Chapter 1

What is Numerical Analysis about ?

Case 1. Minimizer

Theorem 1. *Suppose* $f : [a,b] \to \mathbb{R}$ *is continuous on* $[a,b]$*. Let* $m := \inf\{f(x) \mid x \in [a,b]\}$*. Then there exists a minimizer* $x_* \in [a,b]$ *such that* $f(x_*) = m$*.*

Note: we have used that a continuous function f on $[a,b]$ is bounded.

Proof. Step 1. For each $n = 1, 2, 3, \cdots$ let $x_n \in [a,b]$ be any element such that

$$f(x_n) < m + \frac{1}{n}.$$

For each n, such an element must exist: Otherwise, i.e., if $f(x) \geq m + \frac{1}{n}$ for every $x \in [a,b]$, then $m + \frac{1}{n}$ is a lower bound of the range set that is strictly bigger than m, contradicting to the definition of m.
Step 2. We have defined the sequence (x_n), and let $y_n := f(x_n)$. Then $y_n \to m$ as $n \to \infty$ because

$$0 \leq |y_n - m| < \frac{1}{n}.$$

Step 3. (x_n) is a sequence of real numbers in $[a,b]$, and thus (x_n) is a bounded sequence. Therefore, there exists a convergent subsequence $(x_{n_k})_{k=1}^{\infty}$ to a limit x_*. Furthermore, because $[a,b]$ is closed, the limit point $x_* \in [a,b]$. We also know that the subsequence (y_{n_k}) of a convergent sequence (y_n) also is convergent to the same limit m.
Step 4. Now, we observe

$$m = \lim_{k\to\infty} y_{n_k} = \lim_{k\to\infty} f(x_{n_k}) = f(x_*),$$

where in the last equality, we used the fact that f is continuous on $[a,b]$. ☐

Case 2 : Riemann Integral or Quadrature

Let us consider high-school series calculations. Let n be a fixed natural number. For each i, we let $x_i = a + i\frac{b-a}{n}$. We let t_i be any point belogs to $[x_{i-1}, x_i]$ for $i = 1, \cdots n$. The definition of the Riemann Integral of f over $[a, b]$ is justified by the existence of a real number I such that

$$\lim_{n\to\infty} \sum_{i=1}^{n} f(t_i)\ \frac{b-a}{n} = I.$$

regardless of choice (t_i) at each n. We examine the proof of the assertion.

Proof. Step 1. By extreme value theorem on each compact interval $[x_{i-1}, x_i]$, there exist the minimizer t_{i*} and t_i^* in $[x_{i-1}, x_i]$. Then we have for each n

$$\sum_{i=1}^{n} f(t_{i*})\ \frac{b-a}{n} \le \sum_{i=1}^{n} f(t_i)\ \frac{b-a}{n} \le \sum_{i=1}^{n} f(t_i^*)\ \frac{b-a}{n}.$$

Step 2. We omit the proof that $\left(\sum_{i=1}^{n} f(t_{i*})\ \frac{b-a}{n}\right)_{n=1}^{\infty}$ and $\left(\sum_{i=1}^{n} f(t_i^*)\ \frac{b-a}{n}\right)_{n=1}^{\infty}$ are bounded monotone sequences so that their respective limits $\alpha, \beta \in \mathbb{R}$, $\alpha \le \beta$ exist.
Step 3. That f is continuous in $[a, b]$ implies the following:

For every $\epsilon > 0$, there exists $\delta > 0$ such that

$$x, x' \in [a, b], |x - x'| < \delta \quad \Longrightarrow \quad |f(x) - f(x')| < \frac{\epsilon}{b-a}.$$

that is to say, f is uniformly continuous in $[a, b]$. For a given ϵ, we take $K \in \mathbb{N}$ so that $\frac{b-a}{K} < \delta$ for each δ. This gives that

for every ϵ, there exists K such that for $n \ge K$,

$$\Big|\sum_{i=1}^{n} f(t_i^*)\ \frac{b-a}{n} - \sum_{i=1}^{n} f(t_{i_*})\ \frac{b-a}{n}\Big| \le \sum_{i=1}^{n} \Big|f(t_i^*) - f(t_{i*})\Big|\ \frac{b-a}{n} \le \sum_{i=1}^{n} \frac{\epsilon}{b-a}\ \frac{b-a}{n} = \epsilon.$$

This implies that the limits $\alpha = \beta$. By the Squeeze theorem, $\left(\sum_{i=1}^{n} f(t_i)\ \frac{b-a}{n}\right)_{n=1}^{\infty}$ is convergent, to the same limit. The limit is defiend to be I. $\square$

We will use quite much better working numerical methods (midpoint rule, simpson rule, gaussian quadrature, $\cdots$) than the one suggested in the proof.

Case 3 : rank of a matrix

Let A be an $n \times k$ real matrix of $\mathbb{R}^{n\times k}$. That the very first fundamental notion for A is the rank of A, the number being defined by the fact that

the number of linearly independent columns = the number of linearly independent rows.

Denote the $(LHS) = r_{col}(A)$ and $(RHS) = r_{row}(A)$. This fundamental equality is a consequence of the inequality that

$$\text{for any matrix } A, \quad r_{row}(A) \le r_{col}(A)$$

because if that is true then $\Big(r_{col}(A) = \Big) r_{row}(A^T) \le r_{col}(A^T)\Big(= r_{row}(A)\Big)$.
The inequality follows from the assertion

if $r_{i_1}, r_{i_2}, \cdots r_{i_\rho}$ are linearly independent rows of A
then $Ar_{i_1}, Ar_{i_2}, \cdots Ar_{i_\rho}$ are linearly independent vectors in $\mathbb{R}^n$.

Here we denote the rows of A from top to bottom by $r_1,\ r_2,\ \cdots,\ r_n$. We examine the proof of the assertion.

Proof. Suppose not, i.e., we have $(\lambda_1, \lambda_2, \cdots, \lambda_\rho) \neq 0$ such that $\sum_{\alpha=1}^{\rho} \lambda_\alpha \big(Ar_{i_\alpha}\big) = 0$. Then,

$$0 = \sum_{i=\alpha}^{\rho} \lambda_\alpha (Ar_{i_\alpha}) = \sum_{\alpha=1}^{\rho} A(\lambda_\alpha r_{i_\alpha}) = A\Big(\sum_{\alpha=1}^{\rho} \lambda_\alpha r_{i_\alpha}\Big).$$

But,

$$Ax = 0 \quad \Longleftrightarrow \quad x \perp \mathsf{span}\, \big\langle r_1, r_2, \cdots, r_n \big\rangle.$$

Combining the two observations,

$$\sum_{\alpha=1}^{\rho} \lambda_\alpha r_{i_\alpha} \in \mathsf{span}\, \big\langle r_1, r_2, \cdots, r_n \big\rangle \cap \mathsf{span}\, \big\langle r_1, r_2, \cdots, r_n \big\rangle^{\perp} = \{0\}.$$

Thus, we conclude that $\sum_{\alpha=1}^{\rho} \lambda_\alpha r_{i_\alpha}$ is a zero vector. Recalling that $r_{i_1}, r_{i_2}, \cdots r_{i_\rho}$ are linearly independent, we conclude $\lambda_1 = \lambda_2 = \cdots = \lambda_\rho = 0$. Contradiction. $\square$

We have to wait until we prove that any A admits a Singular Value Decomposition to compute the rank of A.

Case 4 : Jordan factorization of a square matrix T.

Another striking example where we do not have an algorithm for it is the Jordan factorization.

We will not prove the statement here, that any square matrix T admits a Jordan factorization $T = PJP^{-1}$. But you will see in its proof that it does not provide a constructive method. There is a fundamental obstacle that prevents us to compute it in general case.

Any pair of (theory,method) from variety of mathematical areas.

What will we treat in this course ?

1. Existence of interpolating / approximating polynomials for a class of functions. Topics: spline, Tchebyshev polynomials, $\cdots$. 1d, multi-d.
2. Existence of fixed point and Newton's method on nonlinear system.
3. Existence of solution of initial value problem of nonlinear ode / Runge-Kutta methods, $\cdots$
4. For all of above, error analysis methodology.
5. (2nd semester) Linear algebra theory
6. (2nd semester) more on (partial) differential equations.

Chapter 2

How does my computer throw $\sin(x)$ for an input x?

Case 1. A function from $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ to $\mathbb{R}$

We can define a function $a : \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \mapsto \mathbb{R}$, using the table.

n	1	2	3	4	5	6	7	8	9	10
a_n										

Case 2. A function from $\mathbb{N}$ to $\mathbb{R}$

Are there ways to define a function from countably infinite set to $\mathbb{R}$?

Case 3. A function from $[0, 1] \subset \mathbb{R}$ to $\mathbb{R}$

Are there ways to define a function from this uncountably infinite set $[0, 1]$ to $\mathbb{R}$?

A function on $E \subset \mathbb{R}^n$?

A function $f : X \mapsto Y$ is a correspondence $G \subset X \times Y$ satisfying the rule that

For every $x \in X$, there exists unique $y \in Y$ so that $(x, y) \in G$.

Is a function a black box ?
Is a function a table ?
Can you come up with an example of a function that you can build as a machine ?
How does my computer know which value to return for $\sin(x)$?

I wanted you to see the importance of a class of functions on $\mathbb{R}^n$ that are

(1) piecewise polynomial, and

(2) each piece's domain (the support more precisely) is elementary in the sense that you can finish the membership check of a given x to the domain in a finite procedures.

Examples of sets $E \subset \mathbb{R}^n$, where the membership check is finitely done.

We will call those functions an elementary pp function or a good pp function.

Remark 2.1. The literal definition of the function on an uncountable set, which might suggest the tabularized mapping, indeed is beyond our intuition. If a function is continuous or measurable, we have excuses.

Excuses of not doing Uncountable Table

1. $f(x)$, the value to return, is calculated "on the fly":
   (a) $id(x)$ is identified on the fly, and a few arithmetic calculations involving it are processed afterwards.
   (b) The power p, for the term x^p, is limited to the Natural numbers (or Integers) so that we can operate
   $$id(x) \times id(x) \times \cdots \times id(x) \quad p \text{ times.}$$
   n-th root, in general, cannot be computed in a finite procedure.
2. irrational input, irrational coefficient are not processed.

Excuses by "Limit"

1. For irrational input, irrational coefficient cases, non-elementary piece's domain (support), the value to return can be reached by the limit procedure.

