LECTURE NOTES

on

ELEMENTARY NUMERICAL METHODS

Eusebius Doedel

TABLE OF CONTENTS

VECTOR AND MATRIX NORMS

In later analysis we shall need a quantity (called *vector norm*) that measures the magnitude of a vector.

Let $\mathbf{x} \equiv (x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^n$.

EXAMPLES (of norms) :

$$\| \mathbf{x} \|_1 \equiv \sum_{k=1}^{n} | x_k | \ , \qquad \text{(the “\textit{one-norm} ”)}$$

$$\| \mathbf{x} \|_2 \equiv (\sum_{k=1}^{n} {x_k}^2)^{\frac{1}{2}} \ , \qquad \text{(the “\textit{two-norm} ”, or Euclidean length)}$$

$$\| \mathbf{x} \|_\infty \equiv \max_{1 \ \leq \ k \ \leq \ n} | x_k | \ , \ \text{(the “\textit{infinity-norm} ”, or “max-norm”)}$$

$\| \mathbf{x} \|_1$ and $\| \mathbf{x} \|_2$ are special cases of

$$\| \mathbf{x} \|_p \equiv (\sum_{k=1}^{n} | x_k |^p)^{\frac{1}{p}} , \qquad \text{(where } p \text{ is a positive integer),}$$

while for any fixed vector $\mathbf{x}$ we have

$$\| \mathbf{x} \|_\infty \quad \text{is the limit of} \quad \| \mathbf{x} \|_p \quad \text{as} \quad p \to \infty . \qquad \text{(Check !)}$$

EXAMPLE : If $\mathbf{x} = (1, -2, 4)^T$ then

$$\| \mathbf{x} \|_1 = 7 , \qquad \| \mathbf{x} \|_2 = \sqrt{21} , \qquad \| \mathbf{x} \|_\infty = 4 .$$

Vector norms are required to satisfy

$$(i) \quad \| \mathbf{x} \| \;\geq\; 0, \quad \forall \mathbf{x} \in \mathbb{R}^n \quad \text{and} \quad \| \mathbf{x} \| \;=\; 0 \;\text{ only }\; \text{if } \mathbf{x} \;=\; \mathbf{0},$$

$$(ii) \quad \| \alpha \mathbf{x} \| \;=\; | \, \alpha \, | \;\; \| \mathbf{x} \|, \quad \forall \mathbf{x} \in \mathbb{R}^n, \quad \forall \alpha \in \mathbb{R},$$

$$(iii) \quad \| \mathbf{x} + \mathbf{y} \| \;\leq\; \| \mathbf{x} \| + \| \mathbf{y} \|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad (\textit{Triangle inequality}).$$

All of the examples of norms given above satisfy (i) and (ii). (**Check !**)

To check condition (iii) let

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^T, \qquad \mathbf{y} = (y_1, y_2, \cdots, y_n)^T \ .$$

Then

$$\begin{aligned} \| \mathbf{x} + \mathbf{y} \|_1 &= \textstyle\sum_{k=1}^n \mid x_k + y_k \mid \ \leq \ \sum_{k=1}^n (\ \mid x_k \mid + \mid y_k \mid \ ) \\ &= \textstyle\sum_{k=1}^n \mid x_k \mid + \sum_{k=1}^n \mid y_k \mid \ = \ \| \mathbf{x} \|_1 + \| \mathbf{y} \|_1 \ . \end{aligned}$$

$$\| \mathbf{x} + \mathbf{y} \|_2 \ \leq \ \| \mathbf{x} \|_2 + \| \mathbf{y} \|_2 \qquad \text{“by geometry''} \quad \text{(Proof given later.)}$$

$$\begin{aligned} \| \mathbf{x} + \mathbf{y} \|_\infty &= \max_k \mid x_k + y_k \mid \ \leq \ \max_k \mid x_k \mid + \max_k \mid y_k \mid \\ &= \| \mathbf{x} \|_\infty + \| \mathbf{y} \|_\infty \ . \end{aligned}$$

EXERCISES :

- Let $\mathbf{x} = (1, -2, 3)^T$. Compute $\| \mathbf{x} \|_1$, $\| \mathbf{x} \|_2$, and $\| \mathbf{x} \|_\infty$.

- Graphically indicate all points $\mathbf{x} = (x_1, x_2)^T$ in $\mathbb{R}^2$ for which $\| \mathbf{x} \|_2 = 1$.
  Do the same for $\| \mathbf{x} \|_1$ and $\| \mathbf{x} \|_\infty$.

- Graphically indicate all points $\mathbf{x} = (x_1, x_2)^T$ in $\mathbb{R}^2$ for which $\| \mathbf{x} \|_2 \leq 1$.
  Do the same for $\| \mathbf{x} \|_1$ and $\| \mathbf{x} \|_\infty$.

- Graphically indicate all points $\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ with $\| \mathbf{x} \|_2 = 1$.
  Do the same for $\| \mathbf{x} \|_1$ and $\| \mathbf{x} \|_\infty$.

- Prove that $\| \mathbf{x} \|_1 \quad \leq \quad n \, \| \mathbf{x} \|_\infty$.

- Prove that $\| \mathbf{x} \|_2 \quad \leq \quad \sqrt{n} \, \| \mathbf{x} \|_\infty$.

- Prove that $\| \mathbf{x} \|_2 \quad \leq \quad \| \mathbf{x} \|_1$.

- Prove that $\lim_{p \rightarrow \infty} \| \mathbf{x} \|_p \quad = \quad \| \mathbf{x} \|_\infty$.

We also need a measure of the magnitude of a square matrix (*matrix norm*).

This is defined *in terms of a given vector norm* , namely,

$$\| \mathbf{A} \| \ \equiv \ \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{Ax} \|}{\| \mathbf{x} \|} \ .$$

Thus $\| \ \mathbf{A} \ \|$ measures the *maximum relative stretching* in a given vector norm that occurs when multiplying all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$ by $\mathbf{A}$.

From this definition it follows that for arbitrary $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{y} \neq \mathbf{0}$, we have

$$\frac{\| \mathbf{Ay} \|}{\| \mathbf{y} \|} \ \leq \ \| \mathbf{A} \| \ ,$$

i.e.,

$$\| \mathbf{Ay} \| \ \leq \ \| \mathbf{A} \| \ \ \| \mathbf{y} \| \ .$$

For specific choices of vector norm it is convenient to express the induced matrix norm directly in terms of the elements of the matrix :

For the case of the $\| \cdot \|_\infty$ let

$$\mathbf{A} \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \qquad \text{and let} \qquad R \equiv \max_i \sum_{j=1}^{n} \mid a_{ij} \mid \ .$$

Thus R is the "*maximum absolute row sum*". For $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$, we have

$$\begin{aligned} \frac{\| \mathbf{Ax} \|_\infty}{\| \mathbf{x} \|_\infty} &= \frac{\max_i \mid \sum_{j=1}^n a_{ij} x_j \mid}{\| \mathbf{x} \|_\infty} \\ &\leq \frac{\max_i \sum_{j=1}^n \mid a_{ij} \mid\mid x_j \mid}{\| \mathbf{x} \|_\infty} \\ &\leq \frac{\max_i \{ \sum_{j=1}^n \mid a_{ij} \mid \ \| \mathbf{x} \|_\infty \}}{\| \mathbf{x} \|_\infty} \ = \ R \ . \end{aligned}$$

Next we show that for any matrix $\mathbf{A}$ there always is a vector $\mathbf{y}$ for which

$$\frac{\| \mathbf{Ay} \|_\infty}{\| \mathbf{y} \|_\infty} \ = \ R \ .$$

Let k be the row of $\mathbf{A}$ for which $\sum_{j=1}^n \mid a_{kj} \mid$ is a maximum, *i.e.*,

$$\sum_{j=1}^n \mid a_{kj} \mid \ = \ R \ .$$

Take $\mathbf{y} \ = \ (y_1, y_2, \cdots, y_n)^T$ such that

$$y_j \ = \ \begin{cases} \ \ 1 \ \text{ if } \ a_{kj} \geq 0 \ , \\ \\ -1 \ \text{ if } \ a_{kj} < 0 \ . \end{cases}$$

Then

$$\frac{\| \mathbf{Ay} \|_\infty}{\| \mathbf{y} \|_\infty} \ = \ \| \mathbf{Ay} \|_\infty \ = \ \max_i \mid \sum_{j=1}^n a_{ij} y_j \mid \ = \ \sum_{j=1}^n \mid a_{kj} \mid \ = \ R \ .$$

Thus we have shown that

$$\| \mathbf{A} \|_\infty \text{ is equal to } \textit{the maximum absolute row sum.}$$

EXAMPLE : If

$$\mathbf{A} \ = \ \begin{pmatrix} 1 & 2 & -3 \\ 1 & 0 & 4 \\ -1 & 5 & 1 \end{pmatrix} ,$$

then

$$\| \mathbf{A} \|_\infty \ = \ \max\{6, 5, 7\} \ = \ 7 \ .$$

NOTE : In this example the vector $\mathbf{y}$ is given by $\mathbf{y} = (-1, 1, 1)^T$.

For this vector we have

$$\frac{\| \mathbf{Ay} \|_\infty}{\| \mathbf{y} \|_\infty} \ = \ 7 \ = \ \text{maximum absolute row sum.}$$

Similarly one can show that

$$\| \mathbf{A} \|_1 \equiv \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{Ax} \|_1}{\| \mathbf{x} \|_1} = \max_j \sum_{i=1}^{n} | a_{ij} |$$

$$= \textit{ the maximum absolute column sum.}$$

(**Check !**)

EXAMPLE : For the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 \\ 1 & 0 & 4 \\ -1 & 5 & 1 \end{pmatrix} ,$$

we have

$$\| \mathbf{A} \|_1 = \max\{3, 7, 8\} = 8 .$$

One can also show that

$$\| \mathbf{A} \|_2 \equiv \max_{\mathbf{x}\neq\mathbf{0}} \frac{\| \mathbf{Ax} \|_2}{\| \mathbf{x} \|_2} = \max_i \kappa_i(\mathbf{A}) ,$$

where the $\kappa_i(\mathbf{A})$ are defined to be

the square roots of the eigenvalues of the matrix $\mathbf{A}^T\mathbf{A}$.

(These eigenvalues are indeed nonnegative).

The quantities $\{\kappa_i(\mathbf{A})\}_{i=1}^n$ are called the *singular values* of the matrix $\mathbf{A}$.

EXAMPLE : If

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} ,$$

then

$$\mathbf{A}^T\mathbf{A} \ = \ \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \ = \ \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} .$$

The eigenvalues λ of $\mathbf{A}^T\mathbf{A}$ are obtained from

$$\det(\mathbf{A}^T\mathbf{A}-\lambda\mathbf{I}) = \det\begin{pmatrix} 1-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} = (1-\lambda)(2-\lambda)-1 \ = \ \lambda^2-3\lambda+1 = 0 ,$$

from which

$$\lambda_1 \ = \ \frac{3+\sqrt{5}}{2} \qquad \text{and} \qquad \lambda_2 \ = \ \frac{3-\sqrt{5}}{2} .$$

Thus we have

$$\| \ \mathbf{A} \ \|_2 \ = \ \sqrt{(3+\sqrt{5})/2} \ \cong \ 1.618 .$$

If $\mathbf{A}$ is invertible then we also have

$$\| \mathbf{A}^{-1} \|_2 \;=\; \frac{1}{\min_i \kappa_i(\mathbf{A})} \;.$$

Thus if we order the square roots of the eigenvalues of $\mathbf{A}^T\mathbf{A}$ as

$$\kappa_1 \;\geq\; \kappa_2 \;\geq\; \cdots \kappa_n \;\geq\; 0 \;,$$

then

$$\| \mathbf{A} \|_2 \;=\; \kappa_1, \qquad \text{and} \qquad \| \mathbf{A}^{-1} \|_2 \;=\; \frac{1}{\kappa_n} \;.$$

Thus in the previous example we have

$$\| \mathbf{A}^{\mathbf{-1}} \|_2 \;=\; \frac{1}{\sqrt{(3-\sqrt{5})/2}} \;\cong\; 1.618 \text{ (!)}$$

EXERCISES :

- Let $\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$. Compute $\| \mathbf{A} \|_2$.

- Let $\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$. Compute $\| \mathbf{A} \|_2$.

For a general n by n matrix $\mathbf{A}$:

- Prove that $\| \mathbf{A} \|_2 \ \leq \ \sqrt{n} \ \| \mathbf{A} \|_\infty$.

- Prove that $\| \mathbf{A} \|_1$ is equal to the maximum absolute column sum.

EXERCISES :

- Let $\mathbf{A}$ be any n by n matrix. For each of the following state whether it is true or false. If false then give a counter example.

$$\| \mathbf{A} \|_1 \quad \leq \quad \| \mathbf{A} \|_\infty \quad , \quad \| \mathbf{A} \|_\infty \quad \leq \quad \| \mathbf{A} \|_1 \quad .$$

- Give a *constructive proof* that for any square matrix $\mathbf{A}$

  there is a vector $\mathbf{x} \neq \mathbf{0}$ such that

$$\| \mathbf{Ax} \|_\infty \quad = \quad \| \mathbf{A} \|_\infty \quad \| \mathbf{x} \|_\infty \quad .$$

- Give a *constructive proof* that for any square matrix $\mathbf{A}$

  there is a vector $\mathbf{x} \neq \mathbf{0}$ such that

$$\| \mathbf{Ax} \|_1 \quad = \quad \| \mathbf{A} \|_1 \quad \| \mathbf{x} \|_1 \quad .$$

- Is there a vector $\mathbf{x}$ such that

$$\| \mathbf{Ax} \|_1 \quad > \quad \| \mathbf{A} \|_1 \quad \| \mathbf{x} \|_1 \quad ?$$

All matrix norms defined in terms of (induced by) a given vector norm as

$$\| \mathbf{A} \| \ = \ \max_{\mathbf{x}\neq\mathbf{0}} \frac{\| \mathbf{Ax} \|}{\| \mathbf{x} \|}$$

automatically satisfy

(i) $\quad \| \mathbf{A} \| \ \geq \ 0, \ \text{and} \ \| \mathbf{A} \| \ = \ 0 \ \text{only if} \ \mathbf{A} \ = \ \mathbf{O} \qquad \text{(zero matrix)} \ ,$

(ii) $\quad \| \alpha\mathbf{A} \| \ = \ | \, \alpha \, | \quad \| \mathbf{A} \|, \quad \forall \alpha \in \mathbb{R} \ ,$

(iii) $\quad \| \mathbf{A} + \mathbf{B} \| \ \leq \ \| \mathbf{A} \| + \| \mathbf{B} \| \ .$

Check : Properties (i) and (ii) !

PROOF of *(iii)* :

$$
\begin{aligned}
\| \mathbf{A} + \mathbf{B} \| \quad &= \quad \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| (\mathbf{A} + \mathbf{B})\mathbf{x} \|}{\| \mathbf{x} \|} \\
&= \quad \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x} \|}{\| \mathbf{x} \|} \\
&\leq \quad \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{A}\mathbf{x} \| + \| \mathbf{B}\mathbf{x} \|}{\| \mathbf{x} \|} \\
&\leq \quad \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{A}\mathbf{x} \|}{\| \mathbf{x} \|} + \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{B}\mathbf{x} \|}{\| \mathbf{x} \|} \\
&\equiv \quad \| \mathbf{A} \| + \| \mathbf{B} \| \ .
\end{aligned}
$$

QED !

In addition we have

$$(iv) \qquad \| \mathbf{AB} \| \ \leq \ \| \mathbf{A} \| \ \| \mathbf{B} \| \ .$$

PROOF of (iv) :

$$
\begin{aligned}
\| \mathbf{AB} \| \quad &= \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| (\mathbf{AB})\mathbf{x} \|}{\| \mathbf{x} \|} \\
&= \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{A}(\mathbf{Bx}) \|}{\| \mathbf{x} \|} \\
&\leq \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{A} \| \ \| \mathbf{Bx} \|}{\| \mathbf{x} \|} \\
&\leq \max_{\mathbf{x} \neq \mathbf{0}} \frac{\| \mathbf{A} \| \ \| \mathbf{B} \| \ \| \mathbf{x} \|}{\| \mathbf{x} \|} \ = \ \| \mathbf{A} \| \ \| \mathbf{B} \| \ . \qquad \text{QED !}
\end{aligned}
$$

EXERCISES : Let $\mathbf{A}$ and $\mathbf{B}$ be arbitrary n by n matrices.

- Is it true that
$$\| \mathbf{AB} \|_2 \quad = \quad \| \mathbf{A} \|_2 \quad \| \mathbf{B} \|_2 \quad ?$$
If false then give a counterexample.

- Is it true that
$$\| \mathbf{A} \|_1 \quad \| \mathbf{A}^{-1} \|_1 \quad = \ 1 \ \ ?$$
If false then give a counterexample.

- Let $\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$. Compute spr($\mathbf{A}$) , the "*spectral radius*" of A .

  (Here spr($\mathbf{A}$) is the absolute value of the largest eigenvalue of $\mathbf{A}$.)

  Explain why spr($\mathbf{A}$) is not a matrix norm.

The Banach Lemma.

Let $\mathbf{B}$ be an n by n matrix .

If in some induced matrix norm

$$\| \mathbf{B} \| \ < \ 1 \ ,$$

then

$$\mathbf{I} + \mathbf{B} \ \text{ is nonsingular}$$

and

$$\| \, (\mathbf{I} + \mathbf{B})^{-1} \, \| \ \leq \ \frac{1}{1 - \| \mathbf{B} \|} \ .$$

PROOF :

Suppose on the contrary that $\mathbf{I} + \mathbf{B}$ is singular.

Then

$$(\mathbf{I} + \mathbf{B})\ \mathbf{y} \ = \ \mathbf{0} \ ,$$

for some nonzero vector $\mathbf{y}$.

Hence

$$\mathbf{B}\ \mathbf{y} \ = \ -\,\mathbf{y} \ ,$$

and

$$1 \ = \ \frac{\|\ \mathbf{B}\ \mathbf{y}\ \|}{\|\ \mathbf{y}\ \|} \ \leq \ \frac{\|\ \mathbf{B}\ \| \quad \|\ \mathbf{y}\ \|}{\|\ \mathbf{y}\ \|} \ = \ \|\ \mathbf{B}\ \| \quad ,$$

which contradicts the assumption of the Lemma. (Why ?)

Hence $\mathbf{I} + \mathbf{B}$ is nonsingular.

We now have

$$(\mathbf{I}+\mathbf{B})\ (\mathbf{I}+\mathbf{B})^{-1} \ = \ \mathbf{I}\ ,$$

from which

$$(\mathbf{I}+\mathbf{B})^{-1} \ = \ \mathbf{I}-\mathbf{B}\ (\mathbf{I}+\mathbf{B})^{-1}\ .$$

Hence

$$\|\ (\mathbf{I}+\mathbf{B})^{-1}\ \| \ \leq \ \|\ \mathbf{I}\ \| + \|\ \mathbf{B}\ \| \ \ \|\ (\mathbf{I}+\mathbf{B})^{-1}\ \| \ .$$

Since $\|\ \mathbf{I}\ \|$ is always 1 in any induced matrix norm, we get

$$(1- \|\ \mathbf{B}\ \|)\ \ \|\ (\mathbf{I}+\mathbf{B})^{-1}\ \| \ \leq \ 1\ ,$$

from which

$$\|\ (\mathbf{I}+\mathbf{B})^{-1}\ \| \ \leq \ \frac{1}{1- \|\ \mathbf{B}\ \|}\ .$$

QED !

EXERCISES :

- Consider the n by n tridiagonal matrix $\mathbf{T}_n = \text{diag}[1,3,1]$.
  For example,

$$\mathbf{T}_4 \;=\; \begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 3 \end{pmatrix} \ .$$

  Use the Banach Lemma to show that $\mathbf{T}_n$ is invertible for all positive integers n. Also compute an upper bound on $\| \mathbf{T}_n^{-1} \|_\infty$.

- Let $\mathbf{A}_n$ be the n by n symmetric matrix

$$\mathbf{A}_n \;=\; \begin{pmatrix} 1 & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & \frac{1}{n} \\ \frac{1}{n} & 1 & \frac{1}{n} & \cdots & \frac{1}{n} & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & 1 & \cdots & \frac{1}{n} & \frac{1}{n} \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & 1 \end{pmatrix} \ .$$

  Show $\mathbf{A}_n$ for the case $n = 3$. Prove that $\mathbf{A}_n$ is invertible for any dimension n , and determine an upper bound on $\| \mathbf{A}_n^{-1} \|_\infty$.

EXERCISES :

- Let $\mathbf{A}_n$ be the n by n symmetric matrix

$$\mathbf{A}_n \;=\; \begin{pmatrix} 2n & 1 & 1 & \cdots & 1 & 1 \\ 1 & 2n & 1 & \cdots & 1 & 1 \\ 1 & 1 & 2n & \cdots & 1 & 1 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 1 & 1 & 1 & \cdots & 1 & 2n \end{pmatrix} .$$

  Show $\mathbf{A}_n$ for the cases $n = 2, 3$. Prove that $\mathbf{A}_n$ is invertible for any dimension n , and determine an upper bound on $\| \mathbf{A}_n^{-1} \|_\infty$.

- A square matrix is called *diagonally dominant* if in each row the absolute value of the diagonal element is greater than the sum of the absolute values of the off-diagonal elements. Use the Banach Lemma to prove that a diagonally dominant matrix is invertible.

- Derive an upper bound on $\| \mathbf{T}_n^{-1} \|_\infty$ for the n by n tridiagonal matrix $\mathbf{T}_n = \text{diag}[1\ ,\ 2 + 1/n\ ,\ 1]$.

The Gauss Elimination Method.

EXAMPLE : For given 4 by 4 matrix $\mathbf{A}$ and vector $\mathbf{f} \in \mathbb{R}^4$,

$$\mathbf{A} = \begin{pmatrix} 1 & -2 & -1 & 2 \\ 2 & 0 & 1 & 2 \\ 2 & 0 & 4 & 1 \\ 1 & 6 & 1 & 2 \end{pmatrix} , \quad \mathbf{f} = \begin{pmatrix} -2 \\ 5 \\ 7 \\ 16 \end{pmatrix} ,$$

solve

$$\mathbf{Ax} = \mathbf{f} ,$$

for

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} .$$

$$\begin{pmatrix} \mathbf{1} & -2 & -1 & 2 \\ 2 & 0 & 1 & 2 \\ 2 & 0 & 4 & 1 \\ 1 & 6 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -2 \\ 5 \\ 7 \\ 16 \end{pmatrix}$$

subtract 2 × row 1 from row 2
subtract 2 × row 1 from row 3
subtract 1 × row 1 from row 4

$$\begin{pmatrix} 1 & -2 & -1 & 2 \\ 0 & \mathbf{4} & 3 & -2 \\ 0 & 4 & 6 & -3 \\ 0 & 8 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -2 \\ 9 \\ 11 \\ 18 \end{pmatrix}$$

subtract 1 × row 2 from row 3
subtract 2 × row 2 from row 4

$$\begin{pmatrix} 1 & -2 & -1 & 2 \\ 0 & 4 & 3 & -2 \\ 0 & 0 & \mathbf{3} & -1 \\ 0 & 0 & -4 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -2 \\ 9 \\ 2 \\ 0 \end{pmatrix}$$

$subtract - \frac{4}{3} \times row\ 3\ from\ row\ 4$

$$\begin{pmatrix} 1 & -2 & -1 & 2 \\ 0 & 4 & 3 & -2 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & \mathbf{8/3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -2 \\ 9 \\ 2 \\ 8/3 \end{pmatrix}$$

The bold-face numbers at the top left of each submatrix are the *pivots* :

$$\begin{pmatrix} \mathbf{1} & -2 & -1 & 2 \\ 0 & \mathbf{4} & 3 & -2 \\ 0 & 0 & \mathbf{3} & -1 \\ 0 & 0 & 0 & \mathbf{8/3} \end{pmatrix}$$

The final matrix is an *upper triangular* matrix.

The upper triangular system

$$\begin{pmatrix} 1 & -2 & -1 & 2 \\ 0 & 4 & 3 & -2 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 8/3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -2 \\ 9 \\ 2 \\ 8/3 \end{pmatrix} ,$$

can be solved by *backsubstitution* :

$$\begin{aligned} x_4 &= (8/3)/(8/3) \;=\; 1 , \\ x_3 &= [2 - (-1)1]/3 \;=\; 1 , \\ x_2 &= [9 - (-2)1 - (3)1]/4 \;=\; 2 , \\ x_1 &= [-2 - (2)1 - (-1)1 - (-2)2]/1 \;=\; 1 . \end{aligned}$$

(Of course, actual computer computations use *floating point arithmetic* .)

Operation Count.

Using Gauss elimination for general n by n matrices, *counting multiplications and divisions only* (and treating these as equivalent).

(i) Triangularization (illustrated for $n = 4$) :

$$\begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \bullet \\ \bullet \\ \bullet \\ \bullet \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \circ & \star & \star & \star \\ \circ & \circ & \star & \star \\ \circ & \circ & \circ & \star \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \bullet \\ \star \\ \star \\ \star \end{pmatrix}$$

$$(n+1)(n-1) \ + \ n(n-2) \ + \ \cdots \ + \ (3)(1)$$

$$= \sum_{k=1}^{n-1} k(k+2) \ = \ \sum_{k=1}^{n-1} k^2 \ + \ 2\sum_{k=1}^{n-1} k$$

$$= \frac{(n-1)n(2n-1)}{6} \ + \ n(n-1) \ = \ \frac{n(n-1)(2n+5)}{6} \, . \qquad (\textbf{Check !})$$

(ii) Backsubstitution :

$$\begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \circ & \star & \star & \star \\ \circ & \circ & \star & \star \\ \circ & \circ & \circ & \star \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \bullet \\ \star \\ \star \\ \star \end{pmatrix}$$

$$1 \ + \ 2 \ + \ \cdots \ + \ n \ = \ \frac{n(n+1)}{2} \ .$$

Taking the total of triangularization and backsubstitution we obtain

$$\frac{n(n-1)(2n+5)}{6} + \frac{n(n+1)}{2} = \frac{n^3}{3} + n^2 - \frac{n}{3} \ . \qquad (\textbf{Check !})$$

EXAMPLES :

if $n \ = \ 10$, then the total is 430,

if $n \ = \ 100$, then the total is 343 430,

if $n \ = \ 1000$, then the total is 336 333 430.

For large values of n the *dominant term* in the total operation count is $n^3/3$.

Reconsider the Gauss elimination procedure for solving the system

$$\mathbf{A}\mathbf{x} \ = \ \mathbf{f} \ ,$$

given by

$$\begin{pmatrix} 1 & -2 & -1 & 2 \\ 2 & 0 & 1 & 2 \\ 2 & 0 & 4 & 1 \\ 1 & 6 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -2 \\ 5 \\ 7 \\ 16 \end{pmatrix} .$$

- In each step *retain* the equation that *cancels* the operation performed.

- This is done by storing the multipliers appropriately in the identity matrix.

$$\overset{\mathbf{I}}{\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}}\overset{\mathbf{A}}{\begin{pmatrix}1&-2&-1&2\\2&0&1&2\\2&0&4&1\\1&6&1&2\end{pmatrix}}\overset{\mathbf{x}}{\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}}=\overset{\mathbf{I}}{\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}}\overset{\mathbf{f}}{\begin{pmatrix}-2\\5\\7\\16\end{pmatrix}}$$

$$\begin{pmatrix}1&0&0&0\\2&1&0&0\\2&0&1&0\\1&0&0&1\end{pmatrix}\begin{pmatrix}1&-2&-1&2\\0&4&3&-2\\0&4&6&-3\\0&8&2&0\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1&0&0&0\\2&1&0&0\\2&0&1&0\\1&0&0&1\end{pmatrix}\begin{pmatrix}-2\\9\\11\\18\end{pmatrix}$$

$$\begin{pmatrix}1&0&0&0\\2&1&0&0\\2&1&1&0\\1&2&0&1\end{pmatrix}\begin{pmatrix}1&-2&-1&2\\0&4&3&-2\\0&0&3&-1\\0&0&-4&4\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1&0&0&0\\2&1&0&0\\2&1&1&0\\1&2&0&1\end{pmatrix}\begin{pmatrix}-2\\9\\2\\0\end{pmatrix}$$

$$\underset{\mathbf{L}}{\begin{pmatrix}1&0&0&0\\2&1&0&0\\2&1&1&0\\1&2&-\frac{4}{3}&1\end{pmatrix}}\underset{\mathbf{U}}{\begin{pmatrix}1&-2&-1&2\\0&4&3&-2\\0&0&3&-1\\0&0&0&\frac{8}{3}\end{pmatrix}}\underset{\mathbf{x}}{\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}}=\underset{\mathbf{L}}{\begin{pmatrix}1&0&0&0\\2&1&0&0\\2&1&1&0\\1&2&-\frac{4}{3}&1\end{pmatrix}}\underset{\mathbf{g}}{\begin{pmatrix}-2\\9\\2\\\frac{8}{3}\end{pmatrix}}$$

NOTE :

- Gauss elimination generates an *LU-decomposition* of $\mathbf{A}$:

$$\mathbf{A} \ = \ \mathbf{L}\ \mathbf{U} \ .$$

  $\mathbf{L}$ is *lower triangular* and $\mathbf{U}$ is *upper triangular*.

  The below-diagonal entries of $\mathbf{L}$ are the *multipliers*.

- In addition we have $\mathbf{Lg} \ = \ \mathbf{f}$.

  Furthermore, $\mathbf{LUx} \ = \ \mathbf{Lg}$, where $\mathbf{L}$ is nonsingular.

- Hence we also have $\mathbf{Ux} \ = \ \mathbf{g}$.

Using the LU-decomposition for multiple right hand sides.

Suppose we want to solve

$$\mathbf{A}\mathbf{x}^{(k)} \;=\; \mathbf{f}^{(k)} \ ,$$

with fixed $\mathbf{A}$, but for *multiple right hand side vectors*

$$\mathbf{f}^{(k)} \ , \quad k = 1, 2, \cdots, m \ .$$

Algorithm :

(i) Determine the **LU**-decomposition of $\mathbf{A}$.

(ii) Solve $\left\{ \begin{array}{l} \mathbf{L}\mathbf{g}^{(k)} \;=\; \mathbf{f}^{(k)} \ , \\ \\ \mathbf{U}\mathbf{x}^{(k)} = \mathbf{g}^{(k)} \ , \end{array} \right. \qquad\qquad k = 1, 2, \cdots, m \ .$

Note that the decomposition need only be computed once.

Operation Count.

Multiplications and divisions for an n by n system with m right hand sides :

Step (i) (**LU**-decomposition) :

$$\begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \circ & \star & \star & \star \\ \circ & \circ & \star & \star \\ \circ & \circ & \circ & \star \end{pmatrix}$$

$$n\ (n-1) \ + \ (n-1)\ (n-2) \ + \ \cdots + (2)\ (1)$$

$$= \sum_{k=1}^{n-1} k\ (k+1) \ = \ \sum_{k=1}^{n-1} k^2 \ + \ \sum_{k=1}^{n-1} k$$

$$= \frac{(n-1)\ n\ (2n-1)}{6} + \frac{n\ (n-1)}{2} = \frac{n^3}{3} - \frac{n}{3} \ . \qquad (\textbf{Check !})$$

$$\overset{\mathbf{L}}{\begin{pmatrix} 1 & \circ & \circ & \circ \\ \star & 1 & \circ & \circ \\ \star & \star & 1 & \circ \\ \star & \star & \star & 1 \end{pmatrix}} \overset{\mathbf{g}}{\begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{pmatrix}} = \overset{\mathbf{f}}{\begin{pmatrix} \bullet \\ \bullet \\ \bullet \\ \bullet \end{pmatrix}} \quad , \quad \overset{\mathbf{U}}{\begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \circ & \star & \star & \star \\ \circ & \circ & \star & \star \\ \circ & \circ & \circ & \star \end{pmatrix}} \overset{\mathbf{x}}{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}} = \overset{\mathbf{g}}{\begin{pmatrix} \bullet \\ \bullet \\ \bullet \\ \bullet \end{pmatrix}}$$

Step (ii) (Backsubstitution) :

$$\begin{cases} \mathbf{Lg}^{(k)} = \mathbf{f}^{(k)} \ : \quad m(1+2+\cdots+(n-1)) \ , \\ \\ \mathbf{Ux}^{(k)} = \mathbf{g}^{(k)} \ : \quad m(1+2+\cdots+n) \ . \end{cases} \qquad k = 1, 2, \cdots, m \ .$$

Total Step (ii) : mn^2 (**Check !**).

The total of Steps (i) and (ii) is therefore

$$\frac{n^3}{3} + mn^2 - \frac{n}{3} \ .$$

NOTE : For m small and n large the *dominant term* remains $n^3/3$.

Tridiagonal systems.

For *tridiagonal systems* of linear equations

$$\begin{pmatrix} b_1 & c_1 & & & & & \\ a_2 & b_2 & c_2 & & & & \\ & a_3 & b_3 & c_3 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} & \\ & & & & a_n & b_n & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \cdot \\ f_{n-1} \\ f_n \end{pmatrix} ,$$

Gauss elimination reduces to this *simple algorithm* :

$$\beta_1 = b_1 \ , \quad g_1 = f_1 \ ,$$

$$\left.\begin{aligned} \gamma_k &= a_k/\beta_{k-1} \ , \\ \beta_k &= b_k - \gamma_k c_{k-1} \ , \\ g_k &= f_k - \gamma_k g_{k-1} \ , \end{aligned}\right\} \qquad k = 2, 3, \cdots, n \ .$$

This transform the tridiagonal system into the *upper-triangular* form

$$\begin{pmatrix} \beta_1 & c_1 & & & & \\ & \beta_2 & c_2 & & & \\ & & \beta_3 & c_3 & & \\ & & & \cdot & \cdot & \\ & & & & \beta_{n-1} & c_{n-1} \\ & & & & & \beta_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ \cdot \\ g_{n-1} \\ g_n \end{pmatrix} .$$

The *backsubstitution* algorithm now becomes

$$x_n = \frac{g_n}{\beta_n} ,$$

$$x_k = \frac{g_k - c_k x_{k+1}}{\beta_k}, \qquad k = n-1 , n-2 , \cdots , 1 .$$

The resulting **LU**-*decomposition* is

$$\begin{pmatrix} 1 & & & & & \\ \gamma_2 & 1 & & & & \\ & \gamma_3 & 1 & & & \\ & & \cdot & \cdot & & \\ & & & \gamma_{n-1} & 1 & \\ & & & & \gamma_n & 1 \end{pmatrix} \begin{pmatrix} \beta_1 & c_1 & & & & \\ & \beta_2 & c_2 & & & \\ & & \beta_3 & c_3 & & \\ & & & \cdot & \cdot & \\ & & & & \beta_{n-1} & c_{n-1} \\ & & & & & \beta_n \end{pmatrix}$$

The total number of *multiplications and divisions* is $5n - 4$. (**Check !**)

Inverses.

The *inverse* of a n by n matrix $\mathbf{A}$ is defined to be a matrix $\mathbf{A}^{-1}$ such that

$$\mathbf{A}\ (\mathbf{A}^{-1})\ =\ (\mathbf{A}^{-1})\ \mathbf{A}\ =\ \mathbf{I}\ ,$$

where

$$\mathbf{I}\ \equiv\ \begin{pmatrix} 1 & 0 & \cdot & 0 \\ 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & 1 \end{pmatrix} \qquad \text{(the } \textit{identity matrix}\text{)}\ .$$

$\mathbf{A}$ is *invertible* if and only if

$$\det \mathbf{A}\ \neq\ 0\ .$$

The inverse is then unique.

To compute $\mathbf{A}^{-1}$ we can solve

$$\mathbf{A}\ (\mathbf{A}^{-1}) \quad = \quad \mathbf{I}\ ,$$

which is of the form

$$\overset{\mathbf{A}}{\begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix}} \overset{\mathbf{A}^{-1}}{\begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{pmatrix}} \quad = \quad \overset{\mathbf{I}}{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}} \ .$$

This corresponds to solving a linear system with n right hand sides.

Using the earlier formula, the number of multiplications and divisions is

$$\frac{n^3}{3} \ + \ (n)\ n^2 \ - \ \frac{n}{3} \quad = \quad \frac{4n^3}{3} \ - \ \frac{n}{3}\ .$$

But we can *omit some operations* , because the right hand side vectors, *i.e.*, the columns of $\mathbf{I}$, are very special.

In particular, multiplications by 0 or 1 can be omitted.

The total number of multiplications that can be omitted is seen to be

$$\begin{aligned}(n)\ (n-1)\ \ +\ \ (n-1)\ (n-2) + \cdots + (2)\ (1)\ &= \sum_{k=1}^{n-1} k\ (k+1) \\ &= \sum_{k=1}^{n-1} k^2\ +\ \sum_{k=1}^{n-1} k\ =\ \frac{(n-1)\ n\ (2n-1)}{6} + \frac{n\ (n-1)}{2} \\ &= \frac{n^3}{3} - \frac{n}{3}. \qquad (\textbf{Check !})\end{aligned}$$

Thus to find $\mathbf{A}^{-1}$ we need only

$$(\frac{4n^3}{3} - \frac{n}{3}) - (\frac{n^3}{3} - \frac{n}{3}) \ =\ n^3 \quad \text{operations .}$$

NOTE :

To solve a n by n linear system

$$\mathbf{A}\mathbf{x}^{(k)} \ = \ \mathbf{f}^{(k)} \ ,$$

with m right hand sides, takes

$$\frac{n^3}{3} \ + \ mn^2 \ - \ \frac{n}{3} \qquad \text{operations} \ ,$$

as derived earlier for the **LU**-decomposition algorithm.

One can also find the solution vectors by computing $\mathbf{A}^{-1}$, and setting

$$\mathbf{x}^{(k)} \ = \ \mathbf{A}^{-1} \ \mathbf{f}^{(k)} \ .$$

But this takes

$$n^3 \ + \ mn^2 \qquad \text{operations} \ ,$$

which is *always less efficient* , no matter how big n and m are !

EXERCISES :

- Compute the **LU**-decomposition of the tridiagonal matrix

$$\mathbf{T}_4 \ = \ \begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 3 \end{pmatrix} \ .$$

  Let $\mathbf{f} = (4,5,5,4)^T$. Using the matrices **L** and **U**, solve $\mathbf{Lg} = \mathbf{f}$, followed by $\mathbf{Ux} = \mathbf{g}$. After having computed the vector **x** in this way, check your answer by verifying that **x** satisfies the equation $\mathbf{T}_4\mathbf{x} = \mathbf{f}$.

- How many multiplications and divisions are needed to compute the **LU**-decomposition of the specific tridiagonal matrix $\mathbf{T}_n = \text{diag}[1,3,1]$ as a function of n ? Make sure not to count unnecessary operations.

- If the **LU**-decomposition of this n by n tridiagonal matrix takes 0.01 second on a given computer if $n = 10^5$, then how much time could it take if $n = 10^9$?

EXERCISES :

- Suppose the **LU** decomposition of a matrix **A** is given by

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} .$$

  Using *only* **L** , **U** , and **f** , *i.e.*, without explicitly determining **A** , solve $\mathbf{Ax} = \mathbf{f}$, when $\mathbf{f} = (\,6, 9, 10\,)^T$.

- Suppose that solving a general n by n linear system of the form $\mathbf{Ax} = \mathbf{f}$ by Gauss elimination takes 10 seconds on a given computer if $n = 1000$.

  Estimate how much time it will take to solve a 1000 by 1000 system $\mathbf{Lg} = \mathbf{f}$, followed by solving $\mathbf{Ux} = \mathbf{g}$, where **L** is lower triangular with 1's along its main diagonal, and where **U** is upper triangular?

  Thus you may assume that **L** and **U** have already been computed.

EXERCISES :

- Suppose that multiplying two general n by n matrices takes 3 seconds on a given computer, if $n = 1000$.

  Estimate how much time it will take to compute the **LU**-decomposition of such a matrix.

- Suppose that solving a general system of linear equations of dimension 1000 takes 10 seconds on a given computer.

  Estimate how much time it will take to solve a *tridiagonal* linear system of dimension 10^6 on that computer.

- • How many *divisions* are needed for **LU**-decomposition of an n by n *tridiagonal* matrix (not counting multiplications and additions)?

- • How many *divisions* are needed for **LU**-decomposition of an n by n *general* matrix (not counting multiplications and additions)?

Practical Considerations.

- *Memory reduction.*

In an implementation of the **LU** decomposition algorithm, the multipliers can be stored in the lower triangular part of the original matrix **A**.

In the earlier example, with

$$A \;=\; \begin{pmatrix} 1 & -2 & -1 & 2 \\ 2 & 0 & 1 & 2 \\ 2 & 0 & 4 & 1 \\ 1 & 6 & 1 & 2 \end{pmatrix} ,$$

this function would return the matrix :

$$\begin{pmatrix} \mathbf{1} & -2 & -1 & 2 \\ 2 & \mathbf{4} & 3 & -2 \\ 2 & 1 & \mathbf{3} & -1 \\ 1 & 2 & -4/3 & \mathbf{8/3} \end{pmatrix} .$$

- *Row interchanges.*

Gauss elimination will *fail* for the matrix

$$\begin{pmatrix} 0 & 2 & 1 \\ 1 & 1 & 2 \\ 2 & 3 & -1 \end{pmatrix} ,$$

since the first pivot is zero.

A *division by zero* will occur when the first multiplier is computed !

The *remedy* is to interchange rows to get

$$\begin{pmatrix} 1 & 1 & 2 \\ 0 & 2 & 1 \\ 2 & 3 & -1 \end{pmatrix}$$

Several such *interchanges* may be needed during Gauss elimination.

- *Loss of accuracy.*

More generally, loss of accuracy may occur when there are *large multipliers* .

EXAMPLE : Solve

$$\begin{pmatrix} 0.0000001 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} ,$$

on a "six-decimal-digit computer".

NOTE :

- The solution is $x_1 \cong 1$, $x_2 \cong 1$.

- The multiplier is 10,000,000 .

A "*Six-decimal-digit computer*" :

Assume all arithmetic operations are performed to infinite precision,

but then *truncated* to six decimal digits (plus exponent).

Thus, for example,

$$-100/3$$

is stored as

$$-3.33333 \ 10^1$$

$$\begin{pmatrix} 0.0000001 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

(a) Elimination gives :

$$\begin{pmatrix} 1.00000\mathrm{E}-07 & 1.00000\mathrm{E}+00 \\ 0 & -999999\mathrm{E}+01 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1.00000\mathrm{E}+00 \\ -999999\mathrm{E}+01 \end{pmatrix} .$$

(b) Backsubstitution gives :

$$x_2 \ = \ 1.00000\mathrm{E}+00 \ , \qquad x_1 \ = \ 0.00000\mathrm{E}+00 \ .$$

Clearly this result is *very bad* !

Again, the *remedy* is to interchange rows :

$$\begin{pmatrix} 1 & 1 \\ 0.0000001 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} .$$

Now the multiplier is only $1.00000\text{E} - 07$, and we obtain :

(a) Elimination :

$$\begin{pmatrix} 1.000000\text{E}+00 & 1.000000\text{E}+00 \\ 0 & .999999\text{E}+00 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2.00000\text{E}+00 \\ .999999\text{E}+00 \end{pmatrix} .$$

(b) Backsubstitution : $x_2 = 1.00000\text{E}+00, \;\; x_1 = 1.0000\text{E}+00$.

This solution is *accurate* !

Gauss Elimination with pivoting.

A variant of the Gauss elimination procedure that avoids loss of accuracy due to large multipliers is called

“Gauss elimination with *partial pivoting* (or *row pivoting*)”.

Here *rows are interchanged* each time a pivot element is sought, so that the pivot is as large as possible in absolute value.

(In practice *pointers* to rows are interchanged.)

All multipliers will then be less than 1 in magnitude.

EXAMPLE :

$$\begin{pmatrix} 2 & 2 & 1 \\ 1 & 0 & 1 \\ 4 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \\ 7 \end{pmatrix} \qquad \textit{interchange row 1 and 3}$$

$$\begin{pmatrix} 4 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2 \\ 5 \end{pmatrix} \qquad \begin{array}{l} \textit{subtract } \frac{1}{4} \textit{ row 1 from row 2} \\ \textit{subtract } \frac{2}{4} \textit{ row 1 from row 3} \end{array}$$

$$\begin{pmatrix} 4 & 1 & 2 \\ 0 & -1/4 & 1/2 \\ 0 & 3/2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 1/4 \\ 3/2 \end{pmatrix} \qquad \textit{interchange row 2 and 3}$$

$$\begin{pmatrix} 4 & 1 & 2 \\ 0 & 3/2 & 0 \\ 0 & -1/4 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 3/2 \\ 1/4 \end{pmatrix} \qquad \textit{subtract } \frac{-1}{6} \textit{ row 2 from row 3}$$

$$\begin{pmatrix} 4 & 1 & 2 \\ 0 & 3/2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 3/2 \\ 1/2 \end{pmatrix} \qquad \begin{array}{l} \textit{backsubstitution} \ : \ x_1 = 1 \\ \textit{backsubstitution} \ : \ x_2 = 1 \\ \textit{backsubstitution} \ : \ x_3 = 1 \end{array}$$

EXERCISES :

- Use Gauss elimination with *row pivoting* to solve

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 8 & 11 \\ 3 & 22 & 35 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 10 \end{pmatrix} .$$

  The solution is $\mathbf{x} = (-1\ ,-1\ ,\ 1)$. If done honestly (with pivoting !) then all multipliers will be less than 1 in magnitude.

- Suppose that Gauss elimination with *row pivoting* is used to solve

$$\begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 8 \\ 12 \\ 11 \end{pmatrix} .$$

  Are any rows actually interchanged?

  Can you also answer this question for *general* $\mathbf{T}_n = \text{diag}[1, 2, 1]$?

Error Analysis.

Suppose we want to solve

$$\mathbf{Ax} \ = \ \mathbf{f},$$

where the n by n matrix $\mathbf{A}$ is nonsingular (*i.e.*, invertible).

Suppose a small *error* is made in the right hand side, *i.e.*, instead we solve

$$\mathbf{Ay} \ = \ \mathbf{f} \ + \ \mathbf{r}.$$

What will be the error $\| \mathbf{y} - \mathbf{x} \|$ in the solution ?

From

$$\mathbf{A}\mathbf{y} \;=\; \mathbf{f} \;+\; \mathbf{r} \;,$$

subtract

$$\mathbf{A}\mathbf{x} \;=\; \mathbf{f} \;,$$

to get

$$\mathbf{A}(\mathbf{y}-\mathbf{x}) \;=\; \mathbf{r} \;.$$

Thus

$$\mathbf{y}-\mathbf{x} \;=\; \mathbf{A}^{-1}\mathbf{r} \;,$$

so that

$$\|\,\mathbf{y}-\mathbf{x}\,\| \;=\; \|\,\mathbf{A}^{-1}\mathbf{r}\,\| \;\leq\; \|\,\mathbf{A}^{-1}\,\| \;\; \|\,\mathbf{r}\,\| \;.$$

Hence if $\|\,\mathbf{A}^{-1}\,\|$ *is large* then a *small* perturbation $\mathbf{r}$ in $\mathbf{f}$ may lead to a *large* change in $\mathbf{x}$.

EXAMPLE :

$$\begin{pmatrix} 1 & -1.001 \\ 2.001 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \mathbf{2.001} \\ \mathbf{4.001} \end{pmatrix} ,$$

has *exact solution*

$$x_1 = 1 , \qquad x_2 = -1 .$$

Suppose instead we solve

$$\begin{pmatrix} 1 & -1.001 \\ 2.001 & -2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mathbf{2.000} \\ \mathbf{4.002} \end{pmatrix} .$$

The *exact solution* of this system is

$$y_1 = 2 , \qquad y_2 = 0 .$$

Note that the small change in the right hand side has norm

$$\| \mathbf{r} \|_\infty \quad = \quad 0.001 \ .$$

Also note that the change in the solution is much larger , namely,

$$\| \mathbf{x} - \mathbf{y} \|_\infty \quad = \quad 1 \ .$$

In this example

$$\mathbf{A}^{-1} \cong \begin{pmatrix} -666.44 & 333.55 \\ & \\ -666.77 & 333.22 \end{pmatrix} \ .$$

Hence

$$\| \mathbf{A}^{-1} \|_\infty \quad \cong \quad 1000 \ , \qquad \text{whereas} \qquad \| \mathbf{A} \|_\infty \quad \cong \quad 4 \ .$$

Errors always occur in floating point computations due to finite word length.

For example, on a “six digit computer” $\frac{1}{3}$ is represented by $3.33333\ 10^{-1}$.

Such errors occur in both right hand side and matrix.

Suppose we want to solve

$$\mathbf{A}\mathbf{x} \;=\; \mathbf{f} \;,$$

but instead solve the perturbed system

$$(\mathbf{A}+\mathbf{E})\mathbf{y} \;=\; \mathbf{f}+\mathbf{r} \;.$$

Here the perturbation $\mathbf{E}$ is also a n by n matrix.

THEOREM : Consider

$$\mathbf{A}\mathbf{x} \ = \ \mathbf{f} \ , \qquad \text{and} \qquad (\mathbf{A}+\mathbf{E})\mathbf{y} \ = \ \mathbf{f}+\mathbf{r} \ .$$

Assume that

$\mathbf{A}$ is nonsingular, and that $\parallel \mathbf{E} \parallel < \dfrac{1}{\parallel \mathbf{A}^{-1} \parallel}$, *i.e.* , $\parallel \mathbf{A}^{-1} \parallel \ \parallel \mathbf{E} \parallel \ < \ 1$.

Then

$$\frac{\parallel \mathbf{y}-\mathbf{x} \parallel}{\parallel \mathbf{x} \parallel} \ \leq \ \frac{\text{cond}(\mathbf{A})}{1- \parallel \mathbf{A}^{-1} \parallel \parallel \mathbf{E} \parallel}\Big(\frac{\parallel \mathbf{r} \parallel}{\parallel \mathbf{f} \parallel} + \frac{\parallel \mathbf{E} \parallel}{\parallel \mathbf{A} \parallel}\Big) \ ,$$

where

$$\text{cond}(\mathbf{A}) \ \equiv \ \parallel \mathbf{A}^{-1} \parallel \parallel \mathbf{A} \parallel \ ,$$

is called the *condition number* of $\mathbf{A}$.

PROOF :

First write

$$\mathbf{A} + \mathbf{E} \ = \ \mathbf{A} \ (\mathbf{I} + \mathbf{A}^{-1}\mathbf{E}) \ .$$

Now, using the assumption,

$$\| \ \mathbf{A}^{-1} \ \mathbf{E} \ \| \ \ \leq \ \ \| \ \mathbf{A}^{-1} \ \| \ \ \ \| \ \mathbf{E} \ \| \ \ < \ 1 \ ,$$

it follows from by the Banach Lemma that

$(\mathbf{I} + \mathbf{A}^{-1}\mathbf{E})$ is nonsingular and

$$\| \ (\mathbf{I} + \mathbf{A}^{-1}\mathbf{E})^{-1} \ \| \ \ \leq \ \frac{1}{1- \ \| \ \mathbf{A}^{-1}\mathbf{E} \ \|} \ \leq \ \frac{1}{1- \ \| \ \mathbf{A}^{-1} \ \|\| \ \mathbf{E} \ \|} \ .$$

Next

$$(\mathbf{A}+\mathbf{E})\ \mathbf{y}\ =\ \mathbf{f}+\mathbf{r}\ ,$$

implies

$$(\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})\ \mathbf{y}\ =\ \mathbf{A}^{-1}\ (\mathbf{f}+\mathbf{r})\ ,$$

so that

$$\mathbf{y}\ =\ (\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})^{-1}\ \mathbf{A}^{-1}\ (\mathbf{f}+\mathbf{r})\ .$$

Then

$$\begin{aligned}\mathbf{y}-\mathbf{x}\ &=\ (\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})^{-1}\ \Big(\mathbf{A}^{-1}(\mathbf{f}+\mathbf{r})-(\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})\mathbf{x}\Big)\\ &=\ (\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})^{-1}\ \Big(\mathbf{x}+\mathbf{A}^{-1}\mathbf{r}-\mathbf{x}-\mathbf{A}^{-1}\mathbf{E}\mathbf{x}\Big)\\ &=\ (\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})^{-1}\ \mathbf{A}^{-1}\ (\mathbf{r}-\mathbf{E}\mathbf{x})\ .\end{aligned}$$

Finally,

$$
\begin{aligned}
\frac{\| \mathbf{y}-\mathbf{x} \|}{\| \mathbf{x} \|} &\leq \frac{\| (\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})^{-1} \| \quad \| \mathbf{A}^{-1} \| \ \Big(\| \mathbf{r} \| + \| \mathbf{E} \| \| \mathbf{x} \| \Big)}{\| \mathbf{x} \|} \\
&\leq \frac{\| \mathbf{A}^{-1} \|}{1- \| \mathbf{A}^{-1} \| \quad \| \mathbf{E} \|} \Big(\frac{\| \mathbf{r} \|}{\| \mathbf{x} \|} + \| \mathbf{E} \| \Big) \\
&= \frac{\| \mathbf{A}^{-1} \| \quad \| \mathbf{A} \|}{1- \| \mathbf{A}^{-1} \| \quad \| \mathbf{E} \|} \Big(\frac{\| \mathbf{r} \|}{\| \mathbf{A} \| \quad \| \mathbf{x} \|} + \frac{\| \mathbf{E} \|}{\| \mathbf{A} \|} \Big) \\
&\leq \frac{\| \mathbf{A}^{-1} \| \quad \| \mathbf{A} \|}{1- \| \mathbf{A}^{-1} \| \quad \| \mathbf{E} \|} \ \Big(\frac{\| \mathbf{r} \|}{\| \mathbf{f} \|} + \frac{\| \mathbf{E} \|}{\| \mathbf{A} \|} \Big) \ .
\end{aligned}
$$

The last step uses the fact that

$$\| \mathbf{f} \| \ = \ \| \mathbf{Ax} \| \ \leq \ \| \mathbf{A} \| \| \mathbf{x} \| \ . \qquad \textbf{QED !}$$

From the above theorem we can conclude that :

$$\text{If } \operatorname{cond}(\mathbf{A}) \text{ is large, then the relative error } \frac{\| \mathbf{y} - \mathbf{x} \|}{\| \mathbf{x} \|} \text{ can be large .}$$

Note, however, that cond($\mathbf{A}$) is never less than 1 because

$$1 \ = \ \| \mathbf{I} \| \ = \ \| \mathbf{A}^{-1}\mathbf{A} \| \ \leq \ \| \mathbf{A}^{-1} \| \ \| \mathbf{A} \| \ \equiv \ \operatorname{cond}(\mathbf{A}) \ ,$$

in any induced matrix norm.

A matrix having a large condition number is said to be *ill-conditioned*.

EXAMPLE : The 2 by 2 matrix

$$\begin{pmatrix} 1 & -1.001 \\ 2.001 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} ,$$

from the earlier example, with inverse

$$\mathbf{A}^{-1} \cong \begin{pmatrix} -666.44 & 333.55 \\ -666.77 & 333.22 \end{pmatrix} ,$$

has condition number

$$\text{cond}(\mathbf{A}) \ \cong \ (4)(1000) \ = \ 4000 \ ,$$

in the matrix infinity norm.

We may say that this matrix is ill-conditioned.

Solving ill-conditioned systems numerically, if they can be solved at all on a given computer, normally requires pivoting.

However, systems that need pivoting need not be ill-conditioned .

Reconsider, for example, the 2 by 2 matrix

$$\mathbf{A} = \begin{pmatrix} .0000001 & 1 \\ 1 & 1 \end{pmatrix} , \qquad \text{with} \quad \mathbf{A}^{-1} \cong \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix} \quad (\textbf{Check !}) \ ,$$

for which the condition number is approximately 4 using the infinity norm.

But solving a linear system with the above **A** as coefficient matrix requires pivoting, at least on a six (decimal) digit computer.

EXERCISE :

- Let

$$0 \ = \ t_0 \ < \ t_1 \ < \ t_2 \ < \ \cdots \ < \ t_n \ = \ 1 \ ,$$

and let

$$h_i \ = \ t_i - t_{i-1} \ , \qquad i = 1, 2, \ \cdots \ , n \ .$$

The following tridiagonal matrix arises in *cubic spline interpolation*

(to be discussed later) :

$$\mathbf{S}_{n-1} = \begin{pmatrix} 2(h_1 + h_2) & h_2 & & & \\ h_2 & 2(h_2 + h_3) & h_3 & & \\ & h_3 & 2(h_3 + h_4) & h_4 & \\ & & \cdot & \cdot & \cdot \\ & & & h_{n-1} & 2(h_{n-1} + h_n) \end{pmatrix} .$$

Prove that $\mathbf{S}_{n-1}$ is invertible and determine a bound on cond$(\mathbf{S}_{n-1})$.

Are there situations where $\mathbf{S}_{n-1}$ could be ill-conditioned?

EXERCISES :

- Consider the n by n matrix :

$$\mathbf{C}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdots & 1 \end{pmatrix} ,$$

  and let $\mathbf{I}_n$ be the n by n identity matrix.

  For what ϵ does the Banach Lemma guarantee $\mathbf{I}_n + \epsilon\mathbf{C}_n$ is invertible?

  Also determine a bound on $\text{cond}(\mathbf{I}_n + \epsilon\mathbf{C}_n)$ in this case.

- Use the Banach Lemma to prove that the five-diagonal matrix

$$\mathbf{T}_n \;=\; \text{diag}[1, 1, 5, 1, 1] \ ,$$

  is invertible for all $n \geq 1$.

  Derive an upper bound on $\text{cond}(\mathbf{T}_n)$ using the matrix infinity norm.

EXERCISES :

For each of the following statements, state whether it is true or false. If true then explain why; if false then give a counterexample.

- A condition number of 10^6 is large.
- All large matrices are ill-conditioned.
- All ill-conditioned matrices have small determinants.
- Only ill-conditioned matrices require pivoting.
- If pivoting is needed then the matrix is ill-conditioned.
- The condition number of a matrix is never smaller than 1.
- Tridiagonal matrices are never ill-conditioned.

EXERCISES :

For each of the following statements about matrices, say whether it is true or false. Explain your answer.

- Two n by n matrices can be multiplied using n^2 multiplications.

- **LU**-decomposition of the n by n tridiagonal matrix $\text{diag}[1, 2, 1]$ can be done using only $n - 1$ divisions and zero multiplications.

- **LU**-decomposition of a general n by n tridiagonal matrix requires $2n-2$ multiplications and divisions.

- The n by n tridiagonal matrix $\mathbf{T}_n = \text{diag}[1, 2 + 1/n, 1]$ is nonsingular for any positive integer n.

EXERCISES :

For each of the following statements about matrices, say whether it is true or false. Explain your answer.

- For large n, the **LU**-decomposition of a general n by n matrix requires approximately $n^3/3$ multiplications and divisions.

- The inverse of a general, nonsingular n by n matrix can be computed using n^3 multiplications and divisions.

- If **D** is a diagonal matrix (*i.e.*, its entries d_{ij} are zero if $i \neq j$), then

$$\| \mathbf{D} \|_1 \quad = \quad \| \mathbf{D} \|_2 \quad = \quad \| \mathbf{D} \|_\infty \quad .$$

- If $\| \mathbf{A}^{-1} \|$ is large then cond$(\mathbf{A})$ is large.

THE NUMERICAL SOLUTION OF NONLINEAR EQUATIONS

Introduction.

For a system of n *linear* equations in n unknowns

$$\mathbf{A}\mathbf{x} \ = \ \mathbf{f} \ ,$$

where $\mathbf{x}, \mathbf{f} \in \mathbb{R}^n$, and $\mathbf{A}$ an n by n matrix , we have these possibilities :

(i) $\mathbf{A}$ is *nonsingular* : In this case there is a unique solution.

(ii) $\mathbf{A}$ is *singular* : There are no solutions or infinitely many. (Examples?)

Usually only case (i) is of interest.

The solution can be computed in a finite number of steps by Gauss Elimination (with pivoting, if necessary).

We can write a *system* of n *nonlinear* equations in n unknowns as

$$\mathbf{G}(\mathbf{x}) \ = \ \mathbf{0} \ ,$$

where

$$\mathbf{x} \ , \ \mathbf{0} \ \in \ \mathbb{R}^n \ ,$$

$$\mathbf{x} \ = \ (x_1, x_2, \cdots, x_n)^T \ ,$$

and where $\mathbf{G}$ is a vector-valued function of $\mathbf{x}$ having n component functions,

$$\mathbf{G}(\mathbf{x}) \ = \ (g_1(\mathbf{x}) \ , \ g_2(\mathbf{x}) \ , \ \cdots \ , \ g_n(\mathbf{x}))^T \ .$$

EXAMPLES : (of possible situations) :

- $x^2 - 1 = 0$ has two solutions : $x = 1$, $x = -1$.

- $x^2 + 1 = 0$ has two solutions : $x = i$, $x = -i$.

- $e^{-x} - \sin(x) = 0$ has a countably infinite number of solutions.

- The system $(2x_1 - x_2 = 0 \ , \ x_1^3 + x_1 - x_2 = 0)$ has three solution pairs, namely, $(x_1, x_2) \ = \ (0,0) \ , \ (1,2) \ , \ (-1,-2)$.

- The system $(x_1 x_2 - 1 = 0 \ , \ x_1 x_2 - 2 = 0)$ has no solutions.

- The system $(e^{x_1 - x_2} - 1 = 0 \ , \ x_1 - x_2 = 0)$ has a *continuum* of solutions.

NOTE :

For nonlinear equations :

- There can be $0,\ 1,\ 2,\ 3,\ 4,\ \cdots\ \infty$ solutions.

- A solution can usually not be computed in a finite number of steps.

- Instead, *iterative methods* will be used.

- We will not consider the case of a continuum of solutions.

Some Methods for Scalar Nonlinear Equations.

Consider the *scalar* equation (one equation, one unknown)

$$g(x) \ = \ 0 \ ,$$

and let x^* denote a *solution* (or *zero* , or *root*) of this equation.

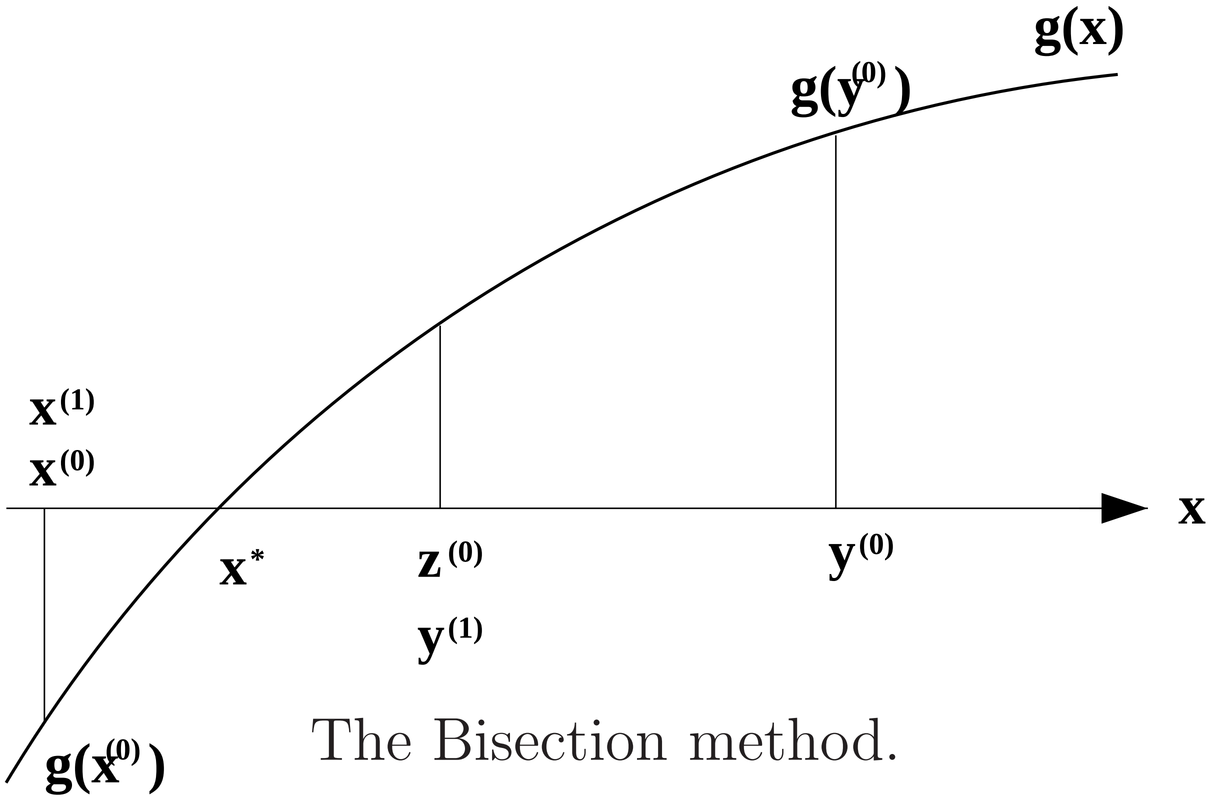
The Bisection Method.

This method requires two initial points :

$$x^{(0)} \ , \ y^{(0)} \qquad \text{with} \qquad g(x^{(0)}) < 0 \quad \text{and} \quad g(y^{(0)}) > 0 \ .$$

Algorithm : For $k = 0, \ 1, \ 2, \ \cdots$:

- Set $z^{(k)} \ = \ \frac{1}{2}(x^{(k)} + y^{(k)})$,
- If $g(z^{(k)}) < 0$ set $x^{(k+1)} = z^{(k)}$, $y^{(k+1)} = y^{(k)}$,
- If $g(z^{(k)}) > 0$ set $x^{(k+1)} = x^{(k)}$, $y^{(k+1)} = z^{(k)}$.



The Bisection method.

The bisection method works if $g(x)$ is continuous in the interval $[x^{(0)}, y^{(0)}]$.

In fact we have

$$| \, x^{(k)} - x^* \, | \quad \leq \quad \frac{1}{2^k} \, | \, x^{(0)} - y^{(0)} \, | \quad .$$

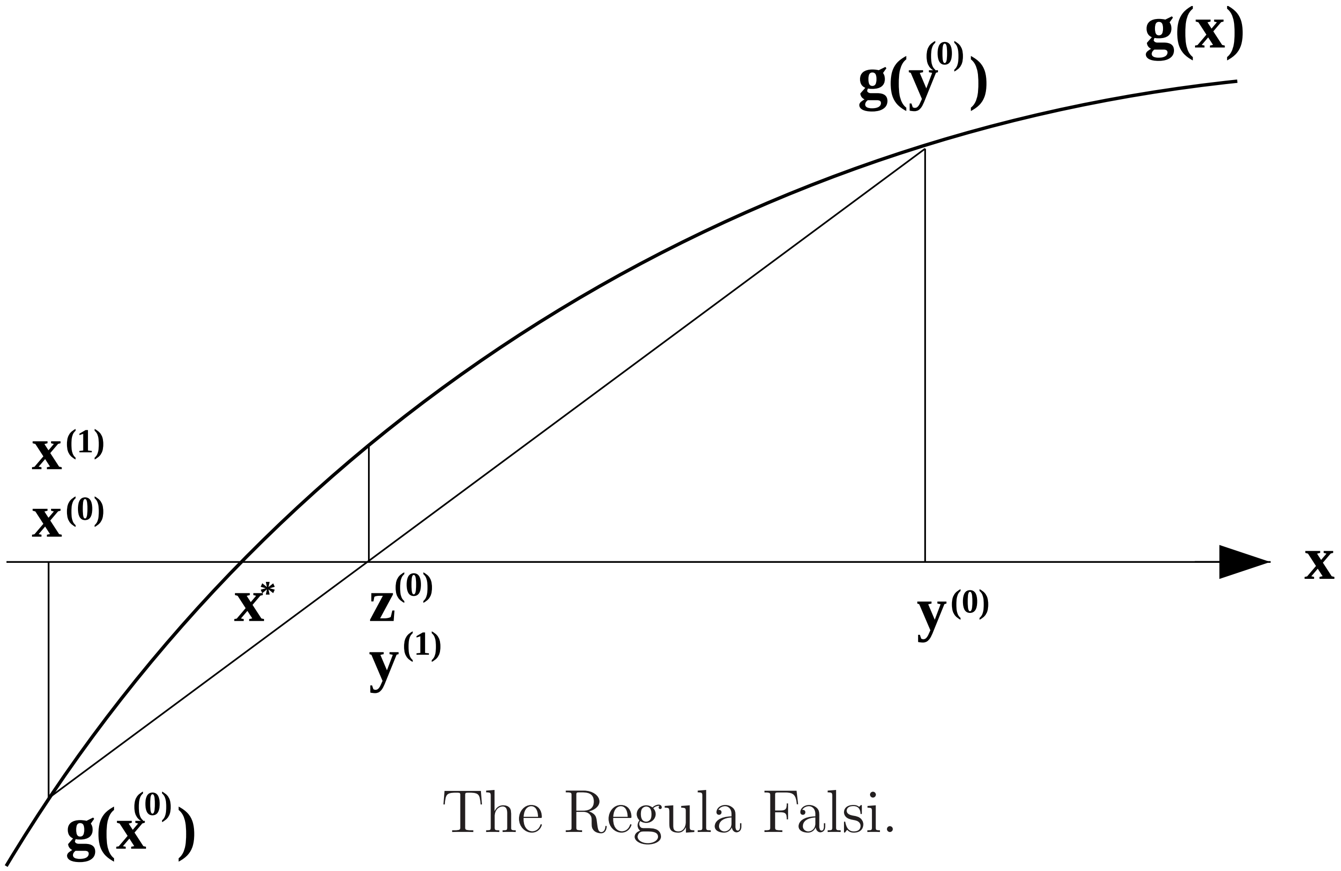
The method does not readily generalize to systems of nonlinear equations.

