

# 1 Taylor

Let  $f \in C^n[a, b]$  and assume  $f^{(n+1)}$  exists in  $(a, b)$ . Then for any  $c, x \in [a, b]$  there is some  $\zeta$  between  $c$  and  $x$  s.t.

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(c)(x-c)^k}{k!} + E_n(x) \quad (1)$$

where

$$E_n(x) = \frac{f^{(n+1)}(\zeta)(x-c)^{n+1}}{(n+1)!}$$

Equation (1) is called the Taylor expansion of  $f$  around  $c$ .

**Observation.** The famous *mean value theorem* is simply the case  $n = 0$  of Taylor's expansion: if  $f \in C[a, b]$  and  $f'$  exists on  $(a, b)$ , then for  $x, c \in [a, b]$

$$f(x) = f(c) + f'(\zeta)(x-c)$$

where  $\zeta$  is between  $c$  and  $x$ . Take  $x = b, c = a$  and the theorem appears:

$$f(b) - f(a) = f'(\zeta)(b-a)$$

We typically extend the Taylor approximation of  $f$  around a point  $r$ , where  $r = x + h$  is an approximation some value of interest  $x$ . This is useful because said approximation gives

$$f(r) = f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \dots + \frac{f^{(n)}(x)}{n!}h^n + E_n(h)$$

In other words, this strategy allows us to extend  $f(r)$  in terms of  $x$  and  $h$ , the approximation and its error. Usually,  $r, h$  are unknown but  $h$  can be bounded.

## 2 Alg. de Horner: Polynomial evaluation

Consider

$$p(x) = \sum_{i=0}^n a_i x^i$$

We wish to compute  $p(k)$  for a given  $k \in \mathbb{R}$  minimizing the number of operations. Directly computing  $a_0 + a_1 k + \dots$  leads to  $n$  sums. The  $i$ th term requires computing  $k^i$ , which means  $i$  product operations, for a total of  $\sum_{i=1}^n i = \frac{n(n+1)}{2}$  products. The total number of operations is then

$$\Theta = n + n(n+1)/2$$

The associated complexity is  $O(n^2)$ .

Horner's method consists of re-writing  $p(x)$  so that the number of products is reduced. One writes

$$p(x) = a_0 + x b_0$$

where  $b_{n-1} = a_n$  and for  $0 \leq i < n-1$ :

$$b_{i-1} = a_i + x b_i$$

Let  $p(x) = 3 + 5x - 4x^2 + 0x^3 + 6x^4$ , giving  $n = 4$ . Then  $b_3 = 6$  and

$$\begin{aligned} b_2 &= a_3 + x b_3 = 6x, & b_1 &= a_2 + x b_2 = -4 + x(6x), \\ b_0 &= a_1 + x b_1 = 5 + x(-4 + x(6x)) \end{aligned}$$

This finally gives

$$p(x) = 3 + x b_0 = 3 + x(5 + x(-4 + x(6x)))$$

Here, one must perform  $n$  sums again but only  $n$  products. Thus, there are  $\Theta = n + n = 2n$  operations, giving a complexity of  $O(n)$  (in the operation space). See the algorithm below:

```

input  $n; a_i, i = 0, \dots, n; x$ 
 $b_{n-1} \leftarrow a_n$ 
for  $i = n - 2$  to  $i = 0$ 
     $b_i = a_{i+1} + x * b_{i+1}$ 
od
 $y \leftarrow a_0 + x * b_0$ 
return  $y$ 

```

It is easy to see in this code that the **for** loop performs  $n - 1$  iterations, in each of which a single sum and a single product are computed. The  $n$ th sum and  $n$ th product are performed in the computation of  $y$ , the final result.

A more polished version includes the last computation (the one in the assignment of  $y$ ) within the loop and makes no use of indexes:

```

input  $n; a_i, i = 0, \dots, n; x$ 
 $b \leftarrow a_n$ 
for  $i = n - 2$  to  $i = -1$ 
     $b = a_{i+1} + x * b$ 
od
return  $b$ 

```

In Python,

```

def horner(coefs, x):
    n = len(coefs)-1
    b = coefs[n]

    for i in reversed(range(-1, n-1)):
        b = coefs[i+1] + x*b

    return b

```

It is trivial to adapt the code so that it returns the coefficients  $b_0, \dots, b_{n-1}$  and not the final result, if needed.

### 3 Error

Let  $r, \bar{r}$  be two real numbers s.t. the latter is an approximation of the first. We define the **error** of the approximation to be  $r - \bar{r}$ , and

$$\Delta r = |r - \bar{r}|, \quad \delta r = \frac{\Delta r}{|r|}$$

With  $r$  unknown the strategy is to work with a known bound of  $r$ .

## 4 Non-linear equations

The general problem is to find members of the set  $\mathcal{R}_f$  of roots of  $f \in \mathbb{R} \rightarrow \mathbb{R}$ . The numerical strategy is to iteratively approximate some  $r \in \mathcal{R}_f$  until some pre-established threshold in the error of approximation is met.

More formally, the numerical strategy produces a sequence  $\{x_k\}_{k \in \mathbb{N}}$  which satisfies

- $\lim_{k \rightarrow \infty} \{x_k\} = r$  for some  $r \in \mathcal{R}_f$
- Either  $e(x_k) < e(x_{k-1})$  or, more strongly,  $\lim_{k \rightarrow \infty} e(x_k) = 0$ , where  $e(x_k)$  is some appropriate measure of the error of approximation.

### 4.1 Bisection

A very simple procedure: if a root exists in  $[a, b]$ , it iteratively shrinks  $[a, b]$  in halves (keeping the halves which contain the root) until the interval is of sufficiently small length or the root is found.

**Theorem 1** (Intermediate value). If  $f$  is continuous in  $[a, b]$  and  $f(a)f(b) < 0$ , then  $\exists r \in \mathcal{R}_f$  s.t.  $r \in [a, b]$ .

Assume  $f$  is continuous. A root exists in  $[a, b]$  if  $f(a)f(b) < 0$  (**Theorem 1**). If that is the case, the midpoint  $(a + b)/2$  is taken as the approximation  $x_0$ . It is also trivial to observe that  $x_0$  is *at most* at a distance of  $(b - a)/2$  from the real root, so  $e_0 = |x_0 - r| \leq (b - a)/2$ .

If  $f(x_0) = 0$  the procedure must end because a root was found. Otherwise, suffices to find which half of the interval contains a root computing  $f(a)f(c)$  and, if needed,  $f(c)f(b)$ .

The iterations may stop after reaching a maximum number of steps, when  $|f(c)|$  is sufficiently close to zero, or when the error bound  $|e_k| \leq (b_k - a_k)/2$  (where  $[a_k, b_k]$  is the interval of this iteration) is sufficiently small.

(!) The algorithm not always converges. Take  $f(x) = 1/x$ . Clearly, it has no root. Yet setting  $a = -1, b = 1$  in the initial iteration falsely passes the test. (The problem obviously is that  $f$  is not continuous in  $[-1, 1]$ .) If one sets

**Input :**  $a, b, \delta, M, f$

**Output :** Tupla de la forma:  $(r, \text{cota de error})$

$f_a \leftarrow f(a)$

$f_b \leftarrow f(b)$

**if**  $f_a * f_b > 0$

**return** ?

**fi**

**for**  $i = 1$  **to**  $i = M$  **do**

$c \leftarrow a + (b - a)/2$

$f_c \leftarrow f(c)$

**if**  $f_c = 0$  **then**

**return**  $(c, 0)$

**fi**

$\epsilon = \frac{b - a}{2}$

**if**  $\epsilon < \delta$  **then**

**break**

**fi**

**if**  $f_a * f_c < 0$  **then**

$b \leftarrow c$

$f_b = f(b)$

**else**

$a \leftarrow c$

$f_a = f(a)$

**fi**

**od**

**return**  $(c, \epsilon)$

```

def bisection(f : callable, a : float, b : float, delta : float, M : int):

    s, e = f(a), f(b) # function values at (s)tart, (e)nd of interval

    if s*e > 0:
        raise ValueError("Interval [a, b] contains no root.")

    for i in range(M):

        c = a + (b-a)/2
        m = f(c) # value of f at (m)idpoint

        if m == 0:
            return c, 0

        e = (b-a)/2
        if e < delta:
            return c, e

        if s*m < 0:
            b = c
            e = f(b)
        else:
            a = c
            s = f(a)

    return c, e

```

**Theorem 2.** If  $\{[a_i, b_i]\}_{i=0}^{\infty}$  are the intervals generated by the bisection method on iterations  $i = 0, 1, \dots$ , then:

1.  $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n$  is a member of  $\mathcal{R}_f$ .
2. If  $c_n = \frac{1}{2}(a_n + b_n)$ ,  $r = \lim_{n \rightarrow \infty} c_n$ , then  $|r - c_n| \leq \frac{1}{2^{n+1}}(b_0 - a_0)$

**Proof. (1)** It is clear that  $a_i \leq a_{i+1}$  and  $b_i \geq b_{i+1}$ , since the interval on each iteration shrinks in one direction.

$\therefore a_n, b_n$  are monotonous.

But clearly  $a_n$  is bounded by  $b_0$  and  $b_n$  is bounded by  $a_0$ .

$\therefore a_n, b_n$  are monotonous and bounded.

$\therefore$  Their limits exist.

It is also clear that the interval shrinks to half its size on each iteration:

$$b_n - a_n = \frac{1}{2}(b_{n-1} - a_{n-1}), \quad n \geq 1 \quad (1)$$

By recurrence on (1),

$$b_n - a_n = \frac{1}{2^n}(b_0 - a_0), \quad n \geq 0 \quad (2)$$

Then

$$\lim_{n \rightarrow \infty} a_n - \lim_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} (a_n - b_n) = \lim_{n \rightarrow \infty} \frac{1}{2^n}(b_0 - a_0) = 0 \quad (3)$$

$\therefore \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n$ .

Since the limit of  $a_n, b_n$  exists and  $f$  is by assumption continuous, the composition limit theorem applies and:

$$\begin{aligned} & \lim_{n \rightarrow \infty} (f(a_n) \cdot f(b_n)) \\ &= \lim_{n \rightarrow \infty} f(a_n) \cdot \lim_{n \rightarrow \infty} f(b_n) \quad \{\text{Product of limits}\} \\ &= f\left(\lim_{n \rightarrow \infty} a_n\right) \cdot f\left(\lim_{n \rightarrow \infty} b_n\right) \quad \{\text{Composition limit theorem}\} \\ &= [f(r)]^2 \quad \left\{r = \lim_{n \rightarrow \infty} a_n\right\} \end{aligned} \quad (4)$$

The invariant of the algorithm is  $f(a_n)f(b_n) < 0$ . But due to the last result,

$$\lim_{n \rightarrow \infty} f(a_n)f(b_n) \leq 0 \iff [f(r)]^2 \leq 0 \iff f(r) = 0$$

$\therefore r = \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n$  is a root.



(2) Follows directly from result (2)

$$\begin{aligned} |r - c_n| &= \left| r - \frac{1}{2}(b_n - a_n) \right| \\ &\leq \left| \frac{1}{2}(b_n - a_n) \right| \\ &= \left| \frac{1}{2^{n+1}}(b_0 - a_0) \right| \end{aligned} \quad \{\text{Result (2)}\}$$

## 4.2 Newton's method

Assume  $r \in \mathcal{R}_f$  and  $r = x + h$ , with  $x$  an approximation of  $r$  and  $h$  its error. Assume  $f''$  exists and is continuous in some  $I$  around  $x$  s.t.  $r \in I$ . What we explained on Taylor expansions around a point gives:

$$0 = f(r) = f(x + h) = f(x) + f'(x)h + O(h^2)$$

If  $x$  is sufficiently close to  $r$ ,  $h$  is small and  $h^2$  even smaller, so that  $O(h^2)$  is unconsiderable:

$$0 \approx f(x) + hf'(x)$$

Therefore,

$$h \approx -\frac{f(x)}{f'(x)} \tag{1}$$

From this follows that  $r = x + h$  is approximated by

$$r \approx x - \frac{f(x)}{f'(x)}$$

Since the approximation in (5) truncated the terms of  $O(h^2)$  complexity, this new approximation is closer to  $r$  than  $x$  originally was. In other words,  $x - f(x)/f'(x)$  is a better approximation to  $r$  than  $x$  itself.

Thus, if  $x_0$  is an original approximation, we can define

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2}$$

to produce a sequence of approximations. This is the fundamental idea of Newton's method.

```

Input:  $x_0, M, \delta, \epsilon;$ 
 $v \leftarrow f(x_0)$ 
if  $|v| < \epsilon$  then return  $x_0$  fi
for  $k = 1$  to  $k = M$  do
     $x_1 \leftarrow x_0 - \frac{v}{f'(x_0)}$ 
     $v \leftarrow f(x_1)$ 
    if  $|x_1 - x_0| < \delta \vee v < \epsilon$  then
        return  $x_1$ 
    fi
     $x_0 \leftarrow x_1$ 
od
return  $x_0$ 

```

The predicate  $|x_1 - x_0| < \delta$  checks whether our algorithm is adjusting  $x$  in a negligible degree. If that is the case, we should stop.

**Theorem 3.** If  $f''$  continuous around  $r \in \mathcal{R}_f$  and  $f'(r) \neq 0$ , then there is some  $\delta > 0$  s.t. if  $|r - x_0| \leq \delta$ , then:

- $|r - x_n| \leq \delta$  for all  $n \geq 1$ .
- $\{x_n\}$  converges to  $r$
- The convergence is quadratic, i.e. there is a constant  $c(\delta)$  and a natural  $N$  s.t.  $|r - x_{n+1}| \leq c |r - x_n|^2$  for all  $n \geq N$ .

**Proof.** Let  $e_n = r - x_n$  be the error in the  $n$ th approximation. Assume  $f''$  is continuous and  $f(r) = 0, f'(r) \neq 0$ . Then

$$\begin{aligned}
 e_{n+1} &= r - x_{n+1} \\
 &= r - \left( x_n - \frac{f(x_n)}{f'(x_n)} \right) \\
 &= r - x_n + \frac{f(x_n)}{f'(x_n)} \\
 &= \frac{e_n f'(x_n) + f(x_n)}{f'(x_n)}
 \end{aligned} \tag{3}$$

Thus, the error at any given iteration is a function of the error at the previous iteration. Now consider the expansion of  $f(r)$  as

$$f(r) = f(x_n - e_n) = f(x_n) + e_n f'(x_n) + \frac{e_n^2 f''(\zeta_n)}{2} \quad (4)$$

for  $\zeta_n$  between  $x_n$  and  $r$ . This equation gives

$$e_n f'(x_n) + f(x_n) = -\frac{1}{2} f''(\zeta_n) e_n^2 \quad (5)$$

The expression in (5) is the numerator in (3), whereby we obtain via substitution:

$$e_{n+1} = -\frac{1}{2} \frac{f''(\zeta_n) e_n^2}{f'(x_n)} \quad (6)$$

Equation (6) ensures that the error scales quadratically. Now we wish to bound the error expression in (6). To bound  $e_{n+1}$ , we take  $\delta > 0$  to define a neighbourhood of length  $\delta$  around  $r$ . For any  $x$  in this neighbourhood, (6) reaches its maximum when the numerator is maximized and the denominator is minimized:

$$c(\delta) = \frac{1}{2} \frac{\max_{|x-r| \leq \delta} |f''(x)|}{\min_{|x-r| \leq \delta} |f'(x)|}$$

In other words,  $c(\delta)$  is the maximum value which  $e_{n+1}$  can take if  $\zeta_n, x_n$  are assumed to belong to the neighbourhood. Now we make two assumptions:

1.  $x_0$  belongs to the neighbourhood, i.e.  $|x_0 - r| \leq \delta$
2.  $\delta$  is sufficiently small so that  $\varrho := \delta c(\delta) < 1$ .