2. More importantaly,

   A continuous function on $\mathbb{R}^n$ is **a limit of continuous elementary piecewise polynomial functions**, where the convergence is uniform in every compact subset of $\mathbb{R}^n$.

   A measurable function on $\mathbb{R}^n$ is **a limit of elementary piecewise polynomial functions**, where the convergence is pointwise, for almost everywhere $x \in \mathbb{R}^n$.

3. Therefore, in principle, a suitable quantity can be computed as accurate as we want.

Objective of our study

For a given function f, we come up with a pp function φ and seek for a formula of the remainder

$$R(x) = f(x) - \varphi(x).$$

1. How do we choose a pp function φ? $\rightarrow$ theory and method
2. How accurate is it ? $\rightarrow$ The remainder formula.

For these two questions,

First, we will take one piece, out of piecewise polynomial function, and seek for the remainder formular, restricted on the piece's support. (Local problem)

Second, we assemble pieces into one pp function. (Global problem)

(Local) polynomial interpolation

Let $I = [a, b]$. We consider

$$C^{n+1}([a,b]) = \big\{ f : [a,b] \to \mathbb{R} \mid f \text{ is } n+1 \text{ times differentiable and} \\ f,\, f',\, f'',\, \cdots,\, f^{(n+1)} \text{ are continuous in } [a,b] \big\}.$$

Note: At the endpoint a and b, the differentibility is the right-differentiability and the left-differentiability respectively.

We consider the following interpolation problem:

1. Suppose we sample the data $(x_0, f(x_0))$, $\cdots$, $(x_n, f(x_n))$ for $f \in C^{n+1}([a,b])$ and $x_0, \cdots, x_n \in [a,b]$.
2. We look for a polynomial p on $[a,b]$ with

$$p(x_i) = f(x_i) \quad i = 0, 1, 2, \cdots n.$$

3. We seek for a formula of the remainder $R(x)$,

$$R(x) = f(x) - p(x), \quad x \in [a,b]$$

In fact, the term *interpolation* does not necessarily assume the function f and does not assume that the data are collected by sampling. Nevertheless, it is convenient to speak of the function f.

We are going to prove the far-reaching remainder theorem in the next lecture, which is much stronger than the Taylor's theorem. We recall first the fundamental theorem of calculus:

Theorem 1. *Let $f \in C^1([a,b])$ and $x_0 \in [a,b]$. Then for $x \in [a,b]$,*

$$R(x) = \int_{x_0}^{x} \frac{df}{dx}(x')\,dx' = (x - x_0)\int_{\tau=0}^{1} \frac{df}{dx}\big((1-\tau)x_0 + \tau x\big)\,d\tau = f(x) - f(x_0).$$

Theorem 2 (1d remainder theorem)**.** *Let $f \in C^{n+1}([a,b])$ and $x_0,\ x_1,\ \cdots,\ x_n \in [a,b]$. Then, for $x \in [a,b]$,*

$$\begin{aligned} R(x) &= \frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})(x-x_n)}{\sqrt{n+2}} \int_{\Lambda^{n+1}\subset\mathbb{R}^{n+2}} \frac{d^{n+1}f}{dx^{n+1}}\big(\lambda_0 x_0 + \cdots \lambda_n x_n + \lambda_{n+1}x\big)\,dS(\lambda) \\ &= f(x) - \Big(f[x_0] + f[x_0,x_1](x-x_0) + f[x_0,x_1,x_2](x-x_0)(x-x_1) + \cdots \\ &\quad + f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})\Big) \end{aligned}$$

Remark 2.2. We will introduce the definition, which is the integral expression, of

$$f[x_0,x_1,\cdots,x_k], \Lambda^k, \cdots$$

Remark 2.3. We will see that the Taylor's theorem is a special case where $x_0 = x_1 = \cdots = x_n = \bar{x}$.

Theorem 3 (Taylor's theorem)**.** *Let $f \in C^{n+1}([a,b])$ and $x_0 \in [a,b]$. Then for $x \in [a,b]$,*

$$\begin{aligned} R(x) &= (x-x_0)^{n+1} \times \frac{1}{n!}\int_{\tau=0}^{1} (1-\tau)^n \frac{d^{n+1}f}{dx^{n+1}}\Big((1-\tau)x_0 + \tau x\Big)\,d\tau \\ &= f(x) - \Big(f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n\Big). \end{aligned}$$

Chapter 3

Divided Differences

Let $I = [a, b]$, $n \in \mathbb{N}$, and consider $k + 1$ points ($k \in \{0, 1, 2, \cdots, n + 1\}$)

$$x_0, x_1, x_2, \cdots, x_k \in [a, b] \quad \text{(they may be repeated).}$$

We will denote the choice of $(k + 1)$ points by

$$[x_0, x_1, \cdots, x_k].$$

A choice $[x_0, \cdots, x_k]$ acts on the class $C^{n+1}([a, b])$: For $f \in C^{n+1}([a, b])$,

$$\begin{aligned} &\text{if } k = 0 \qquad && \big([x_0], f\big) = f(x_0), \\ &\text{if } k \geq 1 && \big([x_0, x_1, \cdots, x_k], f\big) := \frac{1}{\sqrt{k+1}} \int_{\Lambda^k \subset \mathbb{R}^{k+1}} \frac{d^k f}{dx^k} \big(\lambda_0 x_0 + \cdots \lambda_k x_k\big)\, dS(\lambda). \end{aligned}$$

$$\text{The set } \Lambda^k = \Big\{(\lambda_0, \lambda_1, \cdots, \lambda_k) \in \mathbb{R}^{k+1} \mid \forall e \ \ \lambda_e \geq 0, \quad \sum_{e=0}^{k} \lambda_e = 1\Big\} \subset \mathbb{R}^{k+1}.$$

The integral is denoted by $\big([x_0, x_1, \cdots, x_k], f\big)$, or $[x_0, \cdots, x_k]f$, or $f[x_0, x_1, \cdots, x_k]$.

Properties of the Integral $f[x_0, x_1, \cdots, x_k]$

(1) The integral is independent of orders in the $[x_0, x_1, \cdots, x_k]$.

(2) $k = 1$ cases:

$$\begin{aligned} &\text{if } x_0 \neq x_1, && f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0) - f(x_1)}{x_0 - x_1}, \\ &\text{if } x_0 = x_1 = \bar{x}, && f[\bar{x}, \bar{x}] = f'(\bar{x}). \end{aligned}$$

(3) For $k \geq 1$, the equality always holds:

$$f[x_0, x_1, \cdots, x_k]\ (x_k - x_0) = f[x_1, x_2, \cdots, x_k] - f[x_0, x_1, \cdots, x_{k-1}].$$

Indeed, above two $k = 1$ cases lead to that

$$f[x_0, x_1](x_1 - x_0) = f[x_1] - f[x_0].$$

Change of variables formula for the integral

For $k \geq 1$,

$$\text{The set } \tau^k = \big\{(\tau_1, \tau_2, \cdots, \tau_k) \in \mathbb{R}^k \mid 0 \leq \tau_1 + \cdots + \tau_k \leq 1\big\}.$$

For $k \geq 1$,

$$\begin{aligned}
f[x_0, x_1, \cdots, x_k] =& \frac{1}{\sqrt{k+1}} \int_{\Lambda^k \subset \mathbb{R}^{k+1}} \frac{d^k f}{dx^k}\big(\lambda_0 x_0 + \cdots + \lambda_k x_k\big)\, dS(\lambda) \\
\text{(change of variable)} \quad =& \int_{\tau^k \subset \mathbb{R}^k} \frac{d^k f}{dx^k}\big((1 - \tau_1 - \tau_2 - \cdots - \tau_k)x_0 + \tau_1 x_1 + \cdots + \tau_k x_k\big)\, d\mathcal{L}^k(\tau) \\
\text{(Fubini)} \quad =& \int_{\tau_1=0}^{1} \int_{\tau_2=0}^{1-\tau_1} \int_{\tau_3=0}^{1-\tau_1-\tau_2} \cdots \int_{\tau_k=0}^{1-\tau_1-\cdots-\tau_{k-1}} \\
& \frac{d^k f}{dx^k}\big((1 - \tau_1 - \tau_2 - \cdots - \tau_k)x_0 + \tau_1 x_1 \cdots \tau_k x_k\big)\, d\tau_k d\tau_{k-1} \cdots d\tau_1
\end{aligned}$$

For the change of variable, the mapping from τ^k to Λ^k is

$$(\tau_1, \tau_2, \cdots, \tau_k) \quad \mapsto \quad \big(\lambda_0, \lambda_1, \cdots, \lambda_k\big) = \big((1 - \tau_1 - \tau_2 - \cdots - \tau_k), \tau_1, \cdots, \tau_k\big).$$

The linear mapping from Λ^k to $[a, b]$ is

$$x(\lambda) = \lambda_0 x_0 + \lambda_1 x_1 + \cdots + \lambda_k x_k \in [a, b].$$

The linear mapping from τ^k to $[a, b]$ is

$$\tilde{x}(\tau) = (1 - \tau_1 - \tau_2 - \cdots - \tau_k)x_0 + \tau_1 x_1 + \cdots + \tau_k x_k \in [a, b].$$

We can also compute

$$\int_{\tau_1=0}^{1} \int_{\tau_2=0}^{1-\tau_1} \int_{\tau_3=0}^{1-\tau_1-\tau_2} \cdots \int_{\tau_k=0}^{1-\tau_1-\cdots-\tau_{k-1}} 1\, d\tau_k d\tau_{k-1} \cdots d\tau_1 = \frac{1}{k!}.$$

Proof of the Key equality

Lemma 1. *Let* $f \in C^{n+1}([a,b])$*. Then for* $k \in \{1, 2, \cdots, n+1\}$ *and* $x_0,\ x_1,\ \cdots,\ x_k \in [a,b]$,

$$f[x_0, x_1, \cdots, x_k]\ (x_k - x_0) = f[x_1, x_2, \cdots, x_k] - f[x_0, x_1, \cdots, x_{k-1}].$$