The Regula Falsi.

This method is similar to the bisection method.

However, in each step we now let

$$z^{(k)} \;=\; \frac{x^{(k)}g(y^{(k)}) - y^{(k)}g(x^{(k)})}{g(y^{(k)}) - g(x^{(k)})} \; .$$

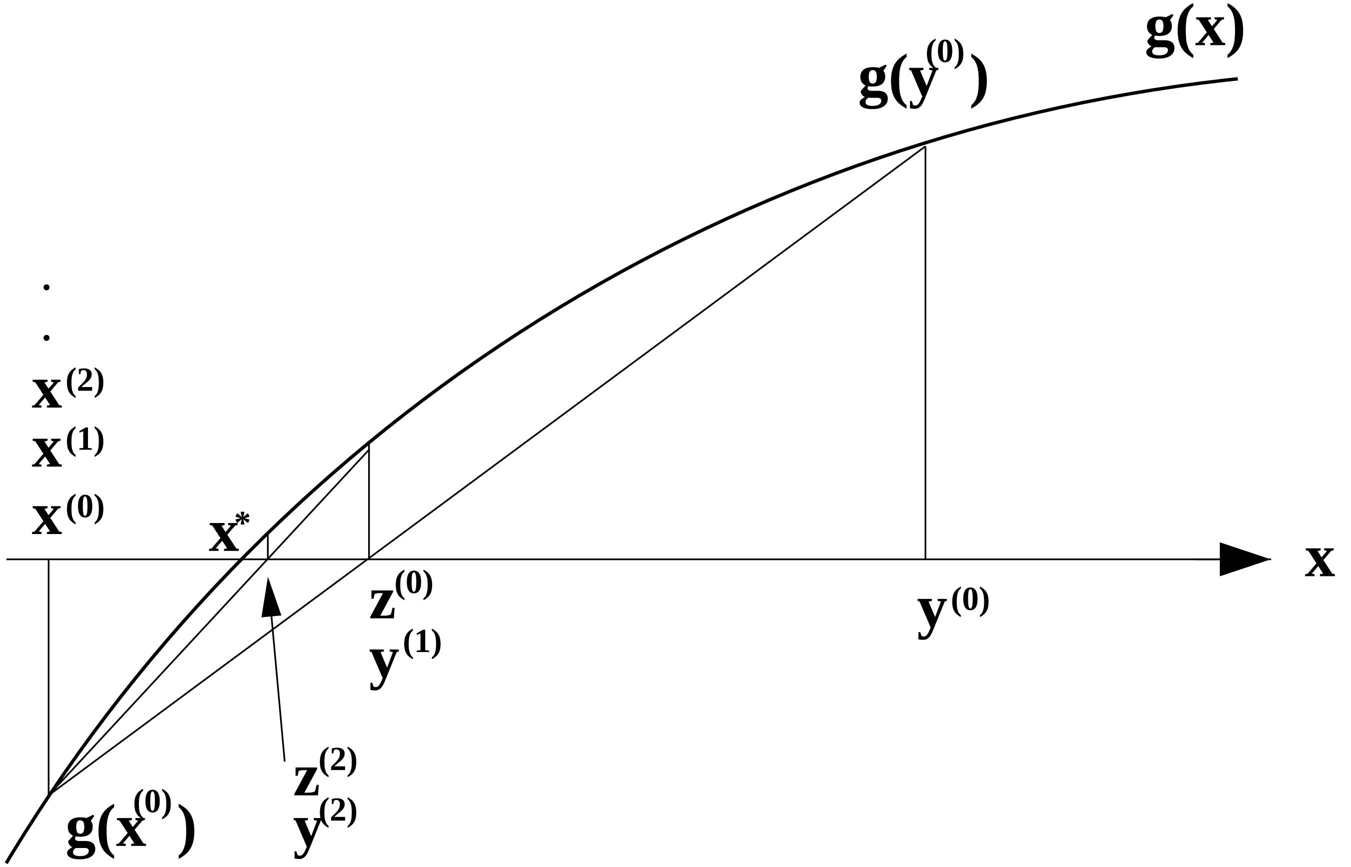


The Regula Falsi.

$$z^{(k)} \quad = \quad \frac{x^{(k)} g(y^{(k)}) - y^{(k)} g(x^{(k)})}{g(y^{(k)}) - g(x^{(k)})} \; .$$

$z^{(k)}$ is the zero of the line from $(x^{(k)}, g(x^{(k)}))$ to $(y^{(k)}, g(y^{(k)}))$. (**Check !**)

Unlike the bisection method, not both $x^{(k)}$ *and* $y^{(k)}$ need converge to x^* :



The Regula Falsi : Nonconvergence of $x^{(k)}$ to x^* .

The Regula Falsi does not readily generalize to nonlinear systems.

Newton's Method.

Let $x^{(0)}$ be an initial guess for a zero x^* of $g(x) = 0$.

The line $p_0(x)$ that satisfies

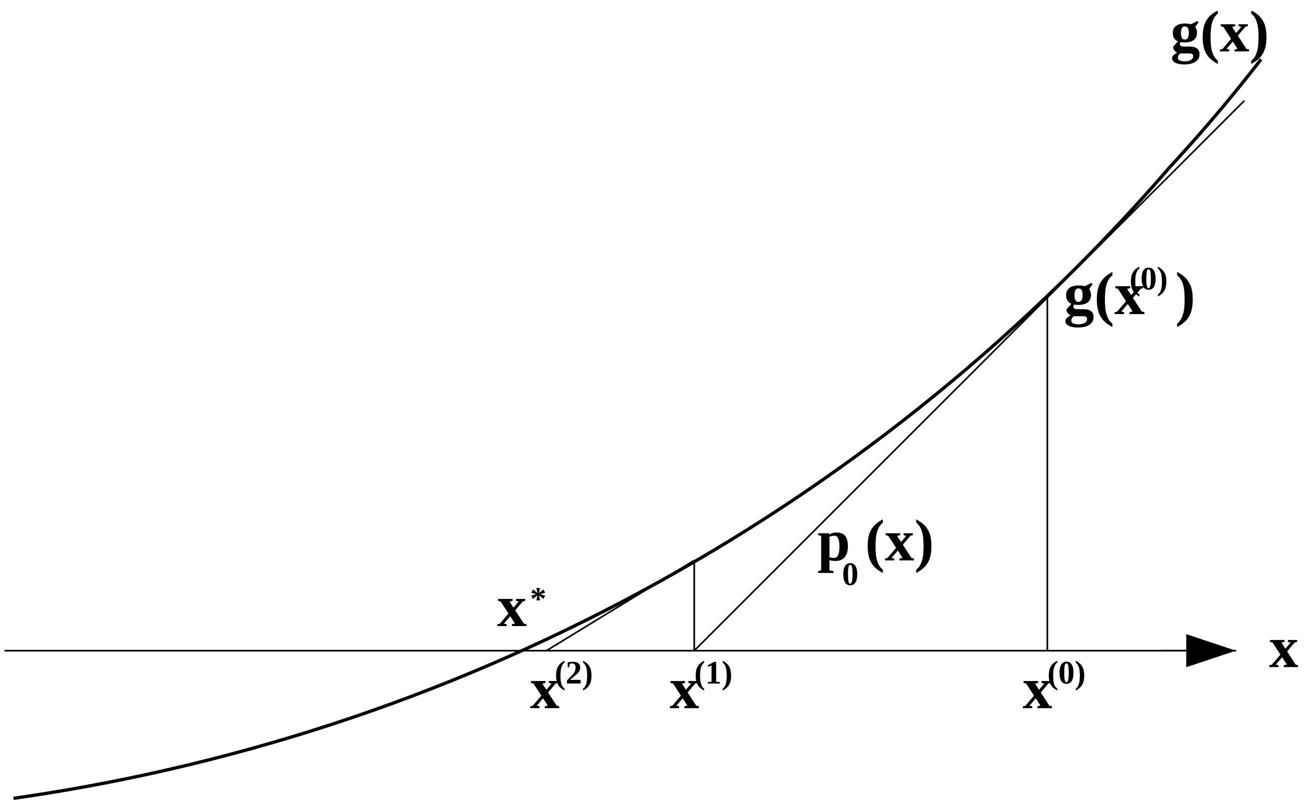
$$p_0(x^{(0)}) \;=\; g(x^{(0)})$$

and

$$p_0'(x^{(0)}) \;=\; g'(x^{(0)}) \; ,$$

is given by

$$p_0(x) \;=\; g(x^{(0)}) + (x - x^{(0)}) \; g'(x^{(0)}) \; . \qquad \textbf{(Check !)}$$



Newton's method

If g is sufficiently smooth and $x^{(0)}$ is close to x^* then we expect the zero

$$x^{(1)} = x^{(0)} - \frac{g(x^{(0)})}{g'(x^{(0)})} , \qquad (\textbf{Check !})$$

of $p_0(x)$ to be a better approximation to x^* than $x^{(0)}$.

This procedure may now be repeated for the point $x^{(1)}$.

The general algorithm for *Newton's method* can therefore be written as

$$x^{(k+1)} \ = \ x^{(k)} \ - \ \frac{g(x^{(k)})}{g'(x^{(k)})} , \qquad k \ = \ 0, \ 1, \ 2, \ \cdots \ .$$

Later we show that Newton's method converges to a zero x^* of $g(x) = 0$ if

- g has two continuous derivatives near x^*,
- $g'(x^*) \neq 0$,
- $x^{(0)}$ is sufficiently close to x^* .

EXAMPLE :

Use Newton's method to compute the square root of 2 .

Note that this square root is a zero of $g(x) \equiv x^2 - 2$.

Thus the Newton iteration is given by

$$x^{(k+1)} = x^{(k)} - \frac{(x^{(k)})^2 - 2}{2x^{(k)}} = \frac{(x^{(k)})^2 + 2}{2x^{(k)}} .$$

With $x^{(0)} = 1.5$, we get

$$x^{(1)} = 1.41666 , \qquad x^{(2)} = 1.414215 , \qquad etc.$$

Newton's method generalizes to systems of nonlinear equations.

This extension will be considered later.

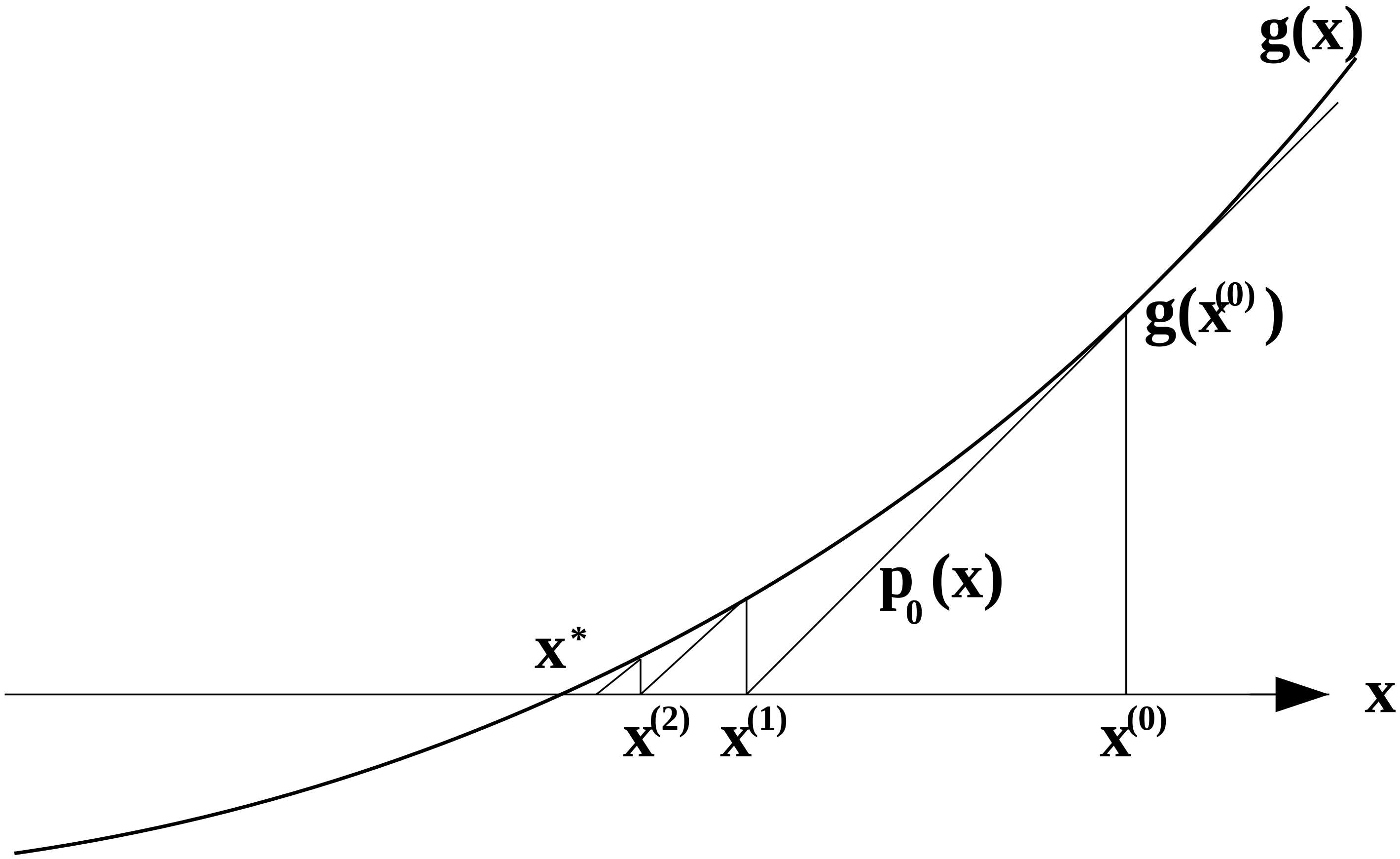
The Chord Method.

This method is similar to Newton's method.

The only difference is that $g'(x)$ is always evaluated at the initial point $x^{(0)}$.

Thus the algorithm is

$$x^{(k+1)} \ = \ x^{(k)} \ - \ \frac{g(x^{(k)})}{g'(x^{(0)})} \ .$$



The Chord method

$$x^{(k+1)} \ = \ x^{(k)} \ - \ \frac{g\big(x^{(k)}\big)}{g'\big(x^{(0)}\big)} \ .$$

Compared to Newton's method :

- The Chord method takes fewer arithmetic operations per iteration.

- The two methods converge under essentially the same conditions.

- The Chord method needs more iterations for a prescribed accuracy.

EXAMPLE :

With $x^{(0)} = 1.5$ the Chord method for solving $x^2 - 2 = 0$ takes the form

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

The first few iterations give

$$x^{(1)} \ = \ 1.416666 \ , \qquad x^{(2)} \ = \ 1.414351 \ , \qquad x^{(3)} \ = \ 1.414221 \ .$$

EXERCISES :

- Show how to use *Newton's method* to compute the cube root of 2. Carry out the first few iterations, using $x^{(0)} = 0.6$.

- Show how to use *the Chord method* to compute the cube root of 2. Carry out the first few iterations, using $x^{(0)} = 0.6$.

- Consider the equation
$$\sin(x) \ = \ 1/x \ .$$
Show the graphs of $\sin(x)$ and $1/x$ in one diagram. How many solutions are there to this equation ? Write down Newton's method for finding a solution. Carry out the first few iterations with $x^{(0)} = \pi/2$.

- Consider the equation $\sin(x) = e^{-x}$. Draw the functions $\sin(x)$ and e^{-x} in one graph. How many solutions are there to the above equation ? Show how one can use Newton's method to find a solution of the equation. Carry out the first few Newton iterations, using $x^{(0)} = 0$.

Newton's Method for Systems of Nonlinear Equations.

First reconsider Newton's method for scalar equations

$$g(x) \ = \ 0 \ .$$

Given $x^{(k)}$, we set $x^{(k+1)}$ to be the zero of

$$p_k(x) \ \equiv \ g(x^{(k)}) \ + \ (x - x^{(k)}) \ g'(x^{(k)}) \ .$$

NOTE :

- $p_k(x)$ is the *tangent line* to $g(x)$ at $x = x^{(k)}$, *i.e.*,
- $p_k(x)$ is the *linear approximation* to $g(x)$ at $x = x^{(k)}$, *i.e.*,
- $p_k(x)$ is the *degree 1 Taylor polynomial* of $g(x)$ at $x = x^{(k)}$.

Similarly for systems of the form

$$\mathbf{G}(\mathbf{x}) \ = \ \mathbf{0} \ ,$$

we have the *linear approximation*

$$\mathbf{P}_k(\mathbf{x}) \ \equiv \ \mathbf{G}(\mathbf{x}^{(k)}) \ + \ \mathbf{G}'(\mathbf{x}^{(k)}) \ (\mathbf{x} - \mathbf{x}^{(k)})$$

of $\mathbf{G}(\mathbf{x})$ about $\mathbf{x} = \mathbf{x}^{(k)}$.

Here $\mathbf{G}'(\mathbf{x}^{(k)})$ is the *Jacobian matrix* of $\mathbf{G}(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^{(k)}$.

Analogous to the scalar case we let $\mathbf{x}^{(k+1)}$ be the zero of $\mathbf{P}_k(\mathbf{x}) = \mathbf{0}$.

Thus $\mathbf{x}^{(k+1)}$ is the solution of the linear system

$$\mathbf{G}'(\mathbf{x}^{(k)})\ (\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \ = \ -\,\mathbf{G}(\mathbf{x}^{(k)})\ ,$$

that is, we can get $\mathbf{x}^{(k+1)}$ by first solving

- $\mathbf{G}'(\mathbf{x}^{(k)})\ \Delta\mathbf{x}^{(k)} \ = \ -\,\mathbf{G}(\mathbf{x}^{(k)})\ ,$

  and then setting

- $\mathbf{x}^{(k+1)} \ = \ \mathbf{x}^{(k)} \ + \ \Delta\mathbf{x}^{(k)}\ .$

EXAMPLE :

Use Newton's method to solve the system

$$\begin{aligned} x_1^2 x_2 \ - \ 1 \ &= \ 0 \ , \\ x_2 \ - \ x_1^4 \ &= \ 0 \ . \end{aligned}$$

Here

$$\mathbf{x} \ = \ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \ , \qquad \mathbf{G}(\mathbf{x}) \ = \ \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \end{pmatrix} \ = \ \begin{pmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{pmatrix} \ = \ \begin{pmatrix} x_1^2 x_2 - 1 \\ x_2 - x_1^4 \end{pmatrix} \ .$$

The Jacobian matrix in this example is given by

$$\mathbf{G}'(\mathbf{x}) \ \equiv \ \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{pmatrix} \ = \ \begin{pmatrix} 2x_1 x_2 & x_1^2 \\ \\ -4x_1^3 & 1 \end{pmatrix} \ .$$

Hence Newton's method for this problem takes the form

- $$\begin{pmatrix} 2x_1^{(k)}x_2^{(k)} & (x_1^{(k)})^2 \\ -4(x_1^{(k)})^3 & 1 \end{pmatrix} \begin{pmatrix} \Delta x_1^{(k)} \\ \Delta x_2^{(k)} \end{pmatrix} = \begin{pmatrix} 1-(x_1^{(k)})^2 x_2^{(k)} \\ (x_1^{(k)})^4 - x_2^{(k)} \end{pmatrix} ,$$

- $$\begin{cases} x_1^{(k+1)} = x_1^{(k)} + \Delta x_1^{(k)} , \\ x_2^{(k+1)} = x_2^{(k)} + \Delta x_2^{(k)} , \end{cases}$$

for $k = 0, 1, 2, \cdots$.

Thus for each iteration two linear equations in two unknowns must be solved.

With the initial guess

$$x_1^{(0)} = 2 \ , \qquad x_2^{(0)} = 2 \ ,$$

the first iteration consists of solving

$$\begin{pmatrix} 8 & 4 \\ -32 & 1 \end{pmatrix} \begin{pmatrix} \Delta x_1^{(0)} \\ \Delta x_2^{(0)} \end{pmatrix} = \begin{pmatrix} -7 \\ 14 \end{pmatrix} ,$$

which gives

$$\Delta x_1^{(0)} = -0.463 \ , \qquad \Delta x_2^{(0)} = -0.823 \ ,$$

and then setting

$$x_1^{(1)} = x_1^{(0)} + \Delta x_1^{(0)} \ = \ 1.537 \ ,$$

$$x_2^{(1)} = x_2^{(0)} + \Delta x_2^{(0)} \ = \ 1.177 \ .$$

After a second iteration what will $x_1^{(2)}$ and $x_2^{(2)}$ be ?

EXERCISES :

- Describe in detail how Newton's method can be used to compute solutions (x_1, x_2) of the system of two nonlinear equations

$$\begin{aligned} x_1^2 + x_2^2 - 1 &= 0 \ , \\ x_2 - e^{x_1} &= 0 \ . \end{aligned}$$

- Describe in detail how Newton's method can be used to compute a solution (x_1, x_2, x_3) of the system of three nonlinear equations

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 - 1 &= 0 \ , \\ x_3 - e^{x_1} &= 0 \ , \\ x_3 - e^{x_2} &= 0 \ . \end{aligned}$$

Residual Correction.

Suppose we use Newton's method for a *linear system* $\mathbf{A}\mathbf{x} = \mathbf{f}$,

that is, we let

$$\mathbf{G}(\mathbf{x}) \ \equiv \ \mathbf{A}\mathbf{x} \ - \ \mathbf{f} \ .$$

Then

$$\mathbf{G}'(\mathbf{x}) \ = \ \mathbf{A} \ ,$$

so that Newton's method

$$\mathbf{G}'(\mathbf{x}^{(k)}) \ \Delta\mathbf{x}^{(k)} \ = \ - \, \mathbf{G}(\mathbf{x}^{(k)}) \ ,$$

$$\mathbf{x}^{(k+1)} \ = \ \mathbf{x}^{(k)} \ + \ \Delta\mathbf{x}^{(k)} \ ,$$

becomes

$$\mathbf{A} \ \Delta\mathbf{x}^{(k)} \ = \ - \, (\mathbf{A}\mathbf{x}^{(k)} \ - \ \mathbf{f}) \ ,$$

$$\mathbf{x}^{(k+1)} \ = \ \mathbf{x}^{(k)} \ + \ \Delta\mathbf{x}^{(k)} \ .$$

$$\mathbf{A}\ \Delta\mathbf{x}^{(k)} = -\left(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{f}\right) ,$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)} .$$

NOTE :

- the Jacobian needs to be LU-decomposed only once.

- With *exact arithmetic* , the *exact solution* is found in *one iteration* :

$$\Delta\mathbf{x}^{(0)} = -\mathbf{A}^{-1}(\mathbf{A}\mathbf{x}^{(0)} - \mathbf{f}) = -\mathbf{x}^{(0)} + \mathbf{x} ,$$

so that

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \Delta\mathbf{x}^{(0)} = \mathbf{x}^{(0)} - \mathbf{x}^{(0)} + \mathbf{x} = \mathbf{x} .$$

$$\mathbf{A}\ \Delta\mathbf{x}^{(k)} \ = \ -\,(\mathbf{A}\mathbf{x}^{(k)}\ -\ \mathbf{f})\ ,$$

$$\mathbf{x}^{(k+1)} \ = \ \mathbf{x}^{(k)}\ +\ \Delta\mathbf{x}^{(k)}\ .$$

NOTE :

- For *inexact arithmetic* , this iteration is called *residual correction* .

- Residual correction can improve the accuracy of the solution of $\mathbf{A}\mathbf{x} = \mathbf{f}$.

- Residual correction is valuable for mildly ill-conditioned linear systems.

- The "*residual*" $\mathbf{A}\mathbf{x}^{(k)} - \mathbf{f}$ should be computed with high precision.

Convergence Analysis for Scalar Equations.

Most iterative methods for solving a scalar equation $g(x) = 0$ can be written

$$x^{(k+1)} \ = \ f(x^{(k)}) \ , \qquad k = 0, 1, 2, \cdots , \qquad x^{(0)} \ \text{ given} \ .$$

EXAMPLE : In Newton's method

$$f(x) \ = \ x \ - \ \frac{g(x)}{g'(x)} \ ,$$

and in the Chord method

$$f(x) \ = \ x \ - \ \frac{g(x)}{g'(x^{(0)})} \ .$$

Sometimes the iteration $x^{(k+1)} \ = \ x^{(k)} - g(x^{(k)})$ also works. In this method

$$f(x) \ = \ x - g(x) \ .$$

NOTE :

Iterations of the form

$$x^{(k+1)} \ = \ f(x^{(k)}) \ , \qquad k \ = \ 0, \ 1, \ 2, \ \cdots \ , \qquad x^{(0)} \ \text{ given} \ ,$$

also arise independently, *e.g.*, as models of “population growth” :

$$x^{(k+1)} \ = \ c \ x^{(k)}, \qquad k = 0, 1, 2, \cdots, \qquad x^{(0)} > 0 \ \text{ given} \ ,$$

models exponential population growth.

EXERCISES : What happens to the sequence $x^{(k)}$, $k = 0, 1, 2, \cdots$,

- when $x^{(0)} \ > \ 0$ and $c \ > \ 1$?
- when $x^{(0)} \ > \ 0$ and $c \ < \ 1$?

The iteration

$$x^{(k+1)} \ = \ c\, x^{(k)}\ (1 - x^{(k)})\ , \qquad k = 0, 1, 2, \cdots ,$$

known as the *logistic equation*, models population growth when there are limited resources.

EXERCISES : For $0 \ < \ x^{(0)} \ < \ 1$:

What happens to the sequence $x^{(k)}$, $k = 0, 1, 2, \cdots$,

- when $0 \ \le \ c \ < \ 1$?
- when $1 \ \le \ c \ < \ 2$?
- when $2 \ \le \ c \ < \ 3$?
- when $3 \ \le \ c \ \le \ 4$?

In general, an iteration of the form

$$x^{(k+1)} \ = \ f(x^{(k)}) \ , \qquad k = 0, 1, 2, \cdots ,$$

is sometimes called a *fixed point iteration* (or a *recurrence relation* , or a *discrete dynamical system*) .

Suppose the sequence $x^{(k)}$, $k = 0, 1, 2, \cdots$, converges, *i.e.*,

$$\lim_{k \to \infty} x^{(k)} \ = \ x^* \ .$$

Then x^* satisfies the equation

$$x \ = \ f(x) \ ,$$

(assuming that f is continuous near x^*).

We call x^* a *fixed point* of f.

EXAMPLE :

In Newton's method

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

Thus a fixed point x^* satisfies

$$x^* \ = \ x^* \ - \ \frac{g(x^*)}{g'(x^*)} \ ,$$

that is,

$$g(x^*) \ = \ 0 \ ,$$

(assuming that $g'(x^*) \ \neq \ 0$.)

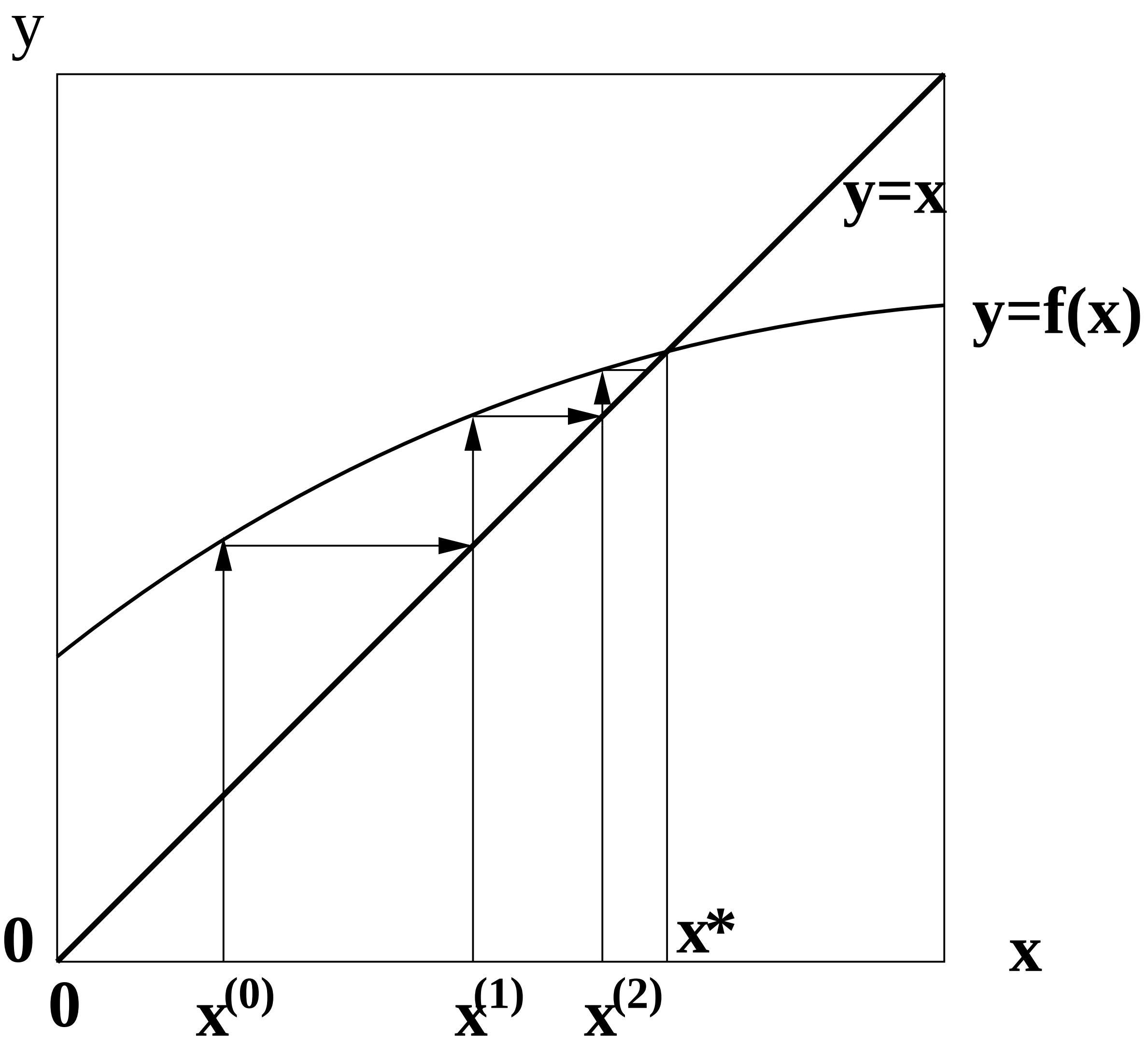
Thus x^* is a solution of $g(x) = 0$.

Assuming that f has a fixed point, when does the fixed point iteration

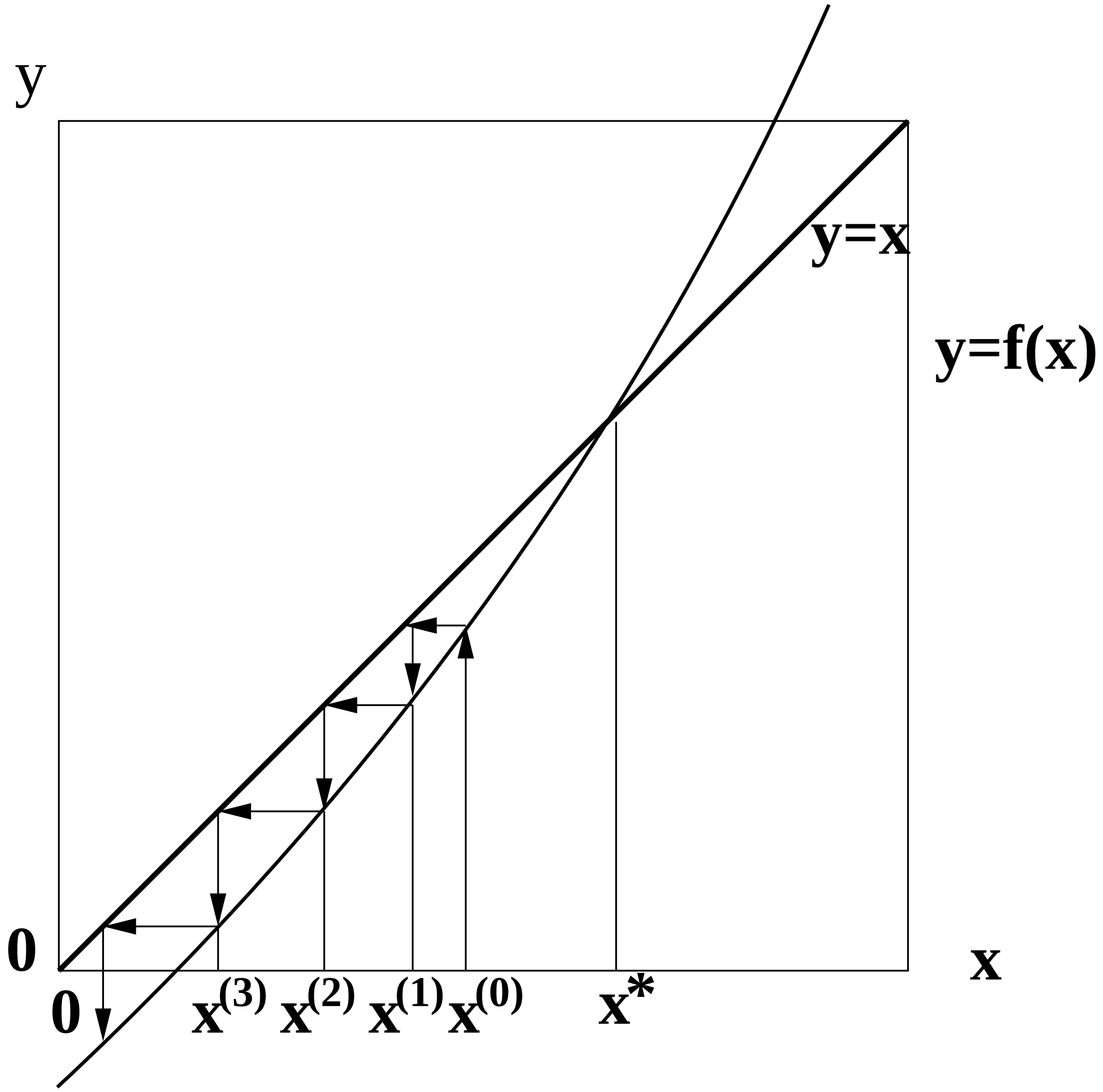
$$x^{(k+1)} \; = \; f(x^{(k)}) \; ,$$

converge ?

The answer is suggested in the following two diagrams :



A *convergent* fixed point iteration.



A *divergent* fixed point iteration.

THEOREM :

Let $f'(x)$ be continuous near a fixed point x^* of $f(x)$, and assume that

$$| \, f'(x^*) \, | \quad < 1 \, .$$

Then the fixed point iteration

$$x^{(k+1)} \;=\; f(x^{(k)}) \; , \qquad k = 0, 1, 2, \cdots \; ,$$

converges to x^*, whenever the initial guess $x^{(0)}$ is sufficiently close to x^* .

PROOF :

Let $\alpha \equiv | f'(x^*) |$.

Then $\alpha < 1$.

Choose β such that $\alpha < \beta < 1$.

Then, for some $\epsilon > 0$, there exists an interval

$$I_\epsilon \equiv [x^* - \epsilon, x^* + \epsilon] ,$$

such that

$$| f'(x) | \leq \beta \quad \text{in} \quad I_\epsilon ,$$

(because f' is continuous near x^*).

Let $x^{(0)} \in I_\epsilon$.

By Taylor's Theorem (or by the Mean Value Theorem)

$$x^{(1)} \ \equiv \ f(x^{(0)}) \ = \ f(x^*) \ + \ (x^{(0)} - x^*) \ f'(\eta_0) \ ,$$

for some η_0 between $x^{(0)}$ and x^*.

Since $f(x^*) = x^*$ it follows that

$$|\, x^{(1)} - x^* \,| \ = \ |\, (x^{(0)} - x^*) f'(\eta_0) \,| \ = \ |\, x^{(0)} - x^* \,| \ \ |\, f'(\eta_0) \,| \ \leq \ \beta \,|\, x^{(0)} - x^* \,| \ .$$

Thus $x^{(1)} \in I_\epsilon$, (because $0 < \beta < 1$) .

Again by Taylor's Theorem (or the Mean Value Theorem)

$$x^{(2)} \equiv f(x^{(1)}) = f(x^*) + (x^{(1)} - x^*)\ f'(\eta_1)\ ,$$

for some η_1 between $x^{(1)}$ and x^*.

Hence

$$|\ x^{(2)} - x^*\ | \leq \beta\ |\ x^{(1)} - x^*\ | \leq \beta^2\ |\ x^{(0)} - x^*\ |\ .$$

Thus $x^{(2)} \in I_\epsilon$, (because $0 < \beta < 1$) .

Proceeding in this manner we find

$$|\ x^{(k)} - x^*\ | \leq \beta^k\ |\ x^{(0)} - x^*\ |\ .$$

Since $0 < \beta < 1$ this implies that

$$x^{(k)} \to x^* \quad \text{as} \quad k \to \infty\ .$$

COROLLARY : Let

$$I_\epsilon \ \equiv \ [\, x^* - \epsilon \ , \ x^* + \epsilon \,] \ ,$$

and assume that for some $\epsilon > 0$ we have :

- $f(x)$ has a fixed point $x^* \in I_\epsilon$,
- $f(x)$ is a smooth function in I_ϵ ,
- $|\, f'(x) \,| \ < \ 1$ everywhere in I_ϵ ,
- $x^{(0)} \in I_\epsilon$.

Then the fixed point iteration

$$x^{(k+1)} \ = \ f(x^{(k)}) \ , \qquad k = 0, 1, 2, \cdots \ .$$

converges to x^*.

PROOF : This follows from the proof of the Theorem.

COROLLARY :

If

- x^* is a zero of $g(x) = 0$,
- $g(x)$ has two continuous derivatives near x^* ,
- $g'(x^*) \neq 0$,
- $x^{(0)}$ is sufficiently close to x^* .

then *Newton's method* for solving $g(x) = 0$, *i.e.*,

$$x^{(k+1)} \; = \; x^{(k)} - \frac{g(x^{(k)})}{g'(x^{(k)})} \; ,$$

converges to x^* .

PROOF :

In Newton's method

$$f(x) \;=\; x \;-\; \frac{g(x)}{g'(x)} \;.$$

Hence

$$f'(x^*) \;=\; 1 \;-\; \frac{g'(x^*)^2 \;-\; g(x^*)\; g''(x^*)}{g'(x^*)^2} \;=\; \frac{g(x^*)\; g''(x^*)}{g'(x^*)^2} \;=\; 0 \;.$$

Therefore, certainly $\mid f'(x^*) \mid < 1$.

EXAMPLE :

The fixed points of the *logistic equation* ,

$$x^{(k+1)} \;=\; c\, x^{(k)}\; (1 - x^{(k)})\;,$$

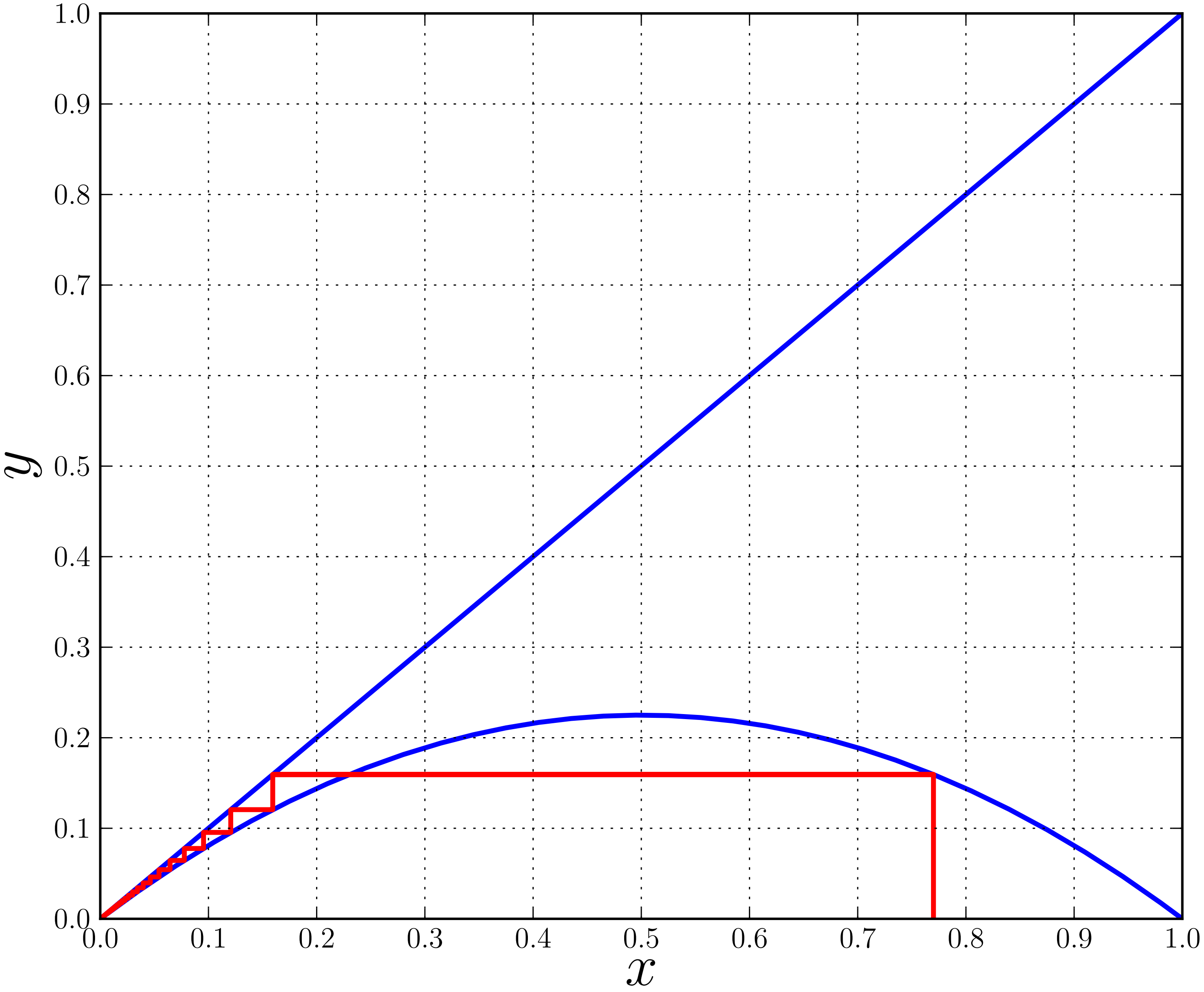
satisfy

$$x^* \;=\; c\, x^*\; (1 - x^*)\;.$$

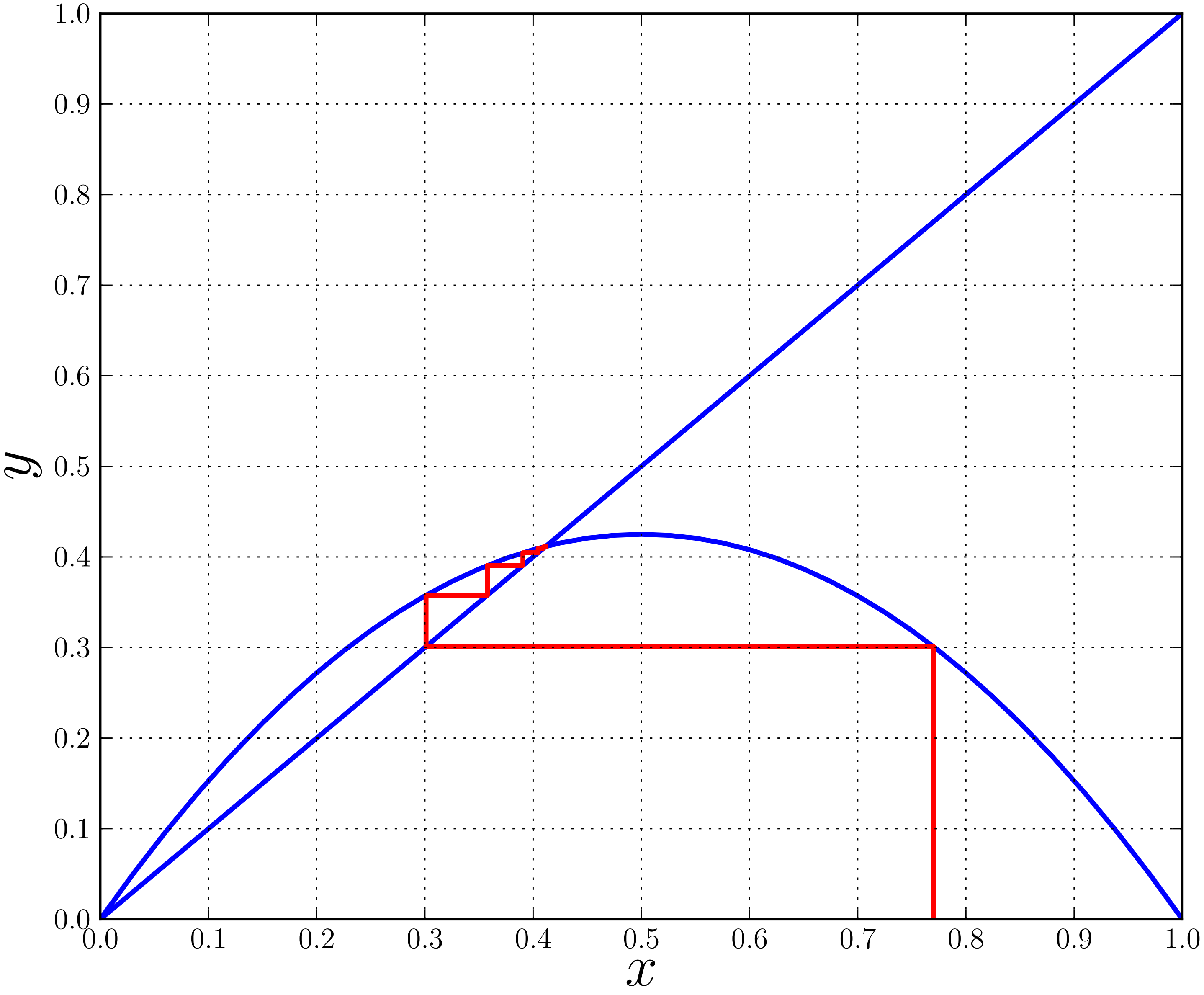
We see that the fixed points are given by

$$x^* \;=\; 0\;, \qquad \text{and} \qquad x^* \;=\; 1 \;-\; \frac{1}{c}\;.$$

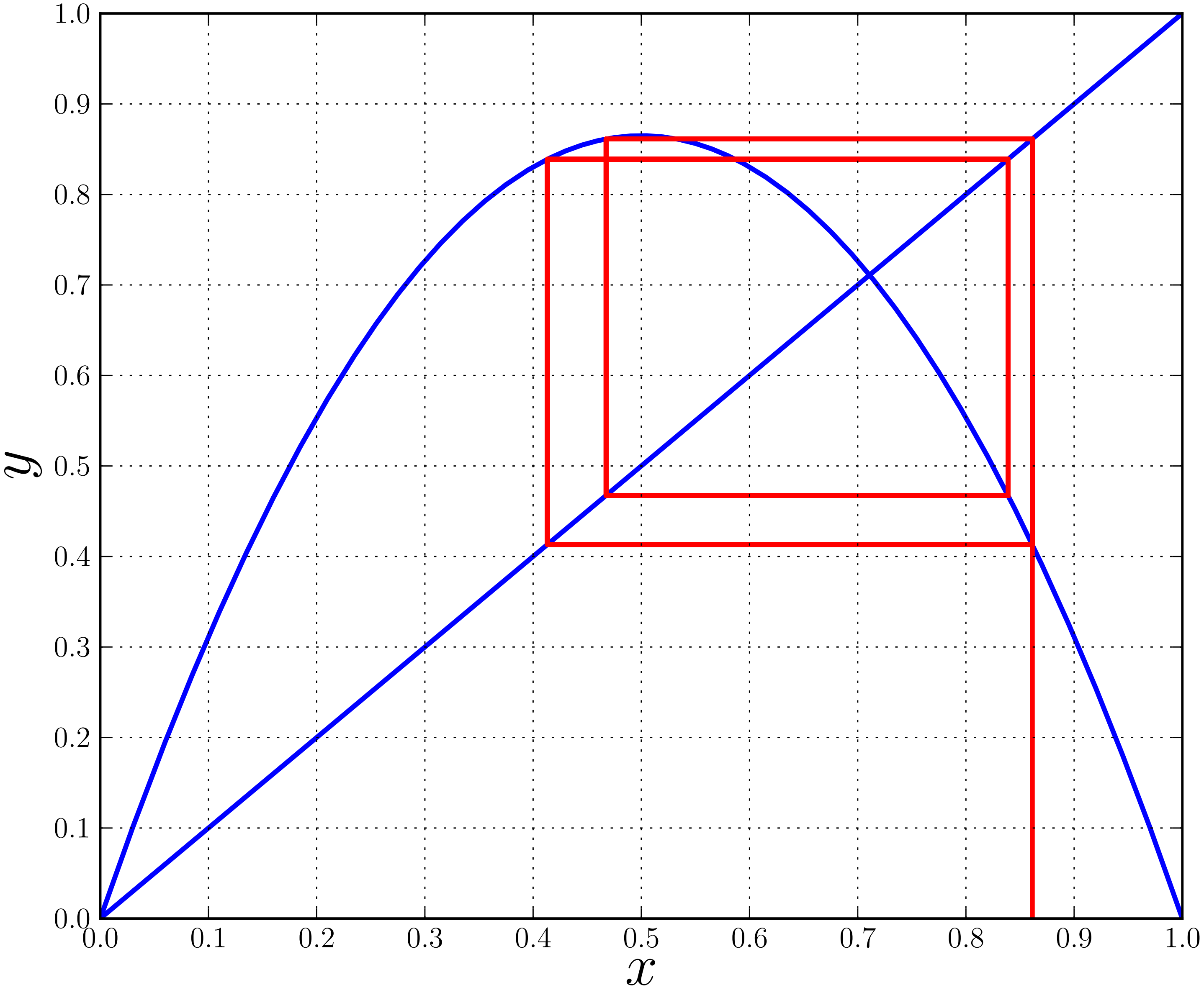
EXERCISE : ● Determine, for all values of c , $(0 < c \leq 4)$, whether these fixed points are *attracting* $(\mid f'(x^*) \mid < 1)$, or *repelling* $(\mid f'(x^*) \mid > 1)$.



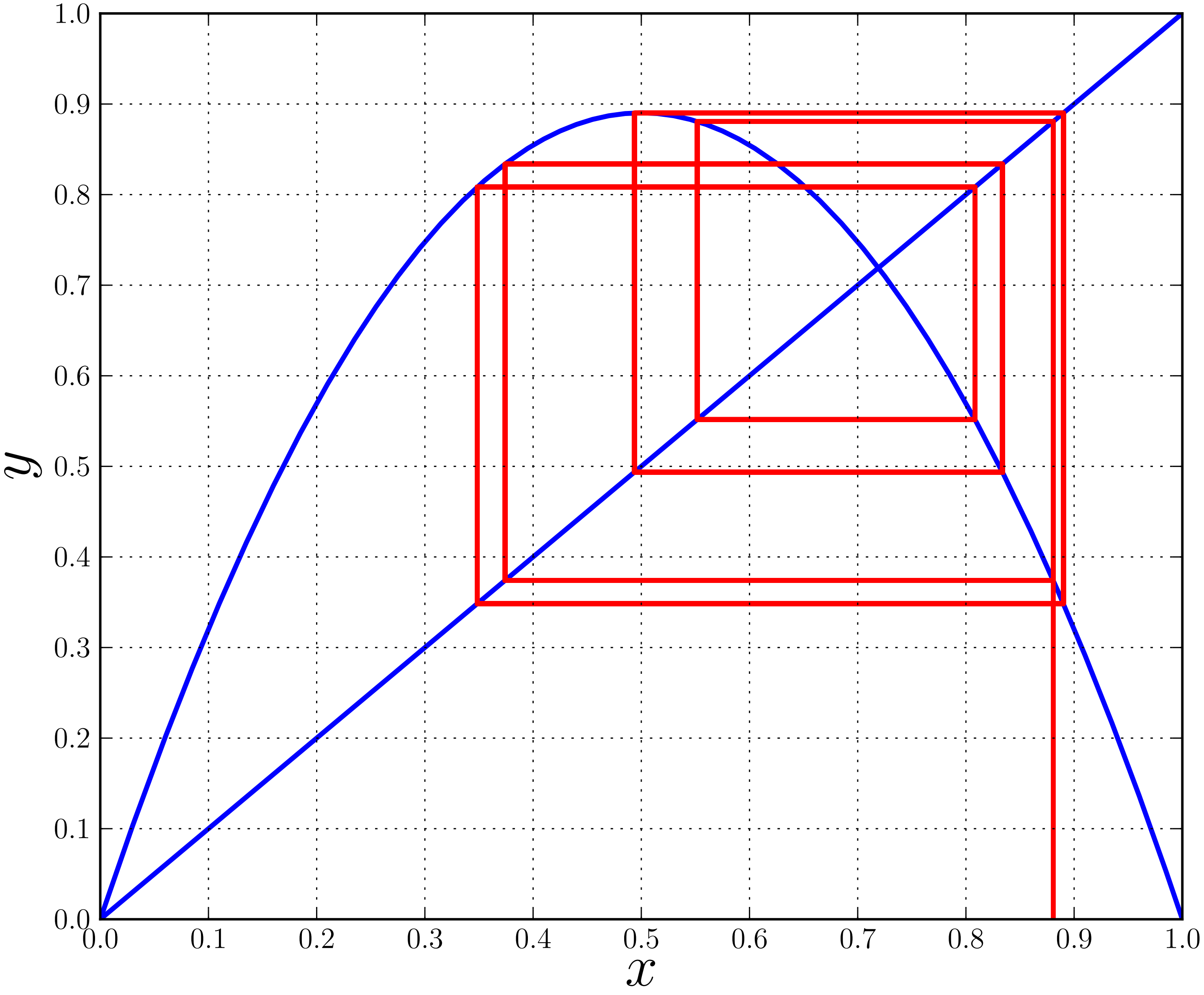
The logistic equation : $c = 0.9$.



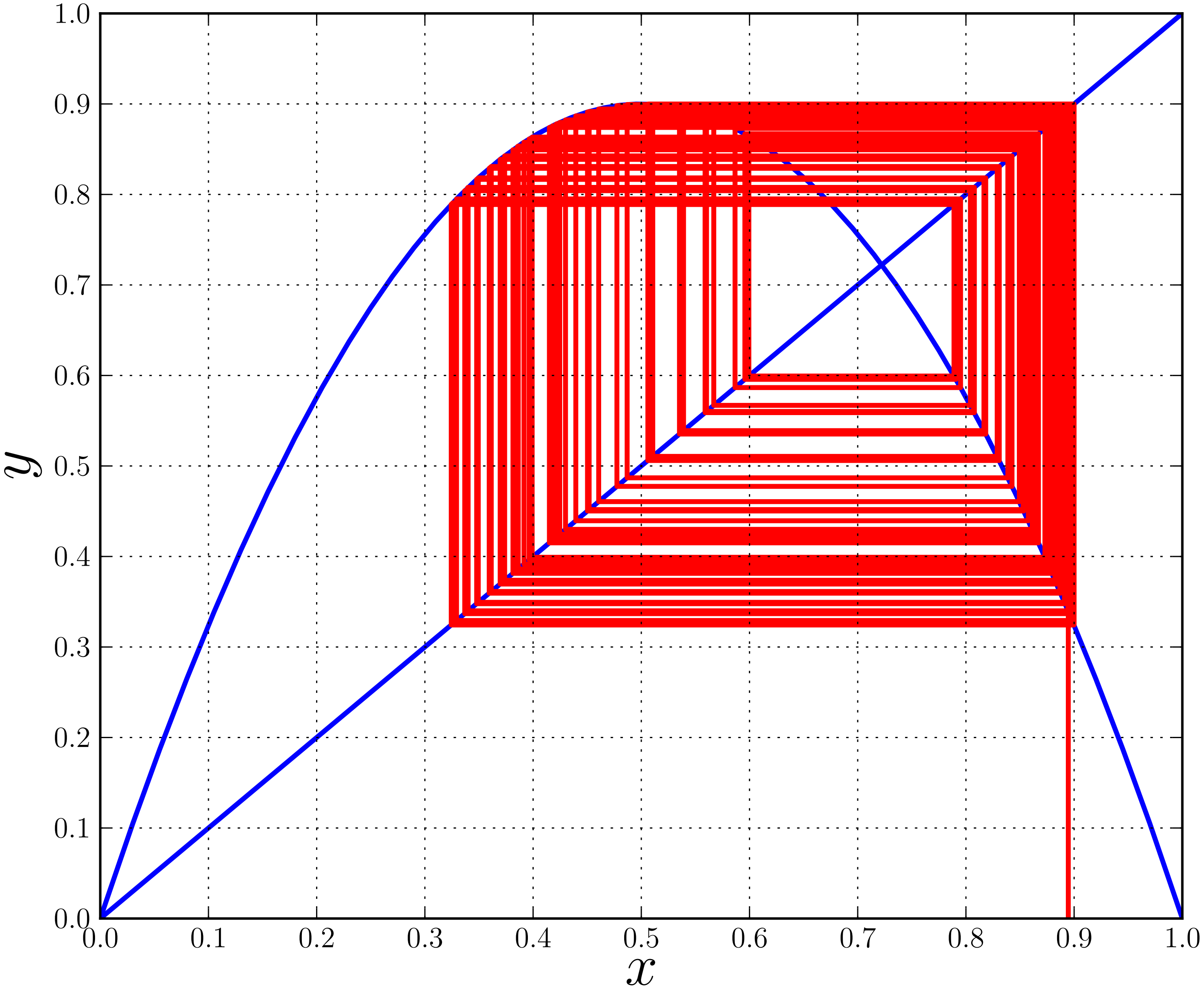
The logistic equation : $c = 1.7$.



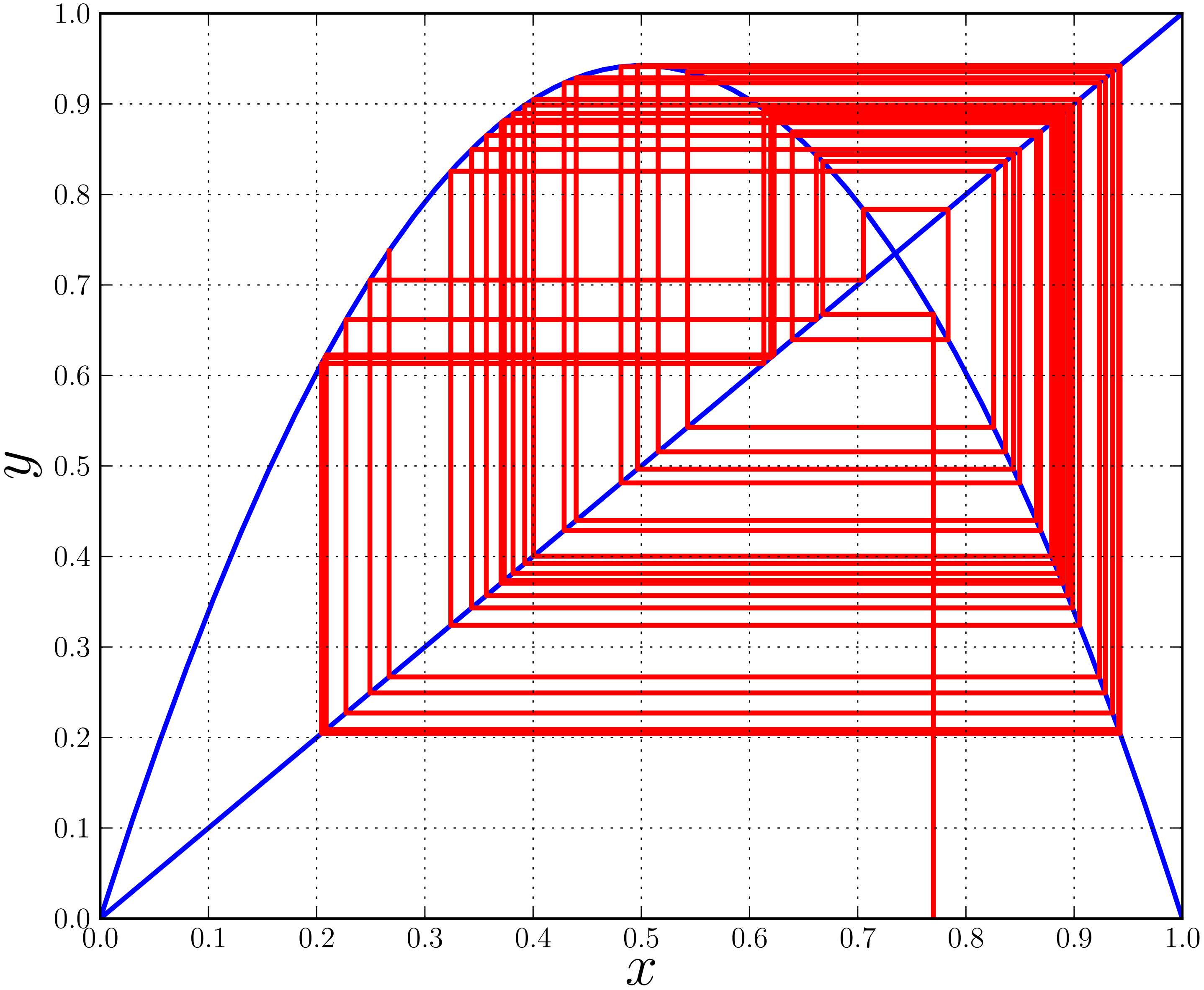
The logistic equation : $c = 3.46$.



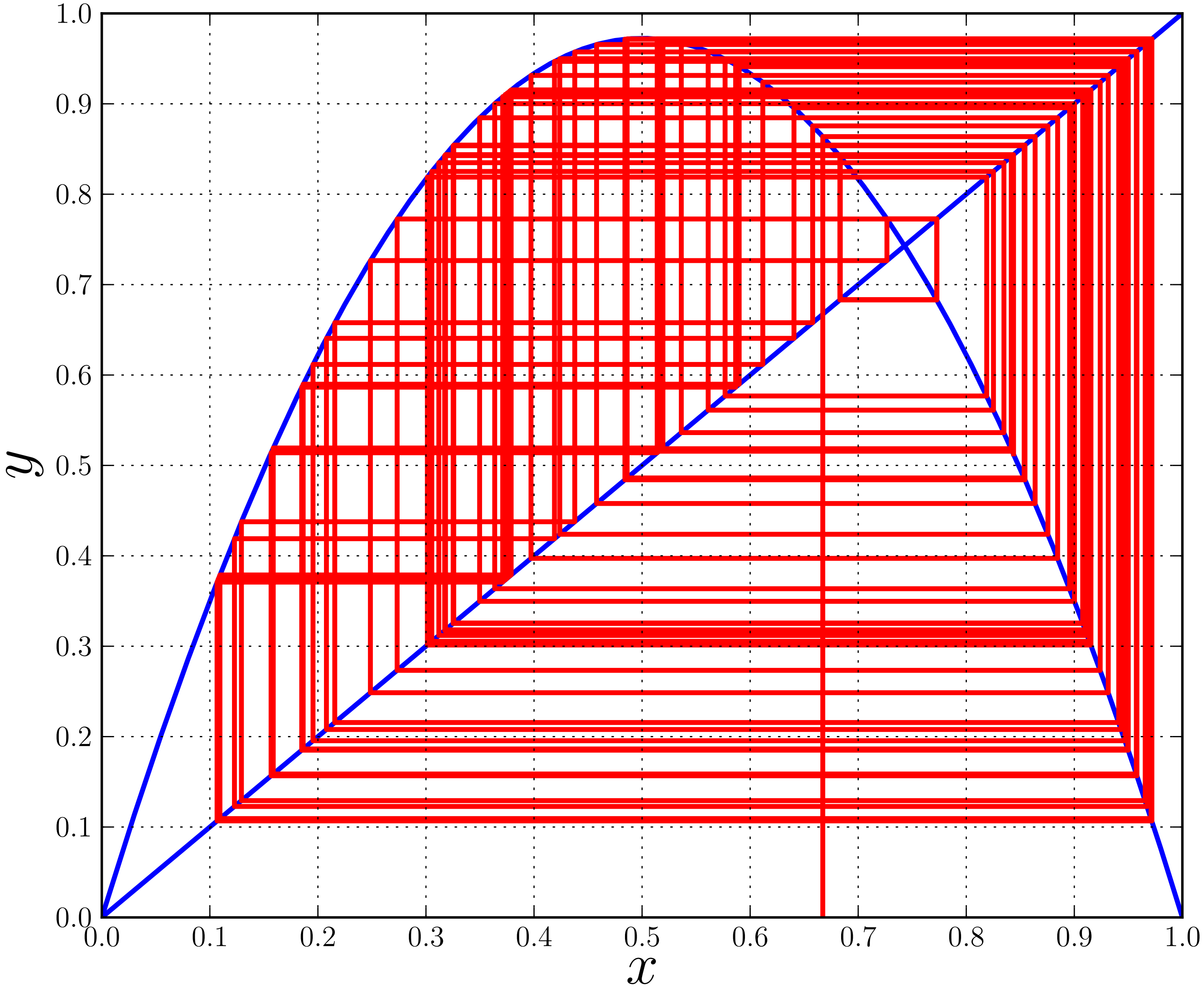
The logistic equation : $c = 3.561$.



