_i^3) + x_{i+1}(x_{i+2}^2 - x_i^2) + x_{i+1}^2(x_{i+2} - x_i)}{x_{i+2} - x_i} \\
=&(x_{i+2}^2 + x_{i+2}x_i + x_i^2) + x_{i+1}(x_{i+2} + x_i) + x_{i+1}^2 \\
=&\tau_2(x_i, x_{i+1} + x_{i+2}).
\end{aligned}$$

(b) *By the lemma on recursive relations of complete symmetric polynomials, we have*

$$\begin{aligned}
&(x_{i+n+1} - x_i)\tau_{m-n-1}(x_i, ..., x_{i+n+1}) \\
=&\tau_{m-n}(x_i, ..., x_{i+n+1}) - \tau_{m-n}(x_i, ..., x_{i+n}) - x_i\tau_{m-n-1}(x_i, ..., x_{i+n+1}) \\
=&\tau_{m-n}(x_{i+1}, ..., x_{i+n+1}) + x_i\tau_{m-n-1}(x_i, ..., x_{i+n+1}) - \tau_{m-n}(x_i, ..., x_{i+n}) - x_i\tau_{m-n-1}(x_i, ..., x_{i+n+1}) \\
=&\tau_{m-n}(x_{i+1}, ..., x_{i+n+1}) - \tau_{m-n}(x_i, ..., x_{i+n}).
\end{aligned}$$

Now we prove the theorem by induction. For $n = 0$*, clearly*

$$\tau_m(x_i) = [x_i]x^m = x_i^m.$$

Now we suppose the theorem is true for some $0 \le n < m$*. Then for* $n + 1$*, we have*

$$\begin{aligned}
\tau_{m-n-1}(x_i, ..., x_{i+n+1}) &= \frac{\tau_{m-n}(x_{i+1}, ..., x_{i+n+1}) - \tau_{m-n}(x_i, ..., x_{i+n})}{x_{i+n+1} - x_i} \\
&= \frac{[x_{i+1}, ..., x_{i+n+1}]x^m - [x_i, ..., x_{i+n}]}{x_{i+n+1} - x_i} \\
&= [x_i, ..., x_{i+n+1}]x^m
\end{aligned}$$

Then the theorem is proved by induction.

Chapter 4 Computer Arithmetic

Problem 4.1 Convert the decimal integer 477 to a normalized FPN with $\beta = 2$.

Solution $477 = (111011101)_2 = (1.11011101)_2 \times 2^8$.

Problem 4.2 Convert the decimal fraction $\frac{3}{5}$ to a normalized FPN with $\beta = 2$.

Solution *Calculate by the following table.*

Arithmetic	Decimal Part	Integer Part
$\frac{3}{5} \times 2 = \frac{6}{5}$	$\frac{1}{5}$	1
$\frac{1}{5} \times 2 = \frac{2}{5}$	$\frac{2}{5}$	0
$\frac{2}{5} \times 2 = \frac{4}{5}$	$\frac{4}{5}$	0
$\frac{4}{5} \times 2 = \frac{8}{5}$	$\frac{3}{5}$	1
$\vdots$	$\vdots$	$\vdots$

Hence we got that

$$\frac{3}{5} = (1.0011001\cdots)_2 \times 2^{-1}$$

Problem 4.3 Let $x = \beta^e,\ e \in \mathbb{Z},\ L < e < U$ be a normalized FPN in $\mathbb{F}$ and $x_L, x_R \in \mathbb{F}$ the two normalized FPNs adjacent to x such that $x_L < x < x_R$. Prove $x_R - x = \beta(x - x_L)$.

Solution *We represent x, x_L, x_R in the form of normalized FPN as following.*

$$\begin{aligned} x &= (1.00\cdots0)_\beta \times \beta^e \\ x_L &= ([\beta-1].[\beta-1]\cdots[\beta-1])_\beta \times \beta^{e-1} \\ x_R &= (1.00\cdots01)_\beta \times \beta^e \end{aligned}$$

And hence we have:

$$\begin{aligned} x_R - x &= (0.00\cdots01)_\beta \times \beta^e = \beta^{e-p+1} \\ x - x_L &= (0.00\cdots01)_\beta \times \beta^{e-1} = \beta^{e-p} \end{aligned}$$

That is $x_R - x = \beta(x - x_L)$.