Note that, since  $\zeta_0$  is between  $x_0$  and  $r$ , assumption (1) ensures that  $\zeta_0$  is also in the neighbourhood, i.e.  $|r - \zeta_0| \leq \delta$ . Then we have:

$$|e_0| = \frac{1}{2} |f''(\zeta_0)/f'(x_0)| \leq c(\delta)$$

Then:

$$\begin{aligned} |x_1 - r| &= |e_1| \\ &= \left| e_0^2 \cdot \frac{1}{2} f''(\zeta_0)/f'(x_0) \right| \\ &\leq |e_0^2| c(\delta) && \left\{ \frac{1}{2} f''(\zeta_0)/f'(x_0) \leq c(\delta) \right\} \\ &\leq |e_0| \delta c(\delta) && \{|e_0| \leq \delta\} \\ &= |e_0| \varrho && \{\varrho = \delta c(\delta)\} \\ &< |e_0| && \{\varrho < 1\} \\ &\leq \delta \end{aligned}$$

$\therefore |e_1| < |e_0| \leq \delta$ , which means the error decreases. This argument may be repeated inductively, giving:

$$\begin{aligned}
|e_1| &\leq \varrho |e_0| \\
|e_2| &\leq \varrho |e_1| \leq \varrho^2 |e_0| \\
|e_3| &\leq \varrho |e_2| \leq \varrho^3 |e_0| \\
&\vdots
\end{aligned}$$

In general,  $|e_n| \leq \varrho^n |e_0|$ . And since  $0 \leq \varrho < 1$ , we have  $\varrho^n \rightarrow 0$  when  $n \rightarrow \infty$ , entailing that  $|e_n| \rightarrow 0$  when  $n \rightarrow \infty$ .

**Theorem 4.** If  $f''$  is continuous in  $\mathbb{R}$ , and if  $f$  is increasing, convex, and has a root, then said root is unique and Newton's method converges to it from any starting point.

Recall that  $f$  is convex if  $f''(x) > 0$  for all  $x$ . Graphically, it is convex if the line connecting two arbitrary points of  $f$  lies above the curve of  $f$  between those two points.

### 4.3 Secant method

In Newton's method,

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

The function of interest is  $f$ . We cannot escape computing  $f(x_n)$ , but it would be desirable to avoid the computation of  $f'(x_n)$ , which may potentially be expensive. Since

$$f'(x) = \lim_{h \rightarrow x} \frac{f(x) - f(h)}{x - h}$$

it is natural to suggest

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \tag{1}$$

Graphically, this means we are not using the line tangent to the point  $(x_n, f(x_n))$  but the line secant to the points  $(x_n, f(x_n))$  and  $(x_{n-1}, f(x_{n-1}))$ . The point  $x_{n+1}$  is then the value of  $x$  where this secant line has a root.

## 4.4 Fixed point iteration

The key observation is this: if  $r \in \mathcal{R}_f$ , then  $g(x) = x - kf(x)$  has  $r$  as fixed point, for any  $k \in \mathbb{R}$ . Inversely, if  $g$  has a fixed point in  $r$ , then  $r \in \mathcal{R}_f$ .

**Theorem 5.** (1) Let  $g \in C[a, b]$  and assume  $g(x) \in [a, b]$  for all  $x \in [a, b]$ . Then there is a fixed point of  $g$  in  $[a, b]$ .

(2) If, on top of previous conditions,  $g$  is differentiable in  $(a, b)$  and there is some  $k < 1$  s.t.  $|g'(x)| \leq k$  for all  $x \in (a, b)$ , then the fixed point referred in (1) is unique.

**Theorem 6** (Mean value theorem). Let  $f : [a, b] \rightarrow \mathbb{R}$  continuous and differentiable on  $(a, b)$  with  $a < b$ . Then there is some  $c \in (a, b)$  s.t.

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

The interpretation is simple: consider the line secant to  $f$  on  $a, b$ . The theorem ensures that there is some point  $c$  s.t. the line tangent to  $c$  is parallel to said secant (equal slopes).

**Proof.** (1) If  $a$  or  $b$  are fixed points the proof is done so assume otherwise. Since  $g(x) \in [a, b]$ , we have  $g(a) > a$  and  $g(b) < b$ .

Take  $\varphi(x) = g(x) - x$ , which is continuous and defined in  $[a, b]$ . Then

$$\varphi(a) = g(a) - a > 0, \quad \varphi(b) = g(b) - b < 0$$

Then  $\varphi(a)\varphi(b) < 0$ . Then, by the intermediate value theorem,  $\varphi$  has a root in  $(a, b)$ . In other words, there is at least one  $p$  s.t.

$$\varphi(p) = g(p) - p = 0$$

$\therefore g(p) = p$  is a fixed point of  $g$ .

(2) Assume two distinct fixed points  $p, q$  exist in  $[a, b]$ . The mean value theorem ensures the existence of some  $\zeta$  between  $p, q$  (and thus in  $[a, b]$ ) s.t.

$$g'(\zeta) = \frac{g(a) - g(b)}{a - b} \iff g'(\zeta)(a - b) = g(a) - g(b) \quad (1)$$

By hypothesis,  $|g'(x)| \leq k < 1$ . Since  $p, q$  are assumed to be fixed points, equation (1) gives:

$$\begin{aligned} |p - q| &= |g(p) - g(q)| \\ &= |g'(\zeta)| |p - q| \\ &\leq k |p - q| < |p - q| \end{aligned}$$

But this is absurd. The contradiction arises from assuming  $p, q$  to be distinct. Therefore, the fixed point is unique.

The fixed point algorithm begins with an approximation  $p_0$ . Then,

$$p_n = g(p_{n-1})$$

If  $g$  is continuous and the sequence converges, then it converges to a fixed point, since:

$$p := \lim_{n \rightarrow \infty} p_n = \lim_{n \rightarrow \infty} g(p_{n-1}) = g\left(\lim_{n \rightarrow \infty} p_{n-1}\right) = g(p)$$

```

Input:  $p, M, \delta$ 
 $p_{\text{previous}} = p$ 
for  $i = 1$  to  $i = M$  do
     $p \leftarrow g(p)$ 
    if  $|p - p_{\text{previous}}| < \delta$  then
        return  $p$ 
    fi
     $p_{\text{previous}} = p$ 
od
return  $p$ 

```

**Theorem 7.** Let  $g \in C[a, b]$  be a self-map of  $[a, b]$  differentiable in  $(a, b)$ . Assume there is a constant  $0 < k < 1$  s.t.  $|g'(x)| \leq k$  for all  $x \in (a, b)$ .

For all  $p_0 \in [a, b]$ , the sequence  $p_n = g(p_{n-1})$  converges to the unique fixed point  $p$  in  $(a, b)$ .

**Proof.** The mean value theorem ensures that

$$\begin{aligned}
 |p_n - p| &= |g(p_{n-1}) - g(p)| \\
 &= |g'(\zeta_n)| |p_{n-1} - p| \\
 &\leq k |p_{n-1} - p|
 \end{aligned}$$

with  $\zeta_n \in (a, b)$ . More succinctly, with  $e_n := p_n - p$ ,

$$|e_n| \leq k |e_{n-1}| \leq k |e_{n-2}| \leq \dots \leq k |e_0|$$

By recurrence,

$$|e_n| \leq k^n |e_0|$$

Since  $0 < k < 1$ ,  $k^n \rightarrow 0$  when  $n \rightarrow \infty$ , which entails  $|e_n| \rightarrow 0$  when  $n \rightarrow \infty$ . It follows that  $\{p_n\} \rightarrow p$  when  $n \rightarrow \infty$ .

Now let us consider the error of this method. Take  $p_n = p + e_n$  and consider the Taylor expansion of  $g$  around  $p$  evaluated at  $p_n = p + e_n$ :

$$g(p_n) = g(p + e_n) = \sum_{i=1}^{m-1} \frac{g^{(i)}(p)}{i!} e_n^i + \frac{f^{(m)}(\zeta_n)}{(n+1)!} e_n^m \quad (2)$$

See that in (2),  $n$  corresponds to the iteration we are dealing with, and thus  $\zeta_n$  and  $e_n$  depend on it. On the contrary,  $m$  is the degree to which we expand the series of  $g$  around  $p$  evaluated at  $p_n$ . We also assume that  $\zeta_n$  lies between  $p_n$  and  $p$ .

By definition,  $g(p_n) = p_{n+1}$  so (2) is nothing but an expression for this value. Assume  $g^{(k)}(p) = 0$  for  $k = 1, 2, \dots, m-1$ , but  $g^{(m)}(p) \neq 0$ . Then

$$\begin{aligned} e_{n+1} &= p_{n+1} - p \\ &= g(p_n) - g(p) \\ &= \frac{g^{(m)}(\zeta_n)}{m!} e_n^m \end{aligned}$$

More succintly,

$$e_{n+1} = \frac{g^{(m)}(\zeta_n)}{m!} e_n^m$$

Then

$$\lim_{n \rightarrow \infty} \left| \frac{e_{n+1}}{e_n^m} \right| = \frac{|g^{(m)}(p)|}{m!}$$

which is a constant. In conclusion, if the derivatives of  $g$  are null in  $p$  up to the order  $m-1$ , the method has an order of convergence of at least  $m$ . Three results follow from this fact.



## 4.5 Exercises

(1) Let  $f(x) = (x+2)(x+1)^2x(x-1)^3(x-2)$ . To which root does the bisection method converge on the following intervals?

$$[-1.5, 2.5], \quad [-0.5, 2.4], \quad [-0.5, 3], \quad [-3, -0.5]$$

(a) The midpoint of  $I_0 = [-1.5, 2.5]$  is  $c_0 := (2.5 - 1.5)/2 = 1/2$ . Since  $f(a)f(c) < 0$ , we have  $I_1 = [-1.5, 0.5]$ . The midpoint of  $I_1$  is  $c_1 = -0.5$ , so  $I_2$  will be  $[-0.5, 0.5]$ . The only root in this interval is  $r = 0$ , so the algorithm converges to it.

(b) The midpoint of  $I_0 = [-0.5, 2.4]$  is  $c := (2.4 - 0.5)/2 = 0.95$ . Then  $I_1 = [-1.5, 0.95]$ . Same logic gives  $c_1 = -0.725$  and then  $I_2 = [-0.725, 0.95]$ . The only root here is zero again.

(c, d) Same.

(2) We wish to find a root of  $f$  in  $[a, b]$  using bisection method and ensuring that the error is not greater than  $\epsilon \in \mathbb{R}^+$ .

(a) Estimate the number of iterations sufficient to meet the criterion.

(b) What is the number of iterations for  $a = 0, b = 1, \epsilon = 10^{-5}$ ?

Let  $e_n = x_n - r$ . It is trivial to note that  $|e_n| \leq \frac{b-a}{2^n}$ . Furthermore, the length of  $I_1$  is half the length of  $I_0$ , that of  $I_2$  is half that of  $I_1$ , etc. In other words,

$$|e_0| \leq \frac{b-a}{2}, \quad |e_1| \leq \frac{b-a}{2^2}, \quad |e_2| \leq \frac{b-a}{2^3}, \dots$$

In general,

$$|e_n| \leq \frac{b-a}{2^{n+1}}$$

Imposing

$$|e_n| \leq \frac{b-a}{2^{n+1}} \leq \epsilon$$

we satisfy our criterion, but we wish to express this bound in terms of  $n$ . Now, clearly,

$$\begin{aligned} \frac{b-a}{2^{n+1}} &\leq \epsilon \\ \iff \frac{b-a}{\epsilon} &\leq 2^{n+1} \\ \iff \log_2 \left( \frac{b-a}{\epsilon} \right) - 1 &\leq n \\ \iff \log_2 \left( \frac{b-a}{\epsilon} \right) &\leq n \\ \iff \frac{\ln \left( \frac{b-a}{\epsilon} \right)}{\ln 2} &\leq n \end{aligned}$$

which is our final answer.

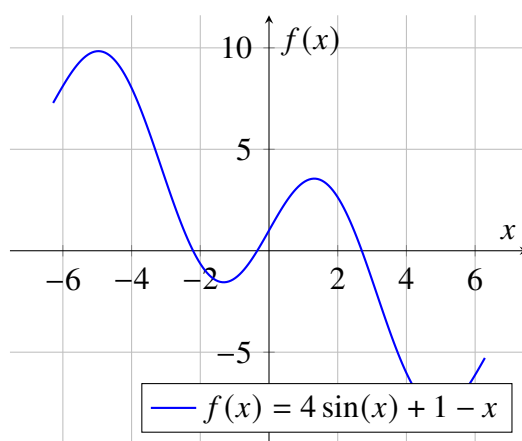
(b) For  $a = 0, b = 1, \epsilon = 10^{-5}$ , we need

$$n \geq \frac{\ln\left(\frac{1}{10^{-5}}\right)}{\ln 2} \approx 16.609$$

so  $n = 17$  would suffice.

(3) Determine graphically some root of  $f(x) = 4 \sin x + 1 - x$  and perform three iterations of the bisection method to approximate. How many steps are needed to ensure an error less than  $10^{-3}$ ?

Let us unveil the full power of LaTeX:



I'm too lazy to perform the steps of the algorithm. The number of steps needed again are given by

$$n \geq \frac{\ln\left(\frac{4-2}{10^{-3}}\right)}{\ln 2} \approx 10.96$$

so taking  $n = 11$  suffices.

(4) Let  $a > 0$ . Computing  $\sqrt{a}$  is equivalent to finding the root of  $f(x) = x^2 - a$ .

(a) Show that Newton's sequence for this case is

$$x_{n+1} = \frac{1}{2} \left( x_n + \frac{a}{x_n} \right)$$

(b) Prove that for any  $x_0 > 0$ , the approximations  $\{x_n\}$  satisfy  $x_n \geq \sqrt{a}$  for  $n \geq 1$ .

(c) Prove  $\{x_n\}$  is decreasing.

(d) Conclude that the sequence converges to  $\sqrt{a}$

(a) In Newton's algorithm,

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

Clearly,

$$f'(x) = \frac{d}{dx}(x^2 - a) = 2x$$

Therefore,

$$\begin{aligned} x_{n+1} &= x_n - \frac{x_n^2 - a}{2x_n} \\ &= x_n - \frac{1}{2} \left( x_n - \frac{a}{x_n} \right) \\ &= \frac{1}{2}x_n + \frac{1}{2} \frac{a}{x_n} \\ &= \frac{1}{2} \left( x_n + \frac{a}{x_n} \right) \quad \blacksquare \end{aligned}$$

(b) Let  $x_0 > 0$ . Recall that, among all Pythagorean means, the arithmetic mean is the greatest, assuming positively-valued vectors. In particular, it is greater or equal to the geometric mean:

$$\frac{1}{N} \sum_{i=1}^n y_i \geq \sqrt[n]{\prod_{i=1}^n y_i}$$

for any set of points  $y_1, \dots, y_n$  all positive. In particular,

$$x_{n+1} = \frac{1}{2} \left( x_n + \frac{a}{x_n} \right) \geq \sqrt{x_n \frac{a}{x_n}} = \sqrt{a} \quad \blacksquare$$

(c)

$$\begin{aligned} \frac{1}{2} \left( x_n + \frac{a}{x_n} \right) &\leq x_n \\ \iff x_n + \frac{a}{x_n} &\leq 2x_n \\ \iff \frac{a}{x_n} &\leq x_n \\ \iff a &\leq x_n^2 \\ \iff \sqrt{a} &\leq x_n \end{aligned}$$

which is true due to point (b).