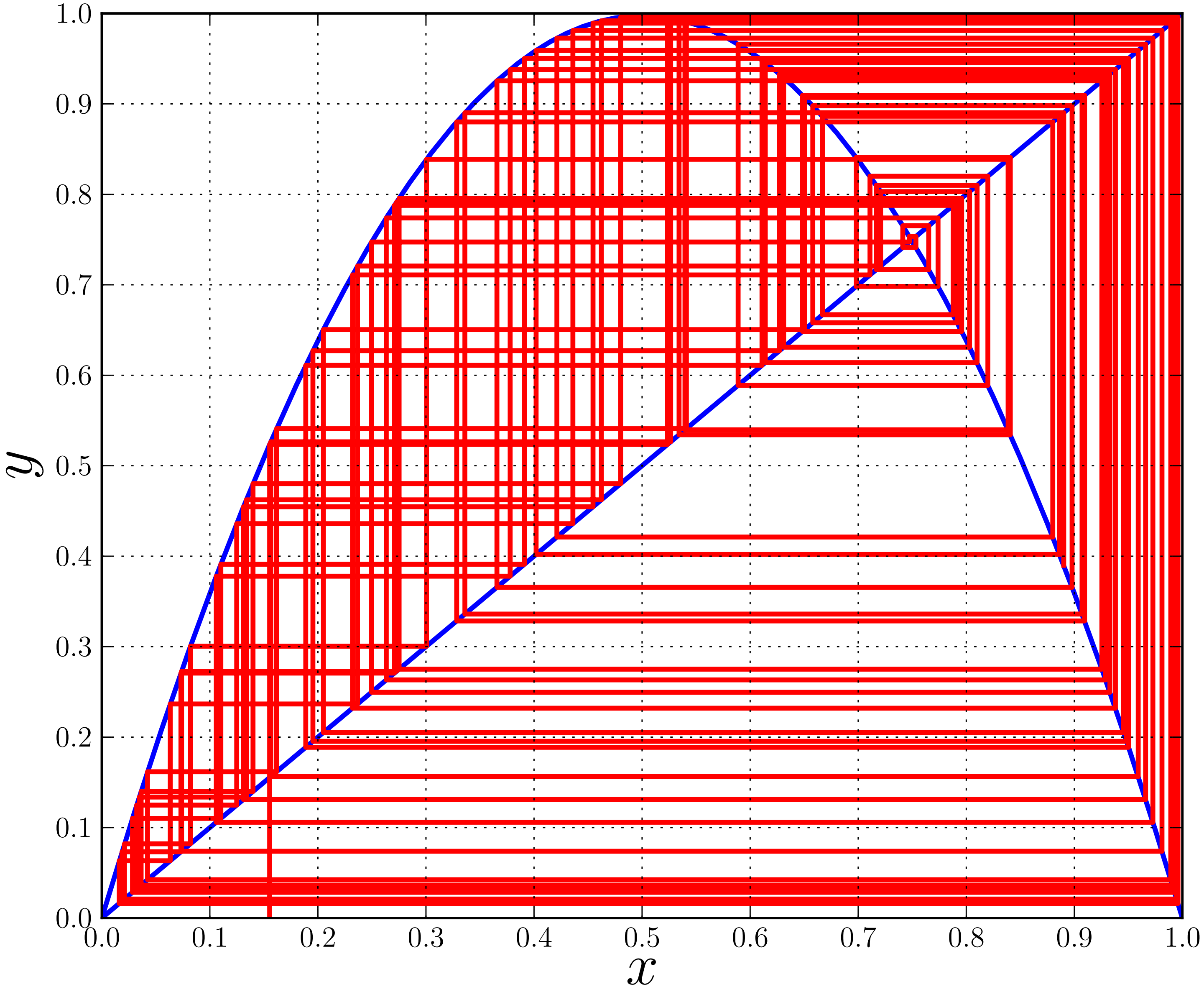
The logistic equation : $c = 3.6$.



The logistic equation : $c = 3.77$.



The logistic equation : $c = 3.89$.



The logistic equation : $c = 3.99$.

If a fixed point iteration

$$x^{(k+1)} = f(x^{(k)}) \ ,$$

converges to a fixed point x^* of $f(x)$, then *how fast* does it converge ?

To answer this we let

$$e_k \equiv \ | \, x^{(k)} - x^* \, | \ .$$

Thus e_k is the error after the kth iteration.

We can now show the following :

THEOREM :

Let $f'(x)$ be continuous near x^* with $| f'(x^*) | < 1$.

Assume that $x^{(0)}$ is sufficiently close to x^*, so that the fixed point iteration

$$x^{(k+1)} = f(x^{(k)})$$

converges to x^*.

Then
$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k} = | f'(x^*) | \qquad (\textit{linear convergence}) .$$

(The value of $| f'(x^*) |$ is then called *the rate of convergence*.)

If in addition $f'(x^*) = 0$ and if $f''(x)$ is continuous near x^* then

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k^2} = \frac{1}{2} | f''(x^*) | \qquad (\textit{quadratic convergence}) .$$

PROOF :

Case : $f'(x^*) \neq 0$:

$$\begin{aligned} e_{k+1} &= \mid x^{(k+1)} - x^* \mid \\ &= \mid f(x^{(k)}) - x^* \mid \\ &= \mid f(x^*) + (x^{(k)} - x^*)\ f'(\eta_k) - x^* \mid \\ &= \mid x^{(k)} - x^* \mid \ \mid f'(\eta_k) \mid \\ &= e_k \ \mid f'(\eta_k) \mid \ , \end{aligned}$$

where η_k is some point between $x^{(k)}$ and x^*.

Hence

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k} = \lim_{k \to \infty} \mid f'(\eta_k) \mid = \mid f'(x^*) \mid \ .$$

Case : $f'(x^*) \ = \ 0$:

$$\begin{aligned} e_{k+1} &= \mid x^{(k+1)} - x^* \mid \quad = \quad \mid f(x^{(k)}) - x^* \mid \\ &= \mid f(x^*) \ + \ (x^{(k)} - x^*) f'(x^*) \ + \ \frac{1}{2} \, (x^{(k)} - x^*)^2 \ f''(\eta_k) \ - \ x^* \mid \\ &= \frac{1}{2} e_k^2 \mid f''(\eta_k) \mid . \end{aligned}$$

where η_k is some point between $x^{(k)}$ and x^*,

Thus

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k^2} \quad = \quad \lim_{k \to \infty} \frac{1}{2} \mid f''(\eta_k) \mid \quad = \quad \frac{1}{2} \mid f''(x^*) \mid \qquad \textbf{QED !}$$

COROLLARY :

If

- $g(x)$ has three continuous derivatives near a zero x^* of $g(x) = 0$,
- $g'(x^*) \neq 0$,
- $x^{(0)}$ is sufficiently close to x^* ,

then Newton's method for solving $g(x) = 0$ *converges quadratically* .

PROOF : In Newton's method

$$f(x) \ = \ x \ - \ \frac{g(x)}{g'(x)} \ ,$$

and we have already shown that

$$f'(x^*) \ = \ 0 \ .$$

EXAMPLE :

Newton's method for computing $\sqrt{2}$, *i.e.*, for computing a zero of

$$g(x) \ = \ x^2 - 2 \ ,$$

is given by

$$x^{(k+1)} \ = \ x^{(k)} \ - \ \frac{(x^{(k)})^2 \ - \ 2}{2x^{(k)}} \ ,$$

that is,

$$x^{(k+1)} \ = \ \frac{(x^{(k)})^2 \ + \ 2}{2x^{(k)}} \ ,$$

that is,

$$x^{(k+1)} \ = \ f(x^{(k)}) \ ,$$

where

$$f(x) \ = \ \frac{x^2 + 2}{2x} \ .$$

$$x^{(k+1)} = f(x^{(k)}) \ , \quad \text{where} \quad f(x) = \frac{x^2+2}{2x} \ .$$

We observe that :

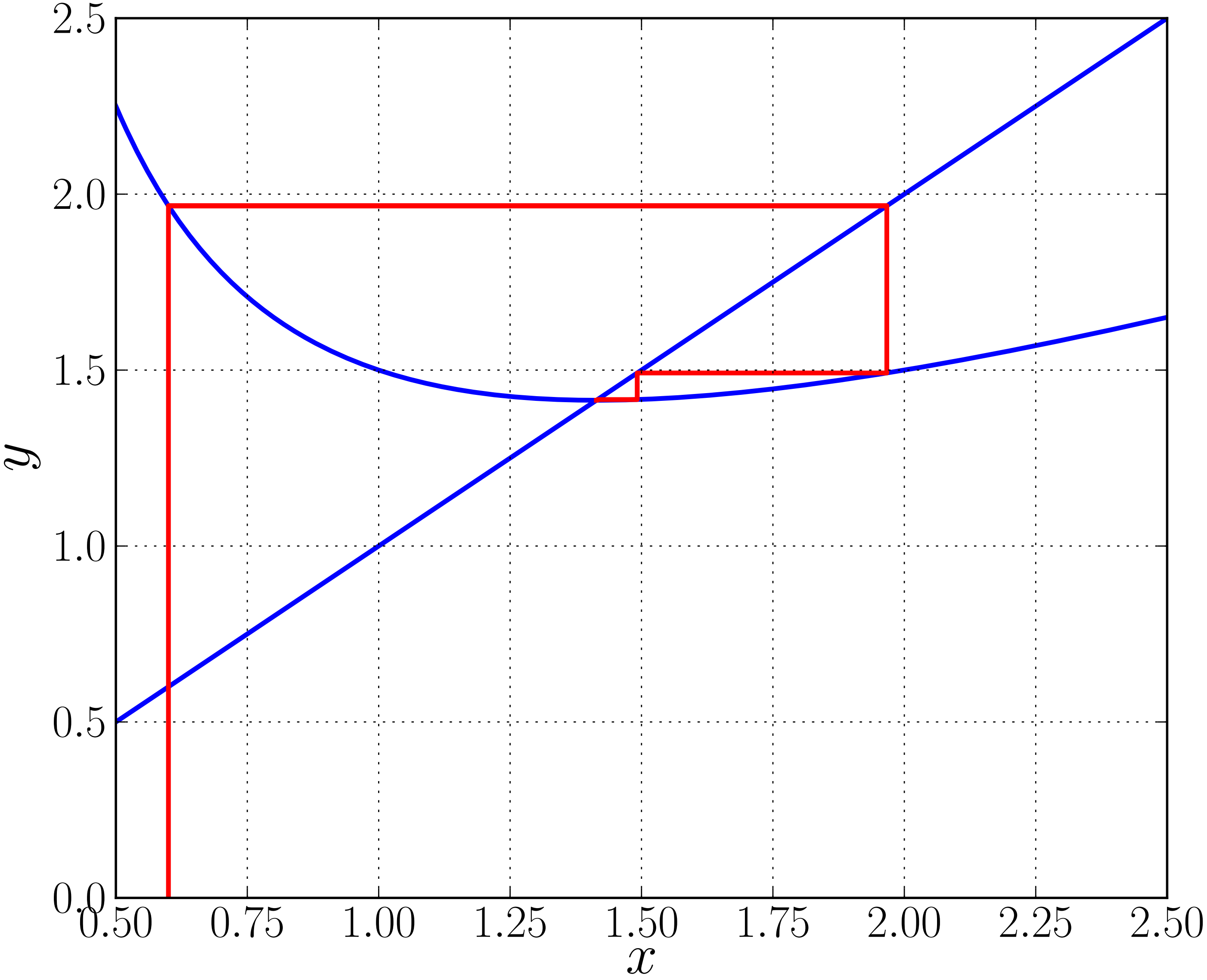
- The fixed points of f are $x^* = +\sqrt{2}$ and $x^* = -\sqrt{2}$.

- $f'(x^*) = 0$. (Hence quadratic convergence).

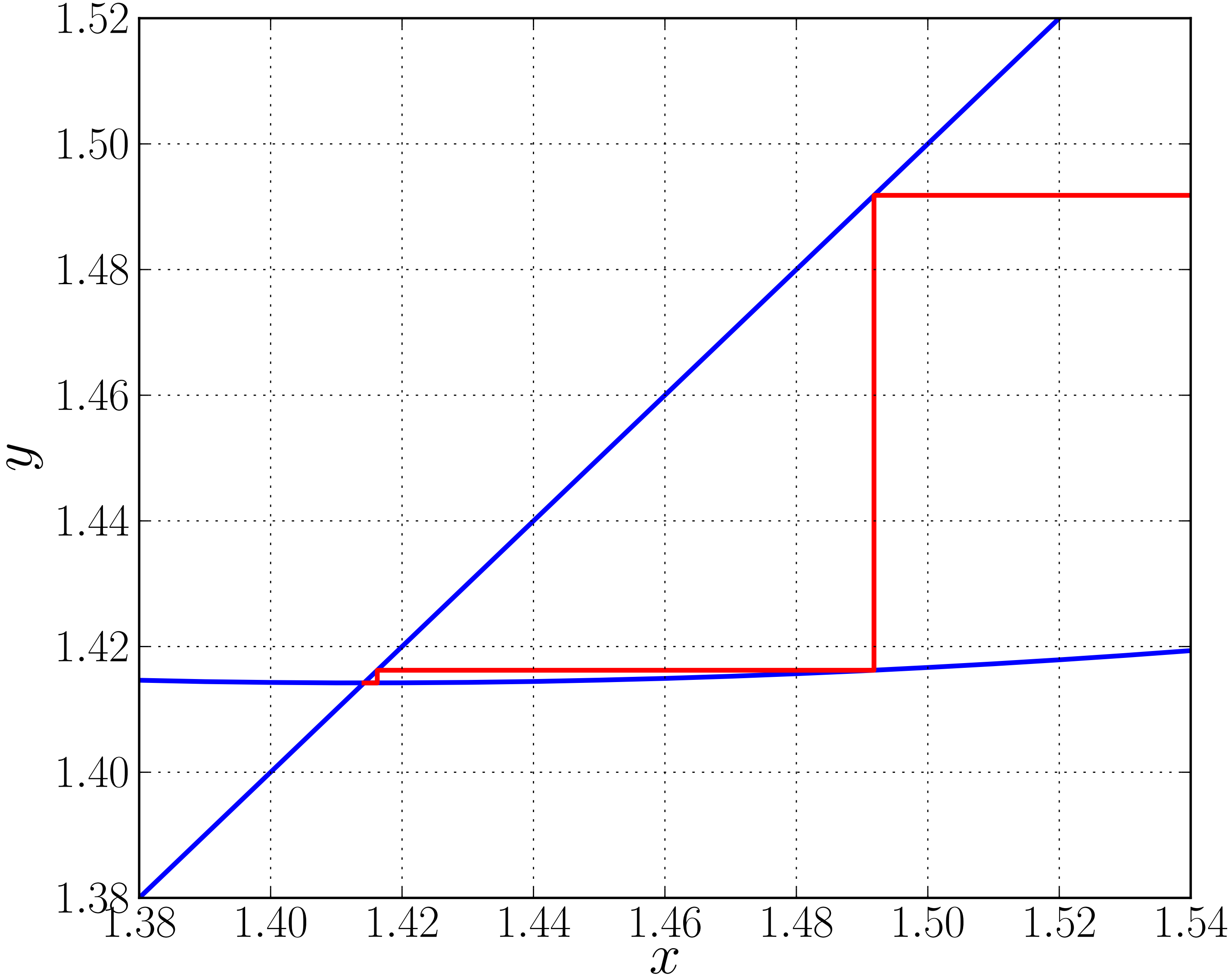
- $f(x) \to \infty$ as $x \downarrow 0$, and $f(x) \to -\infty$ as $x \uparrow 0$. (Vertical asymptotes.)

- $f(x) \cong x/2$ as $| x | \to \infty$.

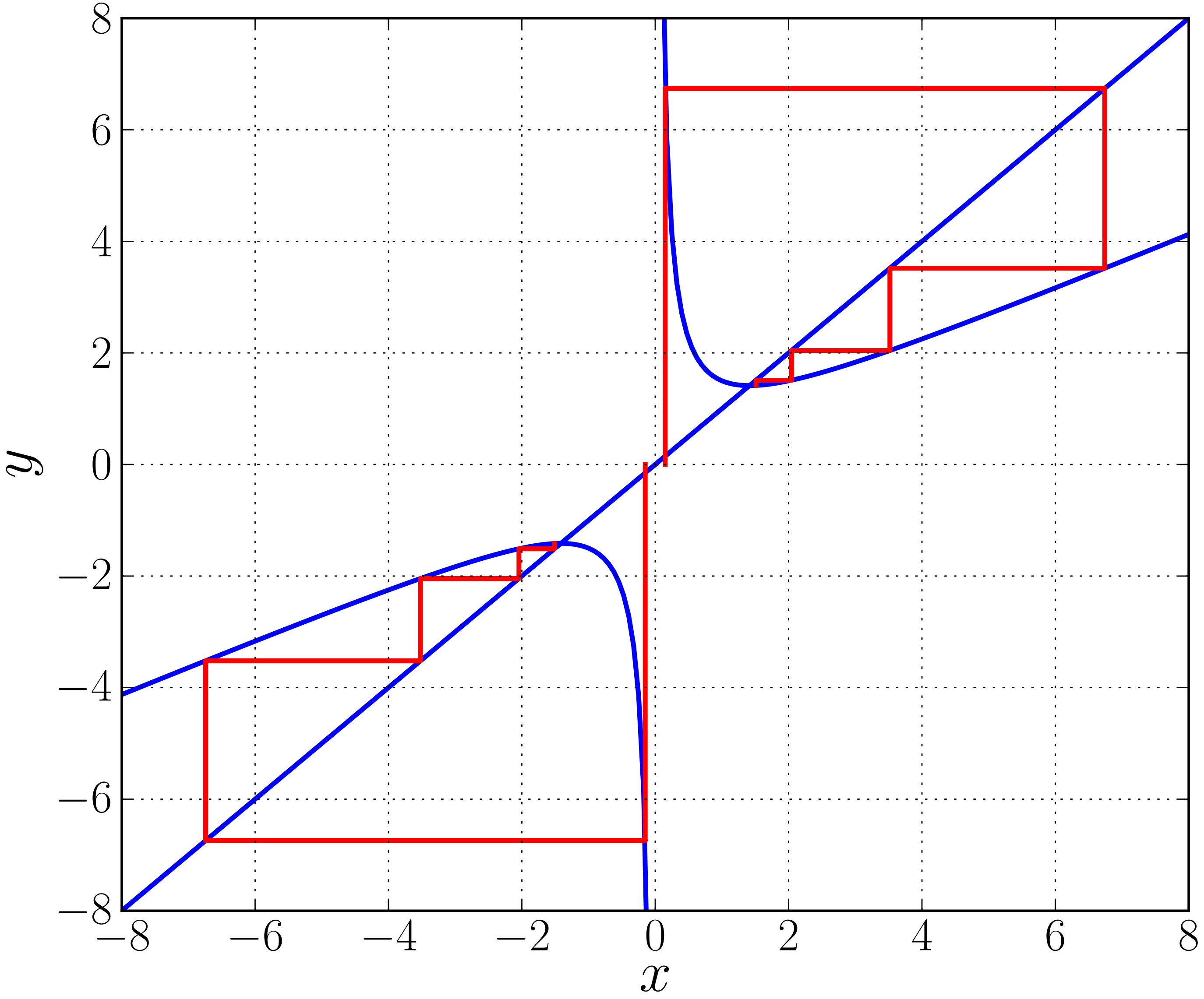
(**Check !**)



Newton's Method for $\sqrt{2}$ as a fixed point iteration



Newton's Method for $\sqrt{2}$ as a fixed point iteration (blow-up)



Newton's Method for $\pm\sqrt{2}$ as a fixed point iteration

From the last graph we see that :

- The iteration converges to $x^* = +\sqrt{2}$ for any $x^{(0)} > 0$.

- The iteration converges to $x^* = -\sqrt{2}$ for any $x^{(0)} < 0$.

(**Check !**)

EXAMPLE :

Consider the fixed point iteration

$$x^{(k+1)} \;=\; x^{(k)} \;-\; \gamma \; g(x^{(k)}) \; ,$$

for computing a zero of $g(x) = 0$.

Indeed, a fixed point x^* satisfies

$$x^* \;=\; x^* \;-\; \gamma \; g(x^*) \; ,$$

that is,

$$g(x^*) \;=\; 0 \; . \qquad (\text{Assuming } \gamma \;\neq\; 0 \; .)$$

In this example

$$f(x) \ = \ x \ - \ \gamma \, g(x) \ .$$

A fixed point x^* is *attracting* if

$$| \, f'(x^*) \, | \ < \ 1 \ ,$$

i.e., if

$$| \, 1 \ - \ \gamma \, g'(x^*) \, | \ < \ 1 \ ,$$

i.e., if

$$-1 \ < \ 1 \ - \ \gamma \, g'(x^*) \ < \ 1 \ ,$$

i.e., if

$$-2 \ < \ - \gamma \, g'(x^*) \ < \ 0 \ ,$$

i.e., if

$$0 \ < \ \gamma \, g'(x^*) \ < \ 2 \ .$$

The convergence is *quadratic* if

$$f'(x^*) \ = \ 1 - \gamma \ g'(x^*) \ = \ 0 \ ,$$

that is, if

$$\gamma \ = \ \hat{\gamma} \ \equiv \ \frac{1}{g'(x^*)} \ .$$

Now x^* is unknown beforehand and therefore $\hat{\gamma}$ is also unknown.

However, after the kth iteration an approximation to $\hat{\gamma}$ is given by

$$\hat{\gamma} \ \cong \ \gamma_k \ \equiv \ \frac{1}{g'(x^{(k)})} \ .$$

This leads to the iteration

$$x^{(k+1)} \ = \ x^{(k)} - \gamma_k \ g(x^{(k)}) \ ,$$

where $\gamma_k = 1/g'(x^{(k)})$,

i.e., we have rediscovered *Newton's method* !

EXERCISES :

- If the following fixed point iteration converges, then what number will it converge to? Is the convergence quadratic?

$$x^{(k+1)} = f(x^{(k)}) \ , \qquad \text{where} \qquad f(x) = \frac{2x^3+3}{3x^2} \ .$$

- Analytically determine all fixed points of

$$x^{(k+1)} = 2(x^{(k)})^2 - 2x^{(k)} + 1 \ , \quad k = 0, 1, 2, \cdots \ .$$

  Are the fixed points attracting or repelling? If attracting, then is the convergence linear or quadratic? Also draw a graphical interpretation.

- Analytically determine all fixed points of $x^{(k+1)} = 2x^{(k)}(1 - x^{(k)})$. Are these attracting or repelling? If attracting then is the convergence linear or quadratic? Also give a graphical interpretation.

EXERCISES :

- Give a graphical interpretation of the fixed point iteration.
$$x^{(k+1)} \;=\; \sin(x^{(k)}) \; .$$
What are the fixed points? Does the derivative test give conclusive evidence whether the fixed point $x = 0$ is attracting or repelling? Based on the graphical interpretation, can one conclude whether $x = 0$ is attracting or repelling?

- Consider the fixed point iteration $x^{(k+1)} \;=\; f(x^{(k)})$, where
$$f(x) \;=\; \begin{cases} 2x \; , & \text{when} \;\; x \le \frac{1}{2} \; , \\[2ex] 2(1-x) \; , & \text{when} \;\; x > \frac{1}{2} \; . \end{cases}$$
Give an accurate graphical interpretation in the interval $[0,1]$, with $x^{(0)} \cong 0.1$, showing enough iterations to illustrate the behavior of this fixed point iteration. Analytically determine all fixed points, and for each fixed point determine whether it is attracting or repelling.

EXERCISES :

- Show how to use the Chord method to compute the cube root of 5.

  Carry out the first two iterations of the Chord method, using $x^{(0)} = 2$.

  Analytically determine all fixed points of this Chord iteration.

  For each fixed point, determine whether it is attracting or repelling.

- Draw the graph of $g(x) = x^3 - 2$, clearly showing its zero.

  Write down Newton's method for finding a zero of g , and simplify the expression for the Newton iteration as much as possible.

  Will Newton's method converge if the initial guess is sufficiently close?

  If yes, then what is the rate of convergence?

  Will Newton's method converge for *any positive* initial guess ?

  Will Newton's method converge for *negative* initial guesses ?

EXERCISES :

- Suppose you enter any number on a calculator and then keep pushing the *cosine* button. (Assume the calculator is in "radian-mode".)

  What will happen in the limit to the result shown in the display?

  Give a full mathematical explanation and a graphical interpretation.

  Do the same for $\sin(x)$ and $\tan(x)$.

- Consider the fixed point iteration

  $$x^{(k+1)} = x^{(k)} \left(1 - x^{(k)}\right) .$$

  Does the derivative test give conclusive evidence whether or not the fixed point $x = 0$ is attracting?

  Give a careful graphical interpretation of this fixed point iteration.

  What can you say about the convergence of the fixed point iteration?

EXERCISES :

- Consider the fixed point iteration

$$x^{(k+1)} = \frac{1}{\sqrt{x^{(k)}}} , \quad k = 0, 1, 2, \cdots .$$

Give a careful graphical interpretation of this fixed point iteration.

Determine all fixed points and whether they are attracting or repelling.

Does the iteration converge for all positive initial points $x^{(0)}$?

- Consider the fixed point iteration

$$x^{(k+1)} = f(x^{(k)}) , \quad \text{where} \quad f(x) = \frac{1 + x^2}{1 + x} .$$

Determine all fixed points and whether they are attracting or repelling.

If attracting determine whether the convergence is linear or quadratic.

Give a graphical interpretation of the first few iterations, with $x^{(0)} = 2$.

Convergence Analysis for Systems.

Again most iterative methods for solving

$$\mathbf{G}(\mathbf{x}) \ = \ \mathbf{0} \ ,$$

can be written as

$$\mathbf{x}^{(k+1)} \ = \ \mathbf{F}(\mathbf{x}^{(k)}) \ , \qquad k =, 1, 2, \cdots ,$$

where the function $\mathbf{F}$ should be chosen such that

$$\mathbf{x}^* \text{ is a root of } \mathbf{G}(\mathbf{x}) = \mathbf{0} \quad \text{if and only if} \quad \mathbf{x}^* \text{ is a fixed point of } \mathbf{F}.$$

EXAMPLE :

Newton's method for systems is

$$\mathbf{G}'(\mathbf{x}^{(k)})\ (\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \ = \ - \mathbf{G}(\mathbf{x}^{(k)}) \ .$$

Thus

$$\mathbf{x}^{(k+1)} \ = \ \mathbf{x}^{(k)} - \mathbf{G}'(\mathbf{x}^{(k)})^{-1}\ \mathbf{G}(\mathbf{x}^{(k)}) \ ,$$

assuming $\mathbf{G}'(\mathbf{x})^{-1}$ to exist near $\mathbf{x} = \mathbf{x}^*$.

So here

$$\mathbf{F}(\mathbf{x}) \ = \ \mathbf{x} - \mathbf{G}'(\mathbf{x})^{-1}\ \mathbf{G}(\mathbf{x}) \ .$$

NOTE : Fixed point iterations also arise as models of physical processes, where they are called *difference equations* or *discrete dynamical systems* .

EXAMPLE :

The equations

$$\begin{aligned} x_1^{(k+1)} &= \lambda\, x_1^{(k)}(1 - x_1^{(k)}) \ - \ c_1\, x_1^{(k)}\, x_2^{(k)} \ , \\ x_2^{(k+1)} &= c_2\, x_2^{(k)} \ + \ c_1\, x_1^{(k)}\, x_2^{(k)} \ , \end{aligned}$$

model a "*predator-prey*" system, where, for example,

$x_1^{(k)}$ denotes the biomass of "fish" in year k ,

and

$x_2^{(k)}$ denotes the biomass of "sharks" in year k ,

and where λ , c_1 , and c_2 are constants.

Derivatives :

For scalar functions

$$f'(x) \equiv \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} ,$$

or, equivalently, $f'(x)$ is the number such that

$$\frac{f(x+h) - f(x) - f'(x)h}{h} \rightarrow 0 \quad \text{as} \quad h \rightarrow 0 .$$

If $f'(x)$ exists then f is said to be *differentiable* at x .

Similarly for vector valued functions $\mathbf{F}(\mathbf{x})$ we say that $\mathbf{F}$ is *differentiable* at $\mathbf{x}$ if there exists a *matrix* $\mathbf{F}'(\mathbf{x})$ such that

$$\frac{\| \mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x})\mathbf{h} \|}{\| \mathbf{h} \|} \rightarrow 0 \quad \text{as} \quad \| \mathbf{h} \| \rightarrow 0 \, .$$

The matrix $\mathbf{F}'(\mathbf{x})$ is the *Jacobian matrix* introduced earlier :

If $\mathbf{F}(\mathbf{x})$ has component functions

$$(f^1(\mathbf{x}) \, , \, f^2(\mathbf{x}) \, , \, \cdots \, , \, f^n(\mathbf{x}))^T \, ,$$

and if

$$\mathbf{x} \equiv (x_1 \, , \, x_2 \, , \, \cdots \, , \, x_n)^T \, ,$$

then

$$\{\mathbf{F}'(\mathbf{x})\}_{i,j} \equiv \frac{\partial f^i}{\partial x_j} \, .$$

THEOREM :

Let $\mathbf{F}'(\mathbf{x})$ be continuous near a fixed point $\mathbf{x}^*$ of $\mathbf{F}(\mathbf{x})$ and

$$\| \mathbf{F}'(\mathbf{x}^*) \| \ < \ 1 \ ,$$

in some induced matrix norm.

Then the fixed point iteration

$$\mathbf{x}^{(k+1)} \ = \ \mathbf{F}(\mathbf{x}^{(k)}) \ , \qquad k = 0, 1, 2, \cdots ,$$

converges to $\mathbf{x}^*$ whenever the initial guess $\mathbf{x}^{(0)}$ is sufficiently close to $\mathbf{x}^*$.

NOTE :

It can be shown that sufficient condition for

$$\| \mathbf{F}'(\mathbf{x}^*) \| \; < \; 1 \, ,$$

in some matrix norm, is that

$$\mathrm{spr}(\mathbf{F}'(\mathbf{x}^*)) \; < \; 1 \, .$$

Here

$\mathrm{spr}(\mathbf{F}'(\mathbf{x}^*))$ is the *spectral radius* of $\mathbf{F}'(\mathbf{x}^*)$,

that is, the size of the largest eigenvalue of $\mathbf{F}'(\mathbf{x}^*)$.

NOTE : Eigenvalues may be complex numbers.

PROOF (of the Theorem) :

(Similar to the proof of the scalar case.)

Let $\alpha \equiv \| \mathbf{F}'(\mathbf{x}^*) \|$. Then $\alpha < 1$.

By definition of $\mathbf{F}'$, given any $\epsilon > 0$, in particular

$$\epsilon \equiv \frac{1-\alpha}{2} ,$$

there exists a $\delta > 0$ such that

$$\| \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^*) - \mathbf{F}'(\mathbf{x}^*) \ (\mathbf{x} - \mathbf{x}^*) \| \ \leq \ \epsilon \| \mathbf{x} - \mathbf{x}^* \| \ ,$$

whenever $\mathbf{x} \in \mathcal{B}_\delta(\mathbf{x}^*)$.

Here

$$\mathcal{B}_\delta(\mathbf{x}^*) \equiv \{\mathbf{x} \ : \ \| \mathbf{x} - \mathbf{x}^* \| \ \leq \ \delta\} \ .$$

Let $\mathbf{x}^{(0)} \in \mathcal{B}_\delta(\mathbf{x}^*)$. Then

$$\begin{aligned}
\| \mathbf{x}^{(1)} - \mathbf{x}^* \| &= \| \mathbf{F}(\mathbf{x}^{(0)}) - \mathbf{F}(\mathbf{x}^*) \| \\
&= \| \mathbf{F}(\mathbf{x}^{(0)}) - \mathbf{F}(\mathbf{x}^*) \ - \ \mathbf{F}'(\mathbf{x}^*) \ (\mathbf{x}^{(0)} - \mathbf{x}^*) \ + \ \mathbf{F}'(\mathbf{x}^*) \ (\mathbf{x}^{(0)} - \mathbf{x}^*) \| \\
&\le \| \mathbf{F}(\mathbf{x}^{(0)}) - \mathbf{F}(\mathbf{x}^*) - \mathbf{F}'(\mathbf{x}^*) \ (\mathbf{x}^{(0)} - \mathbf{x}^*) \| + \| \mathbf{F}'(\mathbf{x}^*) \ (\mathbf{x}^{(0)} - \mathbf{x}^*) \| \\
&\le \epsilon \| \mathbf{x}^{(0)} - \mathbf{x}^* \| \ + \ \alpha \| \mathbf{x}^{(0)} - \mathbf{x}^* \| \\
&\le (\epsilon + \alpha) \| \mathbf{x}^{(0)} - \mathbf{x}^* \| \\
&= (\frac{1-\alpha}{2} + \alpha) \| \mathbf{x}^{(0)} - \mathbf{x}^* \| \\
&= \frac{1+\alpha}{2} \| \mathbf{x}^{(0)} - \mathbf{x}^* \| \ \le \ \beta \ \delta \ ,
\end{aligned}$$

where

$$\beta \ \equiv \ \frac{1+\alpha}{2} \ < \ 1 \ .$$

Thus $\mathbf{x}^{(1)} \in \mathcal{B}_\delta(\mathbf{x}^*)$.

Since $\mathbf{x}^{(1)} \in \mathcal{B}_\delta(\mathbf{x}^*)$, we also have

$$\| \mathbf{x}^{(2)} - \mathbf{x}^* \| \quad \leq \quad \beta \, \| \mathbf{x}^{(1)} - \mathbf{x}^* \| \quad \leq \quad \beta^2 \, \delta \ .$$

Thus, since $\beta < 1$, we have $\mathbf{x}^{(2)} \in \mathcal{B}_\delta(\mathbf{x}^*)$.

Continuing in this fashion, we find

$$\| \mathbf{x}^{(k)} - \mathbf{x}^* \| \quad \leq \quad \beta^k \, \delta \ .$$

Thus, since $\beta < 1$, we see that $\mathbf{x}^{(k)}$ converges to $\mathbf{x}^*$. **QED !**

EXAMPLE : In Newton's method

$$\mathbf{F}(\mathbf{x}) \ = \ \mathbf{x} \ - \ (\mathbf{G}'(\mathbf{x})^{-1}) \ \mathbf{G}(\mathbf{x}) \ .$$

Hence

$$\mathbf{G}'(\mathbf{x})\mathbf{F}(\mathbf{x}) = \mathbf{G}'(\mathbf{x})\mathbf{x} - \mathbf{G}(\mathbf{x}) \ .$$

$$\Rightarrow \quad \mathbf{G}''(\mathbf{x})\mathbf{F}(\mathbf{x}) + \mathbf{G}'(\mathbf{x})\mathbf{F}'(\mathbf{x}) = \mathbf{G}''(\mathbf{x})\mathbf{x} + \mathbf{G}'(\mathbf{x}) - \mathbf{G}'(\mathbf{x}) = \mathbf{G}''(\mathbf{x})\mathbf{x}$$

$$\Rightarrow \quad \mathbf{G}'(\mathbf{x})\mathbf{F}'(\mathbf{x}) = \mathbf{G}''(\mathbf{x})(\mathbf{x} - \mathbf{F}(\mathbf{x})) = \mathbf{G}''(\mathbf{x})(\mathbf{G}'(\mathbf{x}))^{-1}\mathbf{G}(\mathbf{x})$$

$$\Rightarrow \quad \mathbf{F}'(\mathbf{x}) \ = \ (\mathbf{G}'(\mathbf{x}))^{-1}\mathbf{G}''(\mathbf{x})(\mathbf{G}'(\mathbf{x}))^{-1}\mathbf{G}(\mathbf{x})$$

$$\Rightarrow \quad \mathbf{F}'(\mathbf{x}^*) \ = \ (\mathbf{G}'(\mathbf{x}^*))^{-1}\mathbf{G}''(\mathbf{x}^*)(\mathbf{G}'(\mathbf{x}^*))^{-1}\mathbf{G}(\mathbf{x}^*) = \mathbf{O} \ (\textit{zero matrix}) \ .$$

because $\mathbf{G}(\mathbf{x}^*) = \mathbf{0}$.

So $\parallel \mathbf{F}'(\mathbf{x}^*) \parallel = 0$, and therefore certainly $\parallel \mathbf{F}'(\mathbf{x}^*) \parallel < 1$.

Thus if

- $\mathbf{G}''(\mathbf{x})$ is continuous near $\mathbf{x}^*$,

- $(\mathbf{G}'(\mathbf{x}^*))^{-1}$ exists ,

- $\mathbf{x}^{(0)}$ is sufficiently close to $\mathbf{x}^*$,

then Newton's method converges.

Again this convergence can be shown to be *quadratic* , *i.e.*,

$$\lim_{k\to\infty} \frac{\parallel \mathbf{x}^{(k+1)} - \mathbf{x}^* \parallel}{\parallel \mathbf{x}^{(k)} - \mathbf{x}^* \parallel^2} \ \leq \ C \ , \qquad \text{for some constant } C.$$

EXERCISE :

- Consider the fixed point iteration

$$\begin{aligned} x_1^{(k+1)} &= \lambda\, x_1^{(k)}(1 - x_1^{(k)}) \ - \ 0.2\, x_1^{(k)}\, x_2^{(k)} \ , \\ x_2^{(k+1)} &= 0.9\, x_2^{(k)} \ + \ 0.2\, x_1^{(k)}\, x_2^{(k)} \ . \end{aligned}$$

This is a "*predator-prey*" model, where, for example, $x_1^{(k)}$ denotes the biomass of "fish" and $x_2^{(k)}$ denotes the biomass of "sharks" in year k .

Numerically determine the long-time behavior of $x_1^{(k)}$ and $x_2^{(k)}$ for the following values of λ :

$$\lambda \ = \ 0.5,\ 1.0,\ 1,5,\ 2.0,\ 2.5 \ ,$$

taking, for example, $x_1^{(0)} = 0.1$ and $x_2^{(0)} = 0.1$.

What can you say *analytically* about the fixed points of this system?

THE APPROXIMATION OF FUNCTIONS.

Function Norms.

To measure how well a given function $f \in \mathbb{C}[a,b]$ is approximated by another function we need a quantity called *function norm*.

Examples of these are :

$$\begin{aligned}
\| \, f \, \|_1 \quad &\equiv \quad \int_a^b \mid f(x) \mid \; dx \ , \\
\| \, f \, \|_2 \quad &\equiv \quad \{ \int_a^b f(x)^2 \; dx \}^{\frac{1}{2}} \ , \\
\| \, f \, \|_\infty \quad &\equiv \quad \max_{[a,b]} \mid f(x) \mid \quad .
\end{aligned}$$

Note the similarity of these norms to the corresponding vector norms.

A function norm must satisfy :

$$(i) \qquad \| \, f \, \| \;\; \geq \; 0 \, , \quad \forall f \in \mathbb{C}[a,b] \, , \qquad \| \, f \, \| \;\; = \;\; 0 \;\; \text{iff} \;\; f \; \equiv \; 0 \, ,$$

$$(ii) \qquad \| \, \alpha f \, \| \;\; = \;\; | \, \alpha \, | \;\; \| \, f \, \| \;\; , \quad \forall \alpha \in \mathbb{R} \, , \quad \forall f \in \mathbb{C}[a,b] \, ,$$

$$(iii) \qquad \| \, f + g \, \| \;\; \leq \;\; \| \, f \, \| \;\; + \;\; \| \, g \, \| \;\; , \quad \forall f, g \in \mathbb{C}[a,b] \, .$$

All of the norms above satisfy these requirements. (**Check !**)

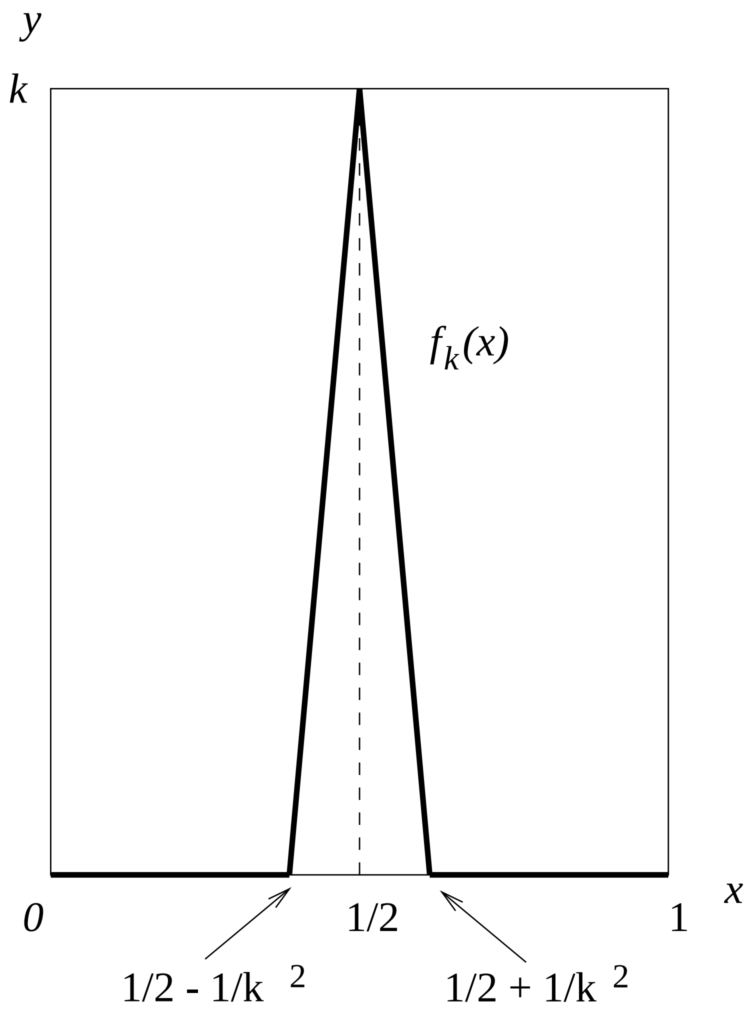
EXAMPLE :

$$
\begin{aligned}
\| \, f + g \, \|_1 \quad &= \quad \int_a^b \mid f(x) + g(x) \mid \; dx \\
&\leq \quad \int_a^b \mid f(x) \mid \; + \; \mid g(x) \mid \; dx \\
&= \quad \int_a^b \mid f(x) \mid \; dx \; + \; \int_a^b \mid g(x) \mid \; dx \\
&= \quad \| \, f \, \|_1 \; + \; \| \, g \, \|_1 \quad .
\end{aligned}
$$

(For the $\| \cdot \|_2$ we shall verify (iii) later.)

NOTE : If a function is "small" in a given function norm then it need not be small in another norm.

EXAMPLE : Consider $f_k(x)$, $k = 2, 3, \cdots$, as shown below :



Then

$$\| \, f_k \, \|_\infty \;=\; k \;\to\; \infty \quad \text{as} \quad k \to \infty \; ,$$

while

$$\| \, f_k \, \|_1 \;=\; \int_0^1 | \, f_k(x) \, | \;\; dx \;=\; \frac{1}{k} \;\to\; 0 \quad \text{as} \quad k \to \infty \; ,$$

and

$$\| \, f_k \, \|_2 \;=\; \{ \int_0^1 f_k(x)^2 \; dx \}^{\frac{1}{2}} \;=\; \sqrt{2/3} \qquad (\textbf{Check !}) \; .$$

EXAMPLE :

Approximate

$$f(x) \;=\; x^3$$

by

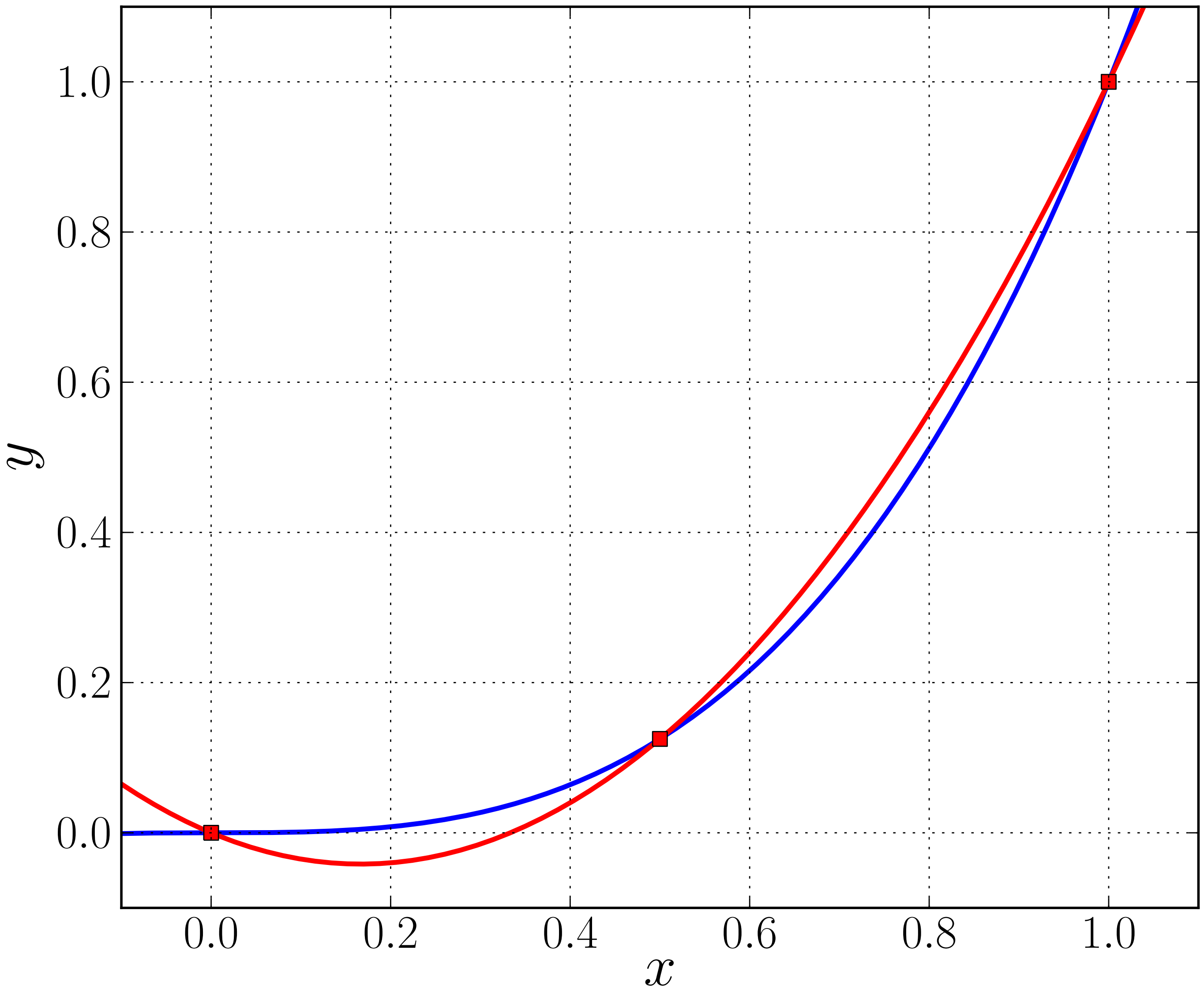
$$p(x) \;=\; \frac{3}{2}\,x^2 \;-\; \frac{1}{2}\,x \ ,$$

on the interval $[0, 1]$.

Then

$$p(x) \;=\; f(x) \qquad \text{for} \quad x = 0 \ , \ \frac{1}{2} \ , \ 1 \ ,$$

that is, $p(x)$ *interpolates* $f(x)$ at these points.



Graph of $f(x) = x^3$ (blue) and its interpolant $p(x) = \frac{3}{2}x^2 - \frac{1}{2}x$ (red) .

A measure of "how close" f and p are is then given by, for example,

$$\| f - p \|_2 \ = \ \{ \int_0^1 (f(x) - p(x))^2 \ dx \}^{\frac{1}{2}} \ .$$

We find that

$$\| f - p \|_2 \ = \ \frac{\sqrt{210}}{420} \ \cong \ 0.0345. \qquad\qquad (\textbf{Check !})$$

The Lagrange Interpolation Polynomial.

Let f be a function defined on $[a, b]$.

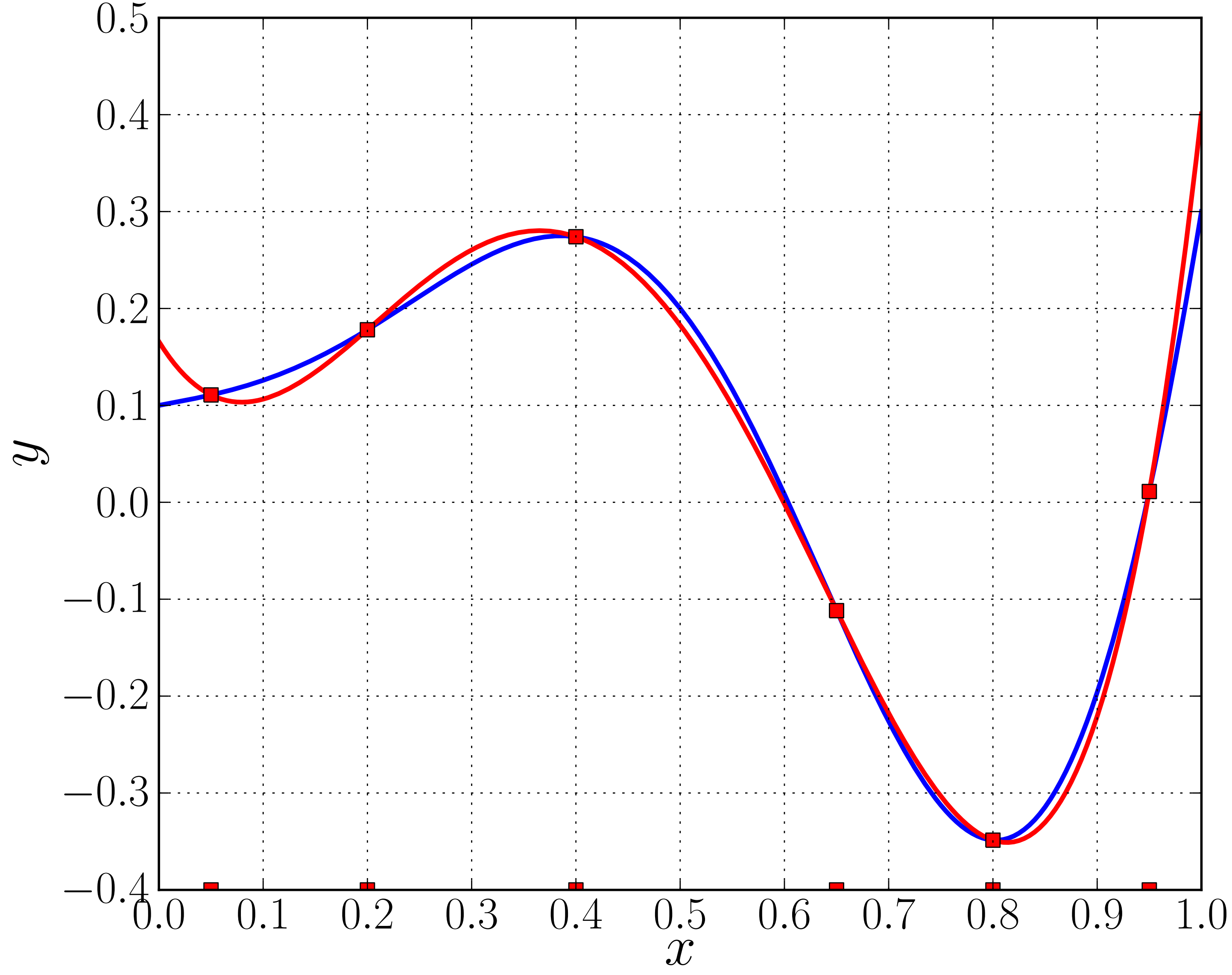
Let $\mathbb{P}_n$ denote all polynomials of degree less than or equal to n .

Given points $\{x_k\}_{k=0}^n$ with

$$a \leq x_0 < x_1 < \cdots < x_n \leq b ,$$

we want to find $p \in \mathbb{P}_n$ such that

$$p(x_k) = f(x_k) , \qquad k = 0, 1, \cdots, n .$$



Graph of $f(x) = \frac{1}{10} + \frac{1}{5}x + x^2 \sin(2\pi x)$ (blue)
and its Lagrange interpolant $p(x) \in \mathbb{P}_5$ (red)
at six interpolation points $(n = 5)$.

The following *questions* arise :

(i) Is $p(x)$ uniquely defined ?

(ii) How well does p approximate f ?

(iii) Does the approximation get better as $n \rightarrow \infty$?

To answer the above questions let

$$\ell_i(x) \equiv \prod_{k=0,k\neq i}^{n} \frac{(x - x_k)}{(x_i - x_k)} , \qquad i = 0, 1, \cdots, n ,$$

be the *Lagrange interpolating coefficients*, or *Lagrange basis functions*.

Then each $\ell_i \in \mathbb{P}_n$. (**Check !**)

EXAMPLE : If $n = 2$ we have

$$\ell_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} ,$$

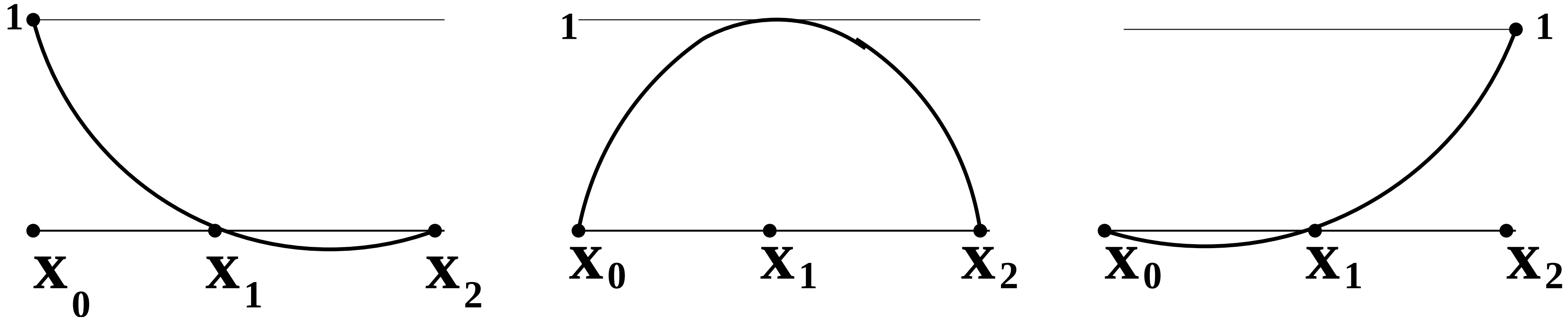
$$\ell_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} ,$$

and

$$\ell_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} .$$

NOTE : $\ell_i \in \mathbb{P}_2$, $i = 0, 1, 2$, and

$$\ell_i(x_k) = \begin{cases} 0 & \text{if} \quad k \neq i , \\ 1 & \text{if} \quad k = i . \end{cases}$$



$$\ell_0(x) \; = \; \tfrac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \; , \quad \ell_1(x) \; = \; \tfrac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \; , \quad \ell_2(x) \; = \; \tfrac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \; .$$

Lagrange basis functions (case $n \; = \; 2$) .

Now given $f(x)$ let

$$p(x) \ = \ \sum_{k=0}^{n} f(x_k) \ \ell_k(x) \ .$$

Then

$$p \ \in \ \mathbb{P}_n \ ,$$

and

$$p(x_i) \ = \ \sum_{k=0}^{n} f(x_k) \ \ell_k(x_i) \ = \ f(x_i) \ ,$$

that is, $p(x)$ *interpolates* $f(x)$ at the points $x_0, x_1, \cdots, x_n$.

THEOREM : Let $f(x)$ be defined on $[a, b]$ and let

$$a \leq x_0 < x_1 < \cdots < x_n \leq b .$$

Then there is a *unique* polynomial $p \in \mathbb{P}_n$ that interpolates f at the $\{x_k\}_{k=0}^n$.

PROOF :

We have already demonstrated the existence of $p(x)$.

Suppose $q \in \mathbb{P}_n$ also interpolates f at the points $\{x_k\}_{k=0}^n$.

Let $r(x) \equiv p(x) - q(x)$.

Then $r \in \mathbb{P}_n$ and $r(x_k) = 0$, $k = 0, 1, \cdots, n$.

But $r \in \mathbb{P}_n$ can have at most n zeroes, unless $r(x) \equiv 0$.

Hence $r(x) \equiv 0$.

Thus $p(x) \equiv q(x)$, *i.e.*, p is unique. **QED !**

EXAMPLE : Let $f(x) = e^x$.

Given $f(0) = 1$, $f(1) = 2.71828$, $f(2) = 7.38905$, we want to approximate $f(1.5)$ by polynomial interpolation at $x = 0, 1, 2$.

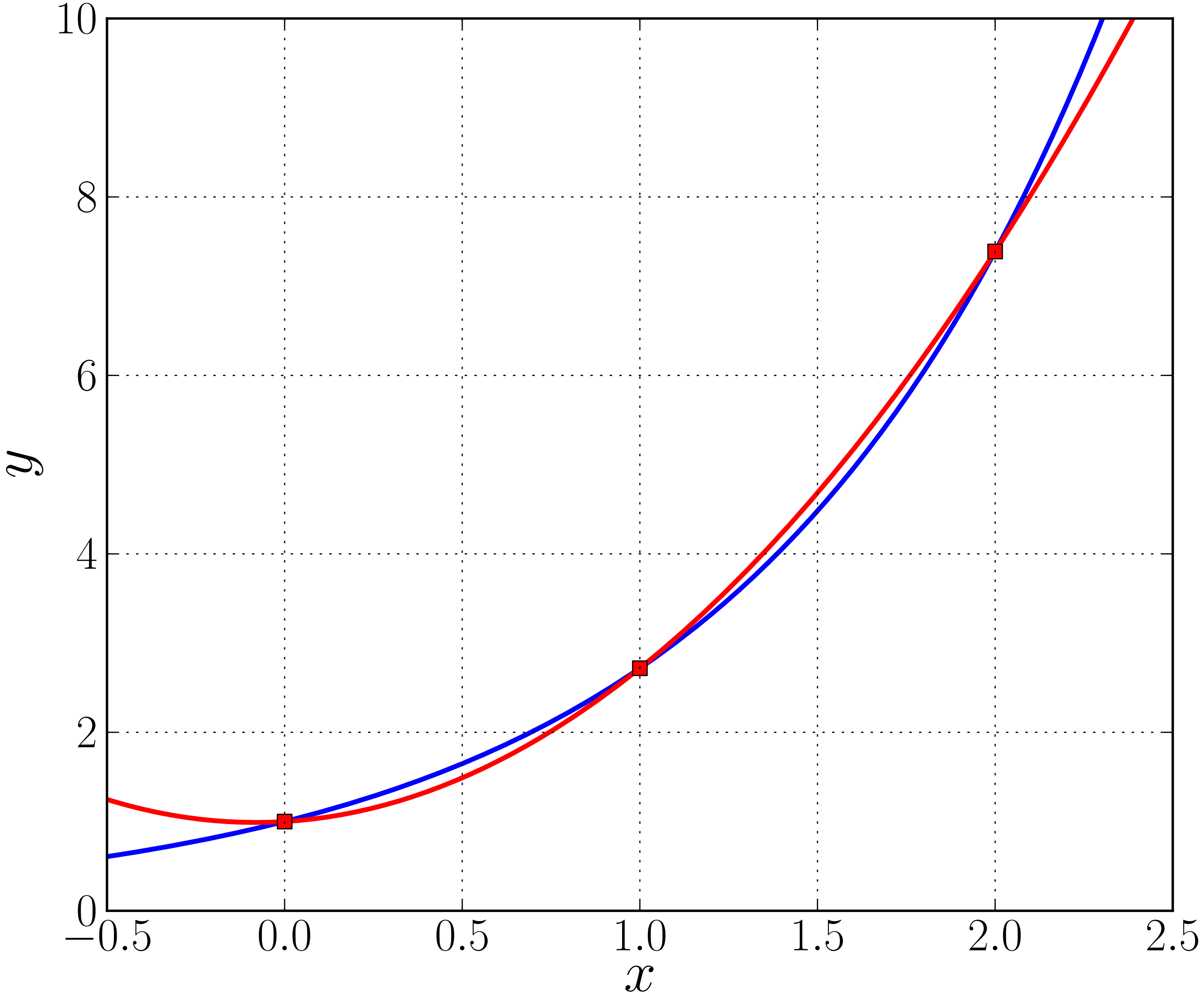
Here

$$\begin{aligned}
\ell_0(1.5) &= \frac{(1.5-1)\ (1.5-2)}{(0-1)\ (0-2)} = -\frac{1}{8}\ , \\
\ell_1(1.5) &= \frac{(1.5-0)\ (1.5-2)}{(1-0)\ (1-2)} = \frac{6}{8}\ , \\
\ell_2(1.5) &= \frac{(1.5-0)\ (1.5-1)}{(2-0)\ (2-1)} = \frac{3}{8}\ ,
\end{aligned}$$

so that

$$\begin{aligned}
p(1.5) &= f(0)\ \ell_0(1.5) + f(1)\ \ell_1(1.5) + f(2)\ \ell_2(1.5) \\
&= (1)\ (-\frac{1}{8}) + (2.71828)\ (\frac{6}{8}) + (7.38905)\ (\frac{3}{8}) = 4.68460\ .
\end{aligned}$$

The exact value is $f(1.5) = e^{1.5} = 4.48168$.



Graph of $f(x) = e^x$ (blue) and its Lagrange interpolant $p(x) \in \mathbb{P}_2$ (red).

THE LAGRANGE INTERPOLATION THEOREM :

Let

$$x_0 \ < \ x_1 \ < \ \cdots \ < \ x_n \ , \qquad \text{and let} \qquad x \in \mathbb{R} \ .$$

Define

$$a \ \equiv \ \min\{x_0, x\} \qquad \text{and} \qquad b \ \equiv \ \max\{x_n, x\} \ .$$

Assume that $f \in \mathbb{C}^{n+1}[a,b]$.

Let $p \in \mathbb{P}_n$ be the unique polynomial that interpolates $f(x)$ at $\{x_k\}_{k=0}^n$.

Then

$$f(x) - p(x) \ = \ \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{k=0}^{n} (x - x_k) \ ,$$

for some point $\xi \ \equiv \ \xi(x) \in [a,b]$.

PROOF :

If $x = x_k$ for some k then the formula is clearly valid.

So assume that $x \neq x_k$, for $k = 0, 1, \cdots, n$.

Let

$$w(z) \equiv \prod_{k=0}^{n} (z - x_k) \qquad \text{and} \quad c(x) \equiv \frac{f(x) - p(x)}{w(x)} .$$

Then $c(x)$ is well defined since $w(x) \neq 0$.

We want to show that

$$c(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} . \qquad \text{(Why ?)}$$

Let

$$F(z) \;\equiv\; f(z) \;-\; p(z) \;-\; w(z)\; c(x) \;.$$

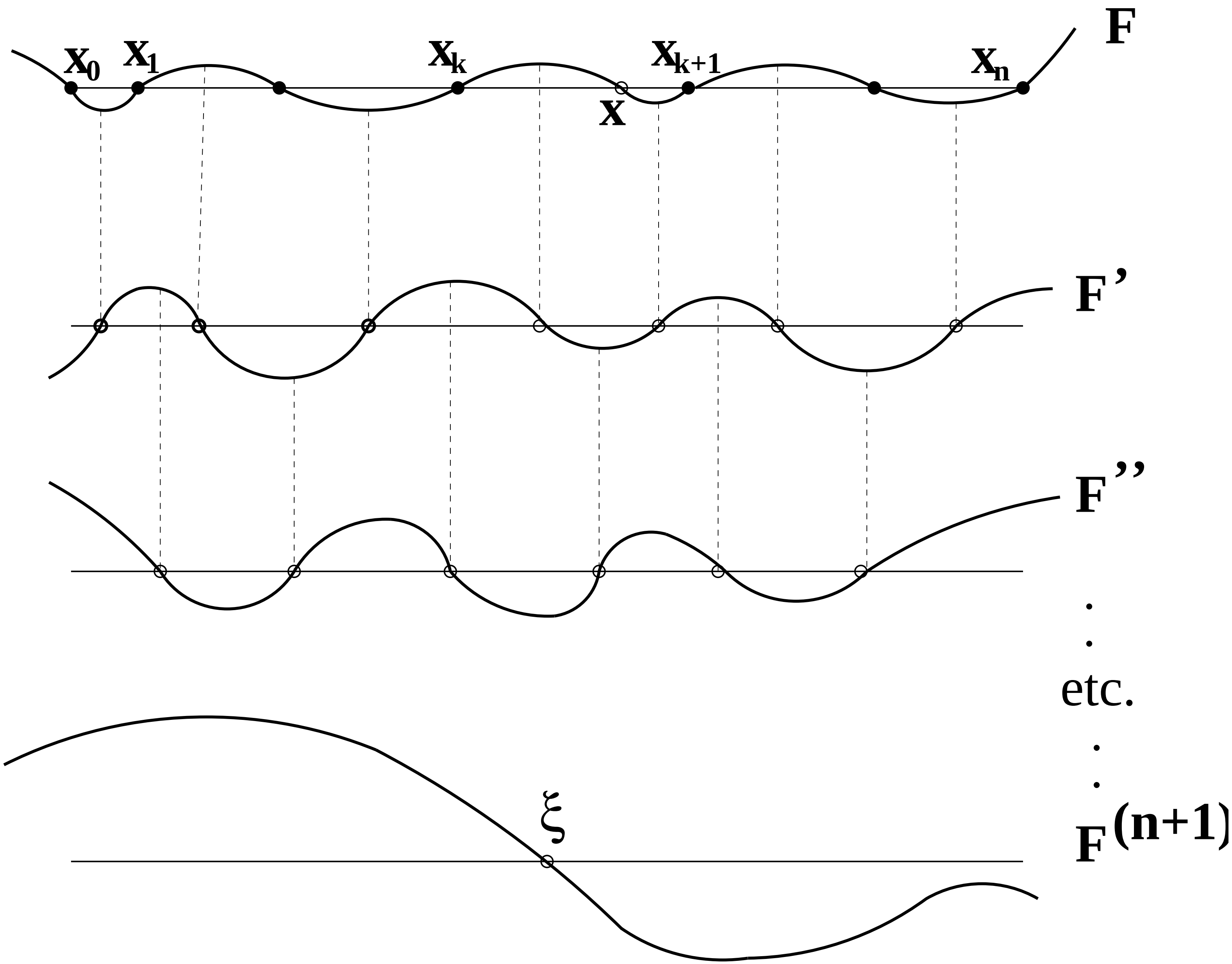
Then

$$F(x_k) \;=\; f(x_k) \;-\; p(x_k) \;-\; w(x_k)\; c(x) \;=\; 0, \qquad k = 0, 1, \cdots, n \;,$$

and

$$F(x) \;=\; f(x) \;-\; p(x) \;-\; w(x)\; \frac{f(x) - p(x)}{w(x)} \;=\; 0 \;.$$

Thus $F(z)$ has (at least) $n+2$ *distinct zeroes* in $[a, b]$.



The *zeroes* of $F(x)$ and of its derivatives.

Hence, by Rolle's Theorem, $F'(z)$ has $n+1$ distinct zeroes in $[a,b]$,

$F''(z)$ has n distinct zeroes in $[a,b]$,

$F'''(z)$ has $n-1$ distinct zeroes in $[a,b]$, *etc.*

We find that $F^{(n+1)}(z)$ has (at least) one zero in $[a,b]$, say,

$$F^{(n+1)}(\xi) \ = \ 0 \ , \qquad \xi \in [a,b] \ .$$

But

$$F^{(n+1)}(z) \ = \ f^{(n+1)}(z) \ - \ p^{(n+1)}(z) \ - \ w^{(n+1)}(z) \ c(x) \ .$$

Hence

$$F^{(n+1)}(\xi) \ = \ f^{(n+1)}(\xi) \ - \ (n+1)! \ c(x) \ = \ 0 \ .$$

It follows that

$$c(x) \ = \ \frac{f^{(n+1)}(\xi)}{(n+1)!} \ .$$

EXAMPLE : In the last example we had

$$n = 2 \ , \qquad f(x) = e^x \ , \qquad x_0 = 0 \ , \quad x_1 = 1, \quad x_2 = 2 \ ,$$

and we computed the value of $p(x)$ at $x = 1.5$.

