Numerical Root-finding

Introduction to root-finding algorithms (1/2)

Problem: Given three real numbers a, b, c solve this equation:

$$ax^2+bx+c = 0$$

We know a closed-form solution:

$$x_1, x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Now solve this other equation: $e^{ax}=e^{bx}$ $(a \neq b)$

Again we can find a closed-form solution: $x = 0$

However, let try to solve: $x^2+\ln(x)=0$

We cannot find a closed-form solution to this equation.

Introduction to root-finding algorithms (2/2)

$x^2+\ln(x)=0$ f(x)=0

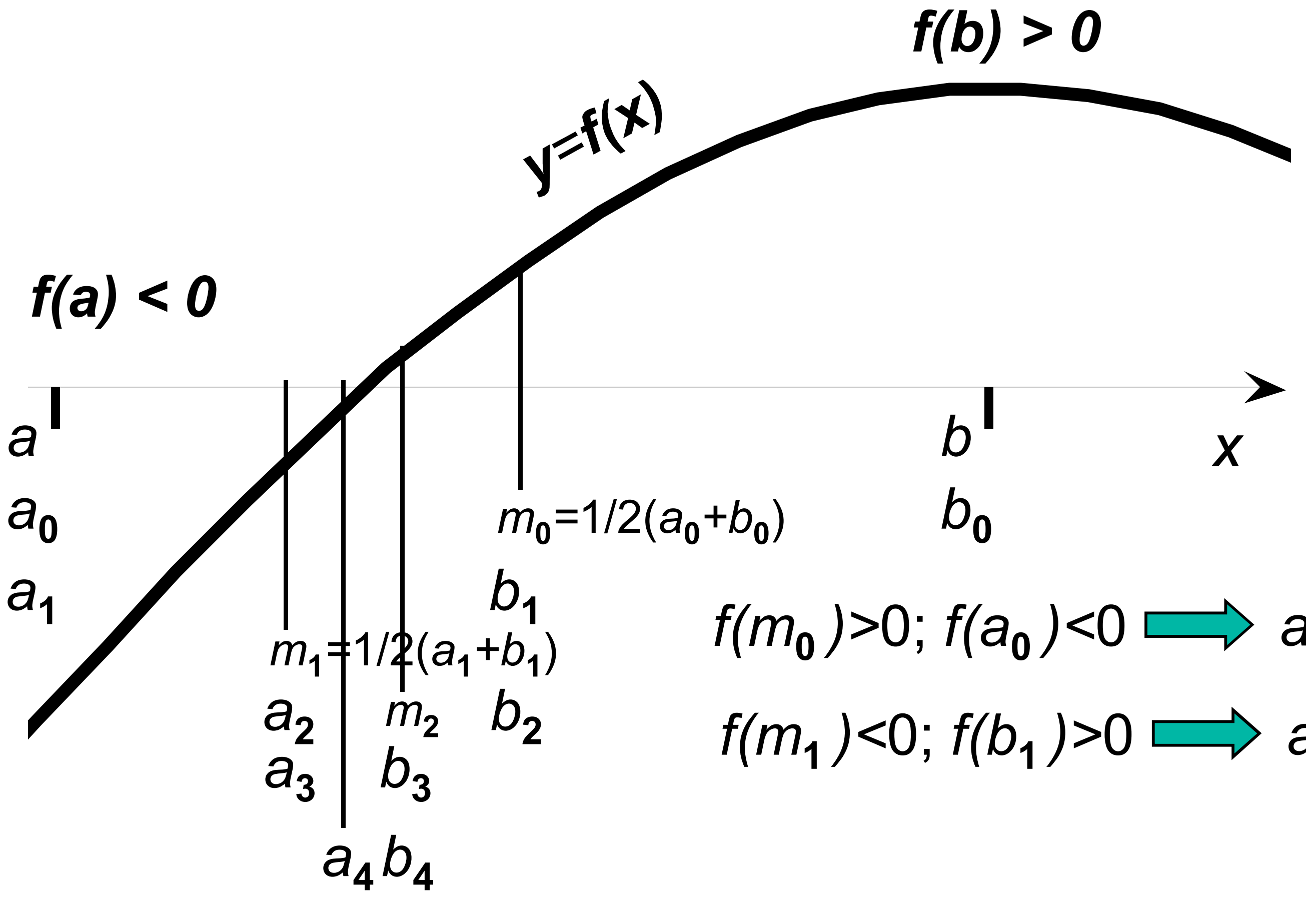
Let search for a solution by means of an iterative method:

- **we graphically find the following interval: [1/e,1].**
- **f(1/e)<0; f(1)>0. Then there is a solution in this interval.**
- **let try a smaller subinterval: [(1+e)/(2e),1]**
- **this interval has an amplitude ½ the previous one**
- **a solution is still comprised between (1+e)/(2e) and 1**
- **we go further, each time subdividing the last interval**

By doing so, we never get a precise solution, but each time we reduce the interval the solution is included in.

This is the concept of the bisection method.

The bisection method



$f(m_0)>0;\ f(a_0)<0 \Rightarrow a_1 = a_0;\ b_1 = m_0$

$f(m_1)<0;\ f(b_1)>0 \Rightarrow a_2 = m_1;\ b_2 = b_1$

Bisection algorithm

$f(a)<0$ and $f(b)>0$ (or $f(a)>0$ and $f(b)<0$)

```
FIRST STEP:
    $a_0=a;\ b_0=b$
    if $f((a_0+b_0)/2)f(a_0)>0$
    then $a_1=((a_0+b_0)/2);\ b_1=b_0$
    else
        $a_1=a_0;\ b_1=((a_0+b_0)/2)$

GENERIC STEP:
    if $f((a_{n-1}+b_{n-1})/2)f(a_{n-1})>0$
    then $a_n=((a_{n-1}+b_{n-1})/2);\ b_n=b_{n-1}$
    else
        $a_n=a_{n-1};\ b_n=((a_{n-1}+b_{n-1})/2)$
```

Error in the bisection method

a_0, a_1, a_2,…, a_n,… b_0, b_1, b_2,…, b_n,…

are two convergent sequences.

