THREE LEVELS OF NUMERICAL MATHEMATICS

Numerical mathematics $\approx$ practical (physical, machinane aided)
evaluation of theoretical formulas

1) Pure mathematical level

Example. Riemann conjecture $\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}$

$$\zeta(z) = 0,\ z \neq -2, -4, \ldots \Longrightarrow \exists\ t \in \mathbb{R}\ \ z = \tfrac{1}{2} + it$$

Compromise-free axionmatic poof is required.

Alan TURING: looking for counter-examples with (ancient) computer.

2) Theoretical analysis of computation algorithms.

Theoretical accuracy, reliability of guesses for solutions. ("Usual" university texts)

Investigation of finite computations with "Infinitely long arithmetics".

Example. Convergence rate estimates for Newton iteration.

We are going farther: investigation of possible chaos in calculations.

3) "Engineering" level.

Use of machine architectures and algorithms without theoretical criticism.

Not for disprising even from the view points of 1-2):

Learning- and genetic algorithms — not yet controlled, but extremally successful.

Topics for 2). Quantum computers

Example for the cooperation of 1-2-3).

Computations in Quantum Chemistry (Gaussian package, Nobel Prize 2000).

1) Schrödinger-equation,

2) Approximating solutions techniques ($\longrightarrow$ conjugate gradient methods).

3) Imitation with computer of 2) (far from financial success in the moment)

BASIC CONCEPTS IN NUMERICAL MATHEMATICS

Task: $F(x, d) = 0$ where $d :=$ [data]

Direct problem: given $F, d,\ \ x =?$

Inverse problem: given $F, x,\ \ d \in \{?\}$;

Identification problem: given $x, d,\ \ F =?$.

Well-posedness: existing unique solution with continuous dependence on the data

Relative condition number (for $F(x + \delta x, d + \delta d) = 0$): $K(d) := \sup_{\delta d} \frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|}$.

Condition number (when $d = 0$ or $x = 0$): $K_{\text{abs}}(d) := \sup_{\delta d} \frac{\|\delta x\|}{\|\delta d\|}$.

Stability for approximating methods with $F_n(x_n, d_n) = 0$:

$$F_n(x, d) \to 0 \ \ (n \to \infty) \ \ \text{whenever} \ \ F(x, d) = 0.$$

A proiori analysis: Investigation of direct problems.

A posteriori analysis: Investigation of indirect problems.

NEWTON ITERATION

Recall the classical 1D case.

Mesopotamia ($\approx$ 1500 BC). $\sqrt{2}$: $1+\frac{30}{60}=1.5$ $1+\frac{25}{60}\approx 1.41666$ $1+\frac{24}{60}+\frac{51}{3600}\approx 1.1.414667$

$2,\ \frac{1}{2}\Big[2+\frac{2}{2}\Big],\quad \frac{1}{2}\Big[1.S+\frac{2}{1.5}\Big],\ \frac{1}{2}\Big[1.41\dot{6}+\frac{2}{1.41\dot{6}}\Big]$

$$x^2=2 \quad x=\frac{2}{x} \quad \left.\begin{matrix} x=\frac{2}{x} \\ x=x \end{matrix}\right\}\begin{matrix}\cdot 1/2 \\ \frac{\cdot 1/2}{+}\end{matrix} \quad x=\frac{1}{2}\Big(x+\frac{2}{x}\Big)$$

$$\left.\begin{matrix} \sqrt{a}-\text{ra}-\text{is} & x=\frac{a}{x} \\ x^2=a & x=x \end{matrix}\right\} \quad x=\frac{1}{2}\Big[x+\frac{a}{x}\Big]$$

Questions

1) Why does it converge

2) How to proceed with other numbers, e.g. $\sqrt[3]{a}$ =?

$$\text{KEPLER} \qquad \left.\begin{matrix} x^3=a \\ x=x \end{matrix}\right\}\begin{matrix}\frac{1}{3} \\ \cdot\frac{2}{3}\end{matrix} \quad x=\frac{2}{3}x+\frac{1}{3}\frac{a}{x^2}$$

Conjecture $\sqrt[N]{a}$-hz $x_0:=a>1$

$$x_{n+1}=\Big(1-\frac{1}{N}\Big)x_n+\frac{1}{N}\frac{a}{x_n^{N-1}}$$

1) NEWTON's answer: in general for the solution of $f(x)=0$ [in particular $x^2-2=0$

ITERATION WITH TANGENT LINES

Slope $=f'(x_n)$ FIGURE

$$(x_{n+1}-x_n)\cdot f'(x_n)=f(x_n)$$

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

Example. $\sqrt[N]{a}$ $x^N-a=0$ $f(x):=x^N-a$ $f'(x)=Nx^{N-1}$

$$x_{n+1}=x_n-\frac{x_n^N-a}{Nx_n^{N-1}}=\Big(1-\frac{1}{N}\Big)x_n+\frac{1}{N}\frac{a}{x_n^{N-1}}$$

Reasons for convergence? Not always feasable:
$f(x) := x^3 - x - 1/\sqrt{3}$ "finding 0" may be CHAOTIC [details later]

TAYLOR FORMULA + LAGRANGE REMAINDER TERM

$$f(x+h) = f'(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \ldots + \frac{1}{(d-1)!}f^{(d-1)}(x)h^{d-1} +$$
$$+ \frac{1}{d!}f^{(d)}(t_{x,h})h^d \qquad \exists\, t_{x,h} \in [x, x+h]\ ,$$

if $f : \mathbb{R} \to \mathbb{R}$ is d times continuously differentiable.

$d = 2$: $\qquad f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(t_{x,h})h^2$

NEWTON's HEURISTICS

$x_n \approx x_* \quad f(x_*) = 0$

$f(x_n + h) \approx f(x_n) + f'(x_n)h$

$0 = f(x_*) \approx f(x_n) + f'(x_n) \cdot (x_* - x_n)$
$x_* \approx x_n - f(x_n)/f'(x_n)$

ESTIMATE
(1) $\quad 0 = f(x_*) = f(x_n) + f'(x_n)(x_* - x_n) + \frac{1}{2}f''(t_{x_n, x_n - x_*})(x_* - x_n)^2$

(2) $\quad 0 = [f$ replaced with its linear approximation $](x_{n+1}) = f(x_n) + f'(x_n)(x_{n+1} - x_n)$

(2) - (1) $\Rightarrow \qquad 0 = f'(x_n)(x_{n+1} - x_*) - \frac{1}{2}f''(t_{x_n, x_* - x_n})(x_n - x_*)^2$

$$|x_{n+1} - x_*| = \frac{1}{2} \underbrace{\frac{|f''(t_{x_n, x_n - x_*})|}{|f'(x_n)|}} |x_n - x_*|^2$$
$$\leq M\ , \quad \text{ha} \quad x_n \in (x_* - \varepsilon, x_* + \varepsilon).$$

$h_n := |x_n - x_*| = [$ distance of x_n from the solution $]$

$h_{n+1} \leq M h_n^2\ , \quad$ but M can be very large

Iterating this formally yields

$$\begin{aligned} h_1 &\le Mh_0^2 \\ h_2 &\le Mh_1^2 \le M(Mh_0^2)^2 = M^3h_0^4 \\ h_3 &\le Mh_2^2 \le M(M^3h_0^4)^2 = M^7h_0^8 \\ &\vdots \\ h_{n+1} &\le M^{2^n-1}h_0^{2^n} = \frac{1}{M}[Mh_0]^{2^n} \end{aligned}$$

If $Mh_0 < 1$ (x_0 lies "very near" to x_*) then $x_n \to x_*$, moreover $|x_n - x_*| \le \text{const.}(1-\delta)^{2^n}$

NEWTON ITERATION IN SEVERAL VARIABLES

$$F : \mathbb{R}^N \to \mathbb{R}^N, \quad x_i : \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_N \end{bmatrix} \mapsto \xi_i \quad \text{coordinates (coordinate functions)}, \quad F = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix}$$

$$F_H'(A) := \lim_{\tau \to 0} \frac{1}{\tau}\big[F(A+\tau H) - F(A)\big] = \frac{\partial F(A+\tau H)}{\partial \tau}\Big|_{\tau=0} \quad \textbf{directional derivative}$$

Recall. F continuously differentiable $\Rightarrow$ $H \mapsto F_H'(A)$ is a linear mapping $\mathbb{R}^N \to \mathbb{R}^N$

F d times continuously differentiable $\Rightarrow$ $(H_1, \dots, H_d) \mapsto F_{H_1 \cdots H_d}''\cdots'(A)$
is a *symmetric* d-linear $\big[\mathbb{R}^N\big]^d \to \mathbb{R}^N$ mapping

Notation: $F^{(d)}(A)H_1 \cdots H_d := F_{H_1 \cdots H_d}''\cdots'(A), \quad F^{(d)}(A)H^d := F^{(d)}(A)\underbrace{H \cdots H}_{d \text{ times}}$

Tensorial form: $F^{(d)} \sim \left[\dfrac{\partial^d F}{\partial x_{k_1} \cdots \partial x_{k_d}}\right]_{k_1,\dots,k_d=1}^{N}$.

Taylor formulas with remainder terms

$$F^{(d)}_{[A,A+H]} := d! \int_{t_1=0}^{1} \int_{t_2=0}^{1} \cdots \int_{t_d=0}^{1} F^{(d)}\big(A + t_d H\big)\, dt_d\, dt_{d-1} \cdots dt_1;$$

$$\begin{aligned} F(A+H) &= \sum_{k=0}^{d-1} \frac{1}{k!} F^{(k)}(A) H^k + \frac{1}{d!} F^{(d)}_{[A,A+H]} H^d = \\ &= \sum_{k=0}^{d-1} \frac{1}{k!} F^{(k)}(A) H^k + \frac{1}{d!} \int_{\tau=0}^{1} \underbrace{w_d(\tau)}_{d(1-\tau)^{d-1}} F^{(d)}(A+\tau H) H^d \, d\tau, \quad \int_0^1 w_d = 1; \\ &= \Big[F^{(d)}\big(A+\vartheta_1 H\big), \dots, F^{(d)}\big(A+\vartheta_N H\big) \text{ convex lin. combination} \Big] = \\ &= \sum_{k=0}^{d-1} \frac{1}{k!} F^{(k)}(A) H^k + \frac{1}{d!} \Big\langle \phi \Big| F^{(d)}\big(A+\vartheta_{A,H} H\big) H^d \Big\rangle U, \quad 1 = \langle \phi | U \rangle = \|\phi\| = \|U\|. \end{aligned}$$

Iteration step

$$X_{n+1} := X_n - \big[F'(X_n)\big]^{-1} F(X_n)$$

That is X_{n+1} is the solution of the equation $\big[$Taylor polynomial of 1st order of F around $X_n\big] = 0$:

$$(*) \qquad F(X_n) + F'(X_n)(X_{n+1} - X_n) = 0\ .$$

This is well-defined and unique if the linear map $F'(X_n) : \mathbb{R}^N \to \mathbb{R}^N$ is *invertible*.

Convergence estimate: Similarly as with 1 variable, but using Taylor formula with remainder in integration:

$$\begin{aligned} 0 &= F(X_n) + F'(X_n)(X_{n+1} - X_n), \\ 0 &= F(X_*) = \\ &= F(X_n) + F'(X_n)(X_* - X_n) + \frac{1}{2} \int_{\tau=0}^{1} w_2(\tau) F''\big(X_n + \tau(X_* - X_n)\big)(X_* - X_n)^2 \, d\tau. \end{aligned}$$

Taking the difference of these equations we get

$$\begin{aligned} 0 &= F'(X_n)(X_{n+1} - X_*) - \frac{1}{2} \int_{\tau=0}^{1} w_2(\tau) F''\big(X_n + \tau(X_* - X_n)\big)(X_* - X_n)^2 \, d\tau, \\ X_{n+1} - X_* &= \frac{1}{2} \int_{\tau=0}^{1} w_2(\tau) \big[F'(X_n)\big]^{-1} F''\big(X_n + \tau(X_* - X_n)\big)(X_* - X_n)^2 \, d\tau, \\ \|X_{n+1} - X_*\| &\le \frac{1}{2} \max_{\tau \in [0,1]} \Big\| \big[F'(X_n)\big]^{-1} F''\big(X_n + \tau(X_* - X_n)\big)(X_* - X_n)^2 \Big\| \le \\ &\le \frac{1}{2} \Big[\max_{\substack{X,Y \in [X_n, X_*] \\ \|H\|=1}} \Big\| \big[F'(X)\big]^{-1} F''(Y) H^2 \Big\| \Big] \|X_* - X_n\|^2. \end{aligned}$$

Theorem. *Assume $F : \mathbb{R}^N \to \mathbb{R}^N$ is a 2 times continuously differntiable mapping, such that $F(X_*) = 0$ at the point $X_* \in \mathbb{R}^N$. Suppose furthermore that K is a bounded sphericak neighvorhood of X_* with the property that $\det\big(F'(X)\big) \neq 0$ $(X \in K)$. Then the constant*

$$M := \frac{1}{2} \max_{\substack{X,Y \in K \\ \|H\|=1}} \Big\| \big[F'(X)\big]^{-1} F''(Y) H^2 \Big\|$$

is a well-defines finite number. If $X_0 \in K$ lies so near to X_* that we have

$$\lambda := M\|X_0 - X_*\| < 1 \ ,$$

then all the points $X_1, X_2, \ldots$ are (uniquely) well-defined by the steps $(*)$ of the Newton iteration starting from X_0 and they remain in K (i.e. $X_1, X_2, \ldots \in K$) and $\|X_n - X_*\| = O\big(\lambda^{2^n}\big)$. Namely

$$\|X_n - X_*\| \leq \lambda^{2^n - 1} \|X_0 - X_*\| \qquad (n = 0, 1, 2, \ldots).$$

Proof. Verification that the constant M is well-defined: since F is 2 times continuously differentiable, the function $(X, Y, H) \mapsto \Big\| \big[F'(X)\big]^{-1} F''(Y) H^2 \Big\|$ is continuous and hence it assumes it maximum felveszi in the closed bounded figure $K \times K \times \{H : \ \|H\| = 1\} \subset \big[\mathbb{R}^N\big]^3$.

Let $K := \{X : \ \|X\| \leq \varepsilon\}$, $X_0 \in K$ és $\lambda = M\|X_0 - X_*\| < 1$. Hnceforth we can use analogous arguments as in the case of 1 variable.

We have $\|X_0 - X_*\| = \lambda^{2^0 - 1} \|X_0 - X_*\| \leq \varepsilon$. Induction step in accordance with the convegence estimate:

$$\begin{aligned} X_n \in K \ + \ \|X_n - X_*\| \leq \lambda^{2^n - 1} \|X_0 - X_*\| \implies \\ \|X_{n+1} - X_*\| \leq M\|X_n - X_*\|^2 \leq M\big[\lambda^{2^n - 1} \|X_0 - X_*\|\big]^2 =^{M\|X_0 - X_*\| = \lambda} \\ = \lambda^{2(2^n - 1) + 1} \|X_0 - X_*\| = \lambda^{2^{n+1} - 1} \|X_0 - X_*\| \leq \varepsilon \ \ \big(\Rightarrow X_{n+1} \in K\big). \end{aligned}$$

Procedure with matrix calculus

$$\nabla f_i(x) = \Big[\frac{\partial f_i}{\partial x_1} \ \frac{\partial f_i}{\partial x_2} \ \ldots \ \frac{\partial f_i}{\partial x_N}\Big] \qquad \textit{gradient} \quad \text{(row vector)}$$

$$\nabla^2 f_i(x) := \Big[\frac{\partial^2 f_i}{\partial x_k \partial x_\ell}\Big]_{k,\ell=1}^{N} \qquad \textit{Hessian} \text{ (matrix)}$$

$$f_i(x+h) = f_i(x) + \big[\nabla f_i(x)\big]h\big\rangle + \frac{1}{2}h^*\big[\nabla^2 f_i(t_{x,h})\big]h$$

Suppose the point $X_n \in \mathbb{R}^N$ is a given $\qquad \big[\ n\text{-th approximation of } X_* \text{ where } F(X_*) = 0\ \big]$

$$\begin{aligned} f_i(X) \approx f_i^{[n]}(X) &:= \big[\text{ approximation in 1-st order of } f_i \text{ around } X_n\big] = \\ &= f_i(X_n) + \big[\nabla f_i(X_n)\big](X - X_n) = \\ &= f_i(x_n) + \Big\langle [\nabla f_i(x_n)]^* \Big| \ X - X_n \Big\rangle \end{aligned}$$

$F(X) = 0 \quad \approx \quad f_i^{[n]}(x_1, \ldots, x_n) = 0 \ (i = 1, 2, \ldots, N)$ system of equations

Example. $N = 2, \quad X = \begin{bmatrix} x \\ y \end{bmatrix}, \quad F = \begin{bmatrix} f \\ g \end{bmatrix}, \quad X = \begin{bmatrix} x_n \\ y_n \end{bmatrix}$

$$\left.\begin{aligned} F(X_n) + \frac{\partial f(X_n)}{\partial x}(X_{n+1} - X_n) + \frac{\partial f(X_n)}{\partial y}(y_n, -y_n) = 0 \\ g(X_n) + \frac{\partial g(X_n)}{\partial x}(x_{n+1} - x_n) + \frac{\partial g(X_n)}{\partial y}(y_{n+1} - y_n) = 0 \end{aligned}\right\}$$

$$f(X_n) + \underbrace{\begin{bmatrix} \nabla f \\ \nabla g \end{bmatrix}}_{\substack{2 \times 2 \text{ matrix} \\ \text{the Jacobian of } F}} (X_{n+1} - X_n) = 0$$

In general: (in arbitrary dimensions $N \geq 1$)

$$F'(X) = \big[\text{ Jacobian of } F \text{ at the point } X\big] := \underbrace{\begin{bmatrix} \nabla f_1(X) \\ \vdots \\ \nabla f_N(X) \end{bmatrix}}_{N \times N\text{-es mátrix}} .$$

We have to find the solution X_{n+1} of

$$F(X_n) + F'(X_n)(X_{n+1} - X_n) = 0.$$

Hence

$$X_{n+1} := X_n - \underbrace{F'(X_n)}_{\text{Jacobi}}^{-1} F(X_n) = X_n - \begin{bmatrix} \nabla f_1(X_n) \\ \vdots \\ \nabla f_N(X_n) \end{bmatrix}^{-1} F(X_n)$$

Estimate. Since

$$f_i(X_n + H) = f_i(X_n) + [\nabla f_i(X_n)]H + \frac{1}{2}H^*[\nabla^2 f_i(T_i)]H \qquad \exists\, T_i \in [X_n, X_n + H],$$

by setting $H := X_* - X_n$ we have

$$0 = f_i(X_*) = f_i(X_n) + [\nabla f_i(X_n)](X_* - X_n) + \frac{1}{2}(X_* - X_N)^*[\nabla^2 f_i(T_i)](X_* - X_N),$$

$$0 = F(X_*) = F(X_n) + F'(X_n)(X_* - X_n) + \frac{1}{2}\begin{bmatrix} (X_* - X_N)^*[\nabla^2 f_1(T_i)](X_* - X_N) \\ \vdots \\ (X_* - X_N)^*[\nabla^2 f_1(T_i)](X_* - X_N) \end{bmatrix}$$

for suitable $T_1, \ldots, T_N \in [X_n, X_*]$. On the other hand, by the definition of X_{n+1} we have

$$0 = F(X_n) + F'(X_n)(X_{n+1} - X_n).$$

Subtracting the previous equations (vectorially) from this, we get

$$X_{n+1} - X_* = \frac{1}{2}F'(X_n)^{-1}\begin{bmatrix} (X_n - X_*)^*\nabla^2 f_1(T_1)(X_n - X_*) \\ \vdots \\ (X_n - X_*)^*\nabla^2 f_N(T_N)(X_n - X_*) \end{bmatrix}.$$

Theorem. *If F is $\mathcal{C}^2$-smooth, $F(X_*) = 0$ and the Jacobian matrix $F'(X_*)$ is invertible then there is a (small) convex neighborhood U of the point X_* and a (large) constant M such that starting from $X_0 \in U$ each step X_n is well-defined, belongs to U and*

$$\big\|X_{n+1} - X_*\big\| \le M\big\|X_n - X_*\big\|^2.$$

In particular

$$\|X_n - X_*\| \le \mu^{2^n}/M \to 0 \quad \text{if} \quad \mu := M\|X_0 - X_*\| < 1.$$

Example: 2D GPS-problem. *Let* $S_1, S_2, S_3 \in \mathbb{R}^2$ *be three given distinct points. Determine the coordinates of a point* $P \in \mathbb{R}^2$ *which satisfies the equiations*

$$\delta_k := d(P, S_k) - d(P, S_{k+1})) = \big\|P - S_k\big\| - \big\|P - S_k\big\| \qquad (k = 1, 2)$$

concerning distance differences.

Newton iteration for solving the equation $F(P) = 0$ where

$$F(X) := \begin{bmatrix} d(X, S_1) - d(X, S_2) - \delta_1 \\ d(X, S_2) - d(X, S_3) - \delta_2 \end{bmatrix} \qquad (X \in \mathbb{R}^2).$$

In terms of the canonical coordinates $X \equiv (x, y)$, $S_k \equiv (p_k, q_k)$ we can write

$$F(X) = \begin{bmatrix} \sqrt{(x-p_1)^2 + (y-q_1)^2} - \sqrt{(x-p_2)^2 + (y-q_2)^2} - \delta_1 \\ \sqrt{(x-p_2)^2 + (y-q_2)^2} - \sqrt{(x-p_3)^2 + (y-q_3)^2} - \delta_2 \end{bmatrix}.$$

Notice that the firsr order Taylor approximation of F around the point $X_n \equiv (x_n, y_n)$ of the iteration is

$$F(X_n + V) \approx^{(1)} F(X_n) + F'(X_n)V = F(X) + \frac{d}{d\tau}\Big|_{\tau=0} F(X_n + \tau V) =$$
$$= F(x_n, y_n) + \frac{d}{d\tau}\Big|_{\tau=0} F(x_n + \tau v_n, y_n + \tau w_n).$$

In general,

$$\frac{d}{d\tau}\Big|_{\tau=0} F(x + \tau v, y + \tau w) = \begin{bmatrix} \frac{(x-p_1)v + (y-q_1)w}{\sqrt{(x-p_1)^2+(y-q_1)^2}} - \frac{(x-p_2)v + (y-q_2)w}{\sqrt{(x-p_2)^2+(y-q_2)^2}} \\ \frac{(x-p_2)v + (y-q_2)w}{\sqrt{(x-p_2)^2+(y-q_2)^2}} - \frac{(x-p_3)v + (y-q_3)w}{\sqrt{(x-p_3)^2+(y-q_3)^2}} \end{bmatrix}.$$

If X_n is known then

$X_{n+1} = X_n + V$, where V is the solution of the linear equation $F(X_n) + F'(X_n)V = 0$.