Problem 4.4 By reusing your result of II, find out the two normalized FPNs adjacent to $x = \frac{3}{5}$ under the IEEE 754 single-precision protocol. What is $\mathrm{fl}(x)$ and the relative roundoff error?

Solution *Recall that $x = (1.0011001\cdots)_2 \times 2^{-1}$, find x_L and x_R under IEEE 754 single-precision protocol:*

$$\begin{aligned} x_L &= (1.0011001\ 10011001\ 1001100)_2 \times 2^{-1}, \\ x_R &= (1.0011001\ 10011001\ 1001101)_2 \times 2^{-1}. \end{aligned}$$

We calculate that:

$$\begin{aligned} x - x_L &= (1.10011001\cdots)_2 \times 2^{-23} = \frac{8}{5} \times 2^{-23}, \\ x_R - x_L &= 2^{-22}, \\ x_R - x &= (x_R - x_L) - (x - x_L) = 2^{-22} - \frac{8}{5} \times 2^{-23} = \frac{2}{5} \times 2^{-23}. \end{aligned}$$

Clearly that $x_R - x < x - x_L$, hence $fl(x) = x_R$ and the relative roundoff error is $\frac{|x_R - x|}{|x|} = \frac{2}{3} \times 2^{-23}$.

Problem 4.5 If the IEEE 754 single-precision protocol did not round off numbers to the nearest, but simply dropped excess bits, what would the unit roundoff be?

Solution *It would be* $\epsilon_u^* = \epsilon_M = \beta^{1-p} = 2^{-23}$.

To prove it, we should prove that for $x \in \mathcal{R}(\mathcal{F})$, *we have*

$$fl^*(x) = x(1+\delta), \qquad |\delta| < \epsilon_u^* \tag{4.1}$$

where $fl^*(x)$ *is the approximate of* x *got by the discription of the problem.*

we could find $x_L, x_R \in \mathcal{F}$ *s.t.*

- x_L *and* x_R *are adjacent.*
- $x_L \le x \le x_R$.

If $x = x_L$ *or* $x = x_R$, *then* $fl^*(x) - x = 0$ *and (4.1) clearly holdes. Otherwise* $x_L < x < x_R$. *Then Lemma 4.23, Definition 4.22 yield*

$$|fl^*(x) - x| \le |x_R - x_L| \le \epsilon_u^* \min(|x_L|, |x_R|) < \epsilon_u^* |x|.$$

which yields (4.1). And the upper bound of the error can be reached as $x \to x_{R-}$. *Hence* ϵ_u^* *is the unit roundoff.*

Problem 4.6 How many bits of precision are lost in the subtraction $1 - \cos x$ when $x = \frac{1}{4}$?

Solution *For* $x = \frac{1}{4}$, *we know that* $1 > \cos x$, *hence by the theorem on the loss of most significant digits, we should calculate:*

$$1 - \frac{\cos x}{1} = 0.0310875783 \cdots \in [2^{-6}, 2^{-5}].$$

(The result above is calculated with `long double`, *which is accurate enough.)*

Hence we lost at most 6 *and at least* 5 *significant bits.*

Problem 4.7 Suggest at least two ways to compute $1 - \cos x$ to avoid catastrophic cancellation caused by subtraction.

Solution

(1) We can use Taylor series:

$$\begin{aligned} 1 - \cos x &= 1 - \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\right) \\ &= \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} - \cdots. \end{aligned}$$

(2) We can use sum-to-product identities:

$$1 - \cos x = \cos 0 - \cos x = 2\sin^2\left(\frac{x}{2}\right).$$

Problem 4.8 What are the condition numbers of the following functions? Where are they large?

- $f_1(x) = (x-1)^\alpha$,
- $f_2(x) = \ln x$,
- $f_3(x) = e^x$,
- $f_4(x) = \arccos x$.

Solution

- *We should discuss the value of* α.

  (i) $\alpha \ne 0$. *The condition number of* f_1 *is* $C_{f_1}(x) = \left|\frac{x f_1'(x)}{f_1(x)}\right| = \left|\frac{\alpha x (x-1)^{\alpha-1}}{(x-1)^\alpha}\right| = \left|\frac{\alpha x}{x-1}\right|$.
  Hence $C_{f_1}(x) \to +\infty$ *as* $x \to 1$.