They both converge to the solution of the equation f(x)=0.

If we use either a_n or b_n as an approximate solution for f(*x*)=0 we introduce an error depending on the amplitude of the [a_n,b_n] interval, according to the following relationship:

$$E \le |b_n - a_n| = \frac{|b - a|}{2^n}$$

Bisection method: number of steps

$$E \leq \frac{|b-a|}{2^n} < \text{precision}$$

$$2^n > \frac{|b-a|}{\text{precision}}$$

$$n > \log_2\left(\frac{|b-a|}{\text{precision}}\right)$$

Bisection method: decimal precision

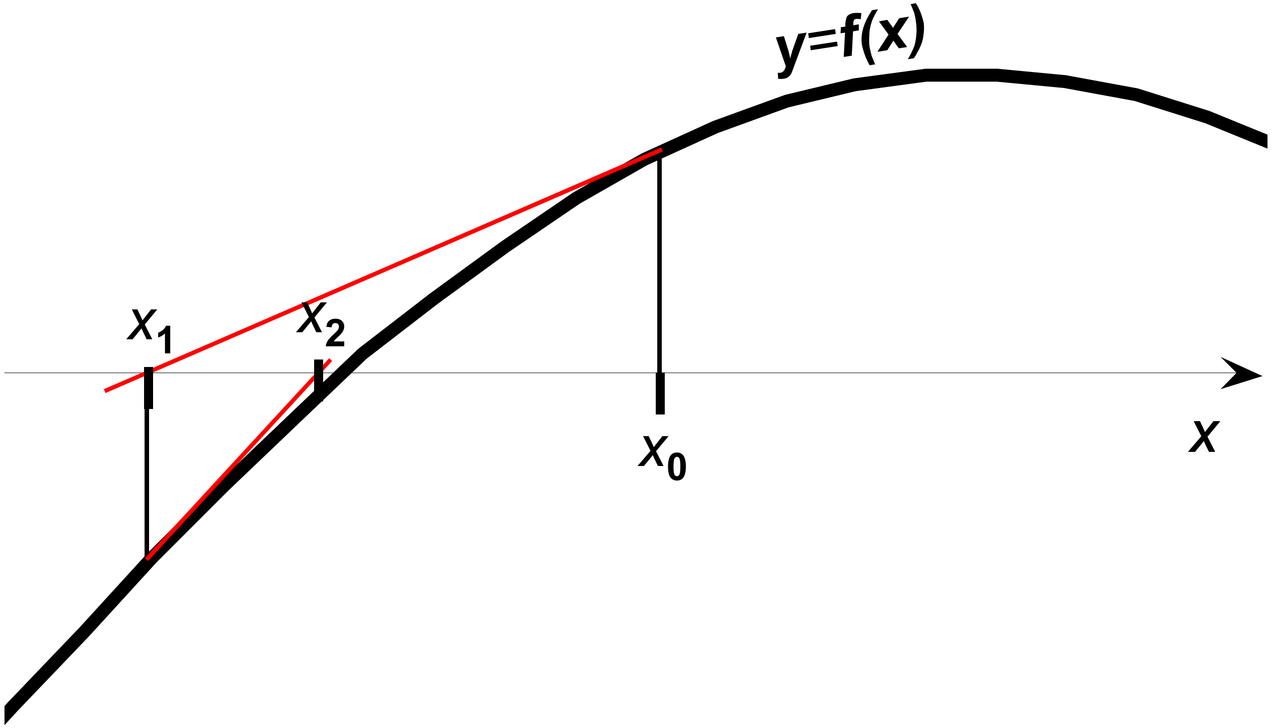
$$E \leq \frac{|b - a|}{2^n} < \text{precision}$$

From this equation we see that we need four consecutive steps to get a further correct decimal digit ($(1/2)^4 < 1/10$).

This property classifies the bisection method as a slowly convergent one.

On the positive side, the method is guaranteed to converge.

The Newton's method



Approximation in the Newton's method

$x_0, x_1, x_2, \ldots, x_n, \ldots$

is a sequence of approximate values.

x_0 is an initial guess that should not be too far from the solution.

This sequence converges to the solution of the equation f(x)=0.

The convergence is quadratic: at each step the error is basically squared.

Under this aspect, the Newton's method is faster than the bisection method (it requires less steps). The bisection method's convergence is linear (at each step the error is multiplied by a constant factor).

Newton's method: caveats and practical considerations

If x_0 is not sufficiently close to the solution, the method can fail to converge.

For this reason, most practical implementations put an upper limit on the number of iterations.

By contrast, the bisection method always converges.