With given data: Let $S_1 \equiv (0, 10), S_2 \equiv (8, 6), S_3 \equiv (10, 0)$; assume P lies near to the point $X_0 \equiv (\mathbf{4.8}, \mathbf{6.4})$ (whose respective distances from S_1, S_2, S_3 are $6, 3.225, 8.246$), but $\delta_1 = d(P, S_1) - d(P, S_2) = 3$, $\delta_2 = d(P, S_2) - d(P, S_3) = -5$. Then

$$X_1 \equiv (4.8 + v, 6.4 + w), \quad \text{where} \quad F(4.8, 6.4) + \frac{d}{d\tau}\Big|_{\tau=0} F(4.8 + \tau u, 6 - 4 + \tau w) = 0,$$

Here we have

$$F(4.8,6.4)=\begin{bmatrix}\sqrt{4.8^2+(6.4-10)^2}-\sqrt{(4.8-8)^2+(6.4-6)^2}-3\\ \sqrt{(4.8-8)^2+(6.4-6)^2}-\sqrt{(4.8-10)^2+6.4^2}+5\end{bmatrix}=\begin{bmatrix}-0.225\\ -0.021\end{bmatrix},$$

$$\tfrac{d}{d\tau}\big|_{\tau=0}F(4.8+\tau u,6.4+\tau w)=\begin{bmatrix}\frac{4.8v+(6.4-10)w}{6}-\frac{(4.8-8)v+(6.4-6)w}{3.225}\\ \frac{(4.8-8)v+(6.4-6)w}{3.225}-\frac{(4.8-10)v+6.4w}{8.246}\end{bmatrix}.$$

Hence $v=0.092,\ \ w=-0.083,\ \ X_1\equiv(\mathbf{4.892},\mathbf{6.317})$.

Then $d(X_1,S_1)=6.123,\ \ d(X_1,S_2)=3.124,\ \ d(X_1,S_3)=8.124,$

The distance differences became improved: $\overline{\delta}_1=2.999,\ \ \overline{\delta}_2=-5.124.$

Inner convergence estimate for Newton iteration

$D\subset\mathbb{R}^N$ compact region, $f:D\to\mathbb{R}^N$ C^2-smooth, $f'(x)$ invertible $(x\in D)$

$$M:=\max\Big\{\frac{1}{2}\big\|f'(x_1)^{-1}f''(x_2)\big\|:x_1,x_2\in D\Big\}.$$

$x_0\in D$, $x_{k-1}:=x_k-f'(x_k)f(x_k)$ well-defined $(k=1,\dots,n)$.

Question. If x_n and x_{n+1} are "close", how far can be a root of f from x_{n+1}?

Assumption: $f(x_*)=0,\quad [x_n,x^*]\subset D$

$$0=f(x_*)=f(x_n)+f'(x_n)(x_n-x_*)+\underbrace{\frac{1}{2}f''_{[x_n,x_*]}}_{\int_{t_1=0}^1\int_{t_2=0}^1 f''(x_n+t_2(x_*-x_n))\,dt_2\,dt_1}(x_*-x_n)^2$$

$$y_*:=x_n-x_*$$

$$y_*=-f'(x_n)^{-1}f(x_n)-\frac{1}{2}f'(x_n)^{-1}f''_{[x_n,x_*]}y_*^2$$

$$y_*=T(y_*),\qquad T(y):=\underbrace{-f'(x_n)^{-1}f(x_n)}_{x_{n+1}-x_n}-\frac{1}{2}f'(x_n)^{-1}f''_{[x_n,x_*]}y^2$$

Recall. According to Brower's Fixed Point Theorem, if a continuous mapping T maps a closed ball (or a topologically equivalent figure to a ball) into itself then T admits a fixed point $(T(x)=x)$. Hence
if for some $\delta>0$ *we have* $T:\overline{B(\delta)}=\{y\in\mathbb{R}^N:\|y\|\le\delta\}\to\overline{B(\delta)}$,
then $\exists\ y_*\in\overline{B(\delta)}\qquad y_*=T(y_*)$.

Thus if there exists $\delta > 0$, such that $x_n + \overline{B(\delta)} \subset D$ and $\|T(y)\| \le \delta$ whenever $\|y\| \le \delta$, then, for any value δ with the above property, we can find a point $x_* \in D$, such that $f(x_*) = 0$ and $\|x_n - x_*\| \le \delta$.

How large can be the value of such a δ?

$$\begin{aligned} \|T(y)\| \le \delta \quad (\|y\| \le \delta), \quad \Leftarrow \quad & \|x_{n+1} - x_n\| + M\delta^2 \le \delta, \\ & M\delta^2 - \delta + \|x_{n+1} - x_n\| \le 0, \\ & \delta \in \left[\frac{1-\sqrt{1-4M\|x_{n+1}-x_n\|}}{2M}, \frac{1+\sqrt{1-4M\|x_{n+1}-x_n\|}}{2M} \right]. \end{aligned}$$

Theorem. *If we have* $4M\|x_{n+1} - x_n\| \le 1$ *and* D *contains the closed ball of radius*

$$\delta_n := \left[1-\sqrt{1-4M\|x_{n+1}-x_n\|}\right]/(2M) \quad \textit{centered at } x_n$$

then f *admits a root* $x_* \in D$ *lying within the distance* δ_n *from* x_n.

Corollary. *Since* $M\delta_n \le 1/2 < 1$ *and* $\|x_n - x_*\| \le \delta_n$, *we have*

$$\|x_{n+k} - x_*\| \le M^{2^k-1}\delta_n^{2^k} \searrow 0 \qquad (k \to \infty)$$

during the continuation of the Newton iteration.

CHAOS WITH NEWTON ITERATION

Recall. $f : \mathbb{R} \to \mathbb{R}$ C^2-smooth, $f(x_*) = 0$, $f'(x_*) \neq 0$.

Starting point x_0,

x_{n+1}: location of the zero of the 1st-order
Taylor polynomial of f around x_n

Heuristics: $f \approx \big[x \mapsto f'(x_n)(x - x_n)\big]$ around x_n.

IN SEVERAL DIM as well: $x_{n+1} := x_n - f'(x_n)^{-1} f(x_n)$,

If $f : \mathbb{R}^N \to \mathbb{R}^N$ is C^2-smooth, the same notations can be used (FRÉCHET).

Discussion: Assume U is a convex neighborhood of x_* and
$f'(x)$ is invertible for all $x \in U$.

Then

$$\left.\begin{aligned} 0 &= f(x_*) \\ 0 &= f(x_n) + f'(x_n)(xn + 1 - x_n) \end{aligned}\right\} \quad \begin{aligned} f(x_*) &=^{h:=x_*-x_n} = f(x_n + h) = \\ &= f(x_n) + f'(x_n)h + \int_{t=0}^{1} \int_{s=0}^{t} f''(x_n + sh)h^2 \, dsdt \end{aligned}$$

$$x_{n+1} - x_* = f'(x_n)^{-1} \int_{t=0}^{1} \int_{s=0}^{t} f''(x_n + sh)(x_n - x_*)^2 \, dsdt.$$

if $x_n \in U$ and ,

$$\|x_{n+1} - x_*\| \leq M_U \|x_n - x_*\|^2, \quad \text{where} \quad M_U := \frac{1}{2} \max_{x \in U} \|f'(x)^{-1} \max_{y \in U} \|f''(y).$$

Good starting guess: $x_0 \in M_U$ where $M_U \|x_0 - x_*\| < 1$.

Then *induction* yields that

$$\|x_n - x_*\| \leq \big[M_U \|x_0 - x_*\|\big]^{2^n - 1} \|x_0 - x_*\| \searrow 0 \quad (n = 1, 2, \ldots).$$

Case of arbitrary starting point.

We can encounter chaos even if looking for roots of polynomals (in 1 variable).

Observation. *Regarding a holomorphic function $f : \mathbb{C} \to \mathbb{C}$ as a real mapping $\mathbb{R}^2 \to \mathbb{R}^2$, the steps of the Newton iteration can be written as*

$$z_{n+1} := N(z_n), \quad \text{where} \quad N(z) := z - \frac{f(z)}{f'(z)} \quad \text{compl. diff.}$$

Recall. (A celebrated theorem of **Sharkovsky-Lee-York**)

The iterations of a non 3-periodic continuous self mapping of $[-\infty, \infty]$ are chaotic provided it admits a fixed point of 3rd order.

[Namely fixed points of all orders appear and there is an uncountably infinite set whose points never map to fixed points of higher order (chaotic points).

Let us apply the above mapping $N(\cdot)$ of the Newton iteration to a real polynomial

$$f(z) := \prod_{k=1}^{N} (z - \omega_k), \quad \omega_k \neq \omega_\ell \quad (k \neq \ell)$$

with N *simple roots* and extend it continuously to the *Riemann sphere* $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$ by setting $N(\infty), N(\eta_1), \ldots, N(\eta_{N-1}) := \infty$ at the roots $\eta_k (\in \mathbb{R})$ of the derivative $N'(\cdot)$. Given a fixed point z of 3rd order of $N(\cdot)$, we have

$$z = N^{\circ 3}(z) = N\Big(N\big(N(z)\big)\Big), \quad N(u) = u - \frac{1}{\sum_{k=1}^{N} \frac{1}{u - \omega_k}}.$$

Since $N(\cdot)$ is a rational function with real coefficients, the function $z \mapsto z - N^{\circ 3}(z)$ is also rational with real coefficients. Thus we can write the equation for the fixed point of 3rd order in the polynomial form

$$0 = z - N^{\circ 3}(z) = \frac{P(z)}{Q(z)} \iff P(z) = 0.$$

This admits complex solutions by the Fundamental Theorem of the Algebra. Moreover, since $P(\cdot)$ (and also $Q(\cdot)$) is a real polynomial, $z - N^{\circ 3}(z)$ has a real root provided its degree is odd. Therefore $N(\cdot)$ admits a fixed point of 3rd order in such a case.

Thus we obtained:

Theorem. *The Newton iteration algorithm applied to a real polynomial is either 3-periodic or chaotic with some starting point, provided it has only simple real roots and the numerator of the rational polynomial $z - N^{\circ 3}(\cdot)$ admits a real root (e.g. in case of odd degree).*

Remark. 3-periodicity can be checked with *computer algebra* if the coefficients are given numbers in the form algebraic formulas with integers. [There are not too many such cases].

The theoretical investigation of which cases are 3-periodic, is harder.

Example. Determine the unique real root of the polynomial $f(x) := x^3 - x - \frac{1}{\sqrt{3}}$ with Newton iteration. Thus

$$x_{n+1} = N(x_n), \quad \text{where} \quad N(x) := x - \frac{f(x)}{f'(x)} = x - \frac{x^3 - x - \frac{1}{\sqrt{3}}}{3x^2 - 1} \quad (n = 0, 1, \ldots).$$

Observation. Given any point x the half line $(-\infty, 0] \times \times 0$ in the X-axis, we can draw exactly two tangent straight lines from x to the graph of f. Therefore

$$z < 0 \Longrightarrow N^{-1}\{z\} = \{x : N(x) = z\} = \left\{N_-^{-1}(z), N_+^{-1}(z)\right\},$$

$$\text{where} \quad N_-^{-1}(z) = \left[x < z : N(x) = z\right], \; N_+^{-1}(z) = \left[x > z : N(x) = z\right] .$$

Since f is concave on the interval $(-\infty, 0]$ and its local maximum is attained at the point $-1/\sqrt{3}$, furthermore since the tangent straight line to the graph of f at the point of inflection $\left(1/\sqrt{3}, f(1/\sqrt{3})\right)$ passes through the origin, we have

$$N_-^{-1} : \mathbb{R} \leftrightarrow \mathbb{R}, 0 \mapsto -1/\sqrt{12} \text{ increasing}, \quad N_+^{-1} : (-\infty, -1/\sqrt{3}] \leftrightarrow (-\infty, 0] \text{ increasing}$$

and hence

$$I_0 := [-1/\sqrt{3}, 0], \quad I_1 := N_-^{-1}1(I_0) \cap (-\infty, 0], \quad I_2 := N_-^{-1}(I_1) \cap (-\infty, -1/2]$$

are well-defined intervals with the ordering $I_2 < I_1 < I_0$, furthermore $N : I_2 \leftrightarrow I_1 \leftrightarrow I_0$ in increasing manner. The properties of the inverse N_+^{-1} on the right hand side give rise to the construction of the intervals

$$J_0 := N_+^{-1}(I_2), \quad J_1 := N_-^{-1}(J_0), \quad J_2 := N_-^{-2}(J_0) = N_-^{-1}(J_1)$$

where $J_k \subset I_k$ $(k = 1, 2, 3)$ and hence $N : J_0 \leftrightarrow J_1 \leftrightarrow J_2$. That is

$$J_2 \subset I_0, \quad N^3 = N \circ N \circ N : J_2 \leftrightarrow I_2.$$

Since the functions N, N^2, N^3 are increasing on the interval $I_0 = [-1/\sqrt{3}, 0]$, the relation $J_2 \subset I_2$ implies that the function $x - N^3(x)$ is ≤ 0 at the left end point of J_2 and it is ≥ 0 at the right end point. Thus it admits a root in J_2 by Bolzano's classical theorem, that is exists a point $z_* \in J_2$, which is a fixed point of 3rd order of the Newton iteration map N: $\left(z_* = N^3(z_*) = N(N(N(z_*))\right)$. This fact implies the (*Sharkovsky* type) chaotic behaviour of the Newton iteration.