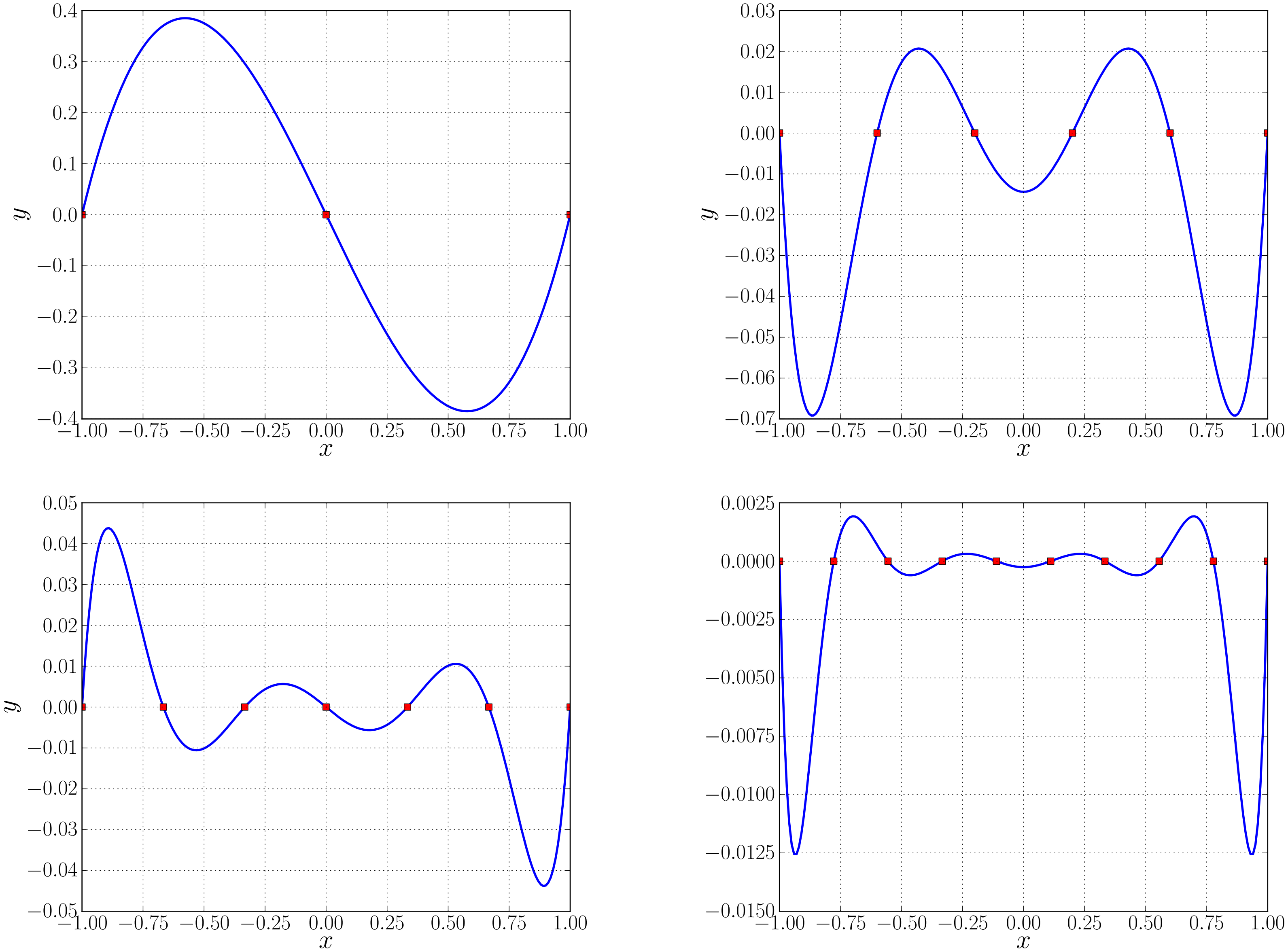
By the Theorem

$$f(x) - p(x) \ = \ \frac{f^{(3)}(\xi)}{3!} \ (x-0)(x-1)(x-2) \ , \qquad \xi \in [0,2] \ .$$

Since $f^{(3)}(\xi) \ \leq \ e^2 < 7.4$ we find that

$$|\ f(1.5) - p(1.5)\ | \ < \ \frac{7.4}{6} \ (1.5) \ (0.5) \ (0.5) \ = \ \frac{7.4}{16} \ < \ 0.47 \ .$$

The actual error is $|\ p(1.5) - e^{1.5}\ | \ \cong \ |\ 4.68460 - 4.48168\ | \ \cong \ 0.2$.



The graph of $w_{n+1}(x) = \prod_{k=0}^{n}(x - x_k)$ for equally spaced interpolation points in the interval $[-1, 1]$, for the cases $n+1 = 3$, 6 , 7 , 10 .

n	max	n	max	n	max	n	max
1	1.00000	5	0.06918	9	0.01256	13	0.00278
2	0.38490	6	0.04382	10	0.00853	14	0.00193
3	0.19749	7	0.02845	11	0.00584	15	0.00134
4	0.11348	8	0.01877	12	0.00400	16	0.00095

The maximum value of $\mid w_{n+1}(x) \mid$ in the interval $[-1, 1]$ for the case of $n+1$ equally spaced interpolation points .

EXERCISES :

- Consider the polynomial $p_n(x)$ of degree n or less that interpolates $f(x) = \sin(x)$ at $n+1$ distinct points in $[-1,1]$. Write down the general error formula for $|\sin(x) - p_n(x)|$. For distinct, but otherwise *arbitrary interpolation points*, how big should n be to guarantee that the maximum interpolation error in $[-1,1]$ is less than 10^{-2} ?

- Also answer the above question for *equally spaced interpolation points* in $[-1,1]$, using the Table on the preceding page .

- Also answer the above questions for the case of $f(x) = e^x$ in $[-1,1]$.

- Consider the problem of interpolating a smooth function $f(x)$ at two points, $x_0 = -h/2$ and $x_1 = h/2$, by a polynomial $p \in \mathbb{P}_3$ such that
$$p(x_0) = f(x_0), \quad p'(x_0) = f'(x_0), \quad p(x_1) = f(x_1), \quad p'(x_1) = f'(x_1).$$
Prove that this interpolation problem has one and only one solution.

Chebyshev Polynomials.

From the preceding Theorem it follows that

$$\| f - p \|_\infty \;\leq\; \frac{1}{(n+1)!} \; \| f^{(n+1)} \|_\infty \; \| w_{n+1} \|_\infty \; ,$$

where

$$w_{n+1}(x) \;\equiv\; \prod_{k=0}^{n} (x - x_k) \; ,$$

and where

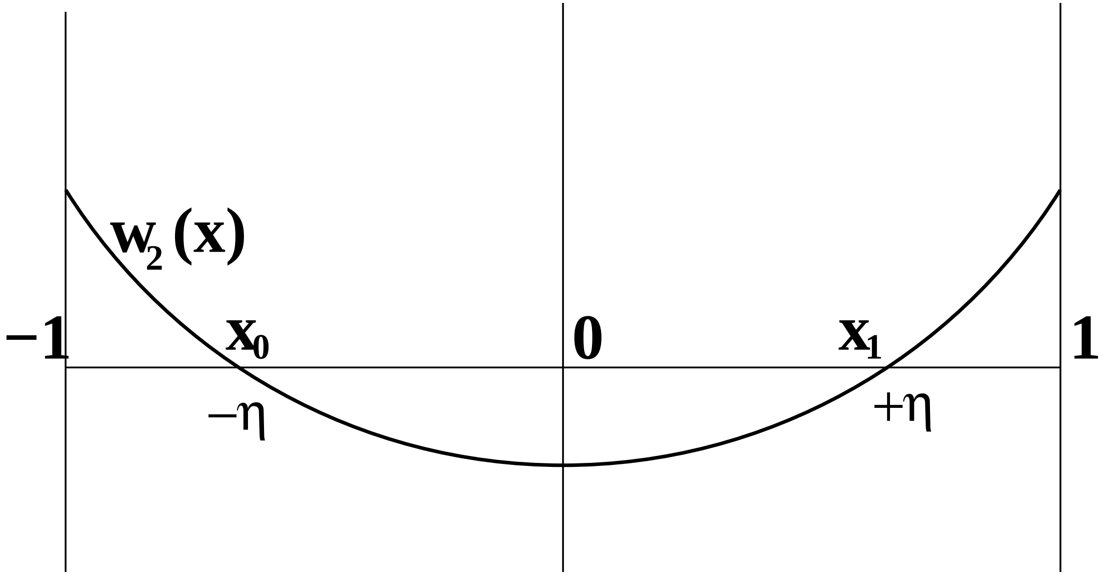
$$\| w_{n+1} \|_\infty \;\equiv\; \max_{[a,b]} | \, w_{n+1}(x) \, | \; .$$

NOTE :

- It does *not* follow that $\| f - p \|_\infty \;\to\; 0$ as $n \to \infty$.
- There are examples where $\| f - p \|_\infty \to \infty$ as $n \to \infty$.
- For given n, we can *choose* the points $\{x_k\}_{k=0}^{n}$ so $\| w_{n+1} \|_\infty$ is *minimized* .

EXAMPLE :

Let $n = 1$ and place the x_0 and x_1 symmetrically in $[-1, 1]$:



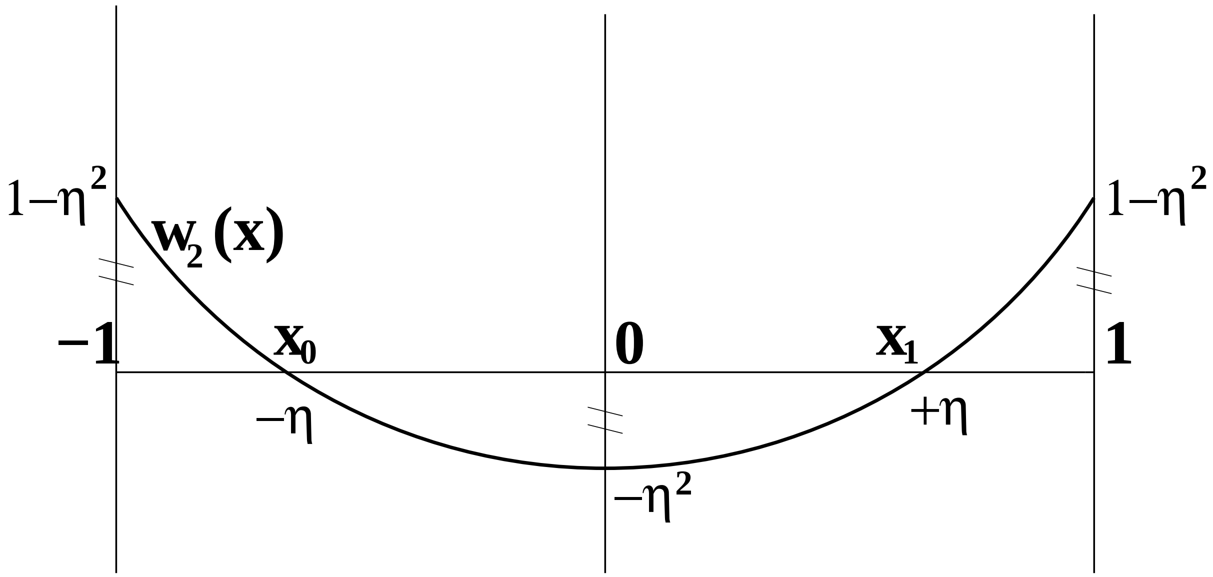
Then

$$w_2(x) \;=\; (x - x_0)\,(x - x_1) \;=\; (x + \eta)\,(x - \eta) \;=\; x^2 - \eta^2 \;.$$

We want to *choose* η such that

$$\|\, w_2 \,\|_\infty \;\equiv\; \max_{[-1,1]} |\, w_2(x) \,|$$

is *minimized* .



At the critical point : $w_2(0) \;=\; -\,\eta^2\ .$

At the endpoints : $w_2(-1) \;=\; w_2(1) \;=\; 1 - \eta^2\ .$

We see that $\|\, w_2 \,\|_\infty$ is *minimized* if we take η such that

$$|\, w_2(-1) \,| \;=\; |\, w_2(0) \,| \;=\; |\, w_2(1) \,| \ , \qquad i.e., \text{ if } \qquad \eta^2 \;=\; 1 - \eta^2\ ,$$

Thus $\qquad \eta \;=\; \frac{1}{2}\sqrt{2} \qquad$ and $\qquad \|\, w_2 \,\|_\infty \;=\; \frac{1}{2}\ .$

In general, the points

$$\{x_k\}_{k=0}^{n} \text{ that } \textit{minimize} \ \parallel w_{n+1} \parallel_\infty \text{ on } [-1,1]$$

are the zeroes of the *Chebyshev Polynomial* T_{n+1} of degree $n+1$.

These polynomials are defined as

$$T_k(x) \ \equiv \ \cos(\ k \cos^{-1}(x)\ )\ , \qquad k = 0, 1, 2, \cdots , \qquad x \in [-1,1]\ ,$$

where $\cos^{-1}(x)$ is used to denote $\arccos(x)$.

The T_k are indeed *polynomials* :

First $T_0(x) \ \equiv \ 1$ and $T_1(x) = x$.

Also

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

$$T_{k+1}(x) \ = \ 2x \ T_k(x) \ - \ T_{k-1}(x)$$

To derive this *recurrence formula* we use the identity

$$\cos(\ (k+1)\theta\ ) + \cos(\ (k-1)\theta\ ) \ = \ 2\cos(k\theta)\ \cos(\theta)\ .$$

which we rewrite as

$$\cos(\ (k+1)\theta\ ) \ = \ 2\cos(k\theta)\ \cos(\theta) \ - \ \cos(\ (k-1)\theta\ )\ .$$

so that, taking $\theta = \cos^{-1}(x)$, we have

$$\begin{aligned} T_{k+1}(x) \ &\equiv \ \cos(\ (k+1)\cos^{-1}(x)\ ) \\ &= \ 2\cos(\ k\cos^{-1}(x)\ )\ \cos(\cos^{-1}(x)) \ - \ \cos(\ (k-1)\cos^{-1}(x)\ ) \\ &= \ 2x\ T_k(x) \ - \ T_{k-1}(x)\ . \end{aligned}$$

Thus, with $T_0(x) \equiv 1$ and $T_1(x) = x$, we obtain

$$T_2(x) = 2x\, T_1(x) - T_0(x) = 2x^2 - 1 ,$$

$$T_3(x) = 2x\, T_2(x) - T_1(x) = 4x^3 - 3x ,$$

$$T_4(x) = 2x\, T_3(x) - T_2(x) = 8x^4 - 8x^2 + 1 ,$$

$$\begin{array}{c} \cdot \\ \cdot \\ \cdot \end{array}$$

$$T_{n+1}(x) = 2^n\, x^{n+1} + \cdots .$$

THE CHEBYSHEV THEOREM :

Let

$$w_{n+1}(x) \ \equiv \ \prod_{k=0}^{n} (x - x_k) \ .$$

Then for fixed n the value of

$$\| \, w_{n+1} \, \|_\infty \ \equiv \ \max_{[-1,1]} | \, w_{n+1}(x) \, |$$

is *minimized* if the points $\{x_k\}_{k=0}^{n}$ are the zeroes of $T_{n+1}(x)$.

For these points the value of $\| \, w_{n+1} \, \|_\infty$ is equal to 2^{-n} .

PROOF :

$$T_{n+1}(x) \;=\; \cos(\;(n+1)\cos^{-1}(x)\;)\;.$$

$$T_{n+1}(x) \;=\; 0 \qquad \text{if} \qquad (n+1)\cos^{-1}(x) \;=\; (2k+1)\frac{\pi}{2}\;,$$

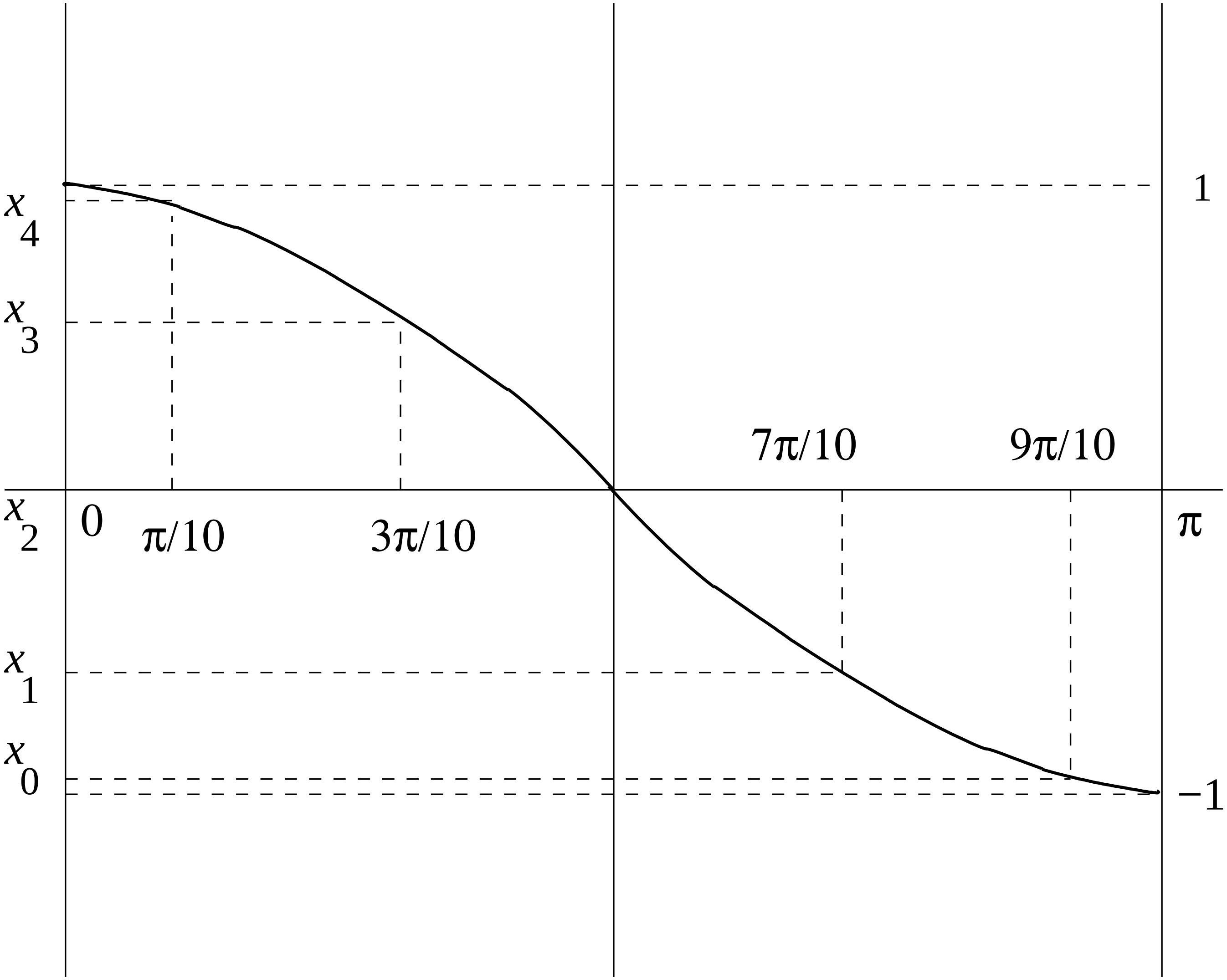
i.e., $T_{n+1}(x) \;=\; 0$ if

$$x \;=\; \cos(\frac{2k+1}{2(n+1)}\pi)\;, \qquad k = 0, 1, 2, \cdots, n\;.$$

Hence the zeroes of $T_{n+1}(x)$ lie indeed in $[-1, 1]$.

There are $n+1$ such zeroes.

$$x \;=\; \cos\left(\frac{2k+1}{2(n+1)}\pi\right)\ , \qquad k = 0, 1, 2, \cdots, n\ .$$



The *Chebyshev points* x_k , $(k = 0, 1, 2, \cdots, n)$, for the case $n = 4$.

$$T_{n+1}(x) \ = \ \cos(\ (n+1)\cos^{-1}(x)\ )\ .$$

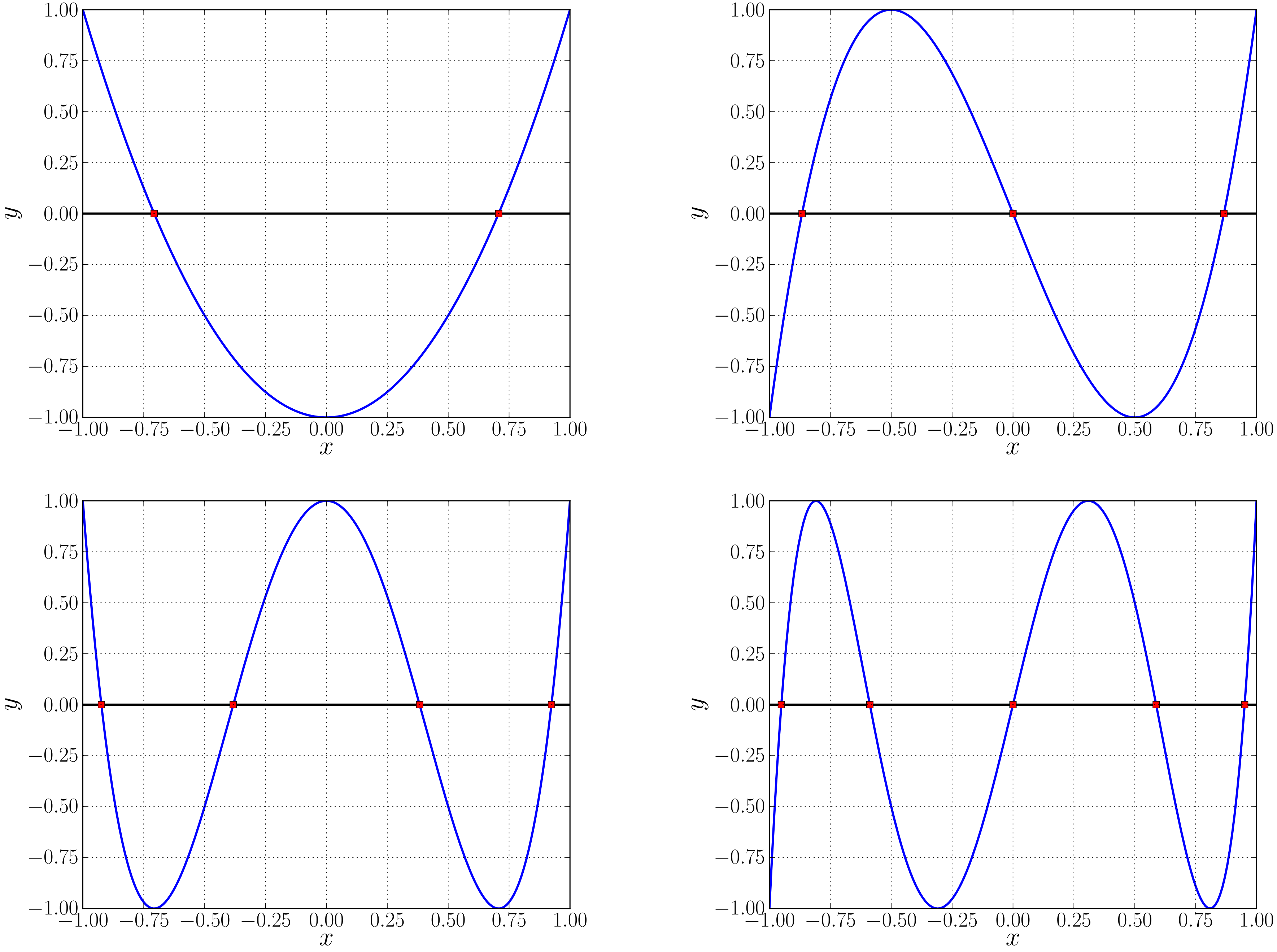
$T_{n+1}(x) \ = \ \pm 1$ if

$$(n+1)\cos^{-1}(x) \ = \ k\pi\ ,$$

that is, if,

$$x \ = \ \cos(\frac{k}{n+1}\pi)\ , \qquad k = 0, 1, 2, \ \cdots\ , n+1\ .$$

We can now draw the *graph* of T_{n+1} :



The graph of T_n for the cases $n = 2, 3, 4, 5$.

Recall that from the *recurrence relation*

$$T_0(x) \ \equiv \ 1 \ , \qquad T_1(x) = x \ , \quad T_{k+1}(x) \ = \ 2x\,T_k(x) \ - \ T_{k-1}(x) \ ,$$

we have

$$T_{n+1}(x) \ = \ 2^n \ x^{n+1} + \cdots \ .$$

Thus we can also write

$$T_{n+1}(x) \ = \ 2^n \ (x - x_0) \ (x - x_1) \ \cdots \ (x - x_n) \ ,$$

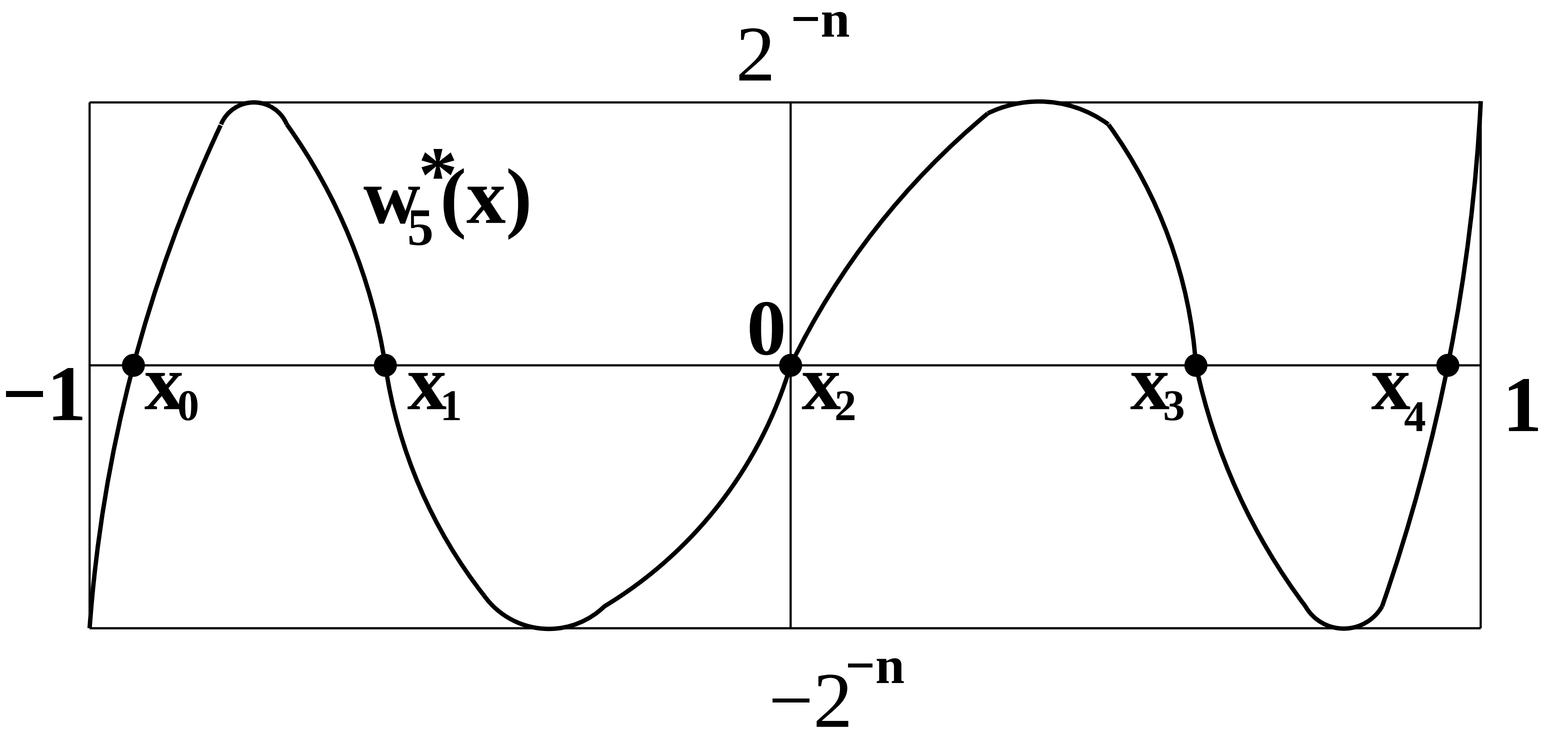
where the x_k are the zeroes of $T_{n+1}(x)$.

Let

$$w^*_{n+1}(x) \ \equiv \ 2^{-n} \ T_{n+1}(x) \ = \ (x - x_0) \ (x - x_1) \ \cdots \ (x - x_n) \ ,$$

Then

$$\| \ w_{n+1}{}^* \ \|_\infty \ = \ \| \ 2^{-n} T_{n+1} \ \|_\infty \ = \ 2^{-n} \ \| \ T_{n+1} \ \|_\infty \ = \ 2^{-n} \ .$$



The graph of $w_{n+1}^* = \prod_{k=0}^{n}(x - x_k)$ for the case $n = 4$.

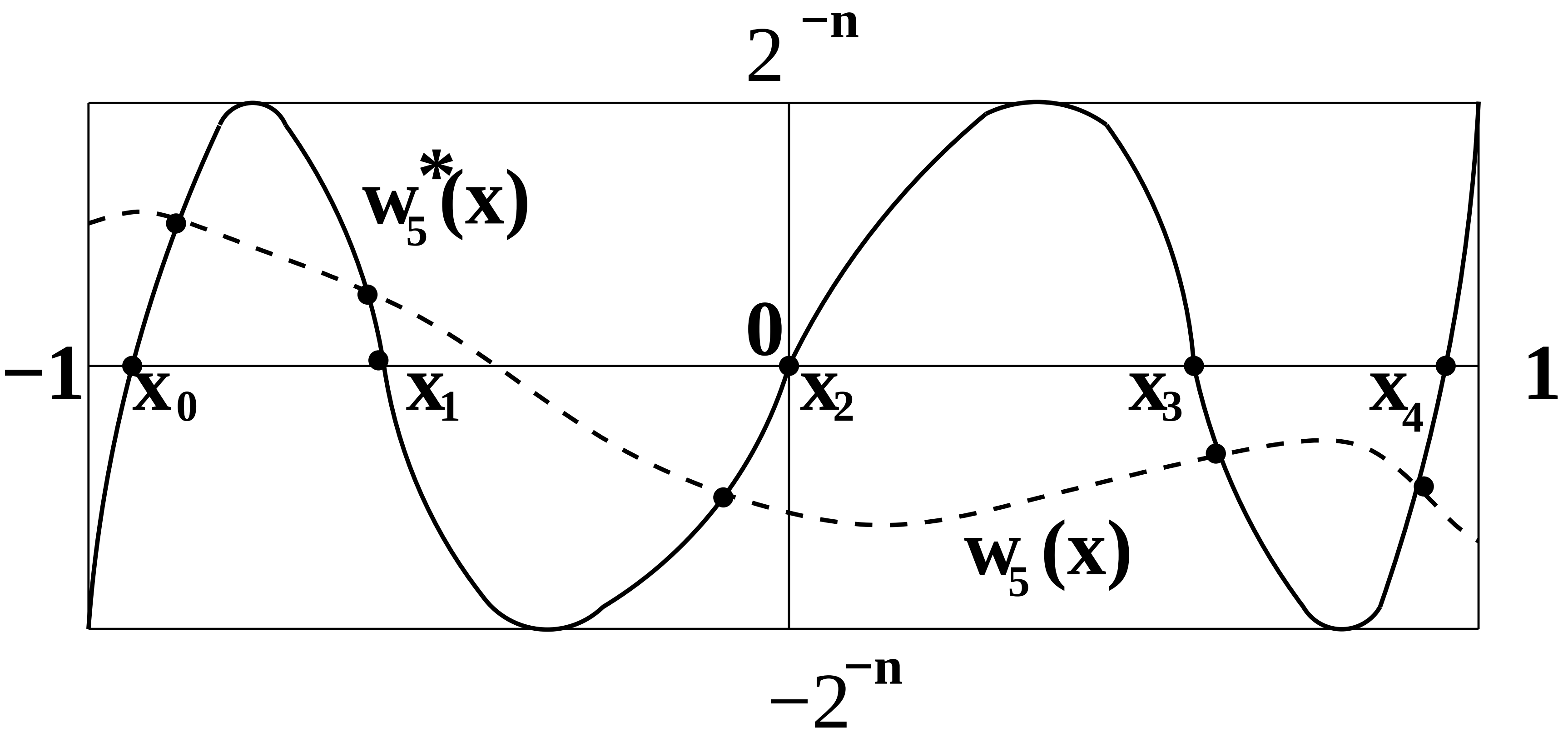
Claim :

There does not exist $w \in \mathbb{P}_{n+1}$, with leading coefficient 1, such that

$$\| \, w \, \|_\infty \ < \ 2^{-n} \ .$$

Suppose there does exist a $w_{n+1} \in \mathbb{P}_{n+1}$, with leading coefficient 1, such that

$$\| w_{n+1} \|_\infty \quad < \quad \| {w_{n+1}}^* \|_\infty \ = \ 2^{-n} \ .$$



Then w_{n+1} must intersect ${w_{n+1}}^*$ at least $n+1$ times in $[-1,1]$.

Thus $(w_{n+1} - {w_{n+1}}^*)$ has $n+1$ zeroes in $[-1,1]$.

But

$$(w_{n+1} - {w_{n+1}}^*) \in \mathbb{P}_n$$

since both w_{n+1} and ${w_{n+1}}^*$ have leading coefficient 1 .

Hence $w_{n+1} - {w_{n+1}}^* \ \equiv \ 0$.

Thus $w_{n+1} \ = \ {w_{n+1}}^*$. **QED !**

n	uniform	Chebyshev	n	uniform	Chebyshev
1	1.00000	0.50000	5	0.06918	0.03125
2	0.38490	0.25000	6	0.04382	0.01563
3	0.19749	0.12500	7	0.02845	0.00782
4	0.11348	0.06250	8	0.01877	0.00391

The maximum of $\mid w_{n+1}(x) \mid$ in the interval $[-1, 1]$
for uniformly spaced points and for Chebyshev points .

EXAMPLE :

Let $f(x) = e^x$ on $[-1,1]$ and take $n = 2$.

$T_3(x) = 4x^3 - 3x$ has zeroes

$$x_0 = -\frac{1}{2}\sqrt{3} , \qquad x_1 = 0 , \qquad x_2 = \frac{1}{2}\sqrt{3} .$$

Approximate $f(0.5)$ by polynomial interpolation at $x_0, \ x_1, \ x_2$:

$$\ell_0(0.5) = \frac{(0.5 - x_1)(0.5 - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{1 - \sqrt{3}}{6} ,$$

$$\ell_1(0.5) = \frac{(0.5 - x_0)(0.5 - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{4}{6} ,$$

$$\ell_2(0.5) = \frac{(0.5 - x_0)(0.5 - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{1 + \sqrt{3}}{6} .$$

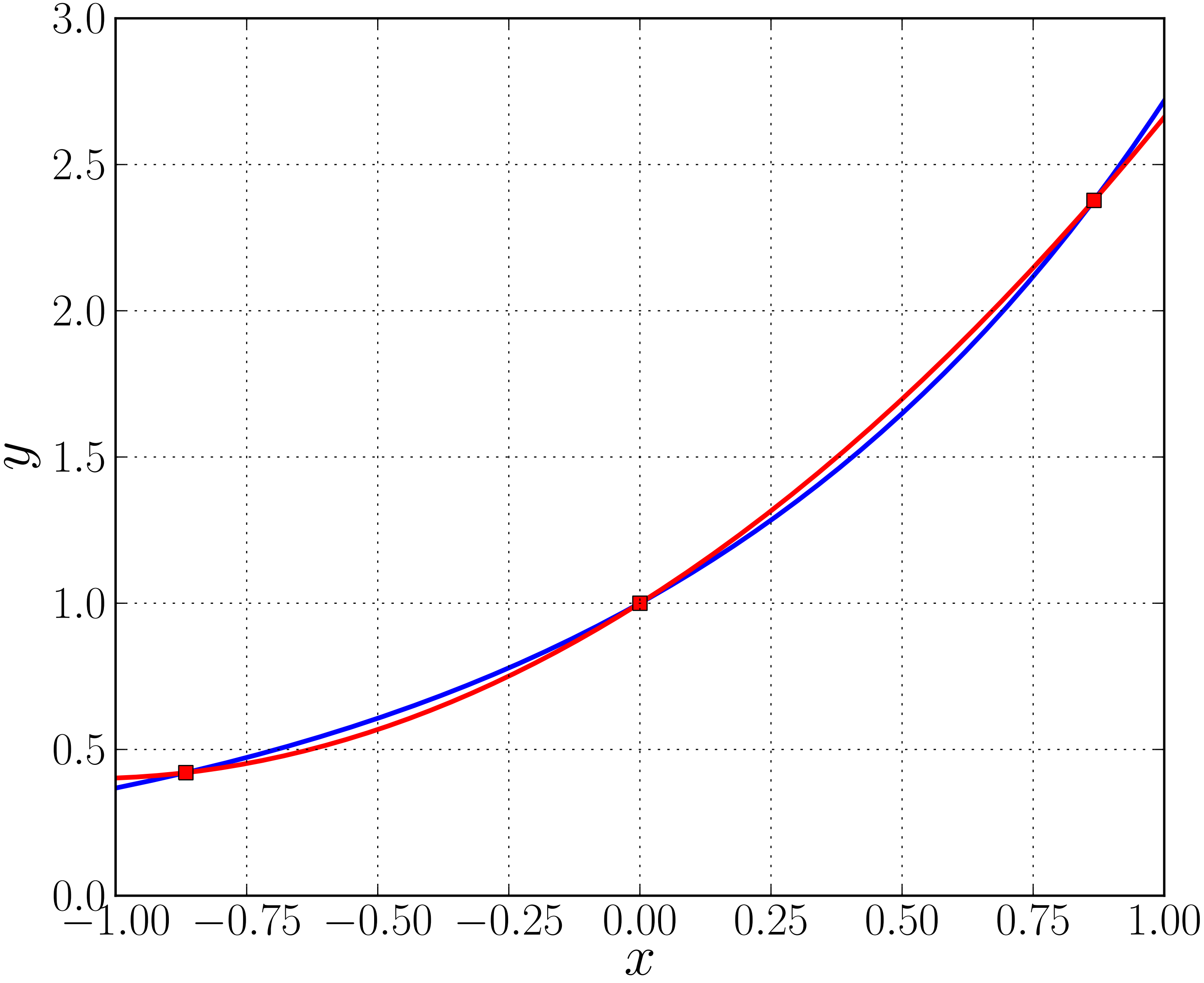
Thus

$$\begin{aligned} p(0.5) &= f(x_0)\ l_0(0.5)\ +\ f(x_1)\ l_1(0.5)\ +\ f(x_2)\ l_2(0.5) \\ &= e^{(-0.5\sqrt{3})}\ \frac{(1-\sqrt{3})}{6}\ +\ e^0\ (\frac{4}{6})\ +\ e^{(0.5\sqrt{3})}\ \frac{(1+\sqrt{3})}{6} \\ &\cong 1.697\ . \end{aligned}$$

The exact value is $e^{0.5} = 1.648 \cdots$.

Thus the exact absolute error is

$$|\ e^{0.5} - p(0.5)\ |\ \cong\ |\ 1.648 - 1.697\ |\ =\ 0.049\ .$$



Graph of $f(x) = e^x$ (blue) on the interval $[-1, 1]$,

and its Lagrange interpolating polynomial $p(x) \in \mathbb{P}_2$ (red)

at three Chebyshev interpolation points $(n = 2)$.

EXAMPLE : More generally, if we interpolate

$$f(x) = e^x \quad \text{by} \quad p \in \mathbb{P}_n \quad \text{at} \quad n+1 \quad \text{Chebyshev points in} \quad [-1,1] \ ,$$

then for $x \in [-1,1]$ we have

$$| \ e^x \ - \ p(x) \ | \quad = \quad | \ \frac{f^{(n+1)}(\xi)}{(n+1)!} \ w_{n+1}(x) \ | \quad ,$$

where $\xi \ \equiv \ \xi(x) \ \in \ [-1,1]$, and where

$$w_{n+1}(x) \ = \ \prod_{k=0}^{n} (x - x_k) \ , \qquad (\text{ Chebyshev points } x_k \ ),$$

Thus

$$\begin{aligned} \max_{x \in [-1,1]} | \ e^x - p(x) \ | \quad & \leq \quad \frac{\| \ f^{(n+1)} \ \|_\infty}{(n+1)!} \ \| \ w_{n+1} \ \|_\infty \\ & \leq \quad \frac{e}{(n+1)!} \ \| \ w_{n+1} \ \|_\infty \\ & = \quad \frac{e}{(n+1)!} \ 2^{-n} \ . \end{aligned}$$

NOTE :

Let f be a sufficiently smooth function.

Let p_U be the polynomial that interpolates f at $n{+}1$ uniformly spaced points.

Let p_C denote the polynomial that interpolates f at $n+1$ Chebyshev points.

Although the Theorem does not *guarantee* that

$$\| \, p_C - f \, \|_\infty \quad \leq \quad \| \, p_U - f \, \|_\infty \quad ,$$

this inequality is "usually" valid.

EXERCISES :

- Suppose $p \in \mathbb{P}^n$ interpolates $\sin(x)$ at $n+1$ distinct points in $[-1,1]$. For the case of *Chebyshev points*, how big should n be for the error to be less than 10^{-4} everywhere in $[-1,1]$?

- Suppose that $p \in \mathbb{P}^n$ interpolates e^x at $n+1$ distinct points in $[-1,1]$. For the case of *Chebyshev points*, how big should n be for the error to be less than 10^{-4} everywhere in $[-1,1]$?

- Suppose $p \in \mathbb{P}^n$ interpolates x^{n+1} at $n+1$ distinct points in $[-1,1]$. For the case of *Chebyshev points*, how big should n be for the maximum interpolation error in $[-1,1]$ to be less than 10^{-4} ?

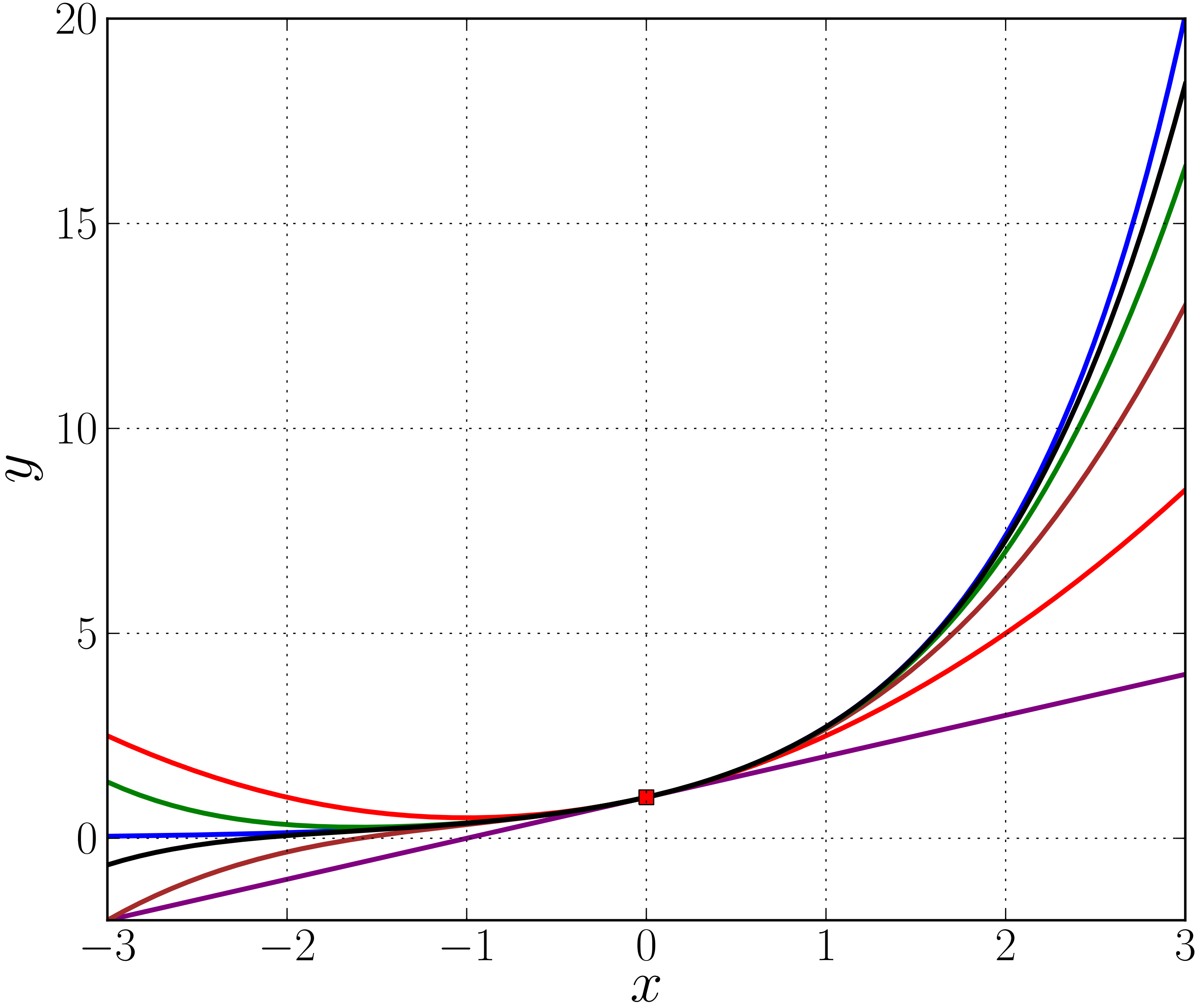
The Taylor Polynomial.

Let $f \in C^n[a,b]$.

Let $\mathbb{P}_n$ denote all polynomials of degree less than or equal to n.

Given the point $x_0 \in [a,b]$, we want to find $p \in \mathbb{P}_n$ such that

$$p^{(k)}(x_0) \ = \ f^{(k)}(x_0) \ , \qquad k \ = \ 0, 1, \cdots, n \ .$$



The function e^x (blue) and its *Taylor polynomials* $p_k(x)$ about $x_0 = 0$:
$k = 1$: purple, $k = 2$: red, $k = 3$: brown, $k = 4$: green, $k = 5$: black .

As for Lagrange interpolation, we have the following *questions* :

- Is the polynomial $p(x) \in \mathbb{P}_n$ uniquely defined ?

- How well does p approximate f ?

- Does the approximation get better as $n \to \infty$?

Existence :

$$p(x) \;=\; \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}\,(x-x_0)^k \;.$$

Clearly

$$p^{(k)}(x_0) \;=\; f^{(k)}(x_0)\;, \qquad k \;=\; 0,1,\cdots,n\;. \qquad (\textbf{Check !})$$

DEFINITION :

$p(x)$ is called the *Taylor polynomial* of degree n for $f(x)$ about the point x_0.

TAYLOR's THEOREM :

Let $f \in C^{n+1}[a,b]$, $x_0 \in [a,b]$.

Let $p(x) \in \mathbb{P}_n$ be the *Taylor polynomial* for f about the point x_0 , *i.e.*,

$$p(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k .$$

Then, for $x \in [a,b]$,

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1} ,$$

for some point $\xi = \xi(x)$ that lies between x_0 and x .

DEFINITION : $R_n(x) \equiv \dfrac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1}$

is called the *Taylor remainder* .

PROOF : (**EXERCISE**)

The steps are similar to those in the Lagrange Interpolation Theorem :

- First show that the Theorem holds if $x = x_0$.
- Next assume x is arbitrary, but $x \neq x_0$. (Consider x as fixed.)
- Define
$$c(x) \;=\; \frac{f(x) - p(x)}{(x - x_0)^{n+1}} \;.$$
- Define $\quad F(z) \;=\; f(z) \;-\; p(z) \;-\; c(x)\;(z - x_0)^{n+1}$.
- Show that $F^{(k)}(x_0) = 0$, $\quad k = 0, 1, \cdots, n$, and that $F(x) = 0$.
- Give a qualitative graph of $F(z)$.

- Show $F'(\xi_0) = 0$ for some ξ_0 between x_0 and x . Graph $F'(z)$.

- Show $F''(\xi_1) = 0$ for some ξ_1 between x_0 and ξ_0 . Graph $F''(z)$.

- *etc.*

- Show that $F^{(n+1)}(\xi_n) = 0$ for some ξ_n between x_0 and ξ_{n-1} .

- From this derive that

$$c(x) \;=\; \frac{f^{(n+1)}(\xi)}{(n+1)!} \;, \qquad (\xi = \xi_n) \;.$$

- Show how Taylor's Theorem follows from this last step. **QED !**

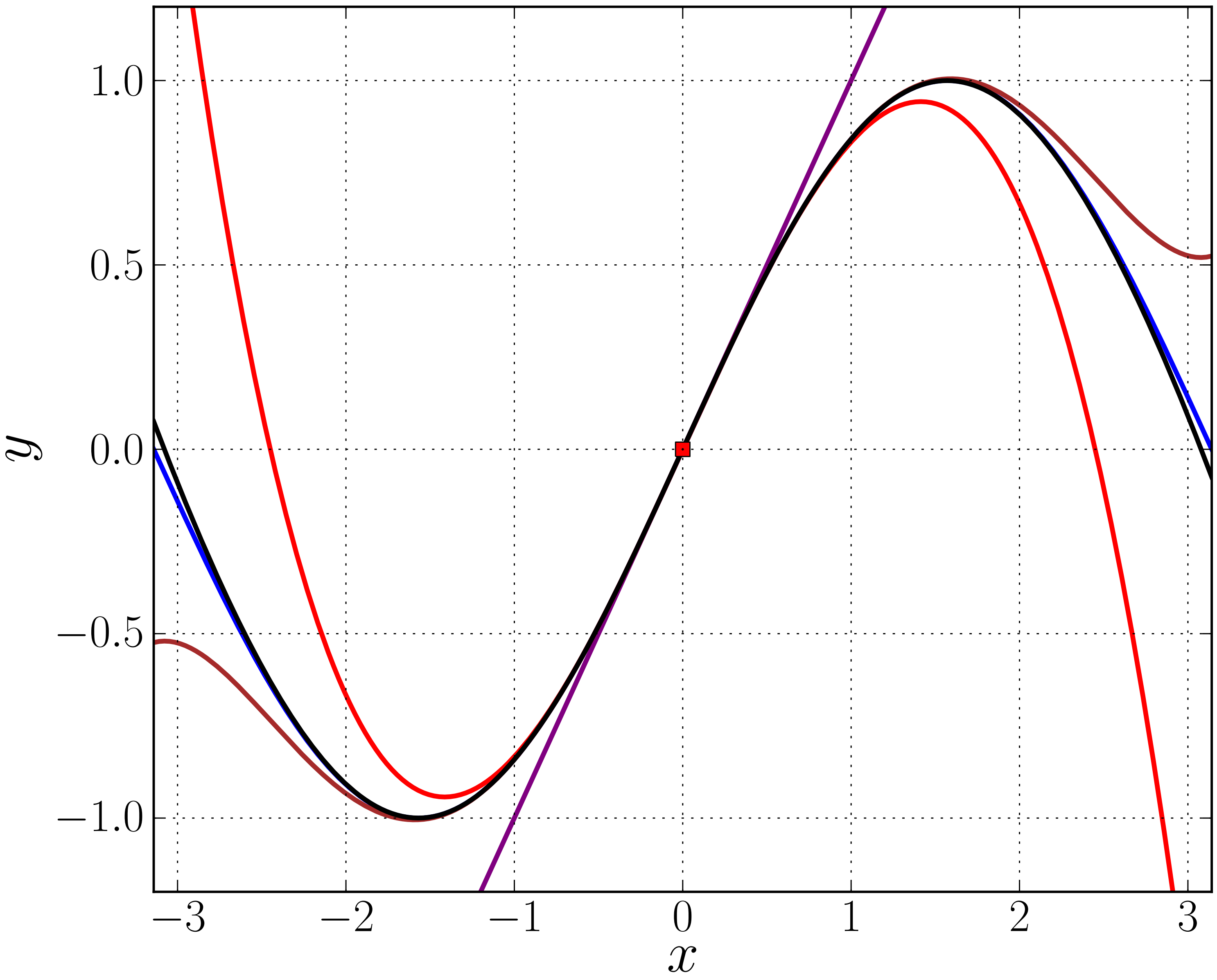
EXERCISES :

- Write down the Taylor polynomials $p_n(x)$ of degree n (or less) for $f(x) = e^x$ about the point $x_0 = 0$, for each of the following cases : $n = 1, 2, 3, 4$.

- How big should n be so that $\mid e^x - p_n(x) \mid < 10^{-4}$ everywhere in the interval $[-1, 1]$?

- Do the same for $f(x) = \sin(x)$ in $[0, 1]$ about the point $x_0 = 0$.

- Do the same for $f(x) = \ln(x)$ in $[\frac{1}{2}, \frac{3}{2}]$ about the point $x_0 = 1$.



Graph of the function $\sin(x)$ (blue) and its *Taylor polynomials* $p_k(x)$ about $x_0 = 0$: $k = 1$: purple, $k = 3$: red, $k = 5$: brown, $k = 7$: black .

Local Polynomial Interpolation.

Let $f \in \mathbb{C}^{n+1}[a,b]$.

Let $p \in \mathbb{P}_n$ interpolate f at $n+1$ distinct points $\{x_k\}_{k=0}^{n}$ in $[a,b]$.

Does

$$\| f - p \|_{\infty} \to 0 \quad \text{as} \quad n \to \infty \ ?$$

The answer is often **NO** !

EXAMPLE : If

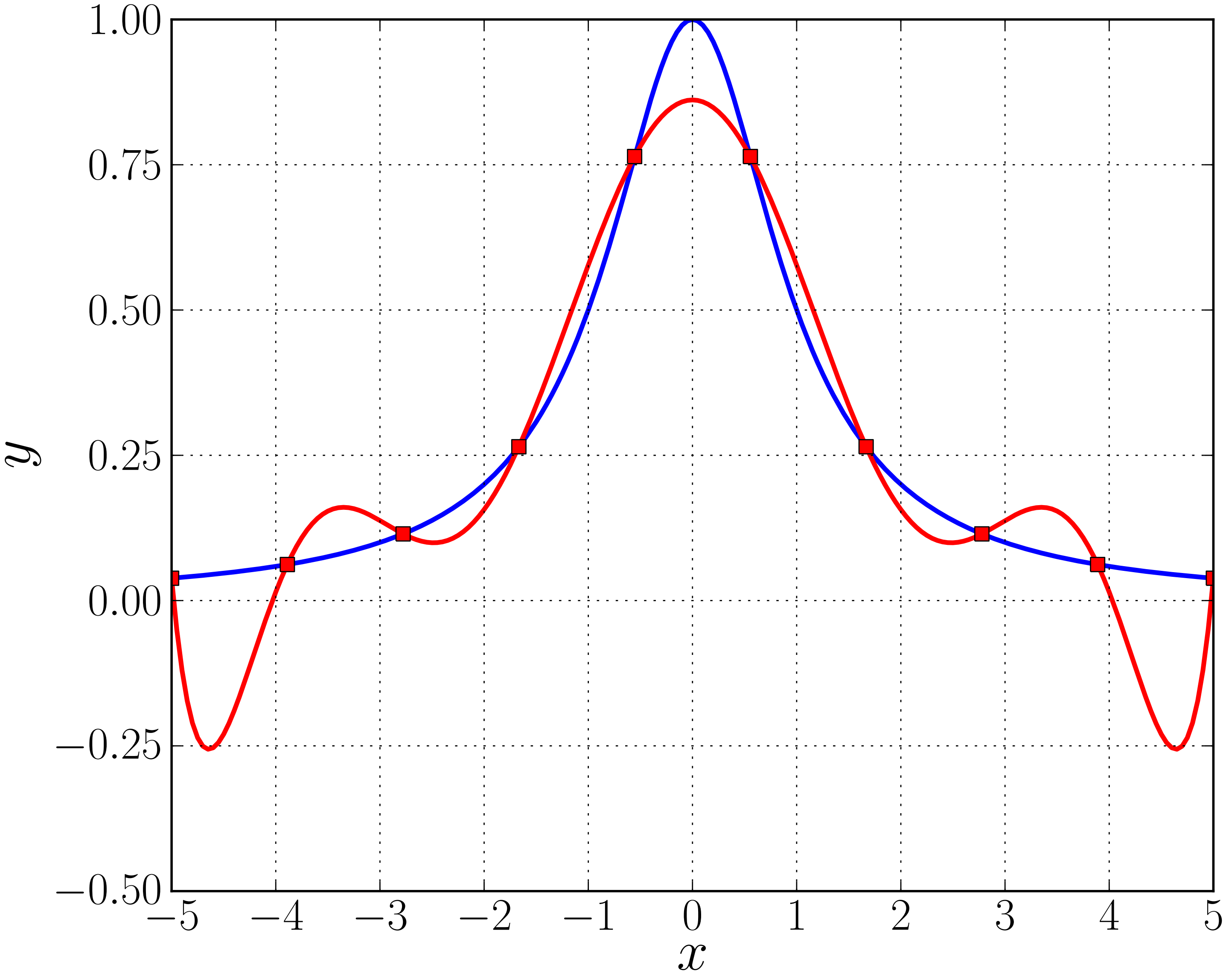
$$f(x) \ = \ \frac{1}{1 \ + \ x^2} \quad \text{on} \quad [-5,5] \ ,$$

and if $p \in \mathbb{P}_n$ interpolates f at the $n+1$ equally spaced points $\{x_k\}_{k=0}^n$ with

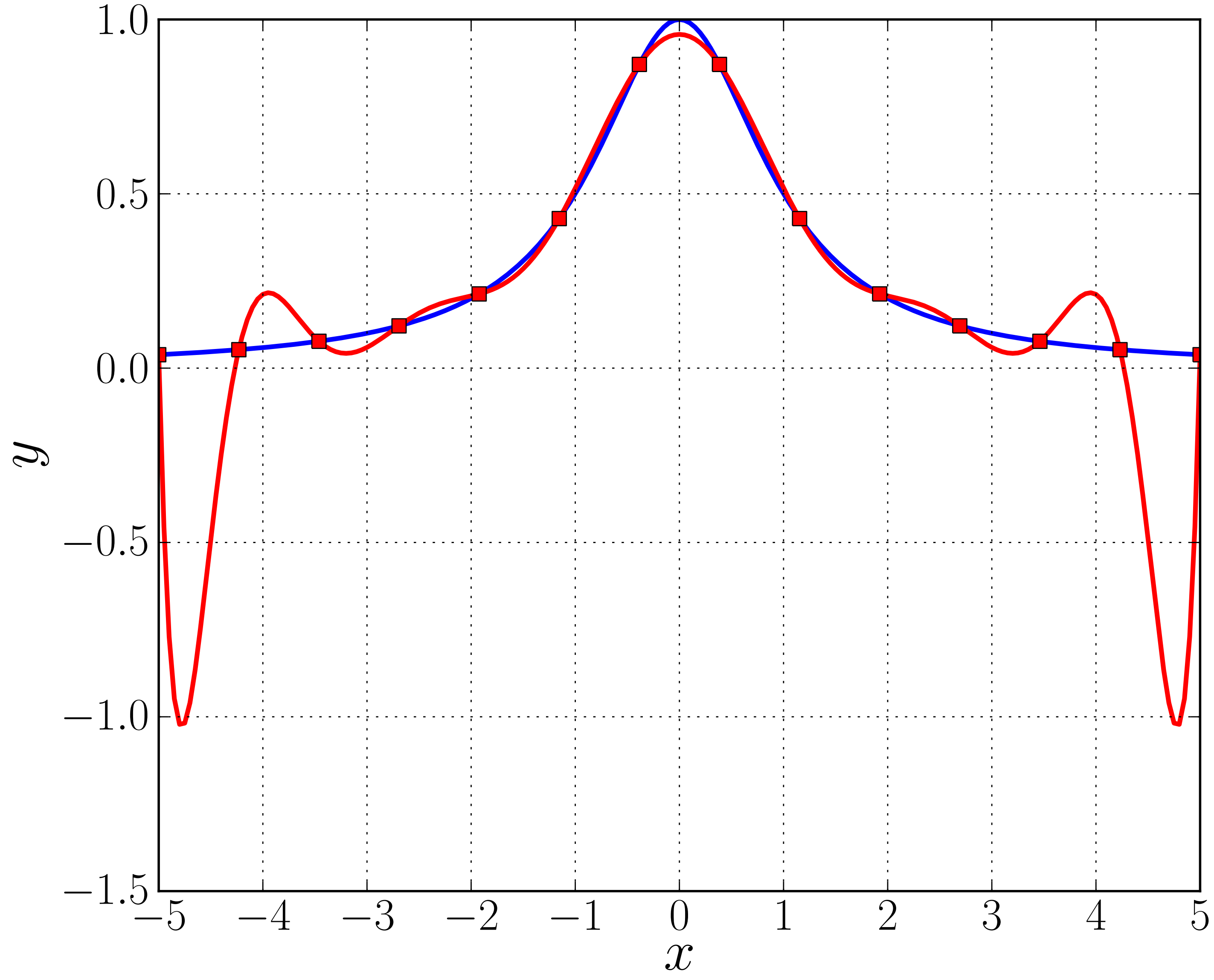
$$x_k \ = \ -5 \ + \ k \ \Delta x \ , \qquad k = 0, 1, \cdots, n \ , \qquad \Delta x \ = \ \frac{10}{n} \ ,$$

then it can be shown that

$$\| \ f \ - \ p \ \|_\infty \ \to \ \infty \qquad \text{as} \qquad n \ \to \ \infty \ .$$



Graph of $f(x) = \frac{1}{1+x^2}$ on the interval $[-5, 5]$
and its Lagrange interpolant $p(x) \in \mathbb{P}_9$ (red)
at ten equally spaced interpolation points $(n = 9)$.



Graph of $f(x) = \frac{1}{1+x^2}$ on the interval $[-5, 5]$
and its Lagrange interpolant $p(x) \in \mathbb{P}_{13}$ (red)
at fourteen equally spaced interpolation points $(n = 13)$.

Conclusion :

- Interpolating a function by a high degree polynomial is *not a good idea* !

Alternative :

- Interpolate the function *locally* by polynomials of relatively low degree .

