Numerical methods

(1) What are floating point numbers $\mathbb{F}$?

- $\mathbb{F} \subset \mathbb{Q}$
- rounding function $\varsigma : \mathbb{R} \to \mathbb{F}$
- properties:
$$|\varsigma(x) - x| \leq \varepsilon |x|$$
for some ε, mostly $\varepsilon \approx 10^{-14}$
- "computer arithmetic":
$$x \oplus y = \varsigma(x+y)$$
- error $|x \oplus y - x + y| \leq \varepsilon |x+y|$
- computers use dyadic fractions
$$x = \frac{n}{2^e} \quad \text{where } n, e \in \mathbb{Z}$$
- domains of n & e are finite
- this model is usually quite accurate
- some results are exact:
  - if $x \in \mathbb{F}$ the $2^k \odot x = \varsigma(2^k x) = 2^k x$
  - if $|x-y|$ small $\Rightarrow$ $x \ominus y = \varsigma(x-y) = x-y$
- error propagation can be a problem
  - let $x, y \in \mathbb{R}$
  - computed result

$$\begin{aligned} x - y &\approx \varsigma(\varsigma(x) - \varsigma(y)) \\ &= (x(1+\delta_1) - y(1+\delta_2))(1+\delta_3) \\ &= (x-y)\cdot\left(1 + \frac{x\delta_1 - y\delta_2}{x-y}\right)(1+\delta_3) \end{aligned}$$

here: $\delta_i = \frac{\varsigma(x)-x}{|x|}$ etc — $\frac{x\delta_1 - y\delta_2}{x-y}$ can be large (eg. if $|x| \gg |x-y|$)

(2) The numerical solution of univariate polynomials

Definition (numerical solution)

A numerical solution of $f(\xi) = 0$ consists of 2 components:

a) an approximation ξ_0 of the solution ξ^*

b) an algorithm which produces a sequence $\xi_0, \xi_1, \xi_2, \ldots$ which converges to ξ^*

- there are many algorithms ....
- a non-solution: (Newton's method)

$\left(\xi_0,\ \xi_{n+1} = \xi_n - \frac{f(\xi_n)}{f'(\xi_n)}\right)$ may not converge

Example (from stackexchange)

- $\xi_0 = 5/9$
- $f(\xi) = -0{,}74 + 0{,}765x + 1{,}1x^2 - 3{,}55x^3$

$\Rightarrow$ solution oscillates

Simple example

$$\begin{cases} f(\xi) = \xi^2 - 2 \\ \xi_0 = 0 \end{cases} \Rightarrow f'(\xi_0) = 0 \text{ and thus } \xi_1 = \infty$$

Example which works

if we know bounds $a \leq \xi^* \leq b$
and $f(a) \cdot f(b) < 0$
then choose

bisection method

- $\xi_0 = a$
- $\xi_1 = b$
- $\xi_n = \frac{\xi_{n-1} + \xi_{n-2}}{2}$ for n even
- $\xi_{n+1} = \begin{cases} \xi_{n-1} & \text{if } f(\xi_{n-1}) \cdot f(\xi_n) < 0 \\ \xi_{n-2} & \text{else} \end{cases}$

- error $|\xi_n - \xi^*| \leq C \cdot 2^{-n}$
  (slow) convergence (one binary digit/iteration)
- requires upper & lower bounds
- for real zeros of real polynomials

- **Simple example with complex zeros**

- $f(\xi) = \xi^2 + 2$
- $\xi_0 \in \mathbb{R} \Rightarrow$ Newton's method does not work

$\rightsquigarrow$ choose ξ_0 complex or ξ_1 complex

- The number of zeros

- Cauchy

$$\frac{1}{2\pi i}\oint_{\Omega}\frac{f'(z)}{f(z)}\,dz = N_{zeros} - N_{poles}$$

in simply connected sets $\Omega \subset \mathbb{C}$

for meromorphic fcts
$\Rightarrow$ rational fcts

- Sturm sequence for real roots

$p_0(x) = p(x)$, $p_1(x) = p'(x)$

$p_n(x) = -\,p_{n-2}(x) \,\%\, p_{n-1}(x)$ (remainder)

for $n = 2, 3, \dots$ $\quad$ s.t. $p_{M+1}(x) = 0$

Sturm's theorem: If $a < b$ & $p(a)\,p(b) \neq 0$ then

sign changes $\{p_0(a), p_1(a), \dots, p_M(a)\}$
$-$ # sign changes $\{p_0(b), p_1(b), \dots, p_M(b)\}$
$= \#\{x \in (a,b) \mid p(x) = 0\}$