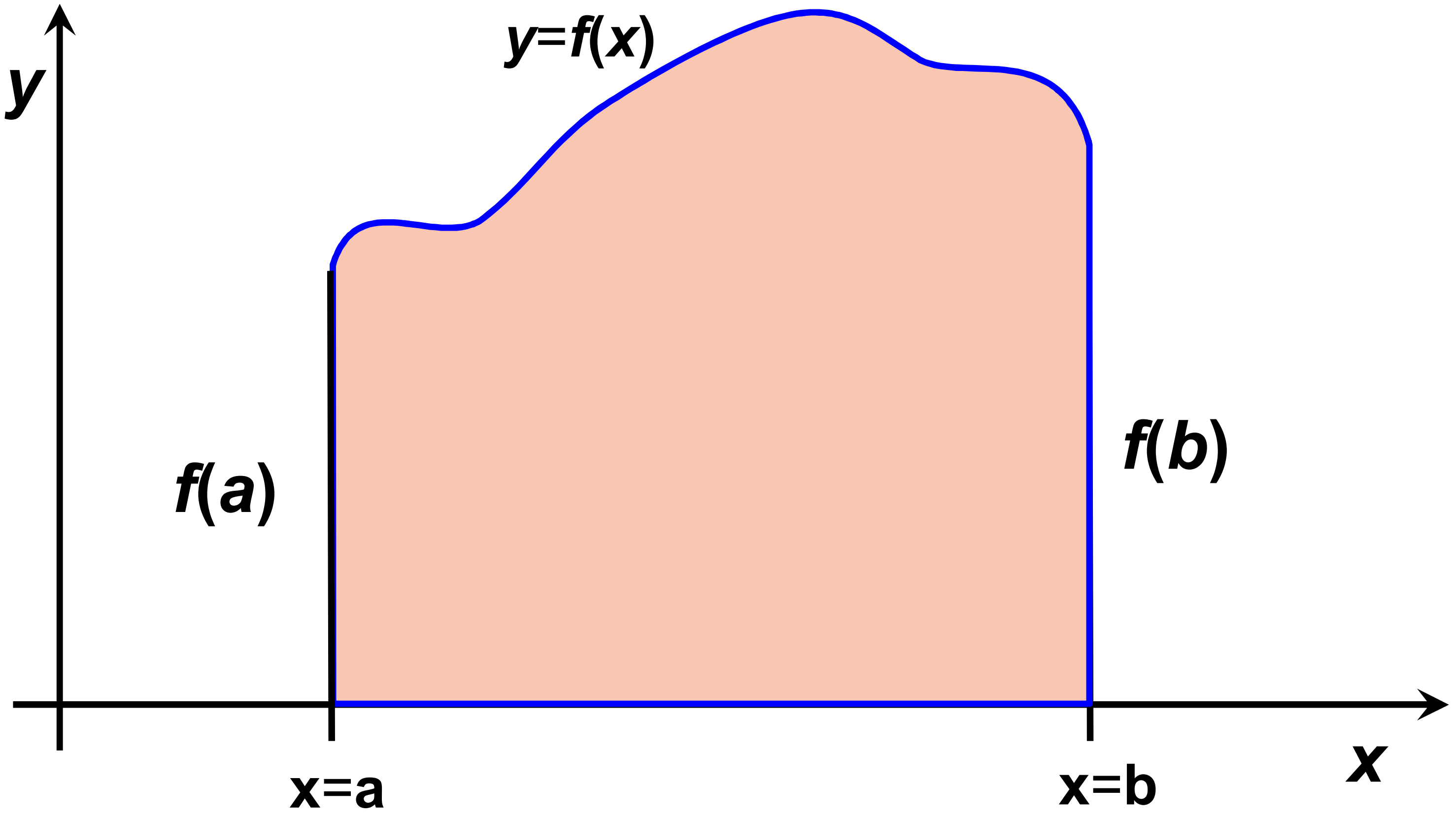
Given a required precision, the number of iterations cannot be determined a priori.

At each step one must evaluate not only the function, but also its derivative.

If the root being sought has multiplicity greater than one, the method's convergence is linear.

Integration Problems

Area of a flat zone in plane *xy*



Determine the area of the plane portion of the *xy*-plane delimited by:

- curve *y*=*f*(*x*);
- *x*-axis;
- straight lines *x*=a, *x*=b,

$$A = \int_a^b f(x)dx$$

Area of a plane zone: example #1

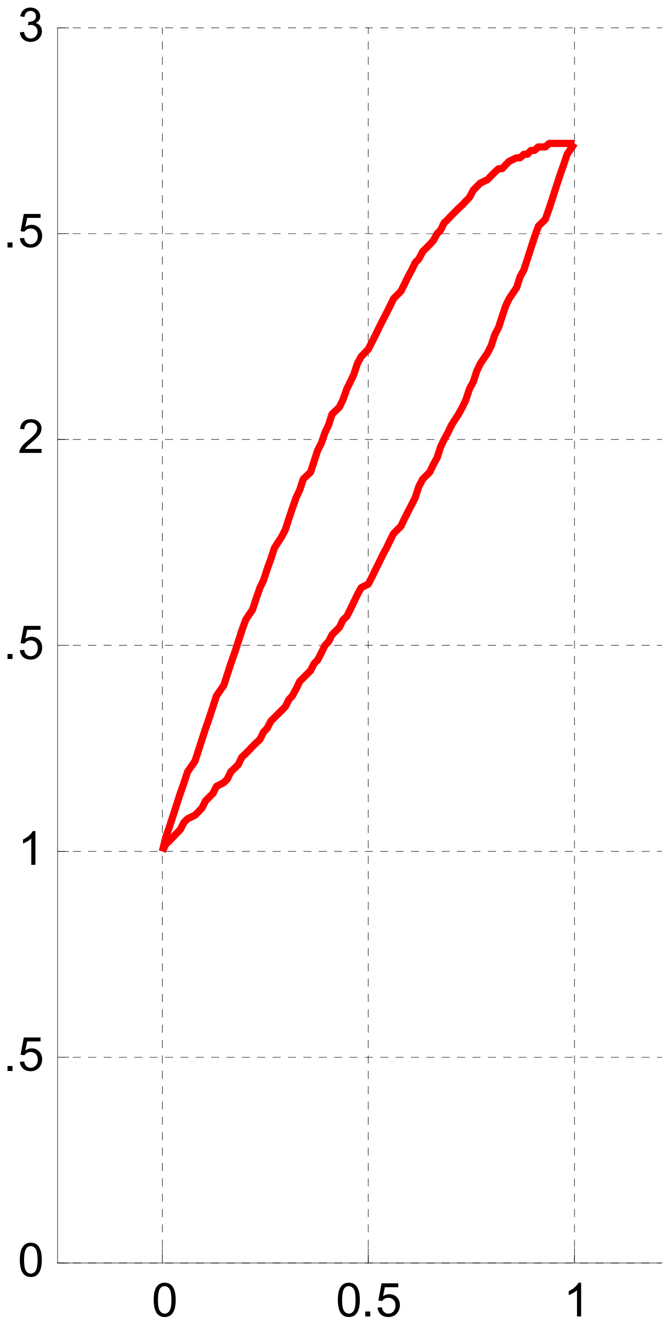
Determine the area of the plane zone comprised between curves:

- $y=(e-1)\sin(\pi/2x)+1;$
- $y=e^x$

$$A=\int_0^1((e-1)\sin(\frac{\pi}{2}x)+1-e^x)dx$$

$$A=\left[-(e-1)\frac{2}{\pi}\cos(\frac{\pi}{2}x)+x-e^x\right]_0^1$$

$$A=1-e+(e-1)\frac{2}{\pi}+1\cong 0.3756$$

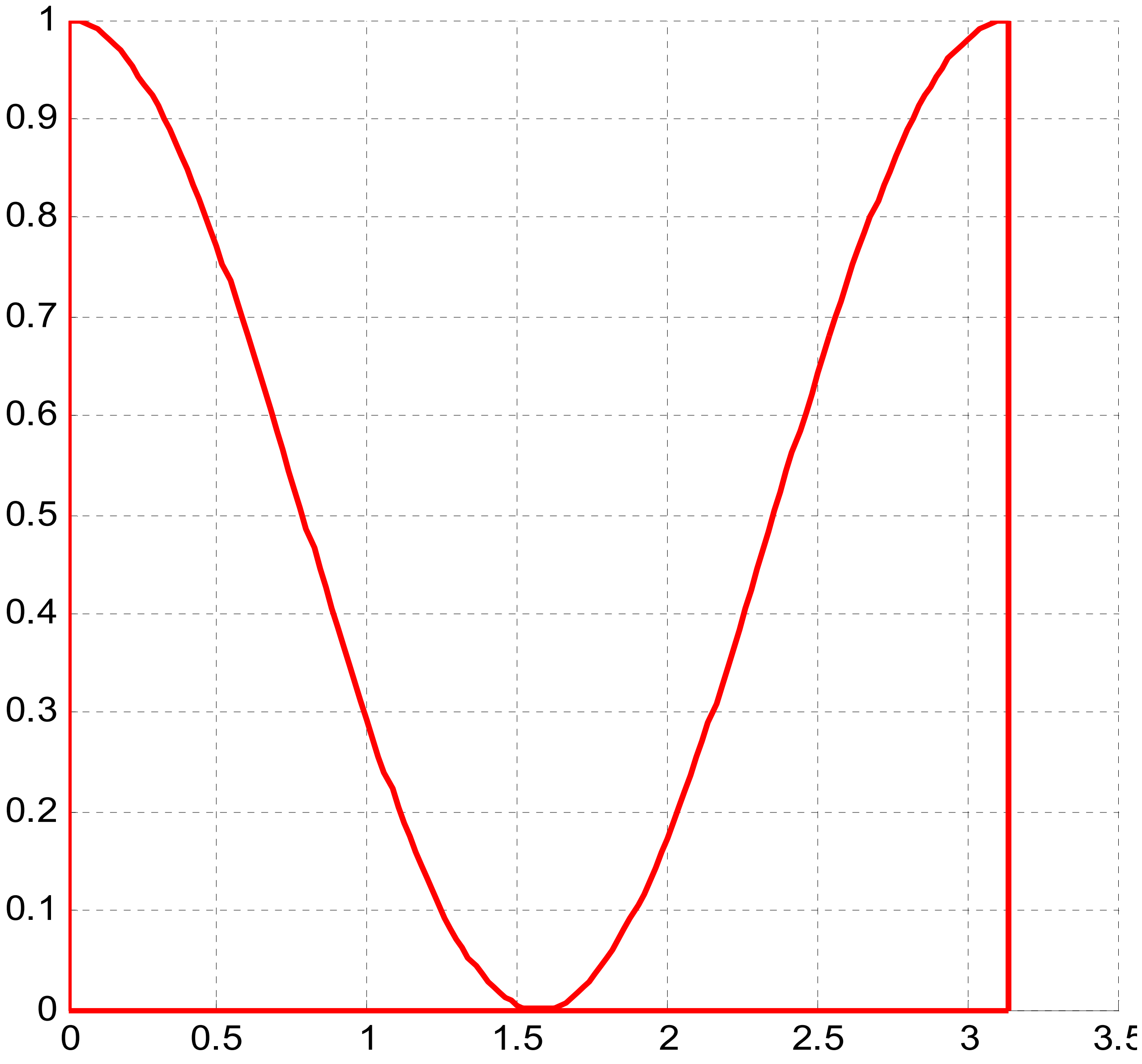


Area of a plane zone: example #2

Determine the area of the zone bounded by curve y=cos2x, the x-axis, straight lines x=0 and x=π

$$A = \int_0^{\pi} \cos^2 x dx$$

How can we compute this integral?