For given integer N let

$$h \equiv \frac{b-a}{N} ,$$

and partition $[a,b]$ into

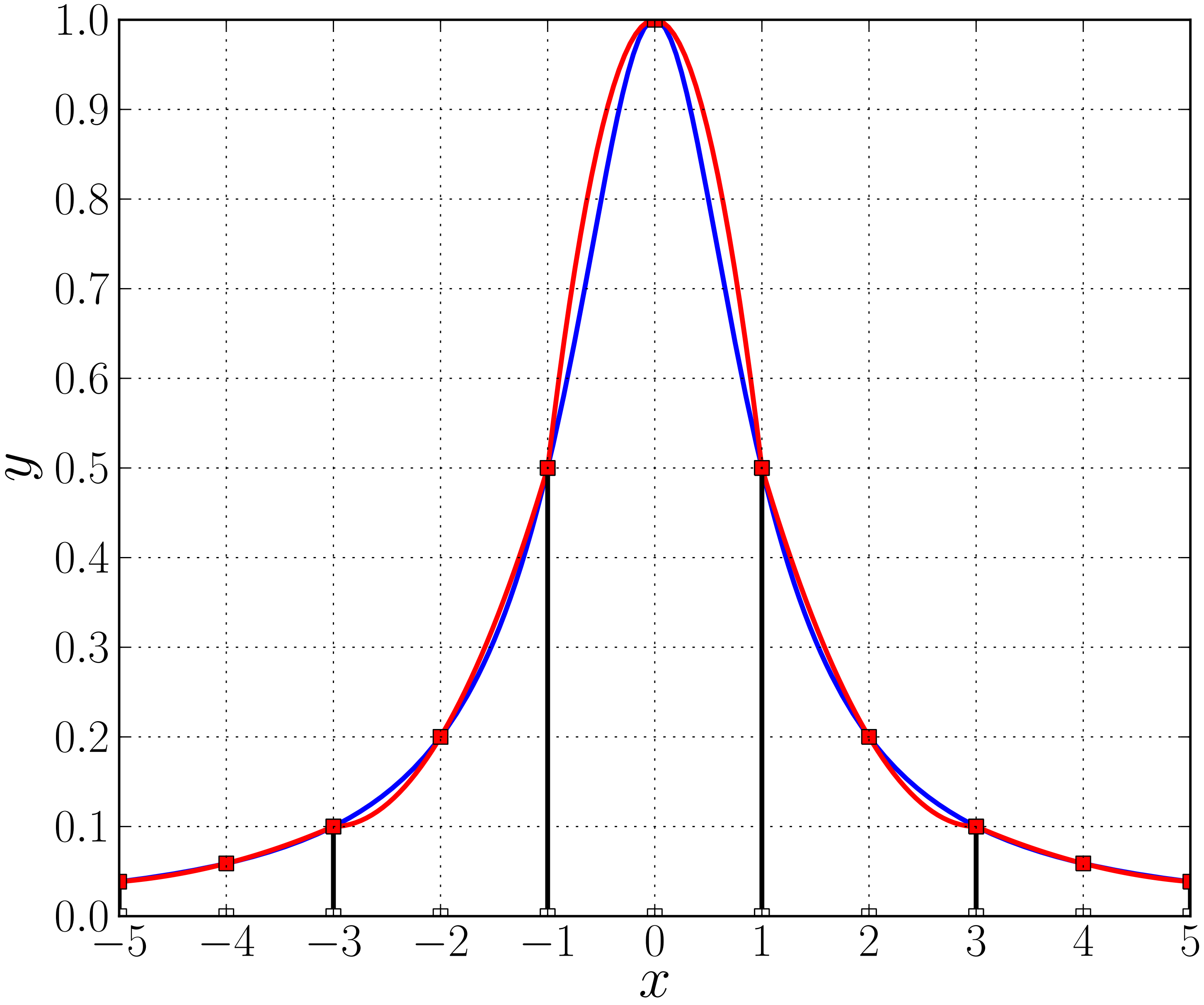
$$a = t_0 < t_1 < \cdots < t_N = b ,$$

where

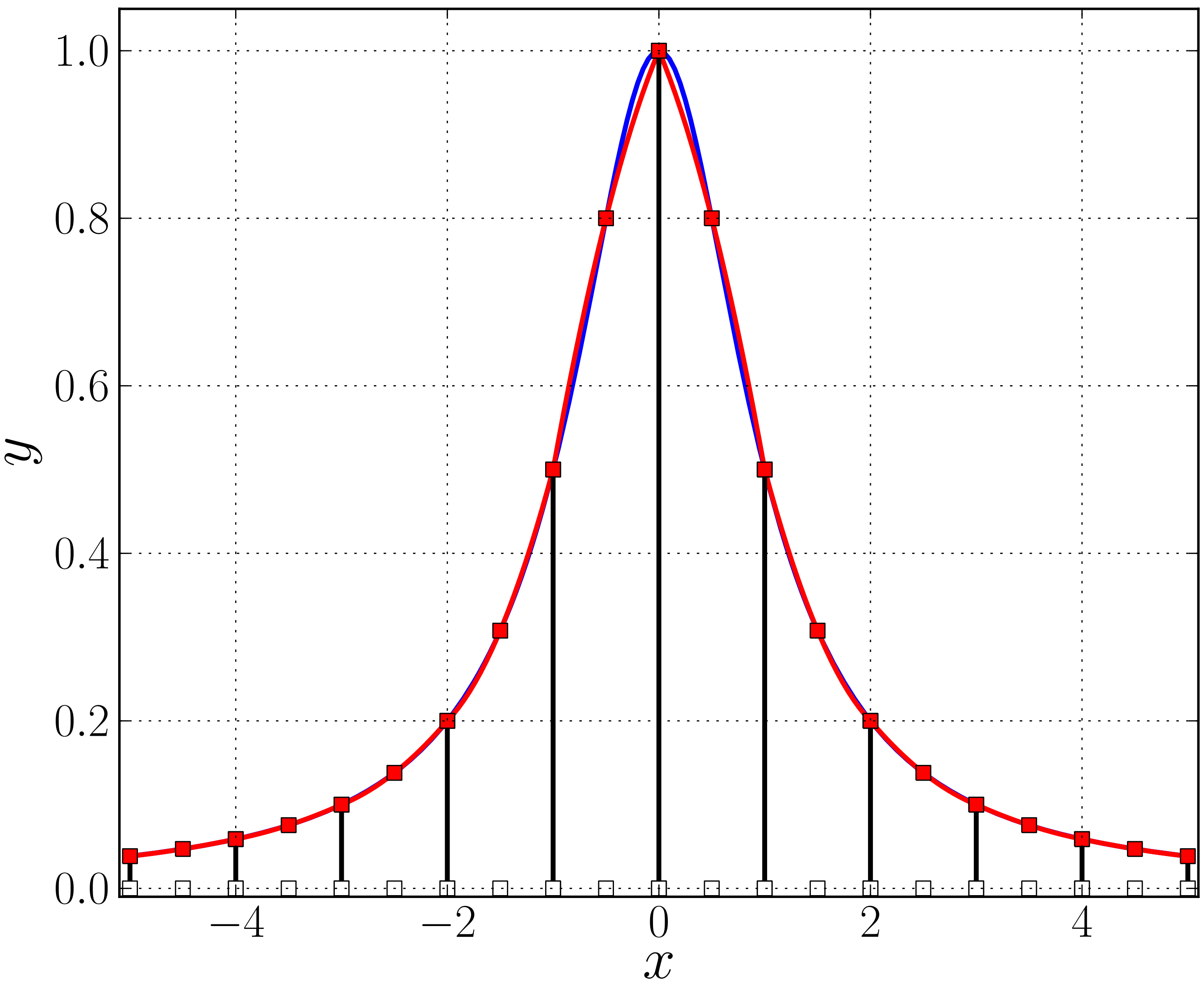
$$t_j = a + jh , \qquad j = 0,1,\cdots,N .$$

In each subinterval $[t_{j-1},t_j]$:

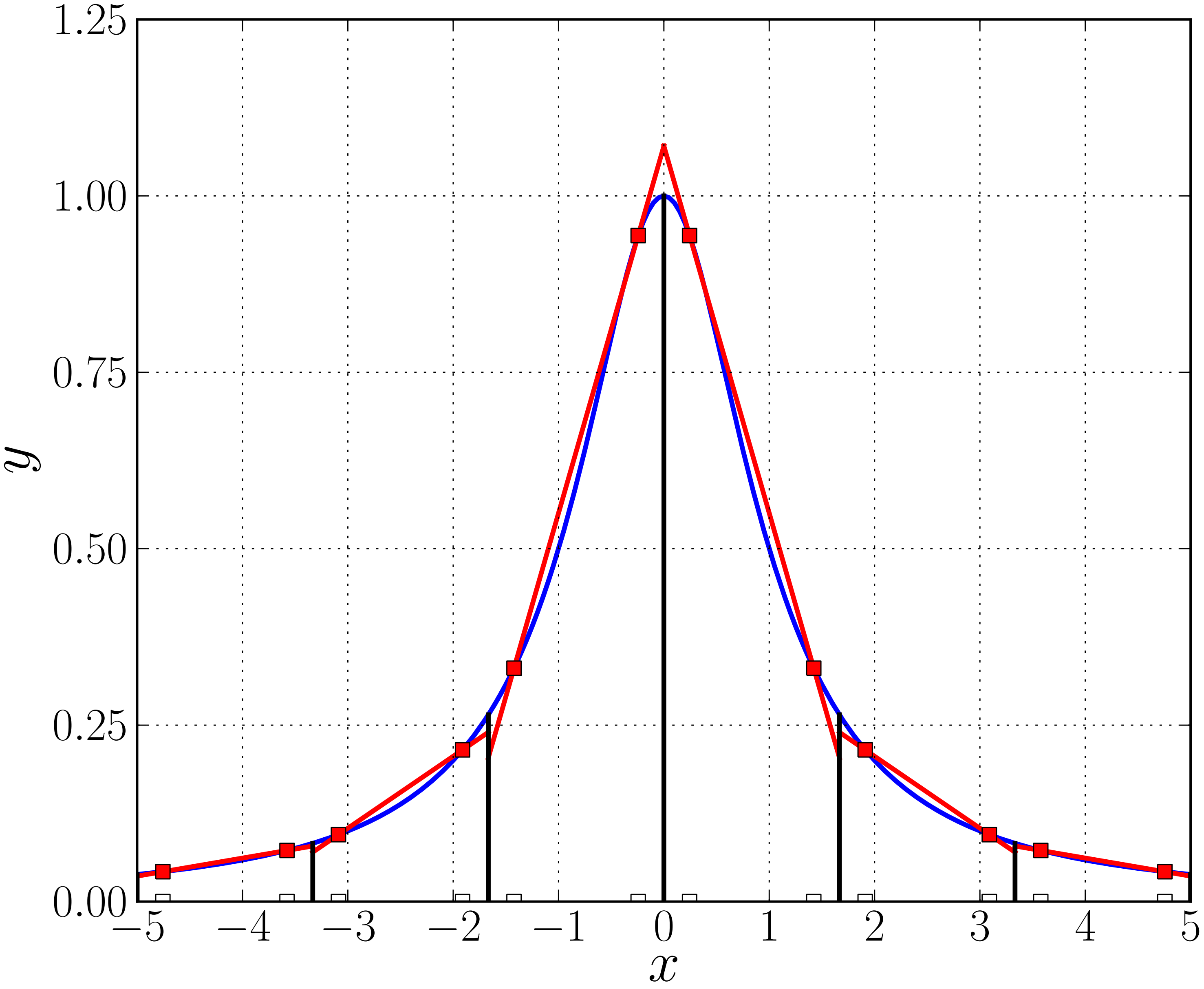
interpolate f by a *local polynomial* $p_j \in \mathbb{P}_n$ at $n+1$ distinct points $\{x_{j,i}\}_{i=0}^{n}$.



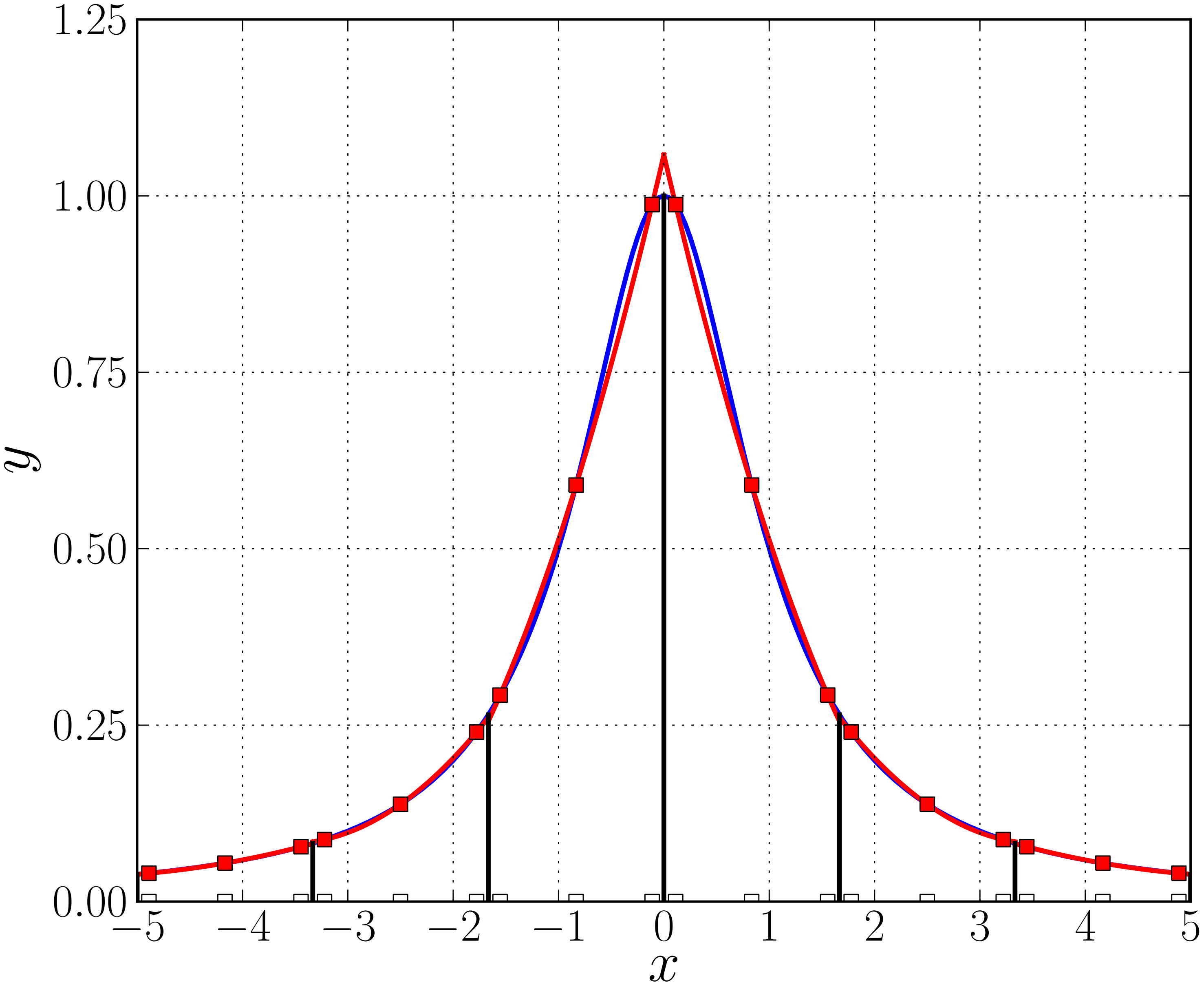
Local polynomial interpolation of $f(x) = \frac{1}{1+x^2}$ at 3 points in 5 intervals.



Local polynomial interpolation of $f(x) = \frac{1}{1+x^2}$ at 3 points in 10 intervals.



Local polynomial interpolation of $f(x) = \frac{1}{1+x^2}$ at 2 Chebyshev points.



Local polynomial interpolation of $f(x) = \frac{1}{1+x^2}$ at 3 Chebyshev points.

By the Lagrange Interpolation Theorem

$$\max_{[t_{j-1},t_j]} \mid f(x) - p_j(x) \mid \ \leq \ \frac{1}{(n+1)!} \max_{[t_{j-1},t_j]} \mid f^{(n+1)}(x) \mid \ \max_{[t_{j-1},t_j]} \mid w_{n+1}(x) \mid$$

where

$$w_{n+1}(x) \ = \ \prod_{i=0}^{n} (x - x_{j,i}) \ , \qquad h \ \equiv \ t_j - t_{j-1} \ = \ \frac{b-a}{N} \ .$$

The Tables on Page 183, 200 show values of $C_n \equiv \max_{[-1,1]} \mid w_{n+1}(x) \mid$ for uniform and for Chebyshev interpolation points.

A scaling argument shows that for *uniformly spaced* local interpolation points

$$\max_{[t_{j-1},t_j]} \mid w_{n+1}(x) \mid \ \leq \ (\frac{h}{2})^{n+1} \ C_n \ ,$$

while for local *Chebyshev points* we have

$$\max_{[t_{j-1},t_j]} \mid w_{n+1}(x) \mid \ \leq \ (\frac{h}{2})^{n+1} \ 2^{-n} \ .$$

NOTE :

- *Keeping* n *fixed* , p_j converges to f as $h \to 0$, ($i.e.$, as $N \to \infty$) .

- To get more accuracy, *increase* N , keeping the degree n *fixed* .

EXAMPLE : If we approximate

$$f(x) \;=\; \cos(x) \qquad \text{on} \qquad [\,0\;,\;\frac{\pi}{2}\,]\;,$$

by local interpolation at 3 *equally spaced* local interpolation points

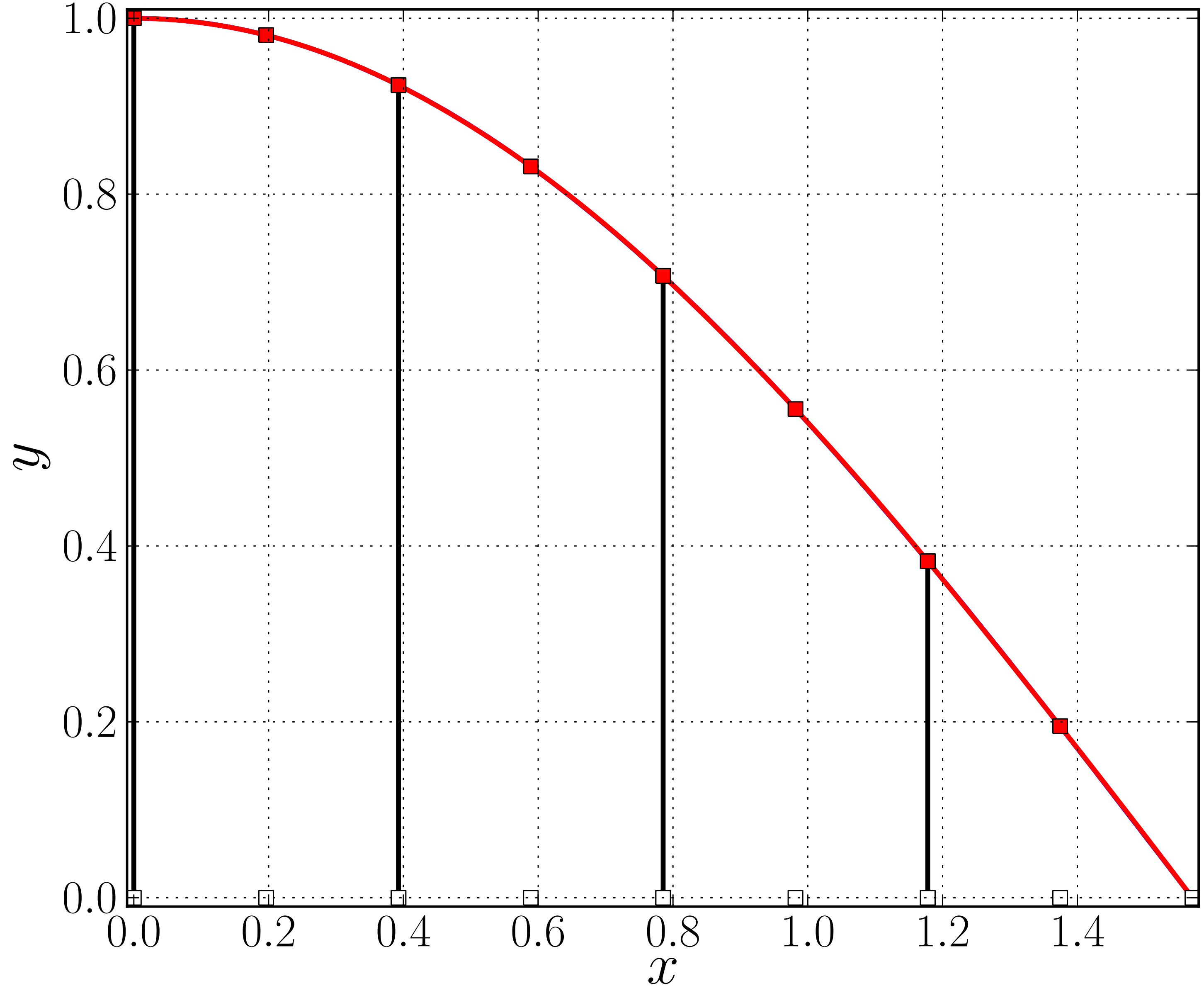
$$x_{j,0} \;=\; t_{j-1}\;, \qquad x_{j,1} \;=\; \frac{t_{j-1}+t_j}{2}\;, \qquad x_{j,2} \;=\; t_j\;,$$

then $n = 2$, $h = \pi/(2N)$, and, using the Table on Page 183,

$$\max_{[t_{j-1},t_j]} \mid f(x) - p_j(x) \mid \;\leq\; \frac{\parallel f^{(3)} \parallel_\infty}{3!}\,(\frac{h}{2})^3\, C_2 \;\leq\; \frac{1}{6}\;\frac{h^3}{8}\;\; 0.3849\;.$$

Specifically, if $N = 4$ (four intervals), then $h = \pi/8$, so that

$$\max_{[t_{j-1},t_j]} \mid f(x) - p_j(x) \mid \;\leq\; \frac{1}{6}\;\frac{1}{8}\;(\frac{\pi}{8})^3\;\; 0.3849 \;=\; 0.000486.$$



Local polynomial interpolation at 3 points in 4 intervals.

EXERCISES :

If we approximate a function $f(x)$ on a given interval by local interpolation with cubic polynomials, then how many intervals of equal size are needed to ensure that the maximum error is less than 10^{-4} ? Answer this question for each of the following cases :

- $f(x) = \sin(x)$ on $[0, 2\pi]$, with arbitrary local interpolation points.
- $f(x) = \sin(x)$ on $[0, 2\pi]$, with equally spaced local points.
- $f(x) = \sin(x)$ on $[0, 2\pi]$, with local Chebyshev interpolation points.
- $f(x) = e^x$ on $[-1, 1]$, with equally spaced local points.
- $f(x) = e^x$ on $[-1, 1]$, with local Chebyshev interpolation points.

NUMERICAL DIFFERENTIATION.

Numerical differentiation formulas can be derived from local interpolation :

Let $p \in \mathbb{P}_n$ interpolate $f(x)$ at points $\{x_i\}_{i=0}^n$. Thus

$$p(x) \ = \ \sum_{i=0}^{n} f(x_i) \ \ell_i(x) \ ,$$

where

$$\ell_i(x) \ = \ \prod_{k=0,k\neq i}^{n} \frac{(x - x_k)}{(x_i - x_k)} \ .$$

For $m \ \leq \ n$ we can approximate $f^{(m)}(x)$ by

$$f^{(m)}(x) \ \cong \ p^{(m)}(x) \ = \ \sum_{i=0}^{n} f(x_i) \ \ell_i^{(m)}(x) \ .$$

EXAMPLE :

Consider the case

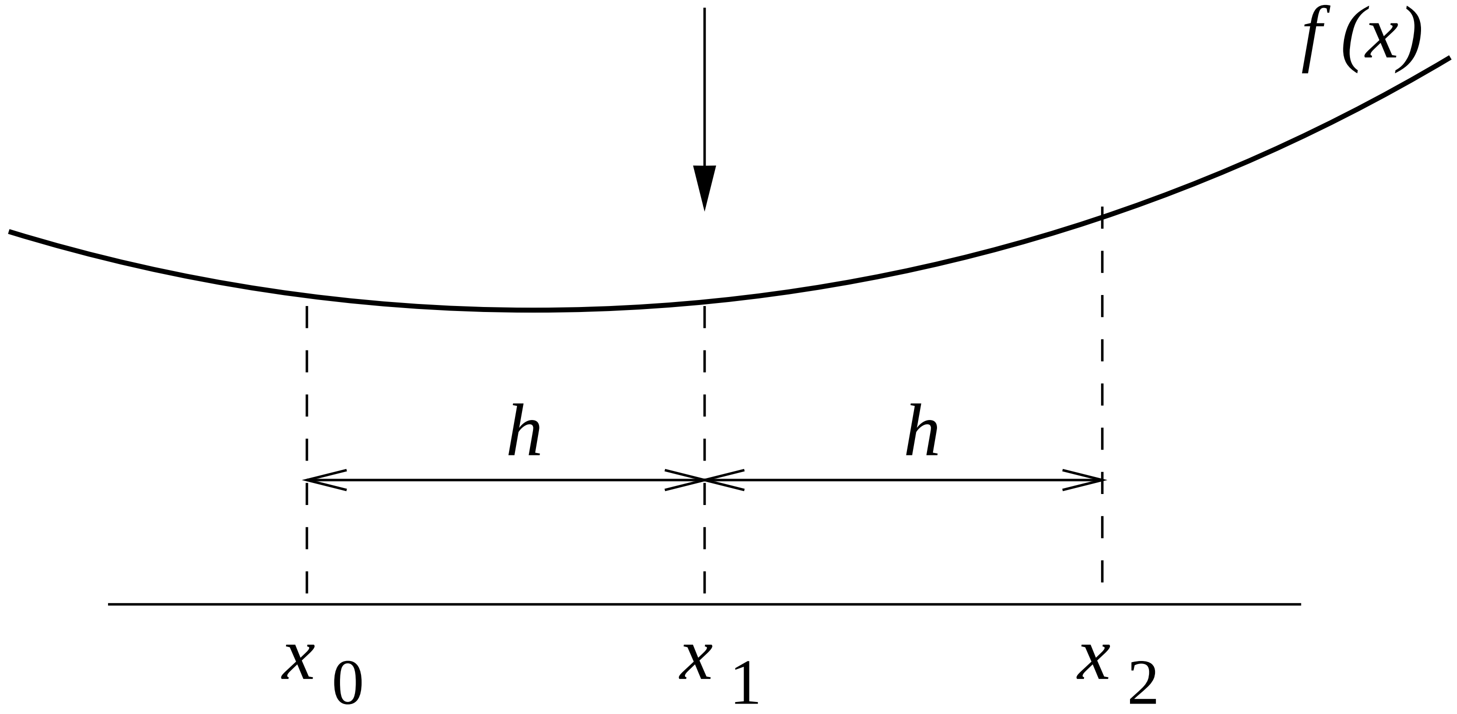
$$n = 2 , \qquad m = 2 , \qquad x = 0 ,$$

for the *reference interval* $[-h, h]$:

$$x_0 = -h , \qquad x_1 = 0 , \qquad x_2 = h .$$

Thus we want to approximate $f''(x_1)$ in terms of

$$f_0 , \quad f_1 , \quad \text{and} \quad f_2 , \qquad (f_i \equiv f(x_i)) .$$



In this case

$$f''(x_1) \;\cong\; p''(x_1) \;=\; f_0\,\ell_0''(x_1) \;+\; f_1\,\ell_1''(x_1) \;+\; f_2\,\ell_2''(x_1)\;.$$

$$f''(x_1) \ \cong \ f_0 \ \ell_0''(x_1) \ + \ f_1 \ \ell_1''(x_1) \ + \ f_2 \ \ell_2''(x_1) \ .$$

Here

$$l_0(x) \ = \ \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} \ ,$$

so that

$$\ell_0''(x) \ = \ \frac{2}{(x_0 - x_1)(x_0 - x_2)} \ = \ \frac{1}{h^2} \ .$$

In particular,

$$\ell_0''(x_1) \ = \ \frac{1}{h^2} \ .$$

Similarly

$$\ell_1''(x_1) \ = \ - \frac{2}{h^2} \ , \qquad \ell_2''(x_1) \ = \ \frac{1}{h^2} \ . \qquad (\textbf{Check !})$$

Hence

$$f''(x_1) \ \cong \ \frac{f_0 - 2f_1 + f_2}{h^2} \ .$$

To derive an optimal error bound we use Taylor's Theorem :

$$\frac{f_2 - 2f_1 + f_0}{h^2} - f_1''$$

$$\begin{aligned}
= \frac{1}{h^2} \Big(\; & f_1 + hf_1' + \frac{h^2}{2} f_1'' + \frac{h^3}{6} f_1''' + \frac{h^4}{24} f''''(\zeta_1) \\
& - 2f_1 \\
& + f_1 - hf_1' + \frac{h^2}{2} f_1'' - \frac{h^3}{6} f_1''' + \frac{h^4}{24} f''''(\zeta_2) \Big) - f_1''
\end{aligned}$$

$$= \frac{h^2}{24} \Big(f''''(\zeta_1) + f''''(\zeta_2) \Big) = \frac{h^2}{12} f''''(\eta) ,$$

where $\eta \in (x_0, x_2)$.

(In the last step we used the *Intermediate Value Theorem* .)

EXAMPLE : With $n = 4$, $m = 2$, and $x = x_2$, and reference interval

$$x_0 \; = \; -\,2h \; , \qquad x_1 \; = \; -\,h \; , \qquad x_2 \; = \; 0 \; , \qquad x_3 \; = \; h \; , \qquad x_4 \; = \; 2h \; ,$$

we have

$$f''(x_2) \;\cong\; \sum_{i=0}^{4} f_i \; \ell_i''(x_2) \; .$$

Here

$$l_0(x) \;=\; \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)(x_0-x_4)}$$

$$= \; \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{24h^4} .$$

Differentiating, and setting x equal to x_2, we find

$$\ell_0''(x_2) \;=\; \frac{-1}{12h^2} \; . \qquad (\textbf{Check !})$$

Similarly

$$\ell_1''(x_2) = \frac{16}{12h^2}, \quad \ell_2''(x_2) = \frac{-30}{12h^2}, \quad \ell_3''(x_2) = \frac{16}{12h^2}, \quad \ell_4''(x_2) = \frac{-1}{12h^2} .$$

(**Check !**)

Hence we have the *five point* finite difference approximation

$$f''(x_2) \cong \frac{-f_0 + 16f_1 - 30f_2 + 16f_3 - f_4}{12h^2} .$$

By Taylor expansion one can show that the leading error term is

$$\frac{h^4 \ f^{(6)}(x_2)}{90} . \qquad (\textbf{Check !})$$

We say that the *order of accuracy* of this approximation is equal to 4 .

EXERCISES :

- Derive a formula for the error in the approximation formula

$$f''(0) \cong \frac{f(2h) - 2f(h) + f(0)}{h^2} .$$

What is the order of accuracy?

- Do the same for $f''(0) \cong \dfrac{f(h) - 2f(0) + f(-h)}{h^2}$.

- Derive the approximation formula

$$f'(0) \cong \frac{-3f(0) + 4f(h) - f(2h)}{2h} .$$

and determine the order of accuracy.

EXERCISES :

- For the reference interval $[0, 3h]$, give complete details on the derivation of the four weights in the numerical differentiation formula

$$f'(0) \cong \frac{-11f(0) + 18f(h) - 9f(2h) + 2f(3h)}{6h} .$$

Use Taylor expansions to determine the leading error term.

- For the reference interval $[-3h/2, 3h/2]$, give complete details on the derivation of the weights in the numerical differentiation formula

$$f'''(0) \cong \frac{-f(-3h/2) + 3f(-h/2) - 3f(h/2) + f(3h/2)}{h^3} .$$

Use Taylor expansions to determine the leading error term.

BEST APPROXIMATION IN THE $\| \cdot \|_2$.

Introductory Example : Best approximation in $\mathbb{R}^3$.

Recall (from Linear Algebra) :

- A *vector* $\mathbf{x} \in \mathbb{R}^3$ is an ordered set of three numbers, $\mathbf{x} = (x_1, x_2, x_3)^T$.

- We can think of $\mathbf{x}$ as a *point* or an *arrow* .

- The dot product or *inner product* of two vectors **x** and **y** is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle \quad \equiv \quad x_1 \; y_1 + x_2 \; y_2 + x_3 \; y_3 \ .$$

- The length or *norm* of a vector is defined in terms of the inner product :

$$\| \mathbf{x} \|_2 \quad \equiv \quad \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}} \quad = \quad \sqrt{x_1^2 + x_2^2 + x_3^2} \ .$$

- Then $\| \mathbf{x}_1 - \mathbf{x}_2 \|_2$ denotes the *distance* between $\mathbf{x}_1$ and $\mathbf{x}_2$.

- Two vectors are *perpendicular* if $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle \ = \ 0$.

Let

$$\mathbf{e}_1 \equiv (1,0,0)^T , \qquad \mathbf{e}_2 \equiv (0,1,0)^T , \qquad \text{and} \qquad \mathbf{e}_3 \equiv (0,0,1)^T .$$

The set $\{\mathbf{e}_k\}_{k=1}^3$ is a *basis* of $\mathbb{R}^3$.

This basis is *orthogonal* because

$$\langle \mathbf{e}_1, \mathbf{e}_2 \rangle = \langle \mathbf{e}_1, \mathbf{e}_3 \rangle = \langle \mathbf{e}_2, \mathbf{e}_3 \rangle = 0 ,$$

and *normal* since

$$\| \mathbf{e}_1 \|_2 = \| \mathbf{e}_2 \|_2 = \| \mathbf{e}_3 \|_2 = 1 ,$$

i.e., the basis is *orthonormal* .

Let $\mathcal{S}_2$ denote the x_1, x_2-plane .

Then

$$\mathcal{S}_2 \; = \; \text{Span}\{\mathbf{e}_1, \mathbf{e}_2\} \; .$$

$\mathcal{S}_2$ is a 2-dimensional *subspace* of $\mathbb{R}^3$.

Suppose we want to find the *best approximation*

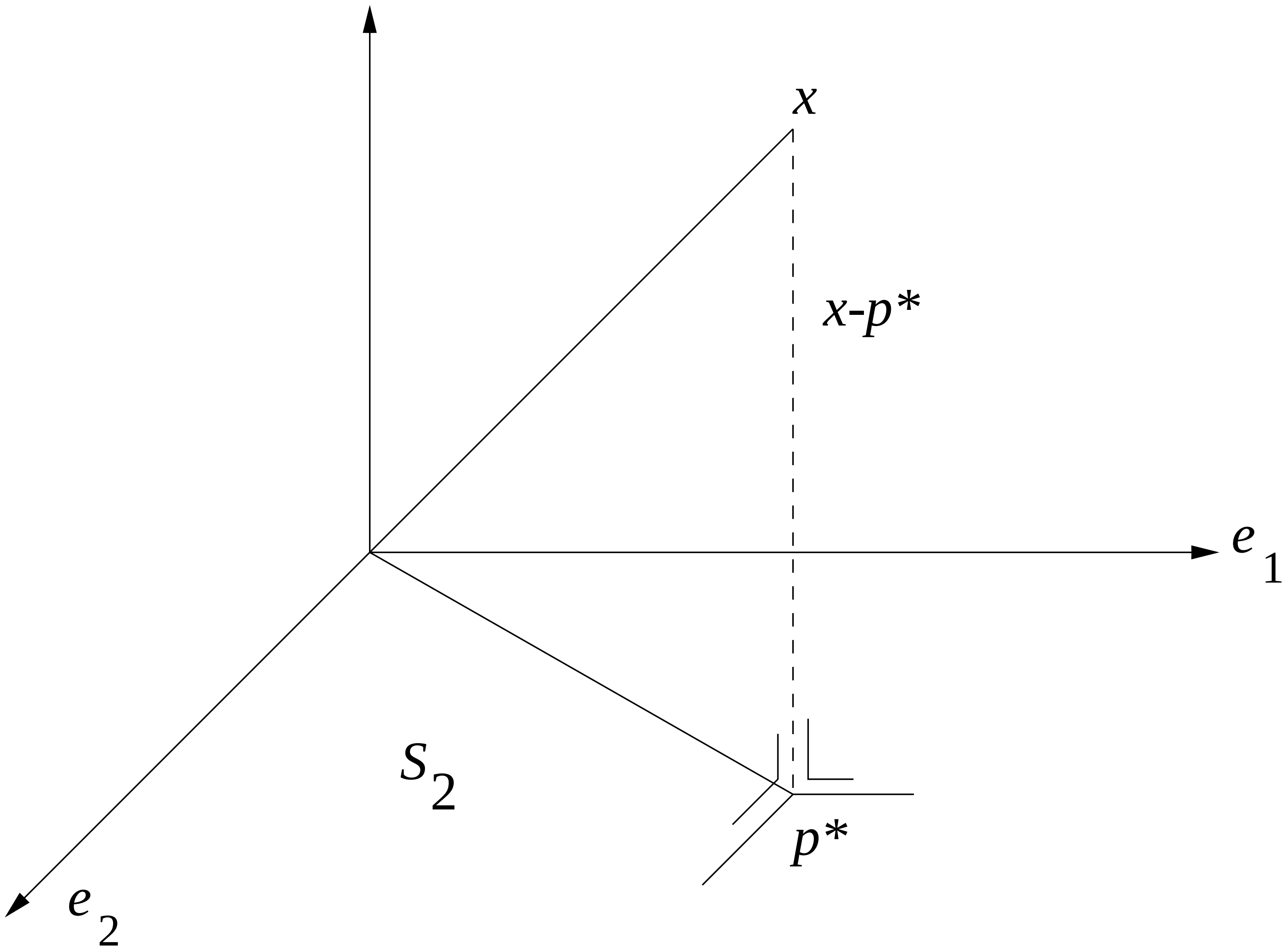
$$\mathbf{p}^* \in \mathcal{S}_2 \; ,$$

to a given vector $\mathbf{x} \in \mathbb{R}^3$.

Thus we want $\mathbf{p}^* \in \mathcal{S}_2$ that minimizes

$$\| \, \mathbf{x} - \mathbf{p} \, \|_2 \quad ,$$

over all $\mathbf{p} \in \mathcal{S}_2$.


x
x-p*
e 1
S 2
p*
e 2

Geometrically we see that $\| \mathbf{x} - \mathbf{p} \|_2$ is minimized if and only if

$$(\mathbf{x} - \mathbf{p}) \perp \mathcal{S}_2 \, ,$$

i.e., if and only if

$$\langle \mathbf{x} - \mathbf{p} \, , \, \mathbf{e}_1 \rangle = 0 \, , \qquad \text{and} \qquad \langle \mathbf{x} - \mathbf{p} \, , \, \mathbf{e}_2 \rangle = 0 \, ,$$

i.e., if and only if

$$\langle \mathbf{x} \, , \, \mathbf{e}_1 \rangle = \langle \mathbf{p} \, , \, \mathbf{e}_1 \rangle \, , \qquad \text{and} \qquad \langle \mathbf{x} \, , \, \mathbf{e}_2 \rangle = \langle \mathbf{p} \, , \, \mathbf{e}_2 \rangle \, .$$

Since $\mathbf{p} \in \mathcal{S}_2$ we have

$$\mathbf{p} \ = \ c_1 \ \mathbf{e}_1 \ + \ c_2 \ \mathbf{e}_2 \ ,$$

for certain constants c_1 and c_2 .

Thus $\| \ \mathbf{x} - \mathbf{p} \ \|_2$ is minimized if and only if

$$\langle \mathbf{x}, \mathbf{e}_1 \rangle \ = \ \langle \mathbf{p}, \mathbf{e}_1 \rangle \ = \ \langle c_1\mathbf{e}_1 + c_2\mathbf{e}_2 \ , \ \mathbf{e}_1 \rangle \ = \ c_1 \ ,$$

$$\langle \mathbf{x}, \mathbf{e}_2 \rangle \ = \ \langle \mathbf{p}, \mathbf{e}_2 \rangle \ = \ \langle c_1\mathbf{e}_1 + c_2\mathbf{e}_2 \ , \ \mathbf{e}_2 \rangle \ = \ c_2 \ .$$

Hence

$$\mathbf{p}^* \ = \ c_1 \ \mathbf{e}_1 + c_2 \ \mathbf{e}_2 \ ,$$

with

$$c_1 \ = \ \langle \mathbf{x}, \mathbf{e}_1 \rangle \qquad \text{and} \qquad c_2 = \langle \mathbf{x}, \mathbf{e}_2 \rangle \ .$$

Best Approximation in General.

Let $\mathbf{X}$ be a (possibly infinite-dimensional) real vector space,

with an inner product satisfying :

for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{X}$, and for all $\alpha \in \mathbb{R}$:

$$(i) \qquad \langle \mathbf{x}, \mathbf{x} \rangle \ \geq \ 0 \ , \qquad \langle \mathbf{x}, \mathbf{x} \rangle \ = \ 0 \ \text{ only if } \mathbf{x} \ = \ \mathbf{0} \ ,$$

$$(ii) \qquad \langle \alpha\mathbf{x}, \mathbf{y} \rangle \ = \ \langle \mathbf{x}, \alpha\mathbf{y} \rangle \ = \ \alpha\langle \mathbf{x}, \mathbf{y} \rangle \ ,$$

$$(iii) \qquad \langle \mathbf{x}, \mathbf{y} \rangle \ = \ \langle \mathbf{y}, \mathbf{x} \rangle \ ,$$

$$(iv) \qquad \langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle \ = \ \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle \ .$$

THEOREM :

Let $\mathbf{X}$ be a vector space with an inner product satisfying the properties above.

Then

$$\| \, \mathbf{x} \, \| \quad \equiv \quad \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}} \ ,$$

defines a *norm* on $\mathbf{X}$.

PROOF : We must show that $\| \cdot \|$ satisfies the usual properties :

(i) Clearly $\| \, \mathbf{x} \, \| \ \geq \ 0$, and $\| \, \mathbf{x} \, \| \ = \ 0$ only if $\mathbf{x} = \mathbf{0}$.

(ii) $\| \, \alpha\mathbf{x} \, \| \ = \ \langle \alpha\mathbf{x}, \alpha\mathbf{x} \rangle^{\frac{1}{2}} \ = \ (\ \alpha^2 \langle \mathbf{x}, \mathbf{x} \rangle \ )^{\frac{1}{2}} \ = \ | \, \alpha \, | \ \ \| \, \mathbf{x} \, \|$.

(iii) The triangle inequality is also satisfied :

Let

$$\alpha \equiv \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\| \mathbf{y} \|^2}, \quad \text{where} \quad \mathbf{x}, \mathbf{y} \in \mathbf{X} .$$

Then

$$\begin{aligned} 0 \leq \| \mathbf{x} - \alpha \mathbf{y} \|^2 &= \langle \mathbf{x} - \alpha \mathbf{y}, \mathbf{x} - \alpha \mathbf{y} \rangle \\ &= \| \mathbf{x} \|^2 - 2\alpha \langle \mathbf{x}, \mathbf{y} \rangle + \alpha^2 \| \mathbf{y} \|^2 \\ &= \| \mathbf{x} \|^2 - 2 \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\| \mathbf{y} \|^2} \langle \mathbf{x}, \mathbf{y} \rangle + \frac{\langle \mathbf{x}, \mathbf{y} \rangle^2}{\| \mathbf{y} \|^4} \| \mathbf{y} \|^2 \\ &= \| \mathbf{x} \|^2 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle^2}{\| \mathbf{y} \|^2} . \end{aligned}$$

Hence

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \| \mathbf{x} \|^2 \ \| \mathbf{y} \|^2 ,$$

or

$$| \langle \mathbf{x}, \mathbf{y} \rangle | \leq \| \mathbf{x} \| \ \| \mathbf{y} \| \qquad (\mathit{Cauchy - Schwartz\ Inequality}) .$$

Now

$$\begin{aligned} \| \mathbf{x} + \mathbf{y} \|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\ &= \| \mathbf{x} \|^2 \ + \ 2\langle \mathbf{x}, \mathbf{y} \rangle \ + \ \| \mathbf{y} \|^2 \\ &\leq \| \mathbf{x} \|^2 \ + \ 2 \, | \, \langle \mathbf{x}, \mathbf{y} \rangle \, | \ + \ \| \mathbf{y} \|^2 \\ &\leq \| \mathbf{x} \|^2 \ + \ 2 \, \| \mathbf{x} \| \| \mathbf{y} \| \ + \ \| \mathbf{y} \|^2 \\ &= (\ \| \mathbf{x} \| \ + \ \| \mathbf{y} \| \ )^2 \ . \end{aligned}$$

Hence

$\| \mathbf{x} + \mathbf{y} \| \ \leq \ \| \mathbf{x} \| \ + \ \| \mathbf{y} \| \ .$ (*Triangle Inequality*) **QED !**

Suppose $\{\mathbf{e}_k\}_{k=1}^n$ is an *orthonormal* set of vectors in $\mathbf{X}$, *i.e.*,

$$\langle \mathbf{e}_\ell, \mathbf{e}_k \rangle \ = \ \begin{cases} 0 \ , & \text{if} \ \ l \ \neq \ k \ , \\ \\ 1 \ , & \text{if} \ \ l \ = \ k \ . \end{cases}$$

Let $\mathcal{S}_n \ \subset \ \mathbf{X}$ be defined by

$$\mathcal{S}_n \ = \ \text{Span}\{\mathbf{e}_k\}_{k=1}^n \ .$$

We want the *best approximation* $\mathbf{p}^* \in \mathcal{S}_n$ to a given vector $\mathbf{x} \in \mathbf{X}$.

Thus we want to find $\mathbf{p}^* \in \mathcal{S}_n$ that *minimizes* $\| \, \mathbf{x} - \mathbf{p} \, \|$ over all $\mathbf{p} \in \mathcal{S}_n$.

THEOREM :

The *best approximation* $\mathbf{p}^* \in \mathcal{S}_n$ to $\mathbf{x} \in \mathbf{X}$ is given by

$$\mathbf{p}^* \ = \ \sum_{k=1}^{n} c_k \ \mathbf{e}_k \ ,$$

where the *Fourier Coefficients* c_k , $(k = 1, 2, \cdots, n)$, are given by

$$c_k \ = \ \frac{\langle \mathbf{x}, \mathbf{e}_k \rangle}{\langle \mathbf{e}_k, \mathbf{e}_k \rangle} \ , \qquad \text{if the basis is } \textit{orthogonal} \ ,$$

and

$$c_k \ = \ \langle \mathbf{x}, \mathbf{e}_k \rangle \ , \qquad \text{if the basis is } \textit{orthonormal} \ .$$

PROOF : Let

$$F(c_1, c_2, \cdots, c_n) \quad \equiv \quad \| \, \mathbf{x} - \sum_{k=1}^{n} c_k \mathbf{e}_k \, \|^2 \quad .$$

Thus we want to find the $\{c_k\}_{k=1}^{n}$ that minimize F.

Now

$$\begin{aligned} F(c_1, c_2, \cdots, c_n) \quad &= \quad \langle \mathbf{x} - \sum_{k=1}^{n} c_k \mathbf{e}_k \ , \ \mathbf{x} - \sum_{k=1}^{n} c_k \mathbf{e}_k \rangle \\ &= \quad \langle \mathbf{x}, \mathbf{x} \rangle \ - \ 2 \, \langle \sum_{k=1}^{n} c_k \mathbf{e}_k, \mathbf{x} \rangle \ + \ \langle \sum_{k=1}^{n} c_k \mathbf{e}_k \ , \ \sum_{k=1}^{n} c_k \mathbf{e}_k \rangle \\ &= \quad \| \, \mathbf{x} \, \|^2 \ - \ 2 \sum_{k=1}^{n} c_k \langle \mathbf{x}, \mathbf{e}_k \rangle \ + \ \sum_{k=1}^{n} c_k^2 \langle \mathbf{e}_k, \mathbf{e}_k \rangle \ . \end{aligned}$$

For F to be *minimized* we must have

$$\frac{\partial F}{\partial c_\ell} \ = \ 0 \ , \qquad \ell = 1, 2, \cdots, n \ .$$

We had

$$F(c_1, c_2, \cdots, c_n) \quad = \quad \| \, \mathbf{x} \, \|^2 \quad - \; 2\sum_{k=1}^{n} c_k \langle \mathbf{x}, \mathbf{e}_k \rangle \; + \; \sum_{k=1}^{n} c_k^2 \langle \mathbf{e}_k, \mathbf{e}_k \rangle \; .$$

Setting $\frac{\partial F}{\partial c_\ell} \; = \; 0$ gives

$$-2\langle \mathbf{x}, \mathbf{e}_\ell \rangle \; + \; 2c_\ell \langle \mathbf{e}_\ell, \mathbf{e}_\ell \rangle \; = \; 0 \; .$$

Hence, for $\ell = 1, 2, \cdots, n$, we have

$$c_\ell \quad = \quad \frac{\langle \mathbf{x}, \mathbf{e}_\ell \rangle}{\langle \mathbf{e}_\ell, \mathbf{e}_\ell \rangle} \; ,$$

$$c_\ell \quad = \quad \langle \mathbf{x}, \mathbf{e}_\ell \rangle \; , \qquad \text{if the basis is orthonormal .}$$

QED !

NOTE :

- The proof uses the fact that $\mathbf{X}$ is an *inner product space* , with norm defined in terms of the inner product.

- In normed vector spaces *without* inner product, *e.g.*,

$$\mathbb{C}[0,1] \qquad \text{with} \qquad \| \cdot \|_\infty \ ,$$

it is *more difficult* to find a best approximation.

Gram-Schmidt Orthogonalization.

To construct

$$\text{an } \textit{orthogonal basis} \quad \{\mathbf{e}_k\}_{k=1}^n \quad \text{of a subspace} \quad \mathcal{S}_n \quad \text{of} \quad \mathbf{X} \ ,$$

we have the *Gram-Schmidt* Orthogonalization Procedure :

- Take any nonzero $\mathbf{e}_1 \in \mathcal{S}_n$.

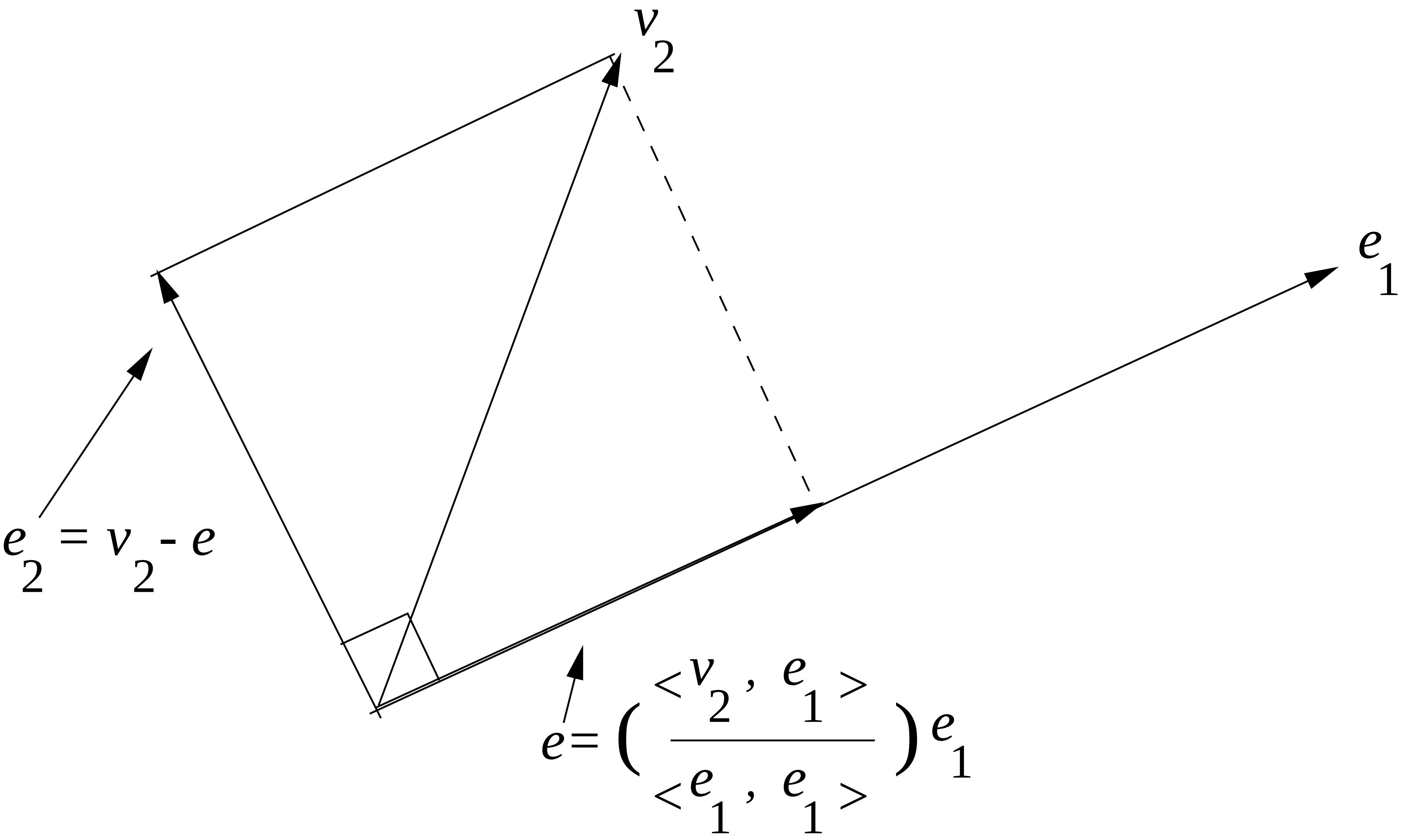
- Choose any $\mathbf{v}_2 \in \mathcal{S}_n$ that is linearly independent from $\mathbf{e}_1$.

- Set

$$\mathbf{e}_2 \ = \ \mathbf{v}_2 \ - \ \frac{\langle \mathbf{v}_2, \mathbf{e}_1 \rangle}{\langle \mathbf{e}_1, \mathbf{e}_1 \rangle} \ \mathbf{e}_1 \ .$$

Then

$$\langle \mathbf{e}_2, \mathbf{e}_1 \rangle \ = \ 0 \ . \qquad (\textbf{Check !})$$


v_2
e_1
$e_2 = v_2 - e$
$e = \left(\frac{<v_2, e_1>}{<e_1, e_1>} \right) e_1$

Inductively, suppose we have mutually orthogonal $\{\mathbf{e}_k\}_{k=1}^{m-1}$, $(m \leq n)$.

- Choose $\mathbf{v}_m \in \mathcal{S}_n$ linearly independent from the $\{\mathbf{e}_k\}_{k=1}^{m-1}$.

- Set

$$\mathbf{e}_m \ = \ \mathbf{v}_m \ - \ \sum_{k=1}^{m-1} \frac{\langle \mathbf{v}_m, \mathbf{e}_k \rangle}{\langle \mathbf{e}_k, \mathbf{e}_k \rangle} \ \mathbf{e}_k \ .$$

Then

$$\langle \mathbf{e}_m, \mathbf{e}_\ell \rangle \ = \ 0 \ . \qquad \ell = 1, 2, \cdots, m-1 \ . \qquad (\textbf{Check !})$$

An orthonormal basis can be obtained by *normalizing* :

$$\hat{\mathbf{e}}_k \ = \ \frac{\mathbf{e}_k}{\| \ \mathbf{e}_k \ \|} \ , \qquad k = 1, 2, \cdots, n \ .$$

Best Approximation in a Function Space.

We now apply the general results the special case where

$$\mathbf{X} \ = \ \mathbb{C}[-1,1] \ ,$$

with *inner product*

$$\langle f,g \rangle \ \equiv \ \int_{-1}^{1} f(x) \ g(x) \ dx \ .$$

This definition satisfies all conditions an inner product must satisfy. (**Check !**)

Hence, from the Theorem it follows that

$$\| \ f \ \|_2 \ \ \equiv \ \ \langle f,f \rangle^{\frac{1}{2}} \ \ = \ \ (\int_{-1}^{1} f(x)^2 \ \ dx \ )^{\frac{1}{2}} \ ,$$

is a *norm* on $\mathbb{C}[-1,1]$.

Suppose we want to find $p^* \in \mathbb{P}_n$ that best approximates a given function

$$f \in \mathbb{C}[-1, 1] ,$$

in the $\| \cdot \|_2$.

Here $\mathbb{P}_n$ is the $(n+1)$-dimensional subspace of $\mathbb{C}[-1, 1]$ consisting of all polynomials of degree less than or equal to n.

By the Theorem we have

$$p^*(x) \ = \ \sum_{k=0}^{n} c_k \ e_k(x) \ ,$$

where

$$c_k \ = \ \frac{\langle f, e_k \rangle}{\langle e_k, e_k \rangle} \ = \ \frac{\int_{-1}^{1} f(x) \ e_k(x) \ dx}{\int_{-1}^{1} e_k^2(x) \ dx} \ , \qquad k = 0, 1, \cdots, n \ ,$$

and where the $\{e_k\}_{k=0}^{n}$ denote the first $n+1$ *orthogonal polynomials* .

Use the Gram-Schmidt procedure to construct an *orthogonal basis* of $\mathbb{P}_n$:

(These basis polynomials are called the *Legendre polynomials*.)

Take $e_0(x) \equiv 1$ and $v_1(x) = x$.

Then

$$\frac{\langle v_1, e_0 \rangle}{\langle e_0, e_0 \rangle} = \frac{\int_{-1}^{1} x \ dx}{\int_{-1}^{1} 1^2 \ dx} = 0 \ .$$

Hence

$$e_1(x) = v_1(x) - 0 \cdot e_0(x) = x \ .$$

Take $v_2(x) = x^2$.

Then

$$\frac{\langle v_2, e_0 \rangle}{\langle e_0, e_0 \rangle} = \frac{\int_{-1}^{1} x^2 \, dx}{\int_{-1}^{1} 1^2 \, dx} = \frac{1}{3} ,$$

and

$$\frac{\langle v_2, e_1 \rangle}{\langle e_1, e_1 \rangle} = \frac{\int_{-1}^{1} x^3 \, dx}{\int_{-1}^{1} x^2 \, dx} = 0 .$$

Hence

$$e_2(x) = v_2(x) - \frac{1}{3} \, e_0(x) - 0 \cdot e_1(x) = x^2 - \frac{1}{3} .$$

Take $v_3(x) = x^3$. Then

$$\frac{\langle v_3, e_0\rangle}{\langle e_0, e_0\rangle} = \frac{\int_{-1}^{1} x^3\ dx}{\int_{-1}^{1} 1^2\ dx} = 0\ ,$$

and

$$\frac{\langle v_3, e_1\rangle}{\langle e_1, e_1\rangle} = \frac{\int_{-1}^{1} x^4\ dx}{\int_{-1}^{1} x^2\ dx} = \frac{3}{5}\ ,$$

and

$$\frac{\langle v_3, e_2\rangle}{\langle e_2, e_2\rangle} = \frac{\int_{-1}^{1} x^3(x^2 - \frac{1}{3})\ dx}{\int_{-1}^{1}(x^2 - \frac{1}{3})^2\ dx} = 0\ .$$

Hence

$$e_3(x) = v_3(x) - 0 \cdot e_0(x) - \frac{3}{5}\ e_1(x) - 0 \cdot e_2(x) = x^3 - \frac{3}{5}x\ .$$

etc.

EXAMPLE :

The polynomial $p^* \in \mathbb{P}_2$ that best approximates

$$f(x) = e^x \ , \qquad \text{on} \qquad [-1,1] \ , \qquad \text{in} \qquad \| \cdot \|_2 \ ,$$

is given by

$$p^*(x) = c_0 \ e_0(x) + c_1 \ e_1(x) + c_2 \ e_2(x) \ ,$$

where

$$c_0 = \frac{\langle f, e_0 \rangle}{\langle e_0, e_0 \rangle} \ , \qquad c_1 = \frac{\langle f, e_1 \rangle}{\langle e_1, e_1 \rangle} \ , \qquad c_2 = \frac{\langle f, e_2 \rangle}{\langle e_2, e_2 \rangle} \ .$$

We find that

$$c_0 \;=\; \frac{\langle f, e_0\rangle}{\langle e_0, e_0\rangle} \;=\; \frac{\int_{-1}^{1} e^x \; dx}{\int_{-1}^{1} 1^2 \; dx} \;=\; \frac{1}{2}(e - \frac{1}{e}) \;=\; 1.175 \;,$$

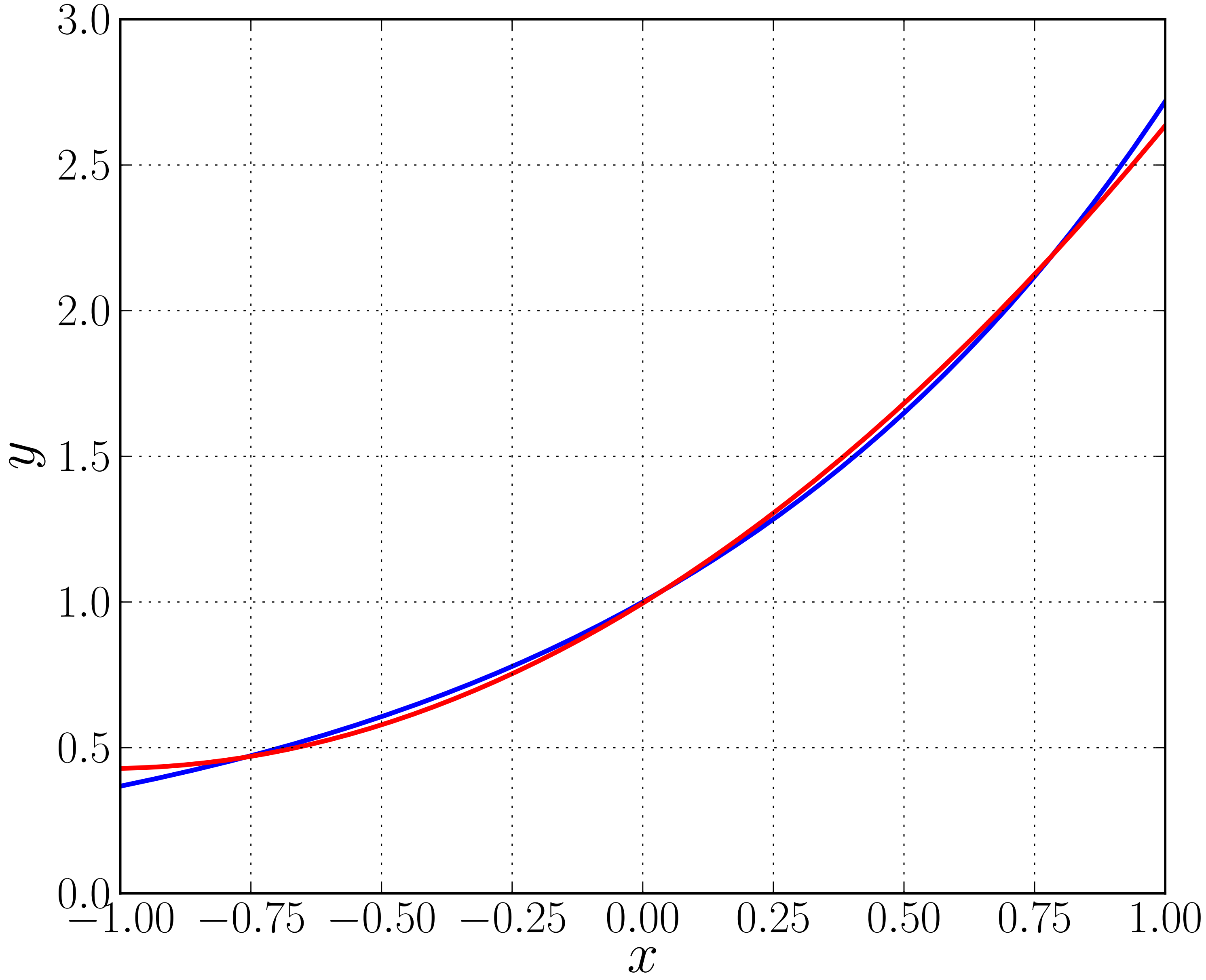
$$c_1 \;=\; \frac{\langle f, e_1\rangle}{\langle e_1, e_1\rangle} \;=\; \frac{\int_{-1}^{1} e^x \; x \; dx}{\int_{-1}^{1} x^2 \; dx} \;=\; \frac{3}{2}(x-1)e^x|_{-1}^{1} \;=\; 1.103 \;,$$

$$c_2 \;=\; \frac{\langle f, e_2\rangle}{\langle e_2, e_2\rangle} \;=\; \frac{\int_{-1}^{1} e^x \; (x^2 - \frac{1}{3}) \; dx}{\int_{-1}^{1} (x^2 - \frac{1}{3})^2 \; dx} \;=\; \frac{45}{8}(x^2 - 2x + \frac{5}{3})e^x|_{-1}^{1} \;=\; 0.536 \;.$$

Therefore

$$p^*(x) \;=\; 1.175\;(1) \;+\; 1.103\;(x) \;+\; 0.536\;(x^2 - \frac{1}{3})$$

$$= \; 0.536\;x^2 \;+\; 1.103\;x \;+\; 0.996 \;. \qquad (\textbf{Check !})$$



Best approximation of $f(x) = e^x$ in $[-1, 1]$ by a polynomial $p \in \mathbb{P}_2$.

EXERCISES :

- Use the Gram-Schmidt procedure to construct an orthogonal basis of the polynomial space $\mathbb{P}_4$ on the interval $[-1,1]$, by deriving $e_4(x)$, given $e_0(x) = 1$, $e_1(x) = x$, $e_2(x) = x^2 - \frac{1}{3}$, and $e_3(x) = x^3 - \frac{3}{5}x$.

- Use the Gram-Schmidt procedure to construct an orthogonal basis of the linear space Span$\{1, x, x^3\}$ for the interval $[-1,1]$. Determine the best approximation in the $\| \cdot \|_2$ to $f(x) = x^5$.

- Use the Gram-Schmidt procedure to construct an orthogonal basis of the linear space Span$\{1, x^2, x^4\}$ for the interval $[-1,1]$. Determine the best approximation in the $\| \cdot \|_2$ to $f(x) = x^6$.

- Show that the functions $e_0(x) \equiv 1$, $e_1(x) = \sin(x)$, $e_2(x) = \cos(x)$, $e_3(x) = \sin(2x)$, $e_4(x) = \cos(2x)$, are mutually orthogonal with respect to the inner product $< f, g >= \int_0^{2\pi} f(x)g(x)\, dx$. Also show how one can determine the coefficients c_k, $k = 0,1,2,3,4$, of the trigonometric polynomial $p(x) = c_0 + c_1 \sin(x) + c_2 \cos(x) + c_3 \sin(2x) + c_4 \cos(2x)$ that minimizes $\int_0^{2\pi} (p(x) - f(x))^2\, dx$, when $f(x) = e^x$.

NUMERICAL INTEGRATION

- Many definite integrals, *e.g.*,

$$\int_0^1 e^{(x^2)} \, dx \ ,$$

are difficult or impossible to evaluate analytically.

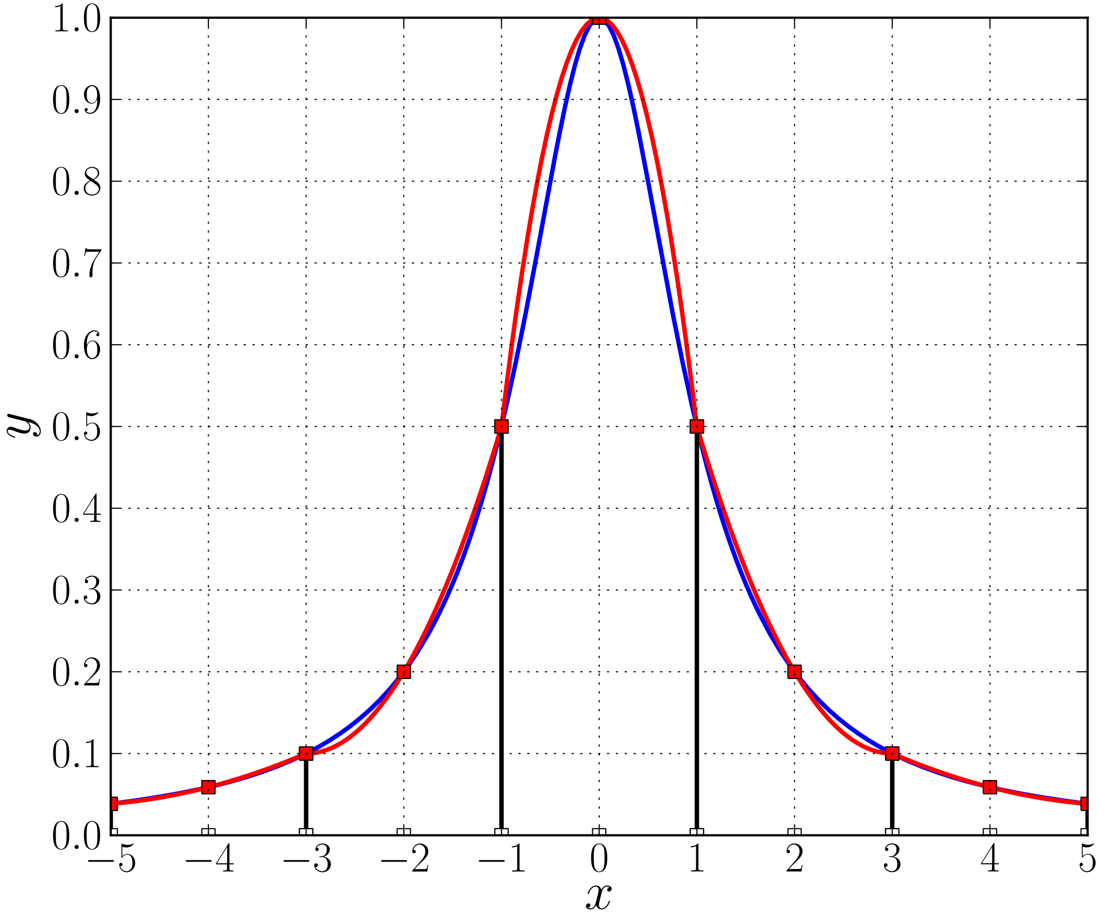
- In such cases we can use *numerical integration*.

- There are many *numerical integration* (or *quadrature*) formulas.

Most formulas are based on integrating local interpolating polynomials of f :

$$\int_a^b f(x)\ dx \ \cong \ \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} p_j(x)\ dx\ ,$$

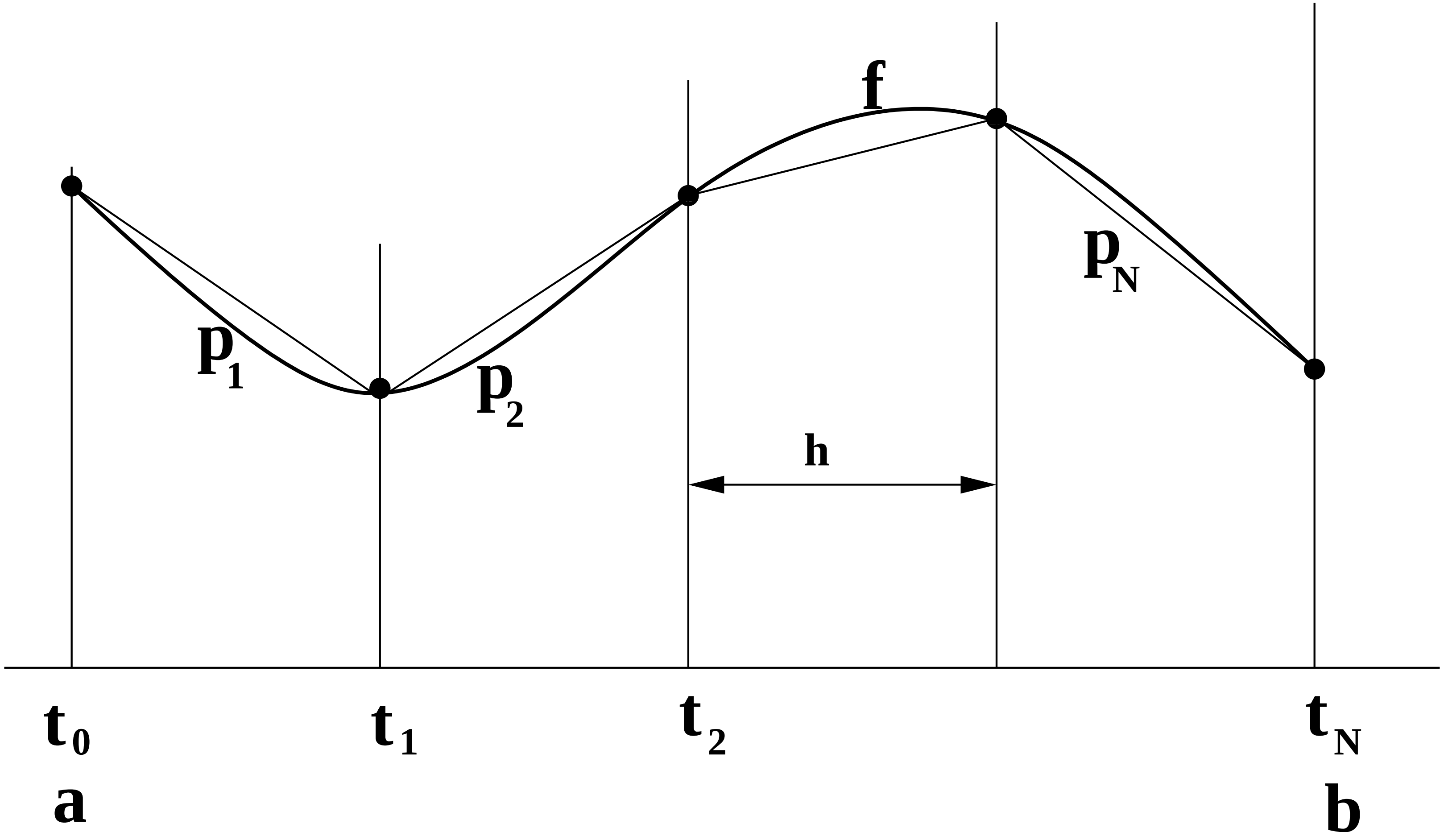
where $p_j \in \mathbb{P}_n$ interpolates f at $n+1$ points in $[t_{j-1}, t_j]$.