Example $\quad p(x) = x^3 - x$ $\quad$ (zeros $= \pm 1, 0$)

let $a = -2$, $b = 2$

$p_0(x) = x^3 - x$, $p_1(x) = 3x^2 - 1$, $p_2(x) = -\,p_0 \,\%\, p_1 = +\frac{2}{3}x$

$p_3(x) = +1$ $\Rightarrow m = 3$

$\Rightarrow p_0(a) \dots p_3(a) = -14, 11, -\frac{4}{3}, 1 \rightarrow 3$ sign changes

$p_0(b) \dots p_3(a) = 6, 11, 4/3, 1 \rightarrow 0$ sign changes

$\Rightarrow$ V.

Application: combine with bisection to get all zeros in (a,b)

③ Newtons method

$$x^{(n+1)} = x^{(n)} - \frac{p(x^{(n)})}{p'(x^{(n)})}$$

how to analyse convergence

$x^{(n+1)} = F(x^{(n)})$ with $F(x) = x - \frac{p(x)}{p'(x)}$

- is fixed point iteration

$\Rightarrow$ sequence convergent if F contractive

Lipschitz constant of F:

$$L = \|F'\|_\infty = \sup_{x \in [a,b]} |F(x)|$$

error $\|x^{(n)} - x^*\| \leq L^n \|x^{(0)} - x^*\|$

where $x^* = F(x^*)$

$\Rightarrow$ $x^{(n)} \to x^*$ for $n \to \infty$ (convergence)

- apply to Newton's method

$$F'(x) = \cancel{1} - \cancel{\frac{p'(x)}{p'(x)}} + \frac{p(x)p''(x)}{p'(x)^2}$$

$\Rightarrow$ convergence if $\underbrace{\left|\frac{p(x)p''(x)}{p'(x)^2}\right|}_{= L(x)} < 1$ for $x \in U_\varepsilon(x^*)$ (Lipschitz constant)

but $p(x^*) = 0$

so if $p'(x^*) \neq 0 \Rightarrow L(x) < 1$ if x sufficiently close to x^*

- also: $L(x^{(n)}) \to 0$ for $x^{(n)} \to x^*$

$\Rightarrow ?$

$$|x^{(n)} - x^*| \leq C |x^{(n-1)} - x^*|^2$$

quadratic convergence

- case of zeros with multiplicity > 1

example: $p(x) = (x-1)^2$

$$\Rightarrow L(x) = \frac{(x-1)^2 \cdot 2}{4(x-1)^2} = \frac{1}{2}$$

$\Rightarrow$ slow convergence

- how to deal with singular points

transform $$q(x) = \frac{p(x)}{\gcd(p(x), p'(x))}$$

maps $\langle p \rangle \to \sqrt{\langle p \rangle}$ radical ideal

and q has only simple zeros

proof: if x^* zero of p with multiplicity α then it is a zero of p' with multiplicity $\alpha - 1$ ... □

- deflation or avoiding revisiting same zero

- if $\tilde{x}$ approximates zero of $p(x)$ then the remaining zeros are obtained by solving $q(x)$ with $q(x) = \frac{p(x)}{x - \tilde{x}}$
- in practice $p(\tilde{x}) \approx 0$ - we compute an approximation

$\Rightarrow q(x)$ is a rational function

④ polynomials and matrices

- **companion matrix** of $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$

$$C = \begin{bmatrix} 0 & & & -a_0 \\ 1 & \ddots & & -a_1 \\ & \ddots & 0 & \vdots \\ & & 1 & -a_{d-1} \end{bmatrix}$$

$\Rightarrow p(x) = \det(C - xI)$

thus zeros of $p(x)$ are eigenvalues of C

the

- **Evaluation of $p(x)$**

Horners scheme: $p(x) = (\cdots(x + a_{n-1})\cdot x + a_{n-2})\cdots) + a_0$

or

$$\boxed{p_k(x) = x\,p_{k-1}(x) + a_{n-k}\,p_0(x), \quad p_0(x) = 1}$$

$\Rightarrow$ companion matrix C

- recursions for polynomials $\Rightarrow$ matrices

- more generally, eigenvalues of any matrix A are the zeros of the characteristic polynomial

$$p(x) = \det(A - xI)$$

- in practice one uses methods related to the power method to compute eigenvalues

$v^{(n+1)} = Av^{(n)} \Rightarrow v^{(n)} \to v_\lambda$ with $Av_\lambda = \lambda v_\lambda$

λ is eigenvalue with $|\lambda|$ maximal ($v^{(0)} \neq 0$)

- numerically, computing the eigenvalues is a much better conditioned problem than solving $p(x) = 0$.