Proof. Thanks to the order independence,

$$\begin{aligned}
(RHS) &= f[x_k, x_1, x_2, \cdots, x_{k-1}] - f[x_0, x_1, x_2, \cdots, x_{k-1}] \\
&= \int_{\tau_1=0}^{1} \int_{\tau_2=0}^{1-\tau_1} \int_{\tau_3=0}^{1-\tau_1-\tau_2} \cdots \int_{\tau_{k-1}=0}^{1-\tau_1-\cdots-\tau_{k-2}} \\
&\quad \frac{d^{k-1} f}{dx^{k-1}} \big((1-\tau_1-\tau_2-\cdots-\tau_{k-1})x_k + \tau_1 x_1 + \tau_2 x_2 + \cdots + \tau_{k-1} x_{k-1}\big) \\
&\quad - \frac{d^{k-1} f}{dx^{k-1}} \big((1-\tau_1-\tau_2-\cdots-\tau_{k-1})x_0 + \tau_1 x_1 + \tau_2 x_2 + \cdots + \tau_{k-1} x_{k-1}\big)\, d\tau_{k-1} \cdots d\tau_1, \\
&=: A.
\end{aligned}$$

The integrand is

$$\left[\frac{d^{k-1} f}{dx^{k-1}} \big((1-\tau_1-\tau_2-\cdots-\tau_{k-1}-\tau_k)x_0 + \tau_1 x_1 + \tau_2 x_2 + \cdots + \tau_{k-1} x_{k-1} + \tau_k x_k\big)\right]_{\tau_k=0}^{1-\tau_1-\tau_2-\cdots-\tau_{k-1}}$$

For each fixed $\tau_1, \tau_2, \cdots, \tau_{k-1}$, the map

$$\tau_k \mapsto (1-\tau_1-\tau_2-\cdots-\tau_{k-1}-\tau_k)x_0 + \tau_1 x_1 + \tau_2 x_2 + \cdots + \tau_{k-1} x_{k-1} + \tau_k x_k$$

from the interval $[0, 1-\tau_1-\tau_2-\cdots-\tau_{k-1}]$ to $[a,b]$ is a linear map, and $f \in C^{n+1}[a,b]$. Thus, for $k \in \{1, 2, \cdots, n+1\}$ the composition function

$$\tau_k \mapsto \frac{d^{k-1} f}{dx^{k-1}} \big((1-\tau_1-\tau_2-\cdots-\tau_{k-1}-\tau_k)x_0 + \tau_1 x_1 + \tau_2 x_2 + \cdots + \tau_{k-1} x_{k-1} + \tau_k x_k\big)$$

is a continuously differentiable function in the interval $[0, 1-\tau_1-\tau_2-\cdots-\tau_{k-1}]$, fulfilling the assumption of the Fundamental Theorem of Calculus. Therefore, the integrand is

$$\begin{aligned}
&\int_{\tau_k=0}^{1-\tau_1-\tau_2-\cdots-\tau_{k-1}} \frac{d}{d\tau_k} \frac{d^{k-1} f}{dx^{k-1}} \big((1-\tau_1-\tau_2-\cdots-\tau_{k-1}-\tau_k)x_0 + \tau_1 x_1 + \cdots + \tau_k x_k\big)\, d\tau_k \\
&= \int_{\tau_k=0}^{1-\tau_1-\tau_2-\cdots-\tau_{k-1}} \frac{d}{dx} \frac{d^{k-1} f}{dx^{k-1}} \big((1-\tau_1-\tau_2-\cdots-\tau_{k-1}-\tau_k)x_0 + \tau_1 x_1 + \cdots + \tau_k x_k\big)\, d\tau_k (x_k - x_0).
\end{aligned}$$

Substituting this expression in place of the integrand in A, we see that

$$A = f[x_0, x_1, \cdots, x_k](x_k - x_0).$$

$\square$

Proof of 1d remainder theorem

All we need is the key equality

$$f[x_0, x_1, \cdots, x_k]\ (x_k - x_0) = f[x_1, x_2, \cdots, x_k] - f[x_0, x_1, \cdots, x_{k-1}].$$

For $k \in \{0, 1, 2, \cdots, n\}$, we let

$$\begin{aligned} p_k(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ &\quad + f[x_0, x_1, \cdots, x_k](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{k-1}). \end{aligned}$$

Theorem 2 (1d remainder theorem)**.** *Let $f \in C^{n+1}([a, b])$ and $x_0,\ x_1,\ \cdots,\ x_n \in [a, b]$. Then, for $x \in [a, b]$,*

$$\begin{aligned} R(x) &= f[x_0, x_1, x_2, \cdots, x_n, x](x - x_0)(x - x_1) \cdots (x - x_{n-1})(x - x_n) \\ &= f(x) - \Big(f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ &\quad + f[x_0, x_1, \cdots, x_n](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{n-1}) \Big). \end{aligned}$$

Proof. We assert that for $k \in \{0, 1, 2, \cdots, n\}$,

$$f(x) - p_k(x) = f[x_0, x_1, \cdots, x_k, x](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_k).$$

At $k = 0$, by the equality

$$f(x) - p_0(x) = f(x) - f(x_0) = f[x_0, x](x - x_0).$$

Now, if the assertion is true for $0, 1, 2, \cdots, k - 1$,

$$\begin{aligned} f(x) - p_k(x) &= f(x) - p_{k-1}(x) - f[x_0, x_1, \cdots, x_k](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{k-1}) \\ &= f[x_0, x_1, \cdots, x_{k-1}, x](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{k-1}) \\ &\quad - f[x_k, x_0, x_1, \cdots, x_{k-1}](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{k-1}) \\ &= f[x_0, x_1, \cdots, x_{k-1}, x_k, x]\Big((x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{k-1})(x - x_k)\Big). \end{aligned}$$

$\square$

Mean Value type Theorem

Although we will not use this much later, it is worth to know that the mean value of the integrand is attained.

Theorem 3. *Let $f \in C^{n+1}([a,b])$, $k \in \{0,1,2,\cdots,n+1\}$, and $x_0, x_1, x_2, \cdots, x_k \in [a,b]$. Then there exists $\xi \in [a,b]$ such that*

$$f[x_0, x_1, \cdots, x_k] = \frac{1}{k!}\frac{d^k f}{dx^k}(\xi).$$

Remark 3.1. In other words, the attained value $\beta = \frac{d^k f}{dx^k}(\xi)$ at $\xi \in [a,b]$ is the mean value so that

$$\int_{\tau^k \subset \mathbb{R}^k} \frac{d^k f}{dx^k}\big(\tilde{x}(\tau)\big)\, d\mathcal{L}^k(\tau) = \left(\int_{\tau^k \subset \mathbb{R}^k} 1\, d\mathcal{L}^k(\tau)\right) \times \beta.$$

Proof. For $\lambda \in \Lambda^k$, the linear function

$$x(\lambda) = \lambda_0 x_0 + \lambda_1 x_1 + \cdots + \lambda_k x_k \in [a,b]$$

is continuous and thus $\lambda \mapsto \frac{d^k f}{dx^k} \circ x(\lambda)$ is a continuous function on the compact set Λ^k. Hence, there exist λ^*, $\lambda_* \in \Lambda^k$, the maximizer and the minimizer.
If $\lambda_* = \lambda^*$, then $\frac{d^k f}{dx^k} \circ x$ must be a constant function over Λ^k. Thus, at $\xi = x(\bar{\lambda})$ for any $\bar{\lambda} \in \Lambda^k$, the integral

$$(LHS) = \frac{d^k f}{dx^k}(\xi)\frac{1}{\sqrt{k+1}}\int_{\Lambda^k} 1\, dS(\lambda) = \frac{1}{k!}\frac{d^k f}{dx^k}(\xi)$$

and we are done.
Now, we assume $\lambda_* \neq \lambda^*$. Note that the convex set Λ^k contains the line segment joining λ_* and λ^*. Therefore, we can define a function

$$h(t) : [0,1] \ni t \mapsto \frac{d^k f}{dx^k} \circ x\big((1-t)\lambda_* + t\lambda^*\big).$$

h is a continuous function on $[0,1]$. By intermediate value theorem, every values in the range $[h(0), h(1)] = \big[\frac{d^k f}{dx^k}(\lambda_*), \frac{d^k f}{dx^k}(\lambda^*)\big]$ is attained at some $t \in [0,1]$.
Certainly,

$$h(0) = \frac{d^k f}{dx^k}\big(x(\lambda_*)\big) \leq \frac{d^k f}{dx^k}\big(x(\lambda)\big) \leq \frac{d^k f}{dx^k}\big(x(\lambda^*)\big) = h(1). \quad \text{This implies}$$

$$\frac{k!}{\sqrt{k+1}}\int_{\Lambda^k} \frac{d^k f}{dx^k}\big(x(\lambda_*)\big)\, dS(\lambda) \leq \frac{k!}{\sqrt{k+1}}\int_{\Lambda^k} \frac{d^k f}{dx^k}\big(x(\lambda)\big)\, dS(\lambda) \leq \frac{k!}{\sqrt{k+1}}\int_{\Lambda^k} \frac{d^k f}{dx^k}\big(x(\lambda^*)\big)\, dS(\lambda).$$

But note the leftmost term is $h(0)$ and the rightmost term is $h(1)$. Therefore, for some $\bar{t} \in [0,1]$

$$\frac{k!}{\sqrt{k+1}}\int_{\Lambda^k \subset \mathbb{R}^{k+1}} \frac{d^k f}{dx^k}\big(x(\lambda)\big)\, dS(\lambda) = h(\bar{t}) = \frac{d^k f}{dx^k}(\xi) \quad \text{at } \xi = x(\bar{\lambda}),\ \bar{\lambda} = (1-\bar{t})\lambda_* + \bar{t}\lambda^*.$$

$\square$

Reduction to the Taylor's Theorem

Theorem 4 (Taylor's theorem)**.** *Let $f \in C^{n+1}([a,b])$ and $x_0 \in [a,b]$. Then for $x \in [a,b]$,*

$$\begin{aligned}
R(x) &= \int_{x_0}^{x} \frac{(x-x')^n}{n!} \frac{d^{n+1}f}{dx^{n+1}}(x')\,dx' \\
&= (x-x_0)^{n+1} \times \int_{\tau=0}^{1} \frac{(1-\tau)^n}{n!} \frac{d^{n+1}f}{dx^{n+1}}\Big((1-\tau)x_0 + \tau x\Big)\,d\tau \\
&= (x-x_0)^{n+1} \times \frac{1}{(n+1)!} \frac{d^{n+1}f}{dx^{n+1}}(\xi) \quad \textit{for some } \xi \in [a,b] \\
&= f(x) - \Big(f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n\Big).
\end{aligned}$$

The reduction is justified by following calculations: If $x_0 = x_1 = \cdots = x_n$,