(d) Let  $e_n = x_n - \sqrt{a}$ . We have shown  $\{x_n\}$  to be decreasing and bounded below by  $\sqrt{a}$ . Therefore, it converges to a limit  $L$  (with  $L$  the infimum of  $\{x_n\}$ ). Then

$$\lim_{n \rightarrow \infty} x_n = \frac{1}{2} \lim_{n \rightarrow \infty} \left( x_{n-1} + \frac{a}{x_{n-1}} \right) = \frac{1}{2} L + \frac{a}{2L}$$

This induces the equation

$$\begin{aligned} L = \frac{L}{2} + \frac{a}{2L} &\iff \frac{L}{2} = \frac{a}{2L} \\ &\iff L^2 = a \\ &\iff L = \sqrt{a} \quad \blacksquare \end{aligned}$$

**(5)** Propose an iteration formula to approximate  $\frac{1}{\sqrt{a}}$ , with  $a > 0$ , using Newton's method. Decide the number of iterations needed so that the relative error in the approximation is less than  $10^{-4}$  when starting from  $x_0 = 1$  and taking  $a = 5$ .

Error:  $e_n = r - x_n$ , quadratic, i.e.  $|r - x_{n+1}| \leq c|r - x_n|^2$ .

(a. Iteration formula) Let  $a > 0$  and assume we wish to approximate  $1/\sqrt{a}$ . Let  $\varphi = \frac{1}{a}$ , so that  $\frac{1}{\sqrt{a}} = \sqrt{\varphi}$ . We see that we can express the problem of finding the reciprocal of a root in terms of a simple root.

We know from the previous exercise that the iteration formula for  $\sqrt{\varphi}$  is

$$x_{n+1} = \frac{1}{2} \left( x_n + \frac{\varphi}{x_n} \right)$$

Now take  $x_0 = 1$  and  $a = 5$ , so that  $\varphi = \frac{1}{5}$ . The relative error of approximation on iteration  $n$  is

$$e_n = \frac{\left| x_n - \frac{1}{\sqrt{5}} \right|}{\frac{1}{\sqrt{5}}}$$

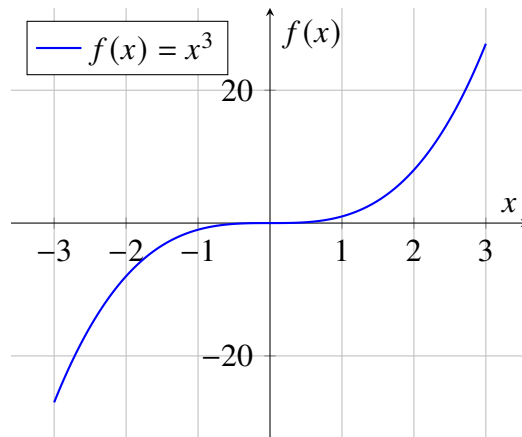
Brute-forcing allows us to see that  $x_0, x_1, x_2, x_3$  do not meet the criterion, but

$$x_4 = 0.4472137791286728 \text{ ( jaja )}$$

has  $e_4 < 10^{-4}$ .

(6) Propose an iteration formula for  $\sqrt[3]{R}$  where  $R > 0$ . Plot the function to see where the procedure converges.

Observe that finding  $\sqrt[3]{R}$  is equivalent to finding a root of  $f(x) = x^3 - R$ .



But  $f(x)$  is simply a vertical displacement of  $x^3$ , so  $\frac{d}{dx}x^3 = \frac{d}{dx}f(x)$  (which holds algebraically). In particular, the derivative of  $x^3$  approaches 0 as  $x \rightarrow 0$ , meaning that Newton's method will fail to converge for intervals of length  $L$  around 0 (with  $L$  unspecified). The graph suggests that an appropriate value for  $L$  is 1.

That said, since  $\frac{d}{dx}f(x) = \frac{d}{dx}x^3$  (in other words, since the derivative of the function is independent of  $R$ ), and  $\frac{d}{dx}x^3 = 3x^2$ , we propose

$$x_{n+1} = x_n - \frac{x_n^3}{3x_n^2} = x_n - \frac{x_n}{3} = \frac{2x_n}{3}$$



(7) (a) Utilizando el teorema del valor intermedio, demostrar que  $g(x) = \arctan(x) - \frac{2x}{1+x^2}$  tiene raíz  $\alpha \in [1, \sqrt{3}]$ .

(b) Then show that if  $\{x_n\}$  is the sequence generated by Newton's method for  $f(x) = \arctan(x)$ , with  $x_0 = \alpha$ , it is the case that  $x_n = (-1)^n \alpha$ .

(a) It is known that  $\arctan x$  is continuous in  $\mathbb{R}$ . Since  $1 + x^2 > 0$  for all  $x$ ,  $2x/(1 + x^2)$  is also continuous in  $\mathbb{R}$ .  $\therefore g$  is continuous in  $\mathbb{R}$ . And it is easy to verify as well that  $g(1)g(\sqrt{3}) < 0$ .

$\therefore$  By virtue of the intermediate value theorem, there is a root  $\alpha$  of  $g$  in  $[1, \sqrt{3}]$ .

(b) Let  $g_1(x) = \arctan x$ ,  $g_2(x) = \frac{2x}{1+x^2}$ , so that  $g = g_1 - g_2$ . Since  $\alpha > 0$ , we have  $g_1(\alpha) > 0$ ,  $g_2(\alpha) > 0$ . And since  $g(\alpha) = 0$  if and only if  $g_1(\alpha) - g_2(\alpha) = 0$ , we conclude that  $g_1(\alpha) = g_2(\alpha)$ . In other words,

$$\arctan \alpha = \frac{2\alpha}{1 + \alpha^2} \quad (1)$$

Since the derivative of  $\arctan x$  is  $1/(1 + x^2)$ , equation (1) may be expressed as follows:

$$\arctan \alpha = 2\alpha \arctan'(\alpha) \quad (2)$$

This entails that

$$\arctan' \alpha = \frac{\arctan \alpha}{2\alpha} \quad (3)$$

Now take  $x_0 = \alpha$  and consider Newton's sequence for  $f(x) = \arctan x = g_1(x)$ . Clearly,

$$\begin{aligned} x_1 &= \alpha - \frac{f(\alpha)}{f'(\alpha)} \\ &= \alpha - \arctan \alpha \times \frac{2\alpha}{\arctan \alpha} \quad \{\text{Eq. (3)}\} \\ &= \alpha - 2\alpha \\ &= -\alpha \end{aligned}$$

Same logic gives  $x_2 = \alpha$ ,  $x_3 = -\alpha$ , ... and the result should be easy to generalize.

(8) Consider for the fixed-point iteration the following functions, whose least positive root we wish to find:

$$\phi(x) = x^3 - x - 1, \quad \psi(x) = 2x - \tan x, \quad \varphi(x) = \exp(-x) - \cos x$$

Find an iteration function and an interval which guarantees the method's convergence.

( $\phi$ ) Let us analyze  $\phi$  in order to ascertain where its roots are.

Consider that  $\phi'(x) = 3x^2 - 1$ , which means  $\phi'$  has roots wherever  $3x^2 = 1$ , which holds if and only if  $x^2 = \frac{1}{3}$ , or equivalently  $x = \pm \frac{\sqrt{3}}{3}$ . Furthermore,  $\phi'(x) < 0$  in the region  $(-\sqrt{3}/3, \sqrt{3}/3)$  and  $\phi'(x) > 0$  elsewhere. In conclusion,  $\phi$  is decreasing in  $(-\sqrt{3}/3, \sqrt{3}/3)$  and increasing everywhere else.

Now, observe that  $\phi(\sqrt{3}/3) < 0$ . Combined with the fact that  $\phi$  is increasing in  $(\sqrt{3}/3, \infty)$ , this means there is a root of  $\phi$  in this interval. (Note that  $\phi$  is a polynomial without asymptotic behavior.) Furthermore,  $\phi(-\sqrt{3}/3) < 0$ . Again, this means there is no root in  $(\infty, \sqrt{3}/3)$ .

$\therefore \phi$  has one and only one root and it belongs to  $(\sqrt{3}/3, \infty)$ .

Now, suffices to note that  $f(1.3) < 0$ ,  $f(1.4) > 0$ , and the intermediate value theorem ensures that there is a root in  $(1.3, 1.4)$ .  $\therefore$  The only root of  $\phi$  lies within  $(1.3, 1.4)$ .

Now, we need only propose a function  $f$  s.t.  $r$  is a fixed-point of  $f$  and  $f(x) \in (1.3, 1.4)$  for all  $x \in (1.3, 1.4)$ . Consider that

$$\phi(x) = 0 \iff x^3 = x + 1 \iff x = \sqrt[3]{x+1} \quad (4)$$

So letting  $f(x) := \sqrt[3]{x+1}$  ensures that the fixed point of  $f$  is the root of  $\phi$ . Furthermore,  $f(1.3) \approx 1.32$ ,  $f(1.4) \approx 1.33$ . Now,

$$f'(x) = \frac{1}{\sqrt[3]{(x+1)^2}}$$

Since  $f'(x) > 0$  (as is simple to note), we know  $f$  is increasing, which means all  $f(x) \in (1.32, 1.33)$  for  $x \in [1.3, 1.4]$ . Furthermore,  $f'(x) \in (0, 1)$  and  $f'(x)$  is clearly decreasing. This means that in  $[1.3, 1.4]$ ,  $f'$  has its maximum at  $f'(1.4) \approx 0.573$ . In other words, if we let  $k = 0.573$ , we know  $|g'(x)| = f'(x) < k$  for all  $x \in [1.3, 1.4]$ .

$\therefore f$  is a self-map of  $[1.3, 1.4]$ , differentiable in  $(1.3, 1.4)$ , and there is a constant  $k \in (0, 1)$  s.t.  $|g'(x)| < k$  for all  $x \in (1.3, 1.4)$ —where incidentally this constant is  $g'(1.3)$ . ]

$\therefore$  By virtue of **Theorem 7**, the fixed-point algorithm will converge to the unique root  $r \in (1.3, 1.4)$  if using the iteration function  $f(x) = \sqrt[3]{x+1}$  and the interval  $[1.3, 1.4]$ .

( $\psi$ ) Let  $\psi(x) = 2x - \tan x$ . A root exists for  $\psi(x)$  whenever

$$x = \frac{\tan x}{2} = \frac{2 \sin x}{\cos x}$$

So we may define  $g(x) := \tan x/2$  guarantying that any fixed point of  $g$  is a root of  $\psi$ . Now,  $\tan 0 = 0$  entails that  $g(0) = 0$ . Furthermore,  $g(\pi/4) = 1/2$ . Since  $g'(x) = \sec^2(x)/2$  is strictly positive,  $g$  is strictly increasing and this means for  $x \in [0, \frac{\pi}{4}]$  we have  $g(x) \in [0, 1/2] \subseteq [0, \frac{\pi}{4}]$ .

$\therefore g$  is a self-map in  $[0, \pi/4]$ .

$\therefore$  There is a fixed-point of  $g$  in  $[0, \pi/4]$ .

Consider now  $g'(x) = \frac{1}{2} \sec^2(x) = \frac{1}{2 \cos^2 x}$ . This is clearly bounded in  $(0, 1]$ . To be more precise, it is geometrically obvious that, for all  $x \in [0, \pi/4]$ ,  $\sqrt{2}/2 \leq \cos x \leq 1$ , which means  $1/2 \leq \cos^2 x \leq 1$ . In particular,  $g'(x)$  reaches its maximum when  $\cos^2 x$  reaches its minimum, so  $g'(x)$  reaches its maximum at  $x = \frac{\pi}{4}$ :

$$g'(\pi/4) = \frac{1}{2 \cos^2 \frac{\pi}{4}} = \frac{1}{2 \cdot 1/2} = 1$$

It follows that there is some constant  $k \in (0, 1)$  such that  $|g'(x)| \leq k$  for all  $x \in (0, \pi/4)$ .

$\therefore$  There is a unique fixed point of  $g$  in  $[0, \pi/4]$ .

$\therefore$  There is a unique root of  $\psi(x)$  in  $[0, \pi/4]$  and the iteration method converges to it using this interval and the iteration function  $g$ .

( $\varphi$ ) Consider  $\varphi(x) = \exp(-x) - \cos x$ . This function is zero if and only if  $e^{-x} = \cos x$ , which may be expressed as  $x = -\ln(\cos x)$ . In other words, the roots of  $\varphi$  correspond to the fixed points of  $f(x) = -\ln(\cos x)$ .

Now,  $-1 \leq \cos x \leq 1$  but  $\ln$  is defined only in  $\mathbb{R}^+$ . From this follows that  $f$  is defined only when  $\cos x > 0$ , i.e. in the right-hand half of the unit circle. This corresponds to values of  $x$  in  $[0, \pi/2)$  or  $(3\pi/2, 2\pi]$  (extended by any factor  $2\pi k$ ,  $k \in \mathbb{Z}$ ).

Take  $I := [0, \pi/4] \subseteq \text{Dom}(f)$ . See that  $f(0) = -\ln(1) = 0$  and  $f(\pi/4) = -\ln(\sqrt{2}/2) \approx 0.346 < \pi/4$ . Furthermore, with  $u = \cos x$ ,

$$\frac{df}{dx} = -\frac{d}{du} \ln(u) \times \frac{d}{dx} \cos x = \frac{\sin x}{\cos x} = \tan x$$

which is strictly positive in  $[0, \pi/4]$ . This suffices to prove that  $f(x) \in [0, \pi/4]$  for all  $x \in [0, \pi/4]$ .

$\therefore f$  is a self-map of  $[0, \pi/4]$ .

$\therefore$  There is a fixed point of  $f$  in  $[0, \pi/4]$ .

Now,  $\tan x$  is increasing in  $[0, \pi/4]$  and, in particular,  $\tan 0 = 0$ ,  $\tan \frac{\pi}{4} = 1$ . This suffices to show that  $|g'(x)| < 1$  for all  $x \in (0, \pi/4)$ .

$\therefore$  There is a unique fixed point of  $f$  in  $[0, \pi/4]$  and the fixed point iteration algorithm converges to it when starting from said interval with  $f$  as iteration function.

**(10)** Let  $x_{n+1} = 2^{x_n-1}$  the formula used to solve  $2x = 2^x$ . What interval should be chosen to ensure  $\{x_n\}$  is convergent? Calculate its limit.

The fixed-point algorithm uses the formula  $p_n = g(p_{n-1})$  where  $g$  is a function s.t. the fixed points of  $g$  are roots of some original function of interest  $f$ . In this case, clearly  $g(x) = 2^{x-1}$ . To ensure convergence, we must find an interval  $I$  s.t.  $g$  is a self-map of  $I$  and  $g'$  lies within a unit neighbourhood of 0.

Now, clearly the equation  $2x = 2^x$  has solutions  $x = 1, x = 2$ , and no other. So whatever self-map  $I$  we build must contain either 1 or 2. So take  $I = [0, 1]$ .

Clearly, if  $x \in I$ , then  $-1 \leq x - 1 \leq 0$ . This means  $2^{x-1}$  has exponent at least  $-1$ , when  $g(0) = 2^{-1} = \frac{1}{2}$ . Furthermore,  $2^{x-1}$  has exponent at most 0, when  $g(1) = 2^0 = 1$ . This suffices to show that  $g(x) \in I$  for all  $x \in I$ .