(ii) $\alpha = 0$. *The condition number of* f_1 *is* $C_{f_1}(x) = \left|\frac{xf_1'(x)}{f_1(x)}\right| = \left|\frac{0}{1}\right| = 0$.
Hence $C_{f_1}(x)$ *will never be large.*

- *The condition number of* f_2 *is* $C_{f_2}(x) = \left|\frac{xf_2'(x)}{f_2(x)}\right| = \left|\frac{x\cdot\frac{1}{x}}{\ln x}\right| = \left|\frac{1}{\ln x}\right|$.
  Hence $C_{f_2}(x) \to +\infty$ *as* $x \to 0_+$.
- *The condition number of* f_3 *is* $C_{f_3}(x) = \left|\frac{xf_3'(x)}{f_3(x)}\right| = \left|\frac{xe^x}{e^x}\right| = |x|$.
  Hence $C_{f_3}(x) \to +\infty$ *as* $x \to \pm\infty$.
- *The condition number of* f_4 *is* $C_{f_4}(x) = \left|\frac{xf_4'(x)}{f_4(x)}\right| = \left|\frac{x}{\sqrt{1-x^2}\arccos x}\right|$
  Hence $C_{f_4}(x) \to +\infty$ *as* $x \to \pm 1$.

Problem 4.9 Consider the function $f(x) = 1 - e^{-x}$ for $x \in [0, 1]$.

- Show that $C_f(x) \le 1$ for $x \in [0, 1]$.
- Let A be the algorithm that evaluates $f(x)$ for the machine number $x \in \mathbb{F}$. Assume that the exponential function is computed with relative error within machine roundoff. Estimate $\text{cond}_A(x)$ for $x \in [0, 1]$.
- Plot $\text{cond}_f(x)$ and the estimated upper bound of $\text{cond}_A(x)$ as a function of x on $[0, 1]$. Discuss your results.

Solution

(1) *The condition number of* f *is* $C_f(x) = \left|\frac{xe^{-x}}{1-e^{-x}}\right| = \left|\frac{x}{e^x-1}\right|$.
Notice that $C_f(x)$ *decreases in* $x \in [0, 1]$, *and* $\lim_{x\to 0} C_f(x) = 1$. *Hence* $C_f(x) \le 1$ *for* $x \in [0, 1]$.

(2) *See that*

$$\epsilon_u > |f_A(x) - f(x)| = |f(x_A) - f(x)| = |f'(\xi)| \cdot |x - x_A|, \quad \textit{for } \xi \textit{ between } x \textit{ and } x_A.$$

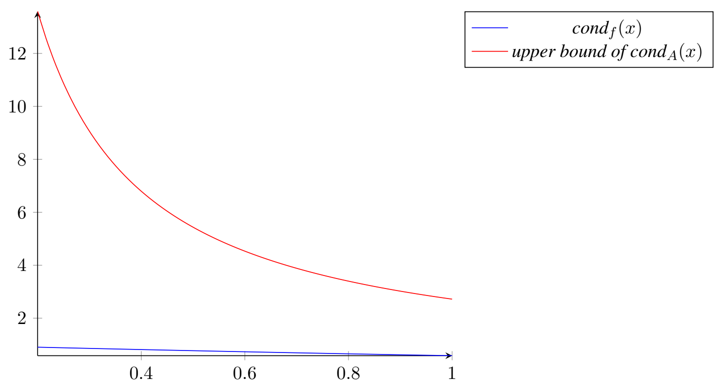
So we have

$$|x - x_A| < \frac{\epsilon_u}{|f'(\xi)|} = \frac{\epsilon_u}{e^{-\xi}} \le e\epsilon_u.$$

Hence by the definition,

$$\textit{cond}_A(x) = \frac{1}{\epsilon_u} \min_{\{x_A\}} \frac{|x - x_A|}{x} < \frac{e}{x}.$$

(3) *See the figure of* $C_f(x)$ *and the upper bound of* $\textit{cond}_A(x)$ *here.*



The upper bound of $\textit{cond}_A(x)$ *goes large as* $x \to 0_+$, *which means calculating* f *with algorithm* A *as* x *is small will cause catastrophic cancellation.*

Problem 4.10 The math problem of root finding for a polynomial

$$q(x) = \sum_{i=0}^{n} a_i x^i, \quad a_n = 1, a_0 \ne 0, a_i \in \mathbb{R} \tag{4.2}$$

can be considered as a vector function $f : \mathbb{R}^n \to \mathbb{C}$:

$$r = f(a_0, a_1, ..., a_{n-1}).$$

Derive the componentwise condition number of f base on the 1-norm. For the Wilkinson example, conpute your condition number, and compare your result with that in the Wilkinson Example. What does the comparision tell you?