The Trapezoidal Rule.

If $n = 1$, and if $p_j \in \mathbb{P}_1$ interpolates f at t_{j-1} and t_j , then

$$\int_{t_{j-1}}^{t_j} p_j(x)\,dx \;=\; \frac{h}{2}\,(f_{j-1} + f_j)\;, \qquad (\textit{local integration formula})\;.$$



The *composite* integration formula then becomes

$$\begin{aligned}
\int_a^b f(x)\, dx \;&\cong\; \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} p_j(x)\, dx \\
&=\; \sum_{j=1}^{N} \frac{h}{2} \left(f_{j-1} + f_j\right) \\
&=\; h \left(\frac{1}{2} f_0 \;+\; f_1 \;+\; \cdots \;+\; f_{N-1} \;+\; \frac{1}{2} f_N \right) ,
\end{aligned}$$

where $f_j \;\equiv\; f(t_j)$.

This is the well-known *Trapezoidal Rule*.

In general

$$p_j(x) \quad = \quad \sum_{i=0}^{n} f(x_{ji})\ \ell_{ji}(x)\ ,$$

where

$$\ell_{ji}(x) \quad = \quad \prod_{k=0,k\neq i}^{n} \frac{x - x_{jk}}{x_{ji} - x_{jk}}\ .$$

Thus we have the approximation

$$\int_{t_{j-1}}^{t_j} f(x)\ dx \ \ \cong\ \ \int_{t_{j-1}}^{t_j} p_j(x)\ dx \ = \ \sum_{i=0}^{n} f(x_{ji}) \int_{t_{j-1}}^{t_j} \ell_{ji}(x)\ dx\ .$$

The integrals

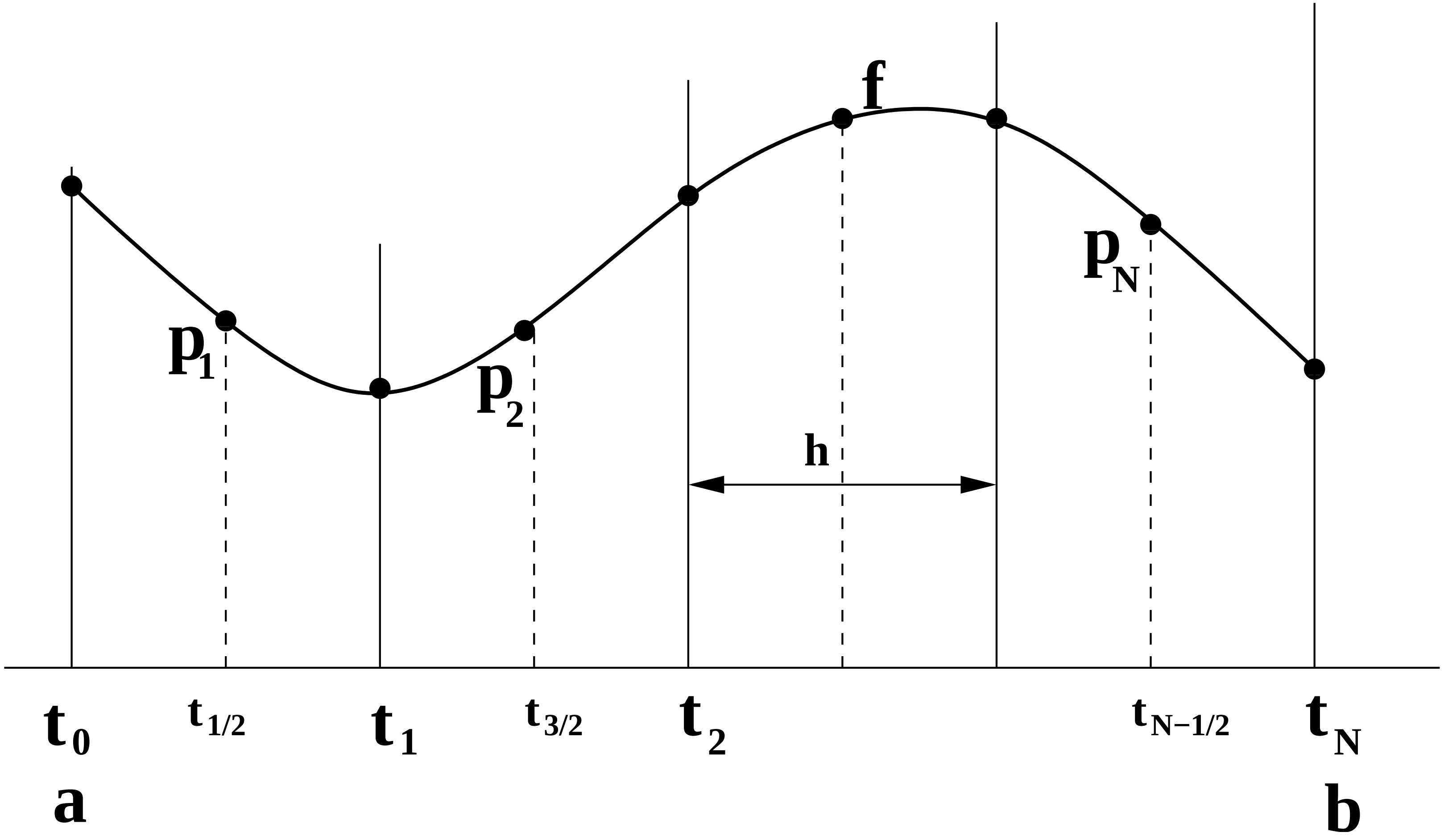
$$\int_{t_{j-1}}^{t_j} \ell_{ji}(x)\ dx\ ,$$

are called the *weights* in the local integration formula.

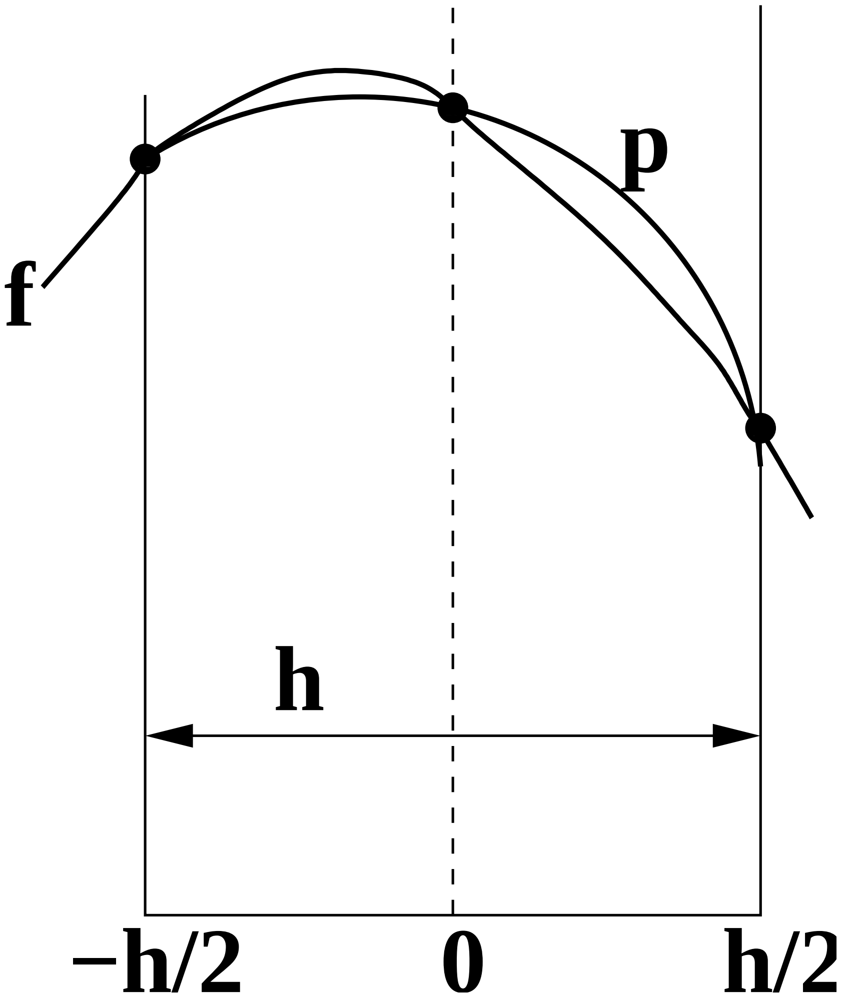
Simpson's Rule.

Let $n = 2$, and in each subinterval $[t_{j-1}, t_j]$ choose the interpolation points

$$t_{j-1} \ , \qquad t_{j-\frac{1}{2}} \ \equiv \ \frac{1}{2}(t_{j-1} + t_j) \ , \quad \text{and} \quad t_j \ .$$



It is convenient to derive the weights for the *reference interval* $[-h/2\ ,\ h/2]$:



The *weights* are

$$\int_{-h/2}^{h/2} \frac{(x-0)\ (x-\frac{h}{2})}{(-\frac{h}{2}-0)\ (-\frac{h}{2}-\frac{h}{2})}\ dx \ = \ \frac{h}{6}\ ,$$

$$\int_{-h/2}^{h/2} \frac{(x+\frac{h}{2})\ (x-\frac{h}{2})}{(0+\frac{h}{2})\ (0-\frac{h}{2})}\ dx \ = \ \frac{4h}{6}\ ,$$

$$\int_{-h/2}^{h/2} \frac{(x+\frac{h}{2})\ (x-0)}{(\frac{h}{2}+\frac{h}{2})\ (\frac{h}{2}-0)}\ dx \ = \ \frac{h}{6}\ .$$

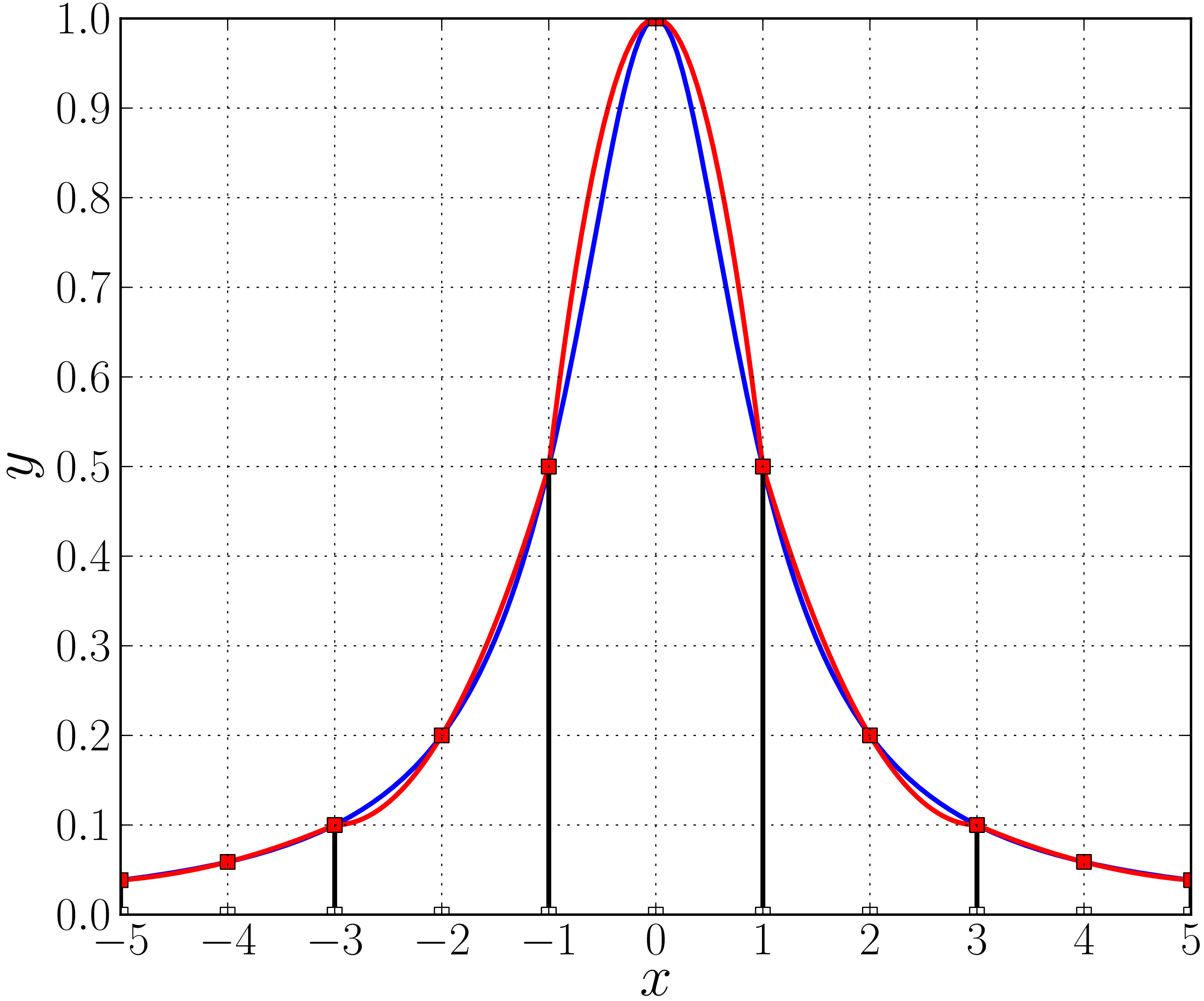
(**Check !**)

With uniformly spaced $\{t_j\}_{j=0}^N$, the *composite* integration formula becomes

$$\int_a^b f(x)\,dx \;\cong\; \sum_{j=1}^N \frac{h}{6}\,(f_{j-1} \;+\; 4f_{j-\frac{1}{2}} \;+\; f_j)$$

$$= \frac{h}{6}\,(\,f_0 + 4f_{\frac{1}{2}} + 2f_1 + 4f_{1\frac{1}{2}} + 2f_2 + \cdots + 2f_{N-1} + 4f_{N-\frac{1}{2}} + f_N\,)\ .$$

This formula is known as *Simpson's Rule*.



The local polynomials (red) in Simpson's Rule for numerically integrating $f(x) = \frac{1}{1+x^2}$ (blue) .

THEOREM :

The error in the composite integration formula, based on local polynomial interpolation at $n+1$ *equally spaced local points* , satisfies the estimate

$$| \int_a^b f(x)\ dx \ - \ \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} p_j(x)\ dx \ | \quad \leq \quad \frac{\| f^{(n+1)} \|_\infty \ \ h^{n+1}\ C_n\ (b-a)}{(n+1)!\ 2^{n+1}} \ ,$$

where

$$h \ = \ \frac{b-a}{N} \ ,$$

and where the value of

$$C_n \ \equiv \ \max_{[-1,1]} \ | \ w_{n+1}(x) \ | \ \ ,$$

can be found in the Table on Page 183.

PROOF : The *local error* is

$$
\begin{aligned}
| \int_{t_{j-1}}^{t_j} f(x)\ dx \ - \ \int_{t_{j-1}}^{t_j} p_j(x)\ dx \ | \ &= \ | \int_{t_{j-1}}^{t_j} f(x) - p_j(x)\ dx \ | \\
&= \ | \int_{t_{j-1}}^{t_j} \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \ \prod_{i=0}^{n} (x - x_{ji})\ dx \ | \\
&\leq \ | \ t_j - t_{j-1} \ | \ \frac{\| \ f^{(n+1)} \ \|_\infty}{(n+1)!} \ \left(\frac{h}{2}\right)^{n+1} C_n \\
&= \ \frac{\| \ f^{(n+1)} \ \|_\infty \ \ h^{n+2} \ C_n}{(n+1)! \ 2^{n+1}} \ .
\end{aligned}
$$

The error in the *composite formula* is now easily determined :

$$\begin{aligned} \mid \int_a^b f(x)\ dx - \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} p_j(x)\ dx \mid \quad &= \quad \mid \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} f(x) - p_j(x)\ dx \mid \\ &\leq \quad \sum_{j=1}^{N} \mid \int_{t_{j-1}}^{t_j} f(x) - p_j(x)\ dx \mid \\ &= \quad N \ \frac{\parallel f^{(n+1)} \parallel_\infty \ h^{n+2}\ C_n}{(n+1)!\ 2^{n+1}} \\ &= \quad \frac{\parallel f^{(n+1)} \parallel_\infty \ h^{n+1}\ C_n\ (b-a)}{(n+1)!\ 2^{n+1}} \ , \end{aligned}$$

where the last step uses the fact that

$$h = \frac{b-a}{N} \ , \qquad i.e. \ , \qquad N = \frac{b-a}{h} \ .$$

QED !

$$| \int_a^b f(x)\ dx \ - \ \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} p_j(x)\ dx\ | \quad \leq \quad \frac{\| f^{(n+1)} \|_\infty \ h^{n+1}\ C_n\ (b-a)}{(n+1)!\ 2^{n+1}}\ .$$

NOTE :

- The *order of accuracy* is at least $n+1$.
- We say that the method is $\mathcal{O}(h^{n+1})$.
- For equally spaced local interpolation points we find C_n in the Table.
- For local Chebyshev points we have $C_n = 2^{-n}$.
- The actual order may be higher.
- For *local Gausss points* we have much higher order of accuracy !

EXAMPLES :

For the *Trapezoidal Rule* , where $n = 1$, $C_1 = 1$, we get the error bound

$$\frac{h^2}{8} \ \| f'' \|_\infty \ (b - a) \ .$$

Indeed the Trapezoidal Rule is $\mathcal{O}(h^2)$.

For *Simpson's Rule* , where $n = 2$, $C_2 = 0.3849$, we get the bound

$$8.01875 \ \cdot \ 10^{-3} \ h^3 \ \| f^{(3)} \|_\infty \ (b - a) \ .$$

The actual order of Simpson's Rule is higher, namely, $\mathcal{O}(h^4)$.

EXAMPLE :

Taylor expand for the precise local error in Simpson's Rule :

$$\int_{-h/2}^{h/2} f(x)\ dx \ - \ \frac{h}{6}\ \Big(\ f(-\frac{h}{2})\ +\ 4f(0)\ +\ f(\frac{h}{2})\ \Big)$$

$$= \ \int_{-h/2}^{h/2} f_0 \ + \ xf_0' \ + \ \frac{x^2}{2}f_0'' \ + \ \frac{x^3}{6}f_0''' \ + \ \frac{x^4}{24}f_0'''' \ + \ \cdots \ dx$$

$$- \ \frac{h}{6} \ \Big(\ f_0 \ - \ (\frac{h}{2})f_0' \ + \ \frac{1}{2}(\frac{h}{2})^2 f_0'' \ - \ \frac{1}{6}(\frac{h}{2})^3 f_0''' \ + \ \frac{1}{24}(\frac{h}{2})^4 f_0'''' \ + \ \cdots$$

$$+ \ 4f_0$$

$$+ \ f_0 \ + \ (\frac{h}{2})f_0' \ + \ \frac{1}{2}(\frac{h}{2})^2 f_0'' \ + \ \frac{1}{6}(\frac{h}{2})^3 f_0''' \ + \ \frac{1}{24}(\frac{h}{2})^4 f_0'''' \ + \ \cdots \ \Big)$$

where $f_0 \equiv f(0)$, *etc.*

$$= \left(x f_0 \;+\; \frac{x^2}{2} f_0' \;+\; \frac{x^3}{6} f_0'' \;+\; \frac{x^4}{24} f_0''' \;+\; \frac{x^5}{120} f_0'''' \;+\; \cdots\right)\Big|_{-h/2}^{h/2} \\ -\; \frac{h}{6}\left(6 f_0 \;+\; \frac{h^2}{4} f_0'' \;+\; \frac{h^4}{192} f_0'''' \;+\; \cdots\right)$$

$$= \left(h f_0 \;+\; \frac{(h/2)^3}{3} f_0'' \;+\; \frac{(h/2)^5}{60} f_0'''' \;+\; \cdots\right) \\ -\; \left(h f_0 \;+\; \frac{h^3}{24} f_0'' \;+\; \frac{h^5}{1152} f_0'''' \;+\; \cdots\right)$$

$$= -\frac{h^5}{2880}\, f_0'''' \;+\; \text{higher order terms.}$$

Thus the leading error term of the composite Simpson's Rule is bounded by

$$\frac{h^4}{2880} \;\| \, f'''' \, \|_\infty \;(b-a)\ .$$

EXERCISE :

- The *Local Midpoint Rule*, for numerically integrating a function $f(x)$ over the reference interval $[-h/2, h/2]$, is given by

$$\int_{-h/2}^{h/2} f(x)\, dx \ \cong \ hf(0) \ .$$

Use Taylor expansion to determine the error in this local formula.

Write down the formula for the *Composite Midpoint Rule* for integrating $f(x)$ over a general interval $[a, b]$.

Derive an error formula for the composite formula.

How big must N be for the global error to be less than 10^{-6}, when integrating $f(x) = \sin(x)$ over the interval $[0, 1]$?

EXERCISE :

- The *local Trapezoidal Rule* for the reference interval $[-h/2, h/2]$ is

$$\int_{-h/2}^{h/2} f(x)\ dx \ \cong\ \frac{h}{2}\Big[f(-h/2) + f(h/2)\Big]\ .$$

Use Taylor expansions to derive the local error formula.

Let $h = (b-a)/N$ and $x_k = a + k\ h$, for $k = 0, 1, 2, 3, \ \cdots\ , N$.
Then the *composite Trapezoidal Rule* is given by

$$\int_a^b f(x)dx \ \cong\ \frac{h}{2}\Big[f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{N-1}) + f(x_N)\Big].$$

Based on the local error, derive an upper bound on the global error.

How big must N be for the global error to be less than 10^{-6}, when integrating $f(x) = \sin(x)$ over the interval $[0, 1]$?

THE GAUSS QUADRATURE THEOREM :

If in each subinterval $[t_{j-1}, t_j]$ the interpolation points $\{x_{ji}\}_{i=0}^{n}$ are taken as

the *zeroes* of the $(n+1)$st *orthogonal polynomial* $e_{n+1}(x)$,

(relative to $[t_{j-1}, t_j]$) ,

then the composite integration formula is $\mathcal{O}(h^{2n+2})$.

NOTE :

- Such integration formulas are known as *Gauss Quadrature* formulas.

- The points $\{x_{ji}\}_{i=0}^{n}$ are the *Gauss points* .

- The order improves from $\mathcal{O}(h^{n+1})$ to $\mathcal{O}(h^{2n+2})$ for Gauss points.

EXAMPLE : The case $n = 1$:

Relative to the interval $[-1, 1]$, the second degree orthogonal polynomial is

$$e_2(x) \quad = \quad x^2 \ - \ \frac{1}{3} \ .$$

The two Gauss points relative to $[-1, 1]$ are the zeroes of $e_2(x)$, *i.e.*,

$$x_0 \ = \ - \frac{\sqrt{3}}{3} \qquad \text{and} \qquad x_1 \ = \ \frac{\sqrt{3}}{3} \ .$$

Relative to $[t_{j-1}, t_j]$ the *Gauss points* are

$$x_{j,0} \ = \ t_{j-\frac{1}{2}} \ - \ h\frac{\sqrt{3}}{6} \ , \qquad \text{and} \qquad x_{j,1} \ = \ t_{j-\frac{1}{2}} \ + \ h\frac{\sqrt{3}}{6} \ ,$$

where $t_{j-\frac{1}{2}} \ \equiv \ \frac{1}{2}(t_{j-1} + t_j)$.

Relative to the *reference interval* $I_h \equiv [-h/2\ ,\ h/2]$ the Gauss points are

$$x_0 = -\frac{h\sqrt{3}}{6} \qquad \text{and} \qquad x_1 = \frac{h\sqrt{3}}{6}\ ,$$

with interpolating polynomial

$$p(x) = f(x_0)\,\ell_0(x) + f(x_1)\,\ell_1(x)\ ,$$

where

$$\ell_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{x - h\sqrt{3}/6}{-h\sqrt{3}/3}\ ,$$

and

$$\ell_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x + h\sqrt{3}/6}{h\sqrt{3}/3}\ .$$

The *local* integration formula is

$$\int_{-h/2}^{h/2} f(x)\ dx \ \cong \ f(x_0) \int_{-h/2}^{h/2} \ell_0(x)\ dx \ + \ f(x_1) \int_{-h/2}^{h/2} \ell_1(x)\ dx \ .$$

Integrating $\ell_0(x)$ and $\ell_1(x)$, we find

$$\int_{-h/2}^{h/2} f(x)\ dx \ \cong \ \frac{h}{2}\ f(x_0) \ + \ \frac{h}{2}\ f(x_1) \ .$$

Hence the *composite* two point Gauss quadrature formula is

$$\int_a^b f(x)\ dx \ \cong \ \frac{h}{2} \sum_{j=1}^{N} [f(x_{j,0}) \ + \ f(x_{j,1})] \ .$$

By the Theorem this integration formula is $\mathcal{O}(h^4)$.

PROOF (of the Gauss Quadrature Theorem.)

The local error for the reference interval $I_h \equiv [-h/2 \ , \ h/2]$ is

$$\int_{-h/2}^{h/2} f(x) - p(x) \ dx \ ,$$

where $p \in \mathbb{P}_n$ interpolates $f(x)$ at Gauss points $\{x_i\}_{i=0}^{n}$ (relative to I_h) .

By the Lagrange Interpolation Theorem

$$\begin{aligned} \int_{-h/2}^{h/2} f(x) - p(x) \ dx \ &= \ \int_{-h/2}^{h/2} \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{i=0}^{n}(x - x_i) \ dx \\ &= \ \int_{-h/2}^{h/2} c(x) \ e_{n+1}(x) \ dx \ , \end{aligned}$$

where

$$c(x) \ \equiv \ \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \qquad \text{and} \qquad e_{n+1}(x) \ = \ \prod_{i=0}^{n}(x - x_i) \ .$$

Note that e_{n+1} is the $(n+1)$*st* orthogonal polynomial (relative to I_h) .

FACT : If $f(x)$ is very smooth then $c(x)$ has $n+1$ continuous derivatives.

Thus we can Taylor expand :

$$c(x) \;=\; \sum_{k=0}^{n} \frac{x^k}{k!}\, c^{(k)}(0) \;+\; \frac{x^{n+1}}{(n+1)!}\, c^{(n+1)}(\eta(x)) \; .$$

Call the remainder $r(x)$ and use the fact that each summation term is in $\mathbb{P}_n$:

$$c(x) \;=\; \sum_{k=0}^{n} c_k \; e_k(x) \;+\; r(x) \; ,$$

where e_k is the k^{th} orthogonal polynomial relative to I_h .

(Recall that the $\{e_k\}_{k=0}^{n}$ form an orthogonal basis of $\mathbb{P}_n$.)

We have

$$\int_{-h/2}^{h/2} f(x) - p(x)\ dx \ = \ \int_{-h/2}^{h/2} c(x)\ e_{n+1}(x)\ dx\ ,$$

and

$$c(x) \ = \ \sum_{k=0}^{n} c_k\ e_k(x) \ + \ r(x)\ .$$

It follows that

$$|\int_{-h/2}^{h/2} f(x) \ - \ p(x)\ dx\ | \ = \ |\int_{-h/2}^{h/2} [\ \sum_{k=0}^{n} c_k\ e_k(x) \ + \ r(x)]\ e_{n+1}(x)\ dx\ |$$

$$= \ |\sum_{k=0}^{n} c_k \int_{-h/2}^{h/2} e_k(x)\ e_{n+1}(x)\ dx \ + \ \int_{-h/2}^{h/2} r(x)\ e_{n+1}(x)\ dx\ |\ .$$

Note that all terms in the summation term are zero by orthogonality, so that

$$\begin{aligned}
\mid \int_{-h/2}^{h/2} f(x) \quad - \quad p(x)\ dx \mid \quad &= \quad \mid \int_{-h/2}^{h/2} r(x)\ e_{n+1}(x)\ dx \mid \\
&= \ \mid \int_{-h/2}^{h/2} \frac{x^{n+1}}{(n+1)!} \ \ c^{(n+1)}\ (\eta(x)) \ \ \prod_{i=0}^{n}(x - x_i)\ dx \mid \\
&\leq \ h \ \max_{x \in I_h} \mid \frac{x^{n+1}}{(n+1)!} \ \ c^{(n+1)}(\eta(x)) \ \ \prod_{i=0}^{n}(x - x_i) \ \mid \\
&\leq \ h \ \frac{(h/2)^{n+1}}{(n+1)!} \ \ \max_{x \in I_h} \mid c^{(n+1)}(x) \mid \ \ h^{n+1} \\
&= \ \frac{h^{2n+3}}{2^{n+1}(n+1)!} \ \ \max_{x \in I_h} \mid c^{(n+1)}(x) \mid .
\end{aligned}$$

Hence the local integration formula is $\mathcal{O}(h^{2n+3})$.

As before, this implies that the composite formula is $\mathcal{O}(h^{2n+2})$. **QED !**

EXERCISE :

- Give complete details on the derivation of the *local* 3-point Gauss integration formula. Also write down the *composite* 3-point Gauss formula for integrating a function $f(x)$ over a general interval $[a, b]$.

- Are the following True or False for any sufficiently smooth $f(x)$?

  - The order of accuracy of a general *composite* $(n+1)$-point integration formula for $f(x)$ is at least $\mathcal{O}(h^{n+1})$.

  - The order of accuracy of the *composite* $(n+1)$-point Gauss formula for integrating $f(x)$ is $\mathcal{O}(h^{2n+4})$.

  - The order of accuracy of the composite 2-Point Gauss formula is the same as the order of accuracy of the composite Simpson formula.

DISCRETE LEAST SQUARES APPROXIMATION

We have solved the *continuous least squares problem* :

Given $f(x)$ on $[-1,1]$, find a polynomial $p(x) \in \mathbb{P}_n$ that minimizes

$$\| \, p - f \, \|_2^2 \quad \equiv \quad \int_{-1}^{1} [p(x) - f(x)]^2 \; dx \; .$$

Next we solve the *discrete least squares problem* :

Given a set of discrete data points

$$\{ \ (x_i, y_i) \ \}_{i=1}^N \ ,$$

find $p \in \mathbb{P}_n$ such that

$$e_L \quad \equiv \quad \sum_{i=1}^N \ [p(x_i) - y_i]^2$$

is minimized.

More generally, find

$$p(x) \ = \ \sum_{i=0}^n \ a_i \ \phi_i(x) \ ,$$

(not necessarily a polynomial), such that e_L is minimized.

Linear Least Squares

- Suppose we have data on the *daily high temperature* in March.

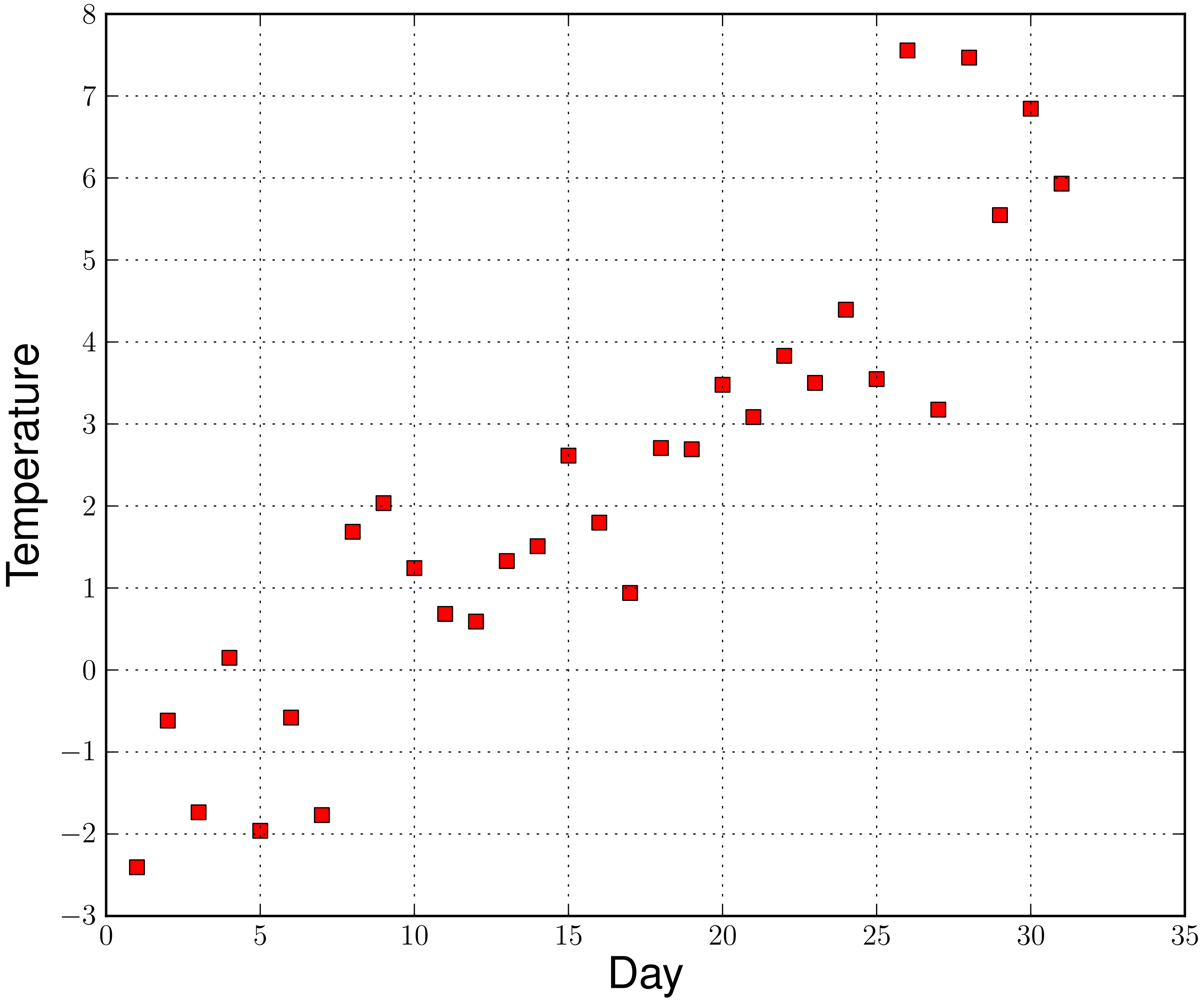
- For each day we compute the *average* high temperature.

- Each average is taken over *a number of years* .

- The (fictitious) data are given in the Table below.

1	-2.4	2	-0.6	3	-1.7	4	0.1	5	-2.0	6	-0.6	7	-1.8
8	1.7	9	2.0	10	1.2	11	0.7	12	0.6	13	1.3	14	1.5
15	2.6	16	1.8	17	0.9	18	2.7	19	2.7	20	3.5	21	3.1
22	3.8	23	3.5	24	4.4	25	3.5	26	7.6	27	3.2	28	7.5
29	5.5	30	6.8	31	5.9								

Average daily high temperature in Montreal in March



Average daily high temperature in Montreal in March

Suppose that :

- We believe these temperatures basically increase *linearly* .

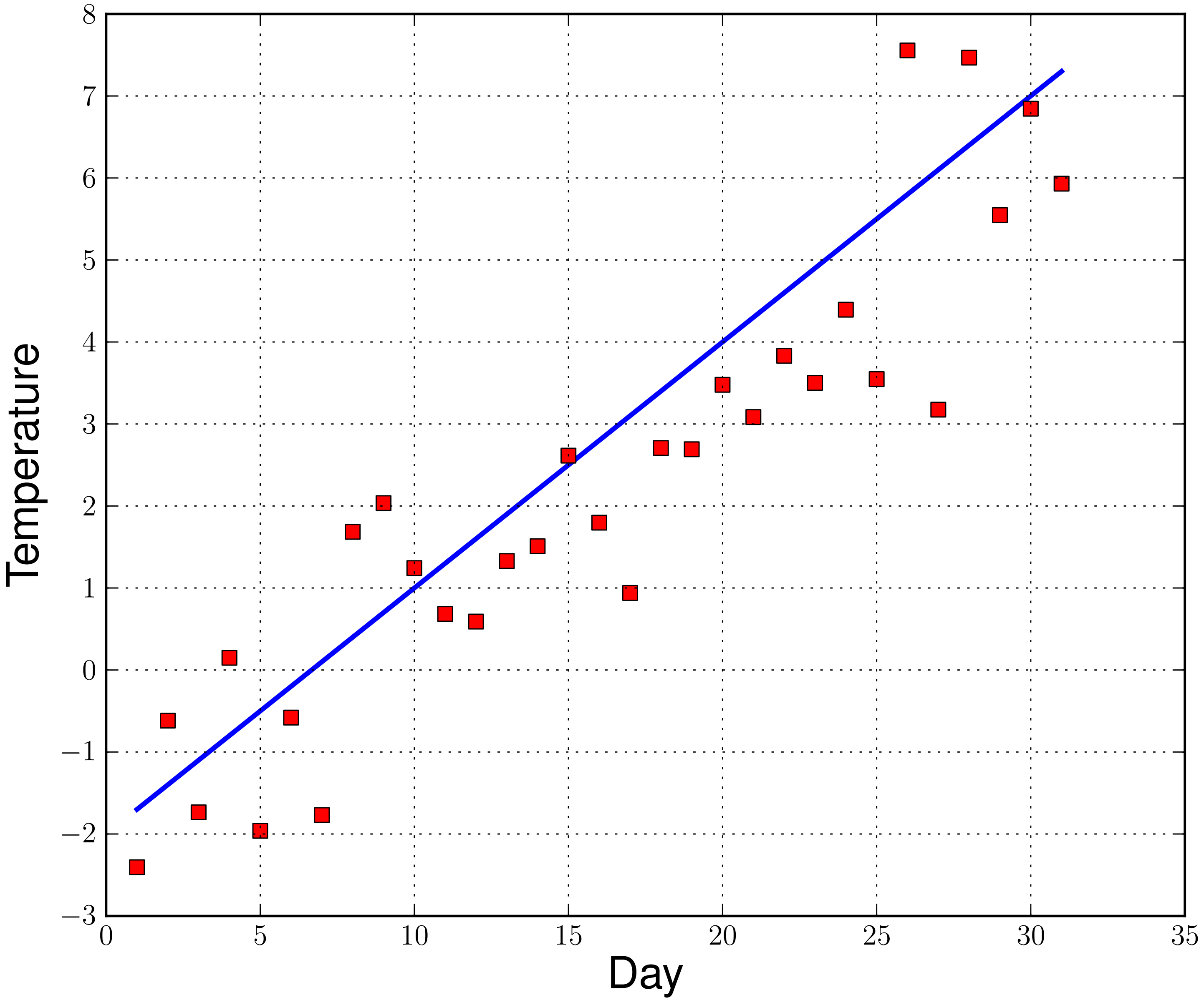
- Thus we believe in a relation

$$T_k \quad = \quad c_1 \ + \ c_2 \, k \ , \qquad k \ = \ 1, \ 2, \ \cdots \ , 31 \ .$$

- The *deviations* from linearity come from *random influences* .

- These random influences can be due to *many factors* .

- We want to determine ”*the best* ” linear approximation.



Average daily high temperatures, with a *linear approximation* .

- There are many ways to determine such a linear approximation.

- Often used is the *least squares method* .

- This method determines c_1 and c_2 that *minimize*

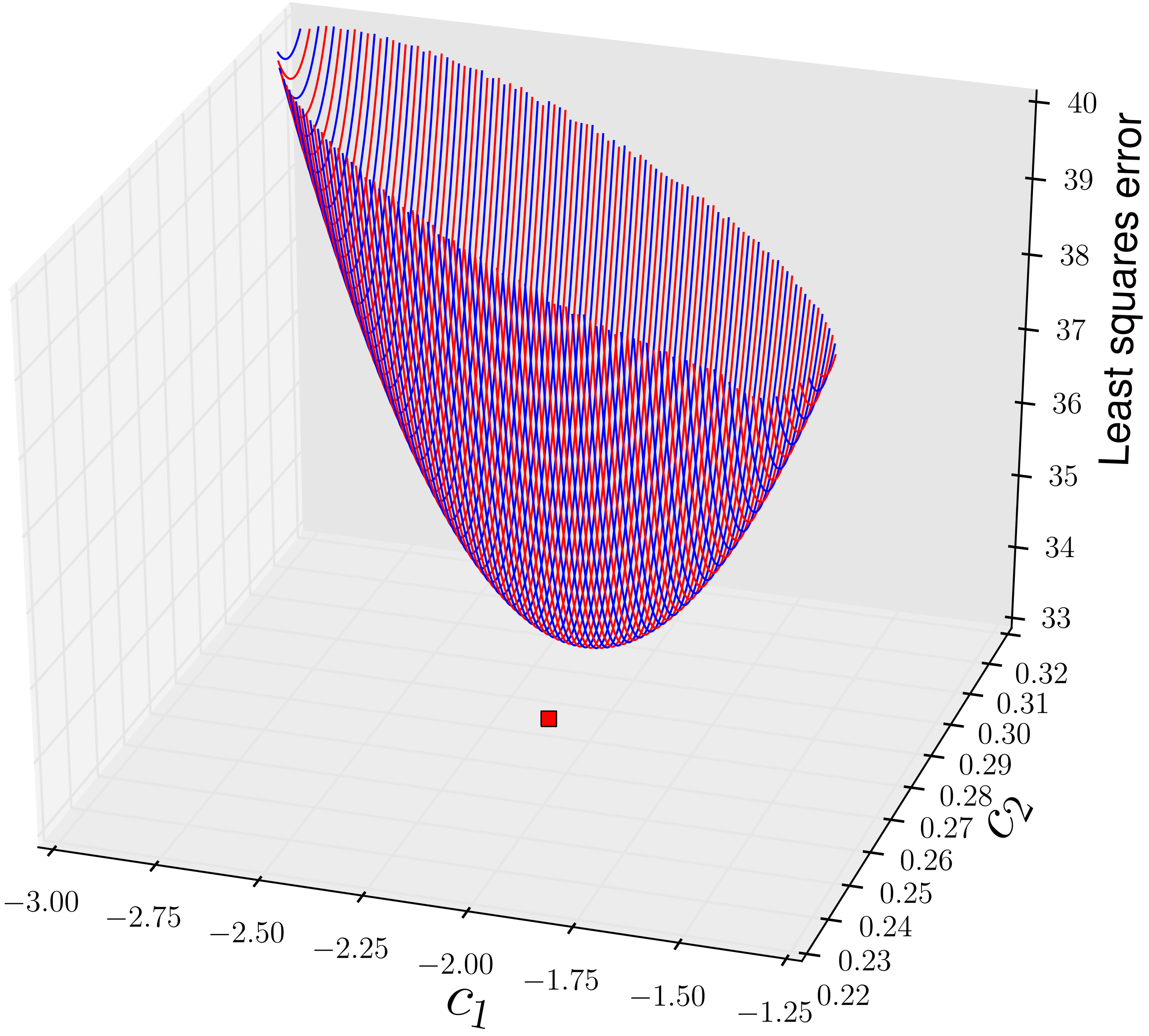
$$\sum_{k=1}^{N} (\ T_k \ - \ (c_1 \ + \ c_2 x_k) \ )^2 \ ,$$

where, in our example, $N = 31$ and $x_k = k$.

- To do so set the *partial derivatives* w.r.t. c_1 and c_2 to zero :

$$\text{w.r.t. } c_1 \ : \qquad -\,2 \sum_{k=1}^{N} (\ T_k \ - \ (c_1 \ + \ c_2 x_k) \ ) \ = \ 0 \ ,$$

$$\text{w.r.t. } c_2 \ : \qquad -\,2 \sum_{k=1}^{N} x_k \ (\ T_k \ - \ (c_1 \ + \ c_2 x_k) \ ) \ = \ 0 \ .$$



The least squares error versus c_1 and c_2 .

From setting the partial derivatives to zero, we have

$$\sum_{k=1}^{N}(\ T_k \ - \ (c_1 \ + \ c_2 x_k) \ ) \ = \ 0 \quad , \quad \sum_{k=1}^{N} x_k \ (\ T_k \ - \ (c_1 \ + \ c_2 x_k) \ ) \ = \ 0 \ .$$

Solving these two equations for c_1 and c_2 gives

$$c_2 \ = \ \frac{\sum_{k=1}^{N} x_k T_k \ - \ \bar{x} \sum_{k=1}^{N} T_k}{\sum_{k=1}^{N} x_k^2 \ - \ N\bar{x}^2} \ ,$$

and

$$c_1 \ = \ \bar{T} \ - \ c_2 \ \bar{x} \ ,$$

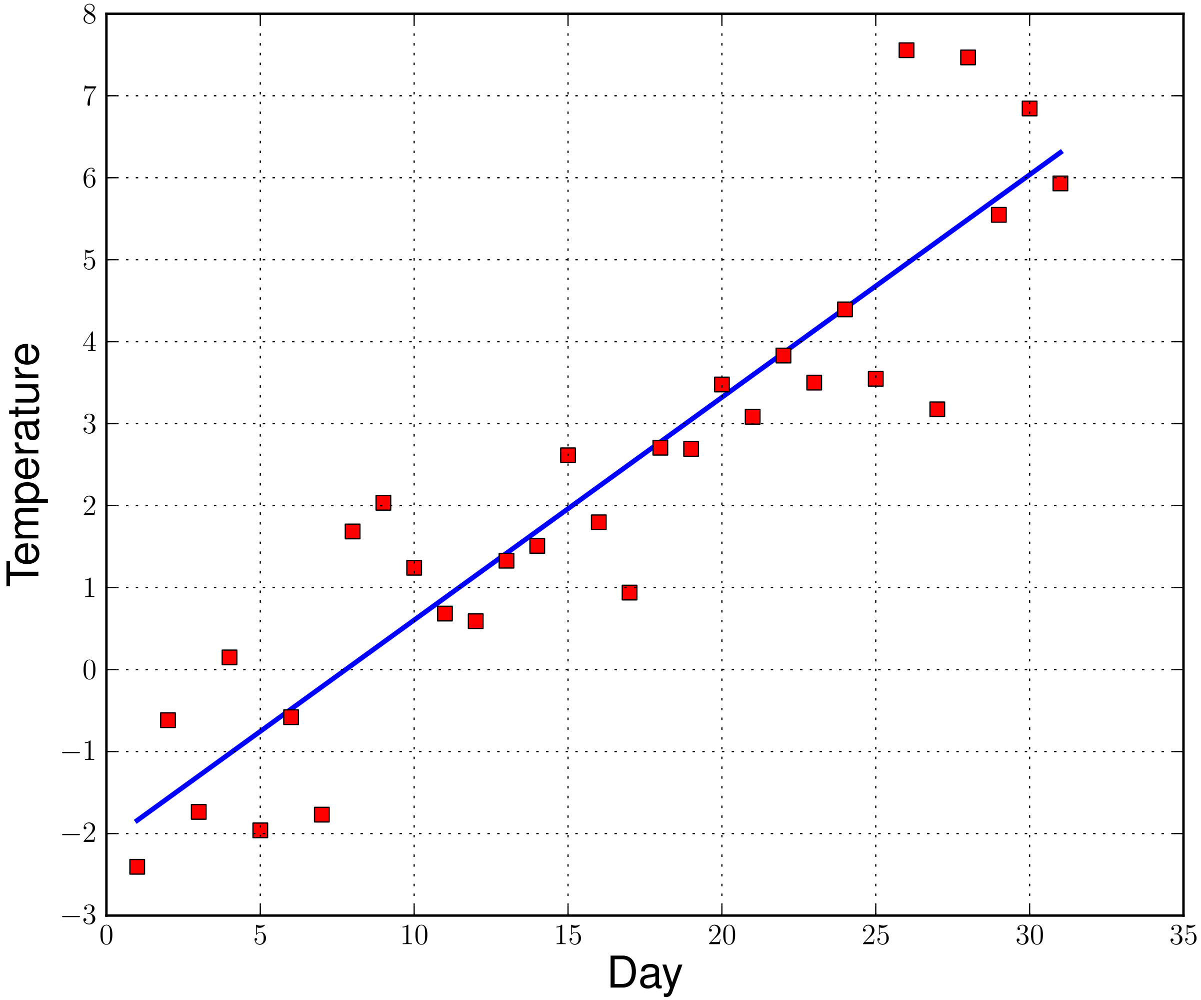
where

$$\bar{x} \ = \ \frac{1}{N} \sum_{k=1}^{N} x_k \quad , \quad \bar{T} \ = \ \frac{1}{N} \sum_{k=1}^{N} T_k \ .$$

EXERCISE : Check these formulas !

EXAMPLE : For our "March temperatures" example, we find

$$c_1 = -\,2.111 \quad \text{and} \quad c_2 = 0.272 \; .$$



Average daily high temperatures, with linear *least squares approximation* .

General Least Squares

Given discrete data points

$$\{\ (x_i, y_i)\ \}_{i=1}^{N}\ ,$$

find the coefficients c_k of the function

$$p(x) \quad \equiv \quad \sum_{k=1}^{n} c_k\ \phi_k(x)\ ,$$

that *minimize* the *least squares error*

$$E_L \quad \equiv \quad \sum_{i=1}^{N} \left(p(x_i) - y_i\right)^2$$

EXAMPLES :

- $p(x) \quad = \quad c_1 \ + \ c_2\ x\ .$ (Already done !)

- $p(x) \quad = \quad c_1 \ + \ c_2\ x \ + \ c_3\ x^2\ .$ (Quadratic approximation)

For any vector $\mathbf{x} \in \mathbb{R}^N$ we have

$$\| \mathbf{x} \|_2^2 \quad \equiv \quad \mathbf{x}^T \mathbf{x} \quad \equiv \quad \sum_{i=1}^{N} x_k^2 \, . \qquad (T \text{ denotes } \textit{transpose}).$$

Then

$$E_L \equiv \sum_{i=1}^{N} [p(x_i) - y_i]^2 = \| \begin{pmatrix} p(x_1) \\ \cdot \\ \cdot \\ p(x_N) \end{pmatrix} - \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{pmatrix} \|_2^2$$

$$= \| \begin{pmatrix} \sum_{i=1}^n c_i \phi_i(x_1) \\ \cdot \\ \cdot \\ \sum_{i=1}^n c_i \phi_i(x_N) \end{pmatrix} - \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{pmatrix} \|_2^2$$

$$= \| \begin{pmatrix} \phi_1(x_1) & \cdot & \phi_n(x_1) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \phi_1(x_N) & \cdot & \phi_n(x_N) \end{pmatrix} \begin{pmatrix} c_1 \\ \cdot \\ c_n \end{pmatrix} - \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{pmatrix} \|_2^2 \quad \equiv \quad \| \mathbf{A}\mathbf{c} - \mathbf{y} \|_2^2 \, .$$

THEOREM :

For the least squares error E_L to be *minimized* we must have

$$\mathbf{A}^T \mathbf{A} \mathbf{c} = \mathbf{A}^T \mathbf{y} .$$

PROOF :

$$\begin{aligned} E_L &= \| \mathbf{Ac} - \mathbf{y} \|_2^2 \\ &= (\mathbf{Ac} - \mathbf{y})^T (\mathbf{Ac} - \mathbf{y}) \\ &= (\mathbf{Ac})^T \mathbf{Ac} - (\mathbf{Ac})^T \mathbf{y} - \mathbf{y}^T \mathbf{Ac} + \mathbf{y}^T \mathbf{y} \\ &= \mathbf{c}^T \mathbf{A}^T \mathbf{Ac} - \mathbf{c}^T \mathbf{A}^T \mathbf{y} - \mathbf{y}^T \mathbf{Ac} + \mathbf{y}^T \mathbf{y} . \end{aligned}$$

PROOF : continued $\cdots$

We had

$$E_L \;=\; \mathbf{c}^T\mathbf{A}^T\mathbf{A}\mathbf{c} \;-\; \mathbf{c}^T\mathbf{A}^T\mathbf{y} \;-\; \mathbf{y}^T\mathbf{A}\mathbf{c} \;+\; \mathbf{y}^T\mathbf{y}\ .$$

For a *minimum* we need

$$\frac{\partial E_L}{\partial \mathbf{c}} \;=\; 0, \qquad i.e., \qquad \frac{\partial E_L}{\partial c_i} \;=\; 0\ , \qquad i \;=\; 0, 1, \cdots, n\ ,$$

which gives

$$\mathbf{c}^T\mathbf{A}^T\mathbf{A} \;+\; (\mathbf{A}^T\mathbf{A}\mathbf{c})^T \;-\; (\mathbf{A}^T\mathbf{y})^T \;-\; \mathbf{y}^T\mathbf{A} \;\;=\;\; 0\ , \qquad \text{(Check !)}$$

i.e.,

$$2\mathbf{c}^T\mathbf{A}^T\mathbf{A} \;-\; 2\mathbf{y}^T\mathbf{A} \;\;=\;\; 0\ ,$$

or

$$\mathbf{c}^T\mathbf{A}^T\mathbf{A} \;=\; \mathbf{y}^T\mathbf{A}\ .$$

Transposing gives

$$\mathbf{A}^T\mathbf{A}\mathbf{c} \;=\; \mathbf{A}^T\mathbf{y}\ . \qquad\qquad \textbf{QED !}$$

EXAMPLE : Given the data points

$$\{ \ (x_i, y_i) \ \}_{i=1}^4 \quad = \quad \{ \ (0,1) \ , \ (1,3) \ , \ (2,2) \ , \ (4,3) \ \} \ ,$$

find the coefficients c_1 and c_2 of $p(x) \ = \ c_1 \ + \ c_2 x$, that minimize

$$E_L \quad \equiv \quad \sum_{i=1}^{4} \ [\ (c_1 + c_2 x_i) \ - \ y_i]^2 \ .$$

SOLUTION : Here $N = 4$, $n = 2$, $\phi_1(x) = 1$, $\phi_2(x) = x$.

Use the Theorem :

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad = \quad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \\ 3 \end{pmatrix} ,$$

or

$$\begin{pmatrix} 4 & 7 \\ 7 & 21 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad = \quad \begin{pmatrix} 9 \\ 19 \end{pmatrix} ,$$

with solution $c_1 = 1.6$ and $c_2 = 0.371429$.

EXAMPLE : Given the same data points, find the coefficients of

$$p(x) \quad = \quad c_1 \ + \ c_2 x \ + \ c_3 x^2 \ ,$$

that minimize

$$E_L \quad \equiv \quad \sum_{i=1}^{4} \ [\ (c_1 \ + \ c_2 \ x_i \ + \ c_3 \ x_i^2) \ - \ y_i \ ]^2 \ .$$

SOLUTION : Here

$$N = 4 \quad , \quad n = 3 \quad , \quad \phi_1(x) = 1 \quad , \quad \phi_2(x) = x \quad , \quad \phi_3(x) = x^2 \ .$$

Use the Theorem :

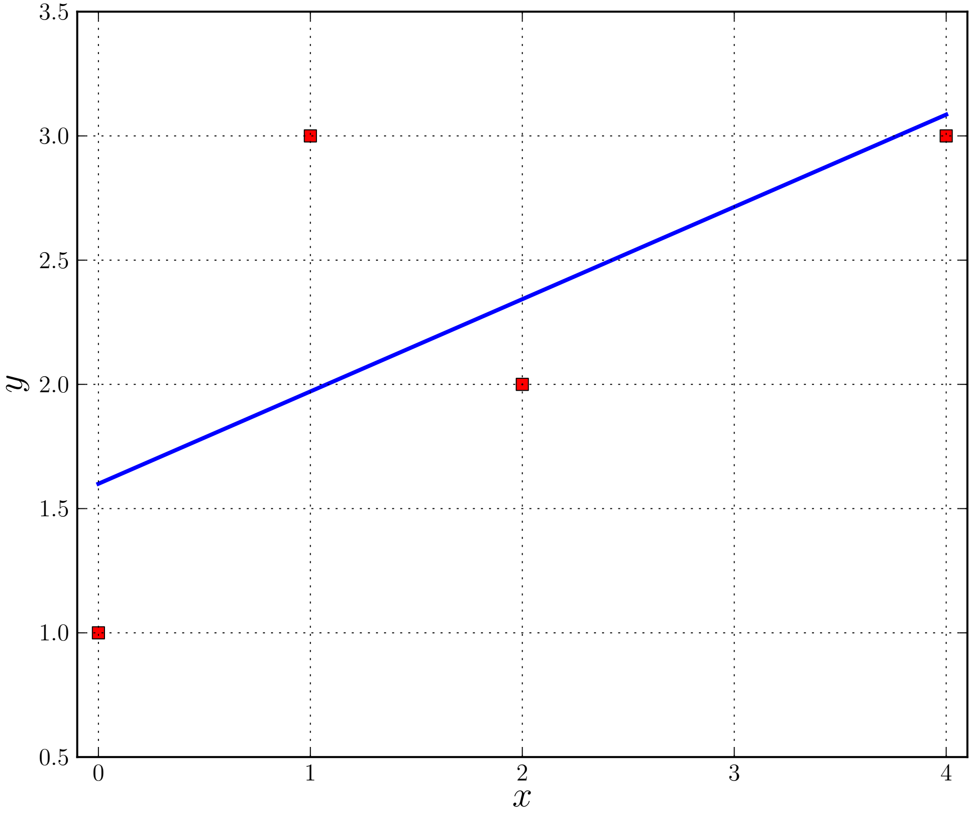
$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \\ 0 & 1 & 4 & 16 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \quad = \quad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \\ 0 & 1 & 4 & 16 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \\ 3 \end{pmatrix} ,$$

or

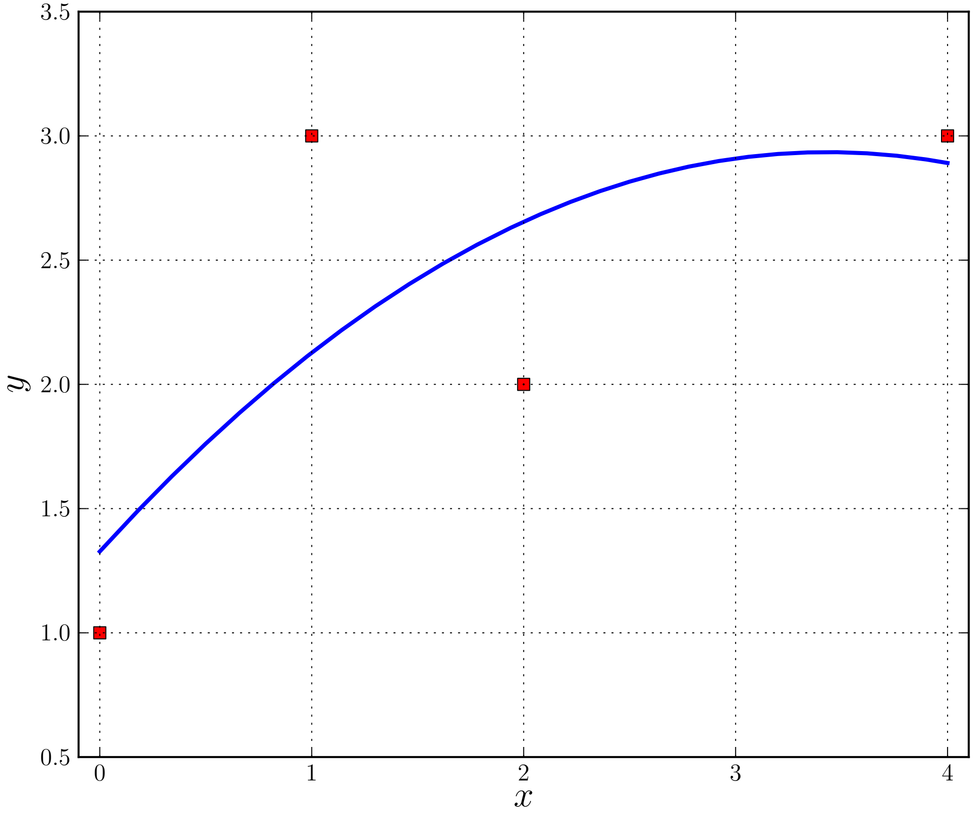
$$\begin{pmatrix} 4 & 7 & 21 \\ 7 & 21 & 73 \\ 21 & 73 & 273 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \quad = \quad \begin{pmatrix} 9 \\ 19 \\ 59 \end{pmatrix} ,$$

with solution $c_1 = 1.32727$, $c_2 = 0.936364$, $c_3 = -0.136364$.

The least squares approximations from the preceding two examples :



$$p(x) \;=\; c_1 \;+\; c_2 x$$



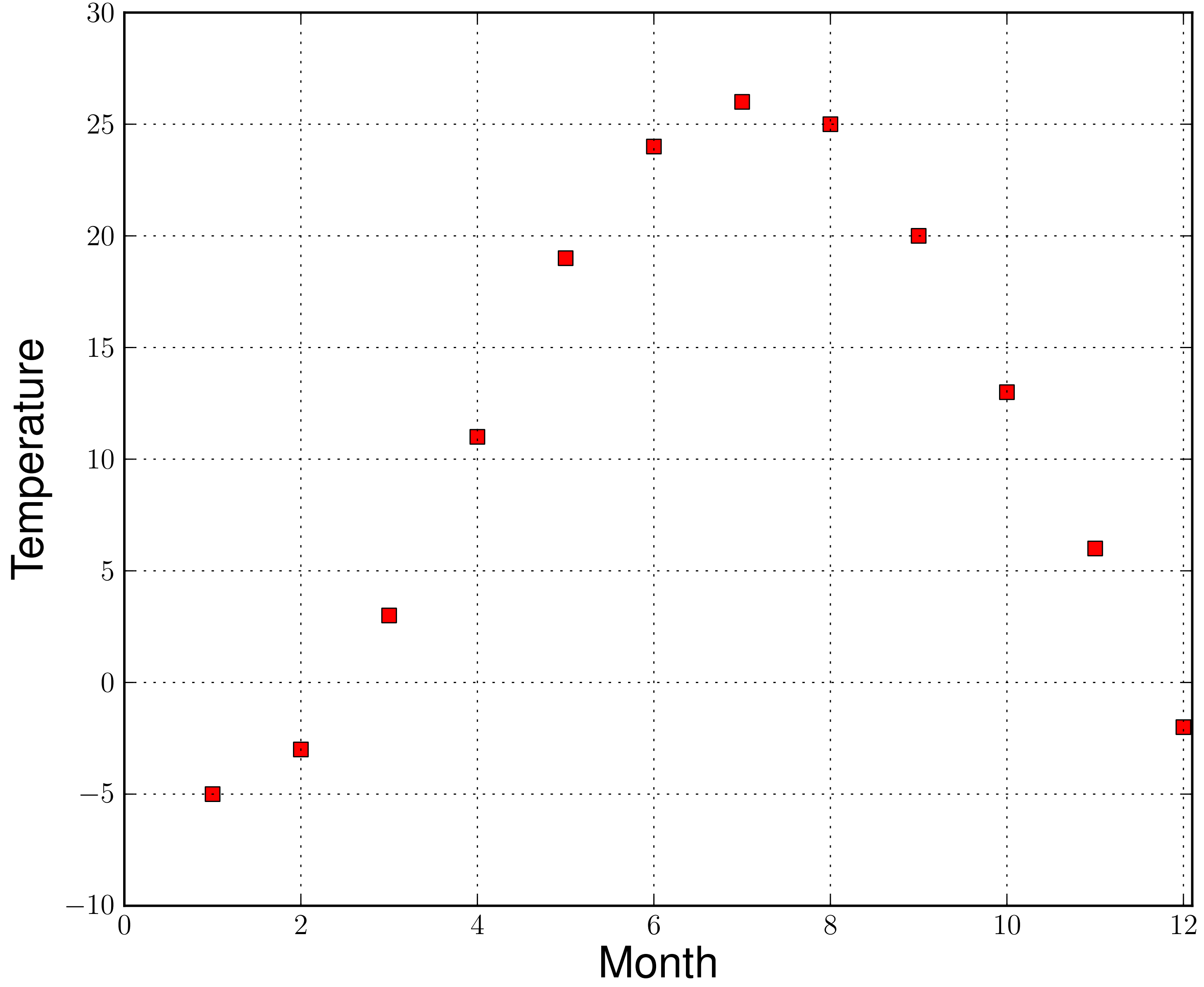
$$p(x) \;=\; c_1 \;+\; c_2 x \;+\; c_3 x^2$$

EXAMPLE : From actual data :

The average daily high temperatures in Montreal (by month) are :

January	-5
February	-3
March	3
April	11
May	19
June	24
July	26
August	25
September	20
October	13
November	6
December	-2

Source : `http://weather.uk.msn.com`



Average daily high temperature in Montreal (by month).

EXAMPLE : continued $\cdots$

The graph suggests using a 3-term *least squares approximation*

$$p(x) \quad = \quad c_1\ \phi_1(x) \quad + \quad c_2\ \phi_2(x) \quad + \quad c_3\ \phi_3(x)\ ,$$

of the form

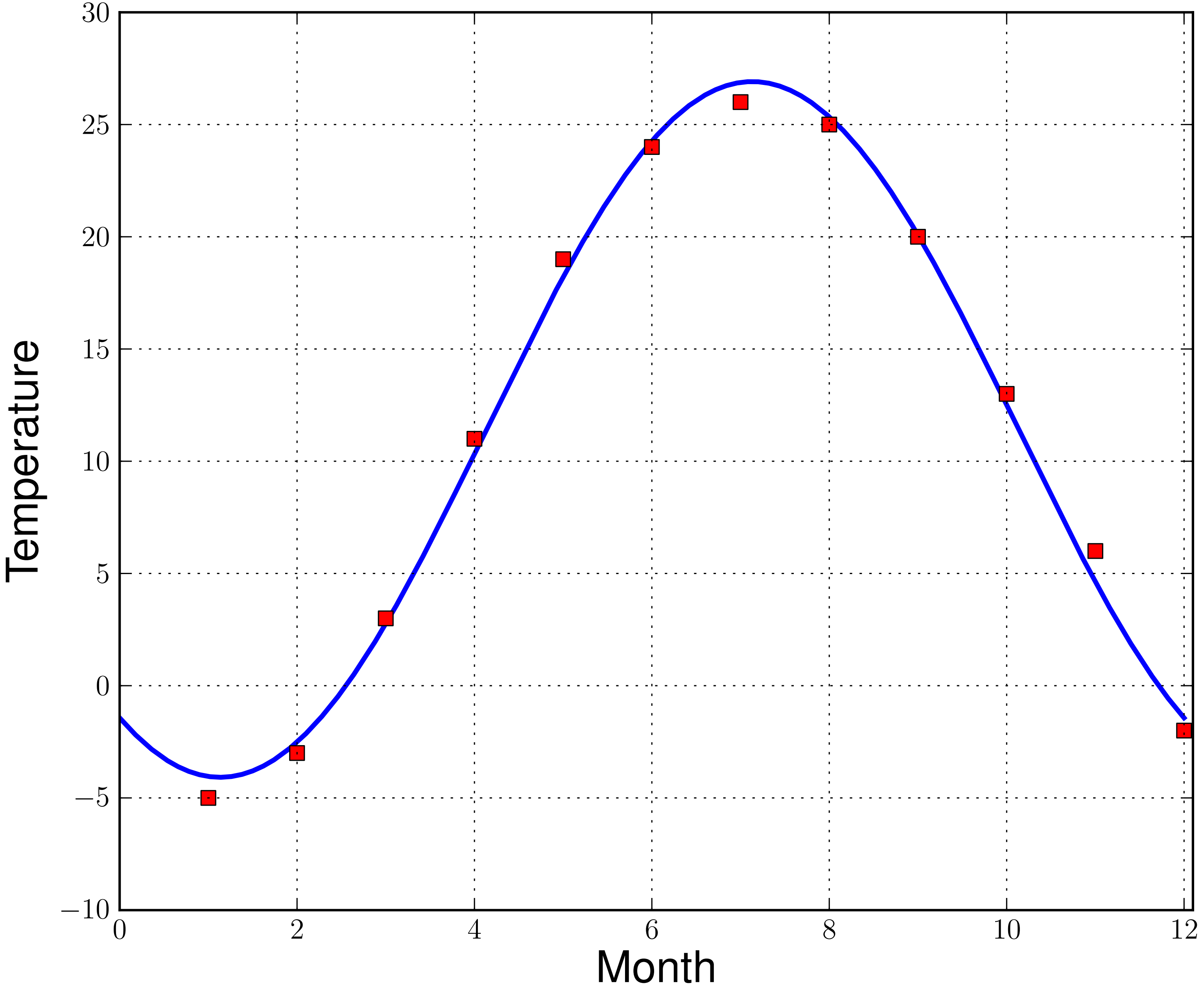
$$p(x) \quad = \quad c_1 \quad + \quad c_2\ \sin(\frac{\pi x}{6}) \quad + \quad c_3\ \cos(\frac{\pi x}{6})\ .$$

QUESTIONS :

- Why include $\phi_2(x) \ = \ \sin(\frac{\pi x}{6})$?
- Why is the argument $\frac{\pi x}{6}$?
- Why include the constant term $\phi_1(x) \ = \ c_1$?
- Why include $\phi_3(x) \ = \ \cos(\frac{\pi x}{6})$?