Now,

$$\frac{d}{dx}2^{x-1} = \frac{d}{du}2^u \times \frac{d}{dx}(x-1) = 2^u \ln u$$

In short,  $g'(x) = 2^{x-1} \ln(2)$ . For  $x \in [0, 1]$ , we have already established that  $0 \leq 2^{x-1} \leq 1$ . Therefore,  $0 \leq g'(x) \leq \ln(2) < 0$  for all  $x \in [0, 1]$ . In other words,  $g'$  lies within a unit-distance of zero when its domain is restricted to  $I$ .

$\therefore$  The algorithm converges to the unique solution of  $2x = 2^x$  in  $[0, 1]$  (which is 1) when starting from said interval with iteration function  $g$ .

**(11)** Suppose  $\{x_n\}$  converges to  $r$  and that  $x_{n+1} = g(x_n)$  where  $|g(y) - g(x)| \leq \lambda|y - x|$  for all  $x, y$  with  $\lambda \in (0, 1)$ . Determine the error bound on each iteration as a function of the difference between the last two iteration values. In other words, find  $C$  s.t.

$$|x_{n+1} - r| \leq C |x_{n+1} - x_n|$$

Recall that  $x_{n+1} = g(x_n)$ . This means

$$|x_{n+1} - r| = |g(x_n) - r|$$

But  $r$  is a fixed-point of  $g$ , i.e.  $r = g(r)$ . Then

$$|g(x_n) - r| = |g(x_n) - g(r)|$$

By assumption, then,

$$\begin{aligned} |x_{n+1} - r| &= |g(x_n) - g(r)| \\ &\leq \lambda |x_n - r| \end{aligned}$$

Recall that  $|e_n| = |x_n - r| \leq k^n |e_{n-1}|$  for some  $k \in (0, 1)$ . Since the property above holds for any  $\lambda \in (0, 1)$ , it holds for said  $k$ .

Since  $|x_n - r| \leq k^n |e_{n-1}|$ , and  $k^n \in (0, 1)$  entails  $k^n |e_{n-1}| < |e_{n-1}|$ , we have  $|x_n - r| < |x_{n-1} - r|$ . In other words, successive approximations in the sequence become increasingly closer to  $r$ . This means

$$|x_{n+1} - r| \leq k |x_n - r|$$

Wtf now?

## 5 Polynomial interpolation

**Teorema fundamental del álgebra.** Every non-zero, single-variable, degree  $n$  polynomial with complex coefficients has, counted with multiplicity, exactly  $n$  complex roots.

**Theorem 8.** Given  $x_0, \dots, x_n, y_0, \dots, y_n$ , there is a unique polynomial  $p_n$  of degree  $\text{gr}(p_n) \leq n$  s.t.  $p_n(x_i) = y_i$  for all  $i$ .

**Proof.** (Existence) If  $n = 0$  simple  $p_0(x) = y_0$  which is trivial. So take as inductive hypothesis the existence of  $p_{k-1}$ , of degree  $\leq k - 1$ , s.t.  $p_{k-1}(x_i) = y_i$  for  $i = 0, \dots, k - 1$ . We will construct a polynomial  $p_k$  of degree  $\leq k$  s.t.  $p_k(x_i) = y_i$  for  $0 \leq i \leq k$ .

Consider

$$p_k(x) = p_{k-1}(x) + c(x - x_0)(x - x_1) \dots (x - x_{k-1})$$

with  $c$  yet to be determined. Its degree is  $\leq k$  and it obviously interpolates the points  $(x_0, y_0), \dots, (x_{k-1}, y_{k-1})$ . Now consider the equation:

$$p_k(x_k) = y_k$$

or equivalently

$$p_{k-1}(x_k) + c(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})$$

Solving for  $c$ , we find

$$c = \frac{y_k - p_{k-1}(x_k)}{(x_k - x_0) \dots (x_k - x_{k-1})}$$

Notice that  $c$  is well defined because each  $x_0, \dots, x_n$  is distinct and therefore the denominator is never zero. This proves the polynomial exists.

(Uniqueness) Assume two interpolating polynomials  $p_n, q_n$  exist. Let  $h = p_n - q_n$ . Clearly, its degree is  $\leq n$  and  $h(x_i) = 0$  for each  $0 \leq i \leq n$ . But this means  $h$  has  $n + 1$  real roots. Then, for the fundamental theorem of algebra,  $h(x) = 0$  for all  $x$  and then  $p_n = q_n$ .

### 5.1 Newton's form

Given  $x_0, \dots, x_n$  we define Newton's basis polynomials:

$$\eta_i(x) = \prod_{j=0}^{i-1} (x - x_j), \quad 0 \leq i \leq n$$



where  $\eta_0(x) := 1$ . Applying the construction seen in the last proof recurrently, we obtain:

$$p_k(x) = \sum_{i=0}^k c_i \eta_i(x) = \sum_{i=0}^k c_i \prod_{j=0}^{i-1} (x - x_j)$$

Here, each  $c_i$  is obtained as in the last proof.

**Example.** Consider the points  $(1, 2)$ ,  $(2, 5)$ ,  $(3, 3)$ . We begin with  $p_0(x) = 2$ , the first interpolation which interpolates  $(1, 2)$ .

Now,

$$\begin{aligned} p_1(x) &= p_0(x) + c_0(x - x_0) \\ &= 2 + c_0(x - 1) \end{aligned}$$

Now, we pose the equation:

$$\begin{aligned} p_1(x_1) &\iff y_1 \equiv p_1(2) = 5 \\ &\iff 2 + c_0(2 - 1) = 5 \\ &\iff c_0 = 3 \end{aligned}$$

Then  $p_1(x) = 2 + 3(x - 1) = 3x - 1$ . Now we repeat:

$$\begin{aligned} p_2(x) &= p_1(x) + c_1(x - x_0)(x - x_1) \\ &= (3x - 1) + c_1(x - 1)(x - 2) \end{aligned}$$

We pose the equation:

$$\begin{aligned} &(3 \cdot 3 - 1) + c_1(3 - 1)(3 - 2) = 3 \\ &\iff 8 + 2c_1 = 3 \\ &\iff 2c_1 = -5 \\ &\iff c_1 = -\frac{5}{2} \end{aligned}$$

from which we have

$$p_2(x) = (3x - 1) - \frac{5}{2}(x - 1)(x - 2)$$

In Newton's form,

$$\begin{aligned} p_2(x) &= 2 + 3(x - 1) - \frac{5}{2}(x - 1)(x - 2) \\ &= \sum_{i=0}^2 c_i \eta_i(x) \end{aligned}$$

with  $c_0 = 2, c_1 = 3, c_2 = -\frac{5}{2}$ .

## 5.2 Lagrange's form

Given  $(x_0, y_0), \dots, (x_n, y_n)$ , we define Lagrange's basic polynomials:

$$\ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{(x - x_j)}{(x_i - x_j)}, \quad i = 0, \dots, n$$

Note that  $\deg(\ell_i) = n$  for all  $i$  and  $\ell_i(x_j) = \delta_{ij}$  for all  $i, j$ . In other words,  $\ell_i(x)$  is nothing but a polynomial that becomes null at each  $x_j$  except at  $x_i$ , where it is one.

The Lagrange form of the interpolating polynomial is

$$p_n(x) = \sum_{i=0}^n y_i \ell_i(x)$$

**Prove that**  $\sum_{i=0}^n \ell_i(x) = 1$ .

Observe that the polynomial  $\sum_{i=0}^n \ell_i(x)$  has roots  $x_0$

Let  $\gamma(x) = \sum_{i=0}^n \ell_i(x) - 1$ . Any root of this polynomial is a value where the sum of each  $\ell_i(x)$  is 1. Since  $\ell_i(x_j) = \delta_{ij}$ , it is the case that each  $x_j$  is a root of  $\gamma$ . Therefore,  $\gamma$  has at least  $n + 1$  roots,

but its degree is  $n$ . By the fundamental theorem of algebra,  $\gamma(x) = 0$ . But then  $\sum_{i=0}^n \ell_i(x) = 1$  necessarily. ■

It should be clear from the fact that  $\ell_i(x_j) = \delta_{ij}$  that Lagrange's form is a valid interpolation. We don't really care what its value is beyond the arguments  $x_0, \dots, x_n$ .

### 5.3 Error of interpolation

**Theorem 9** (Error of interpolation). Let  $f \in C^{n+1}[a, b]$  and  $p$  a polynomial of degree  $\leq n$  which interpolates  $f$  on  $n + 1$  distinct points within  $[a, b]$ . Then for each  $x \in [a, b]$ , there is a number  $\zeta = \zeta_x \in (a, b)$  s.t.

$$f(x) - p(x) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} \eta_{n+1}(x)$$

or equivalently

$$f(x) - p(x) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} \prod_{j=0}^n (x - x_j)$$

### 5.4 Divided differences

In Newton's form, we use  $f[x_0, \dots, x_k]$  to denote  $c_k$ . In other words, we re-write

$$\begin{aligned} p_k(x) &= \sum_{i=0}^k f[x_0, \dots, x_i] \eta_i(x) \\ &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_1)(x - x_2) + \dots \end{aligned}$$

If  $k = 0$ ,  $f[x_0] = f(x_0)$ ; if  $k = 1$ ,

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

where  $f$  is the function being interpolated.

**Theorem 10.** Given  $x_0, \dots, x_n$ ,

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}$$

This allows us to construct a table for the so called divided differences. For instance, with  $n = 3$ :

$x_0$	$f[x_0]$	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
$x_1$	$f[x_1]$	$f[x_1, x_2]$	$f[x_1, x_2, x_3]$	
$x_2$	$f[x_2]$	$f[x_2, x_3]$		
$x_3$	$f[x_3]$			

**Example.** Assume  $f(3) = 1, f(1) = -3, f(5) = 2, f(6) = 4$  where these arguments are  $x_0, \dots, x_3$ . We begin the table thus:

3	1	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
1	-3	$f[x_1, x_2]$	$f[x_1, x_2, x_3]$	
5	2	$f[x_2, x_3]$		
6	4			

This is information sufficient to compute  $f[x_2, x_3]$ , which is

$$f[x_2, x_3] = \frac{f(x_2) - f(x_3)}{x_2 - x_3} = \frac{2 - 4}{5 - 6} = 2$$

giving

3	1	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
1	-3	$f[x_1, x_2]$	$f[x_1, x_2, x_3]$	
5	2	2		
6	4			

Same logic gives  $f[x_1, x_2] = 5/4$  and  $f[x_0, x_1] = 2$ :

3	1	2	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
1	-3	5/4	$f[x_1, x_2, x_3]$	
5	2	2		
6	4			

Now,  $f[x_0, x_1, x_2] = (f[x_1, x_2] - f[x_0, x_1])/(x_2 - x_0)$ . Using the table, this gives  $(5/4 - 2)/(2) = -3/8$ . Similarly,  $f[x_1, x_2, x_3] = (f[x_2, x_3] - f[x_1, x_2])/(x_3 - x_1) = (2 - 5/4)/(5) = 3/20$ .

$$\begin{array}{c|c|c|c|c} 3 & 1 & 2 & -3/8 & f[x_0, x_1, x_2, x_3] \\ 1 & -3 & 5/4 & 3/20 & \\ 5 & 2 & 2 & & \\ 6 & 4 & & & \end{array}$$

The last value is computed in similar fashion.

**Theorem 11** (Error of interpolation with divided differences). Let  $p$  with degree  $\leq n$  an interpolator of  $f$  on nodes  $x_0, \dots, x_n$ . If  $t \neq x_i$  for all  $i$  is a real number, then

$$f(t) - p(t) = f[x_0, \dots, x_n, t] \prod_{j=0}^n (t - x_j)$$

(★) **Observation.** Assume  $p(x)$  is of degree 1 (linear). Then  $f(x) - p(x) = \frac{f^{(2)}(\zeta_x)}{2!}(x - x_0)(x - x_1)$  for some  $\zeta_x$  in  $[x_0, x_1]$ , in accordance with the **Error of interpolation** theorem. See that  $\varphi(x) = (x - x_0)(x - x_1)$  is a quadratic expression with roots  $x_0, x_1$  and minimum at  $x_m = (x_0 + x_1)/2$ . And since  $\varphi(x_m)$  is negative,  $\varphi(x) \geq \varphi(x_m) \Rightarrow |\varphi(x)| \leq |\varphi(x_m)|$ . In consequence,

$$|\varphi(x)| \leq |(x - x_0)(x - x_1)| = \frac{|x_1 - x_0|^2}{4}$$

In consequence,

$$|f(x) - p(x)| \leq \frac{f^{(2)}(\zeta_x)}{8} |x_1 - x_0|^2$$

If we choose  $M$  the maximum of  $f^{(2)}(x)$  in  $[x_0, x_1]$ , then we have

$$|f(x) - p(x)| \leq \frac{f^{(2)}(\zeta_x)}{8} |x_1 - x_0|^2 \leq \frac{M}{8} |x_1 - x_0|^2$$

In consequence,

$$|f(x) - p(x)| \leq \frac{\max_x f^{(2)}_{|[x_0, x_1]}(x)}{8} |x_1 - x_0|^2$$

## 5.5 Hermite interpolation

Hermite interpolation consists of interpolating a function  $f$  and its derivative in certain nodes  $x_0, \dots, x_n$ . For instance, if two points are given, we wish

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i), \quad \text{for } i = 0, 1$$

See that this gives four conditions, which means it is reasonable to seek a solution in  $\Pi_3$  the space of all polynomials of degree  $\leq 3$ . (An element in  $\Pi_3$  has four coefficients.) But instead of writing  $p(x)$  in terms of the coefficients for  $1, x, x^2, x^3$ , we shall write

$$p(x) = a + b(x - x_0) + c(x - x_0)^2 + d(x - x_0)^2(x - x_1)$$

which gives

$$p'(x) = b + 2c(x - x_0) + 2d(x - x_0)(x - x_1) + d(x - x_0)^2$$

Writing the polynomial this way allows us to express the four conditions as follows:

$$\begin{aligned} a &= f(x_0) \\ b &= f'(x_0) \\ f(x_1) &= a + bh + ch^2 \\ f'(x_1) &= b + 2ch + dh^2 \end{aligned}$$

where  $h = x_1 - x_0$ . This approach readily gives  $a$  and  $b$ ,  $c$  can be determined from the third equation, and  $d$  from the fourth equation.

Now, observe that from the third equation,

$$\begin{aligned} c &= \frac{f(x_1) - a - bh}{h^2} = \frac{f(x_1) - f(x_0)}{h^2} - \frac{f'(x_0)h}{h^2} \\ &= \frac{f(x_1) - f(x_0)}{(x_1 - x_0)^2} - \frac{f'(x_0)}{(x_1 - x_0)} \end{aligned}$$

## 5.6 Splines

A spline is an interval-based polynomial approximation. We say  $S(x)$  defined on  $[x_0, x_n]$  is a spline of degree  $k$  if

1.  $S$  is polynomial of degree  $\leq k$  on each sub-interval  $[x_i, x_{i+1})$  para  $i = 0, \dots, n-1$ ;
2. The derivaitves of  $S^{(i)}$  are continuous  $[x_0, x_n]$  for  $i = 0, \dots, k-1$ .