Numerical Integration

Numerical Integration (1/3)

Problem:

Evaluate the following expression

$$\int_a^b f(x)dx$$

where *a* and *b* are constants and *f*(*x*) is a continuous function.

(continued)

Numerical Integration (2/3)

Not all times we can find an antiderivative of the integrand function.

Even if we know it, frequently its computation is very complicated. Moreover, in any case we need to get a numerical value at the end of the process.

Our problem is then to translate the symbolic model of the definite integral into an appropriate numerical model (i.e. a computational procedure).

(continued)

Numerical Integration (3/3)

The numerical method is based on the following steps:

1. **We select some values x_i (knots) and build a table of n+1 pairs (x_i,$f(x_i)$) i=0,1,…,n where $f(x_i)$ are values of the integrand.**
2. **We determine a polynomial interpolating the n+1 pairs (x_i,$f(x_i)$)**
3. **We calculate the definite integral of the interpolating polynomial, assumed as an approximation of the integrand function.**
4. **We use this integral as an approximation of the wanted integral and estimate the error.**

Which polynomial do we use?

The polynomial in its Lagrangian form (Lagrange polynomial).

Lagrange polynomials (1/4)

Given two points in he plane (i.e. two pairs), e.g. (1,2) and (3,4)

The first-degree polynomial:

f(x) = x+1 is such that: f(1)=2, f(3)=4

Then we say that f(x) interpolates the two given pairs, their abscissas are also called interpolation knots

We note that the function f(x) = x+1 may be replaced by an equivalent function (that is by another polynomial that still interpolates the two given pairs), like for instance:

-(x-3) + 2(x-1)

We say that we changed the **polynomial basis**.

(continued)

Lagrange polynomials (2/4)

***In general*: given n+1 pairs we can determine a unique n-degree polynomial interpolating the given n+1 pairs.**

A frequently used expression for such a generic polynomial is the Lagrangian form: this is a combination of some specific base polynomials, named base Lagrange polynomials

Given a set of n+1 values $\mathbf{x_0}$ **,** $\mathbf{x_1}$ **, …,** $\mathbf{x_n}$**, we define a base Lagrange polynomial the following expression:**

$$L_i(x) = \frac{(x - x_0)(x - x_1)\dots(x - x_{i-1})(x - x_{i+1})\dots(x - x_n)}{(x_i - x_0)(x_i - x_1)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n)}$$

(continued)

Lagrange polynomials (3/4)

$$L_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}$$

- is an n-degree polynomial
- it is such that:

$$L_i(x_j) = 1 \qquad i = j$$

$$L_i(x_j) = 0 \qquad i \neq j$$

If we now take the n+1 function values we define an **n-degree polynomial as follows:**

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

(continued)

Lagrange polynomials (4/4)

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

It follows that:

$f_n(x_i) = f(x_i)$ $i = 0,1,\ldots,n$

The function $f_n(x)$

- is an n-degree polynomial;
- interpolates the (n+1) pairs $(x_i, f(x_i))$ $i = 0,1,\ldots,n$

Example (Lagrange polynomial) (1/2)

$$L_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}$$

We are given the three pairs $(-1,2),(4,3),(1,-2)$

$$L_0(x) = \frac{(x-4)(x-1)}{(-1-4)(-1-1)} = \frac{(x-4)(x-1)}{10}$$

$$L_1(x) = \frac{(x+1)(x-1)}{(4+1)(4-1)} = \frac{(x+1)(x-1)}{15}$$

$$L_2(x) = \frac{(x+1)(x-4)}{(1+1)(1-4)} = -\frac{(x+1)(x-4)}{6}$$

(continued)

Example (Lagrange polynomial) (2/2)

$$f_n(x) = \sum_{i=0}^{n} L_i(x) y_i$$

$$(-1,2),(4,3),(1,-2)$$

Now we build the second-degree Lagrange polynomial:

$$f_2(x) = \frac{(x-4)(x-1)}{10}2 + \frac{(x+1)(x-1)}{15}3 - \frac{(x+1)(x-4)}{6}(-2)$$

Error introduced by the polynomial interpolation

By definition, the polynomial gives an exact value when evaluated in the knots.
What happens outside these values?

It can be shown that, if ***I*** is the interval comprising all knots:

$$\boxed{\begin{array}{l} E_n f(x) = f(x) - f_n(x) = \dfrac{f^{n+1}(\xi)}{(n+1)!} \displaystyle\prod_{i=0}^{n} (x - x_i) \\ \xi \in I \end{array}}$$

N.B. The error depends not only on the number of knots, but also on their distribution

Numerical Integration through interpolation (1/2)

Given the integral $\int_a^b f(x)dx$

$$(x_i, f(x_i)) \quad i = 0,1,\ldots,n$$

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

$$\int_a^b f_n(x)dx = \int_a^b \sum_{i=0}^{n} L_i(x) f(x_i) dx$$

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

Numerical Integration through interpolation (2/2)

$$\int_a^b f_n(x)dx = \sum_{i=0}^{n} f(x_i)\int_a^b L_i(x)dx$$

$$\int_a^b f_n(x)dx = \sum_{i=0}^{n} f(x_i)w_i$$

x_i: integration knots

w_i: integration weights

This expression is called a quadrature rule

$$\boxed{\int_a^b f_n(x)dx \cong \int_a^b f(x)dx}$$

The trapezoidal rule (1/2)

As a particular case we use two knots:
the integration limits a and b.