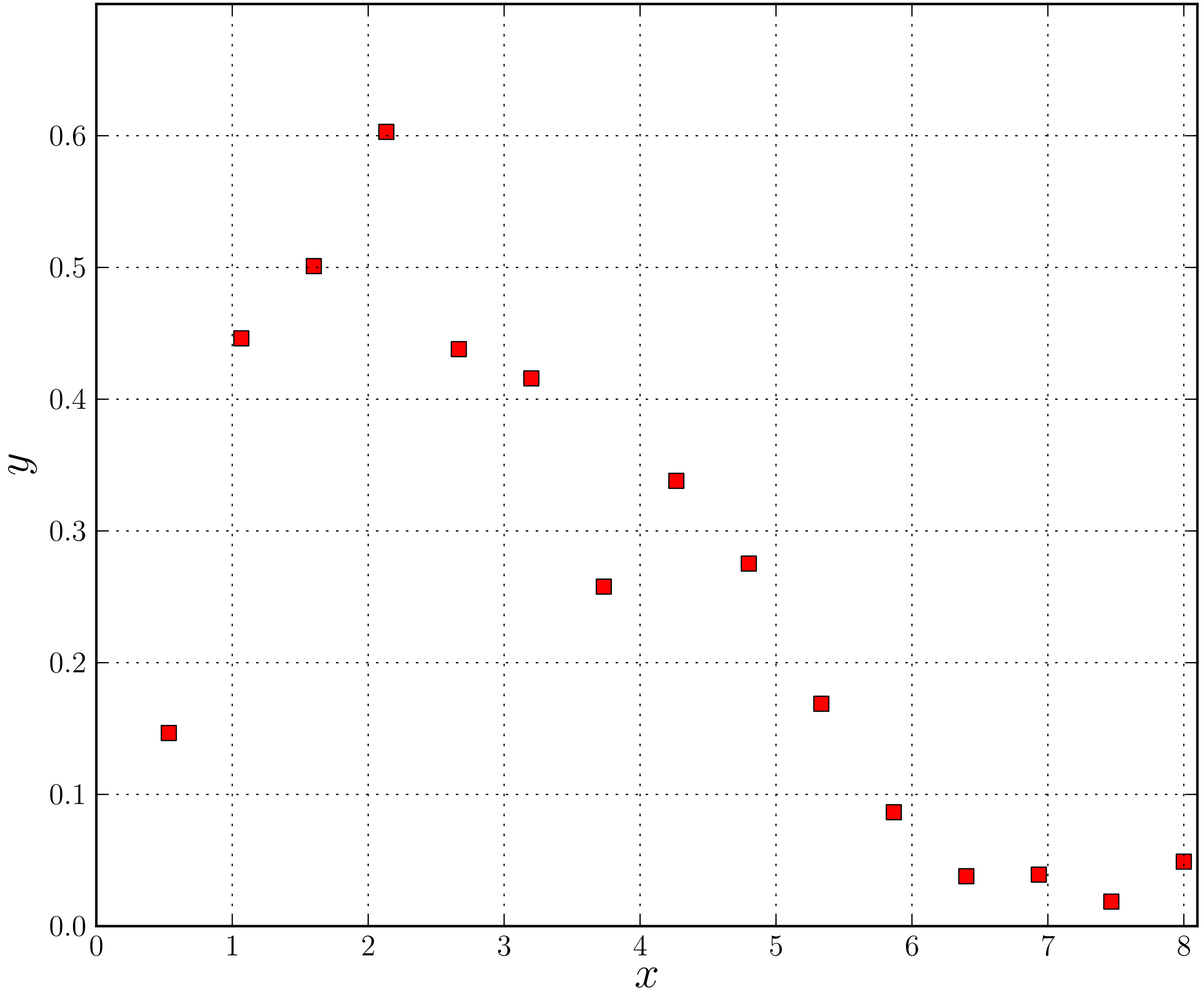
In this example we find the least squares coefficients

$$c_1 = 11.4 \quad , \quad c_2 = -8.66 \quad , \quad c_3 = -12.8\ .$$



Least squares fit of average daily high temperatures.

EXAMPLE : Consider the following *experimental data* :



EXAMPLE : continued $\cdots$

Suppose we are given that :

- These data contain ”*noise*” .

- The underlying physical process is understood.

- The *functional dependence* is *known* to have the form

$$y \quad = \quad c_1 \; x^{c_2} \; e^{-c_3 x} \; .$$

- The values of c_1 , c_2 , c_3 are *not* known.

EXAMPLE : continued $\cdots$

The functional relationship has the form

$$y \quad = \quad c_1 \, x^{c_2} \, e^{-c_3 x} \ .$$

Note that :

- The unknown coefficients c_1 , c_2 , c_3 appear *nonlinearly* !

- This gives *nonlinear equations* for c_1 , c_2 , c_3 !

- Such problems are more *difficult* to solve !

- What to do ?

EXAMPLE : continued ···

Fortunately, in this example we can take the *logarithm* :

$$\log\ y \quad = \quad \log\ c_1 \quad + \quad c_2\ \log\ x \quad - \quad c_3\ x\ .$$

This gives a *linear* relationship

$$\log\ y \quad = \quad \hat{c}_1\ \phi_1(x) \quad + \quad c_2\ \phi_2(x) \quad + \quad c_3\ \phi_3(x)\ ,$$

where

$$\hat{c}_1 \quad = \quad \log\ c_1\ .$$

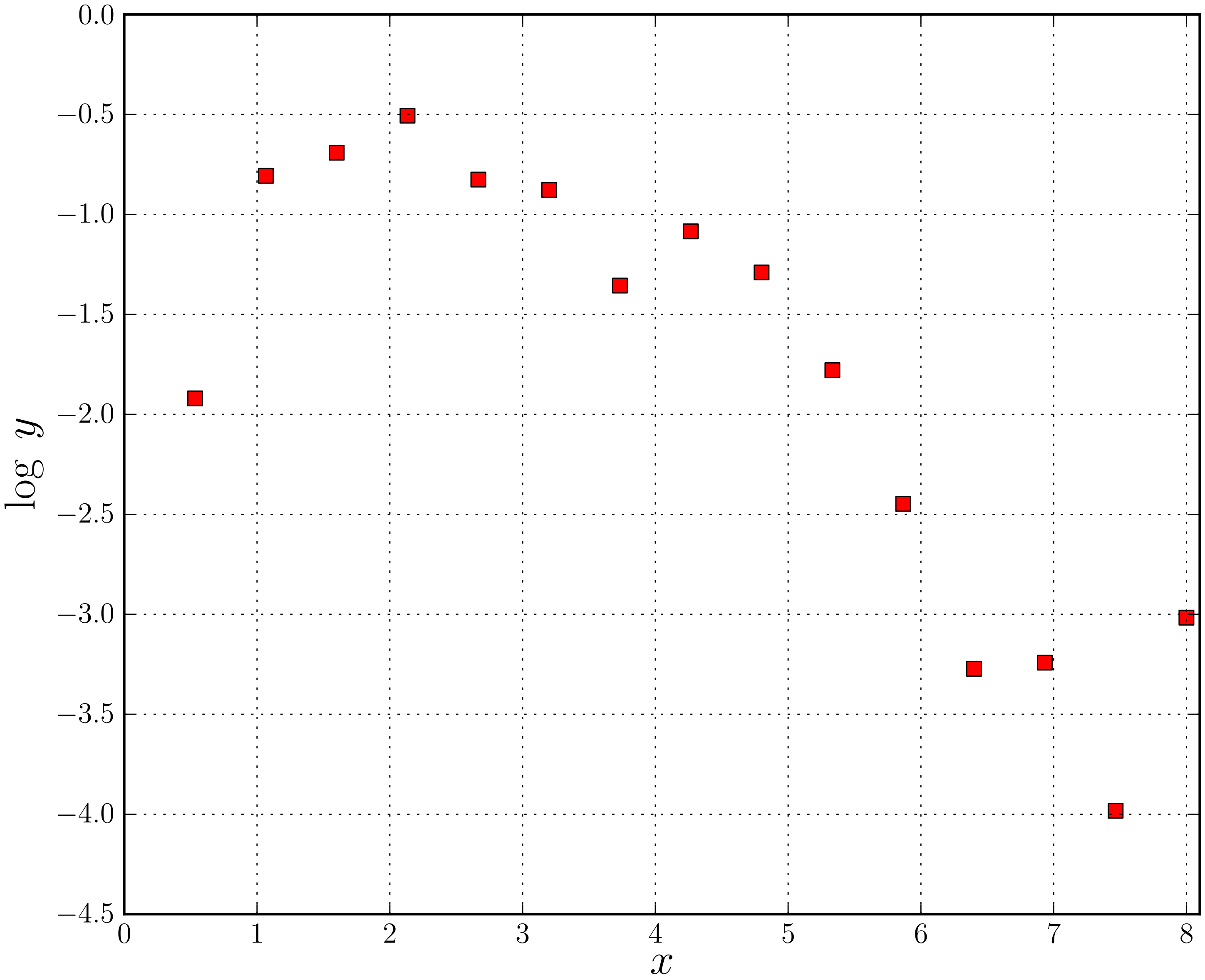
and

$$\phi_1(x) \ = \ 1 \quad , \quad \phi_2(x) \ = \ \log\ x \quad , \quad \phi_3(x) \ = \ -\,x\ .$$

Thus

- We can now use regular least squares.
- We first need to take the logarithm of the data.

EXAMPLE : continued $\cdots$



The logarithm of the original y-values versus x .

EXAMPLE : continued ···

We had

$$y \quad = \quad c_1 \ x^{c_2} \ e^{-c_3 x} \ ,$$

and

$$\log \ y \quad = \quad \hat{c}_1 \ \phi_1(x) \ + \ c_2 \ \phi_2(x) \ + \ c_3 \ \phi_3(x) \ ,$$

with

$$\phi_1(x) \ = \ 1 \quad , \quad \phi_2(x) \ = \ \log \ x \quad , \quad \phi_3(x) \ = \ - \, x \ ,$$

and

$$\hat{c}_1 \quad = \quad \log \ c_1 \ .$$

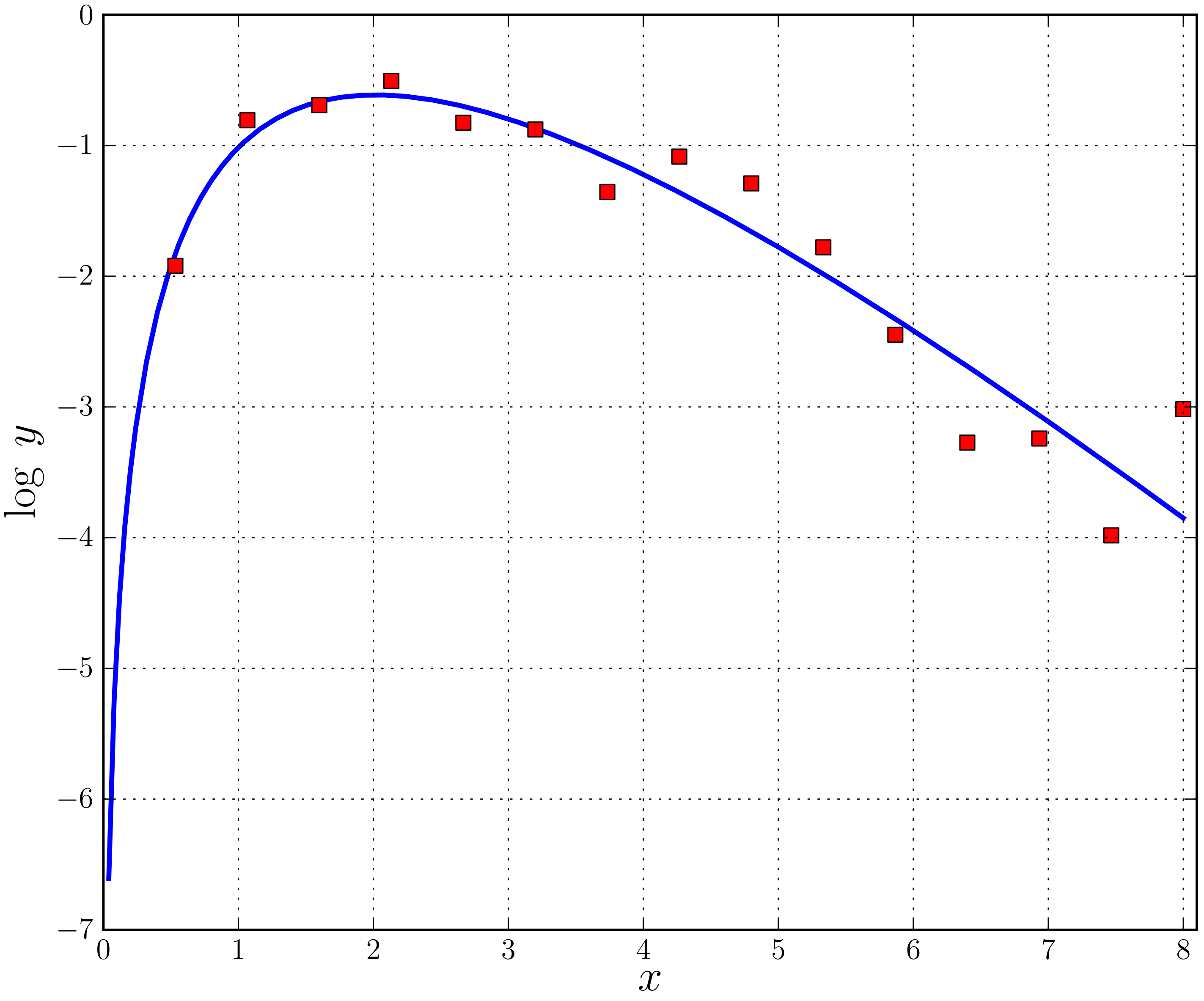
We find the following least squares values of the coefficients :

$$\hat{c}_1 \ = \ - \, 0.00473 \quad , \quad c_2 \ = \ 2.04 \quad , \quad c_3 \ = \ 1.01 \ ,$$

and

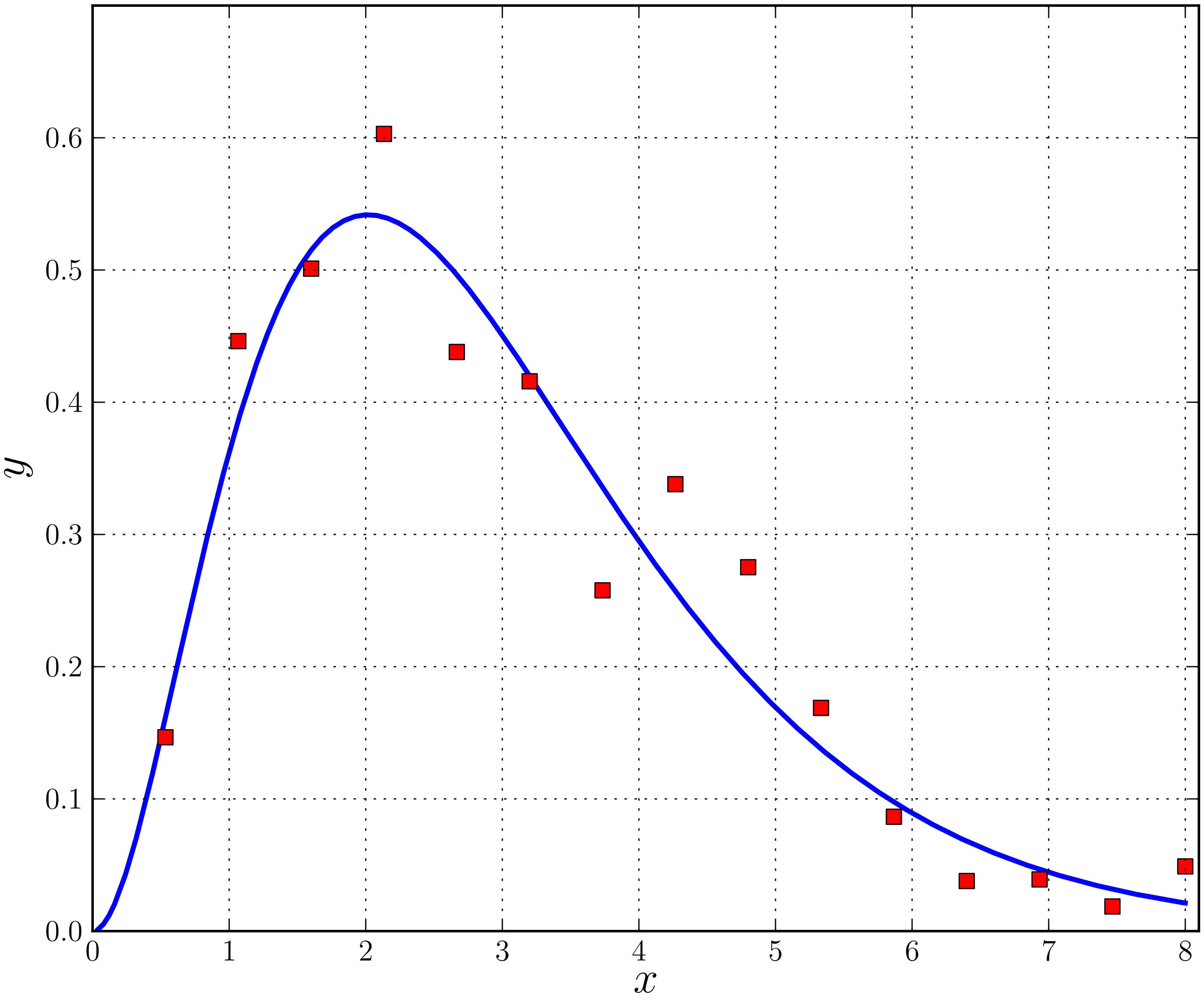
$$c_1 \quad = \quad e^{\hat{c}_1} \quad = \quad 0.995 \ .$$

EXAMPLE : continued $\cdots$



The least squares approximation of the transformed data.

EXAMPLE : continued ···



The least squares approximation shown in the original data.

EXERCISES :

- Compute the discrete least squares approximation of the form $p(x) = c_0 + c_1 x + c_2 x^2$ to the data $\{(0,2), (1,1), (2,1), (3,3)\}$.

- Compute the discrete least squares approximation of the form $p(x) = c_0 + c_1 x + c_2 \, \frac{1}{x}$ to the data $\{(1,5), (2,3), (3,2), (4,3)\}$.

- Derive a formula in terms of N and n for the number of multiplications and divisions needed to solve the linear discrete least squares system

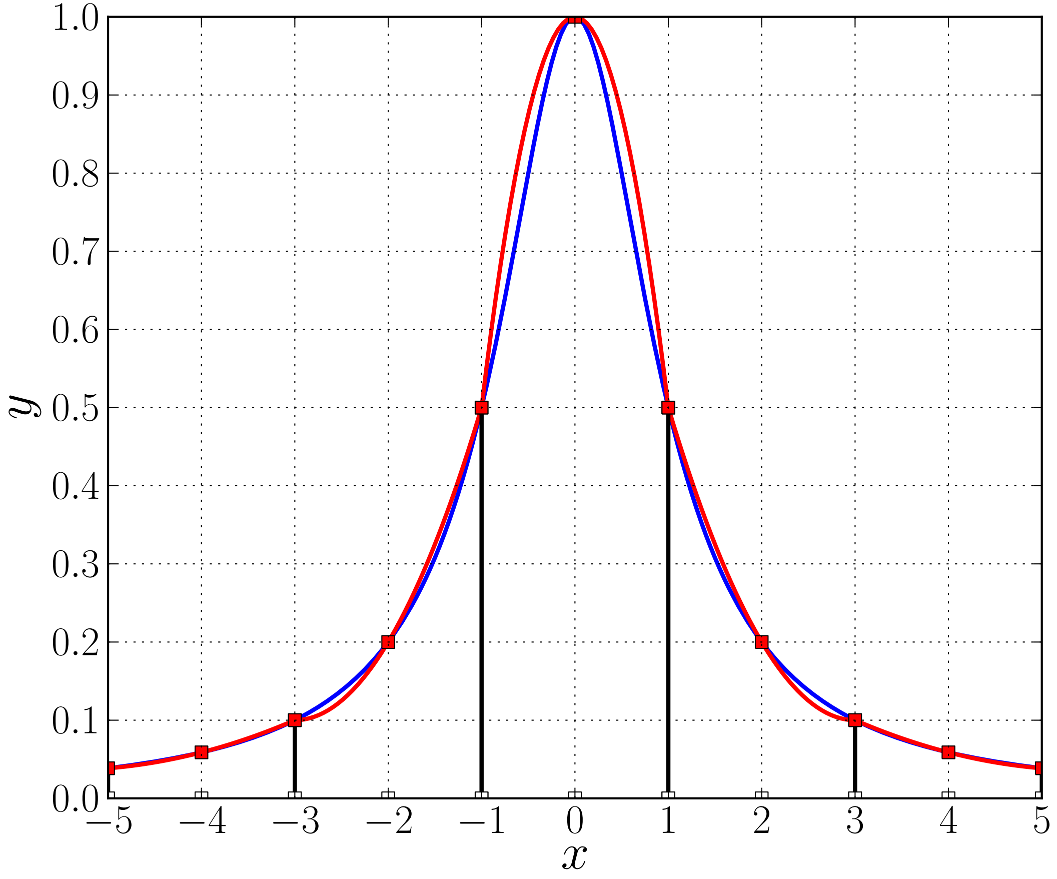
$$\mathbf{A}^T \mathbf{A} \mathbf{c} \;=\; \mathbf{A}^T \mathbf{y} \ ,$$

for $\mathbf{c} \in \mathbb{R}^n$, given the N by n matrix $\mathbf{A}$ and the vector $\mathbf{y} \in \mathbb{R}^N$. Here $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$. What is the total number of multiplications and divisions in terms of N for the special case $n = 2$?

SMOOTH INTERPOLATION BY PIECEWISE POLYNOMIALS

We have already discussed *local* (or *piecewise*) polynomial interpolation :

In each subinterval $[t_{j-1}, t_j]$ the function f is interpolated by a polynomial $p_j \in \mathbb{P}_n$ at interpolation points $\{ x_{j,i} \}_{i=0}^{n}$:



NOTE :

- The collection $\{p_j\}_{j=1}^N$ defines a function $p(t)$.

- $p(t)$ is generally *not smooth*, (not continuously differentiable) .

- In fact, $p(t)$ is not continuous, unless

$$x_{j,0} = t_{j-1} \qquad \text{and} \qquad x_{j,n} = t_j \ ,$$

  i.e., unless in each subinterval the leftmost and rightmost interpolation points are the end points of the subinterval.

NOTE :

- Sometimes a *smooth interpolant* is wanted.

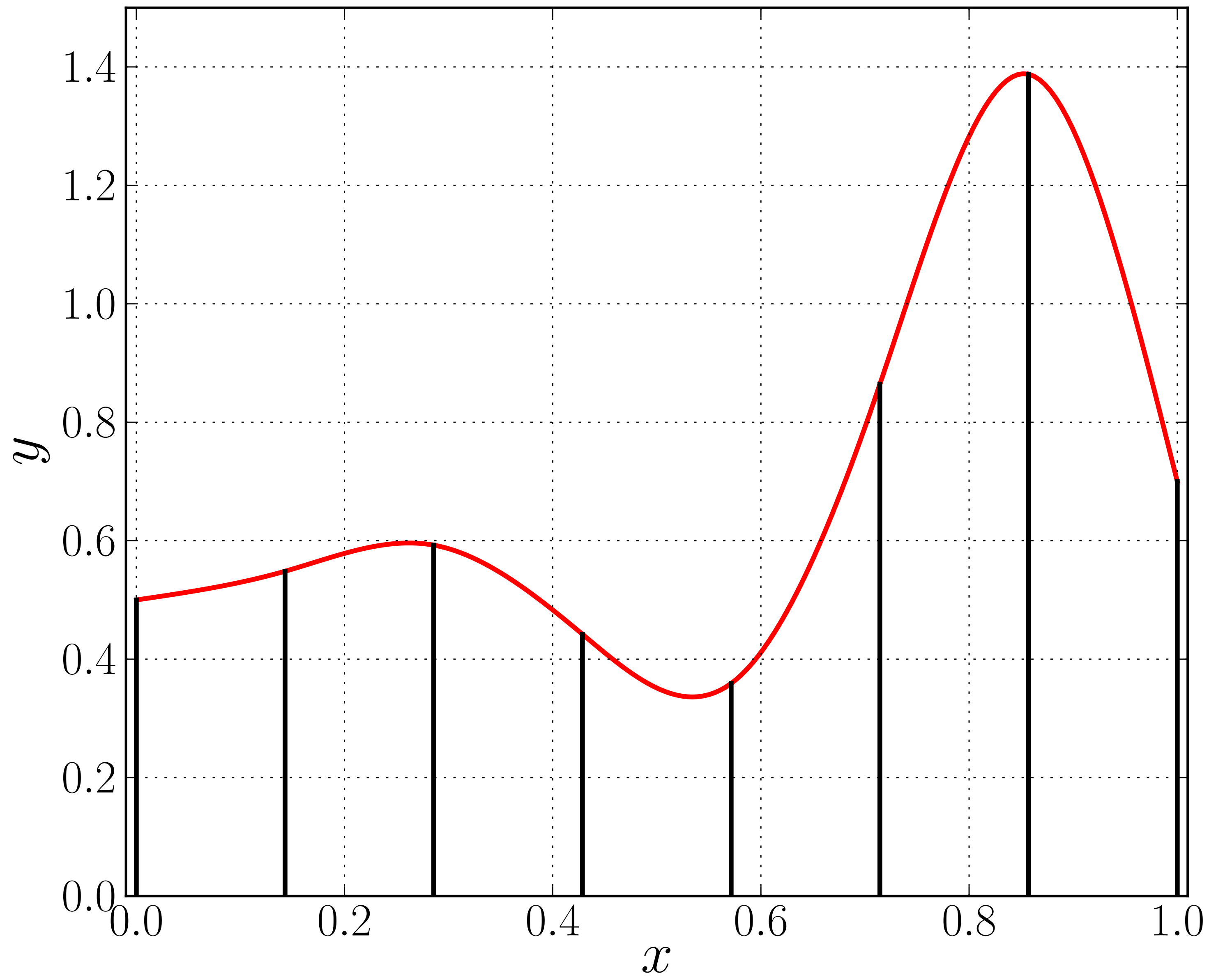
- These can also be constructed using piecewise polynomials.

- One class of smooth piecewise polynomial are called *cubic splines*.

- Cubic splines are piecewise polynomial functions

$$p(t) \ \in \ \mathbb{C}^2[a,b] \ ,$$

for which each component p_j is in $\mathbb{P}_3$.



The cubic spline that interpolates the indicated data points.

Cubic Spline Interpolation.

Given $f(t)$ defined on $[a, b]$ we seek a function $p(t)$ satisfying :

- $p \in \mathbb{C}^2[a, b]$,
- The restriction p_j of p to $[t_{j-1}, t_j]$ lies in $\mathbb{P}_3$,
- $p(t_j) = f(t_j)$, $\qquad j = 0, 1, \cdots, N$,

$\star$ $p''(t_0) = 0$, $\qquad p''(t_N) = 0$.

- There are other possible choices for $\star$.
- With the above choice of $\star$ a spline is called the *natural cubic spline*.
- We may also have *discrete data points* (t_j, f_j) , $j = 0, 1, \cdots, N$.

This spline is ”formally well defined”, because

the total number of *unknowns* is $4N$,

(since each p_j is defined by four coefficients) ,

which is matched by the number of *equations* :

continuity equations	$3(N-1)$
interpolation equations	$N+1$
end point conditions	2
Total	$4N$

NOTE :

- In practice we do not solve these $4N$ equations to find the spline.

- Often we want the values of the spline at a large number of points,

  whereas the actual number of data points

  $$\{ \ (\ t_j \ , \ f_j \ ) \ \}_{j=0}^{N}$$

  is relatively small.

- For this purpose we derive a more *efficient algorithm* below.

Consider the interval $[t_{j-1}, t_j]$ of size h_j .

To simplify notation take the interval $[t_0, t_1]$ of size h_1 .

Corresponding to this interval we have a polynomial $p \in \mathbb{P}_3$.

We can write

$$\begin{aligned} p(t_0) &= p_0 \ , \qquad & p(t_1) &= p_1 \ , \\ p''(t_0) &= p''_0 \ , \qquad & p''(t_1) &= p''_1 \ . \end{aligned}$$

These four equations uniquely define $p \in \mathbb{P}_3$ in terms of the values

$$p_0 \quad , \quad p_1 \quad , \quad p''_0 \quad , \quad p''_1 \ .$$

In fact, for the interval $[t_0, t_1]$, one finds the polynomial

$$p_1(t) \ = \ \frac{p_0''}{6h_1}\,(t_1{-}t)^3 + \frac{p_1''}{6h_1}\,(t{-}t_0)^3 + (\frac{p_1}{h_1} - \frac{p_1''h_1}{6})\,(t{-}t_0) + (\frac{p_0}{h_1}{-}\frac{p_0''h_1}{6})\,(t_1{-}t)\ .$$

Indeed, $p_1 \ \in \ \mathbb{P}_3$, and

$$p(t_0) \ = \ p_0\ , \qquad p(t_1) \ = \ p_1\ ,$$

$$p''(t_0) \ = \ p_0''\ , \qquad p''(t_1) \ = \ p_1''\ .$$

Similarly, for the interval $[t_1, t_2]$, one finds the polynomial

$$p_2(t) \ = \ \frac{p_1''}{6h_2}\,(t_2{-}t)^3 + \frac{p_2''}{6h_2}\,(t{-}t_1)^3 + (\frac{p_2}{h_2} - \frac{p_2''h_2}{6})\,(t{-}t_1) + (\frac{p_1}{h_2}{-}\frac{p_1''h_2}{6})\,(t_2{-}t)\ .$$

EXERCISE : Derive the formulas given above.

By construction the local polynomials p_1 and p_2 connect continuously at t_1.

By construction the second derivatives also connect continuously.

However, *the first derivatives* must also match :

$$p_1'(t_1) \ = \ p_2'(t_1) \ .$$

This requirement leads to the *consistency relation*

$$h_1 \ p_0'' \ + \ 2(h_1 + h_2) \ p_1'' \ + \ h_2 \ p_2'' \ \ = \ \ 6 \ \Big(\frac{p_2 - p_1}{h_2} \ - \ \frac{p_1 - p_0}{h_1} \Big) \ .$$

EXERCISE : Derive this formula.

For consecutive intervals $[t_{j-1}, t_j]$ and $[t_j, t_{j+1}]$, the *consistency relation* is

$$h_j \, p''_{j-1} \ + \ 2(h_j + h_{j+1}) \, p''_j \ + \ h_{j+1} \, p''_{j+1} \ = \ 6 \left(\frac{p_{j+1} - p_j}{h_{j+1}} \ - \ \frac{p_j - p_{j-1}}{h_j} \right) ,$$

where

$$h_j \ \equiv \ t_j - t_{j-1} \qquad \text{and} \qquad h_{j+1} \ \equiv \ t_{j+1} - t_j \ .$$

We have *one* such equation for each *interior mesh point* .

To *interpolate* the data points $\{(t_j, f_j)\}_{j=0}^{N}$, we have

$$p_j \ = \ f_j \ , \qquad j = 0, 1, \cdots, N \ .$$

Furthermore we have the natural spline *endpoint conditions*

$$p''_0 \ = \ p''_N \ = \ 0 \ .$$

This gives a *tridiagonal system* of equations for the unknown values

$$p_j'' \ , \qquad \text{for} \quad j = 1, \cdots, N-1 \ ,$$

namely,

$$\begin{pmatrix} 2(h_1+h_2) & h_2 & & & \\ h_2 & 2(h_2+h_3) & h_3 & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & h_{N-1} & 2(h_{N-1}+h_N) \end{pmatrix} \begin{pmatrix} p_1'' \\ p_2'' \\ \cdot \\ \cdot \\ p_{N-2}'' \\ p_{N-1}'' \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ \cdot \\ F_{N-2} \\ F_{N-1} \end{pmatrix}$$

where

$$F_j \ \equiv \ 6 \left(\frac{f_{j+1} - f_j}{h_{j+1}} - \frac{f_j - f_{j-1}}{h_j} \right) .$$

NOTE :

- In each row the diagonal entry is bigger than the sum of the other entries.

- Such a matrix is called *diagonally dominant*.

- By the Banach Lemma this matrix is nonsingular. (**Check !**)

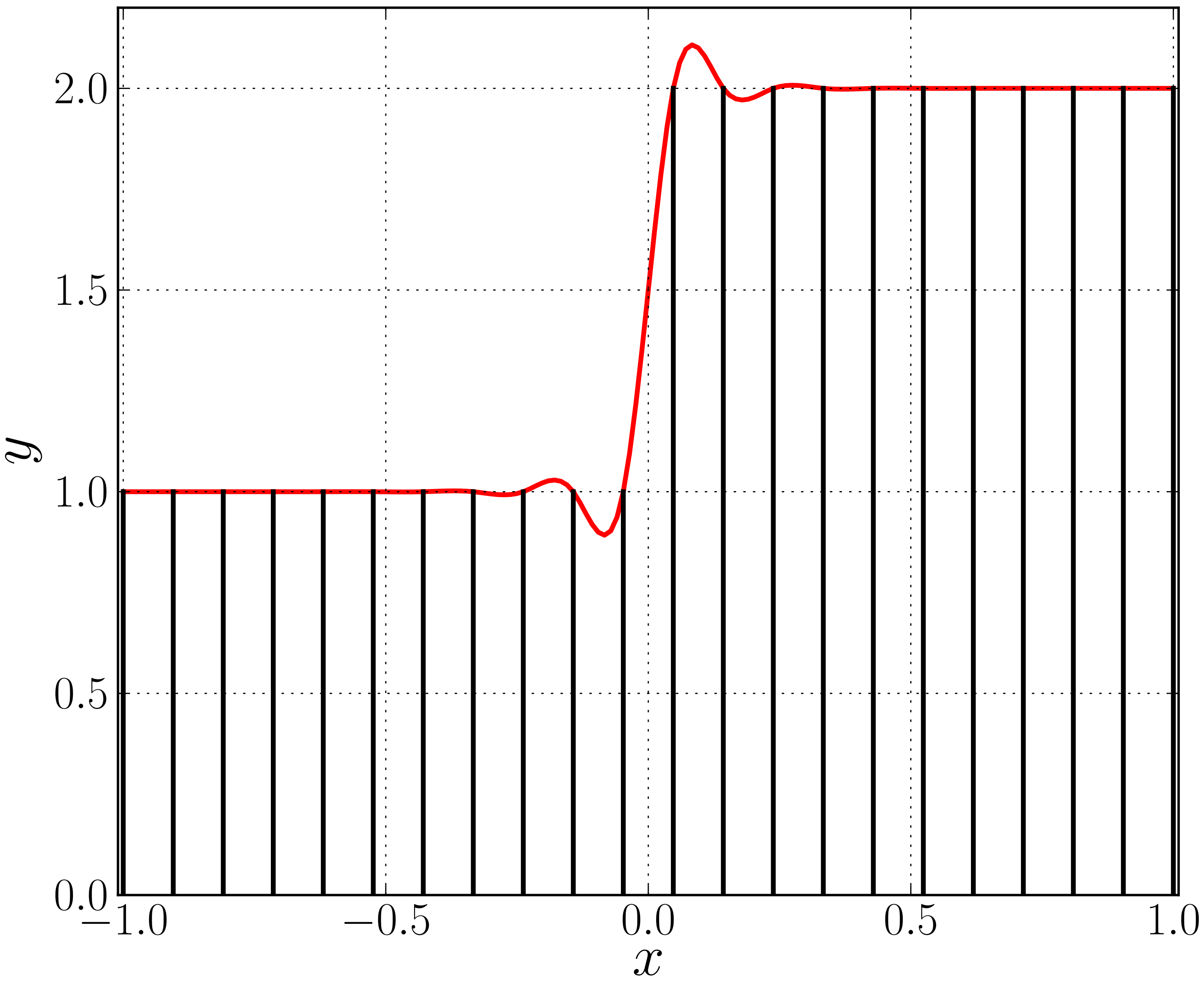
- Thus we can compute the p_j'' using the tridiagonal algorithm.

Thereafter evaluate each local polynomial with the formula

$$\begin{aligned} p_j(t) \ = \ & \frac{p_{j-1}''}{6h_j}\,(t_j - t)^3 \ + \ \frac{p_j''}{6h_j}\,(t - t_{j-1})^3 \\ & + \ (\frac{p_j}{h_j} - \frac{p_j'' h_j}{6})\,(t - t_{j-1}) \ + \ (\frac{p_{j-1}}{h_j} - \frac{p_{j-1}'' h_j}{6})\,(t_j - t) \ . \end{aligned}$$

NOTE :

- The smoothness makes the component polynomials *interdependent*.
- One can not determine each component polynomial individually.
- As seen above, a tridiagonal system must be solved.
- This interdependence can lead to unwanted *oscillations*.



The cubic spline that interpolates the indicated data points.

NUMERICAL METHODS FOR INITIAL VALUE PROBLEMS

Here we discuss some basic concepts that arise in the numerical solution of initial value problems (IVPs) in ordinary differential equations (ODEs) .

Consider the *first order* IVP

$$\mathbf{u}'(t) \; = \; \mathbf{f}(\; \mathbf{u}(t) \;) \; , \qquad \text{for} \;\; t \geq 0 \; ,$$

with *initial conditions*

$$\mathbf{u}(0) \; = \; \mathbf{u}_0 \; .$$

Here $\mathbf{u}$, $\mathbf{f}(\cdot) \in \mathbb{R}^n$.

Many *higher order* ODEs can be rewritten as *first order* systems .

EXAMPLE :

$$\mathbf{u}'' = \mathbf{g}(\ \mathbf{u}(t)\ ,\ \mathbf{u}'(t)\ )\ ,$$

where $\mathbf{u}$, $\mathbf{g}(\cdot,\cdot) \in \mathbb{R}^n$, with *initial conditions*

$$\mathbf{u}(0) = \mathbf{u}_0\ ,$$

$$\mathbf{u}'(0) = \mathbf{v}_0\ ,$$

can be *rewritten* as

$$\mathbf{u}'(t) = \mathbf{v}(t)\ ,$$

$$\mathbf{v}'(t) = \mathbf{g}(\ \mathbf{u}(t)\ ,\ \mathbf{v}(t)\ )\ ,$$

with *initial conditions*

$$\mathbf{u}(0) = \mathbf{u}_0\ ,$$

$$\mathbf{v}(0) = \mathbf{v}_0\ .$$

EXAMPLE :

The equations of motion of a *satellite* in an Earth-Moon-like system are :

$$
\begin{aligned}
x'' &= 2y' + x - (1-\mu)(x+\mu)r_1^{-3} - \mu(x-1+\mu)r_2^{-3} \ , \\
y'' &= -2x' + y - (1-\mu)yr_1^{-3} - \mu yr_2^{-3} \ , \\
z'' &= -(1-\mu)zr_1^{-3} - \mu zr_2^{-3} \ ,
\end{aligned}
$$

where

$$
r_1 = \sqrt{(x+\mu)^2 + y^2 + z^2} \ , \qquad r_2 = \sqrt{(x-1+\mu)^2 + y^2 + z^2} \ .
$$

Rewritten as a *first order system* :

$$\begin{aligned}
x' &= v_x \ , \\
y' &= v_y \ , \\
z' &= v_z \ , \\
v_x' &= 2v_y + x - (1-\mu)(x+\mu)r_1^{-3} - \mu(x-1+\mu)r_2^{-3} \ , \\
v_y' &= -2v_x + y - (1-\mu)yr_1^{-3} - \mu y r_2^{-3} \ , \\
v_z' &= -(1-\mu)zr_1^{-3} - \mu z r_2^{-3} \ ,
\end{aligned}$$

with

$$r_1 = \sqrt{(x+\mu)^2 + y^2 + z^2} \quad \text{and} \quad r_2 = \sqrt{(x-1+\mu)^2 + y^2 + z^2} \ .$$

This system is of the form

$$\mathbf{u}'(t) \ = \ \mathbf{f}(\ \mathbf{u}(t)\ ) \ , \quad \text{with } \textit{initial condition} \quad \mathbf{u}(0) \ = \ \mathbf{u}_0 \ .$$

Here μ is the *mass ratio* , *i.e.*,

$$\mu \equiv \frac{m_2}{m_1 + m_2} ,$$

where m_1 is the mass of the *larger* body, and m_2 of the *smaller* body.

For example,

$$\mu \cong 0.01215 \quad \text{for the Earth Moon system,}$$

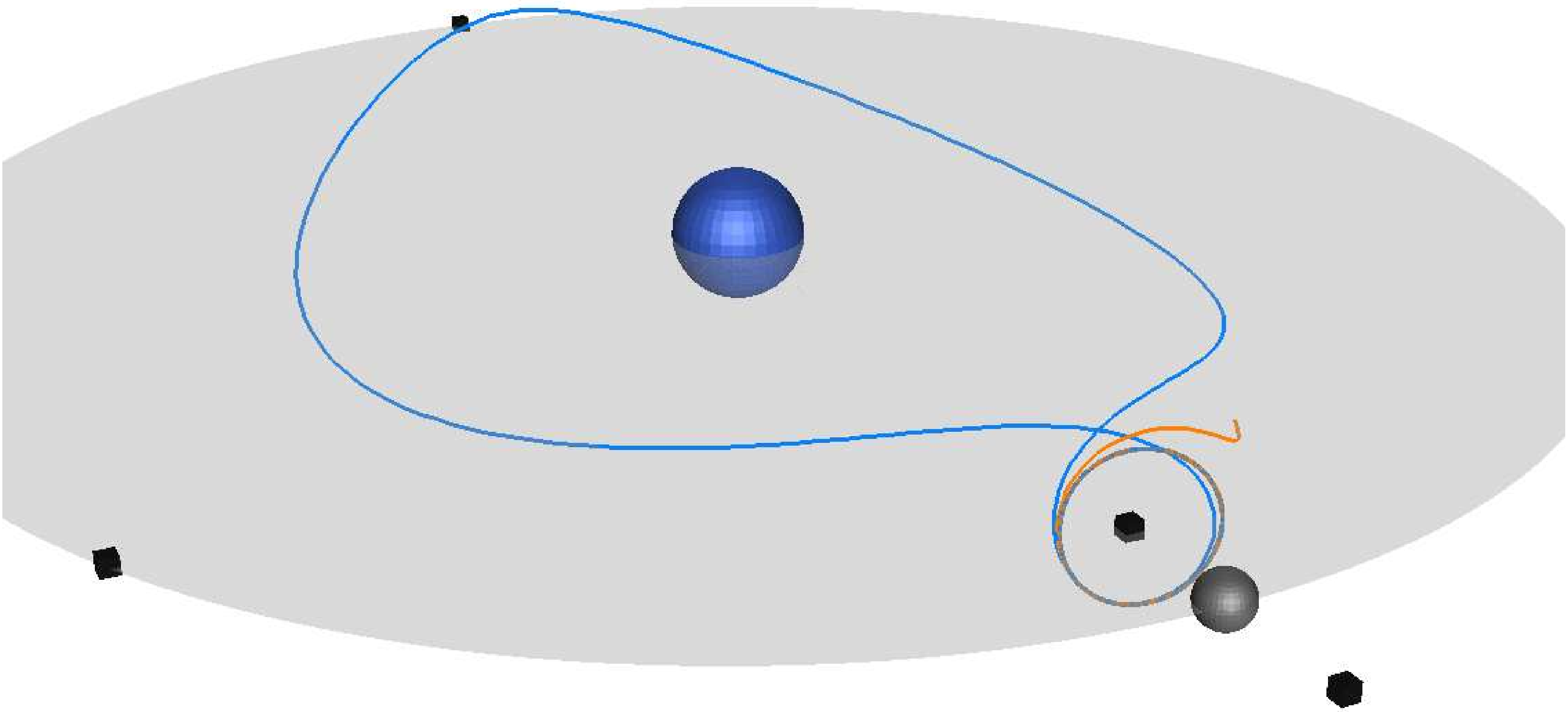
$$\mu \cong 9.53\ 10^{-4} \quad \text{for the Sun Jupiter system,}$$

$$\mu \cong 3.0\ 10^{-6} \quad \text{for the Sun Earth system.}$$

The variables are *scaled* such that

- the *distance* between the two bodies is 1 ,
- the *sum* of their masses is 1 .

The larger body is located at $(-\mu, 0, 0)$, and the smaller body at $(1-\mu, 0, 0)$.



A trajectory connecting a periodic “Halo orbit” to itself.

Numerical Methods.

Let

$$t_j \equiv j \, \Delta t \, , \qquad j = 0, 1, 2, \cdots \, .$$

Below we give several *basic numerical methods* for solving the IVP

$$\begin{aligned} \mathbf{u}'(t) &= \mathbf{f}(\ \mathbf{u}(t)\ ) \, , \qquad \mathbf{u} \, , \ \mathbf{f}(\cdot) \in \mathbb{R}^n \, . \\ \mathbf{u}(0) &= \mathbf{u}_0 \, . \end{aligned}$$

We use the *notation*

$\mathbf{u}(t_j)$ = the *exact solution* of the ODE at time t_j ,

$\mathbf{u}_j$ = the *numerical solution* at time t_j .

Euler's Method :

Using the *first order accurate* numerical differentiation formula

$$\frac{\mathbf{u}(t_{j+1}) \;-\; \mathbf{u}(t_j)}{\Delta t} \;\cong\; \mathbf{u}'(t_j) \;\;=\;\; \mathbf{f}(\;\mathbf{u}(t_j)\;)\;,$$

we have

$$\mathbf{u}_{j+1} \;=\; \mathbf{u}_j \;+\; \Delta t\;\mathbf{f}(\mathbf{u}_j)\;, \qquad j = 0, 1, 2, \cdots \;,$$

$$(\; \textit{explicit, one-step}\;, \;\; \mathcal{O}(\Delta t)\;)\;.$$

(Check the order of accuracy!)

The Trapezoidal Method :

Using the *second order accurate* approximation formula

$$\frac{\mathbf{u}(t_{j+1}) \ - \ \mathbf{u}(t_j)}{\Delta t} \cong \frac{\mathbf{u}'(t_j) + \mathbf{u}'(t_{j+1})}{2} \ = \ \frac{\mathbf{f}(\ \mathbf{u}(t_j)\ ) + \mathbf{f}(\ \mathbf{u}(t_{j+1})\ )}{2} \ ,$$

we have

$$\mathbf{u}_{j+1} \ = \ \mathbf{u}_j \ + \ \frac{\Delta t}{2} \, [\ \mathbf{f}(\mathbf{u}_j) \ + \ \mathbf{f}(\mathbf{u}_{j+1})\ ] \ , \qquad j = 0, 1, 2, \cdots \ ,$$

(*implicit, one-step* , $\mathcal{O}(\Delta t^2)$) .

(Check the order of accuracy!)

NOTE : In each time-step a *nonlinear system* must be solved !

A Two-Step (Three-Point) Backward Differentiation Formula (BDF) :

Using the *second order accurate* approximation formula

$$\frac{3\ \mathbf{u}(t_{j+1})\ -\ 4\ \mathbf{u}(t_j)\ +\ \mathbf{u}(t_{j-1})}{2\Delta t}\ \cong\ \mathbf{u}'(t_{j+1})\ =\ \mathbf{f}(\ \mathbf{u}(t_{j+1})\ )\ ,\qquad (\textbf{Check !})$$

we have

$$\mathbf{u}_{j+1}\ =\ \frac{4}{3}\ \mathbf{u}_j\ -\ \frac{1}{3}\ \mathbf{u}_{j-1}\ +\ \frac{2\Delta t}{3}\ \mathbf{f}(\mathbf{u}_{j+1})\ ,\qquad j=1,2,\cdots\ ,$$

(*implicit, two-step* , $\mathcal{O}(\Delta t^2)$) .

NOTE : In each time-step a *nonlinear system* must be solved!

A Two-Step (Three-Point) Forward Differentiation Formula :

Using the *second order accurate* approximation formula

$$\frac{-\mathbf{u}(t_{j+1}) \ + \ 4\ \mathbf{u}(t_j) \ - \ 3\mathbf{u}(t_{j-1})}{2\Delta t} \cong \mathbf{u}'(t_{j-1}) \ = \ \mathbf{f}(\ \mathbf{u}(t_{j-1})\ )\ , \qquad (\textbf{Check !})$$

we have

$$\mathbf{u}_{j+1} \ = \ 4\ \mathbf{u}_j \ - \ 3\ \mathbf{u}_{j-1} \ - \ 2\Delta t\ \mathbf{f}(\mathbf{u}_{j-1})\ , \qquad j = 1, 2, \cdots\ ,$$

(*explicit*, *two-step* , $\mathcal{O}(\Delta t^2)$) .

(We will show that *this method is useless* !)

The Improved Euler Method :

$$\hat{\mathbf{u}}_{j+1} \; = \; \mathbf{u}_j \; + \; \Delta t \, \mathbf{f}(\mathbf{u}_j) \; ,$$

$$\mathbf{u}_{j+1} \; = \; \mathbf{u}_j \; + \; \frac{\Delta t}{2} \, [\, \mathbf{f}(\mathbf{u}_j) \; + \; \mathbf{f}(\hat{\mathbf{u}}_{j+1}) \,] \; ,$$

for $j = 0, 1, 2, \cdots$.

(*explicit, one-step* , $\mathcal{O}(\Delta t^2)$) .

An Explicit 4th order accurate Runge-Kutta Method :

$$
\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(\mathbf{u}_j) \ , \\
\mathbf{k}_2 &= \mathbf{f}(\mathbf{u}_j + \frac{\Delta t}{2}\ \mathbf{k}_1) \ , \\
\mathbf{k}_3 &= \mathbf{f}(\mathbf{u}_j + \frac{\Delta t}{2}\ \mathbf{k}_2) \ , \\
\mathbf{k}_4 &= \mathbf{f}(\mathbf{u}_j + \Delta t\ \mathbf{k}_3) \ ,
\end{aligned}
$$

$$
\mathbf{u}_{j+1} \ = \ \mathbf{u}_j \ + \ \frac{\Delta t}{6}\ \{\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\} \ ,
$$

for $j = 0, 1, 2, \cdots$.

(*explicit*, *one-step* , $\mathcal{O}(\Delta t^4)$) .

The *order of accuracy* of a local formula can be found by *Taylor expansion* .

EXAMPLE :

For the *two-step BDF* we have the *local discretization error*

$$\begin{aligned}
\tau_j \ &\equiv \ \frac{1}{\Delta t} \ \{ \ \frac{3}{2} u(t_{j+1}) \ - \ 2u(t_j) \ + \ \frac{1}{2} u(t_{j-1}) \ \} \ - \ u'(t_{j+1}) \\
&= \ \frac{1}{\Delta t} \Big\{ \frac{3}{2} \ u(t_{j+1}) \\
&\qquad - 2[\ u(t_{j+1}) \ - \ \Delta t \ u'(t_{j+1}) \ + \ \frac{\Delta t^2}{2} \ u''(t_{j+1}) \ - \ \frac{\Delta t^3}{6} \ u'''(t_{j+1}) \ + \ \cdots \ ] \\
&\qquad + \frac{1}{2} [\ u(t_{j+1}) - 2\Delta t \ u'(t_{j+1}) + \frac{(2\Delta t)^2}{2} u''(t_{j+1}) - \frac{(2\Delta t)^3}{6} u'''(t_{j+1}) + \cdots \ ] \Big\} \\
&- \ u'(t_{j+1}) \\
&= \ -\frac{1}{3} \ \Delta t^2 \ u'''(t_{j+1}) \ + \ \text{higher order terms.}
\end{aligned}$$

The accuracy of this method is of *order* 2 .

Stability of Numerical Approximations.

The very simple *model equation*

$$u'(t) \ = \ 0 \ , \qquad u(0) \ = \ u_0 \ , \qquad u \ , \ 0 \ \in \mathbb{R} \ ,$$

has *solution*

$$u(t) \ = \ u_0 \ , \qquad \text{(constant)} \ .$$

A general *m-step approximation* has the form

$$\alpha_m \ u_{j+1} \ + \ \alpha_{m-1} \ u_j \cdots \ + \ \alpha_0 \ u_{j+1-m} \ = \ 0 \ .$$

Assume that

$$u_0 \text{ is given } ,$$

and (if $m > 1$) that

$$u_1, \ u_2, \ \cdots \ , u_{m-1} \text{ are computed by another method } ,$$

e.g., by a one-step method of the same order of accuracy.

General *m-step approximation* of $u'(t) = 0$, $u(0) = u_0$:

$$\alpha_m \, u_{j+1} \; + \; \alpha_{m-1} \, u_j \cdots \; + \; \alpha_0 \, u_{j+1-m} \;\; = \;\; 0 \; .$$

EXAMPLES :

(1)	$u_{j+1} - u_j \;\; = \;\; 0$,	u_0 given	Euler, Trapezoidal
(2)	$3u_{j+1} - 4u_j + u_{j-1} \;\; = \;\; 0$	u_0, u_1 given	Backward Differentiation
(3)	$-u_{j+1} + 4u_j - 3u_{j-1} \;\; = \;\; 0$	u_0, u_1 given	Forward Differentiation

The *difference equation*

$$\alpha_m \, u_{j+1} \; + \; \alpha_{m-1} \, u_j \; + \; \cdots \; + \; \alpha_0 \, u_{j+1-m} \; = \; 0 \; ,$$

can be solved *explicitly* :

Try solutions of the *form* $u_j \; = \; z^j$.

Then we have

$$\alpha_m \, z^{j+1} \; + \; \alpha_{m-1} \, z^j \; + \; \cdots \; + \; \alpha_0 \, z^{j+1-m} \; = \; 0 \; ,$$

or, multiplying through by z^{m-j-1}

$$\alpha_m \, z^m \; + \; \alpha_{m-1} \, z^{m-1} \; + \; \cdots \; + \; \alpha_1 \, z \; + \; \alpha_0 \; = \; 0 \; .$$

This is the *Characteristic Equation* of the difference equation.

Difference equation :

$$\alpha_m \, u_{j+1} \; + \; \alpha_{m-1} \, u_j \; + \; \cdots \; + \; \alpha_0 \, u_{j+1-m} \; = \; 0 \; .$$

Characteristic Equation :

$$\alpha_m \, z^m \; + \; \alpha_{m-1} \, z^{m-1} \; + \; \cdots \; + \; \alpha_1 \, z \; + \; \alpha_0 \; = \; 0 \; .$$

- If $\alpha_m \neq 0$, then the characteristic equation has m *roots* $\{z_k\}_{k=1}^m$.

- For simplicity we assume here that these roots are *distinct* .

- The *general solution* of the difference equation is then

$$u_j \; = \; \gamma_1 \, z_1^j \; + \; \gamma_2 \, z_2^j \; + \; \cdots + \gamma_m \, z_m^j \; .$$

- The coeficients γ_k are determined by the *initial data* $u_0, u_1, \cdots, u_{m-1}$.

FACT :

If the characteristic equation has one or more zeroes z_k with $| z_k |> 1$ then the numerical method is *unstable* .

In such a case the u_j can become *arbitrarily large* in a fixed time interval by taking Δt sufficiently small.

THEOREM :

A necessary condition for *numerical stability* of a multistep method is that the *characteristic equation*

$$\alpha_m z^{j+1} \ + \ \alpha_{m-1} z^j \ + \ \cdots \ + \ \alpha_0 \ = \ 0 \ ,$$

have *no zeroes outside the unit circle* .

EXAMPLES :

	Formula	Char. Eqn.	Roots	Stability
(1)	$u_{j+1} - u_j = 0$	$z - 1 = 0$	$z = 1$	Stable
(2)	$3u_{j+1} - 4u_j + u_{j-1} = 0$	$3z^2 - 4z + 1 = 0$	$z = 1, \frac{1}{3}$	Stable
(3)	$-u_{j+1} + 4u_j - 3u_{j-1} = 0$	$-z^2 + 4z - 3 = 0$	$z = 1, 3$	Unstable

Consider the last two examples in more detail :

Case (2) : Here the *general solution* is

$$u_j \ = \ \gamma_1 \ (1)^j \ + \ \gamma_2 \ (\frac{1}{3})^j \ .$$

The *initial data* are u_0 and u_1 , so that

$$\gamma_1 \ + \ \gamma_2 \ = \ u_0 \ ,$$

$$\gamma_1 \ + \ \frac{1}{3} \ \gamma_2 \ = \ u_1 \ ,$$

from which

$$\gamma_1 \ = \ \frac{3}{2} \ u_1 \ - \ \frac{1}{2} \ u_0 \ ,$$

$$\gamma_2 \ = \ \frac{3}{2} \ u_0 \ - \ \frac{3}{2} \ u_1 \ .$$

Hence

$$u_j \ = \ (\frac{3}{2}\ u_1 \ - \ \frac{1}{2}\ u_0) \ + \ (\frac{3}{2}\ u_0 \ - \ \frac{3}{2}\ u_1)\ (\frac{1}{3})^j \ .$$

If

$$u_1 \ = \ u_0 \ ,$$

then we see that

$$u_j \ = \ u_0 \ , \qquad \text{for all } j \ .$$

Moreover, if

$$u_1 \ = \ u_0 \ + \ \epsilon \ ,$$

then

$$u_j \ = \ u_0 \ + \ \frac{3}{2}\ \epsilon \ - \ \frac{3\epsilon}{2}\ (\frac{1}{3})^j \ ,$$

so that u_j *stays close* to u_0 if ϵ is small.

Case (3) : Here the *general solution* is

$$u_j \ = \ \gamma_1 \ (1)^j \ + \ \gamma_2 \ (3)^j \ .$$

Using the initial data

$$\gamma_1 \ + \ \gamma_2 \ = \ u_0 \ ,$$

$$\gamma_1 \ + \ 3 \ \gamma_2 \ = \ u_1 \ ,$$

we find

$$\gamma_1 \ = \ \frac{3}{2} \ u_0 \ - \ \frac{1}{2} \ u_1 \quad , \quad \gamma_2 \ = \ \frac{1}{2} \ u_1 \ - \ \frac{1}{2} \ u_0 \ .$$

Hence

$$u_j \ = \ (\frac{3}{2} \ u_0 \ - \ \frac{1}{2} \ u_1) \ + \ (\frac{1}{2} \ u_1 \ - \ \frac{1}{2} \ u_0) \ (3)^j \ .$$

Again, if $u_1 = u_0$ then $u_j = u_0$ for all j .

But if $u_1 \ = \ u_0 \ + \ \epsilon$ then $u_j \ = \ u_0 \ - \ \frac{1}{2} \ \epsilon \ + \ \frac{1}{2} \ \epsilon \ 3^j$.

Hence u_j becomes *arbitrarily large in finite time* by taking small Δt !

THEOREM :

Suppose that the local approximation is accurate.

Then if the zeroes $\{z_k\}_{k=1}^m$ of the *characteristic equation*

$$\alpha_m z^m \ + \ \alpha_{m-1} z^{m-1} \ + \ \cdots \ + \ \alpha_1 z \ + \ \alpha_0 \ = \ 0 \ ,$$

satisfy

$$|\, z_k \,| \ \leq \ 1 \ , \qquad \text{and} \qquad |\, z_k \,| \ = \ 1 \ \Rightarrow \ z_k \ \text{ is simple} \ ,$$

then the *method is stable* and

$$u_j \rightarrow u(t_j) \qquad \text{as} \qquad \Delta t \rightarrow 0 \ .$$

PROOF : Omitted.

Stiff Differential Equations.

There are ODEs for which *explicit* difference approximations require Δt to be *very small* before one gets the convergence guaranteed by the theorem.

To investigate this, we use the *model equation*

$$u'(t) \ = \ \lambda \, u(t) \ , \qquad t \geq 0 \ ,$$

with

$$u(0) \ = \ u_0 \ ,$$

where λ is a constant. (We allow λ to be *complex*.)

The *solution* is

$$u(t) \ = \ e^{\lambda t} \, u_0 \ .$$

Consider the case where

$$Re(\lambda) \ << \ 0 \ ,$$

$i.e.$, λ has *large negative real part* .

Then the *exact solution* of

$$u'(t) \ = \ \lambda \ u(t) \ ,$$

namely, $u(t) \ = \ e^{\lambda t} \ u_0$, *decays* very quickly as $t \to \infty$.

The *numerical solution* u_j again has the form

$$u_j \ = \ \sum_{k=1}^{m} \gamma_k \ z_k^j \ ,$$

and we certainly don't want u_j to increase as $j \to \infty$!

Thus we don't want any z_k *outside* the unit disk in the complex plane.

However, for many difference formulas

$$\Delta t \ \text{ must be } \textit{very small}$$

in order that

$$\text{all} \ \ z_k \ , \quad k = 1, \cdots m, \ \text{ are inside the unit disk} \ .$$

Thus problems with

$$Re(\lambda) \ << \ 0 \ , \qquad (\text{“}\textit{Stiff} \ \ \textit{Problems}\text{”}) \ ,$$

need *special* difference approximations .

More generally the IVP

$$\begin{aligned} \mathbf{u}'(t) &= \mathbf{f}(\ \mathbf{u}(t)\ )\ , \\ \mathbf{u}(0) &= \mathbf{u}_0\ , \end{aligned}$$

is called *stiff* if the Jacobian

$$\mathbf{f}_u(\ \mathbf{u}(t)\ )\ ,$$

has one or more *eigenvalues* $\lambda_i = \lambda_i(t)$, with

$$Re(\lambda_i)\ <<\ 0\ .$$

NOTE : Eigenvalues can be complex.

EXAMPLES :

We will approximate

$$u'(t) \;=\; \lambda\, u(t) \ ,$$

by various discretization formulas and determine the values of $\Delta t\lambda$ in the complex plane for which the solution of the difference formula decays .

Assume

$$\Delta t \;>\; 0 \ , \qquad \text{and} \qquad Re(\lambda) \;<\; 0 \ .$$

Then $\Delta t\lambda$ always lies in the *negative half plane* , *i.e.*,

$$Re(\Delta t\lambda) \;<\; 0 \ .$$

NOTE : Since eigenvalues can be complex, we allow λ to be complex.

Explicit Euler.

Applying Euler's explicit formula to the model equation

$$u'(t) \;=\; \lambda\, u(t)\ ,$$

we get the *difference equation*

$$\frac{1}{\Delta t}\,(u_{j+1} - u_j) \;=\; \lambda\, u_j\ , \qquad i.e., \qquad u_{j+1} \;=\; (1 + \Delta t\lambda)\, u_j\ .$$

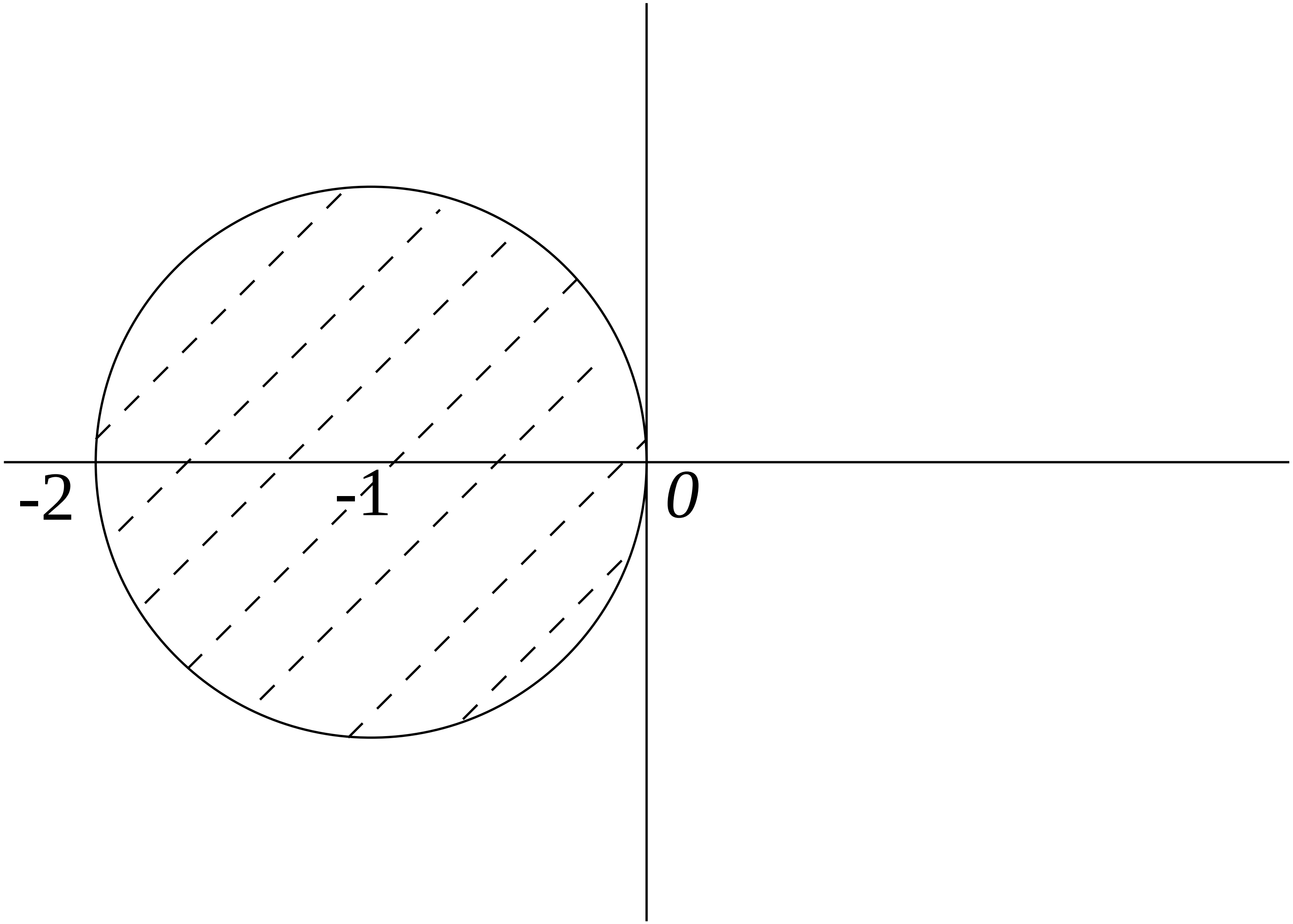
Trying solutions of the form $u_j = z^j$ gives the *characteristic equation*

$$z \;-\; (1 + \Delta t\lambda) = 0\ , \qquad \text{with zero} \qquad z \;=\; 1 + \Delta t\lambda\ .$$

Thus $|\, z \,| \;\leq\; 1$ if and only if $|\, 1 + \Delta t\lambda \,| \;\leq\; 1$, *i.e.*, if and only if

$$|\, \Delta t\lambda - (-1) \,| \;\leq 1\ .$$

Complex $(\Delta t \cdot \lambda)-$ *plane*



Stability region of the Explicit Euler method

EXAMPLE : Take $\lambda \ = \ -10^6$.