num0_eng_fig1.pdf
3 pages

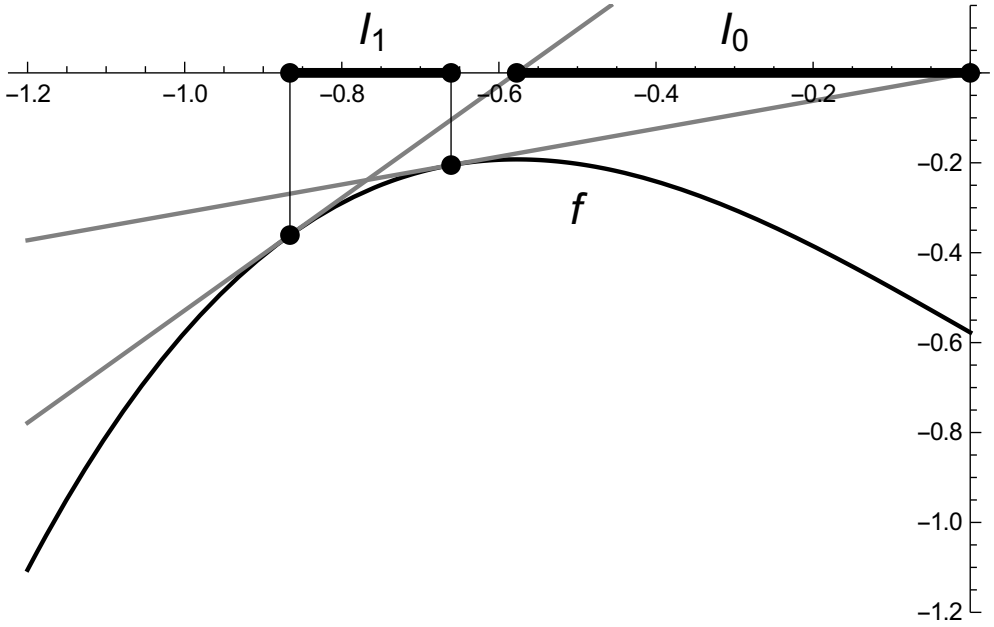


Figure by Prof. R. Vajda

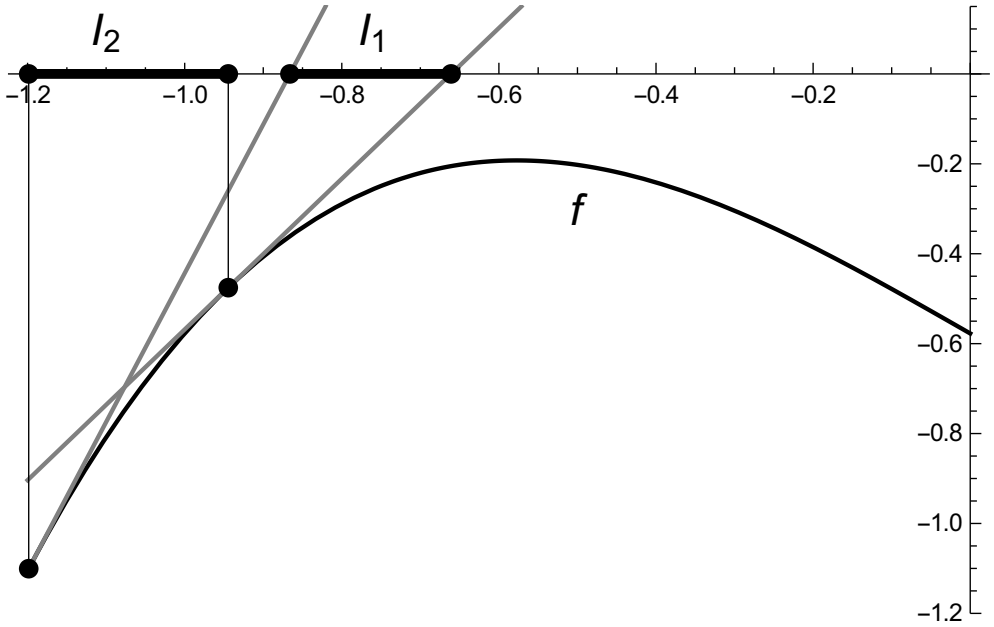


Figure by Prof. R. Vajda constructed with Wolfram Mathemaica 12

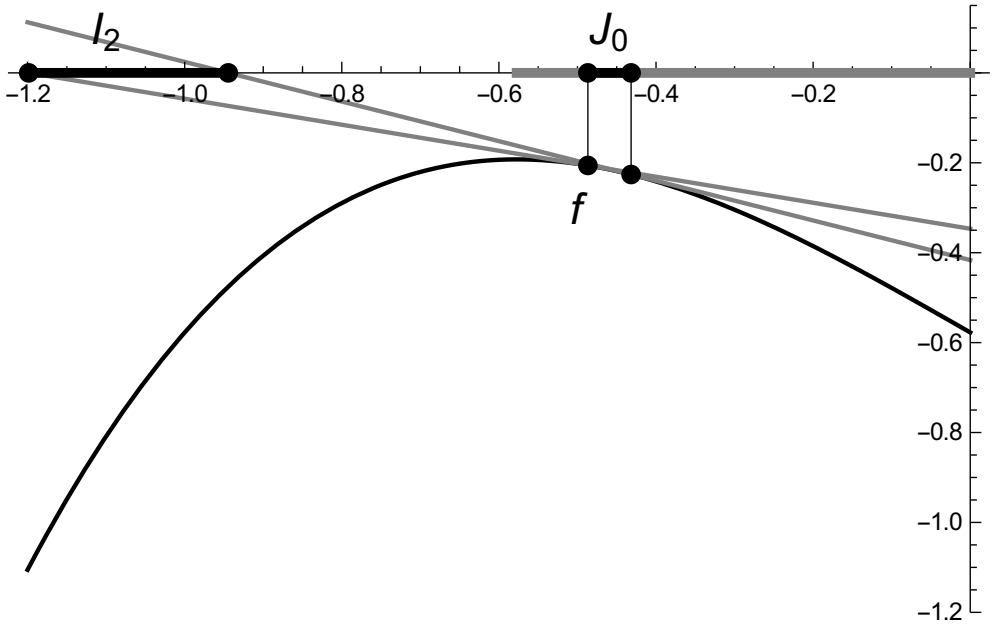


Figure by Prof. R. Vajda constructed with Wolfram Mathemaica 12

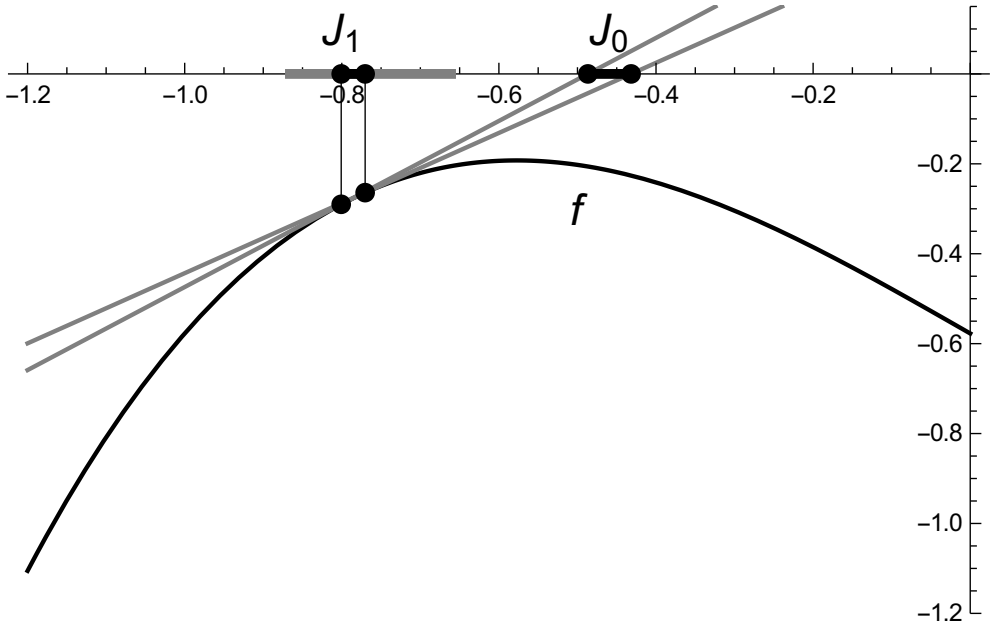


Figure by Prof. R. Vajda constructed with Wolfram Mathemaica 12

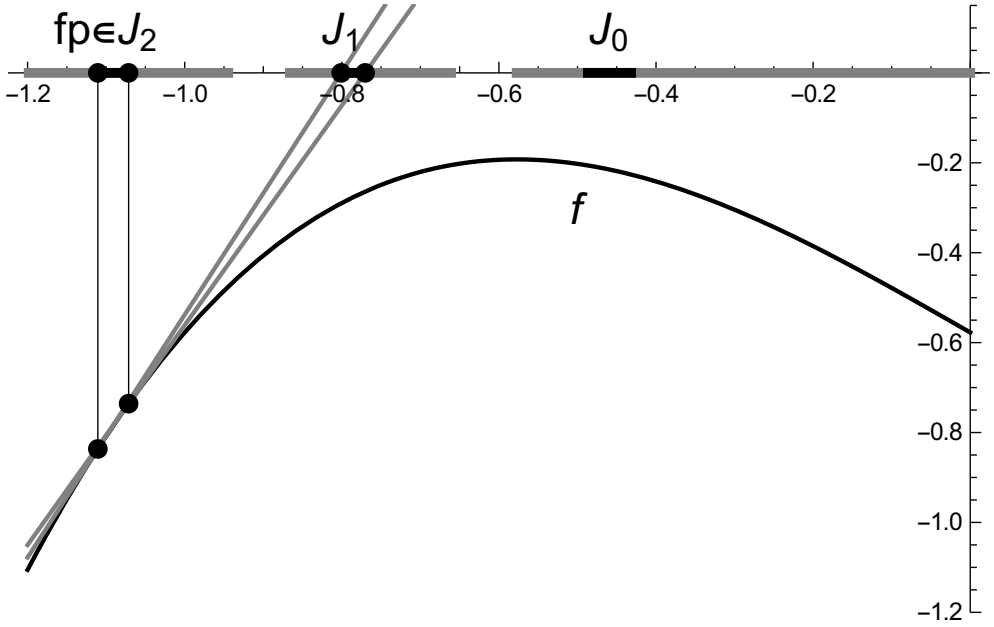


Figure by Prof. R. Vajda constructed with Wolfram Mathemaica 12

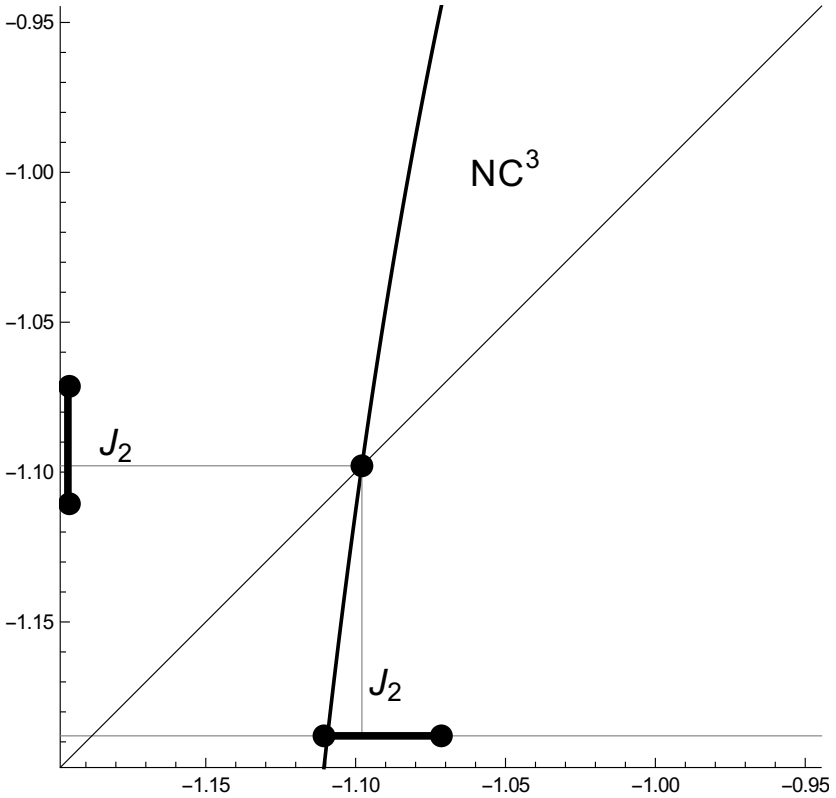


Figure by Prof. R. Vajda constructed with Wolfram Mathemaica 12

CONDITION NUMBERS

Recall. The number $\widetilde{x} \in \mathbb{R}$ approaches the number $x \in \mathbb{R}$
with (*absolute*) *error* $|\widetilde{x} - x|$,
with *relative error* $|\widetilde{x} - x|/|x|$.

In general: In a normed space $(\mathbf{X}, \|.\|)$,
the point $\widetilde{x} \in \mathbf{X}$ approaches the point $x(\in \mathbf{X})$ with *absolute* resp. *relative error]*

$$\mathrm{err}_x(\widetilde{x}) := \|\widetilde{x} - x\|, \qquad \mathrm{rel}_x(\widetilde{x}) := \frac{\|\widetilde{x} - x\|}{\|x\|}.$$

Definition. Let U open $\subset \mathbf{X}$, and $f : U \to \mathbf{X}$. The *condition number* of the function (mapping) f at the point $x \in U$ resp. on the region $U \subset X$ is

$$\mathrm{cond}_x(f) := \limsup_{\widetilde{x} \to x} \frac{\mathrm{rel}_{f(x)}\big(f(\widetilde{x})\big)}{\mathrm{rel}_x(\widetilde{x})}, \qquad \mathrm{cond}_U(f) := \sup_{x \in U} \mathrm{cond}_x(f).$$

Lemma. *If f is C^1-smooth then* $\mathrm{cond}_x = \|x\| \cdot \|f'(x)\|/\|f(x)\|$.

Proof.

$$\begin{aligned}
\mathrm{cond}_x(f) &= \limsup_{v \to 0} \frac{\|f(x+v) - f(x)\|/\|f(x)\|}{\|(x+v) - x\|/\|x\|} = \\
&= \limsup_{v \to 0} = \frac{\|f'(x)v + o(\|v\|)\|/\|f(x)\|}{\|v\|/\|x\|} = \\
&= \limsup_{v \to 0} \frac{\|f'(x)v\|}{\|v\|} \cdot \frac{\|x\|}{\|f(x)\|} = \frac{\|f'(x)\| \cdot \|x\|}{\|f(x)\|}.
\end{aligned}$$

Condition number of a matrix.

As usually, we identify the real $(N \times N)$-es matrix
with the linear mapping $x \mapsto (\mathbb{R}^N \to \mathbb{R}^N)$.

In such manner, if $x \neq 0$ and $0 \neq U$ then

$$\mathrm{cond}_x(A) = \limsup_{v \to 0} \frac{\|Av\|}{\|v\|} \cdot \frac{\|x\|}{\|Ax\|} = \|A\| \frac{\|x\|}{\|Ax\|}.$$

Thus if A is invertible and U is an arbitrary neighborhood of 0 then

$$\mathrm{cond}_{U\setminus\{0\}}(A) = \sup_{x\neq 0} \mathrm{cond}_x(A) = \|A\| \sup_{x\neq 0} \frac{\|x\|}{\|Ax\|} =^{y=Ax}$$

$$= \|A\| \sup_{y\neq 0} \frac{\|A^{-1}y\|}{\|y\|} = \|A\| \cdot \|A^{-1}\|.$$

Definition. The *condition number of the matrix* A is

$$\mathrm{cond}(A) := \|A\| \cdot \|A^{-1}\| \ .$$

We obtain with direct calculation:

Proposition. (1) $\mathrm{cond}_x(f+g) \leq \mathrm{cond}_x(f) + \mathrm{cond}_x(g)$,
(2) $\mathrm{cond}_x(f \circ g) \leq \mathrm{cond}_{g(x)}(f) \cdot \mathrm{cond}_x(g)$,
(3) $1 \geq \mathrm{cond}(A) = \mathrm{cond}(A^{-1}) \leq \mathrm{cond}(B) \cdot \mathrm{cond}(C)$ *ha* $A = BC$,
(4) *With respect to* L^2*-norm, if* Q *is an orthogonal matrix then*
$\mathrm{cond}(Q) = 1$ *and* $\mathrm{cond}(QX) = \mathrm{cond}(XQ) = \mathrm{cond}(X)$.

LINEAR ERROR ANALYSIS

Basic setting:

$A, \delta A \in \mathrm{Mat}(N, N, \mathbb{R})$, $b, \delta b, x, \delta x \in \mathrm{Mat}(N, 1, \mathbb{R})$

are given matrices resp. column vectors

(here δ is no operator, but only a notation for the "error term") where

$$Ax = b \qquad \text{(is the ideal equation)},$$

$$(A + \delta A)(x + \delta x) = b + \delta b \quad \text{(is the computed solution)}.$$

Assumption (technical)**:** $\det(A) \neq 0$, $\|A^{-1}\delta A\| < 1$.

Remark. In the basic setting, $A + \delta A$ is invertible. Namely

$$(A + \delta A)^{-1} = \big[A(1 + A^{-1}\delta A)\big]^{-1} = (1 + A^{-1}\delta A)^{-1}A^{-1} =$$
$$= \sum_{n=0}^{\infty}(-1)^n\big[A^{-1}\delta A\big]^n A^{-1},$$
$$\|(A + \delta A)^{-1}\| \leq \sum_{n=0}^{\infty}\big\|A^{-1}\delta A\big\|^n\|A^{-1}\| = \frac{\|A^{-1}\|}{1 - \|A^{-1}\delta A\|}.$$

Recall: $\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|$ is the condition number of the matrix A.

Fundamental Theorem (on data sensibility)**.**

Under the above hypothesis, we can estimate the relative errors as

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\delta A\|}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).$$

Proof. $(A + \delta A)(x + \delta x) = b + \delta b$, $Ax = b$, $\implies \underline{(\delta A)x + (A + \delta A)\delta x = \delta b}$,

$$\delta x = (A + \delta A)^{-1}\big[\delta b - (\delta A)x\big],$$
$$\|\delta x\| \leq \|(A + \delta A)^{-1}\|\big[\|\delta b\| + \|\delta A\|\,\|x\|\big],$$

$$\frac{\|\delta x\|}{\|x\|} \leq \|(A + \delta A)^{-1}\|\left[\frac{\|\delta b\|}{\|x\|} + \|\delta A\|\right].$$

Observation: $$\frac{1}{\|x\|} = \frac{1}{\|A^{-1}b\|} \leq \frac{1}{\|b\| \inf\{\|A^{-1}e\| : \|e\| = 1\}} =$$
$$= \frac{1}{\|b\|} \frac{1}{\inf\{\|A^{-1}y\|/\|y\| : y \neq 0\}} =$$
$$= \frac{1}{\|b\|} \sup\{\|y\|/\|A^{-1}y\|/ : y \neq 0\} =$$
$$= \frac{1}{\|b\|} \sup\{\|Az\|/\|z\|/ : z \neq 0\} = \frac{1}{\|b\|}\|A\|.$$

Thus
$$\frac{\|\delta x\|}{\|x\|} \leq \|(A+\delta A)^{-1}\| \left(\frac{\|\delta b\|}{\|b\|}\|A\| + \|\delta A\| \right) \leq$$
$$\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\delta A\|}\|A\| \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right). \quad \text{Qu.e.d.}$$

Corollary. $\|A^{-1}\| \, \|\delta A\| < 1 \Longrightarrow$

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\|\delta A\|/\|A\|} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right).$$

Remark. In our above considerations, we can use
any matrix norm defined by any (vector) norm $\|\cdot\|$ on $\mathbb{R}^N \equiv \mathrm{Mat}(N, 1, \mathbb{R})$.

We apply mostly the Euclidean norm

$$\|z\|_2 := \big[\sum_{k=1}^{N} z_k^z \big]^{1/2} \qquad \text{(for vectors)},$$

$$\|B\|_2 := \sup \big\{ \|Bz\|_2 : \|z\|_2 = 1 \big\} \quad \text{(for matrices)}.$$

It is an interesting geometrical fact that

$$\frac{1}{\mathrm{cond}^{\|\cdot\|_2}(A)} = \min \left\{ \frac{\|\delta A\|_2}{\|A\|_2} : A + \delta A \text{ is not invertible} \right\}.$$

Lower estimate for $\dfrac{\|\delta x\|}{\|x\|}$**:** $\quad \delta A = 0 \Longrightarrow \ \dfrac{1}{\mathrm{cond}(A)} \dfrac{\|\delta b\|}{\|b\|} \le \dfrac{\|\delta x\|}{\|x\|}.$

Proof. $\quad \delta A = 0 \Rightarrow A\delta x = \delta b \Rightarrow \|\delta x\| \ge \|\delta b\|/\|A\|.$

Since $\quad \|x\| = \|A^{-1}b\| \le \|A^{-1}\|\,\|b\|, \quad$ we obtain

$$\|\delta x\|/\|x\| \ge [\|\delta b\|/\|A\|]/[\|A^{-1}\|\,\|b\|] = \|\delta b\|/[\mathrm{cond}(A)\|b\|].$$

Estimation componentwise in case of ($\delta b = 0$ resp. $\delta A = 0$)**.**

Notation. $\quad |Z| := \left[|z_{ij}|\right]_{i=1\,j=1}^{K\quad M}$ for any $K \times M$ matrix;
$e^{[i]} := [i.$ unit vector$]$.

Theorem. *Assume we have*

$$|\delta A| \le \gamma |A|, \quad |\delta b| \le \gamma |b|$$

in the equation $(A + \delta A)(x + \delta x) = b + \delta b$ *on data sensibility.*

Then, in terms of the column vectors

$$s^{[i]} := [e^{[i]}]^{\mathrm{T}} A^{-1}$$

of the matrix A^{-1} *we can conclude the following estimates.*

(1) *In case of* $\delta b = 0$ *and given* $\widehat{x} := x + \delta x$ *we have* $\quad |\delta x_i| \le \gamma |s^{[i]}|\,|A|\,|\widehat{x}|$*;*

(2) *In case of* $\delta A = 0$ *we have* $\quad |\delta x_i| \le \gamma |s^{[i]}||b|, \quad \dfrac{|\delta x_i|}{|x_i|} \le \gamma \dfrac{|s^{[i]}||b|}{|s^{[i]}b|}.$

Proof. $(A + \delta A)(x + \delta x) = b + \delta b \Rightarrow$

$$\begin{aligned} A\delta x &= -(\delta A)(x + \delta x) + \delta b, \\ \delta x &= A^{-1}[-(\delta A)(x + \delta x) + \delta b], \\ \delta x_i &= [e^{[i]}]^{\mathrm{T}} \delta x = [e^{[i]}]^{\mathrm{T}} A^{-1}[-(\delta A)(x + \delta x) + \delta b] = \\ &= s^{[i]}[-(\delta A)(x + \delta x) + \delta b]. \end{aligned}$$

(1) IF $\delta b = 0, \quad |\delta x_i| \le |s^{[i]}|\,|\delta A|\,|\underbrace{x + \delta x}_{\widehat{x}}| \le \gamma |s^{[i]}|\,|A|\,|\widehat{x}|;$

(2) IF $\delta A = 0, \quad |\delta x_i| \le |s^{[i]}|\,|\delta b| \le \gamma |s^{[i]}|\,|b|,$

$$\frac{|\delta x_i|}{|x_i|} \le \gamma \frac{|s^{[i]}||b|}{|e^{[i]}A^{-1}b|} \le \gamma \frac{|s^{[i]}||b|}{|s^{[i]}b|}. \qquad \text{Q.e.d.}$$

A priori analysis backwords

Basic setting: $\widehat{x}=x+\delta x=Cb$.

E.g. $C=\big[A^{-1}$ with rounding errors$\big]$ not computed explicitly;

A priori assumption: $C=(A+\delta A)^{-1}$, where δA not know for us.

Lemma. *If the norm of* $R:=I-AC$ *is* <1 *then* C *is invertible and*

$$\|A^{-1}\| \le \frac{\|C\|}{1-\|R\|}, \quad \frac{\|R\|}{\|C\|} \le \|C-A^{-1}\| \le \frac{\|C\|\,\|R\|}{1-\|R\|}.$$

Proof. $R=I-AC \Rightarrow C=A^{-1}(I-R)$ is a product of invertible matrice if, $\|R\|<1$.
$R=I-AC \Rightarrow A=(I-R)C^{-1} \Rightarrow A^{-1}=C(I-R)^{-1}=C\sum\limits_{n=0}^{\infty}R^n \Rightarrow \|A^{-1}\|\le\|C\|\sum\limits_{n=0}^{\infty}\|R\|^n$.
$R=I-AC \Rightarrow R=A(A^{-1}-C) \Rightarrow \|R\|\le\|A\|\,\|A^{-1}-C\|$.
Here we have $A^{-1}-C=C(I-R)^{-1}-C=C\sum\limits_{n=0}^{\infty}R^n \Rightarrow \|A^{-1}-C\|\le\|C\|/(1-\|R\|)$.

Lemma. *In case of* $C=\widehat{A}^{-1}=(A+\delta A)^{-1}$ *and* $\|R\|<1$ *we have*

$$\|\delta A\| \le \|R\|\,\|C^{-1}\| \le \frac{\|R\|\,\|A\|}{1-\|R\|}.$$

Proof. Observation: $\delta A=C^{-1}-A=(I-AC)C^{-1}=RC^{-1}$.
Here we have $C^{-1}=[A^{-1}(I-R)]^{-1}=(1-R)^{-1}A$
whence the norm estimates are immediate.

A posteriori analysis

Basic setting: Given an approximate solution

$$y(=x+\delta x) \quad \text{of the equation } Ax=b \text{ (with } \det(A)\neq 0).$$

Let us estimate the residual- resp error vectors

$$r:=Ay-b\big(=(\delta A)(x+\delta x)-\delta b\big), \quad e:=A^{-1}r.$$

On the basis of our earlier results we can conclude the following.

Lemma. $\|e\| \le \|r\|\,\|A^{-1}\| \le \dfrac{\|r\|\,\|C\|}{1-\|R\|}, \quad \dfrac{\|e\|}{\|x\|} \le \mathrm{cond}(A)\dfrac{\|r\|}{\|b\|}.$

The effect of the inaccuracy in machine representations

Let $\widehat{r} = r + \delta r := \big[\text{machine output for } r(= b - Ay)\big]$.

Assumption. $A \in \mathbb{R}^{N\times N}$, and by writing u for the machine-0,

$$|\delta r| \le \gamma_{N+1}(|A|\,|y| + |b|), \quad \text{ahol } \ \gamma_{N+1} := \frac{(N+1)u}{1-(N+1)u}.$$

Theorem. *Then with the norm* $\|.\| := \|.\|_\infty$ *we have*

$$\frac{\|e\|_\infty}{\|y\|_\infty} \le \frac{\big\||A^{-1}|\big(|\widehat{r}| + \gamma_{n+1}(|A|\,|y| + |b|)\big)\big\|_\infty}{\|y\|_\infty}.$$

Rounding in LU-factorization

Machine rounding: The machine output of the operations $\diamondsuit(= +, -, \cdot, /)$ is

$$\widehat{\big[x \diamondsuit y\big]} = (1 + \delta_{\diamondsuit,x,y})(x \diamondsuit y), \quad |\delta_{\diamondsuit,x,y}| \le u := [\text{machine 0}].$$

Assumption. $(A + \delta A) = (L + \delta L)(U + \delta U)$ where
$A = LU$, L is a lower-triangular matrix with main diagonal 1,
U is an upper-triangular matrix, both of $N \times N$ type,
the machine outputs are $\widehat{A} := A + \delta A$, $\widehat{L} := L + \delta L$,
$\widehat{U} := U + \delta U$ where $u := \big[\text{machine-0}\big]$.

One can establish the following fact.

Theorem. *Under the above assumption,*

$$A = LU,\ \widehat{A} = \widehat{L}\widehat{U},\ [\text{main diagonal of } \widehat{L}\ ] = 1 \quad \Longrightarrow \quad |\delta A| \le \frac{Nu}{1-Nu}|\widehat{L}|\,|\widehat{U}|.$$

Excercise. Verify the numerical solution for a system of linear equations

Theoretical problem: $Ax = b$. We are given a numerical solution $\widetilde{x}$. That is

$$\widetilde{A}\widetilde{x} = \widetilde{b}, \quad x = \widetilde{x} + e, \quad A = \widetilde{A} + E, \quad b = \widetilde{b} + d,$$

where we have estimates for the errors d resp. E. In the most general setting

$$d \in \mathcal{D}, \qquad E \in \mathcal{E},$$

where the family $\mathcal{D}$ of vectors and the family $\mathcal{E}$ of matrices is given.

Question: How much is $\widetilde{x}$ suitable instead of x?

Estimate in terms of $\mathcal{D}, \mathcal{E}$.

$$(\widetilde{A} + E)(\widetilde{x} + e) = \widetilde{b} + d,$$
$$E\widetilde{x} + Ee + \widetilde{A}e = d,$$
$$e = (\widetilde{A} + E)^{-1}(d - E\widetilde{x}).$$

Starting argument: $e \in \Big\{ \big(\widetilde{A} + E\big)^{-1}(d - E\widetilde{x}) : \ E \in \mathcal{E}, \ d \in \mathcal{D} \Big\}$.