Solution *Notice that r satisfies equation*

$$\sum_{i=0}^{n} a_i r^i = 0.$$

Compute partial differetial to each side, we have

$$r^j + \sum_{i=0}^{n} i a_i r^{i-1} \frac{\partial r}{\partial a_j} = 0, \qquad j = 1, 2, ..., n-1.$$

That implies

$$\nabla r = -\frac{(1, r, ..., r^{n-1})}{q'(r)}.$$

Hence we have

$$cond_f(\mathbf{a}) = \frac{||\mathbf{a}||_1 \, ||\nabla r||_1}{|r|} = \frac{\sum_{i=1}^{n-1} |a_i| \cdot \sum_{i=0}^{n-1} |r|^i}{|\sum_{i=0}^{n} i a_i r^i|} \left(\geq \frac{\sum_{i=1}^{n-1} |a_i r^i|}{|\sum_{i=0}^{n} i a_i r^i|} \right)$$

In Wilkinson example, $q(x) = \prod_{i=1}^{n}(x-i)$, $r = n$ is a root. Then we have

$$\sum_{i=0}^{n-1} |a_n r^i| \geq -\sum_{i=0}^{n-1} a_i r^i = r^n = n^n,$$

and

$$\left| \sum_{i=0}^{n} i a_i r^i \right| = |r| \cdot |q'(r)| = n|q'(n)| < n^2 n!.$$

Hence we have $cond_f(\mathbf{a}) \geq \frac{n^{n-2}}{n!}$, which goes $+\infty$ as $n \to \infty$. It supports the Wilkinson Example. And it tells us that finding the root of high-order polynomial equation is very hard.

Problem 4.11 Suppose the division of two FPNs is calculated in a register of precision $2p$. Give an example that contradicts the conclusion of the model of machine arithmetic.

Solution *Consider the FPN system: $\beta = 2, p = 2, L = -1, U = 1$. The number:*

$$a = (1.0)_2 \times 2^0, \quad b = (1.1)_2 \times 2^0.$$

We have the theoretical result:

$$\frac{a}{b} = \frac{2}{3} = (0.101010\cdots)_2.$$

If calculating in a register with precision $2p = 4$, then:

$$fl\left(\frac{a}{b}\right) = fl((0.101)_2) = (0.10)_2 \times 2^0 = 0.5$$

The relative error is

$$\left| \frac{0.5}{\frac{2}{3}} - 1 \right| = 0.25$$

And the machine precision is

$$\epsilon_u = 2^{-2} = 0.25$$

So the result contradicts to the model of machine arithmetic.

Problem 4.12 If the bisection method is used in single precision FPNs of IEEE 754 starting with the interval $[128, 129]$, can we compute the root with absolute accuracy $< 10^{-6}$? Why?

Solution *In this system,* $\epsilon_M = 2^{-23}$. *Hence the distance of two adjacent floating numbers in* $[128, 129]$ *is*

$$2^7 \epsilon_M = 2^{-16} \approx 1.5259 \times 10^{-5} > 2 \times 10^{-6}.$$

Hence we can't compute the root with absolute accuracy less than 10^{-6}.

Problem 4.13 In fitting a curve by cubic splines, one gets inaccurate results when the distance between two adjacent points is much smaller than those of other adjacent pairs. Use the condition number of a matrix to explain this phenomenon.

Solution *Consider calculate* $s(x) = ax^3 + bx^2 + cx + d$ *on* $[x_i, x_{i+1}]$ *by* $s(x_i), s(x_{i+1}), s'(x_i), s'(x_{i+1})$. *We should solve a linear equation where the coefficient matrix is*

$$\begin{pmatrix} x_i^3 & x_i^2 & x_i & 1 \\ x_{i+1}^3 & x_{i+1}^2 & x_{i+1} & 1 \\ 3x_i^2 & 2x_i & 1 & 0 \\ 3x_{i+1}^2 & 2x_{i+1} & 1 & 0 \end{pmatrix}.$$

It has large condition number. So the result will be very inaccurate.