1. We have $f[x_0, \cdots, x_k] = \frac{1}{k!} f^{(k)}(x_0)$ for each k.
2. We have

$$\begin{aligned}
R(x) =& (x-x_0)(x-x_1)\cdots(x-x_n) f[x_0, x_1, \cdots, x_n, x] \\
=& (x-x_0)^{n+1} \times \int_{\tau^{n+1} \subset \mathbb{R}^{n+1}} \frac{d^{n+1}f}{dx^{n+1}}\big((1-\tau_1-\tau_2-\cdots-\tau_{n+1})x_0 + \tau_1 x_1 \cdots \tau_n x_n + \tau_{n+1} x\big)\,d\mathcal{L}^{n+1}(\tau) \\
=& (x-x_0)^{n+1} \times \int_{\tau^{n+1} \subset \mathbb{R}^{n+1}} \frac{d^{n+1}f}{dx^{n+1}}\big((1-\tau_{n+1})x_0 + \tau_{n+1} x\big)\,d\mathcal{L}^{n+1}(\tau) \\
=& (x-x_0)^{n+1} \times \int_{\tau_{n+1}=0}^{1} \frac{d^{n+1}f}{dx^{n+1}}\big((1-\tau_{n+1})x_0 + \tau_{n+1} x\big) \times \\
& \int_{\tau_n=0}^{1-\tau_{n+1}} \int_{\tau_{n-1}=0}^{1-\tau_{n+1}-\tau_n} \cdots \int_{\tau_1=0}^{1-\tau_{n+1}-\cdots-\tau_2} 1\,d\tau_1 d\tau_2 \cdots d\tau_{n+1} \\
=& (x-x_0)^{n+1} \times \int_{\tau_{n+1}=0}^{1} \frac{d^{n+1}f}{dx^{n+1}}\big((1-\tau_{n+1})x_0 + \tau_{n+1} x\big) \times \frac{(1-\tau_{n+1})^n}{n!}\,d\tau_{n+1}.
\end{aligned}$$

   Finally, by the Mean Value type Theorem

$$R(x) = (x-x_0)^{n+1} \times \frac{1}{(n+1)!} \frac{d^{n+1}f}{dx^{n+1}}(\xi) \quad \text{for some } \xi \in [a,b].$$

Chapter 4

(Local) Polynomial Interpolation

Case 1. Data $\big(x_0, f(x_0)\big), \cdots, \big(x_n, f(x_n)\big)$, x_i all distinct.

We definde $\mathcal{P}^{<n+1}([a,b]) \subset C^{n+1}([a,b])$ be the set of all polynomials of order less than $n+1$, tautologically, the set of all polynomials of order at most n.

Theorem 1 (existence)**.** *For any $f \in C^{n+1}([a,b])$ and distinct points $x_0, x_1, \cdots, x_n \in [a,b]$, there exists a polynomial $p \in \mathcal{P}^{<n+1}([a,b])$ such that*

$$p(x_i) = f(x_i) \quad \text{for every } i = 0, 1, 2, \cdots, n.$$

Proof. It suffices to recall that

$$f(x) - p_n(x) = f[x_0, x_1, \cdots, x_n, x](x - x_0)(x - x_1)\cdots(x - x_n),$$

and $p_n \in \mathcal{P}^{<n+1}([a,b])$. By assumption that $\frac{d^{n+1}f}{dx^{n+1}}$ is continuous in $[a,b]$, $f[x_0, x_1, \cdots, x_n, x]$ is a continuous function of x in $[a,b]$ and thus it is bounded in $[a,b]$. The conclusion follows by plugging in x_i in place of x. $\square$

Theorem 2 (uniqueness)**.** *Under the same assumptions, the interpolating polynomial in $\mathcal{P}^{<n+1}([a,b])$ is unique.*

Proof. Suppose $p, q \in \mathcal{P}^{<n+1}([a,b])$ be the two interpolating polynomials. Then $r = p - q \in \mathcal{P}^{<n+1}([a,b])$ too and

$$r(x_i) = 0 \quad \text{for every } i = 0, 1, 2, \cdots, n.$$

Since $x_0, \cdots, x_{n-1}$ are roots of r,

$$r(x) = A(x - x_0)(x - x_1)\cdots(x - x_{n-1}). \quad \text{But}$$

$$r(x_n) = A(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1}) = 0.$$

Since $x_0, x_1, \cdots, x_n$ are all distinct, we conclude $A = 0$, or $p = q$. $\square$

Case 2. Data $\big(x_1, f(x_1)\big), \cdots, \big(x_n, f(x_n)\big)$, $\big(x_1, f'(x_1)\big), \cdots, \big(x_n, f'(x_n)\big)$, x_i all distinct.

Theorem 3 (existence)**.** *For any $f \in C^{2n}([a,b])$ and distinct points $x_1, \cdots, x_n \in [a,b]$, there exists a polynomial $p \in \mathcal{P}^{<2n}([a,b])$ such that*

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i) \quad \text{for every } i = 1, 2, \cdots, n.$$

Proof. We consider a choice of $2n$ points in $[a,b]$ that is

$$[\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_{2n-1}] = [x_1, x_1, x_2, x_2, x_3, x_3, \cdots, x_n, x_n].$$

It suffices to recall that

$$\begin{aligned} f(x) - p_{2n-1}(x) &= f[\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_{2n-1}, x](x - \tilde{x}_0)(x - \tilde{x}_1)\cdots(x - \tilde{x}_{2n-1}) \\ &= f[x_1, x_1, x_2, x_2, \cdots, x_n, x_n, x](x - x_1)^2(x - x_2)^2 \cdots (x - x_n)^2 = R(x), \end{aligned}$$

where $f[x_1, x_1, x_2, x_2, \cdots, x_n, x_n, x]$ is continuous function of x in $[a,b]$ and

$$\begin{aligned} p_{2n-1}(x) = f[x_1] + f[x_1, x_1](x - x_1) + f[x_1, x_1, x_2](x - x_1)^2 + f[x_1, x_1, x_2, x_2](x - x_1)^2(x - x_2) + \cdots \\ + f[x_1, x_1, x_2, x_2, x_3, x_3, \cdots, x_n, x_n](x - x_1)^2(x - x_2)^2 \cdots (x - x_{n-1})^2(x - x_n) \end{aligned}$$

in $\mathcal{P}^{<2n}([a,b])$. Certainly, $R(x_i) = 0$ for $i = 1, \cdots, n$. It is also certain that $R(x) = f(x) - p_{2n-1}(x)$ is differentiable at x_i. We directly compute the derivative at x_1: By writing $g(x) = f[x_1, x_1, x_2, x_2, \cdots, x_n, x_n, x](x - x_2)^2 \cdots (x - x_n)^2$, continuous in $[a,b]$,

$$\frac{R(x_1 + h) - R(x_1)}{h} = \frac{R(x_1 + h)}{h} = \frac{h^2 g(x_1 + h)}{h} = h g(x_1 + h) \to 0 \quad \text{as } h \to 0.$$

Similar calculations can be done at $x = x_i$, to conclude that

$$R'(x_i) = 0 \quad \text{for } i = 1, \cdots, n.$$

□

Theorem 4 (uniqueness)**.** *Under the same assumptions, the interpolating polynomial in $\mathcal{P}^{<2n}([a,b])$ is unique.*

Proof. Suppose $p, q \in \mathcal{P}^{<2n}([a,b])$ be the two interpolating polynomials. Then $r = p - q \in \mathcal{P}^{<2n}([a,b])$ too and x_i, $i = 1, 2, 3, \cdots, n$ are all double roots of r. We write

$$r(x) = A(x - x_1)^2(x - x_2)^2 \cdots (x - x_{n-1})^2(x - x_n). \quad \text{But}$$

$$r'(x_n) = A(x_n - x_1)^2(x_n - x_1)^2 \cdots (x_n - x_{n-1})^2 = 0.$$

Since $x_1, x_2, \cdots, x_n$ are all distinct, we conclude $A = 0$, or $p = q$. □

Remark 4.1. Note that we can in general achieve interpolation with a choice

$$[x_0, x_1, x_2, \cdots, x_r] \quad \text{(points can be repeated in any certain way)}$$

in $\mathcal{P}^{<r+1}([a,b])$.

Chapter 5

Methods for the Lagrange Interpolation

Case 1. Data $\big(x_0, f(x_0)\big), \cdots, \big(x_n, f(x_n)\big)$, x_i all distinct.

1. The same polynomial $p \in \mathcal{P}^{<n+1}([a,b])$ of order at most n can be expanded in many different ways. For instance, the polynomial

$$\begin{aligned} p_n(x) = f[x_0] + f[x_0,x_1](x-x_0) + f[x_0,x_1,x_2](x-x_0)(x-x_1) + \cdots \\ + f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1}), \end{aligned}$$

   which is expanded with polynomials of increasing orders with respect to the sampling point x_i, of course can be re-arranged in the form

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

2. The former is said to be expanded in Newton basis, while the latter is said to be expanded in power basis. Newton basis and power basis both <u>are bad</u>.

Remark 5.1. Although the newton basis expansion of p is bad, at least one can compute it because the coefficient $f[x_0, x_1, \cdots, x_k]$ can be computed by the key recursive equality: because $x_0, x_1, \cdots, x_n$ are distinct

$$\begin{aligned} f[x_0] &= f(x_0) \\ f[x_0, x_1] &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \\ f[x_0, x_1, x_2] &= \frac{\frac{f(x_2)-f(x_1)}{x_2-x_1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_2 - x_0} \\ &\vdots \end{aligned}$$

Remark 5.2. A basis what is called Lagrange basis and another basis what is called the Bernstein basis are in general said to better behave. Here we implement polynomial construction expanded in Lagrange basis.

Definition 1. *Let $x_0, x_1, \cdots, x_n \in [a, b]$ be distinct points. We define for each $i = 0, 1, \cdots, n$,*

$$L_i(x) = \beta_i(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n) \in \mathcal{P}^{<n+1}([a, b]),$$

where $\beta_i = \Big((x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)\Big)^{-1} = (\gamma_i)^{-1}$.

A few observations:

1. By assumption that $x_0, x_1, \cdots, x_n$ are distinct, $\gamma_i \neq 0$ and thus β_i is well-defined.
2. $L_i(x_i) = 1$ and $L_i(x_j) = 0$ if $i \neq j$.
3. For every i, the order of $L_i(x)$ is n.

Remark 5.3. While $L_i(x)$ is of order n for every i, for a combination $\sum_{i=0}^{n} f(x_i)L_i(x)$ terms may cancel so that the resultant polynomial is of order less than n.