Example. $\mathcal{D} := \big\{d' : \ \|b'\| < \delta\big\}, \ \mathcal{E} := \big\{E' : \ \|E'\| < \varepsilon\big\}$. Ekkor

$$\|e\| \le \sup_{\|E\|<\varepsilon} \big\|(\widetilde{A} + E)^{-1}\big\|(\delta + \varepsilon\|\widetilde{x}\|).$$

In general $\big\|B^{-1}\big\| = \sup\limits_{\|z\|=1} \big\|B^{-1}z\big\| = \dfrac{1}{\inf_{\|y\|=1} \big\|By\big\|}$. Hence

$$\big\|(\widetilde{A} + E)^{-1}\big\| \le \frac{1}{\Big[\inf_{\|y\|=1} \|\widetilde{A}y\| - \varepsilon\Big]_+} = \frac{1}{\frac{1}{\|\widetilde{A}^{-1}\|} - \varepsilon},$$

provided $\|\widetilde{A}y\| > \varepsilon$ ($\|y\| = 1$). This latter can be rewritten in terms of te conditional number

$$K(B) := \big\|B^{-1}\big\| \, \|B\|$$

whence

$$\big(\big\|\big(\widetilde{A} + E\big)^{-1}\big\| \le \frac{\widetilde{A}^{-1}}{1 - \varepsilon\big\|\big(\widetilde{A} + E\big)^{-1}\big\|} = \frac{K(\widetilde{A})}{\|\widetilde{A}\| - \varepsilon K(\widetilde{A})},$$
$$\|e\| \le \frac{K\big(\widetilde{A}\big)}{\|\widetilde{A}\| - \varepsilon K\big(\widetilde{A}\big)} \, \big(\delta + \varepsilon\|\widetilde{x}\|\big).$$

INTERVAL ARITHMETICS

Basic concepts. As usually, by *intervals* we mean the subsets $(a,b),[a,b),(a,b],[a,b]$ with $a \le b \in \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. We regard $\overline{\mathbb{R}}$ as the 2-point compactification of $\mathbb{R}$. Namely the neighborhoods of a point $x \in \mathbb{R}$ consist of the sets $W \subset \overline{\mathbb{R}}$ such that $W \supset (x-\varepsilon, x+\varepsilon)$ for some $\varepsilon > 0$ while W is a neighborhood of ∞ resp. $-\infty$ iff $W \supset (z,\infty]$ resp. $W \supset [-\infty, z)$ for some $z \in \mathbb{R}$. In particular the closure of an interval I with endpoints a, b is $\overline{I} = [a,b]$ which is compact in any case with respect to this topology. The set of all subintervals resp. compact subintervals of the interval I will be denoted by

$$\mathrm{Intv}(I) := \{J \text{ interval: } J \subset I\}, \quad \overline{\mathrm{Intv}}(I) := \{[a,b] : a,b \in I,\ -\infty < a \le b < \infty\}.$$

In order to control the accuracy of calculating the value of a formula Φ at a given argument x, like $\Phi(x) = \dfrac{e^x}{\sin^2(x) + 2\cos^2 x}$, we consider x as belonging to some interval I containing the set of all representations of x with error (like 3.14 fo π) and, for each elementary step of the calculation, try to find an interval which includes the partial result.

In general, if $\mathbf{X_1}, \ldots, \mathbf{X_n}, \mathbf{Z}$ are arbitrary (non-empty) sets and $f : \mathbf{X_1} \times \cdots \times \mathbf{X_n} \to \mathbf{Z}$ is a mapping of n variables, we extend the effect of f to sets $X_k \subset \mathbf{X}_k$ in the natural manner as $f(X_1, \ldots, X_n) := \{f(x_1, \ldots, x_n) : x_k \in X_k \ (k = 1, \ldots, n)\}$. The classical algebraic operations $\diamondsuit = +, -, \cdot, /$ are regarded as functions of two variables with

$$X \diamondsuit Y := \big\{x \diamondsuit y \in \mathbb{R} \cup \{\pm\infty\} : x \in X,\ y \in Y\big\} \quad (X, Y \subset \mathbb{R} \cup \{\pm\infty\}).$$

Inclusion Principle. $X_1 \subset Y_1, \ldots, X_n \subset Y_n \Longrightarrow f(X_1, \ldots, X_n) \subset f(Y_1, \ldots, Y_n)$.

Theorem. $\overline{\mathrm{Intv}}(\mathbb{R}) \diamondsuit \overline{\mathrm{Intv}}(\mathbb{R}) \subset \overline{\mathrm{Intv}}(\mathbb{R}) \quad \big(\diamondsuit = +, -, \cdot\big)$**,**
$\overline{\mathrm{Intv}}(\mathbb{R}) / \overline{\mathrm{Intv}}(0,\infty),\ \overline{\mathrm{Intv}}(\mathbb{R}) / \overline{\mathrm{Intv}}(-\infty, 0) \ \subset \overline{\mathrm{Intv}}(\mathbb{R})$**.**
In any case we can write $[a,b] \diamondsuit [c,d]$ *in the form* $[a,b] \diamondsuit [c,d] = [\varphi(a,b,c,d), \psi(a,b,c,d)]$.

Proof. Recall: continuous functions map compact connected sets into compact connected sets and the compact connected subsets of $\mathbb{R}$ are exactly the closed bounded intervals. Also the operations $\diamondsuit = +, -, \cdot, /$ ($0 \notin$ denominator) are continuous and monotonic in their variables over the plane quartals $\mathbb{R}_\varepsilon \times \mathbb{R}_\sigma$ $(\varepsilon, \sigma = \pm)$. Hence we can write the end points as formulas of the end points a, b, c, d of the two intervals.

Example. $[a,b] \diamondsuit [c,d] = [\min\{a \diamondsuit c, a \diamondsuit d\}, \max\{b \diamondsuit c, b \diamondsuit d\}] \quad (\diamondsuit = \pm)$.

Exercise. 1) End points of $[a,b] \cdot [c,d]$. 2) End points of $[a,b]/[c,d]$.

In the sequel we need extensions for the inclusion property $I \subset J \Rightarrow f(I) \subset f(J)$.

Definition. Let $\mathbf{X}, \mathbf{Z} \neq \emptyset$ arbitrary sets and $F : \mathcal{X} \to \mathcal{Z}$ be a set mapping where $\mathrm{dom}(F) = \mathcal{X} \subset 2^{\mathbf{X}} (:= \{\text{subsets of } \mathbf{X}\})$ and $\mathcal{Z} \subset 2^{\mathbf{Z}}$. We say that F is *isotonic* if

$$F : \mathcal{X} \nearrow \mathcal{Z} \quad \text{that is if} \quad X, Y \in \mathcal{X} \text{ with } Y \subset X \Longrightarrow \quad F(Y) \subset F(X).$$

By an *interval function* we mean a map $F : \mathcal{X} \to \mathrm{Intv}(\overline{\mathbb{R}})$ where $\mathcal{X} \subset \overline{\mathbb{R}}$.
The domain $\mathcal{X}$ of an interval function F is *hereditary* [resp. *compact-hereditary*] if

$$\mathcal{X} \supset \mathrm{Intv}(I) \quad \big[\text{resp. } \mathcal{X} \supset \mathrm{Intv}_0(I)\big] \quad \text{for all } I \in \mathcal{X}.$$

The interval function F is said to be *regular* if F is isotonic with compact-hereditary domain and maps compact intervals to compact intervals.

Lemma. *The composition $F_1 \circ F_2 : I \mapsto F_1\big(F_2(I)\big)$ of two isotonic interval functions is isotonic.*

The family of regular interval functions is closed for composition.

Proof. Straightforward. (Notice: $\mathrm{dom}(F_1 \circ F_2) = \big\{I : F_2(I) \in \mathrm{dom}(F_1)\big\}$.)

Definition. The function $f : \mathbf{X} \supset \mathrm{dom}(f) \to \mathbf{Z}$ is *included in* (or *covered by*) the set function $F : 2^{\mathbf{X}} \supset \mathrm{dom}(F) \to 2^{\mathbf{Z}}$ if

$$f \prec F: \qquad \forall\, x \in \mathrm{dom}(f) \;\; \exists\, X \in \mathrm{dom}(F) \;\; \text{with} \;\; x \in X \text{ and } f(x) \in F(X).$$

We extend the concept of inclusion to set functions. If $F_k : 2^{\mathbf{X}} \supset \mathrm{dom}(F_k) \to 2^{\mathbf{Z}} \;\; (k = 1, 2)$,

$$F_1 \prec F_2: \quad \forall\, X_1 \in \mathrm{dom}(F_1) \;\; \exists\, X_2 \in \mathrm{dom}(F_2) \;\; \text{with} \;\; X_1 \subset X_2 \text{ and } F_1(X_1) \subset F_2(X_2).$$

Remark. 1) In terms of interval functions, $f \prec F$ if and only if $\big[\{x\} \mapsto \{f(x)\}\big] \prec F$.
2) Given a function by a sophisticated formula, we are going to construct inclusions for it with interval functions of simpler formulas on small intervals such that the graphs do not differ too much. This is the trick of classical mathematics in several cases when establishing estimates by constants or simple formulas. With computer, we can enhance this old techniques enormously.

Theorem. *Assume $F : \mathcal{X} \nearrow \mathrm{Intv}(\mathbb{R})$ is an interval function mapping intervals with equal end points to intervals with equal end points. Then the followig statements are equivalent:*
(i) *F admits a regular covering $\overline{F} \succ F$ such that*

$$\overline{F(I)} = \overline{F}[a, b] = [c, d] \;\; \text{whenever} \;\; I \in \mathcal{X} \text{ with } \overline{I} = [a, b] \text{ and } \overline{F(I)} = [c, d];$$

(ii) *F maps intersecting intervals into intersecting or intervals with intersecting closures.*

Proof. (i) $\Rightarrow$ (ii). Assume $\overline{F} \succ F$ and let $I_1, I_2 \in \mathcal{X}$ be intersecting intervals with $x \in I_1 \cap I_2$ and $\text{Endpoints}\big(F(I_k)\big) = \{c_k, d_k\}$ where $c_k \le d_k$. By assumption, the map $\overline{F}$ is isotonic with compact-hereditary domain and $\{x\} = [x,x] \in \text{dom}(\overline{F})$ with $\overline{F}\{x\} \subset \overline{F}(I_k) = [c_k, d_k]$ $(k = 1, 2)$. In particular $\emptyset \neq [c_1, d_1] \cap [c_2, d_2]$ and hence the intervals $F(I_k)$ with endpoints c_k, d_k have a common point or they are disjoint but admit a common endpoint (i.e. their closures have non-empty intersection).

(ii) $\Rightarrow$ (i). Define $\overline{\mathcal{X}} := \bigcup_{I \in \mathcal{X}} \text{Intv}(\overline{I})$ and let

$(*)$ $\qquad \overline{F}(J) := \bigcap \big\{ \overline{F(I)} : I \in \mathcal{X}_J \big\}$ where $\mathcal{X}_J := \big\{ I \in \mathcal{X} : J \subset \overline{I} \big\} \qquad (J \in \mathcal{X})$.

Key observation: for any interval $J \in \overline{\mathcal{X}}$ the family $\mathcal{X}_J$is in non-empty. By assumption, given any pair $I_1, I_2 \in \mathcal{X}_J$, the image intervals $F(I_k)$ have non-empty intersection or touch each other implying that $\overline{F(I_1)} \cap \overline{F(I_2)} \neq \emptyset$.

Recall: by Helly's Theorem (1-dimensional case only), a family of $\mathcal{K}$ compact intervals admits a common point if and only if $K_1 \cap K_2 \neq \emptyset$ for each pair $K_1, K_2 \in \mathcal{K}$.

In view of Helly's Theorem, our observation ensures that $\bigcap \big\{ \overline{F(I)} : I \in \mathcal{X}_J \big\}$ is a non-empty compact interval. We complete the proof with the observation that $\overline{F}[a,b] = [c,d]$ where

$$c = \sup \Big\{ \text{Leftend}\big(F(I)\big) : I \in \mathcal{X},\ \text{Endpints}(I) = \{\alpha, \beta\},\ \alpha \le a \le b \le \beta \Big\},$$
$$d = \inf \Big\{ \text{Rightend}\big(F(I)\big) : I \in \mathcal{X},\ \text{Endpints}(I) = \{\alpha, \beta\},\ \alpha \le a \le b \le \beta \Big\}.$$

Indeed, hence it is immediate that $\overline{F(J)} = \overline{F}[a,b] = [c,d]$ if $J \in \mathcal{X}$ with $\text{Endpoints}(J) = \{a,b\}$ and $\text{Endpoints}\big(F(J)\big) = \{c,d\}$.

Recall. In a metric space (X,d), the *diameter* of a set $A \subset X$ resp. the distance of a point $x \in X$ from a set $A \subset X$ are

$$\text{diam}(A) := \sup\{d(a,b) : a,b \in A\}, \qquad \text{dist}(x,A) := \inf\{d(x,a) : a \in A\}.$$

We shall use the *Hausdorff distance* $\text{dist}(A,B)$ for measuring the difference between the sets $A, B \subset X$ with respect to the underlying distance d:

$$\text{dist}(X,Y) = \big[\text{Hausdorff distance}\big] = \sup\big\{ d(x,Y), d(y,X) : x \in X, y \in Y \big\} =$$
$$= \max\big\{ \sup\{d(x,Y) : y \in Y \setminus X\}, \sup\{d(y,X) : x \in X \setminus Y\}.$$

Excercise. If $I = [a,b], J = [c,d]$ finite intervals, then

(i) $a \le c \le b \Rightarrow \quad \text{dist}(I,J) = \max\big\{ |c-a|, |b-d| \big\}$,

(ii) $b \le c \Rightarrow \quad \text{dist}(I,J) = \max\big\{ |a-d|, |b-c| \big\}$.

Lemma. (Covering Lemma).

Let $f \prec F$ *and* $X \subset \bigcup_{k=1}^{M} X_k$ *where* $X \in \mathrm{dom}(f)$ *and* $X_k \in \mathrm{dom}(F)$ *with* $X_k \cap X \neq \emptyset$. *Then*

(1) $f(X) \subset F(X_1) \cup \cdots \cup F(X_n)$,

(2) $\mathrm{dist}\Big(\mathrm{range}(f|X), \bigcup_{k=1}^{M} F(X_k)\Big) \leq \max_{k=1}^{M} \mathrm{diam}\big(F(X_k)\big)$.

Proof. (1) Given $x \in X$, $x \in X_k$ for some $k = k(x)$. Hence $f(x) \in F(X_{k(x)}) \subset \bigcup_{m=1}^{M} F(X_m)$.

(2) Since, by (1), $f(X) \subset \bigcup_{m=1}^{M} Y_m$, we simply have

$$\mathrm{dist}\big(f(X), \bigcup_{m=1}^{M} Y_m\big) = \sup\big\{\mathrm{dist}\big(f(x), \bigcup_{m=1}^{M} Y_m\big) : x \in X\big\}.$$

Let us write $Y_m := F(X_m)$ with $f(x) \in Y_{k(x)}$ $(x \in X)$ and for short. Observe that

$$\mathrm{dist}\big(f(x), \bigcup_{m=1}^{M} Y_m\big) \leq \mathrm{dist}\big(f(x), Y_{k(x)}\big) \leq \mathrm{diam}(Y_{k(x)}) \quad (x \in X).$$

Hence the statement is immediate.

Next we proceed to show some fundamental applications of interval analysis.

Root detection (solution of $f(x) = 0$ with $x \in I = [a, b]$)

1) "Catching the Lion"

Assume we know $f : I \to \mathbb{R}$ is continuous and $f(x) = 0$ for some $x \in I$. Let $I_0 := I$.

For $n = 1, 2, \ldots$ we construct $I_n = [a_n, b_n]$ with length $\leq 2^{-n}\mathrm{length}(I)$ as follows.

Suppose we have $\emptyset \neq \{x \in I_{n-1} : f(x) = 0\}$. divide I_n into two equal pieces

$$J_{n,0} := [a_{n-1}, c_{n-1}], \quad J_{n,1} := [c_{n-1}, b_{n-1}] \quad \text{with} \quad c_{n-1} := \mathrm{mid}(I_{n-1}) = \tfrac{a_{n-1}+b_{n-1}}{2}.$$

If $0 \notin \mathrm{range} f|J_{n,s}$ for some $s \in \{0, 1\}$ then there must be a root of f belonging to $J_{n,1-s}$.

In this case we define $I_n := J_{n,1-s}$.

Techniques: We try to use a suitable covering $F \succ f$ and testing if $0 \notin \mathrm{range}(F|J_{n,s}$ in order to see whether $0 \notin \mathrm{range}\big({}^{f}\Phi|J_{n,s}\big)$.

Remark. Our test with F is an example for the utility of the Covering Lemma.

As for the name "Catching the Lion": I is the Sahara, a root is lion, we divide the piece I_{n-1} of the Sahara with a fence into two pieces $J_{n-1,0}, J_{n-1,1}$ which become smaller than a lion for suitably large n.

In Classical Analysis, we encounter the following argument: "if there is a root in some interval J then there must be a root in its left or right half J_1 or J_2. *Choose k such that* $\exists\, x \in J_k \ f(x) = 0$." Logically it is impossible to check in finite steps if $k = 1$ or $k = 2$ is a suitable choice. With the application of the enclosure F "we cut this Gordian knot".

2) Interval version of Newton iteration

Setting. $f \in \mathcal{C}^1(I) \quad f(x_*) = 0, \quad x_* \in I = [a,b]$. Let also $F \succ f$ and $F' \succ f'$

$$x \in I, \Longrightarrow \; f(x) = f(x_*) + f'(\vartheta_x) = x - x_*) \quad \theta_x \in \operatorname{conv}\{x, x_*\},$$
$$x_* = x - \frac{f(x)}{f'(\theta_x)} \in x - \frac{f(x)}{F'(I)} \qquad \text{whenever} \quad 0 \notin F'(I).$$

Definition. In case of $F' \succ f'$, we introduce the *interval Newton operation* as

$$N_{f,F'} := x_J - \frac{f(x_J)}{F'(J)} \quad \text{with} \quad x_J := \operatorname{mid}(J) = \frac{a+b}{2}.$$

Theorem. *Assume* $0 \notin F'(I)$. *Then the iteration*

$$I_0 := I, \quad I_{n+1} := N_{f,F'}(I_n) \cap I_n \quad (n = 1, 2, \ldots)$$

provides intervals converging to $x_* \quad$ *with* $\; x_* \in I_n \;$ *and* $\; \operatorname{diam}(I_n) \le \operatorname{diam}(I_0)/2^n$.

Proof. Since $0 \notin F'(I)$, we may assume that f is increasing on I. Observation:
(1) $I_{n+1} \subset \big[$left half of $I_n\big]$ if $x_* \le \operatorname{mid}(I_n)$, (2) $I_{n+1} \subset \big[$right half of $I_n\big]$ if $x_* \ge \operatorname{mid}(I_n)$.

FIGURE Q.e.d.

3) Krawczyk's method

Setting. As in 2). From the relation $f(x) = f(x_*) + f'(\vartheta_x)(x - x_*) = f'(\vartheta_x)(x - x_*)$ with any constant $C \in \mathbb{IR}$ we obtain

$$Cf(x) = Cf'(\theta_x)(x - x_*), \qquad \big| \; + (x - x_*)$$
$$x - x_* + Cf(x) = x - x_* + Cf'(\theta_x)(x - x_*),$$
$$x_* = x - Cf(x) - \big[1 - Cf'(\theta_x)\big](x - x_*),$$
$$x_* \in x - Cf(x) - \big[1 - CF'(I)\big](x - I) =: K_0(I, x, C).$$

Definition. We define the *Krawczyk operator* as

$$K(J) := K_0\Big(J, x_J, 1/f'\big(x_J\big) =$$
$$= -\frac{f(x_J)}{f'(x_J)} - \left(1 - \frac{F'(J)}{f'(x_J)}\right)\big[-r, r\big] \quad \text{where} \quad x_J := \operatorname{mid}(I),\; r := \operatorname{diam}(I)/2.$$

Theorem. *If* $J \in \operatorname{Intv}(I)$ *then we have the alternatives*

(1) f admits a root in $K(J) \cap J$, (2) $K(J) \cap J = \emptyset$ and f has no root in J.

Remark. The (not presented) proof requires the accurate values $f(x_J), f'(x_J)$. This may be impossible in practice. Use the intervals $F\{x_J\}\big(=F([x_J,x_J])\big)$ resp. $F'\{x_J\}$ instead.

Optimization (minimization) with interval method

Task. Let $I=[a,b]\subset\mathbb{IR}$ be a compact interval and $f: I\to\mathbb{IR}$ a continuous function. Determine

$$y_*:=\min(f)\quad\text{and}\quad E_*:=\big\{x\in I: f(x_*)=y_*\big\}.$$

Algorithm. Choose an interval function $F\succ f$, $F:\mathrm{Intv}(I)\to\mathrm{Intv}(\mathbb{IR})$ covering f. With the aid of F, we are going to construct a pair of sequences

$\infty=y^{(0)}\geq y^{(1)}\geq y^{(2)}\geq\cdots$ resp. $I=E^{(0)}\supset E^{(1)}\supset E^{(2)}\supset\cdots$ such that

$$y^{(n)}\geq y_*\quad\text{and}\quad E_*\subset E^{(n)}=\bigcup_{j=1}^{N_n} I_j^{(n)}$$

where $N_0=1$, $I^{(0)}=I$ and, for $n=1,2,\ldots$,

$$I_1^{(n)}<I_2^{(n)}<\ldots<I_{N_n}^{(n)}\quad\text{are subintervals of}\quad I^{(n)}.$$

After having constructed $y^{(n)}, E^{(n)}$, we proceed to $y^{(n+1)}, E^{(n+1)}$ as follows.