A linear spline is a spline of the form:

$$S(x) = \begin{cases} S_2(x) = a_0x + b_0 & x \in [x_0, x_1) \\ S_1(x) = a_1x + b_1 & x \in [x_1, x_2) \\ \vdots & \\ S_{n-1}(x) = a_{n-1}x + b_{n-1} & x \in [x_{n-1}, x_n) \end{cases}$$

where each  $a_i, b_i$  is to be determined. This gives  $2n$  conditions. Clearly, for a fixed  $i$ ,

$$\begin{aligned} a_i x_i + b_i &= S_i(x_i) = f(x_i) \\ a_i x_{i+1} + b_i &= \lim_{x \rightarrow x_{i+1}} S_i(x) = S_{i+1}(x_{i+1}) = f(x_{i+1}) \end{aligned}$$

Subtracting the first equation in the second one,

$$a_i = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}, \quad b_i = f(x_i) - a_i x_i$$

The error of approximation in a linear spline can be determined if we assume each  $x_0, \dots, x_n$  to be equidistant. In other words, assume  $f$  is two times continuously differentiable in  $[a, b]$  and  $x_k = a + kh$  for  $h = (b - a)/n$  the length of each sub-interval. Then on each interval we have a degree 1 polynomial, which means the error of interpolation for each  $x \in [a, b]$  satisfies

$$|e(x)| < \frac{M}{8} h^2$$

where  $|f''(x)| \leq M$  for all  $x \in [x_0, x_n]$ . (See **Observation** marked with ★.)

### 5.6.1 Cubic splines

## 5.7 Exercises

(1) Construct the Lagrange and Newton interpolating polynomials for  $f(x) = 1/x$  taking  $x_0 = 2, x_1 = 2.5, x_2 = 4$ . Compare them and give their degrees. Graph them. Analyze the results (?).

(Newton) Newton's interpolating polynomial has the form

$$\varphi(x) = \sum_{i=0}^n a_i \eta_i(x)$$

where each  $a_i$  is to be determined and  $\eta_i = \prod_{j=0}^{i-1} (x - x_j)$ . We first do a brute construction, then a construction using divided differences.

(Newton, brute) Take  $\varphi_0(x) = \frac{1}{2}$ , a polynomial interpolating  $f$  at  $x_0$ . Now let  $\varphi_1(x) = \varphi_0(x) + c(x - x_0)$  and solve  $\varphi_1(x_1) = f(x_1)$ :

$$\begin{aligned} \frac{1}{2} + c(2.5 - 2) &= f(2.5) \\ \Leftrightarrow c &= 2 \times \left( \frac{2}{5} - \frac{1}{2} \right) \\ \Leftrightarrow c &= \frac{4}{5} - \frac{5}{5} \\ \Leftrightarrow c &= -\frac{1}{5} \end{aligned}$$

$$\therefore \varphi_1(x) = \frac{1}{2} - \frac{1}{5}(x - 2).$$

Now we let  $\varphi_2(x) = \frac{1}{2} - \frac{1}{5}(x - 2) + c(x - 2)(x - 2.5)$  and solve for  $c$  in  $\varphi_2(4) = f(4)$ :

$$\begin{aligned} \frac{1}{2} - \frac{1}{5} \times 2 + 2 \times \frac{3}{2} c &= \frac{1}{4} \\ \Leftrightarrow \frac{1}{2} - \frac{2}{5} + 3c &= \frac{1}{4} \\ \Leftrightarrow c &= \frac{1}{3} \left( \frac{10}{40} + \frac{16}{40} - \frac{20}{40} \right) \\ \Leftrightarrow c &= \frac{1}{3} \left( \frac{3}{20} \right) \\ \Leftrightarrow c &= \frac{1}{20} \end{aligned}$$



So finally we have the following polynomial in Newton's form:

$$\varphi(x) = \frac{1}{2} - \frac{1}{5}(x - 2) + \frac{1}{20}(x - 2.5)(x - 2)$$

which is of degree 2.

(Newton, divided diffs.) The table of divided differences to interpolate  $f$  on  $x_0, x_1, x_2$  is

$$\begin{array}{c|c|c|c} x_0 & f[x_0] & f[x_0, x_1] & f[x_0, x_1, x_2] \\ x_1 & f[x_1] & f[x_1, x_2] & \\ x_2 & f[x_2] & & \end{array} \Rightarrow \begin{array}{c|c|c|c} 2 & 1/2 & f[x_0, x_1] & f[x_0, x_1, x_2] \\ 2.5 & 2/5 & f[x_1, x_2] & \\ 4 & 1/4 & & \end{array}$$

Now,  $f[x_1, x_2] = (f[x_2] - f[x_1])/(x_2 - x_1) = (1/4 - 2/5)/(1.5) = -1/10$ :

$$\begin{array}{c|c|c|c} 2 & 1/2 & f[x_0, x_1] & f[x_0, x_1, x_2] \\ 2.5 & 2/5 & -1/10 & \\ 4 & 1/4 & & \end{array}$$

Now, same rule gives  $f[x_0, x_1] = -1/5$ :

$$\begin{array}{c|c|c|c} 2 & 1/2 & -1/5 & f[x_0, x_1, x_2] \\ 2.5 & 2/5 & -1/10 & \\ 4 & 1/4 & & \end{array}$$

Lastly,  $f[x_0, x_1, x_2] = (f[x_1, x_2] - f[x_0, x_1])/(x_2 - x_0) = 1/20$ :

$$\begin{array}{c|c|c|c} 2 & 1/2 & -1/5 & 1/20 \\ 2.5 & 2/5 & -1/10 & \\ 4 & 1/4 & & \end{array}$$

Then,

$$\begin{aligned} \varphi(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &= \frac{1}{2} - \frac{1}{5}(x - 2) + \frac{1}{20}(x - 2.5)(x - 2) \end{aligned}$$

which is what we had obtained before.

(Lagrange) Lagrange's polynomial has the form

$$\phi(x) = \sum_{i=0}^n y_i \ell_i(x)$$

where

$$\ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$$

Then,

$$\begin{aligned} \ell_0(x) &= \left( \frac{x - 2.5}{2 - 2.5} \right) \left( \frac{x - 4}{2 - 4} \right) \\ &= \left( -\frac{1}{2}(x - 2.5) \right) (-2(x - 4)) \\ &= (x - 2.5)(x - 4) \end{aligned}$$

$$\begin{aligned} \ell_1(x) &= \left( \frac{x - 2}{2.5 - 2} \right) \left( \frac{x - 4}{2.5 - 4} \right) \\ &= -\frac{3}{4}(x - 2.5)(x - 4) \end{aligned}$$

$$\begin{aligned} \ell_2(x) &= \left( \frac{x - 2}{4 - 2} \right) \left( \frac{x - 2.5}{4 - 2.5} \right) \\ &= 3(x - 2)(x - 4) \end{aligned}$$

Therefore,

$$\begin{aligned} \phi(x) &= \frac{1}{2}\ell_0(x) + \frac{2}{5}\ell_1(x) + \frac{1}{4}\ell_2(x) \\ &= \frac{1}{2}(x - 2.5)(x - 4) - \frac{3}{10}(x - 2.5)(x - 4) + \frac{3}{4}(x - 2)(x - 4) \end{aligned}$$

(2) Prove: If  $f$  polynomial of degree  $\leq n$  then the polynomial of degree  $\leq n$  that interpolates  $f$  at  $x_0, \dots, x_n$  is  $f$  itself.

It is a theorem that there exists a polynomial of degree  $\leq n$  that interpolates  $f$  in the points  $x_0, \dots, x_n$ , and that this polynomial is unique.  $f$  is a polynomial of degree  $\leq n$  that interpolates  $f$  at  $x_0, \dots, x_n$ . This concludes the proof.

Given  $x_0, \dots, x_n$ , prove the following properties of Lagrange's basic polynomials  $\ell_k(x)$ :

1. Their sum is 1.
2. Their linear combination with coefficients  $x_0, \dots, x_n$  is  $x$ .
3. Their linear combinations with coefficients  $x_0^m, \dots, x_n^m$ , with  $m \leq n$ , is  $x^m$ .

(1) This was already proven in a previous section.

(2) See that

$$\sum_{k=0}^n x_k \ell_k(x) = x \iff \sum_{k=0}^n x_k \ell_k(x) - x = 0 \quad (1)$$

Let  $\phi(x) = \sum x_k \ell_k(x) - x$ . Since  $\ell_k(x_j) = \delta_{kj}$ ,  $\phi(x_j) = x_j - x_j = 0$ , meaning that each  $x_j$  is root of  $\phi$ . This means  $\phi$  is a polynomial of degree  $\leq n$  with  $n+1$  roots. Then by virtue of the fundamental theorem of algebra, it is necessarily the case that  $\phi(x) = 0$ . Then the RHS of (1) holds, which concludes the proof.

(3) Assume  $m \leq n$ . See that:

$$\sum_{k=0}^n x_k^m \ell_k(x) = x^m \iff \sum_{k=0}^n x_k^m \ell_k(x) - x^m = 0 \quad (2)$$

Let  $\phi(x) = \sum x_k^m \ell_k(x) - x^m$ . The polynomial has degree  $\leq n$  due to the fact that  $m \leq n$ . Once more,  $\phi(x_j) = x_j^m - x_j^m = 0$ . Etc.

(6) Let  $f : [0, 5] \rightarrow \mathbb{R}$ ,  $f(x) = 2^x$ . Let  $P_n$  a polynomial of degree at most  $n$  that interpolates  $f$  at  $n + 1$  distinct points in  $[0, 5]$ . Prove that for any  $x$  in said interval,

$$|P(x) - f(x)| \leq \frac{32 \times 5^{n+1}}{(n+1)!}$$

Recall that for  $x \in [0, 5]$ ,

$$P(x) - f(x) = \frac{f^{(n+1)}(\zeta_x)}{(n+1)!} \eta_{n+1}(x)$$

for some  $\zeta_x \in [0, 5]$ . Now,  $\frac{d}{dx} 2^x = \ln 2 \times 2^x$ , whose derivative is  $\ln^2 2 \times 2^x$ , whose derivative is  $\ln^3 2 \times 2^x$ , etc.  $\therefore f^{(n+1)}(x) = \ln^{n+1} 2 \times 2^x = (n+1) \ln 2 \times 2^x$ .

$$\begin{aligned} \therefore P(x) - f(x) &= \frac{\ln^{n+1}(2) \times 2^{\zeta_x}}{(n+1)!} \prod_{j=0}^n (x - x_j) \\ &= \frac{\ln^{n+1}(2) \times 2^{\zeta_x}}{n!} \prod_{j=0}^n (x - x_j) \end{aligned}$$

Since  $0 < \ln(2) < 1$ , we know  $0 < \ln^{n+1}(2) < 1$ , and therefore

$$\frac{\ln^{n+1}(2) \times 2^{\zeta_x}}{(n+1)!} \prod_{j=0}^n (x - x_j) < \frac{2^{\zeta_x}}{(n+1)!} \prod_{j=0}^n (x - x_j)$$

Necessasrily,  $2^{\zeta_x} \leq 5$  and

$$\prod_{j=0}^n (5 - x_j) \leq \prod_{j=0}^n 5 = 5^{n+1}$$

From this follows that

$$P(x) - f(x) \leq \frac{2^5 \times 5^{n+1}}{(n+1)!}$$

Since the RHS is positive, taking absolute value on both sides gives us the desired result.

(7) Prove that when interpolating  $\cosh(x)$  with a polynomial  $p(x)$  of degree  $\leq 22$  in  $[-1, 1]$ , the error is  $\leq 5 \times 10^{-16}$ .

A polynomial of degree  $n = 22$  has 23 coefficients, corresponding to the need of determining 23 nodes  $x_0, \dots, x_{22}$ . So we let  $n = 23$ . It is known that  $\frac{d}{dx} \cosh x = \sinh x$ , whose derivative is once more  $\cosh x$ . So,  $\cosh^{(n+1)} = \cosh^{(24)} = \cosh$ . In consequence,

$$p(x) - \cosh(x) = \frac{\cosh(\zeta_x)}{(n+1)!} \prod_{j=0}^n (x - x_j)$$

for some  $\zeta_x \in (-1, 1)$ . The graph of  $\cosh$  is symmetric (the function is even). Therefore, it achieves its maximum (restricted to  $[-1, 1]$ ) at  $\cosh(1) = \cosh(-1) = (e^1 + e^{-1})/2 = (e^2 + 1)/2e$ . So

$$\left| \frac{\cosh(\zeta_x)}{(n+1)!} \right| \left| \prod_{j=0}^n (x - x_j) \right| \leq \frac{e^2 + 1}{2e(n+1)!} \left| \prod_{j=0}^n (x - x_j) \right|$$

Now, since  $x \in [-1, 1]$ , it is obvious that the maximum value which the factorial (and its absolute value) can take is 1. So

$$\frac{e^2 + 1}{2e(n+1)!} \left| \prod_{j=0}^n (x - x_j) \right| \leq \frac{e^2 + 1}{2e(n+1)!}$$

Now, with a calculator one can see that  $24! > 10^{16}$ . Furthermore,  $(e^2 + 1)/2e < 5$ . So,

$$\frac{e^2 + 1}{2e(n+1)!} < \frac{5}{10^{16}} = 5 \times 10^{-16} \quad \blacksquare$$

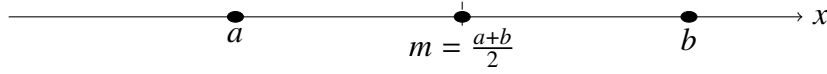
(8) (a) Let  $a < b$ ,  $m$  the midpoint between them,  $p = m - h$  and  $q = m + h$  for  $0 \leq h \leq (b - a)/2$ . Prove that for all  $x \in [a, b]$ ,

$$|(x - p)(x - q)| \leq \frac{(b - a)^2}{4}$$

(b) Let  $x_i = a + i(\frac{b-a}{n})$  for  $i = 0, \dots, n$  equidistant points in  $[a, b]$ . Prove that for all  $x \in [a, b]$ ,

$$|(x - x_0) \dots (x - x_n)| \leq \frac{(b - a)^{n+1}}{2^{n+1}}$$

(a) See the graph below for reference.



Define  $f(x)$  as the quadratic function with roots  $a, b$  and upward tails:  $f(x) = (x - a)(x - b)$ . We know that if  $x_0 = p, x_1 = q$ , then a linear interpolation of  $f$  at nodes  $x_0, x_1$  on the interval  $[a, b]$  satisfies:

$$|f(x) - p(x)| = \left| \frac{f''(\zeta_x)}{2!} \right| |(x - x_0)(x - x_1)| \leq \frac{M}{8} |x_1 - x_0|^2$$

with  $M$  the maximum of  $|f''(x)|$  in  $[a, b]$  and some  $\zeta_x \in (a, b)$ . Now,

$$|x_1 - x_0|^2 = |(m + h) - (m - h)|^2 = 4h^2$$

Therefore,

$$\begin{aligned} & \left| \frac{f''(\zeta_x)}{2} \right| |(x - p)(x - q)| \leq \frac{4h^2 \times M}{8} \\ \Rightarrow & \left| \frac{f''(\zeta_x)}{2} \right| |(x - p)(x - q)| \leq \frac{M}{2} h^2 \\ \Rightarrow & |f''(\zeta_x)| |(x - p)(x - q)| \leq M h^2 \end{aligned}$$

Here's the trick: since  $f$  is quadratic,  $f''$  is constant and therefore  $|f''(\zeta_x)| = |M|$ . So dividing by  $|M|$  on both sides we get

$$|(x - p)(x - q)| \leq h^2$$

Finally, we know  $0 \leq h \leq (b - a)/2$ , so

$$|(x - p)(x - q)| \leq h^2 \leq \left(\frac{b - a}{2}\right)^2 = \frac{(b - a)^2}{4}$$

which is what we wanted to show.