Proposition 2. *The interpolating polynomial $p \in \mathcal{P}^{<n+1}([a, b])$ obtained uniquely by Theorem 1,2 in the preceding section must be*

$$\sum_{i=0}^{n} f(x_i)L_i(x).$$

Proof. We check that $q(x) = \sum_{i=0}^{n} f(x_i)L_i(x)$ is a member of $\mathcal{P}^{<n+1}([a, b])$, and satisfies

$$q(x_i) = f(x_i) \quad i = 0, 1, \cdots, n.$$

Hence q is an interpolating polynomial in $\mathcal{P}^{<n+1}([a, b])$. Since the interpolating polynomial is unique, q coincides with the one obtained by Theorem 1,2. $\square$

Here, we devise three implementations of the function that returns the value $q(x) = \sum_{i=0}^{n} f(x_i)L_i(x)$.

The First Implementation

We consider the literal implementation of

$$q(\bar{x}) = \sum_{i=0}^{n} f(x_i)L_i(\bar{x}).$$

1. We assume the sampling points $x_s = (x_0, x_1, \cdots, x_n)$ and $f_s = (f(x_0), f(x_1), \cdots, f(x_n))$ are given as $n+1$ size vectors.
2. Recalling that

   $$L_i(x) = \beta_i(x - x_0)(x - x_1)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n),$$

   we compute

   $$\beta_i \quad \text{and} \quad m_i(x) = (x - x_0)(x - x_1)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n).$$

   (blank)

   ---

3. We implement a function returning the value $q(\bar{x})$ for a given $\bar{x}$.

   (blank)

   ---

4. This works perfectly fine.

Remark 5.4. We inserted codes block that are in Julia programming language. The labeling of x_i in Julia was adapted so that $i = 1, 2, \cdots, n+1$.

The second Implementation

We consider a few intersting implementational considerations.

1. In computing $L_i(\bar{x}) = \beta_i m_i(\bar{x})$, it is tempting we compute $m_i(\bar{x})$ not making n products, but making one division with $w(\bar{x}) = (\bar{x}-x_0)(\bar{x}-x_1)\cdots(\bar{x}-x_n)$.

$$\text{If } \bar{x} \notin \{x_0, x_1, \cdots, x_n\}, \quad L_i(\bar{x}) = \frac{w(\bar{x})\beta_i}{\bar{x}-x_i},$$

   $w(\bar{x})$ is computed once and is shared afterwards. This is to say that the first implementation is an $O(n^2)$ algorithm while the second implementation is an $O(n)$ algorithm. The difference in order by 1 reflects a very large difference for n large. (I am not sure though if we use such a large n.)

2. Here we are arranging $q(\bar{x})$ in the form

$$q(\bar{x}) = \begin{cases} f(x_i) & \text{if } \bar{x} \in \{x_0, x_1, \cdots, x_n\} \text{ and } \bar{x} = x_i, \\ \displaystyle\sum_{i=0}^{n} f(x_i)\frac{w(\bar{x})\beta_i}{\bar{x}-x_i} & \text{if } \bar{x} \notin \{x_0, x_1, \cdots, x_n\}. \end{cases}$$

(blank)

---

3. One may worry when $\bar{x}$ is very close to some x_j but $\bar{x}-x_j$ is not declared to be 0 by a computer.

4. Practically, this way of computing weights does not cause a problem. Justification?

The third Implementation

We consider another intersting implementation that uses idea to avoid the overflow.

1. Lagrange basis $\big(L_i(x)\big)_{i=0}^n$ has a nice property that
$$\text{for any } x \in [a,b], \quad L_0(x) + L_1(x) + \cdots + L_n(x) = 1$$
i.e., $L_i(x)$, interpreted as weights, add up to 1. (But they are signed weights !).

2. This is because, if we interpolate the function $f(x) \equiv 1$, then $f(x) \in \mathcal{P}^{n+1}([a,b])$ and of course f itself is interpolating the data. By uniqueness, the lagrange basis interpolating polynomial $q(x) = \sum_{i=0}^n f(x_i)L_i(x)$ of 1 must be 1.

3. This in particular gives that
$$\text{If } \bar{x} \notin \{x_0, x_1, \cdots x_n\}, \quad 1 = (\bar{x} - x_0)(\bar{x} - x_1)\cdots(\bar{x} - x_n)\sum_{i=0}^n \frac{\beta_i}{\bar{x} - x_i}.$$

4. This is in fact to conduct the partial fraction. If $\bar{x} \notin \{x_0, x_1, \cdots x_n\}$
$$\frac{1}{(\bar{x} - x_0)(\bar{x} - x_1)\cdots(\bar{x} - x_n)} = \frac{\beta_0}{\bar{x} - x_0} + \frac{\beta_1}{\bar{x} - x_1} + \cdots + \frac{\beta_n}{\bar{x} - x_n}.$$

5. Therefore,
$$\text{If } \bar{x} \notin \{x_0, x_1, \cdots x_n\}, \quad L_i(\bar{x}) = \frac{w(\bar{x})\beta_i}{\bar{x} - x_i} = \frac{\frac{\beta_i}{\bar{x} - x_i}}{\sum_{k=0}^n \frac{\beta_k}{\bar{x} - x_k}}.$$

6. If $\bar{x} \notin \{x_0, x_1, \cdots x_n\}$, define
$$t_k := \frac{\beta_k}{\bar{x} - x_k}, \quad M := \max_{k=0,1,\cdots,n} |t_k|, \quad r_k := \frac{t_k}{M} \in [-1, 1].$$

7. Since $L_i(\bar{x}) = \dfrac{t_i}{\sum_k t_k} = \dfrac{t_i/M}{\sum_k (t_k/M)}$, we compute
$$\text{If } \bar{x} \notin \{x_0, x_1, \cdots x_n\}, \quad L_i(\bar{x}) = \frac{r_i}{\sum_k r_k}.$$

8. What will happen for μ_i if $\bar{x}$ is very close to some x_j ?

(blank)

Chapter 6

The remainder $R(x)$

1. The 1d remainder theorem, stating

$$f(x) - p(x) = f[x_0, x_2, \cdots, x_n, x](x - x_0)(x - x_2) \cdots (x - x_n),$$

   asserts the equality of the left-hand-side and the right-hand-side. (It does not assert the right-hand-side is small.)

2. Of course, we will use the procedure of (sampling $\rightarrow$ interpolation) for the purpose of <u>approximation</u>. How big $|R(x)|$ can be ?

Two general principles

We examine cases where $|R(x)|$ can be so big that (sampling $\rightarrow$ interpolation) procedure fails. Inspecting examples should lead us to the following conclusion to avoid this failure.

Two lessons.

1. For a fixed number of sampling data, use low order polynomials with many small pieces, instead of using high order polynomial with 1 piece.

2. For each small such an interval piece, if possible, choose chebychev zeroes as sampling points, instead of easy equi-distance points.

The Polynomial w of one higher order

1. In our examples,

$$w(x) = (x - x_0)(x - x_2)(x - x_3) \cdots (x - x_n) \notin \mathcal{P}^{<n+1}$$

will play an important role.

2. Note that for this w, the sampling data will be $(x_i, 0)$ for $i = 0, 1, 2, \cdots n$. Hence, the interpolating polynomial p in $\mathcal{P}^{<n+1}$ is the zero function $p \equiv 0$.

3. The 1d remainder theorem is telling that

$$w(x) - 0 = R(x) = w(x).$$

Indeed, in $R(x) = w[x_0, x_1, \cdots, x_n, x](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_n)$, check that $w[x_0, x_1, \cdots, x_n, x]$ is an integral of $(n+1)$-th derivative of w but

$$\frac{d^{n+1}}{dx^n}(x - x_0)(x - x_1)(x - x_2) \cdots (x - x_n) = (n+1)! \quad (const.)$$

and thus

$$\int_{\tau^{n+1}} (n+1)!\, d\mathcal{L}(\tau) = \frac{(n+1)!}{(n+1)!} = 1.$$

The remainder $w[x_0, x_1, \cdots, x_n, x](x - x_0)(x - x_1)(x - x_2) \cdots (x - x_n)$ is indeed $w(x)$ itself.

The constant $M := \sup_{x \in [a,b]} |w(x)|$

1. We define $M := \sup_{x \in [a,b]} |w(x)|$. M is dependent on the $[a, b]$, and the choice $(x_i)_{i=0}^n$.

2. The use of w in remainder analysis for a general f can be put in the following way. The remainder

$$f[x_0, x_1, \cdots, x_n, x](x - x_0)(x - x_1) \cdots (x - x_n)$$

is the product of two expressions $f[x_0, x_1, \cdots, x_n, x]$ and $w(x)$.

For given $x \in [a, b]$, $|w(x)|$ depends only on the choice of sampling points while $\big|f[x_0, x_1, \cdots, x_n, x]\big|$ depends on $(n+1)$-th derivative of f as well. We may estimate size of the remainder separately. In case of $f = w$,

$$|f[x_0, x_1, \cdots, x_n, x]| \equiv 1, \quad |w(x)| \leq M.$$

Problems on M **when not following the two lessons before.**

How does the polynomial $w(x)$ look like in $[a, b]$ for a particular choice $(x_i)_{i=0}^n$?

Calculations of M of for two cases.

1. We adjust the domain in the form $[-a, a]$ and we directly estimate two constants

   $$M_1 := \sup_{x\in[-a,a]} \left|(x-x_0)(x-x_1)\cdots(x-x_n)\right| \quad M_2 := \sup_{x\in[-a,a]} \left|(x-r_0)(x-r_1)\cdots(x-r_n)\right|$$

   where in the first case we use equi-distanced points $x_i = -a + i\frac{2a}{n}$, and in the second case we use r_i the zeroes of $(n+1)$-th order *chebychev polynomial.*

2. The former is a bad case scenario, while the latter is a good case scenario. The estimation of M_1 and M_2 will reveal how big $|w|$ can be, for good and bad cases.

Calculations of M_2: case of chebychev zeroes

A Minimal introduction to Chebychev polynomials.