Given any interval $I_j^{(n)}$, let

$$m_j^{(n)}:=\mathrm{mid}\big(I_j^{(n)}\big),\quad y_j^{(n)}:=\max F\{m_j^{(n)}\}\ \big(\geq f(m_j^{(n)})\big).$$

Observation: $\min(f)\leq\min\big(f|I_j^{(n)}\big)\leq y_j^{(n)}$ in any case. Hence, by writing $L(J), R(J)$ for the left resp right halves of an interval J, the choice

$$y^{(n+1)}:=\min_{j=1}^{N_n} y_j^{(n)},$$
$$\Big\{I_1^{(n+1)},\ldots,I_{N_{n+1}}^{(n+1)}\Big\}=\bigcup_{j=1}^{N_n}\Big[\Big\{J: J=L\big(I_j^{(n)}\big)\text{ with }\min F(J)\leq y^{(n+1)}\Big\}\cup$$
$$\cup\Big\{J: J=R\big(I_j^{(n)}\big)\text{ with }\min F(J)\leq y^{(n+1)}\Big\}\Big]$$

suits our requirements.

Observation: We also have $\min(f)\geq\min_{j=1}^{N_n}\min F(I_j^{(n)})$ for lower estimate.

Next we proceed to the theoretical aspects of programming interval analysis.

Formulas with interval arithmetics

Setting. We regard a family

$$\mathcal{S}:=\Big\{x,\sin,\cos,\log,\ldots\Big\}$$

of classical real functions. We shall call the members of $\mathcal{S}$ *special functions* (playing the role of built in machine functions in a computer). We also fix a set with isotonic interval

extension to each special function for the family

$$\mathfrak{S} := \Big\{X, \mathrm{Sin}, \mathrm{Cos}, \mathrm{Log}, \dots\Big\}, \quad s \prec S \quad (s \in \mathcal{S}).$$

Remark. The family $\mathfrak{S}$ will play the role of the *machine representation* of the interval functions $I \mapsto I$, $I \mapsto \sin(I) = \{\sin x : x \in I\}$ etc. Classical computer arithmetics work only with finitely many number symbols (e.g. with flooting point numbers of 48 digits in many practical cases). In such machines the intervals $X(I)$, $\mathrm{Sin}(I)$ etc. are represented only with endpoints from the given number symbols.
Therefore we do not assume automatically that $X(I) = I$, $\mathrm{Sin}(I) = \{\sin x : x \in I\}$ etc.. Of course $X(I) \subset I$, $\mathrm{Sin}(I) \subset \{\sin x : x \in I\}$ etc. in any case.

Definition. Let $\Phi = \Phi(x)$ denote an elementary function expression of functions from $\mathcal{S}$ with the variable symbol x. That is Φ consists of finitely many function symbols from $\mathcal{S}$ combined with the operations $\diamondsuit \in \{\pm, \cdot, /, \max, \dots\}$ in a syntactically correct manner. [Example: $\Phi = \Phi(x) = \log(\log x)/\sqrt{2 + \sin^3 x}$].

Evaluation. The usual mathematical form is a shorthand way of describing the procedure how Φ is evaluated. We identify Φ with a sequence

$$x = \Phi_0,\ \Phi_1, \Phi_2, \dots,\ \Phi_n = \Phi \quad \text{such that}$$
$$\Phi_i = \Big[x \text{ OR const. OR } s(\Phi_j) \text{ for some } j < i, s \in \mathfrak{S} \text{ OR } \Phi_j \diamondsuit \Phi_k \text{ for some } j, k < i\Big].$$

We introduce the natural *numeric-* resp. *interval realizations* ${}^f\Phi$ resp. ${}^I\Phi$ of Φ with respect to the representation $\mathfrak{S}$ of the special functions as the numeric- resp. interval functions

$${}^f\Phi(x) := \Big[\text{Formal step by step substitution of } x \text{ into } \Phi\Big],$$
$${}^{\mathfrak{S}}\Phi(J) := \Big[\text{Formal step by step substitution wrt. } \mathfrak{S} \text{ of } J \text{ into } \Phi\Big].$$

That is, in terms of the evaluation sequence,

$${}^f\Phi_0 = x, \quad {}^f\Phi_i = \Big[x \ \text{ OR const. OR } \ s\big({}^f\Phi_j(x)\big) \ \text{ OR } \ \big[{}^f\Phi_j(x)\big] \diamondsuit \big[{}^f\Phi_k(x)\big]\Big]$$
$${}^{\mathfrak{S}}\Phi_0 = [\xi \mapsto \xi], \quad {}^{\mathfrak{S}}\Phi_i = \Big[X \ \text{ OR Const. OR } \ S\big({}^{\mathfrak{S}}\Phi_j\big) \ \text{ OR } \ {}^{\mathfrak{S}}\Phi_j \diamondsuit {}^{\mathfrak{S}}\Phi_k\Big].$$

Example. 1) For $\Phi = \log(\log x)/\sqrt{2 + \sin^3 x}$,

$$\Phi_0 = x,\ \Phi_1 = \log(x),\ \Phi_2 = \log(\log(x)),\ \Phi_3 = 2,\ \Phi_4 = x,\ \Phi_5 = \sin(x),$$
$$\Phi_6 = (\sin(x))^3,\ \Phi_7 = \Phi_3 + \Phi_6,\ \Phi_8 = \sqrt{\Phi_7},\ \Phi_9 = \Phi_2/\Phi_8 = \Phi.$$

2) For $\Psi_1(x) := 1 - x \cdot x,\ \ \Psi_2(x) := (1-x)\cdot(1+x)$ we have ${}^f\Psi_1 = {}^f\Psi_2 : x \mapsto 1 - x^2$

$$\mathfrak{S}\Psi_1[-1,1] = 1 - [-1,1]\cdot[-1,1] = [1,1] - [-1,1] = [0,2],$$
$$\mathfrak{S}\Psi_2[-1,1] = \big(([1,1]-[-1,1]\big)\cdot\big(([1,1]+[-1,1]\big) = [0,2]\cdot[0,2] = [0,4].$$

The following important statements have straightforward proofs.

Lemma. *Let $I_1, I_2 \in \mathrm{Intv}(\mathbb{R})$ and, assume that $f_k \prec F_k$ where $f_k : I_k \to \mathbb{R}$ resp. $F_k : \mathrm{Intv}(I_k) \to \mathrm{Intv}(\mathbb{R})$ $(k = 1, 2)$. Then*
(1) *$f_1 \diamondsuit f_2 \prec F_1 \diamondsuit F_2$ whenever* (1a) *$I_1 = I_2$ and $\diamondsuit = +, -, \cdot$ or if* (1b) *$\diamondsuit = /$ and $f_2(I_2) > 0$;*
(2) *$f_1 \circ f_2 \prec F_1 \circ F_2$ whenever $f_2(I_2) \subset I_1$ and $f_1 \circ f_2 \prec F_1 \circ F_2$.*

Hence we can deduce (by induction with respect to formula length) the following statement:

Theorem. (Fundamental Theorem of Interval Analysis).
Assume I is a compact interval and Φ is an S-formula such that all the substitutions ${}^f\Phi_i(x)$ $(x \in I)$ resp. ${}^{\mathfrak{S}}\Phi_i(J)$ $(J \in \mathrm{Intv}(I))$ are well-defined during its evaluation. Then
$${}^f\Phi \prec {}^{\mathfrak{S}}\Phi \quad \textit{i.e. the interval function } {}^{\mathfrak{S}}\Phi \textit{ is an isotonic inclusion of } {}^f\Phi.$$

We continue with pure mathematical considerations before the investigation of an interval representation $\mathfrak{S}$ imitating "inaccurate calculation". We start with the study of the actual version of the Covering Lemma:

Lemma. *Suppose $I, I_1, \ldots, I_M \in \mathrm{Intv}(\mathbb{R})$ with $I \subset J := \bigcup_{k=1}^{M} I_k \in \mathrm{Intv}(\mathbb{R})$, $X_k \cap X \neq \emptyset$ and let Φ be a S-formula whose subformulas Φ_i admit well-defined and continuous numerical realizations $f_i := {}^f\Phi_i$ on J. Assume also that the interval realizations $F_i := {}^{\mathfrak{S}}\Phi_i$ are well-defined with $F_i(K) \in \mathrm{Intv}(J)$ for all subintervals of K of J. Then*

(1) $\mathrm{range}\big(f|I\big) = f(I)$ *and* $\bigcup_{k=1}^{M} F(I_k)$ *are intervals with* $f(I) \subset \bigcup_{k=1}^{M} F(I_k)$,

(2) *by writing* $I = \langle a, b\rangle$, $\bigcup_{k=1}^{M} F(I_k) = \langle A, B\rangle$ *in terms of the end points, we have*
$$A \le a \le b \le B \quad \textit{with} \quad a - A, B - b \le \max_{k=1}^{M} \mathrm{diam}\big(F(J_k)\big).$$

Definition. Given any functon $s : D \to \mathbb{R}\cup\{\pm\infty\}$ with $\emptyset \neq D \subset \mathbb{R}\cup\{\pm\infty\}$, its *idealistic* interval representation is obtained by means of theoretically accurately calculated ranges:
$${}^*s(I) := \big[\text{the minimal interval containing the range } s(I) \text{ of the restrictions} s|I\big]$$
defined for all intervals I contained in $D = \mathrm{dom}(s)$.

We consider the maximal reasonable setting with

$$\mathrm{S}_* := \Big\{ s : \mathrm{dom}(s) = \big[\text{finite union of intervals } I_1 < I_2, \cdots\big] \\ \text{with continuous } s|I_k \text{ and locally Lipschitzian on Interior}(I_k) \Big\},$$

$$\mathfrak{S}_* := \big\{ {}^*s : s \in \mathcal{S}_* \big\} \quad \text{with the short notation} \quad {}^*\Phi := {}^{\mathfrak{S}_*}\Phi.$$

Example. ${}^*\sin[0, 5\pi/4] = [-2^{-1/2}, 1]$ in terms of the (practically symbolic) irrational number $2^{-1/2} = \lim\limits_{n\to\infty} \sum_{k=1}^{n} \binom{-1/2}{k}$ without rounding (as if a machine with infinite arithmetic were used for infinitely long time when calculated this sum).

Remark. The customary special functions as $\sin(x), \log(x), x^p, \arccos(x), \arctan(x), \ldots$ are all continuous and defined on intervals. Moreover most of them are analytic and hence locally Lipschitzian in the inner of their domains. Courious exceptions are $\mathrm{sgn}(x)$ and $\sqrt[3]{x}$ defined on the whole $\mathbb{R}$. In order to avoid technical difficulties caused by oversized inclusions due to such exceptions, like

$$\Phi := \sqrt{\sqrt{x \cdot x} - \sqrt{\sqrt{x \cdot x}} \quad \text{with} \quad {}^f\Phi \equiv 0 \text{ and } {}^{\mathcal{S}_*}\Phi[-\varepsilon, \varepsilon] = \big[-\sqrt{\varepsilon}, \sqrt{\varepsilon}\big],$$

in this note we try to exclude such paradox cases as follows.

Recall. In general, a function $h : X \to Y$ between two metric spaces (X, d) resp. (Y, δ) is *Lipschitzian* if it has finite Lipschitz constant

$$\mathrm{Lip}(h) := \sup\big\{\delta\big(h(x), f(y)\big)/d(x, y) : x \neq y \in X\big\}.$$

The map h is *locally Lipschitzian* if for each point $x \in X$ there is an open subset of X such that $\mathrm{Lip}(h|U) < \infty$. Locally Lipschitzian functions on *compact* sets are Lipschitzian.
In particular, if $h : \mathbb{R} \supset X \to \mathbb{R}$ is a continuously differentiable function and K is a compact interval then

$$\mathrm{Lip}\big(h|K\big) = \max_{x\in K} |h'(x)| \quad \text{and} \quad \mathrm{diam}\, h(Z) \le \mathrm{Lip}\big(h|K\big)\mathrm{diam}\, Z \ (Z \subset K).$$

If K is a compact interval and $f : \mathbb{R} \to \mathbb{R}$ is continuous function then the range $f(K)$ is a compact interval as well. Hence the composite function $\mathrm{inv} \circ f = 1/f$ is well-defined for any point $x \in K$ if and only if $0 \notin f(K)$ that is if either $f(K) \subset (-\infty, 0)$ or $f(K) \subset (0, \infty)$. In the latter cases, $\mathrm{Lip}(1/f|K) \le \mathrm{Lip}(f|K)/\big[\min|f(K)|\big]^2$.

Definition. Henceforth we concentrate to *minimal* special functions forming the family

$$\mathcal{S}_0 := \bigcup\nolimits_{a,b\in[-\infty,\infty]} \big\{\text{locLip functions } (a, b) \to \mathbb{R}\big\}$$

of all locally Lipschitzian functions defined on *open* intervals with idealistic representation

$$\mathfrak{S}_0 = \big\{ {}^*s : s \in \mathcal{S}_0 \big\} \ \big(\subset \mathfrak{S}_*\big), \quad \text{with the abbreviation} \quad {}^0\Phi \ \big(= {}^*\Phi\big) \quad \text{for} \quad {}^{\mathfrak{S}_0}\Phi.$$

Since the functions in S_0 and the operations $\diamondsuit \neq /$ are locally Lipschitzian, also the interval realizations $s_* \in \mathcal{S}^0_*$ are locally Lipschitzian with respect to Hausdorff distance.

The family $\mathcal{S}_0$ includes the functions

$$\mathrm{inv}^{\oplus} : 0 < \xi \mapsto 1/\xi, \quad \mathrm{inv}^{\ominus} : 0 > \xi \mapsto 1/\xi.$$

We are going to consider the evaluation of an $\mathcal{S}_0$-formula Φ/Ψ with division only on intervals I where the denominator in the interval representation does not vanish that is if

$$0 \notin {}^0\Psi(I) \supset {}^f\Psi(I).$$

Namely, by setting $\sigma := \oplus$ if ${}^0\Psi(I) \subset (0,\infty)$ resp. $\sigma := \ominus$ if ${}^{\mathcal{S}^0_*}\Psi(I) \subset (-\infty, 0)$, we define

$$\begin{aligned} {}^f\Phi/\Psi(x) &:= {}^f\Phi(x) \cdot {}^f\mathrm{inv}^{\sigma}\Psi(x) \ \big(= {}^f\Phi(x)/{}^f\Psi(x)\big) && (x \in I), \\ {}^0\Phi/\Psi(J) &:= {}^0\Phi(J) \cdot \mathrm{inv}^{\sigma}_*\Big({}^0\Psi(J)\Big) && (J \in \mathrm{Intv}(I)). \end{aligned}$$

Straightforward step by step checking of the evaluation yields the following observation.

Lemma. *Let Φ be an $\mathcal{S}_0$-formula. Then the set $\mathrm{dom}\big({}^f\Phi\big)$, where the numerical representation is well-defined, is the union of open intervals. As for the interval representation: all the subintervals of an interval belonging to $\mathrm{dom}\big({}^0\Phi\big)$ belong to $\mathrm{dom}\big({}^0\Phi\big)$. For any point $\xi \in \mathrm{dom}\big({}^f\Phi\big)$, we have $\{\xi\} = [\xi,\xi] \in \mathrm{dom}\big({}^0\Phi\big)$, in particular ${}^f\Phi \prec {}^0\Phi$.*

Example. Consider $\Phi := 1/(x - x + x)$, $\big(\Phi \neq 1/x$ as symbolic formula!$\big)$.

Evaluation sequence of ${}^f\Phi(\xi)$ with $\xi \in \mathbb{R}$:

$$\xi, 0 = \xi - \xi, \xi = \xi - \xi + \xi, \big[\mathrm{Inv}^{\ominus}(\xi) = 1/\xi \text{ if } \xi < 0, \ \mathrm{Inv}^{\oplus}(\xi) = 1/\xi \text{ if } \xi > 0, \ 0 \notin \mathrm{dom}\big({}^f\Phi\big)\big].$$

Evaluation sequence of ${}^0\Phi[a,b]$ with $a \leq b \in \mathbb{R}$:

$$[a,b], [a,b] - [a,b] = [a-b, b-a], [a-b, b-a] + [a,b] = [2a-b, 2b-a], \big[\mathrm{Inv}^{\ominus} =$$
$$= [2a–b, 2b–a] \text{ if } [2a–b, 2b–a] \subset (-\infty, 0), \big[\mathrm{Inv}^{\ominus} = [2a–b, 2b–a] \text{ if } [2a–b, 2b–a] \subset (-\infty, 0)\big].$$

Thus $[a,b] \in \mathrm{dom}\big({}^0\Phi\big)$ if and only if $\ 2a - b < 0$ OR $2b - a > 0$. Since $a \leq b$ it follows

$$\mathrm{dom}\big({}^0\Phi[a,b]\big) = \big\{\langle a,b\rangle : 2b < a \leq b < 0\big\} \cup \big\{\langle a,b\rangle : 0 < a \leq b < 2a\big\}.$$

Excercise. $\mathrm{dom}\Big(1/\big(\sin(1/\sin x)\big)\Big) = ?$

Remark. By adding the trivial binary operation $(\xi,\eta) \mapsto \xi$ to the collection of built in operations (like $+, -, \cdot, /$), we can regard the evoluation sequence of Φ in the form

$$x = \Phi_0, \ldots, \Phi_n = \Phi, \qquad \Phi_m = \big[s_{i(m,1)}(\Phi_{j(m,1)})\big] \diamondsuit_m \big[s_{i(m,2)}(\Phi_{j(m,2)})\big]$$

with $\ \diamondsuit_m \in \{+,-,\cdot\}, \quad s_{i(m,1)}, s_{i(m,2)} \in \mathcal{S}, \quad j(m,1), j(m,2) < m.$

This helps to reduce discussions concerning the properties of composite formulas.

Definition. With recursion, we introduce the *distributed form* ${}^{ff}\Phi_0, {}^{ff}\Phi_1, \ldots, {}^{ff}\Phi_n = {}^{ff}\Phi$ of the evaluation sequence $\Phi_0, \Phi_1, \ldots, \Phi_n = \Phi$ above as follows:

$$ {}^{ff}\Phi_m : \big(x_1, x_2, \ldots, x_{N(m)}\big) \mapsto g_m\big(x_1, x_2, \ldots, x_{N(m)}\big) $$

where $\quad N(0) = 1, \;\; g_0(x_1) := x_1,$

$$ \Phi_m = \big[s(\Phi_k)\big] \diamondsuit_m \big[t(\Phi_\ell)\big] \quad \big(s, t \in \mathcal{S},\, k, \ell < m\big) \Longrightarrow \quad N(m) := N(k) + N(\ell), $$

$$ g_m\big(x_1, \ldots, x_{N(m)} := \big[s\big(g_k(x_1, \ldots, x_{N(k)})\big)\big] \diamondsuit_m \big[t\big(g_\ell(x_{N(k)+1}, \ldots, x_{N(m)})\big)\big]. $$

Example. For $\Phi(x) = e^x e^{-x-x^2} e^{x^2}$ we have ${}^{f}\Phi(x) = 1$. Evaluation for ${}^{*}\Phi(I)$:

$I, \exp(I), -I-I\cdot I, -I-I\cdot I, [\exp(I)]\cdot[\exp(-I-I\cdot I)], \big[[\exp(I)]\cdot[\exp(-I-I\cdot I)]\big]\cdot\big[\exp(I\cdot I)\big]$.

Distributed evaluation for ${}^{ff}\Phi$:

$x_1, \exp(-x_1), \exp(-x_1 - x_2^2), \exp(-x_1)\cdot\exp(-x_2 - x_3^2), \exp(-x_1)\cdot\exp(-x_2 - x_3^2)\cdot\exp(x_4^2)$.

Remark. Roughly speaking, we obtain ${}^{ff}\Phi$ from the expression of $\Phi(x)$ by replacing the variable term x with a new symbol x_i at each appearence.

Induction argument with respect to the length of evolution sequences yields the following.

Lemma. *Given an interval I, we have $I \in \mathrm{dom}({}^{*}\Phi)$ if and only if all the expressions ${}^{ff}\Phi\big(x_1, \ldots, x_N\big)$ with $\big(x_1, \ldots, x_N \in I\big)$ are well-defined. In this case*

$$ {}^{*}\Phi(I) = \Big\{ {}^{ff}\Phi\big(x_1, \ldots, x_N\big) : \; \big(x_1, \ldots, x_N\big) \in I \Big\} = {}^{ff}\Phi\big(I, \ldots, I\big) = {}^{ff}\Phi\big(I^N\big). $$

Proposition. *Let Φ be a $\mathcal{S}_0$-formula with the realizations $f := {}^{f}\Phi$ resp. $F := {}^{0}\Phi$. Then, given any inner point x_0 of $\mathrm{dom}(f)$, we can find $\varepsilon, M > 0$ such that*

$$ [x_0 - \varepsilon, x_0 + \varepsilon] \in \mathrm{dom}\big(F\big) \quad \textit{and} \quad \mathrm{diam}\big(F(J)\big) \le M\cdot\mathrm{diam}(J) \;\; \big(J \in \mathrm{Intv}\big([x_0 - \varepsilon, x_0 + \varepsilon]\big)\big). $$

Proof. Notice that all the functions $s \in \mathcal{S}_0$ are locally Lipschitzian and defined on some open interval. Also the operations $\diamondsuit \in \{\pm, \cdot\}$ appearing in Φ are continuous maps $\mathbb{R}^2 \to \mathbb{R}$. Therefore the composite map ${}^{ff}\Phi$ is defined and locally Lipschitzian (as a map of N variables) on some open subset D of $\mathbb{R}^N$. On the other hand we have

$$ f(x_0) = {}^{f}\Phi(x_0) = {}^{ff}\Phi(x_0, \ldots, x_0) \in D. $$

Since $x_0 \in D \,\mathrm{open}\subset \mathbb{R}^N$ we can find $\varepsilon > 0$ such that $[x_0 - \varepsilon, x_0 + \varepsilon]^N \,\mathrm{compact}\subset D$ with $L := \mathrm{Lip}\big({}^{ff}\Phi\big|[x_0 - \varepsilon, x_0 + \varepsilon]^N\big) < \infty$ (with respect to N-dimensional Euclidean distance). Consider any interval $J \subset [x_0 - \varepsilon, x_0 + \varepsilon]$. For any pair of points $a, b \in J$ we have

$$ \big|f(b) - f(a)\big| = \big|{}^{f}\Phi(b) - {}^{f}\Phi(a)\big| = \big|{}^{ff}\Phi(b, \ldots, b) - {}^{ff}\Phi(a, \ldots, a)\big| \le $$
$$ \le L\big\|(b, \ldots, b) - (a, \ldots, a)\big\| = L\sqrt{N}\big|b - a\big|. $$

Hence the choice $M := L\sqrt{N}$ suits our requirements. Q.e.d.