Then

$$u(t) \ = \ e^{(-10^6 t)} \ u_0 \ ,$$

which *decays* very rapidly for increasing t !

However, for u_j to decay, one must take

$$\Delta t \lambda \ > \ -2 \ ,$$

that is,

$$\Delta t \ < \ 2 \cdot 10^{-6} \quad !$$

Thus the explicit Euler method is *useless* for stiff equations !

Implicit Euler.

The difference formula is

$$\frac{1}{\Delta t}\,(u_{j+1} - u_j) \;=\; \lambda\, u_{j+1} \;,$$

that is,

$$u_{j+1} \;=\; \frac{1}{1-\Delta t\lambda}\,u_j \;.$$

The *characteristic equation*

$$z \;-\; \frac{1}{1-\Delta t\lambda} \;=\; 0 \;,$$

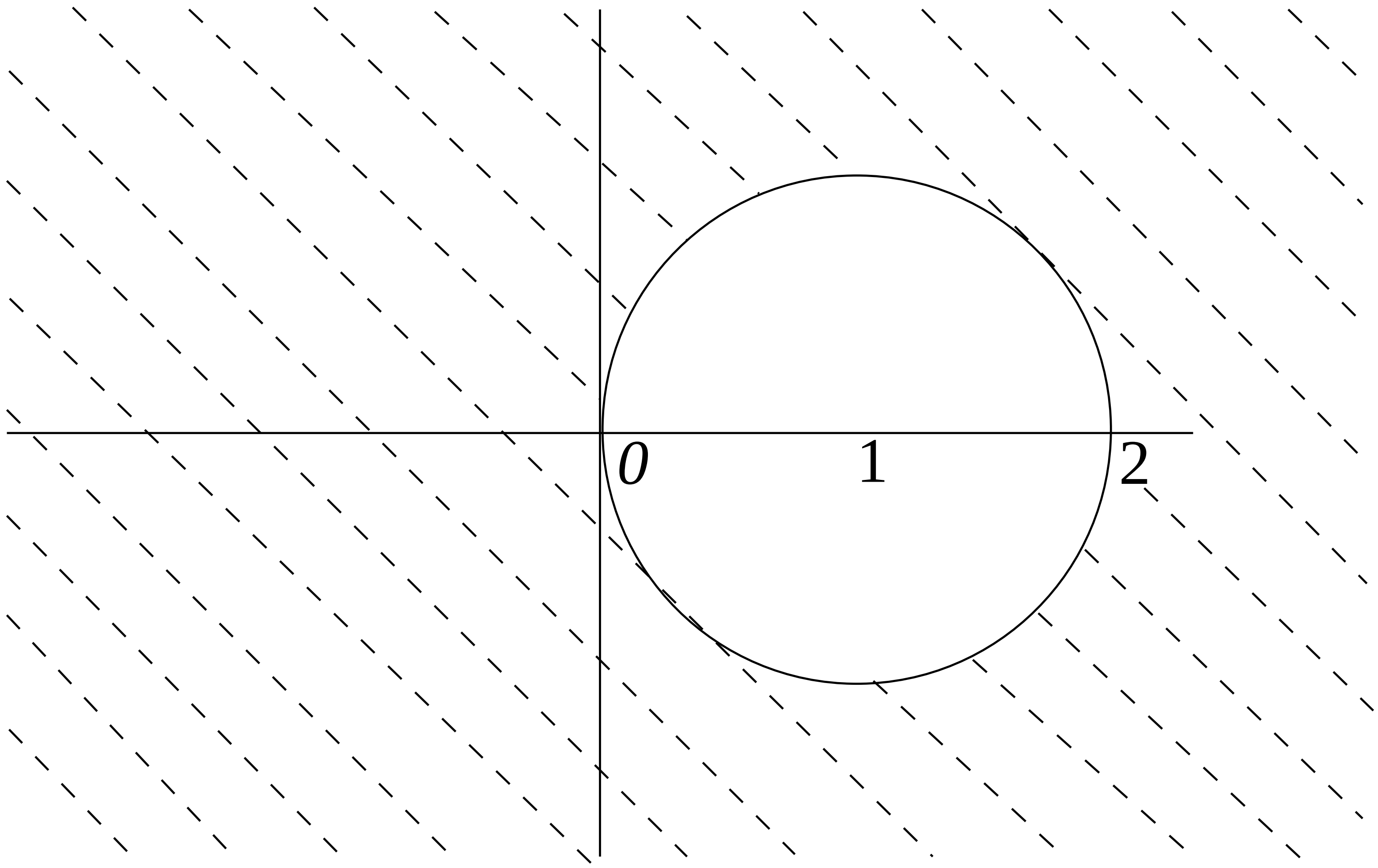
has zero

$$z \;=\; \frac{1}{1-\Delta t\lambda} \;,$$

so that $|\, z \,| \;\leq\; 1$ if and only if $|\, 1 - \Delta t\lambda \,| \;\geq\; 1$, *i.e.*, if and only if

$$|\, \Delta t\lambda - 1 \,| \;\geq\; 1 \;.$$

Complex $(\Delta t \cdot \lambda)$– *plane*



Stability region of the Implicit Euler method

Trapezoidal Method.

When applied to $u'(t) \; = \; \lambda \, u(t)$, the Trapezoidal Method gives

$$\frac{1}{\Delta t} \, (u_{j+1} - u_j) \;\; = \;\; \frac{1}{2} \, \lambda \, (u_j + u_{j+1}) \; .$$

Thus the *characteristic equation* is

$$(1 - \frac{1}{2}\Delta t \lambda) \; z \;\; - \;\; (1 + \frac{1}{2}\Delta t \lambda) \;\; = \;\; 0 \; , \qquad \text{with zero} \qquad z \;\; = \;\; \frac{1 + \frac{1}{2}\Delta t \lambda}{1 - \frac{1}{2}\Delta t \lambda} \; .$$

We find that $z = e^{i\theta}$ if

$$\Delta t \lambda \;\; = \;\; 2 \, (\frac{z-1}{z+1}) \;\; = \;\; 2 \, (\frac{e^{i\theta} - 1}{e^{i\theta} + 1}) \;\; = \;\; 2i \;\; \tan(\frac{\theta}{2}) \; .$$

The *region of stability* is now precisely the entire *negative half plane*.

Thus, $z \; \leq \; 1$ if and only if $Re(\Delta t \lambda) < 0$, which is very *desirable*.

A *disadvantage* is that the decay rate becomes smaller when

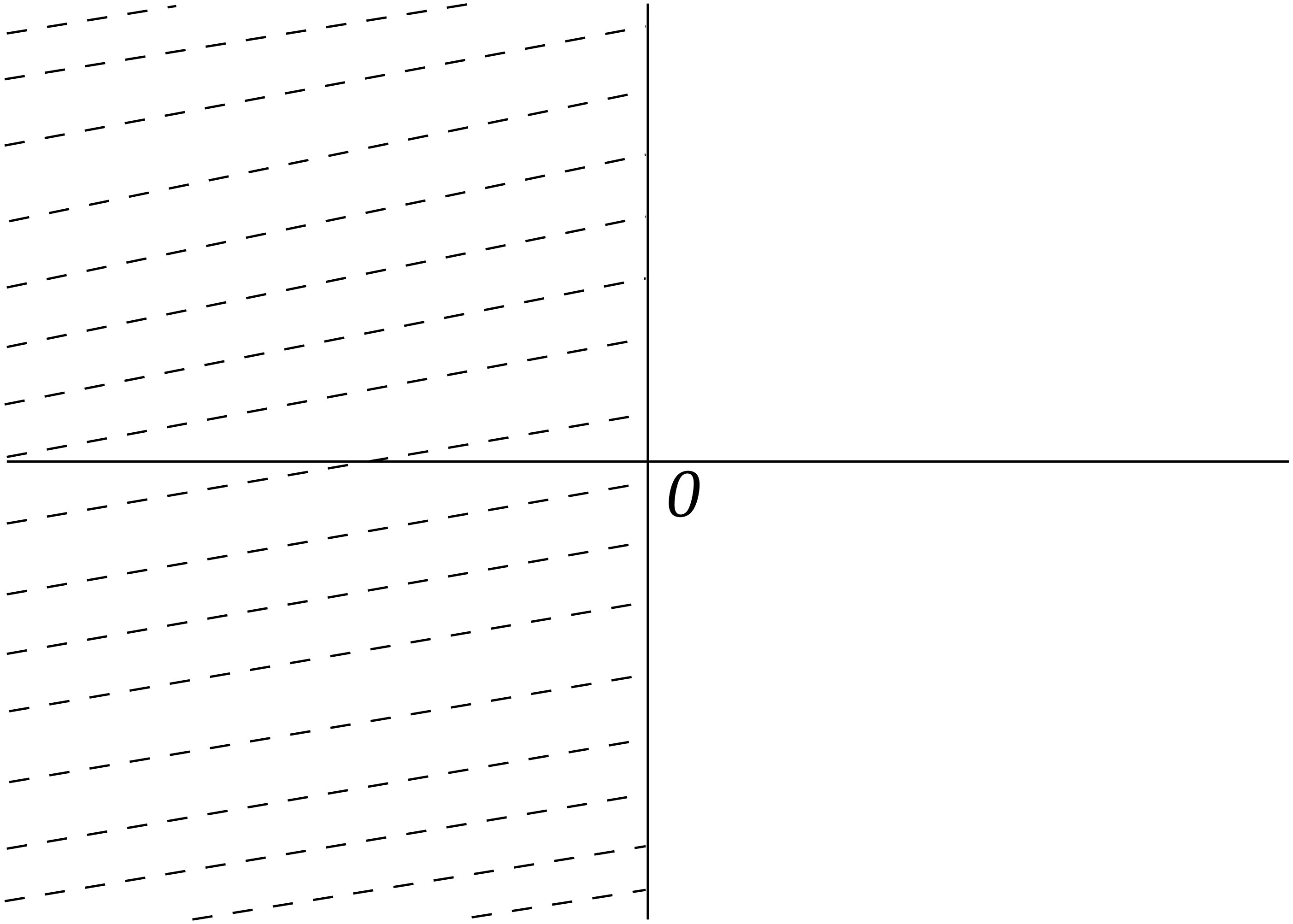
$$Re(\lambda) \ \rightarrow \ -\infty \ ,$$

contrary to the decay rate of the solution of the differential equation.

In fact (thinking of Δt as fixed) we have

$$\lim_{\lambda \rightarrow -\infty} z(\lambda) \quad = \quad \lim_{\lambda \rightarrow -\infty} \frac{1 + \frac{1}{2}\Delta t \lambda}{1 - \frac{1}{2}\Delta t \lambda} \quad = \quad -1 \ .$$

Complex $(\Delta t \cdot \lambda)$– *plane*



Stability region of the Trapezoidal method

Backward Differentiation Formulas (BDF).

For the differential equation $u'(t) = f(u(t))$ the BDF take the form

$$\frac{1}{\Delta t} \sum_{i=0}^{m} \alpha_i \, u_{j+1-i} \;=\; f(u_{j+1}) \; .$$

The $\{\alpha_i\}_{i=0}^{m}$ are chosen so the order is as high as possible, namely, $\mathcal{O}(\Delta t^m)$.

These formulas follow from the *numerical differentiation formulas* that approximate $u'(t_{j+1})$ in terms of

$$u(t_{j+1}) \; , \; u(t_j) \; , \; \cdots \; , \; u(t_{j+1-m}) \; .$$

All of these methods are *implicit* .

The choice $m = 1$ gives the implicit Euler method.

Let $\mathcal{S}_m$ denote the *stability region* of the m-step BDF.

Concerning $\mathcal{S}_m$ one can show the following :

$m = 1, 2$:

$\mathcal{S}_m$ contains the negative half plane.

These methods are called *A-stable* .

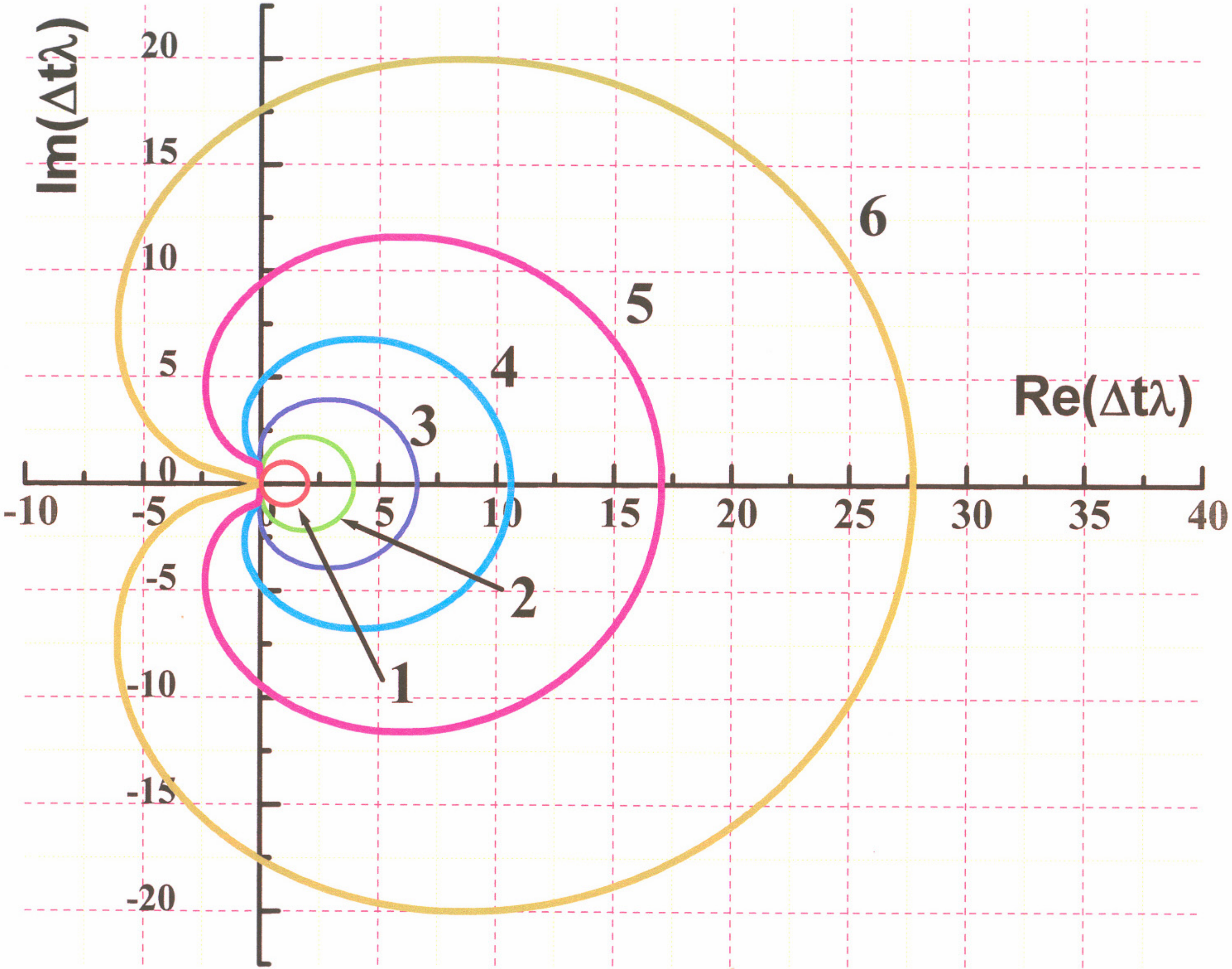
$m = 3, 4, 5, 6$:

$\mathcal{S}_m$ contains the negative axis, but not the entire negative half plane.

These methods are called *A(α)-stable* .

$m \geq 7$:

These methods are *unstable* , even for solving $u'(t) = 0$!



Stability region of Backward Differentiation Formulas.

Collocation at 2 Gauss Points.

The 2-point *Gauss collocation* method for taking a time step for the IVP

$$u'(t) \; = \; f(u(t)) \; , \qquad u(0) \; = \; u_0 \; ,$$

is defined by finding a local polynomial $p \in \mathbb{P}_2$ that satisfies

$$p(t_j) \; = \; u_j \; ,$$

and

$$p'(x_{j,i}) = f(\; p(x_{j,i}) \;) \; , \quad i = 1, 2 \; , \qquad (\textit{collocation}) \; ,$$

where

$$x_{j,i} \; = \; \frac{t_j + t_{j+1}}{2} \; \pm \; \frac{\Delta t \; \sqrt{3}}{6} \; ,$$

and then setting

$$u_{j+1} \; = \; p(t_{j+1}) \; .$$

Applied to the *model equation* $u'(t) \; = \; \lambda \, u(t)$ this gives

$$u_{j+1} \; = \; \frac{1 \; + \; \Delta t \; \lambda \; + \; \frac{1}{12}(\Delta t \; \lambda)^2}{1 \; - \; \Delta t \; \lambda \; + \; \frac{1}{12} \; (\Delta t \lambda)^2} \; u_j \; \equiv \; z(\Delta t \lambda) \; u_j \; .$$

It can be shown that the *stability region*

$$\mathcal{S} \equiv \{ \Delta t\lambda : \frac{1 + \Delta t\lambda + \frac{1}{12}(\Delta t\lambda)^2}{1 - \Delta t\lambda + \frac{1}{12}(\Delta t\lambda)^2} \leq 1 \} ,$$

is the *entire negative half plane*.

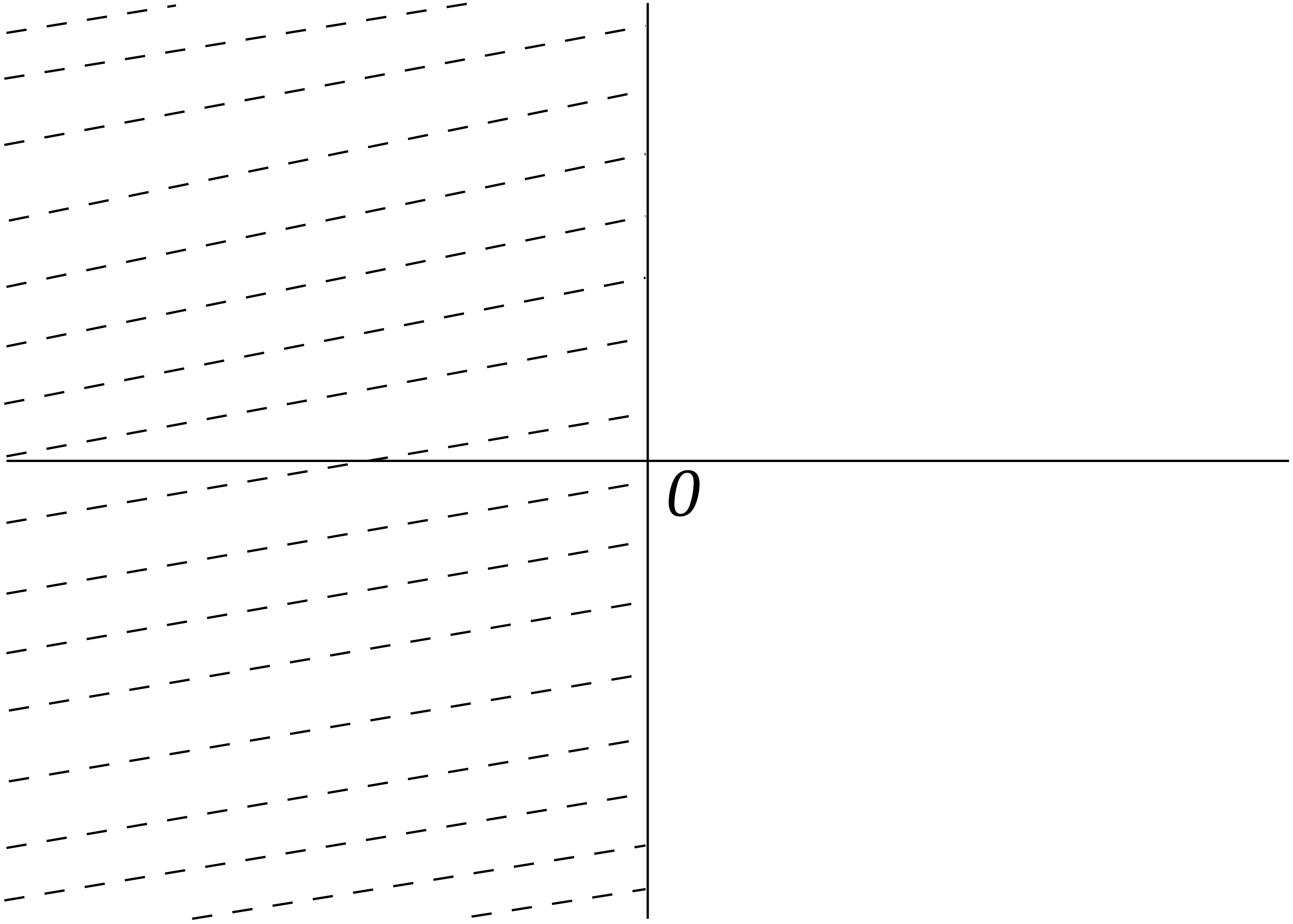
All Gauss collocation methods have this property and thus *are A-stable* .

However,

$$\lim_{\lambda \to -\infty} z(\Delta t\lambda) = 1 ,$$

so that the methods lead to *slow decay* for *stiff problems* .

Complex $(\Delta t \cdot \lambda) -$ *plane*



Stability region of the collocation method

BOUNDARY VALUE PROBLEMS IN ODEs

EXAMPLE : The *boundary value problem* (BVP)

$$y''(x) - y(x) \;=\; -5\sin(2x)\ , \qquad x \in [0,\pi]\ ,$$

$$y(0) \;=\; 0\ , \qquad y(\pi) \;=\; 0\ ,$$

has the exact (and unique) *solution*

$$y(x) \;=\; \sin(2x)\ .$$

- This BVP is a simple example of problems from science and engineering.
- Usually it is difficult or impossible to find an exact solution.
- In such cases numerical techniques can be used.

Partition $[0,\pi]$ into a *grid* or *mesh* :

$$0 \ = \ x_0 \ < \ x_1 \ < \ x_2 \ < \ \cdots \ < \ x_N \ = \ \pi \ ,$$

where

$$x_j \ = \ jh \ , \quad (j = 0,1,2,\cdots,N) \ , \qquad h \ = \ \frac{\pi}{N} \ .$$

We want to find approximations u_j to $y(x_j)$, $j = 0,1,2,\cdots,N$.

A *finite difference approximation* to $y''(x_j)$ is given by

$$y''(x_j) \ \cong \ \frac{\frac{y_{j+1}-y_j}{h} - \frac{y_j-y_{j-1}}{h}}{h} \ = \ \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} \ ,$$

where

$$y_j \ \equiv \ y(x_j) \ .$$

We want to find approximations u_j to $y(x_j)$, $j = 0, 1, 2, \cdots, N$.

The u_j are computed by solving the *finite difference equations* :

$$\begin{aligned}
u_0 &= 0 \ , \\
\frac{u_2 - 2u_1 + u_0}{h^2} - u_1 &= -5\sin(2x_1) \ , \\
\frac{u_3 - 2u_2 + u_1}{h^2} - u_2 &= -5\sin(2x_2) \ , \\
&\cdot \\
&\cdot \\
&\cdot \\
\frac{u_N - 2u_{N-1} + u_{N-2}}{h^2} - u_{N-1} &= -5\sin(2x_{N-1}) \ , \\
u_N &= 0 \ .
\end{aligned}$$

Write the *finite difference equations* as

$$(\frac{1}{h^2})\,u_{j-1} - (1 + \frac{2}{h^2})\,u_j + (\frac{1}{h^2})\,u_{j+1} \;=\; -5\ \sin(2x_j)\ , \qquad j = 1, 2, \cdots, N-1,$$

and put them in *matrix form* :

$$\begin{pmatrix} -1-\frac{2}{h^2} & \frac{1}{h^2} & & & \\ \frac{1}{h^2} & -1-\frac{2}{h^2} & \frac{1}{h^2} & & \\ & \cdot & \cdot & \cdot & \\ & & \frac{1}{h^2} & -1-\frac{2}{h^2} & \frac{1}{h^2} \\ & & & \frac{1}{h^2} & -1-\frac{2}{h^2} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \cdot \\ f_{N-2} \\ f_{N-1} \end{pmatrix} ,$$

where

$$\begin{pmatrix} f_1 \\ f_2 \\ \cdot \\ f_{N-2} \\ f_{N-1} \end{pmatrix} = \begin{pmatrix} -5\sin(2x_1) \\ -5\sin(2x_2) \\ \cdot \\ -5\sin(2x_{N-2}) \\ -5\sin(2x_{N-1}) \end{pmatrix} ,$$

and where the matrix has dimensions $N-1$ by $N-1$.

We found that :

The *finite difference equations* can be written in *matrix form* as

$$\mathbf{A}_h \, \mathbf{u}_h \quad = \quad \mathbf{f}_h \ ,$$

where

$$\mathbf{A}_h \ = \ \text{diag}[\ \frac{1}{h^2}\ , \quad -(1+\frac{2}{h^2})\ , \ \frac{1}{h^2}\ ]\ ,$$

$$\mathbf{u}_h \ \equiv \ (u_1\ ,\ u_2\ ,\ \cdots\ ,\ u_{N-1})^T\ ,$$

and

$$\mathbf{f}_h \ \equiv \ -5\ (\ \sin(2x_1)\ ,\ \sin(2x_2)\ ,\ \cdots\ ,\ \sin(2x_{N-1})\ )^T\ .$$

QUESTIONS :

- How to *solve* the linear systems efficiently, especially when N is large ?

- How to approximate derivatives and find the *error* in the approximation ?

- What is *the actual error* after solving the system,

  i.e. , what is

$$\max_j | u_j - y(x_j) | \quad ?$$

(assuming exact arithmetic)

- How to solve the linear systems efficiently, especially when N is large ?

ANSWER :

The matrix is *tridiagonal* .

Thus the linear system can be solved by the specialized Gauss elimination algorithm for tridiagonal systems.

- How to approximate derivatives and find the *error* in the approximation ?

ANSWER : As done earlier, the *local discretization error* is

$$
\begin{aligned}
\tau_j \quad &\equiv \quad \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - y_j'' \\
&= \quad \frac{1}{h^2} \Big(\; y_j \; + \; hy_j' \; + \; \frac{h^2}{2} y_j'' \; + \; \frac{h^3}{6} y_j''' \; + \; \frac{h^4}{24} y''''(\zeta_1) \\
&\qquad \quad - \; 2y_j \\
&\qquad \quad + \; y_j \; - \; hy_j' \; + \; \frac{h^2}{2} y_j'' \; - \; \frac{h^3}{6} y_j''' \; + \; \frac{h^4}{24} y''''(\zeta_2) \; \Big) \quad - \quad y_j'' \\
&= \quad \frac{h^2}{24} \Big(y''''(\zeta_1) + y''''(\zeta_2) \Big) \\
&= \quad \frac{h^2}{12} \, y''''(\eta_j) \; , \qquad \text{for some } \eta_j \in (x_{j-1}, x_{j+1}) \; ,
\end{aligned}
$$

using Taylor and Intermediate Value Theorem, assuming y'''' is continuous.

We found that

$$\tau_j \quad \equiv \quad \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} \; - \; y_j'' \quad = \quad \frac{h^2}{12} \, y''''(\eta_j) \; .$$

In our BVP, we have

$$y(x) \; = \; \sin(2x) \; , \quad \text{and} \quad y''''(x) \; = \; 16 \sin(2x) \; .$$

Thus $|\, y''''(x) \,| \; \leq \; 16$, and

$$|\, \tau_j \,| \quad \leq \quad \frac{16}{12} \, h^2 \quad = \quad \frac{4}{3} \, h^2 \; , \qquad j = 1, 2, \cdots, N-1 \; .$$

- What is the *actual error* after solving the system ?

i.e., what is

$$\max_j | \, u_j - y(x_j) \, | \quad ?$$

ANSWER :

For this, we will use the *Banach Lemma* .

We already showed that

$$| \tau_j | \quad \equiv \quad | \frac{(y_{j-1} - 2y_j + y_{j+1})}{h^2} \; - \; y_j'' \; | \quad \leq \quad \frac{4h^2}{3} \; , \qquad j = 1, 2, \cdots, N-1 \; .$$

Now

$$\begin{aligned} \frac{1}{h^2} y_{j-1} \; - \; (1 \; + \; \frac{2}{h^2}) y_j \; + \; \frac{1}{h^2} y_{j+1} & \\ &= \; \frac{(y_{j+1} \; - \; 2y_j \; + \; y_{j-1})}{h^2} \; - \; y_j \\ &= \; y_j'' \; + \; \tau_j \; - \; y_j \\ &= \; \tau_j \; - \; 5 \sin(2x_j) \; . \end{aligned}$$

Thus if we define

$$\mathbf{y}_h \; \equiv \; (\; y_1 \; , \; y_2 \; , \; \cdots \; , \; y_{N-1} \;)^T \; , \quad \text{and} \quad \boldsymbol{\tau}_h \; \equiv \; (\; \tau_1 \; , \; \tau_2 \; , \; \cdots \; , \; \tau_{N-1} \;)^T \; ,$$

then

$$\mathbf{A}_h \mathbf{y}_h \; = \; \boldsymbol{\tau}_h \; + \; \mathbf{f}_h \; .$$

We found that

$$\mathbf{A}_h\mathbf{y}_h \ = \ \boldsymbol{\tau}_h \ + \ \mathbf{f}_h \ .$$

Since

$$\mathbf{A}_h\mathbf{u}_h \ = \ \mathbf{f}_h \ ,$$

it follows from subtraction that

$$\mathbf{A}_h \ (\mathbf{y}_h - \mathbf{u}_h) \ = \ \boldsymbol{\tau}_h \ .$$

We found that

$$\mathbf{A}_h(\mathbf{y}_h - \mathbf{u}_h) \ = \ \boldsymbol{\tau}_h \ .$$

Thus if we can show that $\mathbf{A}_h$ has an *inverse* and that

$$\| \, \mathbf{A}_h^{-1} \, \|_\infty \ \leq \ K \ ,$$

for some constant K that does not depend on h, then

$$\begin{aligned} \| \, \mathbf{y}_h \ - \ \mathbf{u}_h \, \|_\infty \ &= \ \| \, \mathbf{A}_h^{-1} \, \boldsymbol{\tau}_h \, \|_\infty \\ &\leq \ \| \, \mathbf{A}_h^{-1} \, \|_\infty \ \ \| \, \boldsymbol{\tau}_h \, \|_\infty \\ &\leq \ K \, \frac{4h^2}{3} \ . \end{aligned}$$

Now

$$\begin{aligned}
\mathbf{A}_h &= \begin{pmatrix} -1-\frac{2}{h^2} & \frac{1}{h^2} & & & \\ \frac{1}{h^2} & -1-\frac{2}{h^2} & \frac{1}{h^2} & & \\ & \cdot & \cdot & \cdot & \\ & & \frac{1}{h^2} & -1-\frac{2}{h^2} & \frac{1}{h^2} \\ & & & \frac{1}{h^2} & -1-\frac{2}{h^2} \end{pmatrix} \\
&= -\frac{h^2+2}{h^2}\ \mathbf{I}_h + \frac{1}{h^2}\begin{pmatrix} 0 & 1 & & \\ 1 & 0 & 1 & \\ & \cdot & \cdot & \cdot \\ & & 1 & 0 \end{pmatrix} \\
&= -\frac{h^2+2}{h^2}\left[\ \mathbf{I}_h - \frac{h^2}{h^2+2}\,\frac{1}{h^2}\begin{pmatrix} 0 & 1 & & \\ 1 & 0 & 1 & \\ & \cdot & \cdot & \cdot \\ & & 1 & 0 \end{pmatrix}\ \right] \\
&= -\frac{h^2+2}{h^2}\left[\ \mathbf{I}_h - \frac{1}{h^2+2}\begin{pmatrix} 0 & 1 & & \\ 1 & 0 & 1 & \\ & \cdot & \cdot & \cdot \\ & & 1 & 0 \end{pmatrix}\ \right] .
\end{aligned}$$

We have

$$\begin{aligned}\mathbf{A}_h &= -\frac{h^2+2}{h^2}\ \Big[\ \mathbf{I}_h - \frac{1}{h^2+2}\begin{pmatrix} 0 & 1 & & \\ 1 & 0 & 1 & \\ & \cdot & \cdot & \cdot \\ & & 1 & 0 \end{pmatrix}\ \Big] \\ &= -\frac{h^2+2}{h^2}\ (\mathbf{I}_h + \mathbf{B}_h)\ ,\end{aligned}$$

where $\mathbf{I}_h$ is the identity matrix and

$$\mathbf{B}_h \equiv \frac{-1}{h^2+2}\begin{pmatrix} 0 & 1 & & \\ 1 & 0 & 1 & \\ & \cdot & \cdot & \cdot \\ & & 1 & 0 \end{pmatrix}\ .$$

Since

$$\|\ \mathbf{B}_h\ \|_\infty = \frac{2}{h^2+2} < 1\ ,$$

it follows by the *Banach Lemma* that $(\mathbf{I}_h + \mathbf{B}_h)^{-1}$ exists and that

$$\|\ (\mathbf{I}_h + \mathbf{B}_h)^{-1}\ \|_\infty \leq \frac{1}{1-\frac{2}{h^2+2}} = \frac{h^2+2}{h^2}\ .$$

We have

$$\mathbf{A}_h \quad = \quad - \frac{h^2+2}{h^2} \; (\mathbf{I}_h + \mathbf{B}_h) \ ,$$

and

$$\| \, (\mathbf{I}_h + \mathbf{B}_h)^{-1} \, \|_\infty \quad \leq \quad \frac{h^2+2}{h^2} \ .$$

Hence

$$\| \, \mathbf{A}_h^{-1} \, \|_\infty \quad = \quad \| \, \frac{-h^2}{h^2+2} \; (\mathbf{I}_h + \mathbf{B}_h)^{-1} \, \|_\infty \quad \leq \quad \frac{h^2}{h^2+2} \; \frac{h^2+2}{h^2} \quad = \quad 1 \ .$$

Thus $K \; = \; 1$, and

$$\| \, \mathbf{y}_h - \mathbf{u}_h \, \|_\infty \quad \leq \quad \frac{4h^2}{3} \ .$$

A Nonlinear Boundary Value Problem.

Consider the *Gelfand-Bratu* problem

$$u''(x) + \lambda e^{u(x)} \ = \ 0 \ , \qquad x \in [0,1] \ ,$$

$$u(0) \ = \ 0 \ , \quad u(1) \ = \ 0 \ , \quad \lambda \text{ is a parameter} \ ,$$

and its *finite difference approximation*

$$\begin{aligned}
g_1(\mathbf{u}) \ &\equiv \ \frac{u_2 - 2u_1}{h^2} \ + \ \lambda e^{u_1} \ = \ 0 \ , \\
g_2(\mathbf{u}) \ &\equiv \ \frac{u_3 - 2u_2 + u_1}{h^2} \ + \ \lambda e^{u_2} \ = \ 0 \ , \\
&\cdot \\
&\cdot \\
g_{N-2}(\mathbf{u}) \ &\equiv \ \frac{u_{N-1} - 2u_{N-2} + u_{N-3}}{h^2} \ + \ \lambda e^{u_{N-2}} \ = \ 0 \ , \\
g_{N-1}(\mathbf{u}) \ &\equiv \ \frac{-2u_{N-1} + u_{N-2}}{h^2} \ + \ \lambda e^{u_{N-1}} \ = \ 0 \ ,
\end{aligned}$$

where $\mathbf{u} \ \equiv \ (u_1 \ , \ u_2 \ , \ \cdots \ , \ u_{N-1})^T$.

If we let

$$\mathbf{G}(\mathbf{u}) \ \equiv \ (\ g_1(\mathbf{u})\ ,\ g_2(\mathbf{u})\ ,\ \cdots\ ,\ g_{N-1}(\mathbf{u})\ )^T\ ,$$

and

$$\mathbf{0} \ \equiv \ (0, 0, \cdots, 0)^T \in \mathbf{R}^{N-1}\ ,$$

then these equations can be compactly written as

$$\mathbf{G}(\mathbf{u}) \ = \ \mathbf{0}\ .$$

The Jacobian matrix is an $N-1$ by $N-1$ tridiagonal matrix :

$$\mathbf{G}'(\mathbf{u}) \ = \ \begin{pmatrix} -\frac{2}{h^2}+\lambda e^{u_1} & \frac{1}{h^2} & & \\ \frac{1}{h^2} & -\frac{2}{h^2}+\lambda e^{u_2} & \frac{1}{h^2} & \\ & \cdot & \cdot & \cdot \\ & & \frac{1}{h^2} & -\frac{2}{h^2}+\lambda e^{u_{N-1}} \end{pmatrix}\ .$$

Each *Newton iteration* for solving the nonlinear system

$$\mathbf{G}(\mathbf{u}) = \mathbf{0} \ ,$$

consists of solving the tridiagonal system

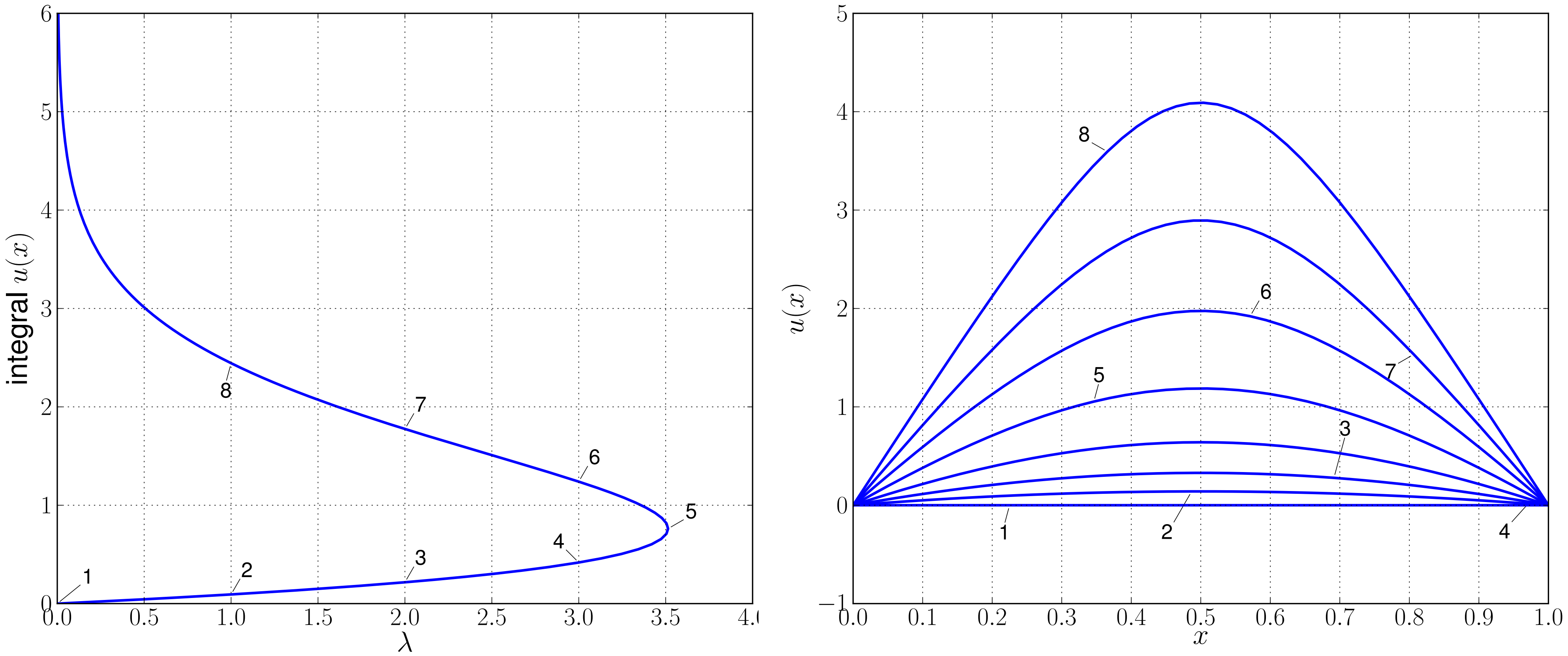
$$\begin{pmatrix} -\frac{2}{h^2} + \lambda e^{u_1^{(k)}} & \frac{1}{h^2} & \\ \frac{1}{h^2} & -\frac{2}{h^2} + \lambda e^{u_2^{(k)}} & \frac{1}{h^2} \\ & \cdot & \cdot \\ & \frac{1}{h^2} & -\frac{2}{h^2} + \lambda e^{u_{N-1}^{(k)}} \end{pmatrix} \Delta\mathbf{u}^{(k)} \ = \ - \mathbf{G}(\mathbf{u}^{(k)}) \ ,$$

where

$$\Delta\mathbf{u}^{(k)} \ \equiv \ (\Delta u_1^{(k)} \ , \ \Delta u_2^{(k)} \ , \ \cdots \ , \ \Delta u_{N-1}^{(k)})^T \ ,$$

and updating

$$\mathbf{u}^{(k+1)} \ = \ \mathbf{u}^{(k)} \ + \ \Delta\mathbf{u}^{(k)} \ .$$



Solutions of the Gelfand-Bratu equations for different values of λ.

DIFFUSION PROBLEMS

Here we consider *parabolic partial differential equations*.

The simplest is the linear diffusion equation or *heat equation* :

$$
\begin{aligned}
u_t(x,t) &= u_{xx}(x,t) \ , \qquad x \in [0,1] \ , \quad t \geq 0 \ , \\
u(x,0) &= g(x) \ , \\
u(0,t) &= u(1,t) \ = \ 0 \ .
\end{aligned}
$$

This equation governs, for example, the temperature in an insulated rod of which the endpoints are kept at the constant temperature zero, and in which the initial temperature distribution is $g(x)$.

First *discretize in space* :

$$u_j'(t) \;=\; \frac{u_{j-1}(t) - 2u_j(t) + u_{j+1}(t)}{\Delta x^2} \;,$$

$$u_j(0) \;=\; g(x_j) \;,$$

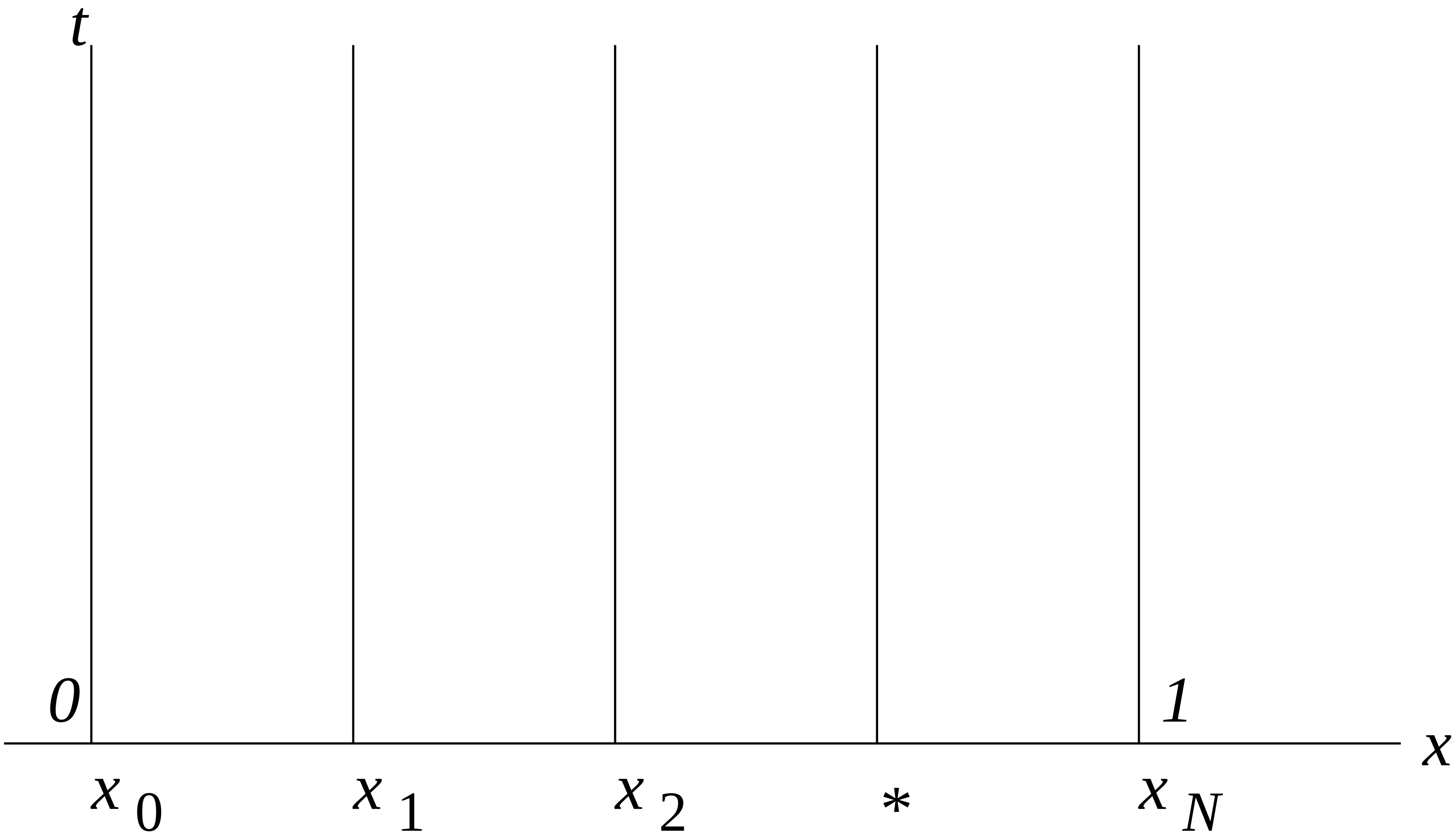
$$u_0(t) \;=\; u_N(t) \;=\; 0 \;,$$

where we have introduced the *notation*

$$u_j(t) \;\equiv\; u(x_j,t) \;,$$

and where $'$ denotes differentiation with respect to t.

These space-discretized equations represents a system of $N-1$ coupled *ordinary* differential equations.


t
0
1
x
x_0
x_1
x_2
*
x_N

In *matrix-vector notation* we can write the space-discretized equations as

$$\mathbf{u}'(t) \ = \ \frac{1}{\Delta x^2} \, \mathbf{D} \, \mathbf{u}(t) \ ,$$

where

$$\mathbf{D} \ \equiv \ \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \cdot & \cdot & \cdot & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \ ,$$

and

$$\mathbf{u} \ \equiv \ \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ u_{N-2} \\ u_{N-1} \end{pmatrix} \ .$$

Now discretize in *time* :

Often used is the *Trapezoidal rule* :

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^{k}}{\Delta t} = \frac{1}{2\Delta x^2} \mathbf{D} \, \{\mathbf{u}^{k+1} + \mathbf{u}^{k}\} \, ,$$

where

$$\mathbf{u}^{k} \equiv \begin{pmatrix} u_1^k \\ u_2^k \\ \cdot \\ u_{N-1}^k \end{pmatrix} ,$$

and

$$u_j^k \text{ approximates } u(x_j, t^k) \, .$$

Assume that the solution has been computed up to time t^k.

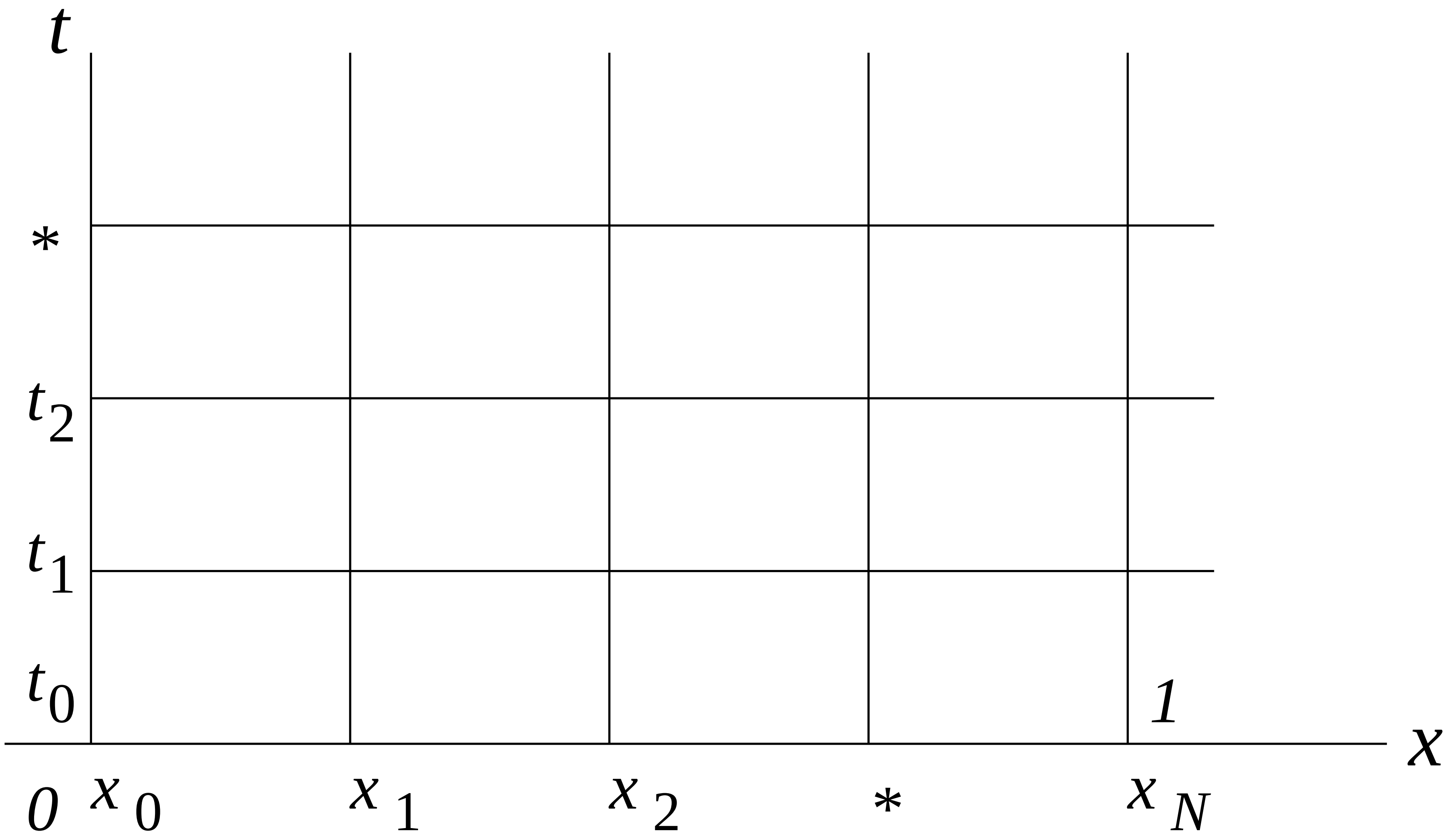
Thus $\mathbf{u}^k$ is *known* , and we want to *solve* for $\mathbf{u}^{k+1}$.

Rewrite the above equation as

$$(\mathbf{I} - \frac{\Delta t}{2\Delta x^2}\mathbf{D})\ \mathbf{u}^{k+1} \ = \ (\mathbf{I} + \frac{\Delta t}{2\Delta x^2}\mathbf{D})\ \mathbf{u}^k \ .$$

Thus to take a step in time we have to solve a *tridiagonal* linear system.

This method is also known as the *Crank-Nicolson scheme*.


t
*
t_2
t_1
t_0
1
x
0
x_0
x_1
x_2
*
x_N

NOTE :

- We can also use an *explicit* method in time, for example explicit Euler.

- But this can be a *bad choice* because the ODE system is *stiff*.

- The time step Δt may have to be *very small* to have stability.

For the system of ODEs

$$\mathbf{u}'(t) \ = \ \frac{1}{\Delta x^2}\,\mathbf{D}\ \mathbf{u}(t)\ , \quad \text{with} \quad \mathbf{D} \ = \ \text{diag}[\ 1\ , -2\ ,\ 1\ ]\ ,$$

we can *demonstrate the stiffness* analytically.

In fact, we can explicitly compute the *eigenvalues* of the matrix

$$\frac{1}{\Delta x^2}\,\mathbf{D}\ ,$$

as follows :

An *eigenvalue-eigenvector* pair λ, $\mathbf{v}$ satisfies

$$\frac{1}{\Delta x^2}\,\mathbf{D}\mathbf{v} \ = \ \lambda\mathbf{v}\ ,$$

that is,

$$\frac{1}{\Delta x^2}(v_{j-1} - 2v_j + v_{j+1}) \ = \ \lambda v_j\ , \qquad j = 1, 2, \cdots, N-1\ , \qquad v_0 \ = \ v_N \ = \ 0\ .$$

We had the *difference equation*

$$\frac{1}{\Delta x^2}(v_{j-1} - 2v_j + v_{j+1}) \ = \ \lambda v_j \ .$$

Try a solution of the form $v_j \ = \ z^j$.

This gives the *characteristic equation*

$$z^2 \ - \ (2 + \Delta x^2 \lambda) z \ + \ 1 \ = \ 0 \ ,$$

or

$$\lambda \ = \ \frac{z + z^{-1} - 2}{\Delta x^2} \ .$$

The characteristic equation has *zeroes*

$$z \ = \ z_1 \qquad \text{and} \qquad z \ = \ z_1^{-1} \ .$$

The *general solution* of the difference equation then has the form

$$v_j \ = \ c_1 z_1^j \ + \ c_2 z_1^{-j} \ .$$

From the *first boundary condition* we have

$$v_0 \;=\; 0 \quad \Rightarrow \quad c_1 + c_2 \;=\; 0 \; .$$

Thus we can take

$$c_1 \;=\; c \qquad \text{and} \qquad c_2 \;=\; -\,c \; .$$

Then

$$v_j \;=\; c\,(z_1^j - z_1^{-j}) \; .$$

From the *second boundary condition* we now find

$$v_N \;=\; 0 \quad \Rightarrow \quad c\,(z_1^N - z_1^{-N}) \;=\; 0 \; ,$$

from which

$$z_1^{2N} \;=\; 1 \qquad \Rightarrow \quad z_1 \;=\; e^{\frac{k2\pi i}{2N}} \; .$$

The *eigenvalues* are therefore

$$
\begin{aligned}
\lambda_k &= \frac{z + z^{-1} - 2}{\Delta x^2} \\
&= \frac{e^{\frac{k2\pi i}{2N}} + e^{-\frac{k2\pi i}{2N}} - 2}{\Delta x^2} \\
&= \frac{2\cos(\frac{k2\pi}{2N}) - 2}{\Delta x^2} \\
&= \frac{2(\cos(\frac{k2\pi}{2N}) - 1)}{\Delta x^2} \\
&= -\frac{4}{\Delta x^2}\ \sin^2(\frac{k\pi}{2N})\ , \qquad k = 1, 2, \cdots, N-1\ .
\end{aligned}
$$

The eigenvalue with *largest negative real part* is

$$\lambda_{N-1} \; = \; - \frac{4}{\Delta x^2} \; \sin^2(\frac{(N-1)\pi}{2N}) \; ,$$

which for large N behaves like

$$\lambda_{N-1} \; \cong \; \lambda^* \; \equiv \; - \frac{4}{\Delta x^2} \; .$$

Thus the system is *stiff* if Δx is small .

EXAMPLE : To make the explicit Euler method stable we need to take the timestep Δt so that $\Delta t \lambda^*$ lies in the circle of radius 1 centered at -1, *i.e.*, we must take

$$\Delta t \; < \; \frac{1}{2} \Delta x^2 \; .$$

Using explicit Euler is often *not a good idea* .

Nonlinear Diffusion Equations.

An example of a nonlinear diffusion equation is *the Fisher equation*

$$u_t(x,t) \;=\; u_{xx}(x,t) \;+\; \lambda\, u(x,t)\ (1-u(x,t))\ ,$$

for

$$x \in [0,1]\ , \qquad t \geq 0\ ,$$

with

$$u(x,0) \;=\; g(x)\ , \qquad u(0,t) \;=\; u(1,t) \;=\; 0\ .$$

This is a simple model of *population growth* with *diffusion* and with *maximal sustainable population* equal to 1 .

Another example is *the time-dependent Gelfand-Bratu equation*

$$u_t(x,t) \;=\; u_{xx}(x,t) \;+\; \lambda\, e^{u(x,t)} \;,$$

for

$$x \in [0,1] \;, \quad t \geq 0 \;,$$

with

$$u(x,0) \;=\; g(x) \;, \qquad u(0,t) \;=\; u(1,t) \;=\; 0 \;,$$

for which we have already considered the *stationary equations*

$$u_{xx}(x) \;+\; \lambda\, e^{u(x)} \;=\; 0 \;, \qquad x \in [0,1] \;,$$

with

$$u(0) \;=\; u(1) \;=\; 0 \;.$$

We illustrate the numerical solution procedure for the *general equation*

$$\begin{aligned} u_t(x,t) &= u_{xx}(x,t) + f(u(x,t)) \ , \qquad x \in [0,1] \ , \quad t \geq 0 \ , \\ u(x,0) &= g(x) \ , \\ u(0,t) &= u(1,t) = 0 \ , \end{aligned}$$

where

$$f(u) = \lambda \, u \, (1-u) \quad \text{for the} \quad \textit{Fisher equation} \ ,$$

and

$$f(u) = \lambda \, e^u \quad \text{for the} \quad \textit{Gelfand-Bratu equation} \ .$$

We approximate this equation as follows :

First *discretize in space* to get a *system of ODEs* :

$$u_j'(t) \;=\; \frac{u_{j-1}(t) - 2u_j(t) \;+\; u_{j+1}(t)}{\Delta x^2} \;+\; f(u_j(t)) \;,$$

for $j = 1, 2, \cdots, N-1$, with

$$u_j(0) \;=\; g(x_j) \;,$$

$$u_0(t) \;=\; u_N(t) \;=\; 0 \;.$$

Then *discretize in time* using *Implicit Euler* :

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} \;=\; \frac{u_{j-1}^{k+1} - 2u_j^{k+1} + u_{j+1}^{k+1}}{\Delta x^2} \;+\; f(u_j^{k+1}) \;.$$

Rewrite these equations as

$$F_j^{k+1} \equiv u_j^{k+1} - u_j^k - \frac{\Delta t}{\Delta x^2}(u_{j-1}^{k+1} - 2u_j^{k+1} + u_{j+1}^{k+1}) - \Delta t \ f(u_j^{k+1}) = 0 \, ,$$

for $j = 1, 2, \cdots, N-1$,

with

$$u_0^{k+1} = 0 \qquad \text{and} \qquad u_N^{k+1} = 0 \, .$$

We can assume that the solution has been computed up to time t^k,

i.e., the u_j^k *are known* and we must *solve for* the u_j^{k+1} .

Since *the equations are nonlinear* we use Newton's method.

As *initial approximation* to u_j^{k+1} in Newton's method use

$$(u_j^{k+1})^{(0)} = u_j^k, \qquad j = 1, 2, \cdots, N-1 .$$

Each *Newton iteration* then consists of solving a linear *tridiagonal system*

$$\mathbf{T}^{k+1,(\nu)} \, \mathbf{\Delta u}^{k+1,(\nu)} = - \mathbf{F}^{k+1,(\nu)} ,$$

where

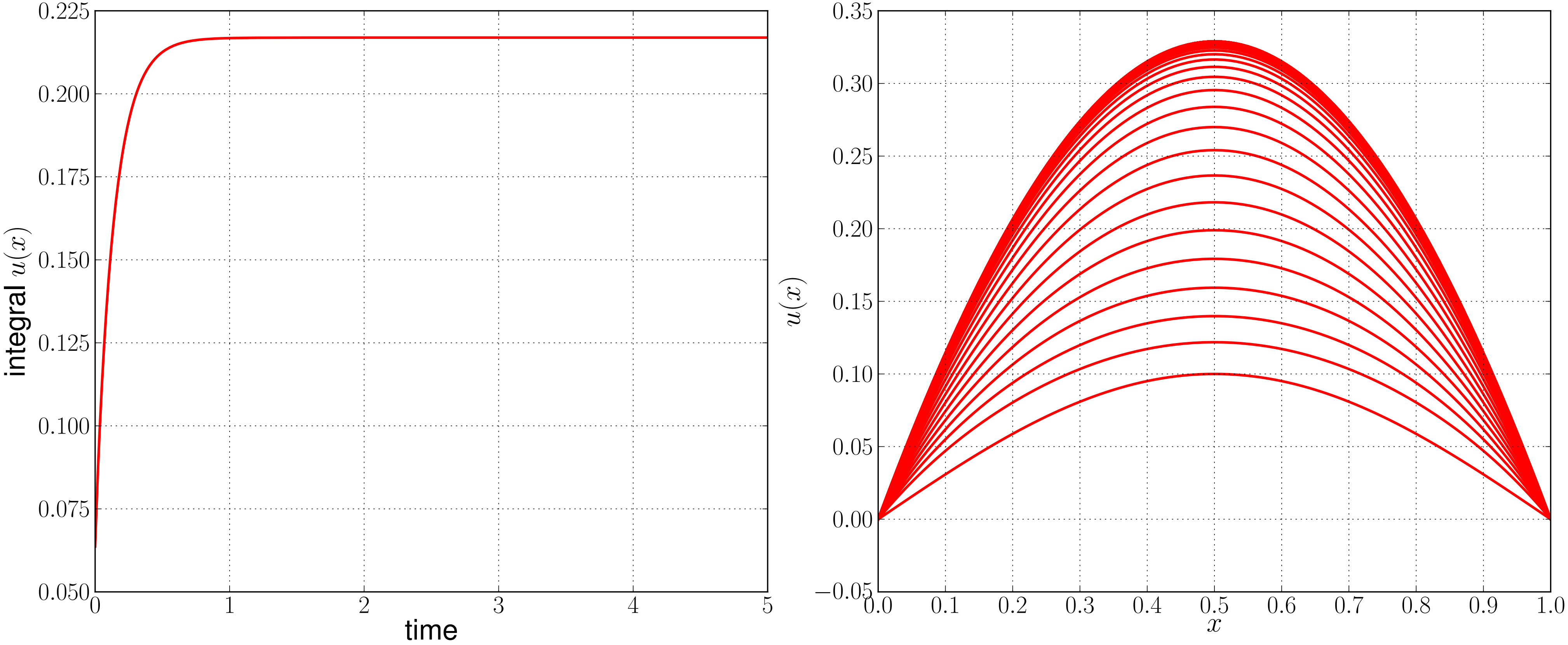
$$\mathbf{T}^{k+1,(\nu)} = \begin{pmatrix} 1+2\frac{\Delta t}{\Delta x^2} - \Delta t f_u(u_1^{k+1,(\nu)}) & -\frac{\Delta t}{\Delta x^2} & \\ -\frac{\Delta t}{\Delta x^2} & 1+2\frac{\Delta t}{\Delta x^2} - \Delta t f_u(u_2^{k+1,(\nu)}) & -\frac{\Delta t}{\Delta x^2} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ & -\frac{\Delta t}{\Delta x^2} & 1+2\frac{\Delta t}{\Delta x^2} - \Delta t f_u(u_{N-1}^{k+1,(\nu)}) \end{pmatrix}$$

and

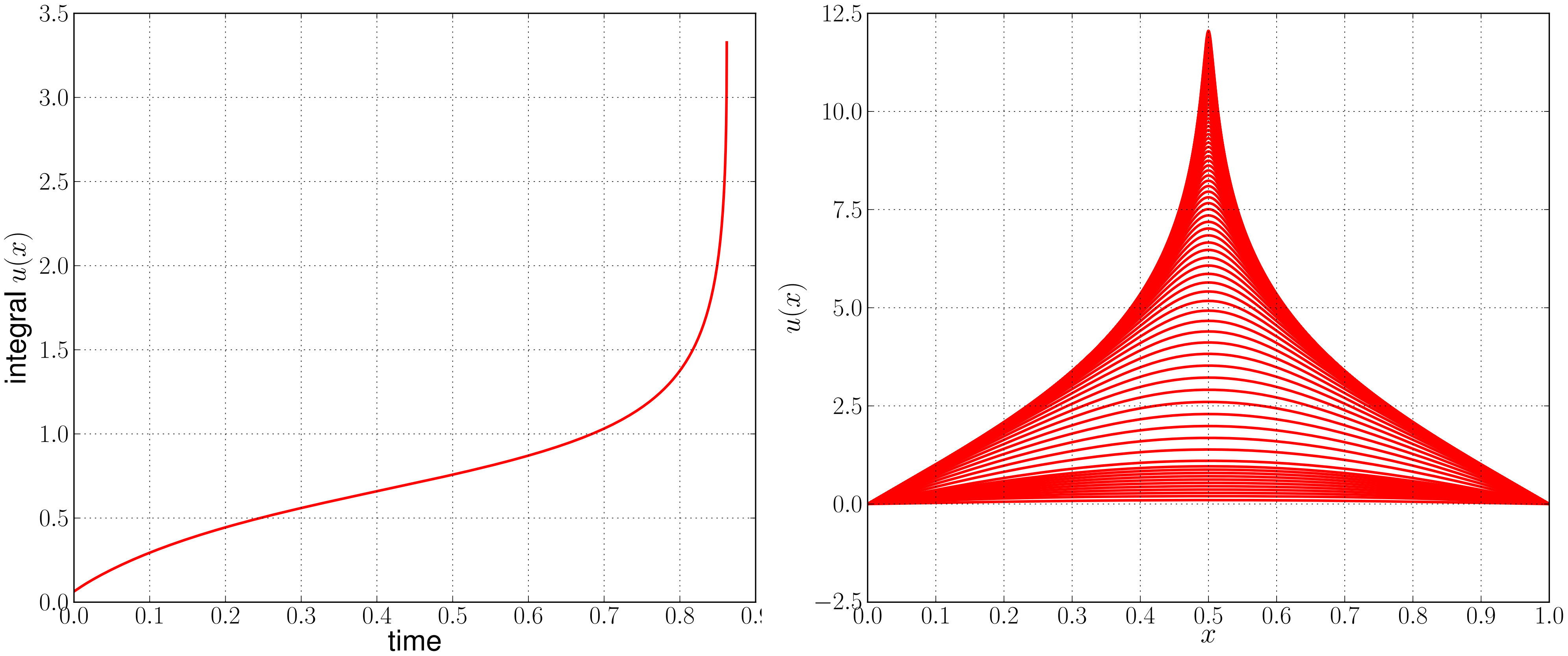
$$\mathbf{\Delta u}^{k+1,(\nu)} = \begin{pmatrix} \Delta u_1^{k+1,(\nu)} \\ \Delta u_2^{k+1,(\nu)} \\ \cdot \\ \Delta u_{N-1}^{k+1,(\nu)} \end{pmatrix} , \quad \mathbf{F}^{k+1,(\nu)} = \begin{pmatrix} F_1^{k+1,(\nu)} \\ F_2^{k+1,(\nu)} \\ \cdot \\ F_{N-1}^{k+1,(\nu)} \end{pmatrix} .$$

Then set the *next approximation* to the solution at time $t = t^{k+1}$ equal to

$$\mathbf{u}^{k+1,(\nu+1)} = \mathbf{u}^{k+1,(\nu)} + \mathbf{\Delta u}^{k+1,(\nu)} .$$



Time-evolution of solutions of the Gelfand-Bratu equations for $\lambda = 2$.



Time-evolution of solutions of the Gelfand-Bratu equations for $\lambda = 4$.