Let $\tilde{I} = [0, \pi]$. We consider its parametrization by a bijection from $[-a, a]$:

$$\theta : [-a, a] \to [0, \pi], \quad \theta(x) = \arccos\left(\frac{x}{a}\right).$$

Definition 1. *For each $k = 0, 1, 2, \cdots$*

$$T_k(x) := \cos\big(k\theta(x)\big).$$

Proposition 2. *For each $k = 0, 1, 2, \cdots$,*

1. *T_k has k zeroes in $[-a, a]$.*
2. *For $k \geq 1$, T_k attains -1, the minimum, and $+1$, the maximum, alternatingly so that $|T_k(x)|$ hits 1 exactly $k+1$ times in $[-a, a]$.*
3. *T_k is a k-th order polynomial.*
4. *For $k \geq 1$, the coefficient of the highest order term of T_k is $\frac{1}{2}\left(\frac{2}{a}\right)^k$.*

Proof. 1. In case $k = 0$, $T_0(x) \equiv 1$ and it has no zero. For $k \geq 1$, T_k attains zero only at x such that

$$k\theta(x) = \frac{\pi}{2} + m\pi$$

for m integer. Those integers so that $\phi_m = \frac{1}{k}\left(\frac{\pi}{2} + m\pi\right) \in [0, \pi]$, i.e. those integers so that $r_m = \theta^{-1}(\phi_m) \in [-a, a]$, are $m = 0, 1, 2, \cdots, k-1$. Note (ϕ_m) is strinctly increasing in m and $\theta(x)$ is a bijection. Thus, $r_0, r_1, \cdots, r_{k-1}$ are distinct k points in $[-a, a]$.

2. $T_k(x) \in [-1, 1]$, the range of the cosine function. For $k \geq 1$, $|T_k(x)|$ hits 1 only at x such that

$$k\theta(x) = m\pi$$

for m integer. Those integers so that $\tilde{\phi}_m = \frac{m\pi}{k} \in [0, \pi]$, i.e. those integers so that $t_m = \theta^{-1}(\tilde{\phi}_m) \in [-a, a]$, are $m = 0, 1, 2, \cdots, k$. Note $(\tilde{\phi}_m)$ is strinctly increasing in m and $\theta(x)$ is a decreasing bijection. Thus, $t_0, t_1, \cdots, t_k$ are distinct $k+1$ points such that

$$t_0 > t_1 > \cdots > t_k$$

and $T_k(t_m) = \cos(m\pi) = (-1)^m$.

3. For $k = 0$, $T_0(x) \equiv 1$ is 0-th order polynomial. For $k = 1$, $T_1(x) = \frac{x}{a}$. Now, we use the trigonometric identity for $k \geq 1$. For notational convenience write $\theta = \theta(x)$, then

$$T_{k+1}(x) = \cos\Big(k\theta + \theta\Big) = \cos k\theta \cos\theta - \sin k\theta \sin\theta$$

$$T_{k-1}(x) = \cos\Big(k\theta - \theta\Big) = \cos k\theta \cos\theta + \sin k\theta \sin\theta$$

$$\implies \quad T_{k+1}(x) + T_{k-1}(x) = 2T_k(x)\cos(\theta(x)) = 2T_k(x)\frac{x}{a}$$

$$\implies \quad T_{k+1}(x) = 2T_k(x)\frac{x}{a} - T_{k-1}(x).$$

Therefore, if T_0, T_1, $\cdots$, T_{k-1} are polynomials of order $0, 1, \cdots, k-1$ respectively, then $T_k(x)$ is a k-th order polynomial.

4. The coefficient of the highest order term of $T_1(x)$ is $\frac{1}{a}$, and from the last recursive formula, that of $T_k(x)$ for $k \geq 1$ is

$$\frac{1}{a}\Big(\frac{2}{a}\Big)^{k-1} = \frac{1}{2}\Big(\frac{2}{a}\Big)^{k}.$$

$\square$

Now, we can compute $M_2 := \sup_{x\in[-a,a]} \big|(x-r_0)(x-r_1)\cdots(x-r_n)\big|$. Since r_i are $(n+1)$ distinctive zeroes of $(n+1)$-th order polynomials T_{n+1} and w, and the coefficient of the highest order of $w(x)$ is 1, $w(x)$ must be

$$w(x) = 2\Big(\frac{a}{2}\Big)^{n+1} T_{n+1}(x).$$

Since the maximum of $|T_{n+1}(x)| = 1$, we conclude that $M_2 = 2\Big(\frac{a}{2}\Big)^{n+1}$.

Corollary 3. *The constant* $M_2 = 2\Big(\frac{a}{2}\Big)^{n+1}$.

Calculations of $\left|w\left(\frac{x_0+x_1}{2}\right)\right|$: equi-distance sampling

Instead of computing the supremum M_1, we go for easier calculations of $K = \left|w\left(\frac{x_0+x_1}{2}\right)\right|$. Note the equi-distance is $\frac{2a}{n}$.

$$\begin{aligned} K &= \frac{a}{n} \times \frac{a}{n} \times \frac{3a}{n} \times \cdots \times \frac{\big(2n-1\big)a}{n} \\ &= \frac{a^{n+1}}{n^{n+1}}\; 1 \times 3 \times \cdots \times \big(2n-1\big). \end{aligned}$$

Because

$$\begin{aligned} 1 \times 2 \times 4 \times \cdots \times 2n-2 \quad &< \quad 1 \times 3 \times \cdots \times \big(2n-1\big) \quad < \quad 2 \times 4 \times \cdots \times 2n, \\ (2n-1)! \quad &< \quad \Big(1 \times 3 \times \cdots \times \big(2n-1\big)\Big)^2 \quad < \quad (2n)! \end{aligned}$$

Thus,

$$\frac{a^{n+1}}{n^{n+1}}\sqrt{(2n-1)!} \quad < \quad K \quad < \quad \frac{a^{n+1}}{n^{n+1}}\sqrt{(2n)!}.$$

It is known that for $n \geq 1$

$$\sqrt{2\pi}\sqrt{2n}\Big(\frac{2n}{e}\Big)^{2n} \quad < \quad (2n)! \quad < \quad e\sqrt{2\pi}\sqrt{2n}\Big(\frac{2n}{e}\Big)^{2n}.$$

This gives that

$$\frac{a^{n+1}}{n^{n+1}}\frac{1}{\sqrt{2n}} \times \sqrt{\sqrt{2\pi}\sqrt{2n}\Big(\frac{2n}{e}\Big)^{2n}} \quad < \quad K \quad < \quad \frac{a^{n+1}}{n^{n+1}} \times \sqrt{e\sqrt{2\pi}\sqrt{2n}\Big(\frac{2n}{e}\Big)^{2n}}$$

This simplifies to that

$$\Big(\frac{2a}{e}\Big)^{n+1} \times n^{-5/4} \times \frac{e}{2\sqrt{2}}(4\pi)^{\frac{1}{4}} \quad < \quad K \quad < \quad \Big(\frac{2a}{e}\Big)^{n+1} \times n^{-3/4} \times \frac{e^{3/2}}{2}(4\pi)^{\frac{1}{4}}.$$

Remark 6.1. The common ratio respectively for two cases:

$$\rho_1 = \frac{2a}{e}, \quad \rho_2 = \frac{2a}{4}, \quad \text{for the interval length } 2a.$$

Thus, if the interval length $2a$ is not smaller than the denominator, the supremum will increases geometrically for both cases.

Runge's example and Hilbert space techniques

Suppose $a = 1$. The interval lenth $2a = 2$. For this interval:

1. One can prove in the chebychev zero case that the interpolating polynomial p_n at chebychev zeroes <u>does converge uniformly</u> to a given function f, if f has a certain differentiability, say $f \in C^2([-1, 1])$. This proof needs Hilbert space techniques.

2. By contrast, in the equi-distance case, Runge found an example of infinitely smooth function on $[-1, 1]$, where the contribution from $\big|f[x_0, x_2, \cdots, x_n, x]\big|$ increases so fast in n that it overwhelms the geometric decrement of the contribution from $|w(x)|$.

$$\begin{aligned}
f(x) &= \frac{1}{1+(5x)^2} = \frac{1}{2i}\Big(\frac{1}{5x-i} - \frac{1}{5x+i}\Big) = \frac{1}{5\times 2i}\Big(\frac{1}{x-i/5} - \frac{1}{x+i/5}\Big) \\
\frac{d^{n+1}f}{dx^{n+1}}(x) &= \frac{1}{5\times 2i}(-1)^{n+1}(n+1)!\Big(\frac{1}{(x-i/5)^{n+2}} - \frac{1}{(x+i/5)^{n+2}}\Big) \\
&= \frac{1}{5\times 2i}(-1)^{n+1}(n+1)!5^{n+2}\Big(\frac{1}{(5x-i)^{n+2}} - \frac{1}{(5x+i)^{n+2}}\Big) \\
&= \frac{(-1)^{n+1}(n+1)!5^{n+1}}{(1+(5x)^2)^{n+2}}\frac{1}{2i}\big(z^{n+2} - \bar{z}^{n+2}\big), \quad z = 5x+i, \quad x \in \mathbb{R} \\
&= \frac{(-1)^{n+1}(n+1)!5^{n+1}}{(1+(5x)^2)^{n+2}}\rho(x)^{n+2}\frac{1}{2i}\big(e^{i(n+2)\theta(x)} - e^{-i(n+2)\theta(x)}\big), \quad z = 5x+i = \rho(x)e^{i\theta(x)} \\
&= \frac{(-1)^{n+1}(n+1)!5^{n+1}}{\big(\sqrt{1+(5x)^2}\big)^{n+2}}\sin\big((n+2)\theta(x)\big).
\end{aligned}$$

3. This derivative formula alone cannot conclude that $\sup_{x\in[-1,1]} |R(x)|$ diverge as n increases, but the factor 5^{n+1} is a conceivable danger: it overwhems the factor $\big(\frac{2}{e}\big)^{n+1}$, i.e., $5 > e$.

At this stage, we wrap it up here, and move on to the topic of piecewise polynomials.

Chapter 7

Nonlinear system of equations and Newton method

We are to solve a system of n equations on n unknowns: for instance

$$\begin{aligned}
\sin(x_1 + x_3^2) + \exp(x_5)\cos(x_2 x_4) &= 3, \\
x_3^2 + x_4^2 &= 1, \\
x_1^4 - x_2^3 - 4x_3 + x_2 &= -7, \\
\sinh(x_3 - x_1) + \cosh(x_2 + x_5) &= 3, \\
x_5 + \log(x_1) &= -2.
\end{aligned}$$

For $\mathbf{f} : E \subset \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ we solve

$$\mathbf{f}(\mathbf{x}) = \mathbf{y}, \quad \mathbf{x} \in \mathbb{R}^n$$

that is a system of equations

$$\begin{aligned}
f_1(x_1, x_2, \cdots, x_n) &= y_1, \\
f_2(x_1, x_2, \cdots, x_n) &= y_2, \\
f_3(x_1, x_2, \cdots, x_n) &= y_3, \\
&\vdots \\
f_n(x_1, x_2, \cdots, x_n) &= y_n.
\end{aligned}$$

We may of course let $\tilde{\mathbf{f}}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{y}$ and solve $\tilde{\mathbf{f}}(\mathbf{x}) = 0$.