Lagrange polynomial interpolating the two pairs (a,*f*(a)); (b,*f*(b)):

$$f_1(x) = L_0(x)f(a) + L_1(x)f(b)$$

$$L_0(x) = \frac{(x-b)}{a-b} \quad L_1(x) = \frac{(x-a)}{b-a}$$

$$f_1(x) = \frac{(x-b)}{a-b}f(a) + \frac{(x-a)}{b-a}f(b)$$

$$\int_a^b f_1(x)dx = \sum_{i=0}^{1} f(x_i)\int_a^b L_i(x)dx = \sum_{i=0}^{1} f(x_i)w_i$$

(continued)

The trapezoidal rule (2/2)

$$\int_a^b f_1(x)dx = \sum_{i=0}^{1} f(x_i)\int_a^b L_i(x)dx = \sum_{i=0}^{1} f(x_i)w_i$$

$$w_0 = \int_a^b L_0(x)dx = \int_a^b \frac{(x-b)}{a-b}dx = \left[\frac{(x-b)^2}{2(a-b)}\right]_a^b = \frac{b-a}{2}$$

$$w_1 = \int_a^b L_1(x)dx = \int_a^b \frac{(x-a)}{b-a}dx = \left[\frac{(x-a)^2}{2(b-a)}\right]_a^b = \frac{b-a}{2}$$

and finally:

$$\int_a^b f_1(x)dx = \sum_{i=0}^{1} f(x_i)w_i = \frac{b-a}{2}(f(a)+f(b))$$

This is the trapezoidal quadrature rule.

Exercise on the trapezoidal rule (1/2)

Integrand function: $f(x) = \frac{1}{x} - \frac{1}{x^2}$

In this case the antiderivative of *f*(*x*) is known, it is:

$$F(x) = \int (\frac{1}{x} dx - \frac{1}{x^2}) dx = \quad ln|x| + \frac{1}{x} + C$$

In order to determine the definite integral we should numerically compute the function *ln*(*x*), e.g.

$$\int_2^4 (\frac{1}{x} - \frac{1}{x^2}) dx = \left[ln|x| + \frac{1}{x} + C \right]_2^4 \cong 0.4431$$

trying to obtain it with the highest precision

(continued)

Exercise on the trapezoidal rule (2/2)

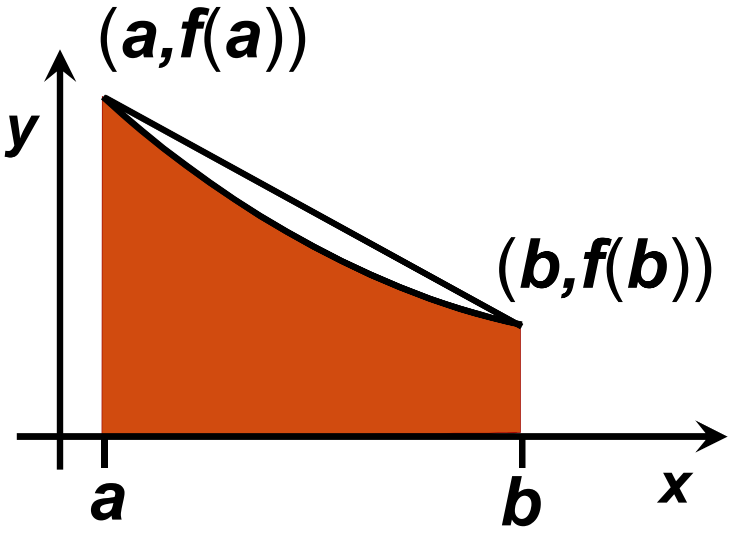
As an alternative, we directly compute the definite integral by means of the trapezoidal rule

$$\int_2^4 \left(\frac{1}{x} - \frac{1}{x^2}\right) dx = \frac{4-2}{2}\left(f(4)+f(2)\right) =$$

$$= \frac{1}{4} - \frac{1}{16} + \frac{1}{2} - \frac{1}{4} =$$

$$= \frac{4-1+8-4}{16} = \frac{7}{16} \cong 0.4375$$

Geometrical interpretation of the trapezoidal rule



$$\int_a^b f(x)dx$$

$$T = \frac{b-a}{2}(f(a)+f(b))$$

If we assume, as in the figure, that the sign of the integrand does not change in [a,b], then the definite integral is equal to the filled area.

The trapezoidal rule gives the area of a trapezoid whose height is (b-a) and parallel sides are *f*(a) e *f*(b).

We approximate the true value with the area of the trapezoid.

Error introduced by the trapezoidal rule

If we calculate the integral of the error due to the Lagrange interpolation, the error introduced by the trapezoidal rule results:

$$E_T = -\frac{(b-a)^3}{12} f''(\eta) \quad \eta \in [a,b]$$

where η is a unpredictable value inside [a,b].

Simpson’s rule (1/2)

In order to improve the accuracy it is possible to use a second degree polynomial over the [a,b] interval.
We use therefore the three following knots:

- **First end a;**
- **Last end b;**
- **Midpoint of [a,b].**

The Lagrange polynomial interpolating the pairs (a,*f*(a)); (b,*f*(b)); ((a+b)/2,*f*((a+b)/2)) is as follows:

$$f_2(x) = L_0(x)f(a) + L_1(x)f\left(\frac{a+b}{2}\right) + L_2(x)f(b)$$

(continued)

Simpson's rule (2/2)

$$f_2(x) = \frac{(x-b)(x-\frac{a+b}{2})}{(a-b)(a-\frac{a+b}{2})}f(a) + \frac{(x-a)(x-\frac{a+b}{2})}{(b-a)(b-\frac{a+b}{2})}f(b) + \frac{(x-a)(x-b)}{(\frac{a+b}{2}-a)(\frac{a+b}{2}-b)}f(\frac{a+b}{2})$$

By determining the definite integral of this polynomial, we obtain the Simpson's quadrature rule:

$$\int_a^b f_2(x)dx = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b))$$

Error introduced by the Simpson’s rule

It can be shown that, by integrating the expression of the Lagrange interpolation error, the Simpson’s rule error is:

$$E_S = -\left(\frac{b-a}{2}\right)^5 \frac{f^{IV}(\eta)}{90} \qquad \eta \in [a,b]$$

Note:

If we increase the number of the (evenly spaced) knots and the corresponding polynomial degree, we obtain a family of rule called **Newton Côtes rules.** This rules are not convergent.