(b) Let  $x_0, \dots, x_n$  s.t.  $x_i = a + i\frac{(b-a)}{n}$  for  $0 \leq i \leq n$ . We wish to show that for all  $x \in [a, b]$ ,

$$|\eta_{n+1}(x)| \leq \frac{(b - a)^{n+1}}{2^{n+1}}$$

Take  $x_i, x_{i+1}$  and let  $f$  be the quadratic function with roots  $x_i, x_{i+1}$  and upward tails. Using the exact same reasoning of (a), we know

$$|(x - x_i)(x - x_{i+1})| \leq \frac{(x_{i+1} - x_i)^2}{2^2} \quad (3)$$

Since the points are equidistant,  $x_{i+1} - x_i = \frac{b-a}{n}$ , as is easy to prove algebraically, so equation (2) is equivalent to:

$$|(x - x_i)(x - x_{i+1})| \leq \left(\frac{b - a}{2n}\right)^2 \quad (4)$$

See that exercise (a) satisfied this formula, where had nodes  $x_0, x_1$  and therefore  $n = 1$ . In other words, exercise (a) was the base case for an inductive proof and we can now assume that the statement holds for  $n = k - 1$  for some  $k > 2$ . Our inductive case consists of having points  $x_0, \dots, x_k$ , where we wish to prove

$$|(x - x_0) \dots (x - x_{k-1})| \leq \frac{(b - a)^k}{2^k} \Rightarrow |(x - x_0) \dots (x - x_{k-1})(x - x_k)| \leq \frac{(b - a)^{k+1}}{2^{k+1}}$$

So assume



$$|\eta_k(x)| \leq \frac{(b-a)^k}{2^k}$$

Now take

$$|\eta_k(x)| |(x - x_{k+1})| \leq \frac{(b-a)^k}{2^k} |(x - x_{k+1})|$$

Since  $x \in [a, b]$  and  $x_{k+1} = b$  is the last node,  $|x - x_{k+1}| \leq b - a$  (i.e. the maximum distance that  $x$  can take from  $b$  is when  $x$  is exactly  $a$ ). So

$$|\eta_k(x)| |(x - x_{k+1})| \leq \frac{(b-a)^k}{2^k} |(x - x_{k+1})| \leq \frac{(b-a)^k}{2^k} (b-a)$$

Something's off. But should be along this lines.

(9) (a) Let  $f(x) = \cos x\pi$ . Find a polynomial of degree  $\leq 3$  that verifies:

$$p(-1) = f(-1), \quad p(0) = f(0), \quad p(1) = f(1), \quad p'(1) = f'(1)$$

(b) Find a polynomial of degree  $\leq 4$  that verifies previous conditions and the added condition  $p''(1) = f''(1)$ .

(a) Let  $x_0 = -1, x_2 = 0, x_3 = 1$ . To keep track of which nodes have double (or more) conditions, let  $z_0 = x_0, z_1 = x_1, z_2 = x_2, z_3 = x_2$  (since  $x_2$  has two conditions, one for  $p$  and one for  $p'$ .) Then, our table of divided differences is

$$\begin{array}{c|c|c|c|c} z_0 & f[z_0] & f[z_0, z_1] & f[z_0, z_1, z_2] & f[z_0, z_1, z_2, z_3] \\ z_1 & f[z_1] & f[z_1, z_2] & f[z_1, z_2, z_3] & \\ z_2 & f[z_2] & f[z_2, z_3] & & \\ z_3 & f[z_3] & & & \end{array}$$

So now we simply compute the first column.

$$\begin{array}{c|c|c|c|c} -1 & -1 & f[z_0, z_1] & f[z_0, z_1, z_2] & f[z_0, z_1, z_2, z_3] \\ 0 & 1 & f[z_1, z_2] & f[z_1, z_2, z_3] & \\ 1 & -1 & f[z_2, z_3] & & \\ 1 & -1 & & & \end{array}$$

Amazing. So now we compute  $f[z_2, z_3] = f[x_2, x_2]$ . Since this is a repeated node, by definition it corresponds to  $f'(x_2) = f'(0)$ . So, we see that  $f'(0) = -\sin(0\pi)\pi = 0$ . Similarly,

$$f[z_1, z_2] = \frac{f[z_2] - f[z_1]}{z_2 - z_1} = \frac{-2}{1} = -2$$

and

$$f[z_0, z_1] = \frac{f[z_1] - f[z_0]}{z_1 - z_0} = 2$$

So,

$$\begin{array}{c|c|c|c|c} -1 & -1 & 2 & f[z_0, z_1, z_2] & f[z_0, z_1, z_2, z_3] \\ 0 & 1 & -2 & f[z_1, z_2, z_3] & \\ 1 & -1 & 0 & & \\ 1 & -1 & & & \end{array}$$

Now,

$$f[z_1, z_2, z_3] = \frac{f[z_2, z_3] - f[z_1, z_2]}{z_3 - z_1} = \frac{2}{1} = 2$$

$$f[z_0, z_1, z_2] = \frac{f[z_1, z_2] - f[z_0, z_1]}{z_2 - z_0} = -\frac{2}{2} = -2$$

$$\begin{array}{c|c|c|c|c|c} -1 & -1 & 2 & -2 & f[z_0, z_1, z_2, z_3] \\ 0 & 1 & -2 & 2 & \\ 1 & -1 & 0 & & \\ 1 & -1 & & & \end{array}$$

At last,

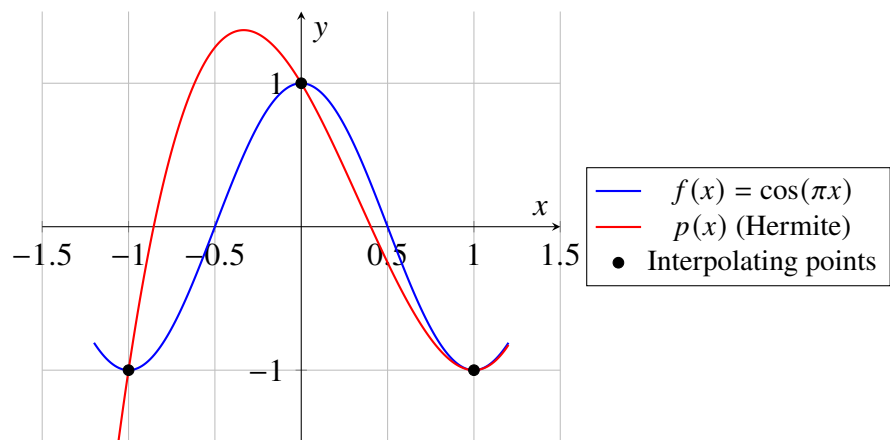
$$f[z_0, \dots, z_3] = \frac{f[z_1, z_2, z_3] - f[z_0, z_1, z_2]}{z_3 - z_0} = 2$$

The interpolating polynomial is built using Newton's form—remember that  $f[x_0], f[x_0, x_1], \dots$  etc are the coefficients  $c_0, c_1, \dots$  in Newton's form  $\sum_{i=0}^n c_i \eta_i(x)$ . So,

$$p(x) = f[z_0] + f[z_0, z_1](x - z_0) + f[z_0, z_1, z_2](x - z_0)(x - z_1) + f[z_0, \dots, z_3](x - z_0)(x - z_1)(x - z_2)$$

Simplifying,

$$p(x) = -1 + 2(x + 1) - 2(x + 1)x + 2(x + 1)x(x - 1)$$



**(10)** We wish to approximate  $f(x) = \sqrt{x}$  with an error of at most  $5 \times 10^{-8}$  using a linear spline and quadratic interpolation every three nodes.

Determine the least number of nodes  $n$  of the form  $x_i = 1 + \frac{i}{n}$ , with  $i = 0, \dots, n$ , and interval length  $h$ , so that the error bound is met.

(Linear spline) See that the desired approximation falls within  $[1, 2]$ . For a linear spline, the error of approximation obeys

$$|e(x)| < \frac{M}{8}h^2, \quad x \in [1, 2] \text{ and } M = \max |f''|$$

where we can think of  $h$  as a function of  $n$ , i.e.  $h(n) = \frac{1}{n}$  for  $n \in \mathbb{N}$ .

$$\begin{aligned} \frac{M}{8}h(n)^2 &\leq 5 \times 10^{-8} \\ \iff \frac{M}{(n)^2} &\leq 40 \times 10^{-8} \\ \iff \frac{10^8 M}{40} &\leq n^2 \\ \iff \sqrt{\frac{10^8 M}{40}} &\leq n \end{aligned}$$

Now, suffices to see that

$$f'(x) = \frac{1}{2\sqrt{x}}, \quad f''(x) = -\frac{1}{4x^{3/2}}$$

Clearly then  $|f''(x)|$  is decreasing and its maximum in  $[1, 2]$  occurs at  $x = 1$  :

$$M = |f''(1)| = \frac{1}{4}$$

So, we require

$$\begin{aligned}
& \sqrt{\frac{10^8}{40 \times 4}} \leq n \\
& \iff \frac{10^4}{\sqrt{160}} \leq n \\
& \iff \frac{10.000}{\sqrt{16 \times 10}} \leq n \\
& \iff \frac{10.000}{4\sqrt{10}} \leq n \\
& \iff \frac{2500}{\sqrt{10}} \leq n \\
& \iff 790.569 \leq n
\end{aligned}$$

So fixing  $n = 791$  suffices.

(Quadratic interpolation) Assume we group  $x_0, \dots, x_n$  into  $x_0, x_1, x_2$ , then  $x_3, x_4, x_5$ , etc. Let  $\vec{x}_i$  denote the  $i$ th grouping of three nodes. We are of course assuming that there are  $n + 1 = 3k$  nodes with  $k \in \mathbb{N}$ . The function  $h(n)$  which specifies the distance between nodes will be specified later.

Assume each  $\vec{x}_i$  is used to fit a quadratic polynomial  $q_i(x) = a_i x^2 + b_i x + c_i$ . The error of interpolation will then be specific to each interval. In other words, for any  $x$  belonging to the interval specified by  $\vec{x}_i$ ,

$$|e(x)| = \left| \frac{f^{(3)}(\zeta_x)}{3!} \prod_{\tilde{x} \in \vec{x}_i} (x - \tilde{x}) \right|$$

for some  $\zeta_x$  in the interval of interest. Taking  $M = \max |f^{(3)}(x)|$ , with  $f^{(3)}(x) = \frac{3}{8x^{\frac{5}{2}}}$ , we obtain  $M = 3/8$ . Now the question becomes whether we can bound the factorial expression. See that

$$\varphi(x) = (x - x_0)(x - x_1)(x - x_2)$$

is a cubic function with distinct roots. (We could have chosen any successive values for  $x_0, x_1, x_2$ .) Consider  $\varphi$  as restricted to the interval  $[x_0, x_2]$ . We know that it will have three roots, a maximum  $x_M$  in  $[x_0, x_1]$  and a minimum  $x_m$  at  $[x_1, x_2]$ . Since the roots are equidistant (root symmetry),  $\varphi$  is symmetric around the mid-root  $x_1$  and  $\varphi(x_m) = \varphi(x_M)$ . But where is the critical point?

(★) Let us consider a centered cubic polynomial, without loss of generality. Let  $\phi(x) = (x - a)x(x - b)$ , with mid-root zero. Assuming root symmetry,  $a = -b$ . So letting  $r = b$  we have

$$\phi(x) = (x - r)x(x + r) = x^3 - xr^2$$

Its derivative  $\phi'(x) = 3x^2 - r^2$  is zero if and only if  $x = \pm \frac{r}{\sqrt{3}}$ . We have already established that these critical points are equal in their absolute values. Now it only suffices to see that

$$\phi\left(\frac{r}{\sqrt{3}}\right) = -\frac{2r^3}{3\sqrt{3}} = -\frac{(c-a)^3}{12\sqrt{3}}$$

(because  $r = (c - a)/2$ ). Therefore,

$$\max |\phi(x)| = \frac{(c-a)^3}{12\sqrt{3}}$$

From (★) readily follows that

$$\left| \prod_{\tilde{x} \in \vec{x}_i} (x - \tilde{x}) \right| \leq \frac{h^3}{12\sqrt{3}}$$

where  $h$  is the distance between the last node in a grouping and the first (i.e.  $x_2 - x_0 = x_5 - x_3 = \dots$  etc.) In consequence,

$$\begin{aligned} |e(x)| &= \left| \frac{f^{(3)}(\zeta_x)}{3!} \prod_{\tilde{x} \in \vec{x}_i} (x - \tilde{x}) \right| \\ &\Rightarrow |e(x)| \leq 3/8 \frac{h^3}{12\sqrt{3}} \end{aligned}$$

Here,  $h$  is a function of  $n$ . If  $n = 2$  (three points), then  $h = 2/3$ . If  $n = 5$  (six points), then each sub-interval is of length  $1/6$  and  $h = 2/6$ . In general, if  $n = 3k$ , then  $h = \frac{2}{n}$ . So we have

$$|e(x)| \leq \frac{3}{8 \times \sqrt{3}} \frac{2}{n} = \frac{3}{4\sqrt{3}n}$$

So now all that is needed is to bound the RHS expression to  $5 \times 10^{-8}$ :

$$\frac{3}{4\sqrt{3}n} \leq 5 \times 10^{-8} \iff \dots$$

bla bla. This is the simplest part of the problem so I skip it.



**(10)** Let  $f(x) = \cos x$ . Determine the step-length  $h$  and minimum number of nodes  $n + 1$  needed to approximate  $f(x)$  via linear spline in  $[0, 2\pi]$  with an error less than or equal to  $5 \times 10^{-7}$ .

We know

$$|e(x)| \leq \frac{M}{8} h(n)^2$$

where  $h(n) = \frac{2\pi}{n}$  and  $M = \max |f''(x)|$ . Since  $f''(x) = -\cos x$ , its maximum in  $[0, 2\pi]$  is 1. From this follows that

$$|e(x)| \leq \frac{4\pi^2}{8n^2} = \frac{1}{2} \left( \frac{\pi}{n} \right)^2$$

and all that is left is bounding said expression to be less than or equal to  $5 \times 10^{-7}$ . If one does the math, we obtain that said condition holds iff  $n \geq \pi \times 10^3 \approx 3141.59$ , which means letting  $n = 3142$  suffices. So the number of nodes needed is  $n + 1 = 3143$ . The step length  $h$  results  $\frac{2\pi}{3142} \approx 0.001999 \approx 0.002$ .

(12) (a) Determine  $\alpha, \beta, \gamma$  such that

$$S(x) = \begin{cases} \alpha x^3 + \gamma x & 0 \leq x \leq 1 \\ -\alpha x^3 + \beta x^2 - 5\alpha x + 1 & 1 \leq x \leq 2 \end{cases}$$

is a cubic spline.

(b) With the determined values, decide if  $S$  interpolates  $f(x) = 2^x + 1/2x^2 - 1/2x - 1$  in  $[0, 2]$  with nodes  $\{0, 1, 2\}$ .

(c) Graph  $f$  and  $S$  in  $[0, 2]$ .