Nonlinear system of equations

For a given nonlinear system of equations, we do not know in advance in general,

1. whether any solution exists or not,
2. if exists, whether there are many solutions or not.

One very crude but reasonable idea is to make this nonlinear problem into an artificial linear problem.

Suppose we want to compute the $\sqrt{5}$, or we want to solve

$$x^2 = 5.$$

Since we can solve $ax = 5$ ($a \neq 0$), namely $x = \frac{5}{a}$, we artificially pose a problem

$$\tilde{x} \times x = 5,$$

where $\tilde{x}$ is to be regarded as a *coefficient*. Now, for given $\tilde{x}$, define

$$\frac{5}{\tilde{x}} =: \eta(\tilde{x}).$$

We now hope that we are very lucky that it happen to be $x = \eta(\tilde{x}) = \tilde{x}$. More rational expectation is to hope the iteration

$$x_{n+1} = \eta(x_n)$$

converges. If so, assuming η is continuous, this implies that at the limit value $\bar{x}$, $\bar{x} = \eta(\bar{x})$ and we found the solution.

The art of choosing $\tilde{x}$ part and x part in nonlinear equations.

The first objective of this crude procedure can be the following.

1. We choose $\tilde{x}$ part and x part in nonlinearities in such a way so that the solution $x = \eta(\tilde{x})$ of the artificial solvable problem falls into a bounded controlled region.

2. More specifically, this first objective is to come up with choice of $\tilde{x}$, x, and the candidate set $K \subset \mathbb{R}^n$ that is compact and convex such that
$$\eta(K) \subset K.$$

3. our example $\eta(\tilde{x}) = \frac{5}{\tilde{x}}$?

Theorem 1. *Suppose $K \subset \mathbb{R}^n$ is a nonempty compact and convex set, η is continuous on K and $\eta(K) \subset K$. Then there exists $\bar{x} \in K$ such that $\eta(\bar{x}) = \bar{x}$.*

We have a few remarks:

1. This theorem gives the existence of the solution.

2. The theorem works merely under the assumption of continuity.

3. Other than the case $n = 1$, it is very difficult to construct an algorithm.

A better iteration?

We return to the problem of solving $x^2 = 5$. For some reason, instead of $\frac{5}{\tilde{x}}$, we set

$$\eta(\tilde{x}) = \frac{1}{2}\Big(\tilde{x} + \frac{5}{\tilde{x}}\Big)$$

and we run the iteration

$$x_{n+1} = \eta(x_n),$$

hoping the generated sequence converges.

In fact, we ran the newton method,

$$x_{n+1} = x_n - \big(f'(x_n)\big)^{-1}(f(x_n)), \quad f(x) = x^2 - 5.$$

When can we come up with a better iteration?

Let us write

$$\mathbf{y}_* - \mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\big(\mathbf{x} - \mathbf{x}\big) = 0.$$

and consider the artificial problem

$$\mathbf{y}_* - \mathbf{f}(\tilde{\mathbf{x}}) - D\mathbf{f}(\tilde{\mathbf{x}})\big(\mathbf{x} - \tilde{\mathbf{x}}\big) = 0$$

that is a linear problem for a given $\tilde{\mathbf{x}}$. In case $D\mathbf{f}(\tilde{\mathbf{x}})$ is invertible, the linear solver

$$\eta(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}} - D\mathbf{f}(\tilde{\mathbf{x}})^{-1}\big(\mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{y}_*\big).$$

By any chance, will the iteration $\mathbf{x}_{n+1} = \eta(\mathbf{x}_n)$ generate a convergent sequence?

Let us suppose the followings:

1. The nonlinearity $\mathbf{f} \in C^2(U)$. $U \subset \mathbb{R}^n$ is a nonempty open set.
2. There exists a solution $\mathbf{x}_* \in U$.
3. $\mathbf{x}_*$ is a simple root, i.e., $D\mathbf{f}(\mathbf{x}_*)$ is invertible.

   Then, the Inverse Function Theorem says that there exist a neighborhood V of $\mathbf{x}_*$ and a neighborhood W of $\mathbf{y}_*$ such that $\mathbf{f}$ is a bijection between V and W.
4. Suppose we do know what V and W is, and where $\mathbf{y}_*$ is. We do not know where $\mathbf{x}_*$ is.

What method can capture the solution $\mathbf{x}_*$?

Define

$$\mathbf{f}(\mathbf{x}_*) - \mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\big(\mathbf{x}_* - \mathbf{x}\big) =: R(\mathbf{x}).$$

By the Taylor's theorem, $|R(\mathbf{x})|$ converges to 0 quadratically as $\mathbf{x} \to \mathbf{x}_*$:

$$\lim_{\mathbf{x}\to\mathbf{x}_*} \frac{|R(\mathbf{x})|}{|\mathbf{x} - \mathbf{x}_*|^2} \le M.$$

5. Suppose $\bar{\mathbf{x}} \in V$ is a reasonably good approximation of $\mathbf{x}_*$, so that $|\bar{\mathbf{x}} - \mathbf{x}_*|^2$ is much smaller than $|\bar{\mathbf{x}} - \mathbf{x}_*|$.

Then, for L, an $n \times n$ invertible matrix,

$$\begin{aligned}
& L^{-1}\big(\mathbf{f}(\mathbf{x}_*) - \mathbf{f}(\bar{\mathbf{x}})\big) - L^{-1}D\mathbf{f}(\bar{\mathbf{x}})\big(\mathbf{x}_* - \mathbf{x}\big) = L^{-1}R(\bar{\mathbf{x}}) \\
\iff\ & L^{-1}\big(\mathbf{f}(\mathbf{x}_*) - \mathbf{f}(\bar{\mathbf{x}})\big) - \big(\mathbf{x}_* - \bar{\mathbf{x}}\big) + \big(\mathrm{I} - L^{-1}D\mathbf{f}(\bar{\mathbf{x}})\big)\big(\mathbf{x}_* - \bar{\mathbf{x}}\big) = L^{-1}R(\bar{\mathbf{x}}) \\
\iff\ & \Big[\bar{\mathbf{x}} - L^{-1}\big(\mathbf{f}(\bar{\mathbf{x}}) - \mathbf{f}(\mathbf{x}_*)\big)\Big] - \mathbf{x}_* = D\mathbf{f}(\mathbf{x})^{-1}R(\bar{\mathbf{x}}) - \big(\mathrm{I} - L^{-1}D\mathbf{f}(\bar{\mathbf{x}})\big)\big(\mathbf{x}_* - \bar{\mathbf{x}}\big)
\end{aligned}$$

If we can take L so that $L^{-1}D\mathbf{f}(\bar{\mathbf{x}})$ is close to the identity, and if we can justify the $(RHS) \le C_0|\bar{\mathbf{x}} - \mathbf{x}_*|^2$, possibly much smaller than $|\bar{\mathbf{x}} - \mathbf{x}_*|$, then the expression

$$\bar{\mathbf{x}} - L^{-1}\big(\mathbf{f}(\bar{\mathbf{x}}) - \mathbf{y}_*\big)$$

could become a better approximation of $\mathbf{x}_*$ than $\bar{\mathbf{x}}$. Thus, we try

$$\eta(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}} - L^{-1}\big(\mathbf{f}(\tilde{\mathbf{x}}) - \mathbf{y}_*\big),$$

expecting the convergence in the iteration $\mathbf{x}_{n+1} = \eta(\mathbf{x}_n)$.

We will justify this convergence in theorem-proof style.

It is important to note that

1. we need to start with a good approximation $\bar{\mathbf{x}}$.
2. $\mathbf{x}_*$ needs to be a simple root. (We will consider later the case $\mathbf{x}_*$ is a multiple root)
3. $\mathbf{f}$ needs to be C^2 on some neighborhood of $\mathbf{x}_*$.

Chapter 8

Convergence of Newton Method in 1d

1. For given $\mathbf{f}$ and $\mathbf{y}_*$, let $\tilde{\mathbf{f}}(x) = \mathbf{f}(x) - \mathbf{y}_*$ and solve zero finding problem
$$\tilde{\mathbf{f}}(\mathbf{x}) = 0.$$
2. We examined that the newton iteration may possibly converge under conditions where $\mathbf{f} \in C^2(U)$ and the zero $\mathbf{x}_*$ has multiplicity 1.
3. Since the Newton Method is convergent under the quite general assumption of multiplicity, we formulate our assumptions with respect to the multiplicity below.

Assumptions with respect to the multiplicity $p \geq 1$ are formulated in (M_p):

(i) if $p = 1$, $f \in C^1\big(I(r)\big)$ for some $r > 0$, where we denote $I(r) = (x_* - r, x_* + r)$.

$(i)'$ if $p > 1$, $f \in C^1\big(I(r)\big)$ for some $r > 0$, where we denote $I(r) = (x_*, x_* + r)$ in this case.

(ii) for $x \in I(r)$, $f(x) = c_p(x - x_*)^p + R(x)$ for some $c_p \neq 0$ and
$$\lim_{x \to x_*} \frac{|R(x)|}{|x - x_*|^p} = 0, \quad \lim_{x \to x_*} \frac{|R'(x)|}{|x - x_*|^{p-1}} = 0, \quad \text{where the limit is taken in } I(r) \setminus \{x_*\}. \tag{M_p}$$

Theorem 1. *Suppose f and $p \geq 1$ satisfy (M_p). Then there exists $0 < r' \leq r$ so that followings are true.*

1. *If $x \in I(r') \setminus x_*$ then $f'(x) \neq 0$.*
2. *If $x \in I(r') \setminus x_*$, define $\eta(x) = x - \frac{f(x)}{f'(x)}$. Then,*

$$\text{for some } 0 < \lambda < 1, \quad |\eta(x) - x_*| < \lambda |x - x_*|,$$
$$\text{in particular if } p > 1, \quad 0 \leq \eta(x) - x_* < \lambda (x - x_*).$$

Remark 8.1. Because of the item 2, $\eta(x)$ is a member of $I(r')$. Hence, either $\eta(x)$ happen to be the solution x_* or the interation can be continued: If $x_0 \in I(r')$, the newton iteration either finds solution in finite steps or can be continued to converge geometrically.