An interesting and very simple case of Newton Côtes rule is the one base on a single knot: the rectangle rule.

The rectangle rule

In this case the approximating expression is:

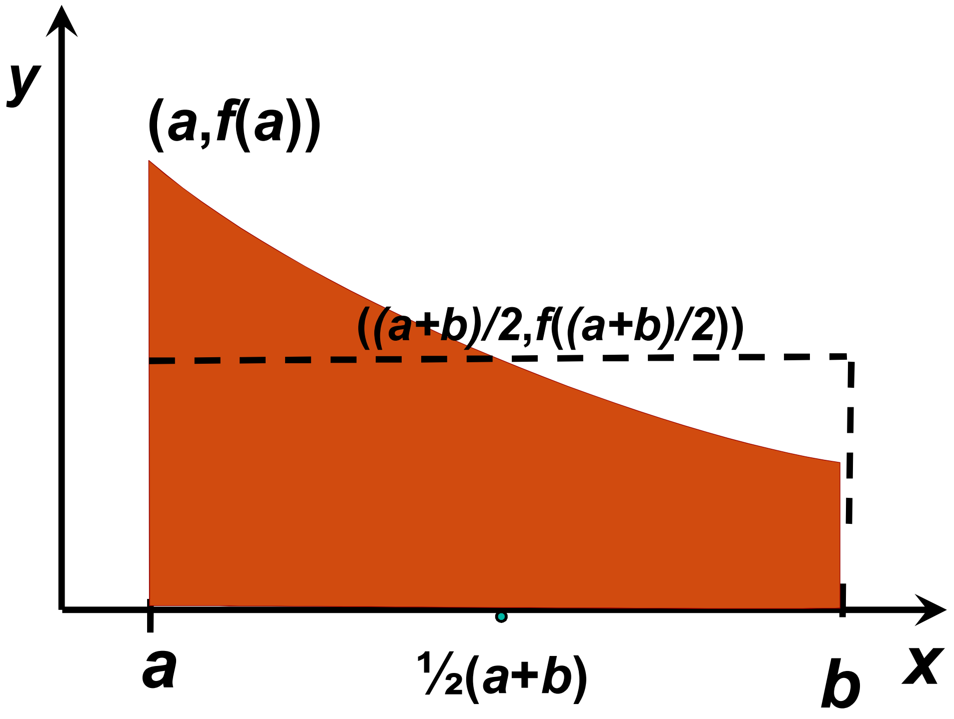
$$\int_a^b f_0(x)dx = (b-a)f(\frac{a+b}{2})$$

giving raise to error: $$E_M = \frac{(b-a)^3}{24} f''(\eta) \quad \eta \in [a,b]$$

The rectangle rule does not use integrand values computed in the interval ends and for this reason is said an open quadrature rule.

It is therefore used to evaluate integrals having singularities in the integration ends.

Geometrical interpretation of the rectangle rule



Assuming *f*(*x*) is positive in the integration interval, then the area of the shaded figure is approximated by the rectangle having base (b-a) and height *f*((a+b)/2).

An application of the rectangle rule

Determining the integral $\int_0^{0.8} \frac{\sin x}{x} dx$

We note that the integrand, even if can be integrated over the interval [0,0.8], is not defined in 0.

Therefore we apply the rectangle rule.

$$\int_a^b f_0(x)dx = (b-a)f(\frac{a+b}{2}) \;=\; 0.8\frac{\sin 0.4}{0.4} \cong 0.77884$$

with error: $E_M = \frac{(b-a)^3}{24} f''(\eta) \quad \eta \in [a,b]$

$$E_M \leq 6.74 \cdot 10^{-3}$$

The composite trapezoidal rule (1/2)

Basic idea: since the error of the trapezoidal rule is strongly dependent on the integration interval (we cannot modify this) we subdivide this interval into many subintervals and we apply the rule to those smaller subintervals.
In this way we obtain a result dependent on the number of subintervals (we can modify this).
Namely:

Be n≥1, h=(b-a)/n, x_j=a+jh j=0,1,…,n, it follows

$$\int_a^b f(x)dx = \sum_{j=0}^{n} \int_{x_j}^{x_{j+1}} f(x)dx = \sum_{j=0}^{n} \frac{h}{2}(f(x_j)+f(x_{j+1})) =$$

$$= \frac{h}{2}(f(x_0)+f(x_n)) + h\sum_{j=1}^{n-1} f(x_j)$$

(continued)

The composite trapezoidal rule (2/2)