Range Enclosure Theorem. *Let $I=[a,b],\Phi$ be a compact interval with an $\mathcal{S}_0$-formula such that $I\subset\mathrm{dom}(f)$ for the numerical realization $f:={}^{f}\Phi$. Then*

(1) $\mathrm{range}\big(f|I\big)=f(I)=[c,d]$ *is a compact interval,*

(2) *there exists $\delta=\delta_{I,\Phi}>0$ along with a constant $K=K_{I,\Phi}\in(0,\infty)$ such that, for the interval realization $F:={}^{\mathcal{S}^0_*}\Phi$ we have*
$J\in\mathrm{dom}(F)$ *whenever* $J\in\mathrm{Intv}[a-\delta,b+\delta]$ *with length* $<\delta$,

(3) *given any covering* $I\subset\bigcup_{k=1}^{n}I_k$, $I_k\cap I\neq\emptyset$ *of* I *with intervals of length* $<\delta$,
$[c,d]\subset\bigcup_{k=1}^{n}F(I_k)\subset\Big[c-K\max_k\mathrm{length}(I_k),d+K\max_k\mathrm{length}(I_k)\Big]$.

Proof. (1) We know already that $\mathrm{dom}(f)$ is an open subset of $\mathbb{R}$ being the disjoint union of a family of open intervals $D_1,D_2,\ldots$ such that the subfunctions $f|D_i$ are continuous. By assumption, $[a,b]\subset\mathrm{dom}(f)$, therefore $[a,b]\subset D_i$ for some (unique) index i and $f(I)=f\big([a,b]\big)$ is a compact interval as being the continuos image of a compact interval.

(2) We also know that for any point $x\in[a,b]$, there are $\varepsilon_x,M_x\in(0,\infty)$ such that

$$[x-\varepsilon_x,x+\varepsilon_x]\in\mathrm{dom}(F),\quad \mathrm{length}F(J)\leq M_x\mathrm{length}(J)\text{ for all }J\in\mathrm{Intv}\big([x-\varepsilon_x,x+\varepsilon_x]\big).$$

Notice that $[x-!\varepsilon_x,x+\varepsilon_x]\subset D_i$ $(x\in I)$. Since trivially $I=[a,b]\subset\bigcup_{x\in I}(x-\varepsilon_x/2,x+\varepsilon_x/2])$, by Borel's Covering Theorem there is a finite sequence

$$a=x_0<x_1<\cdots<x_n=b\quad\text{with}\quad I\subset\textstyle\bigcup_{\ell=1}^{n}\big(x-\varepsilon_{x_\ell}/2,x+\varepsilon_{x_\ell}/2\big).$$

Define $\quad\delta:=\min_{\ell=0}^{n}\varepsilon_{x_\ell}/2,\quad K:=\max_{\ell=0}^{n}M_{x_\ell}$.

Consider an interval $J\subset[a-\delta,b+\delta]$ of length $<\delta$. Observation: we can find points $x\in[a,b]$ and $y\in J$ with $|x-y|<\delta$ and, by construction, $|x-x_\ell|<\varepsilon_{x_\ell}/2$. That is, for some index ℓ we have $J\subset(x_\ell-\varepsilon_{x_\ell}/2-\delta,(x_ell+\varepsilon_{x_\ell}/2+\delta)\subset(x_ell-\varepsilon_{x_\ell},(x_ell+\varepsilon_{x_\ell})$ with $\mathrm{diam}\big(F(J)\big)\leq M_{x_\ell}\mathrm{diam}(J)\leq M_{x_\ell}\mathrm{diam}(J)$.

(3) Immediate consequence of (2) and the covering lemmas.

Computer realization

Arithmetics: $u:=\big[\text{machine }0\big]$, $\quad u^{-1}\geq 1$ being a finite number,

$$\mathfrak{A}_u:=\big\{nu:n\in\mathbb{Z},|n|\leq u^{-1}\big\}\cup\{\pm\infty\}\quad\text{machine numbers.}$$

Intervals are represented with endpoints from $\mathfrak{A}_u$:

$$R_u(I):=\big[\max\{r\in\mathfrak{A}_u:r\leq I\},\min\{r\in\mathfrak{A}_u:r\geq I\}\big]\qquad\big(I\in\mathrm{Intv}(\mathbb{R}).$$

$\mathcal{S}_u=\big\{\text{Special functions available from memory}\big\}$ is a finite subset of $\mathcal{S}_0$.

Given $s \in \mathcal{S}_u$, we suppose that the machine realization of the idealisic function ${}^f s$ is built in with accuracy within u. Hence the machine interval realization ${}^u s$ is constructed by means of including its range in minimal closed intervals with endpoints in $\mathfrak{A}_u$:

$$\mathfrak{S}_u = \big\{{}^u s : s \in \mathcal{S}_u\big\}, \quad {}^{\mathfrak{S}_u} s(I) = {}^u s(I) = R_u\big({}^u s\big(R_u(I)\big)\big) = R_u\big(\mathrm{range}\big(s\big|R_u(I)\big)\big).$$

The binary operations $\diamondsuit \in \{\pm, \cdot\}$ with intervals are represened similarly

$$I\,{}^u\!\diamondsuit J = R_u\big([R_u(I)]\diamondsuit[R_u(I)]\big) \quad (I, J \in \mathrm{Intv}(\mathbb{IR})).$$

Evaluation with rounding. Even with formulas, we also use the shorthand ${}^u\Phi = {}^{\mathfrak{S}_u}\Phi$. Let Φ be a $\mathcal{S}_0$-formula with the evaluation sequence $x = \Phi_0, \Phi_1, \ldots, \Phi_{n-1}, \Phi_n = \Phi$. Given an arithmetics $\mathfrak{A}$ with machine zero u, we define the *u-rounded evaluation sequence* $\big[I \mapsto {}^u I\big] = \big[I \mapsto R_u(I)\big] = {}^u\Phi_0, \ldots, {}^u\Phi_n = {}^u\Phi$ of Φ as

$${}^u\Phi_k : I \mapsto {}^u\big\{s\big({}^u\Phi_i(I)\big)\big\} \qquad \text{if } \Phi_k = s(\Phi_i) \text{ with } s \in \mathcal{S}^0 \text{ and } i < k,$$

$${}^u\Phi_k : I \mapsto {}^u\big\{\big[{}^u\Phi_i(I)\big]\diamondsuit\big[{}^u\Phi_i(I)\big]\big\} \text{ if } \Phi_k = \Phi_i\diamondsuit\Phi_j \text{ with } i, j < k \text{ and } \diamondsuit \in \{+, \cdot\}.$$

Observation: The notation ${}^u\Phi$ with $u > 0$ is not in conflict with the earlier terms ${}^0\Phi$: in any case, ${}^u\Phi$ means that we calculate with accuracy u.

Example. Let only $u := 0.001$. Then ${}^u\sqrt{2} = R_u\big([\sqrt{2}, \sqrt{2}]\big) = [1.414, 1.415]$; also

$${}^u\mathrm{Inv}^{\oplus}\big([1, \sqrt{2}]\big) = {}^u\mathrm{Inv}^{\oplus}\big([1, 1.415]\big) = {}^u[1/1.415, 1] = [0.706, 1] \ ;$$

$${}^u\big([1/\sqrt{2}, 1]/[1, \sqrt{2}]\big) = {}^u\Big({}^u[1/\sqrt{2}, 1]\ {}^u\cdot\ {}^u\mathrm{Inv}^{\oplus}\big([1, \sqrt{2}]\big) =$$

$$= {}^u\big([0.706, 1] \cdot [0.706, 1]\big) = {}^u\big([0.498436, 1] = [0498, 1].$$

Remark. Since the operation R_u increases intervals, morover rounding with smaller machine zero v yields bigger interval, in general we have

$${}^u\Phi \succ {}^v\Phi \succ {}^0\Phi \succ {}^f\Phi \quad (u \geq v > 0).$$

For the same reason, if we calculate the value of a formula precisely (this is mosly possible only in theory) and then we round the result, or even if we calculate precisely with rounded starting data and then round the result, we obtain a not less accurate result than with calculating with rounded data and operations during every evolution step. That is, if $F : I \mapsto R_u\big({}^0\Phi\big(R_u(I)\big)\big)$ and $G := {}^u\Phi$ then we have $F(I) \subset G(I)$ if $I \in \mathrm{dom}(G) \subset \mathrm{dom}(F)$ implying $F \prec G$. Unfortunately, we have no equality in general.

Example. Let $\Phi(x) := x \cdot (x \cdot (x \cdot (x \cdot (x \cdot x))))$, $u := 0.1$ and $I := \big[0.9, 1\big]$. Then, with the usual abbreviation ${}^u J := R_u(J)$ for intervals J, we have ${}^{0.1} I = I = \big[0.9, 1\big]$,

$$F(I) = {}^{0.1}\big\{{}^0\Phi({}^1 I)\big\} = {}^{0.1}\big\{{}^0\Phi[0.9, 1]\big\} = {}^{0.1}\big[0.9^6, 1^6\big] = {}^{0.1}\big[0.531441, 1\big] = \big[0.5, 1\big],$$

$$G(I) = {}^{0.1}I_6 \quad \text{where} \quad I_1 = {}^{0.1}I = I = \big[a_1, 1\big] \text{ with } a_1 = 0.9,$$
$$I_{k+1} = {}^{0.1}\big\{I \cdot I_k\big\} = \big[a_{k+1}, 1\big] \text{ with } a_k = {}^{0.1}[0.9 \cdot a_k],$$
$$a_2 = {}^{0.1}[0.9 \cdot 0.9] = {}^{0.1}0.81 = 0.8, \quad a_3 = {}^{0.1}[0.9 \cdot 0.8] = {}^{0.1}0.72 = 0.7,$$
$$a_4 = {}^{0.1}[0.9 \cdot 0.7] = {}^{0.1}0.63 = 0.6, \quad a_5 = {}^{0.1}[0.9 \cdot 0.6] = {}^{0.1}0.54 = 0.5,$$
$$a_6 = {}^{0.1}[0.9 \cdot 0.5] = {}^{0.1}0.45 = 0.4, \qquad \text{Thus } \; G(I) = \big[0.4, 1\big] \neq F(I).$$

Lemma. *Let $\Psi(x) = [s_1(x)]\diamondsuit[s_2(x)]$ be a $\mathcal{S}_0$-formula, and suppose $I \in \mathrm{dom}\big({}^{u}\Psi\big)$ is an interval such that all the intervals appearing in the evaluation sequence of ${}^{u}\Psi(I)$ are included in $[-u^{-1}, u^{-1}]$.*

Then for any $y \in {}^{u}\Psi(I)$ there exist $x_1, x_2 \in I$, $\theta_1, \theta_2, \zeta_1, \zeta_2, \eta \in [-u, u]$ such that

$$y = \eta + \big\{\big[\zeta_1 + s_1(x_1 + \theta_1)\big]\diamondsuit\big[\zeta_2 + s_2(x_2 + \theta_2)\big].$$

Proof. Recall that, in terms of the rounding operation R_u,

$${}^{u}\Psi\big(I\big) = R_u\big\{\big[R_u\big(s_1(R_u(I))\big]\diamondsuit\big[R_u\big(s_1(R_u(I))\big]\big\}.$$

On the other hand, given any interval $J = [a, b] \in \mathrm{Intv}[-u^{-1}, u^{-1}]$, we have $R_u(J) = [a - \varepsilon_1, b + \varepsilon_2]$ with $\varepsilon_1, \varepsilon_2 \in [0, u]$. That is we can write any point $z \in R_u(J)$ in the form $y = x + \theta$ with some $x \in J$ and $\theta \in [-u, u]$.

Hence the statement is immediate.

Definition. We introduce the *rounded distribution form* ${}^{Rf}\Phi$ for a formula Φ by modifying the construction of ${}^{0}\Phi$ with variables for the errors during each elementary evaluation steps on the basis of the above Lemma.

If $\Phi(x)$ contains N copies of the variable symbol x along with K binary operations $\diamondsuit$ and M appearences of function symbols $s \in \mathcal{S}_0$ then ${}^{Rf}\Psi$ is a $\mathcal{S}_0$-formula of the variables

$$x_1, \ldots, x_N, \quad \theta_1, \ldots, \theta_N, \quad \eta_1, \ldots, \eta_K, \quad \zeta_1, \ldots, \zeta_M.$$

That is, if $x = \Phi_0, \ldots, \Phi_n = \Phi$ is the evaluation sequence of Φ where a generic term ${}^{Rf}\Phi_i$ operates on the variables

$$x_1, \ldots, x_{N(i)},\ \theta_1, \ldots, \theta_{N(i)},\ \eta_1, \ldots \eta_{K(i)},\ \zeta_1, \ldots, \zeta_{M(i)}$$

with the recursion pattern $\Phi_m = [s(\Phi_k)]\diamondsuit t(\Phi_\ell)]$, $k, \ell < m$ then

${}^{Rf}\Phi_0(x_1, \theta_1) = x_1 + \theta_1$, $\quad N(m) = N(k) + N(\ell)$, $\; K(m) = 1 + K(k) + K(\ell)$, $\; M(m) = M(k) + M(\ell)$,

$${}^{Rf}\Phi_m = \eta_{K(m)} + \Big[\big[{}^{Rf}\Phi_k\big]\diamondsuit\big[{}^{Rf}\Phi_\ell\big(x_{1+N(k)}, \ldots, x_{N(m)}, \theta_{1+N(k)}, \ldots, \theta_{N(m)}, \eta_{1+K(k)}, \ldots, \eta_{K(m)}, \zeta_{1+}$$

We denote the rounded function realization of the symbolic formula ${}^{Rf}\Phi$ also with ${}^{Rf}\Phi$ without danger of confusion.

On the basis of the previous Lemma and since compositions of continuous maps defined on open sets are continuous maps defined on open sets as well, we conclude the following.

Proposition. *The distributed function realization* ${}^{Rf}\Phi$ *of a* $\mathcal{S}_0$*-formula* Φ *is a continuous function* $D \to \mathbb{R}$ *with* $D = \mathrm{dom}\big({}^{Rf}\Phi\big)$ open $\subset \mathbb{R}^{2N+K+M}$.
Let $I \subset \mathrm{dom}\big({}^{ff}\Phi\big)$ *be an interval for which the interval realization of* Φ *is well-defined*
As for the function realization of Φ*, we have*

$$ {}^{f}\Phi(x) = {}^{Rf}\Phi\big(\underbrace{x,\ldots,x}_{N},\underbrace{0,\ldots,0}_{N+K+M}\big) \quad \text{whenever} \quad x \in I.$$

Concerning the rounded interval realization, for any point $y \in {}^{u}\Phi(I)$,

$\exists\, x_1,\ldots,x_N \in I \quad \exists\, \theta_1,\ldots,\theta_N \in [-u,u] \quad \exists\, \eta_1,\ldots,\eta_K \in [-u,u] \quad \exists\, \zeta_1,\ldots,\zeta_M \in [-u,u]$
with $\quad y = {}^{Rf}\Phi(x_1,\ldots,x_N,\theta_1,\ldots,\theta_N,\eta_1,\ldots,\eta_K,\zeta_1,\ldots,\ \zeta_M\big)$.
Conversely, if $\quad I^N \times [-u,u]^{N+K+M} \subset \mathrm{dom}\big({}^{Rf}\Phi\big) \quad$ then $\quad I \in \mathrm{dom}\big({}^{u}\Phi\big)$.

Remark. This theoretical model can be regarded as the use of a huge *fixed point arithmetics*. However, most mathematical coprocessors work with numbers in binary floating point form represented as sequeces from $0, 1$ of a given length 2^N.

Definition. We shall write

$$\mathfrak{B}_N := \mathfrak{A}_{u_N} = \Big\{k \cdot u_N : k \in \mathbb{Z} \cap \big[-u_N^{-1}, u_N^{-1}\big]\Big\}, \qquad u_N := 2^{-2^N}$$

Clearly, the family of numbers represented by a binary floating point arithmetics can be included in a some of our arithmetics $\mathfrak{B}_N$.
Given a computer equipped with a coprocessor with machine-zero $u = 2^{-M}$, we can emulate the numbers and operations of $\mathfrak{B}_L$ with $u < u_L$ i.e. $M > 2^L$ in a straightforward manner. (By no means an easy programming task when short run time and effective use of memory is required). Once $\mathfrak{B}_{N-1}$ is available, we can emulate the use of $\mathfrak{B}_N$ as follows: calculate with the available double precision (with sequences of length 2^N), and round back the result in out fixed point form when determining the endpoints of the intervals in calculations.

Corollary. *Let* Φ *be a* $\mathcal{S}_0$*-formula and let* I *be a compact interval for which the precise evaluation* ${}^{0}\Phi(I)$ *is well-defined. Then, for sufficiently large* L*, the rounded evaluation* ${}^{u_L}\Phi(I)$ *is also a well-defined compact interval.*

Proof. According to the Proposition above,

$$I = {}^{Rf}\Phi\big\{c(x) : x \in I\big\} \quad \text{where} \quad c(x) := \big(\underbrace{x,\dots,x}_{N}, \underbrace{0,\dots,0}_{N+K+M}\big),$$

$$c(I) = \big\{c(x) : x \in I\big\} \text{ compact} \subset D := \mathrm{dom}\big({}^{Rf}\Phi\big) \text{ open} \subset \mathbb{R}^{2N+K+M}.$$

Therefore the (Hausdorff) distance $\delta := \mathrm{dist}\big(c(I), D\big)$ between D and the compact segment $c(I)$ is positive, and for the δ-tube set of $c(I)$ we have

$$c(I)_\delta := \big\{z \in \mathbb{R}^{2N+K+M} : d\big(z, c(I)\big) < \delta\big\} = \bigcup_{x\in I} \mathrm{Ball}\big(\underbrace{c(x)}_{\text{center}}, \underbrace{\delta}_{\text{radius}}\big) \subset D.$$

Observation: any cube $c(x) + [-\varepsilon,\varepsilon]^{2N+K+M}$ with center point $c(x)$ and edge length 2ε has diameter $2\varepsilon\sqrt{2N+K+M}$ and hence

$$c(x) + [-\varepsilon,\varepsilon]^{2N+K+M} \subset \mathrm{Ball}\big(c(x),\delta\big) \quad \text{whenever} \quad \varepsilon < \delta/\sqrt{2N+K+M}.$$

From the Proposition we also know that

$${}^{u}\Phi(I) \subset {}^{Rf}\Phi\big(c(I) + [-u,u]^{2N+K+M}\big) \text{ whenever } c(I) + [-u,u]^{2N+K+M} \subset D.$$

Thus any choice for L with $u_L = 2^{2^L} < \delta/\sqrt{2N+K+M}$ suits our requirements.

Lemma. *Let $I_0 = [a,b] \in \mathrm{Intv}(\mathbb{R})$ and assume F is an interval function such that*

$$\mathrm{Intv}(I_0) \subset \mathrm{dom}(F),\ \ F(I_0) \subset [-u^{-1}, u^{-1}],\ \ \mathrm{length}\big(F(I)\big) \le M\mathrm{length}(I)\ \ (I \in \mathrm{Intv}(I_0)).$$

Then, for every subinterval I of $[a+u, b-u]$ we have

$$\mathrm{diam}\big({}^{\mathfrak{A}}F({}^{\mathfrak{A}}I)\big) \le (2M+1)\mathrm{length}(I) \quad \textit{whenever} \quad \mathrm{length}(I) \ge 2u.$$

Proof. Let $I \in \mathrm{Intv}[a+u, b-u]$. Then ${}^{\mathfrak{A}}I \subset [a,b] = I_0$ and

$$\mathrm{length}\big({}^{\mathfrak{A}}I\big) \le \mathrm{length}(I) + 2u, \quad \mathrm{length}\big(F({}^{\mathfrak{A}}I)\big) \le M\big(\mathrm{length}(I) + 2u\big),$$

$$\mathrm{length}\Big({}^{\mathfrak{A}}F\big({}^{\mathfrak{A}}I\big)\Big) \le M\big(\mathrm{length}(I) + 2u\big) + 2u.$$

In particular, if $\mathrm{length}(I) \ge 2u$ then

$$\frac{\mathrm{length}\Big({}^{\mathfrak{A}}F\big({}^{\mathfrak{A}}I\big)\Big)}{\mathrm{length}(I)} \le M\Big(1 + \frac{2u}{\mathrm{length}(I)}\Big) + \frac{2u}{\mathrm{length}(I)} \le M(1+1) + 1. \qquad \text{Q.e.d.}$$

Corollary. (Computer version for the Range Enclosure Theorem.).

Assume that the Arithmetics is accurate in the sense that

$$\mathrm{diam}\Big({}^{\mathfrak{S}}s(J)\Big) \le \ell \cdot u + L \cdot \mathrm{diam}\big(s(J)\big) \textit{ for every } J \in \mathrm{Intv}(I_0) \textit{ and } s \in \mathcal{S}.$$

Then there are finite constants $\widetilde{k}, \widetilde{K}$ depending on n and I_0 such that, for any compact subinterval I of I_0 with a finite open covering $I \subset \bigcup_{k=1}^{M} I_k$ and for any formula $\Phi \in \mathcal{F}_{n,I_0}$,

$${}^{f}\Phi(I) \subset \bigcup_{k=1}^{M} {}^{\mathfrak{S}}\Phi(I_k), \quad \mathrm{diam}\Big(\bigcup_{k=1}^{M} {}^{\mathfrak{S}}\Phi(I_k)\Big) \le \mathrm{diam}\big({}^{\mathfrak{S}}\Phi(I)\big) + \widetilde{k}u + \widetilde{K}\max_{k=1}^{M}\mathrm{diam}I_k\ .$$

Proof. Straightforward imitation of steps (1)-(5) in the idealistic version.