(a)  $S(x)$  must satisfy the interpolation constraint and the continuity constraint. Note that  $S(x)$  are polynomials with continuous first and second derivatives in  $[0, 2]$ . Furthermore,

$$S'(x) = \begin{cases} 3\alpha x^2 + \gamma & 0 \leq x \leq 1 \\ -3\alpha x^2 + 2\beta x - 5\alpha & 1 \leq x \leq 2 \end{cases}, \quad S''(x) = \begin{cases} 6\alpha x & 0 \leq x \leq 1 \\ -6\alpha x + 2\beta & 1 \leq x \leq 2 \end{cases}$$

For all  $f$  in  $\{S, S', S''\}$  we wish that

$$\lim_{x \rightarrow 1(\leftarrow)} f(x) = \lim_{x \rightarrow 1(\rightarrow)} f(x)$$

(a.a) Beginning with  $S''(x)$ , we impose

$$6\alpha = -6\alpha + 2\beta \Rightarrow 12\alpha = 2\beta \Rightarrow \beta = 6\alpha$$

(a.b) Now taking  $S'(x)$ , we impose

$$3\alpha + \gamma = -3\alpha + 2\beta - 5\alpha$$

Substituting with  $\beta = 6\alpha$ , we find

$$\gamma = -6\alpha + 2(6\alpha) - 5\alpha \Rightarrow \gamma = \alpha$$

(a.c) Now taking  $S(x)$ , we impose

$$\alpha + \gamma = -\alpha + \beta - 5\alpha + 1$$

Substituting with  $\gamma = \alpha, \beta = 6\alpha$ , we find

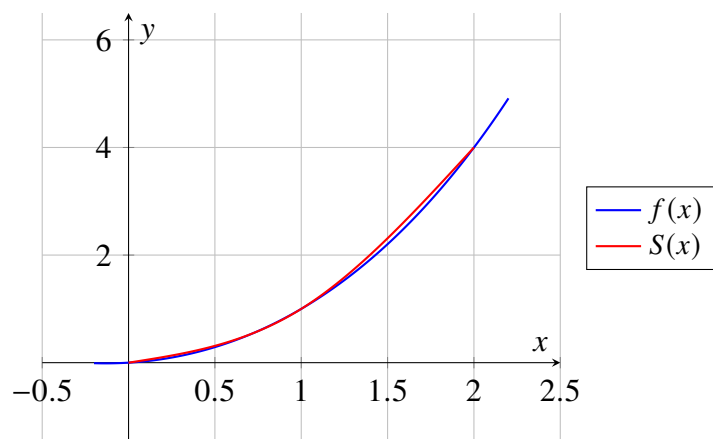
$$2\alpha = 1 \Rightarrow \alpha = \frac{1}{2}$$

So we finally obtain

$$\alpha = \frac{1}{2}, \beta = 3, \gamma = \frac{1}{2}$$

(b) It is simple to see that the spline does interpolate  $f$  simply by evaluating  $S$  and  $f$  on 0, 1, 2.

(c)



## 6 Function approximation

### 6.1 Least squares method

Assume  $(x_1, y_1), \dots, (x_m, y_m)$  are points from a function  $f$  we wish to approximate. A reasonable approach is to find a linear function with coefficients  $a_0, a_1$  that minimize

$$E = \sum_{i=1}^m (y_i - (a_1 x_i + a_0))^2$$

Doing the standard things (take partial derivative, equate it to zero, etc.) we obtain the following normal equations to minimize  $E$ :

$$a_0 = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i y_i)(\sum x_i)}{m \sum x_i^2 - (\sum x_i)^2}, \quad a_1 = \frac{m \sum x_i y_i - (\sum x_i)(\sum y_i)}{m \sum x_i^2 - (\sum x_i)^2}$$

where all sums range from  $i = 1$  to  $m$ .

The general case is when  $f$  is approximated with a polynomial of degree  $\leq n$ , with  $n < m - 1$ . Now, we need to determine the coefficients  $a_0, \dots, a_n$  that minimize

$$\sum_{i=1}^m (y_i - p_n(x_i))^2$$

Again, doing the standard procedure, the following system of equation emerges for variables  $a_0, \dots, a_n$ :

$$\begin{bmatrix} \sum x_i^0 & \sum x_i^1 & \sum x_i^2 & \dots & \sum x_i^n \\ \sum x_i^1 & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{n+1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \sum x_i^2 & \sum x_i^{n+1} & \sum x_i^{n+2} & \dots & \sum x_i^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum y_i x_i^0 \\ \sum y_i x_i^1 \\ \vdots \\ \sum y_i x_i^n \end{bmatrix}$$

## 6.2 Non-polynomial models

It is possible to propose a non-polynomial model for  $f$ , e.g.  $\hat{f}(x) = be^{ax}$  or whatever. This produces a non-linear system of equations which cannot be solved simply. However, taking the logarithm of the proposed model allows for a linear approximation. For example, if  $\hat{f}(x) = be^{ax}$  then

$$\ln \hat{f}(x) = \ln b + ax$$

which is a linear model that can be fitted using least squares. However, it is important to recall that the values  $y_1, \dots, y_m$  must be scaled with  $\ln y_1, \dots, \ln y_m$  to fit the logarithmic model.

## 6.3 Error of approximation

Assume  $f \in C[a, b]$  and  $P_n(x)$  of degree  $\leq n$  an approximation via least squares. Then

$$E = E(a_0, \dots, a_n) = \int_a^b (f(x) - P_n(x))^2 dx = \int_a^b [f(x) - \sum_{k=0}^n a_k x^k]^2 dx$$

In general, we wish to find  $a_0, \dots, a_n$  such that  $E$  is minimal. In general, the partial derivative of  $E$  with respect to  $a_j$  is zero if and only if

$$\sum_{k=0}^n \left( a_k \int_a^b x^{k+j} dx \right) = \int_a^b x^j f(x) dx, \quad j = 0, \dots, n$$

This readily provides a  $(n+1) \times (n+1)$  system of equations which is to be solved to minimize the error:

$$\begin{bmatrix} \int_a^b dx & \int_a^b x dx & \dots & \int_a^b x^n dx \\ \int_a^b x dx & \int_a^b x^2 dx & \dots & \int_a^b x^{n+1} dx \\ \vdots & \vdots & \dots & \vdots \\ \int_a^b x^n dx & \int_a^b x^{n+1} dx & \dots & \int_a^b x^{2n} dx \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \int_a^b x^0 f(x) dx \\ \int_a^b x f(x) dx \\ \vdots \\ \int_a^b x^n f(x) dx \end{bmatrix}$$

Notice that, in LHS matrix, the coefficient  $c_{jk}$  is

$$\int_a^b x^{j+k} dx = \frac{b^{j+k+1} - a^{j+k+1}}{j+k+1}$$

which allows for a more direct computation of the system. The LHS matrix is called **Hilbert matrix** and it is famously ill-conditioned—i.e. small changes in the coefficients greatly affect the final result.

## 6.4 A few polynomial theorems

Recall that vectors  $\phi_1, \dots, \phi_n$  are linearly independent if  $\sum_{i=0}^n c_i \phi_i = 0$  entails that  $c_0 = \dots = c_n = 0$ . In particular, a function space  $\mathcal{F}$  is a vector space such that  $\Pi = \bigcup_{i=0}^{\infty} \Pi_i \subseteq \mathcal{F}$ .

**Theorem 12** (Polynomials of different degrees are linearly independent). Let  $\phi_0, \dots, \phi_n$  such that  $\phi_j \in \Pi_j$ . Then  $\{\phi_i\}_{i=0}^n$  is linearly independent in any interval  $[a, b]$ .

**Proof.** Assume  $c_0, \dots, c_n$  are such that  $P(x) = c_0 \phi_0(x) + \dots + c_n \phi_n(x) = 0$  for any  $x \in [a, b]$ . Since  $P(x)$  is null for all  $x$  in  $[a, b]$ , each exponentiation  $x^k$  must be zero. Since  $c_n \phi_n(x)$  is the only term which includes  $x^n$ , we must have  $c_n = 0$ , which readily entails  $P(x) = c_0 \phi_0(x) + \dots + c_{n-1} \phi_{n-1}(x)$ . By recurrence,  $c_0 = \dots = c_n = 0$ . ■

The previous theorem is highly general, since any polynomial of degree  $n$  is by definition a linear combination of  $n+1$  polynomials of degree  $0, 1, \dots, n$ . For instance,  $1+3x+2x^2$  can be considered a linear combination of the polynomials  $1, x, x^2$  with coefficients  $1, 3, 2$ , or of the polynomials  $1, 3x, 4x^2$  with coefficients  $1, 1, 1/2$ , etc.

**Theorem 13.** Let  $\{\phi_0, \dots, \phi_n\}$  be a set of linearly independent polynomials for any interval  $[a, b]$ , each of degree  $\leq n$ . Then any polynomial of degree  $\leq n$  can be expressed as a linear combination of  $\{\phi_0, \dots, \phi_n\}$ .

## 6.5 Weighted function approximation: Diagonalizing Hilbert matrix

A weight function  $\omega(x)$  on interval  $I$  is a function such that:

1.  $\omega(x) > 0$  for all  $x \in I$ .
2.  $\omega(x) \neq 0$  for all  $x$  in any arbitrary sub-interval of  $I$ .

Condition (2) means that  $\omega$  cannot be constantly zero in any sub-interval, i.e. any zero mapping of  $\omega$  must be a point. Weight function allow one to give more or less importance to approximations in different regions of an interval of interest.

Now assume we have a linearly independent set of functions  $\{\phi_0, \dots, \phi_n\}$  in  $[a, b]$ , a weight function  $\omega$  defined on  $[a, b]$ , and  $f$  continuous in  $[a, b]$  which we wish to approximate. Then there is a set of coefficients  $a_0, \dots, a_n$  of

$$P(x) = \sum_{k=0}^n a_k \phi_k(x)$$

that minimize the following error function:

$$E = E(a_0, \dots, a_n) = \int_a^b \omega(x) [f(x) - P(x)]^2 dx$$

This is the same as we did before: we are approximating  $f$  via a polynomial by finding the coefficients that minimize the error of approximation. But now we are (a) including a weight function and (b) defining  $P(x)$  as a linear combination of independent polynomials.

Once more, taking the derivative of  $E$  with respect to an arbitrary  $a_j$ , we find that said derivative is zero if and only if

$$\sum_{k=0}^n a_k \int_a^b \omega(x) \phi_k(x) \phi_j(x) dx = \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad j = 0, \dots, n$$

which readily provides a system of equations. The system is complex, but it can be greatly simplified if we could choose  $\{\phi_0, \dots, \phi_n\}$  such that

$$\int_a^b \omega(x) \phi_k(x) \phi_j(x) dx = \alpha_j \delta_{jk}$$

for some  $\alpha_j > 0$ . In that case, the equation would simplify to

$$a_j \int_a^b \omega(x) \phi_j^2(x) dx = \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad j = 0, \dots, n$$

By assumption, this would yield

$$a_j \alpha_j = \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad j = 0, \dots, n$$

or equivalently

$$a_j = \frac{1}{\alpha_j} \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad j = 0, \dots, n$$

In short, a relatively simple system of equations emerges if  $\phi_0, \dots, \phi_n$  are intelligently chosen. But how to choose them intelligently?

[Orthogonal set] A set  $\{\phi_0, \dots, \phi_n\}$  of functions defined in  $[a, b]$  is orthogonal with respect to a weight function  $\omega$  iff

$$\int_a^b \omega(x) \phi_k(x) \phi_j(x) dx = \alpha_j \delta_{jk}, \quad j = 0, \dots, n$$

for  $\alpha_j > 0$ . If  $\alpha_j = 1$  for all  $j = 0, \dots, n$  then we say the set is orthonormal.

We can formalize what was said so far as follows:

**Theorem 14** (Orthogonal polynomial approximation theorem). If  $\{\phi_0, \dots, \phi_k\}$  defined in  $[a, b]$  is orthogonal with respect to  $\omega$  defined in  $[a, b]$ , then the least squared approximation of  $f$  weighted by  $\omega$  is

$$P(x) = \sum_{k=0}^n a_k \phi_k(x)$$

with

$$a_k = \frac{1}{\alpha_k} \int_a^b \omega(x) f(x) \phi_k(x) dx$$

where  $\alpha_k = \int_a^b \omega(x) \phi_k^2(x) dx$ .

**Theorem 15** (Orthogonal set generation). The following is an orthogonal set of functions defined in  $[a, b]$  with respect to a weight function  $\omega$ :

$$\phi_0(x) = 1, \quad \phi_1(x) = x - B_1, \quad x \in [a, b]$$

where

$$B_1 = \frac{\int_a^b x \omega(x) \phi_0^2(x) dx}{\int_a^b \omega(x) \phi_0^2(x) dx}$$

For  $k \geq 2$ ,

$$\phi_k(x) = (x - B_k) \phi_{k-1}(x) - C_k \phi_{k-2}(x), \quad x \in [a, b]$$

with

$$B_k = \frac{\int_a^b x \omega(x) \phi_{k-1}^2(x) dx}{\int_a^b \omega(x) \phi_{k-1}^2(x) dx}, \quad C_k = \frac{\int_a^b x \omega(x) \phi_{k-1}(x) \phi_{k-2}(x) dx}{\int_a^b \omega(x) \phi_{k-2}^2(x) dx}$$

In general, we will use a few orthogonal sets which are already known, built recurrently from the previous theorem:

### 1. Legendre polynomials:

$$\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x^2 - \frac{1}{3}, \phi_3(x) = x^3 - \frac{3}{5}x, \phi_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{25}, \dots$$

### 2. More examples in the textbook.



## 6.6 Exercises

(1) Approximate  $f(x)$  with a polynomial of degree 1 and  $g(x)$  with a polynomial of degree two, where:

$x$	0	1	2	3	4	5	6	7	8	9
$\hat{f}(x)$	-0.1	1.1	1.9	3.2	3.8	5.0	6.0	7.3	8.1	8.9

and

$x$	-1	0	1	3	6
$\hat{g}(x)$	6.1	2.8	2.2	6	26.9

(f) We know  $p(x) = a_0 + a_1x + \dots + a_nx^n$  is the least squared approximation of  $f$  if and only if  $a_0, \dots, a_n$  are solutions to the following system of equations:

$$\begin{bmatrix} \sum_{i=0}^m x_i^0 & \sum_{i=0}^m x_i^1 & \dots & \sum_{i=0}^m x_i^m & | & \sum_{i=0}^m x_i^0 y_i \\ \sum_{i=0}^m x_i^1 & \sum_{i=0}^m x_i^2 & \dots & \sum_{i=0}^m x_i^{m+1} & | & \sum_{i=0}^m x_i^1 y_i \\ \vdots & \vdots & \dots & \vdots & & \vdots \\ \sum_{i=0}^m x_i^m & \sum_{i=0}^m x_i^{m+1} & \dots & \sum_{i=0}^m x_i^{2m} & | & \sum_{i=0}^m x_i^m y_i \end{bmatrix}$$

where  $m$  is the number of points  $(x_i, \hat{f}(x_i))$ . Since we wish to approximate  $f$  with a polynomial of degree 1, this gives a  $2 \times 2$  matrix with coefficients

$$\begin{aligned} c_{11} &= \sum_{i=0}^9 (1) = 10, & c_{11} &= \sum_{i=0}^9 x = 45, \\ c_{21} &= \sum_{i=0}^9 x_i = 45, & c_{22} &= \sum_{i=0}^9 x_i^2 = 285 \end{aligned}$$

The solution vector, on the other hand, is:

$$w_1 = \sum_{i=0}^9 y_i = 45.2, \quad w_2 = \sum_{i=0}^9 x_i y_i = 286.7$$

So, suffices to solve the system

$$\begin{bmatrix} 10 & 45 & 45.2 \\ 45 & 285 & 286.7 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 4.5 & 4.52 \\ 1 & 6.33 & 6.371 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 4.5 & 4.52 \\ 0 & 1.83 & 1.85 \end{bmatrix}$$

So  $a_1 = \frac{1.85}{1.83}a_2 = 1.01a_2$ , bla bla. Linear system, cursillo de ingreso.

(g) Not that different from (f): calculate the matrix, which will be  $3 \times 3$  now (degree two polynomial), solve the resulting system. Calculating the matrix is computing sums (stupid), solving the system is solving a linear system (stupid).

(2) Prove that with  $n + 1$  distinct points, the best polynomial approximation (in the sense of least squares) of degree  $n$  coincides with the interpolating polynomial.