- polynomials defined by recursion

$p_0(x) = 1$, $p_1(x) = x - a_1$ and

$$p_k(x) = (x - a_k)p_{k-1}(x) - b_{k-1}^2 p_{k-2}(x), \quad k = 2, \dots, n$$

then $p_n(x) = \det(A - xI)$ where

$$A = \begin{bmatrix} a_1 & b_1 & & \\ b_1 & a_2 & b_2 & \\ & b_2 & a_3 & b_3 \\ & & \ddots & \ddots & \ddots \end{bmatrix} \Rightarrow$$

use matrix A and do not use characteristic polynomials $p_n(x)$ to compute zeros of $p_n(x)$

5. Numerical problems solving $p(x) = 0$

- the numerical analyst Wilkinson studied the polynomial

$$p(x) = (x-1)(x-2)\cdots(x-20)$$

he observed that zeros are hard to compute from polynomial coefficients

- note that this polynomial can be defined recursively as $p_k(x) = (x-k)p_{k-1}(x)$, $p_0(x) = 1$ and the zeros are the eigenvalues of $A = \begin{bmatrix} 1 & & 0 \\ & 2 & \\ 0 & & \ddots \\ & & & 20 \end{bmatrix}$
- it turns out that computing zeros from the polynomial coefficients is ill-conditioned in this case
- condition number of a function f

- compute result y from data x
- condition ~ how change in x affects y

$$y = f(x) \Rightarrow y + \Delta y = f(x + \Delta x)$$

- we are interested in "relative error"

thus $$\frac{\|\Delta y\|}{\|y\|} \leq K \frac{\|\Delta x\|}{\|x\|}, \quad K = \text{cond. number}$$

- linear approximation $\Delta y = J \Delta x$,

and, as, $f(0) = 0 \Rightarrow y \approx Jx$

$$\Rightarrow K = \sup \frac{\|\Delta y\|/\|\Delta x\|}{\|y\|/\|x\|} \lesssim \sup_{\Delta x} \frac{\|\Delta y\|}{\|\Delta x\|} \cdot \sup_{x} \frac{\|x\|}{\|y\|}$$

$$\boxed{K \approx \|J\| \cdot \|J^{-1}\|}$$

- in our case

$$p(x) = x^n + a_{n-1}x^{n-1} + \dots + a_0$$
$$= (x - x_1) \cdots (x - x_n)$$

then $$\boxed{a_k = (-1)^k \sigma_k(x_1, \dots, x_n)} \quad (*)$$ elem. symm. polynomials

data: $a_0, \dots, a_{n-1}$

result: $x_1, \dots, x_n$

$\Rightarrow$ we need to invert polynomial system $(*)$

$\Rightarrow J^{-1} = \left[\frac{\partial \underline{a}}{\partial \underline{x}}\right]$ where $\underline{a} = (a_0 \cdots a_{n-1})^T$, $\underline{x} = (x_1, \dots, x_n)^T$

$\Rightarrow$ compute $K = \|J^{-1}\| \cdot \|J\| = \frac{s_{max}}{s_{min}}$

where s_{max}, s_{min} = max. & min. singular values

- what is the effect of rounding errors

under certain conditions one has

$$\text{error} \approx C_n \cdot K \cdot \varepsilon$$

- result on eigenvalue perturbation.

$$|\mu - \lambda_k| \leq \kappa_2(X)\,\|E\|_2$$

where

- $A \in \mathbb{C}^{n\times n}$ nondefective, nonsingular
- $A = X \Lambda X^{-1}$ eigenvalue decomposition
- μ any eigenvalue of $A+E$
- λ_k eigenvalue of A closest to μ

if eigenvectors of A are orthogonal (i.e. $AA^T = A^TA$, normal matrix) then $\kappa_2(X) = 1$.

N.B. $\kappa_2(X) \neq \kappa_2(A)$!

($\|\cdot\|_2$, κ_2 ~ Spectral norm)