Generalized interval arithmetics

Basic concepts. Intervals in N-dimensisions

$\mathrm{Intv}(\mathbb{R}^N) := \big\{[a_1,b_1]\times\cdots\times[a_N,b_N] :\ -\infty < a_k \le b_k < \infty\ (k=1,\dots,N)\big\}$,

$\mathrm{Intv}(G) := \big\{I \in \mathrm{Intv}(\mathbb{R}^N) :\ I \subset G\big\}$.

Admissible (isotonic) interval functions of (the type $\mathbb{R}^N \to \mathbb{R}^M$)

$F : \mathrm{Intv}(\mathbb{R}^N) \supset \mathcal{G} \to \mathrm{Intv}(\mathbb{R}^M)$, ha mindig $I \subset J \in \mathcal{G} \Rightarrow I \in \mathcal{G}$ és $F(I) \subset F(J)$.

Enclosure of a function with admissible interval function

$f \prec F$ $(F \succ f)$ if $f : \mathbb{R}^N \supset G \to \mathbb{R}^M$, $F : \mathrm{Intv}(\mathbb{R}^N) \supset \mathcal{G} \to \mathbb{R}^M$, and

$\mathrm{Intv}(G) \subset \mathcal{G}\big(=\mathrm{dom}(F)\big)$, $f(I)\big(=\{f(x) : x \in I\}\big) \subset F(I)$ $(I \in \mathrm{Intv}(G))$.

Lipschitz constants (wrt. max-norm): $\mathrm{Lip}(f) := \sup_{x\neq y\in\mathrm{dom}(f)} \|f(x)-f(y)\|/\|x-y\|$,

$\mathrm{Lip}(F) := \sup_{I\in\mathrm{dom}(F)} \mathrm{diam}\big(F(I)\big)/\mathrm{diam}(I)$.

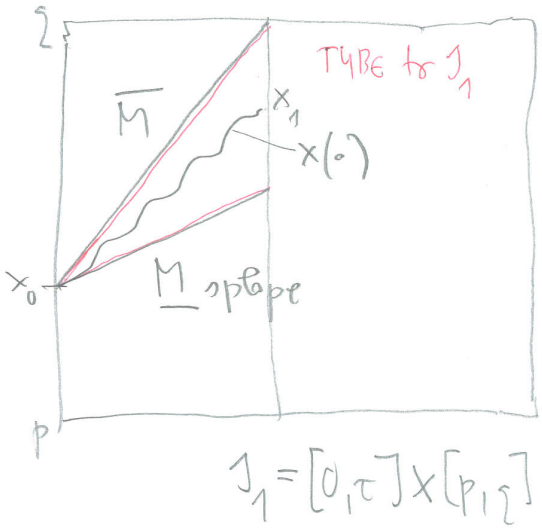
Remark. $f \prec F \ \Rightarrow\ \mathrm{diam}\big(F(I)\big) \le \mathrm{Lip}(f)\cdot\mathrm{diam}(I)$ $\big(I \in \mathrm{Intv}\big(\mathrm{dom}(f)\big)\big)$.

Recall: Piccard–Lindelöf theorem

$\dot{x}(t) = f\big(t,x(t)\big)$, $x(0) = x_0$

$f : [0,T]\times[p,q] \to \mathbb{R}$ continuous, $|f| \le M$,

$\big|f(\tau,\xi_1) - f(\tau,\xi_2)\big| \le L|\xi_1-\xi_2|$ $\big(0 \le \tau \le T,\ \xi_1,\xi_2 \in [p,q]\big)$



CTUBE2

f as a vector field.

The graph of a function passing smoothly along this vector field starting from $(0, x_0)$ remains in the triangle $(0, x_0), (\tau_*, x_0 \pm M\tau_*)$ where

$\tau_* := \min\{(q - x_0)/M, (x_0 - p)/M\}$

$x(t) = x_0 + \int_{s=0}^{t} f\big(t, x(t)\big)\ ds$ equivalent integral equation

$$x_0(t) \equiv x_0, \quad x_{n+1}(t) := x_0 + \int_{s=0}^{t} f\big(t, x_n(t)\big)\ ds \quad (0 \le t \le \tau_*)$$
$$x_n : [0, \tau_*] \to [p, q] \quad \text{well-defined } \forall\ n,$$

If the sequenece $x_0, x_1, \ldots$ is uniformly convergent then its limit x_* is the unique solution

$$x_1(t) - x_0(t) = \int_{s=0}^{t} \big[f(t, x_0)\big]\ ds,$$
$$x_{n+1}(t) - x_n(t) = \int_{s=0}^{t} \big[f\big(t, x_n(s)\big) - f\big(t, x_{n-1}(s)\big)\big]\ ds \quad (n = 1, 2, \ldots)$$

Due to the bounds M, L, we have

$$|x_1(t) - x_0| \le \int_{s=0}^{t} \big|f(t, x_0)\big|\ ds \le \int_{s=0}^{t} M\ ds = Mt,$$
$$\big|x_{n+1}(t) - x_n(t)\big| \le \int_{s=0}^{t} \big|f\big(t, x_n(s)\big) - f\big(t, x_{n-1}(s)\big)\big|\ ds \le$$
$$\le \int_{s=0}^{t} L\big|x_n(s) - x_{n-1}(s)\big|\ ds$$

By induction on n we conclude that

$$\big|x_{n+1}(t) - x_n(t)\big| \le ML^n t^{n+1}/(n+1)! \quad \Big(= \int_{s=0}^{t} L \cdot \big[ML^n s^n/n!\big]\ ds\Big).$$

With respect to $\|\cdot\|_\infty$ norm (max norm), $x_0, x_1, x_2, \ldots$ is a *finite path* in $\mathcal{C}[0, \tau^*]$:

$$\sum_{n=0}^{\infty} \big|x_{n+1}(t) - x_n(t)\big| \le \sum_{n=0}^{\infty} ML^n \tau_*^{n+1}/(n+1)! = M\big[e^{L\tau_*} - 1\big] < \infty$$

Hence the function of solution $x_* := \lim_{n\to\infty} x_n \in \mathcal{C}[0, \tau_*]$ satisfies

$$\big|x_*(t) - x_r(t)\big| \le \sum_{n=r}^{\infty} ML^n t^{n+1}/(n+1)! \qquad (r = 0, 1, \ldots;\ 0 \le t \le \tau_*),$$

and also $x_* : [0, \tau_*] \to [p, q]$ in any case!

Piccard-Lindelöf construction with interval arithmetics

Interval version of $dx/dt = f(x, t)$:

$$F : \mathrm{Intv}\big([0, T] \times [p, q] \to \mathrm{Intv}(\mathbb{IR}) \ \text{ such that } \quad f \prec F$$

with $f(t, \xi) \in F(I \times J)$ whenever $t \in I$, $\xi \in J$.

Initial setting: $x_0 \in X_0 \in \mathrm{Intv}([p, q]); \qquad X_0(t) \equiv X_0.$

Take a partition $0 = \tau_0 < \tau_1 < \cdots < \tau_N = \tau_*$ of $[0, \tau_*]$ and let

$$I_k := [\tau_{k-1}, \tau_k], \quad \ell(t) := [k : t \in I_k, t < \tau_k]$$

$$x_{n+1}(t) = x_0 + \int_{s=0}^{t} f\big(s, x(s)\big)\ ds =$$

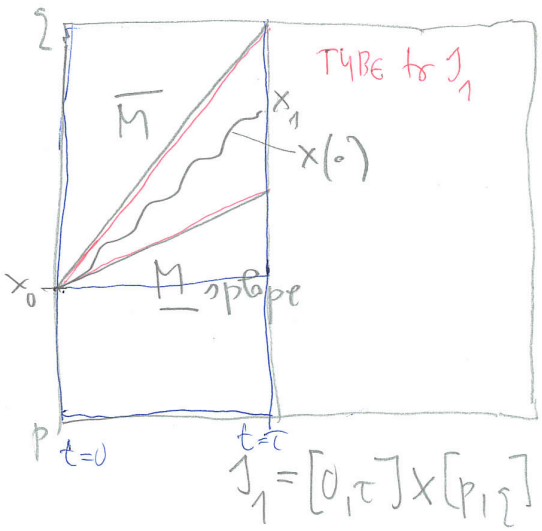
$$= x_0 + \sum_{k:k<\ell(t)} \int_{s\in I_k} f\big(s, x(s)\big)\ ds + \int\limits_{s=\tau_{\ell(t)-1}}^{\tau(\ell(t))} f\big(s, x(s)\big)\ ds \in X_n(t) \quad \text{where}$$

$$X_{n+1}(t) := X_0 + \sum_{k:k<\ell(t)} (\tau_k - \tau_{k-1}) F\Big(I_k \times \bigcup_{t\in I_k} X_n(t)\Big) + (t - \tau_{\ell(t)}) F\Big(I_{\ell(t)} \times \bigcup_{t\in I_k} X_n(t)\Big).$$

Observation. The upper and lower limits of the intervals $X_{n+1}(t)$ intv-ok as functions of t are the *piecewise linear* functions

$$t \mapsto a_{n+1}(t) := t \mapsto \min X_{n+1}(t), \quad t \mapsto b_{n+1}(t) := \max X_{n+1}(t)$$

Hence it is not difficukt to determine the intervals $\bigcup_{t\in I_k} X_n(t)$:



CTUBE 2

$$\bigcup_{t\in I_k} X_n(t) = \big[\min\{a_n(\tau_{k-1}, a_n(\tau_k)\}, \max\{a_n(\tau_{k-1}, a_n(\tau_k)\}\big].$$

$$\text{slope of } \big[a_{n+1} \text{ over } I_k\big] = F\Big(I_k \times \bigcup_{t\in I_k} X_n(t)\Big).$$

Piccard-Lindelöf tubes around the solution.

Since $x_n(t) \in X_n(t)$ and

$\big|x_*(t) - x_n(t)\big| \le \varepsilon_n(t) := \sum_{k:k>n} ML^{k-1}t^n/n!,$

the solution x_* has the property

$$a_n(t) - \varepsilon_n(t) \le x_*(t) \le b_n(t) + \varepsilon_n(t)x_*(t).$$

Thus its graph is contained in the following *Piccard-Lindelöf tubes*

$$\big(t, x(t)\big) \in \text{PL-TUBE}_n := \textstyle\bigcup_{s\in[0,\tau_*]} \big\{s\big\} \times \big[a_n(t) - \varepsilon_n(t), b_n(t) + \varepsilon_n(t)\big] \ .$$

Euler tubes around the solution

Technical assumption (slightly stronger than that in case of PL tubes):

$$f \prec F, \ \ L := \text{Lip}(F) < \infty, \quad -M \le \min f \le F(I) \le \max f \le M \ \ (\forall\, I).$$

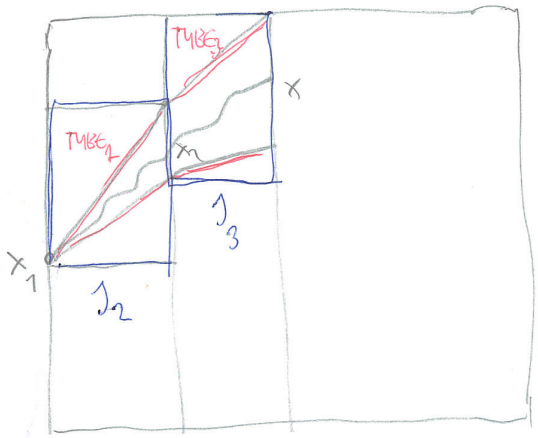
Notations as for case PL:

$x_0 \in (p,q), \ M := \max|f|, \ \tau_* := \min\big\{(q-x_0)/M, (x_0-p)/M\big\}, \ 0 = \tau_0 < \tau_1 < \cdots < \tau_N = \tau_*.$

$$\tfrac{d}{dt}x_*(t) = f\big(t, x_*(t)\big), \quad x_*(t) = x_0 + \int_{s=0}^{t} f\big(s, x_*(s)\big)\ ds \ .$$

(1) $t \in I_1 = [0, \tau_1] \implies x_*(t) \in X_0 + \big[-Mt, x_0 + Mt\big] \subset J_1 := X_0 + \big[-M\tau_1, M\tau_1\big]$

$$\tfrac{d}{dt}x_*(t) = f\big(t, x_*(t)\big) \in F(I_1 \times J_1), \quad x_*(t) \in X_0 + tF(I_1 \times J_1),$$



$x_*(t) \in \big[a(t), b(t)\big]$ $(t \in I_1)$, where

$a(\cdot), b(\cdot)$ are linear functions, $a(0) = \max X_0$, $b(0) := \min X_0$,

$\big[\text{slope of } a\big] = \min F(I_1 \times J_1)$, $\big[\text{slope of } b\big] = \max F(I_1 \times J_1)$.

Geometrically: by setting $X_1 := X_0 + \tau_1 F(I_1 \times J_1)$,

the graph of x_* passes in a trapesoid $\mathcal{T}_1$ whose

parallel edges are X_0 ill. X_1 and its height is $\tau_1 = \operatorname{diam}(I_1)$.

(k) If $a(\cdot), b(\cdot)$ are piecewise linear (continuous) functions on the intervals $I_1, \dots, I_{k-1}$
such that $a(t) \le x_*(t) \le b(t)$ $(0 \le t \le \tau_{k-1})$,
then $\frac{d}{dt}x_*(t) = f\big(t, x_*(t)\big) \in [-M, M] \implies$ over I_k we have
$x_*(t) \in \big[a(\tau_{k-1}) - M(t - \tau_{k-1}), b(\tau_{k-1}) + M(t - \tau_{k-1})\big] =: J_k$.
Hence also $\frac{d}{dt}x_*(t) = f\big(t, x_*(t)\big) \in F(I_k \times J_k)$, $x_*(t) \in \big[a(t), b(t)\big]$,
where $a(t) := a(\tau_{k-1}) + t \min F(I_k \times J_k)$, $b(t) := b(\tau_{k-1}) + t \max F(I_k \times J_k)$.

FIGURE

Geometrically: by setting $X_k := X_{k-1} + (\tau_k - \tau_{k-1})F(I_1 \times J_1)$,
the graph of x_* passes in a trapesoid $\mathcal{T}_k$ whose
parallel edges are X_{k-1} resp. X_k and its height is $\tau_k - \tau_{k-1} = \operatorname{diam}(I_k)$.

Euler tube E-TUBE $= \bigcup_{k=1}^{n} \mathcal{T}_k$.
Constructed to x_* with the time steps $0 = \tau_0, \dots, \tau_n$.

Estimate for the width of the tube.
Assumption: $\tau_k - \tau_{k-1} \equiv \tau_*/n =: \delta$ equidistant partition.
It suffices to obtain an upper estimate for the quantities $\ell_k := \operatorname{diam}(X_k)$ $(k = 1, \dots, n)$.

FIGURE

$\mathcal{T}_k \subset I_k \times \big[a(\tau_{k-1}) - M\delta, b(\tau_{k-1} + M\delta\big] = I_k \times J_k \Longrightarrow \operatorname{diam}(\mathcal{T}_k) \le \delta + \ell_{k-1} + 2M\delta;$

$\ell_k = \operatorname{diam}(X_k) = b(\tau_k) - a(\tau_k) =$

$= \big[b(\tau_{k-1}) + \delta \max F(I_k \times J_k)\big] - \big[a(\tau_{k-1}) + \delta \min F(I_k \times J_k)\big] =$

$= \big[b(\tau_{k-1}) - a(\tau_{k-1})\big] + \delta\big[\max F(I_k \times J_k) - \min F(I_k \times J_k)\big] =$

$= \ell_{k-1} + \delta \operatorname{diam} F(I_k \times J_k) \le \ell_{k-1} + \delta L \operatorname{diam}(I_k \times J_k) =$

$\le \ell_{k-1} + \delta L\big[\delta + \ell_{k-1} + 2M\delta\big] = (1 + \delta L)\ell_{k-1} + (2M+1)L\delta^2.$

We conclude by nduction on k:

$$\ell_k \le \ell_0(1+\delta L)^k + (2M+1)L\delta^2\big[1 + (1+\delta L) + \cdots + (1+\delta L)^{k-1}\big].$$

Remark. With a liear recursion $\ell_k = A\ell_{k-1} + B$ we get

$$\ell_k = A^k\ell_0 + (1 + A + \cdots + A^k)B = A^k\ell_0 + B(A^k - 1)/(A-1).$$

Thus

$$\ell_k \le (1+L\delta)^k\ell_0 + (2M+1)L\delta^2\big[(1+L\delta)^k - 1\big]/\big[(1+L\delta) - 1\big] =$$
$$= (1+L\delta)^k\ell_0 + (2M+1)\delta\big[(1+L\delta)^k - 1\big]\ .$$

Since $\delta = \tau_*/n$,

$$\ell_k \le (1 + L\tau_*/n)^k\ell_0 + (2M+1)(\tau_*/n)\big[(1+L\tau_*/n)^k - 1\big] \le$$
$$\le e^{L\tau_*}\ell_0 + \tau_*[e^{L\tau_*} - 1]/n$$

In particular we have concluded the following:

Therem. *With unlimited refinements of the time partition and starting from an error free data $\ell_0 = 0$, the Euler tubes converge uniformly to the graph of the solution.*

Elementary steps.

Suppose that $0 = \tau_0 < \tau_1 < \cdots < \tau_n = \tau_*$ and

$a, b : [0, \tau_*] \to \big[\min f, \max f\big]$ are such functions that

a, b are continuous and linear on the intervals $[\tau_{k-1}, \tau_k]$, furthermore

$$a(t) \le x_*(t) \le b(t),\ \tfrac{d}{dt}a(t) \le \tfrac{d}{dt}x_*(t) = f\big(t, x_*(t)\big) \le \tfrac{d}{dt}b(t) \quad \big(t \in [0, \tau_*]\big).$$

We insert a new point $\widetilde{\tau} \in (\tau_{\ell-1}, \tau_\ell)$ and then construct a "better" pair

$\widetilde{a}, \widetilde{b} : [0, \tau_*] \to \mathbb{R}$ of linear functions over the parition $I_0, I_1, \ldots, I_{n+1}$

with the endpoints $\tau_0, \ldots, \tau_{\ell-1}, \widetilde{\tau}, \tau_\ell, \ldots, \tau_n$ and satisfying

$(*)$ $\quad a \le \widetilde{a} \le x_* \le \widetilde{b} \le b,\ \tfrac{d}{dt}a \le \tfrac{d}{dt}\widetilde{a} \le \tfrac{d}{dt}x_* \le \tfrac{d}{dt}\widetilde{b} \le \tfrac{d}{dt}b.$

Let $\widetilde{a}(t) \equiv a(t)$, $\widetilde{b}(t) \equiv b(t)$ on the segments $t \le \tau_{\ell-1}$ (i.e. on the intervals $I_1, \ldots, I_{\ell-1}$).

Taking the intervals $I_\ell = [\tau_{\ell-1}, \widetilde{\tau}]$, $I_{\ell+1} = [\widetilde{\tau}, \tau_\ell]$, $I_{\ell+j+1} = [\tau_{\ell+j}, \tau_{\ell+j+1}]$ $(j = 1, \dots, n-\ell)$, consecutively over I_k for $k = \ell, \dots, n+1$, let

$$\big[\text{slope of } \widetilde{a}|I_k\big] := \max\Big\{\big[\text{slope of } a|I_k\big], \min F\big(I_k \times [\min a|I_k, \max b|I_k)\big)\Big\},$$
$$\big[\text{slope of } \widetilde{b}|I_k\big] := \min\Big\{\big[\text{slope of } b|I_k\big], \max F\big(I_k \times [\min a|I_k, \max b|I_k)\big)\Big\}.$$

In terms of the above slope data, consecutively for $k = \ell, \ell+1, \dots, n+1$, we can unambiguously determine the linear functions $\widetilde{a}|I_k, \widetilde{a}|I_k$. This is because we know the value and the slope at the starting point of the interval I_k at the begining of step k. Also the condition $(*)$ is fulfilled automatically due to the relationship $f \prec F$.

FIGURE

Exercise. Write an optimized algorithm for constructiong an Euler tube over uniform time partitions with $n = 1, 2, 4, 8, \dots, 2^r$ terms.

Applying 2D features.

Let $f : [0, T] \times [p, q] \to \mathbb{R}$ be as previously, furthermore $p < \xi_{-1} < \xi_0 < \xi_1 < q$.

Then any solution $\phi_0 : [0, \tau_*] \to [p, q]$ of the differential equation $\frac{d}{dt}x(t) = f\big(t, x(t)\big)$ starting from $(0, \xi_0)$ passes above the solution $\phi_{-1} : [0, \tau_*] \to [p, q]$ starting from $(0, \xi_{-1})$. Hence ϕ_0 passes above the lower limit curve of any Euler tube contructed for ϕ_{-1}.

Thus if such an Euler tube has the form

$$\Big\{(t, \xi) : \ t \in [0, \tau_*], \ \xi \in \big[a_{-1}(t), b_{-1}(t)\big]\Big\},$$

furthermore if the Euler tube of the solution ϕ_1 starting from $(0, \xi_0)$ is of the form

$$\Big\{(t, \xi) : \ t \in [0, \tau_*], \ \xi \in \big[a_1(t), b_1(t)\big]\Big\}$$

then

$$a_{-1}(t) \le \phi_0(t) \le b_1(t) \qquad \big(t \in [0, \tau_*]\big)$$

FIGURE

It may happen that $b_1(t)-a_{-1}(t) < \xi_1-\xi_{-1}$, despite of the fact that the Euler tubes become wider in time.

Example. $\frac{d}{dt}x(t) = -x(t)$, $\xi_k = k$ $(k=-1,0,1)$.

Euler tube with steps of higher order.

Iterating the differential equation $\frac{d}{dt}x(t) = f\big(t,x(t)\big)$ we get

$$\frac{d^k}{dt^k}x(t) = f_k\big(t,x(t)\big) \qquad (k=1,\dots,N)$$

with suitable functions $f_1,\dots,f_N$ whenever $f \in \mathcal{C}^N\big([0,T]\times[p,q]\big)$.