Proof. Let $p \geq 1$. We have that

$$|f'(x)| = \left|pc_p(x-x_*)^{p-1} + R'(x)\right| \geq |pc_p(x-x_*)|^{p-1} - |R'(x)|$$
$$= |x-x_*|^{p-1}\left(|pc_p| - \frac{|R'(x)|}{|x-x_*|^{p-1}}\right).$$

By assumption (M_p), $\frac{|R'(x)|}{|x-x_*|^{p-1}} \to 0$ in $I(r)\backslash\{x_*\}$ as $x \to x_*$. Therefore, for $\epsilon = \frac{|pc_p|}{2} > 0$, there exists $0 < r'' \leq r$ such that

$$x \in I(r'') \setminus \{x_*\} \quad \Longrightarrow \quad \frac{|R'(x)|}{|x-x_*|^{p-1}} < \epsilon.$$

This implies that if $x \in I(r'') \setminus \{x_*\}$ then $f'(x) \neq 0$.

Now, assume $x \in I(r'') \setminus \{x_*\}$ and define $\eta(x) = x - \frac{f(x)}{f'(x)}$. Then,

$$\eta(x) - x_* = x - x_* - \frac{f(x)}{f'(x)}$$
$$= (x-x_*)\left(1 - \frac{f(x)}{f'(x)(x-x_*)}\right) = (x-x_*)\lambda(x).$$

Now,

$$\lambda(x) = 1 - \frac{c_p(x-x_*)^p + R(x)}{pc_p(x-x_*)^p + R'(x)(x-x_*)^p}$$
$$= 1 - \frac{c_p + \frac{R(x)}{(x-x_*)^p}}{pc_p + \frac{R'(x)}{(x-x_*)^{p-1}}}.$$

Note that $\lambda(x) \to 1 - \frac{1}{p}$ as $x \to x_*$ in $I(r'') \setminus \{x_*\}$. Therefore, for any $\epsilon > 0$, there exists $0 < r' \leq r''$ such that

$$x \in I(r') \setminus \{x_*\} \quad \Longrightarrow \quad \left|\lambda(x) - (1 - \frac{1}{p})\right| < \epsilon.$$

In case $p = 1$, $1 - \frac{1}{p} = 0$ and we choose $\epsilon = \frac{1}{2}$ to conclude

$$x \in I(r') \setminus \{x_*\} \quad \Longrightarrow \quad |\eta(x) - x_*| \leq |x-x_*|\ |\lambda(x)| < \frac{1}{2}|x-x_*|.$$

In case $p > 1$, we can take $\epsilon > 0$ so small that

$$0 \leq 1 - \frac{1}{p} - \epsilon \quad \text{and} \quad 1 - \frac{1}{p} + \epsilon < 1.$$

We conclude that with this choice,

$$x \in I(r') \setminus \{x_*\} \quad \Longrightarrow \quad 0 \leq \eta(x) - x_* = \lambda(x)(x-x_*) < \left(1 - \frac{1}{p} + \epsilon\right)(x-x_*).$$

$\square$

1. If the multiplicity $p \geq 1$ is in advance known, one can have hypergeometric convergence by using a variant of newton method:

$$\eta(x) = x - p\frac{f(x)}{f'(x)}.$$

2. Here, we consider this hypergeometric convergence for the case $p = 1$ under the conditions

$$f \in C^2(I(r)), \quad \text{for some } M > 0 \quad \lim_{x \to x_*} \frac{|R(x)|}{|x - x_*|^2} \leq M, \quad \lim_{x \to x_*} \frac{|R'(x)|}{|x - x_*|} \leq M.$$

Chapter 9

Error Analysis framework

1. Many of numerical analysis problem can be formulated as an equation solving problem
$$F(x) = y_*,$$
where $F : X \to Y$ for some nonempty sets X and Y. (Say, metric spaces)

2. For example, consider a problem of finding QR decomposition of a matrix $A \in \mathbb{R}^{n\times k}$.
   (a) We take the vector space $X = X_1 \times X_2$, where X_1 is the space of $n \times k$ orthogonal matrices, X_2 is the space of $k \times k$ upper triangular matrices.
   (b) We take Y as the vector space of $n \times k$ square matrices.
   (c) We take the nonlinear function by the matrix multiplication $F(Q, R) = QR$. Entrywisely, $F_{i,j} = \sum_k Q_{i,k}R_{k,j}$.
   (d) The problem is to solve $F(Q, R) = A$.

To be able to numerically solve the equation, a few minimum requirements should be met.

1. F must be continuous. Otherwise, even if we happen to find a good approximate solution $\tilde{x}$ with $|x - \tilde{x}|$ small, we will not be able to check if $F(\tilde{x}) - y_*$ is close to 0 or not.

One might also not want to assume whether for each member $y \in Y$, there exists a solution $x \in X$ such that $F(x) = y$. However, we take the excuse that we already have a proof of existence theorem, and it has characterized the set Y such that for $y \in Y$, the existence of solution $F(x) = y$ is for sure.

We thus assume that

1. F is continuous and surjective.

We develop an error analysis framework under this assumption.

One typical way of making algorithm.

1. Of course, to obtain the solution is assumed to be more difficult than evaluating F at x.

   For example, if we consider a partial differential equation

   $$-\Delta u(x) = \rho(x), \quad x \in U,$$
   $$u(x) = 0, \quad x \in \partial U,$$

   evaluating the Laplacian for given u is easy, but taking inverse from ρ is difficult.

2. Hence, as we have seen in previous class, we change the problem into easier one that we can solve, namely
   $$\tilde{F}(x) = y.$$
   Here, let us take the solvability of $\tilde{F}$, as we have a unique solution for this problem, and we know howto solve it to have

   $$\tilde{x} = \tilde{M}(y).$$

   For example, a pde problem can be changed to a numerical problem in finite dimensions.

   But have in mind that, in case $F(x) = y$ has many solutions, different algorithm can pick different solution uniquely.

3. In many cases, we have a sequence of modification $\tilde{F}_j$, $j = 1, 2, \cdots$, and thereby a sequence of approximate solver $\tilde{M}_j$.

4. Even though we do not intend to have such parametrized solver $\tilde{M}_j$, our method $\tilde{M}$ will be implemented in a machine, and is affected by rounding-off due to the finite digits of our machine. In this case, the number of digits will do the role of parameter.

Definition 1 (consistency of $(\tilde{F}_j)$ and $\tilde{F}$)**.** .

1. *We say that $(\tilde{F}_j)$ is consistent at x if*

   $$|\tilde{F}_j(x) - F(x)| \to 0 \quad \textit{as } j \to \infty.$$

2. *We say that $\tilde{F}$ is η-consistent at x for $\eta > 0$ if*

   $$|\tilde{F}(x) - F(x)| < \eta.$$

Remark 9.1. The consistency is a necessary requirement: Otherwise, we are solving irrelevant problems.

A consistent algorithm

We have another way of formulating the consistency, the consistency of methods $(\tilde{M}_j)$. We first define the minimal consistency as requirement that $\tilde{M}_j(y) \in X$ at $y \in Y$.

Definition 2 (consistency of $(\tilde{M}_j)$ and $\tilde{M}$)**.** .

1. *We say that $(\tilde{M}_j)$ is consistent at y, if $\tilde{M}_j(y) \in X$ and*

$$\left|y - F(\tilde{M}_j(y))\right| \to 0 \quad \text{as } j \to \infty.$$

2. *We say that $\tilde{M}$ is η-consistent at y for $\eta > 0$ if $\tilde{M}(y) \in X$ and*

$$\left|y - F(\tilde{M}(y))\right| < \eta.$$

We will call $\left|y - F(\tilde{M}_j(y))\right|$ the testing error. This is distinguished from the solution error $\left|x_* - \tilde{M}_j(y)\right|$ for x_* a solution.

1. In many of numerical analysis problem, devising a consistent algorithm can be sufficient, or can be optimal.
2. In that case, we are not aware of that the solution error is small or not.

A convergent algorithm

Definition 3 (convergence of $(\tilde{M}_j)$ and accuracy of $\tilde{M}$)**.** .

1. *We say that a consistent algorithm $(\tilde{M}_j)$ at y is convergent if $\tilde{M}_j(y)$ is convergent to some $\bar{x} \in X$ as $j \to \infty$.*
2. *We say that an η-consistent algorithm $\tilde{M}$ at y is ϵ-accurate for $\epsilon > 0$ if there exists a solution $\bar{x} \in X$ such that*

$$|\bar{x} - \tilde{M}(y)| < \epsilon.$$

A consitent algorithm and additional assumption on F

We now impose additional assumptions on F.

For simpler exposition, the additional assumptions are

A1. F is injective, thus it is bijective between X and Y, and

A2. The inverse R is continuous. For every $y' \in Y$, and every $\epsilon > 0$, there exists $\eta > 0$ such that

$$|y - y'| < \eta \quad \Longrightarrow \quad |x - x'| < \epsilon, \quad \text{where } x = R(y),\, x' = R(y').$$

For more general exposition,

A. F is an open map. For every $x' \in X$, and every $\epsilon > 0$, there exists $\eta > 0$ such that

$$|y - F(x')| < \eta \quad \Longrightarrow \quad \exists x \quad \text{with} \quad F(x) = y \quad \text{and} \quad |x - x'| < \epsilon.$$

In both cases the ϵ is a function of η and y. We write $\epsilon_{\eta,y}$.

Under this additional assumption, a consistent algorithm is an accurate algorithm:

1. Suppose an algorithm $\tilde{M}$ is η-consistent at y_*, i.e.,

$$\left|y_* - F(\tilde{M}(y_*))\right| < \eta, \quad \tilde{y} := F(\tilde{M}(y_*)).$$

2. Suppose $\eta > 0$ and $\epsilon_{\eta,\tilde{y}} > 0$ are chosen as in the above for the map F satisfying the the additional assumption.

3. By the additional assumption on F,

$$\text{there exists } x_* \in X \text{ with } F(x_*) = y_* \text{ and } |x_* - \tilde{M}(y)| < \epsilon_{\eta,\tilde{y}}.$$

This is known as *'a posteriari error analysis'*.

Task of calculating $\epsilon_{\eta,y}$

We bound the tolerable error in the following way.

For the tolerable error $\epsilon > 0$ is imposed, we obtain $\eta > 0$ so that $\epsilon = \epsilon_\eta$ and we increase j until $\tilde{M}_j$ is η-consistent.