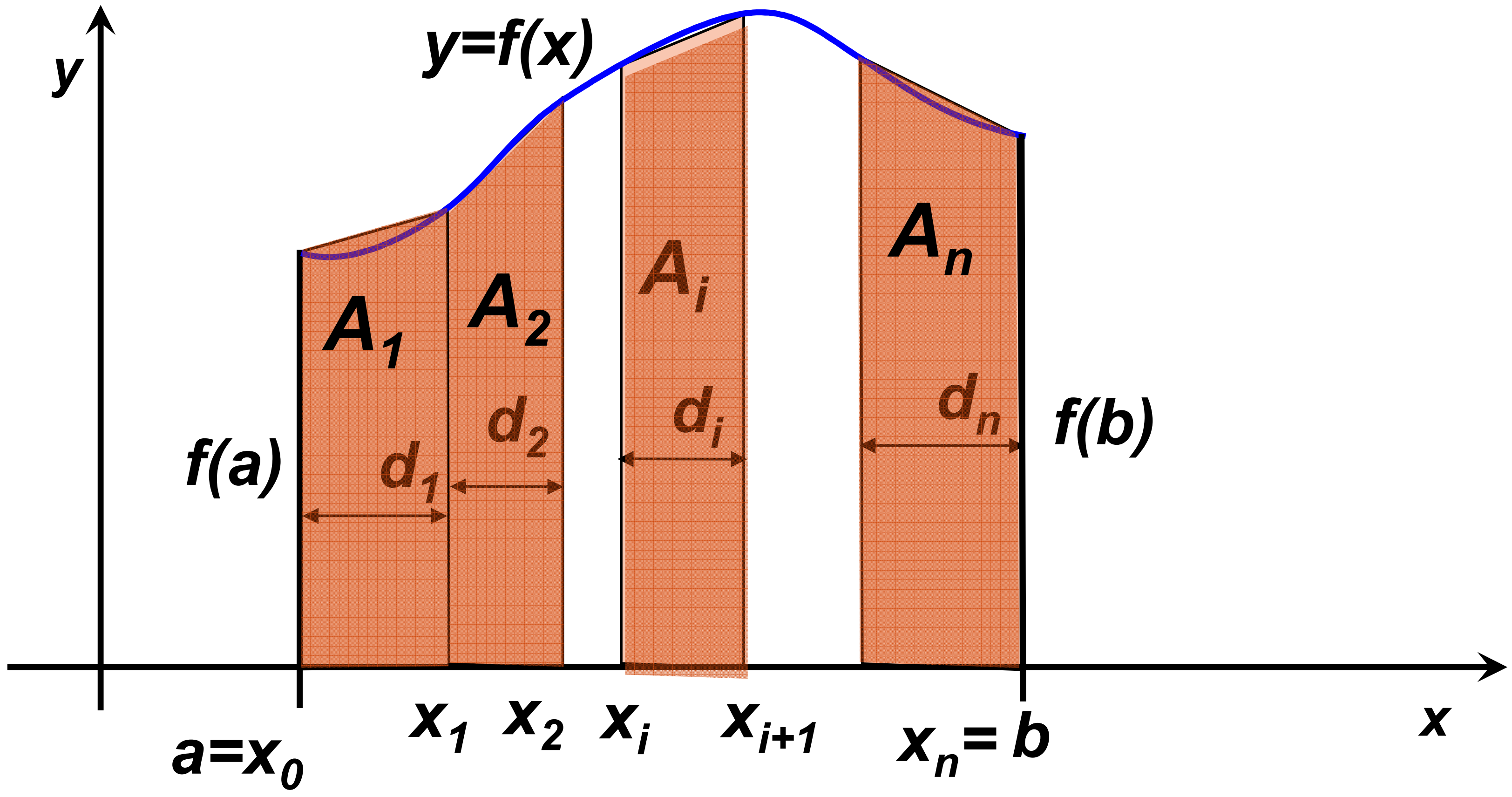
The error now has the following expression:

$$E_{Tc} = -\frac{(b-a)^3}{12n^2} f''(\eta) \quad \eta \in [a,b]$$

From this expression it follows that:

- when n grows the error diminishes
- it is possible to calculate a number of subdivisions for [a,b], in order to get a predetermined accuracy
- small errors on the input values have a moderate effect on the final result .

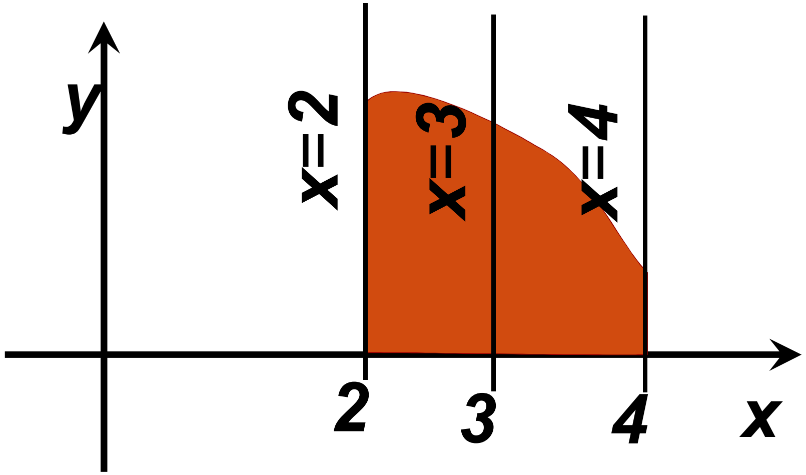
Composite trapezoidal rule: Geometrical interpretation



The shaded area is approximated by the trapezoid whose heights are $(x_{j+1}- x_j)$ and parallel sides $f(x_{j+1})$ and $f(x_j)$.

Exercise on the composite trapezoidal rule

Given the previous integral $$\int_2^4 \left(\frac{1}{x}-\frac{1}{x^2}\right)dx$$



For x≥1 the integrand is positive

We apply the trapezoidal rule over two subintervals

$$(3-2)\frac{f(2)+f(3)}{2} + (4-3)\frac{f(3)+f(4)}{2} =$$

$$=\frac{\left(\frac{1}{2}-\frac{1}{4}\right)+\left(\frac{1}{3}-\frac{1}{9}\right)+\left(\frac{1}{3}-\frac{1}{9}\right)+\left(\frac{1}{4}-\frac{1}{16}\right)}{2} = \frac{127}{288} \cong 0.441$$