According to the Taylor formula,

$$x(t+h) = \sum_{k:k<N} \tfrac{1}{k!}x^{(k)}(t)h^k + \tfrac{1}{N!}x^{(N)}(s)h^N =$$
$$= \sum_{k:k<n} \tfrac{1}{k!}f_k\big(t,x(t)\big)h^k + \tfrac{1}{N!}f_N\big(s,x(s)\big)h^N, \qquad \text{where } \exists\, s\in[t,t+h].$$

Let $f_k \prec F_k$ $(k=1,\dots,N)$ and take a time partition $0=\tau_0<\dots<\tau_n=\tau_*$.

We construct a couple $a,b:[0,\tau_*]\to\mathbb{R}$ of functions ehose restrictions to the intervals $I_m := [\tau_{m-1},\tau_m]$ are polynomials of degree N such that $a(t)\le x_*(t)\le b(t)$ $(0\le t\le\tau_*)$.

Let also $a_0\le x_0\le b_0$, so that $a(0)=a_0$, $x_*(0)=x_0$, $b(0)=b_0$.

Assume that $a|[0,\tau_{m-1}]$, $b|[0,\tau_{m-1}]$ is constructed already.. Then

$$x_*(\tau_{m-1}+h) = \sum_{k:k<n} \tfrac{1}{k!}f_k\big(t,x_*(\tau_{m-1})\big)h^k + \tfrac{1}{N!}f_N\big(s,x_*(s)\big)h^N \in$$
$$\in \big[a(\tau_{m-1}),b(\tau_{m-1})\big] + \sum_{k:k<n} \tfrac{1}{k!}F_k\Big(\{\tau_{m-1}\}\times[a(\tau_{m-1}),b(\tau_{m-1})\Big)h^k +$$
$$+\tfrac{1}{N!}F_N\Big(I_m\times\big[a(\tau_{m-1})-M\mathrm{diam}(I_m),b(\tau_{m-1})+M\mathrm{diam}(I_m)\big]\Big)h^N.$$

Conclusion. *If we do not go beyond the domain of f, the choice below for the times $t\in I_m$ suits our requirements:*

$$a_m(t) := a(\tau_{m-1}) + \sum_{k:k<n} \tfrac{1}{k!}\min F_k\Big(\{\tau_{m-1}\}\times[a(\tau_{m-1}),b(\tau_{m-1})\Big)(t-\tau_{m-1})^k +$$
$$+\tfrac{1}{N!}\min F_N\Big(I_m\times\big[a(\tau_{m-1})-M\mathrm{diam}(I_m),b(\tau_{m-1})+M\mathrm{diam}(I_m)\big]\Big)(t-\tau_{m-1})^N,$$

$$b_m(t) := b(\tau_{m-1}) + \sum_{k:k<n} \tfrac{1}{k!} \max F_k\Big(\{\tau_{m-1}\} \times [a(\tau_{m-1}), b(\tau_{m-1})\Big)(t-\tau_{m-1})^k +$$
$$+ \tfrac{1}{N!} \max F_N\Big(I_m \times \big[a(\tau_{m-1}) - M\mathrm{diam}(I_m), b(\tau_{m-1}) + M\mathrm{diam}(I_m)\big]\Big)(t-\tau_{m-1})^N.$$

In case we go beyond, we insert a new partition piont $\widetilde{\tau} \in I_m$.

FIGURE

Example. ???????

RUNGE-KUTTA TYPE METHODS

Recall. Two functions $f, g : \mathbb{R} \to \mathbb{R}$, differ in *major order* n (*around* 0) if

$$g(h) = f(h) + M(h)h^n$$

for some function $M : \mathbb{R} \to \mathbb{R}$ bounded in some neighborhood of 0.

Similarly, $u, v : \mathbb{R} \to \mathbb{R}$ differ in *minor order* n if

$v(h) = u(h) + m(h)h^n$ for some function $m : \mathbb{R} \to \mathbb{R}$ with $\lim_{h\to 0} m(h) = 0$.

We write $g(h) = f(h) + O(h^n)$ resp. $v(h) = u(h) + o(h^n)$ to indicate such cases.

Setting. Let I, J be open intervals and $f : I \times J \to \mathbb{R}$ be a Lipschitz continuous function of two variables. Suppose the solution

$$(*) \qquad t \mapsto x(t) = {}^{x_0,t_0}x(t)$$

of the differential equation

$$(*) \qquad x'(t) = f\big(x(t), t + t_0\big), \quad x(0) = x_0$$

with initial value is well-defined for $t \in [0, T]$ whenever $x_0 \in [\xi_1, \xi_2] \subset I$ and $[t_0, t_0+T] \subset J$.

For every (small) number $h > 0$, we approximate the the points

$$\big({}^{x_0,t_0}x(0), t_0\big), \big({}^{x_0,t_0}x(h), t_0 + h\big), \ldots, \big({}^{x_0,t_0}x(nh), t_0 + nh\big) \quad \text{where} \quad nh \le T$$

with a sequence

$$\big({}^{x_0,t_0,h}y_0, t_0\big), \big({}^{x_0,t_0,h}y_1, t_0 + h\big), \ldots, \big({}^{x_0,t_0,h}y_n, t_0 + nh\big)$$

where ${}^{x_0,t_0,h}y_0 = x_0$ and

$${}^{x_0,t_0,h}y_{k+1} = Y\big({}^{x_0,t_0,h}y_k, t_0 + kh, h\big) \text{ with some formula } Y : \mathbb{R}^3 \to \mathbb{R}.$$

Definition. The above algorithm is a *forward approximation method of order* N for $(*)$ in the region $I \times J$ with time steps h if

$${}^{x_0,t_0}x(h) = {}^{x_0,t_0,h}y_1 + O(h^N) \quad \text{whenever} \quad \big[x_0 - h, x_0 + h\big] \subset I \text{ and } [t_0, t_0 + T] \subset J.$$

Example. *Euler method is of second order with smooth* f.

In our setting, with simplified notations $\big($omitting the initial value $(x_0, t_0)\big)$,

$$y_0 = x_0, \quad y_{k+1} = f\big(y_k, t_0 + kh\big)h \quad (k = 1, \ldots, n).$$

That is, more precisely,

$${}^{x_0,t_0,h}y_{k+1} = Y\big({}^{x_0,t_0,h}y_k, t_0 + kh, h\big) \quad \text{with} \quad Y\big(y, t, h\big) = y + f\big(y, t\big)h.$$

By the Taylor formula (first order with tail in Lagrange form),

$${}^{x_0,t_0}x(h) = {}^{x_0,t_0}x(0) + h \cdot \tfrac{d}{dt}\big|_{t=0}\,{}^{x_0,t_0}x(t) + \tfrac{1}{2}h^2 \cdot \tfrac{d^2}{dt^2}\big|_{t=\theta_{x_0,t_0,h}}\,{}^{x_0,t_0}x(t) =$$

$$= x_0 + h \cdot f\big(x_0, t_0\big) + O(h^2) \quad \text{provided} \quad t \mapsto {}^{x_0,t_0}x(t) \text{ is } \mathcal{C}^2\text{-smooth.}$$

Hence ${}^{x_0,t_0}x(h) - {}^{x_0,t_0,h}y_1 = \big[x_0 + h\cdot f\big(x_0,t_0\big) + O(h^2)\big] - \big[y_0 + f\big(y_0,t_0\big)h\big] = O(h^2)$.
Since $y_0 = x_0$, it follows ${}^{x_0,t_0}x(h) = {}^{x_0,t_0,h}y_1 + O(h^2)$.

Theorem. *In the setting above,*

$$\big|{}^{x_0,t_0}x(T) - {}^{x_0,t_0,T/n}y_n\big| = O\big(\big[\tfrac{T}{n}\big]^{N-1}\big) \quad \textit{uniformly.}$$

Remark. In more details, the theorem asserts the following:

(i) Given any intervals $I_0 \subset I$ and $J_0 \subset J$ with $\operatorname{dist}(I_0, I), \operatorname{dist}(J_0, J) > 0$, there exists $T > 0$ such that all the terms

$${}^{x_0,t_0}x(t)\ \big(x_0 \in I_0, t_0 \in J_0, t_0 + T \in J\big), \quad {}^{x_0,t_0,h}y_k\ \big(0 \le k \le T/h\big) \quad \text{are well-defined,}$$

(ii) $\big|{}^{x_0,t_0}x(T) - {}^{x_0,t_0,T/n}y_n\big| \le M^*\,\big(\frac{T}{n}\big)^{N-1}$ in all cases of (ii) with a common constant M^*.

Proof. The statements in (i) belong to the standard introductory texts in ODE. In particular we may take $T < \operatorname{dist}(J_0, J)$ with $I_0 + \sup|f(I,J)|\cdot[-1,1] \subset J$ implies that the points ${}^{x_0,t_0}x(t) \in I\ \ (0 \le t \le T)$ are well-defined. In most ODE introductions, the statement concerning the approximating points ${}^{x_0,t_0,h}y_k$ is discussed only for the Euler method $\big(\text{with } Y(y,h) = f(y,t)h\big)$.

(ii) Let us fix x_0, t_0, n with T and $h = T/n$ such that the hypothesis in (i) apply. For short, write

$$x(t) = {}^{x_0,t_0}x(t), \quad , t_k = t_0 + k\tfrac{T}{n}, \quad y_k = {}^{x_0,t_0,T/n}y_k \quad \text{so that}$$
$$x'(t) = f\big(x(t), t_0 + t\big),\ \ x(0) = x_0, \qquad y_0 = x_0,\ \ y_{k+1}(y_k + Y\big(y_k, t_k, \tfrac{T}{n}\big).$$

We have to estimate the difference $\big|x(T) - y_n\big|$.

By assumption, Y provides a method of order N. That is

$$\big|y_1 - x(t_1)\big| \le M\big(\tfrac{T}{n}\big)^N.$$

However, for $k > 0$, we cannot compare $x(t_{k+1})$ with y_{k+1} immediately in the same manner because the point $\big(y_{k+1}, t_{k+1}\big)$ is not the endpoint of the curve $[0, T/n] \ni t \mapsto x\big(x(t{+}t_k), t\big)$.

Observation: $\big(y_{k+1}, t_{k+1}\big)$ is the endpoint of the curve $[0, T/n] \ni t \mapsto \big(z_k(t), t + t_k\big)$ where

$$z_k(t) = {}^{y_k,t_k}x(t) \quad \text{satisfying} \quad z_k'(t) = f\big(z_k(t), t + t_k\big),\ \ z_k(0) = y_k.$$

FIGURE

We can apply the step-formula to x_k by considering the first approximation step with time $\frac{T}{n}$ from (y_k, t_k) along the curve $t \mapsto \big(z_k(t), t + t_k\big)$:

$$y_{k+1} - z_k(T/n) = Y\big(y_k, t_k, T/n\big), \quad \big|y_{k+1} - x_k(T/n)\big| \le M\big(\tfrac{T}{n}\big)^N.$$

On the other hand,

$$z_k(T/n) - x_{k+1} = z_k(T/n) - x(t_k + T/n) = \delta_(T/n) \quad \text{where}$$
$$\delta_k(t) = z_k(t) - x(t + t_k) \quad \text{satisfying}$$
$$\delta_k(0) = y_k - x_k, \;\; \delta_k'(t) = f\big(z_k(t), t + t_k\big) - f\big(x(t + t_k), t + t_k\big).$$

By setting $\widetilde{M} = \sup\big|f(I, J)\big| \,(< \infty)$, we have $\big|\delta'(t)\big| \le 2\widetilde{M}$ and hence

$$\big|\delta_k(t)\big| \le \big|\delta_k(0)\big| \cdot \exp\big(2\widetilde{M}t\big) = |y_k - x_k| \exp\big(2\widetilde{M}t\big).$$

In particular $|x_{k+1} - x_k(T/n)| = \delta_k(T/n) \le |y_k - x_k| \cdot \exp\big(2\widetilde{M}T/n\big)$. It follows

$$\big|y_{k+1} - x_{k+1}\big| \le \big|y_{k+1} - z_k(T/n)\big| + \big|z_k(T/n) - x_{k+1}\big| \le$$
$$\le M\big(\tfrac{T}{n}\big)^N + |y_k - x_k| \cdot \exp\big(2\widetilde{M}T/n\big).$$

By setting $Q_n = \exp\big(2\widetilde{M}T\big)$, we get

$$\big|y_1 - x_1\big| \le M\big(\tfrac{T}{n}\big)^N,$$
$$\big|y_2 - x_2\big| \le M\big(\tfrac{T}{n}\big)^N + |y_1 - x_1| \cdot Q_n \le M\big(\tfrac{T}{n}\big)^N \Big[1 + Q_n\Big],$$
$$\big|y_3 - x_3\big| \le M\big(\tfrac{T}{n}\big)^N + |y_2 - x_2| \cdot Q_n \le M\big(\tfrac{T}{n}\big)^N \Big[1 + Q_n + Q_n^2\Big],$$
$$\vdots \qquad\qquad \vdots$$
$$\big|y_n - x_n\big| \le M\big(\tfrac{T}{n}\big)^N \Big[1 + Q_n + Q_n^2 + \cdots + Q_n^{n-1}\Big].$$

We complete thr proof with the observation that $Q_n^n = \exp\big(2\widetilde{M}T\big)$ and hence

$$\big|y_n - x_n\big| \le M\Big(\frac{T}{n}\Big)^N \frac{\exp(2\widetilde{M}T) - 1}{\exp(2\widetilde{M}T/n) - 1} = MT\Big(\frac{T}{n}\Big)^{N-1} \frac{\exp(2\widetilde{M}T) - 1}{n[\exp(2\widetilde{M}T/n) - 1]} \nearrow$$
$$\nearrow \frac{M}{2\widetilde{M}}\big[\exp(2\widetilde{M}T) - 1\big]\Big(\frac{T}{n}\Big)^{N-1}.$$

Runge-Kutta method RK2 of 2nd order

Consider the solution

$$[0, T] \ni t \mapsto \big[x(t), t\big] \quad \text{with} \quad x(t) = {}^{x_0,t_0}x(t)$$

of the ODE $x'(t) = f\big(x(t), t\big)$ with initial value $x(t_0) = x_0$.

Given a step size $h > 0$, we define a sequence

$$\big[y_0, t_0\big], \big[y_1, t_1\big], \ldots, \big[y_n, t_n\big] \quad \text{with} \quad t_k = t_0 + kh, \;\; nh \le T,$$

of approximating points with $y_k = {}^{x_0,t_0,h}y_k$ for the accurate solution points

$$\big[x_0, t_0\big], \big[x_1, t_1\big], \ldots, \big[x_n, t_n\big] \quad \text{with} \quad x_k = x(t_k) = {}^{x_0,t_0}x(t_k)$$

recursively as follows:

$$y_0 = x_0, \quad y_{k+1} = y_k + \tfrac{1}{2}A_k + \tfrac{1}{2}B_k$$
$$\text{where} \quad A_k = h \cdot f\big(y_k, t_k\big), \quad B_k = h \cdot f\big(y_k + A_k, t_{k+1}\big).$$

Definition. We define the *RK1,RK2 step operators* (first and second order Runge-Kutta step operators) with the formula

$$Y_1\big(y,t,h\big) = h \cdot f\big(x,t\big),$$
$$Y_2\big(y,t,h\big) = \frac{1}{2}Y_1\big(y,z,h\big) + \frac{1}{2}Y_{2,1}\big(y,z,h\big)$$
$$\text{where} \quad Y_{2,1}\big(y,t,h\big) = h \cdot f\big(y + Y_1(y,t,h), t+h\big).$$

Remark. RK1 corresponds simply to Euler's method, while RK2 corresponds to the above approximation procedure (called the *second order Runge-Kutta* method).

Theorem. *Suppose $f : I \times J \to \mathbb{R}$ is has continuous second partial derivatives. Then the sequence $\big[y_k\big]$ constructed with the RK2 operator approximates the accurate solution points $[x_k]$ in 2nd order.*

Proof. It is well-known that the $\mathcal{C}^2$-smoothness of f implies the $\mathcal{C}^2$-smoothness of the solution $t \mapsto x(t) = {}^{x_0,t_0}x(t)$. Hence we can werite it in the Taylor form as

$$x(h) = x(0) + h \cdot x'(0) + \frac{h^2}{2} \cdot x''(0) + O(h^3).$$

Here we have

$$x'(0) = \frac{d}{dt}\Big|_{t=0} = f\big(x(0), t_0\big) = f\big(x_0, t_0\big),$$
$$x''(0) = \frac{d}{dt}\Big|_{t=0} x'(t) = \frac{d}{dt}\Big|_{t=0} f\big(x(t), t_0 + t\big) =$$
$$= \frac{\partial f}{\partial x}\Big|_{(x,t)=(x(0),t_0)} x'(0) + \frac{\partial f}{\partial t}\Big|_{(x,t)=(x(0),t_0)} \cdot 1 = \Big[\frac{\partial f}{\partial x} \cdot f + \frac{\partial f}{\partial t}\Big]_{(x,t)=(x_0,t_0)}.$$

Thus

$$x(h) = x_0 + h \cdot f\big(x_0, t_0\big) + \frac{h^2}{2} \cdot \frac{\partial f}{\partial x}\Big|_{(x_0,t_0)} f\big(x_0, t_0\big) + \frac{h^2}{2} \cdot \frac{\partial f}{\partial t}\Big|_{(x_0,t_0)} + O(h^3).$$

We evaluate y_1 by means of A_0, B_0 expressed in terms of x_0, t_0, f. We simply have $A_0 = h \cdot f(x_0, t_0)$ and

$$B_0 = h \cdot f\big(y_0 + A_0, t_1\big) = h \cdot f\big(x_0 + h \cdot f(x_0,t_0), t_0 + h\big) =$$
$$= h \cdot \Big[f\big(x_0, t_0\big) + \frac{\partial f}{\partial x}\Big|_{(x_0,t_0)} h \cdot f(x_0,t_0) + \frac{\partial f}{\partial t}\Big|_{(x_0,t_0)} h \cdot 1 + O(h^2)\Big],$$
$$y_1 = y_0 + \frac{1}{2}A_0 + \frac{1}{2}B_0 =$$
$$= x_0 + h \cdot f\big(x_0, t_0\big) + \frac{h^2}{2} \cdot \frac{\partial f}{\partial x}\Big|_{(x_0,t_0)} f\big(x_0, t_0\big) + \frac{h^2}{2} \cdot \frac{\partial f}{\partial t}\Big|_{(x_0,t_0)} + O(h^3).$$

We complete the proof by observing that $x_1 - y_1 = x(h) - y_1(h) = O(h^3)$.

Runge-Kutta methods of higher order

Generic pattern. Given $N \in \mathbb{N}$, we construct an approximation step in the form

$$Y_N\big(y,t,h\big) = y + h\cdot\big(b_1K_1 + \cdots + b_NK_M\big) \quad \text{with}$$
$$K_1 = f\big(y,t\big),$$
$$K_2 = f\big(y + h\cdot a_{2,1}K_1, t + h\cdot c_2\big),$$
$$K_3 = f\big(y + h\cdot[a_{3,1}K_1 + a_{3,2}K_2], t + h\cdot c_3\big),$$
$$\vdots$$
$$K_N = f\big(y + h\cdot[a_{N,1}K_1 + a_{N,2}K_2 + \cdots + a_{N,N-1}K_{N-1}], t + h\cdot c_3\big)$$

where the coefficients

$$a_{i,j}\ \ \big(2 \le i \le N, 1 \le j < i\big); \quad b_k\ \ \big(1 \le k < N\big), \quad c_\ell\ \ \big(1 \le \ell \le N\big)$$

are so chosen that we have

$$x(t+h) - Y_N\big(x(t),h,t\big) = O(h^{N+1}) \quad \text{for the solution } t \mapsto x(t) \text{ of } \quad x'(t) = f\big(x(t),t\big)$$

in some open region within the domain of any $(N+1)$-times continuously differentiable function f of 2 variables.

Analogously as in the case of $N = 2$ discussed in details, this means that

the Taylor polynomial of order N for $h \mapsto x(t+h)$ should coincide

with the Taylor polynomial of order N for $h \mapsto Y_N\big(x(t),t,h\big)$

in case of any possibe function f.

That is, for $k = 1,\dots,N,$ we must have

$$\frac{d^k}{dt^k}x(t) = \frac{d^k}{dh^k}\Big|_{h=0} Y_N\big(x(t),t,h\big).$$

As for the beginning, this means that

$k=1)$ $\quad x'(t) = \frac{d}{dh}\big|_{h=0} Y_N\big(x(t),t,h\big),$

$$f\big(x(t),t\big) = [b_1 + \cdots + b_N]f\big(x(t),t\big);$$

$k=2)$ $\quad x''(t) = \frac{d^2}{dh^2}\big|_{h=0} Y_N\big(x(t),t,h\big),$

$$\tfrac{d}{dt}f\big(x(t),t\big) = \tfrac{d^2}{dh^2}\big|_{h=0} h\cdot\Big[b_1 f\big(x(t),t\big) + b_2 f\Big(x(t) + ha_{2,1}f\big(x(t),t\big) + \cdots\Big],$$

$$f^{(1,0)}\big(x(t),t\big)x'(t) + f^{(0,1}\big(x(t),t\big) =$$

$$= 2\Big[\sum_{k=2}^{N}\big(\sum_{\ell<k} a_{k,\ell}\big)b_k\Big]f^{(1,0)}\big(x(t),t\big)x'(t) + 2\Big[\sum_{k=2}^{N} b_kc_k\Big]f^{(0,1)}\big(x(t),t\big),$$

$$f^{(1,0)}\big(x(t),t\big)f\big(x(t),t\big) + f^{(0,1)}\big(x(t),t\big) =$$

$$= 2\Big[\sum_{k=2}^{N}\big(\sum_{\ell<k} a_{k,\ell}\big)b_k\Big]f^{(1,0)}\big(x(t),t\big)f\big(x(t),t\big) + 2\Big[\sum_{k=2}^{N} b_kc_k\Big]f^{(0,1)}\big(x(t),t\big).$$