Let  $p_n(x)$  be the interpolating polynomial of  $f$  on points  $x_0, \dots, x_n$ . Let  $\ell_n(x)$  be the least squares approximation of degree  $n$ . Observe that

$$\sum_{i=0}^n (y_i - p_n(x_i))^2 = 0$$

Since  $\ell_n(x)$  minimizes the error among all polynomials of degree  $n$ , and taking  $p_n$  we have an error of zero, we must have

$$\sum_{i=0}^n (y_i - \ell_n(x_i))^2 = 0$$

which holds if and only if  $\ell_n(x)$  is an interpolating polynomial. But the interpolating polynomial is unique.  $\therefore p_n(x) = \ell_n(x)$ .

(3) Find the polynomial of degree 0 that best approximates  $f : [a, b] \rightarrow \mathbb{R}$  on  $x_1, \dots, x_n$  in  $[a, b]$ .

We wish to find a constant polynomial  $p_0(x) = c$  such that

$$S = \sum_{i=1}^n [y_i - p_0(x_i)]^2 = \sum_{i=1}^n [y_i - c]^2$$

is minimized. Taking

$$\begin{aligned} \frac{\partial S}{\partial c} &= \sum_{i=1}^n \frac{\partial}{\partial c} (y_i - c)^2 \\ &= \sum_{i=1}^n \frac{\partial u^2}{\partial u} \frac{\partial}{\partial c} (y_i - c) \\ &= \sum_{i=1}^n 2(y_i - c)(-1) \\ &= 2 \sum_{i=1}^n (c - y_i) \\ &= 2 \left( cn - \sum_{i=1}^n y_i \right) \end{aligned}$$

The expression above is zero if and only if

$$c = \frac{1}{n} \sum_{i=1}^n y_i$$

Therefore, the constant polynomial which best approximates  $f$  (in the sense of least squares) is the one whose value is the mean of  $y_1, \dots, y_n$ .

(4) Approximate the following data with a model of the form  $f(x) \sim ae^{bx}$ .

$$\begin{array}{c|c|c|c|c} x & -1 & 0 & 1 & 2 \\ \hline y & 8.1 & 3 & 1.1 & 0.5 \end{array}$$

If  $f(x) = ae^{bx}$  then  $g(x) = \ln(f(x)) = \ln a + bx$ . Now take the logarithm of the sampled image of  $f$ :

$$\begin{array}{c|c|c|c|c} x & -1 & 0 & 1 & 2 \\ \hline y & 2.09 & 1.09 & 0.09 & -0.69 \end{array}$$

We can fit the linear model  $g(x) = \ln a + bx$  to this data using the standard procedure. The  $2 \times 2$  matrix associated to the system whose solutions are  $\ln a, b$  is given by:

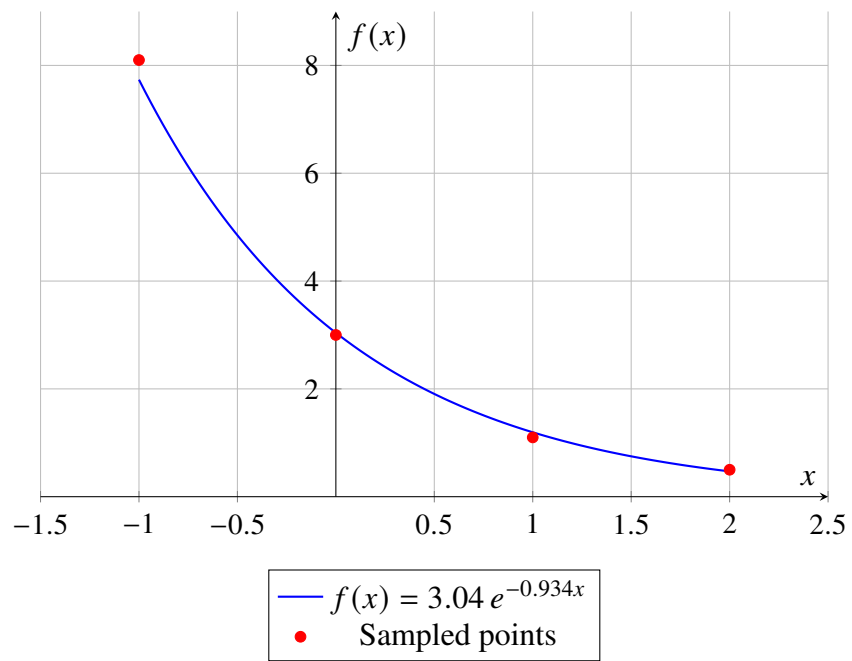
$$\begin{aligned} \sum_{i=0}^3 x_i^0 &= 4, & \sum_{i=0}^3 x_i^1 &= 2, & \sum_{i=0}^3 x_i^2 &= 6, \\ \sum_{i=0}^3 x_i^0 y_i &= 2.58, & \sum_{i=0}^3 x_i^1 y_i &= -3.38 \end{aligned}$$

In other words, the system is

$$\begin{bmatrix} 4 & 2 & 2.58 \\ 2 & 6 & -3.38 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 1.112 \\ 0 & 1 & -0.934 \end{bmatrix}$$

From this follows:

$$\ln a = 1.112 \text{ (or } a = e^{1.112} \approx 3.04), \quad b = -0.934$$



(5) Same, but with  $f(x) \sim -e^{ax^2+bx+c}$  and values

$$\begin{array}{c|c|c|c|c} x & -1 & 0 & 1 & 2 \\ y & -1.1 & -0.4 & -0.9 & -0.5 \end{array}$$

An obvious problem is that  $\mathcal{D}(\ln) = \mathbb{R}^+$ . So take

$$\phi(x) = \ln(-f(x)) = ax^2 + bx + c$$

so that the transformation we need to apply to the data is  $\ln \circ \psi$  with  $\psi$  the negation function:

$$\begin{array}{c|c|c|c|c} x & -1 & 0 & 1 & 2 \\ y & 0.09 & -0.91 & -0.1 & -0.69 \end{array}$$

Since  $n = 2$ , we need to compute a  $3 \times 3$  matrix with coefficients

$$\mathbf{A} = \begin{bmatrix} \sum x_i^0 & \sum x_i^1 & \sum x_i^2 \\ \sum x_i^1 & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}$$

to create the system

$$\mathbf{A} \begin{bmatrix} a & b & c \end{bmatrix}^\top = \begin{bmatrix} \sum y_i & \sum x_i y_i & \sum x_i^2 y_i \end{bmatrix}^\top$$

Now,

$$\begin{aligned} \sum_{i=0}^3 x_i^0 &= 4, & \sum_{i=0}^3 x_i &= 2, & \sum_{i=0}^3 x_i^2 &= 6, & \sum_{i=0}^3 x_i^3 &= 8, & \sum_{i=0}^3 x_i^4 &= 18 \\ \sum_{i=0}^3 x_i^0 y_i &= -1.61, & \sum_{i=0}^3 x_i y_i &= -1.57, & \sum_{i=0}^3 x_i^2 y_i &= -2.77 \end{aligned}$$

So the system can be expressed as:

$$\begin{bmatrix} 4 & 2 & 6 & -1.61 \\ 2 & 6 & 8 & -1.57 \\ 6 & 8 & 18 & -2.77 \end{bmatrix}$$

which solves to

$$a = -0.4285, \quad b = -0.2555, \quad c = 0.1025$$

Now, since

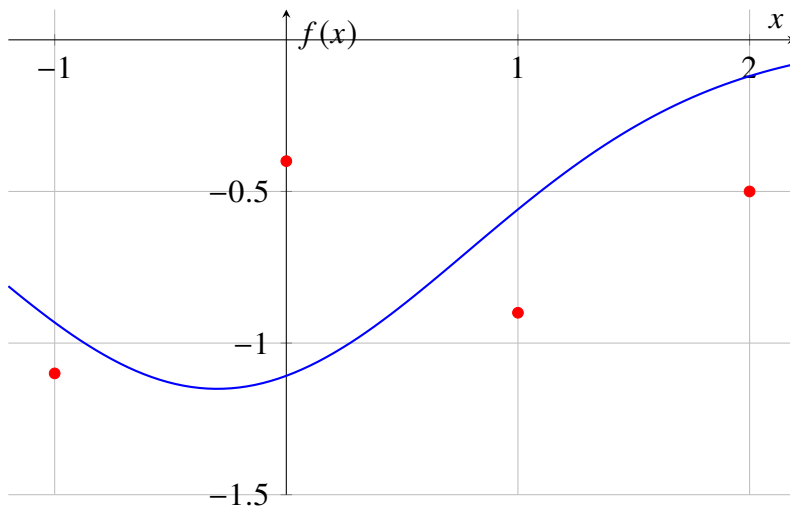
$$\phi(x) = \ln(-f(x)) = ax^2 + bx + c$$

we find

$$f(x) = -\exp(ax^2 + bx + c) = -e^{-0.4285x^2 - 0.2555x + 0.1025}$$

to be the least square fit of our sampled values with model  $f$ .

—  $f(x) = -\exp(-0.4285x^2 - 0.2555x + 0.1025)$  • Sampled values





(8) Approximate the data with  $f(x) \sim a \cos x + b \sin x$ .

$x$	0	1	2	3	4	5	6	7	8	9	10
$y$	1.8	3.5	2.1	-1.0	-3.3	-2.7	0.9	3.3	2.8	-0.1	-3.0

There is no way to transform  $f$  into a polynomial model (that I know of). So let's brute-force this:

$$\begin{aligned}
 E = E(a, b) &= \sum_{i=0}^n (y_i - f(x_i))^2 \\
 &= \sum_{i=0}^n (y_i - (a \cos x_i + b \sin x_i))^2 \\
 &= \sum_{i=0}^n (y_i - a \cos x_i - b \sin x_i)^2
 \end{aligned}$$

Now, it is easy to see that

$$\frac{\partial E}{\partial a} = 2 \sum_{i=0}^n (a \cos x_i + b \sin x_i - y_i) \cos x_i, \quad \frac{\partial E}{\partial b} = 2 \sum_{i=0}^n (a \cos x_i + b \sin x_i - y_i) \sin x_i$$

Equating both expressions to zero, one obtains the system:

$$\begin{aligned}
 a \sum_{i=0}^n \cos^2 x_i + b \sum_{i=0}^n \sin x_i \cos x_i &= \sum_{i=0}^n y_i \cos x_i \\
 a \sum_{i=0}^n \cos x_i \sin x_i + b \sum_{i=0}^n \sin^2 x_i &= \sum_{i=0}^n y_i \sin x_i
 \end{aligned}$$

In other words, we readily obtain the system

$$\begin{bmatrix} \sum_{i=0}^n \cos^2 x_i & \sum_{i=0}^n \sin x_i \cos x_i & \sum_{i=0}^n y_i \cos x_i \\ \sum_{i=0}^n \cos x_i \sin x_i & \sum_{i=0}^n \sin^2 x_i & \sum_{i=0}^n y_i \sin x_i \end{bmatrix}$$

Suffices to evaluate these sums using a calculator and solving the system to obtain the coefficients  $a, b$  which minimize the squared distances from the model to the set of data points.

(9) Consider Legendre's polynomials  $\mathcal{P} := \{P_0, P_1, P_2\}$  on  $[-1, 1]$  given by

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = x^2 - 1/3$$

Verify that the set  $\mathcal{P}$  is orthogonal.

The set  $\mathcal{P}$  is orthogonal (in  $[a, b]$ ) if and only if there is a weight function  $\omega$  such that

$$\int_a^b \omega(x) P_k(x) P_j(x) dx = \alpha_j \delta_{jk}, \quad j = 0, 1, 2$$

for some  $\{\alpha_0, \alpha_1, \alpha_2\}$  all positive. In particular, if  $\alpha_0 = \alpha_1 = \alpha_2 = 1$ , the set is orthonormal. Now, consider that  $\Omega(x) = 1$  is a weight function. It is easy to verify that

$$\int_{-1}^1 \Omega(x) P_j^2(x) dx > 0$$

for  $j = 0, 1, 2$ . Now,  $\int_{-1}^1 \Omega(x) P_0(x) P_1(x) dx = \int_{-1}^1 x dx = 0$ . Similarly,

$$\int_{-1}^1 \Omega(x) P_0(x) P_2(x) dx = \int_{-1}^1 x^2 - 1/3 dx = \frac{2}{3} - \frac{2}{3} = 0$$

Etc. This exercise is silly as it amounts to solving integrals.

**(10)** Determine the linear and quadratic approximations of  $f(x) = e^x$  in the least squares sense using Legendre's polynomials in  $[-1, 1]$ .

Let  $\Phi = \{1, x, x^2 - 1/3\}$ , and use  $\phi_0, \phi_1, \phi_2$  to refer to the elements in this set in the order I wrote them. We know  $\Phi$  to be orthogonal.

(Linear approximation) There is a set of coefficients  $a_0, a_1$  of

$$P(x) = \sum_{k=0}^n a_k \phi_k(x)$$

which minimizes

$$\int_{-1}^1 [f(x) - P(x)]^2 dx$$

Since  $\Phi$  is orthonormal, we have

$$a_j = \frac{1}{\alpha_j} \int_a^b f(x) \phi_j(x) dx, \quad j = 0, 1$$

Recall that

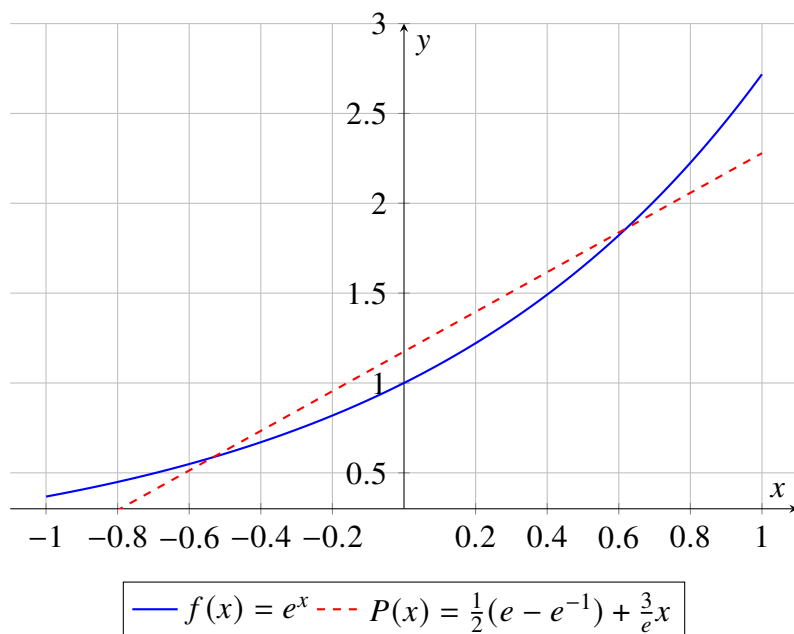
$$\alpha_0 = \int_{-1}^1 \phi_0^2(x) dx = 2, \quad \alpha_1 = \int_{-1}^1 \phi_1^2(x) dx = \frac{2}{3}, \quad \alpha_2 = \int_{-1}^1 \phi_2^2(x) dx = \frac{8}{45}$$

This readily gives the system of equations:

$$\begin{aligned} a_0 &= \frac{1}{2} \int_{-1}^1 e^x dx = \frac{1}{2}(e - e^{-1}) \\ a_1 &= \frac{3}{2} \int_{-1}^1 x e^x dx = \frac{3}{2} [e^{-1} + e^{-1}] = \frac{3 \cdot 2}{2e} = \frac{3}{e} \end{aligned}$$

Hence, the desired polynomial approximation of  $f(x) = e^x$  on  $[-1, 1]$  is:

$$P(x) = \frac{1}{2}(e - e^{-1}) + \frac{3}{e} \cdot x$$

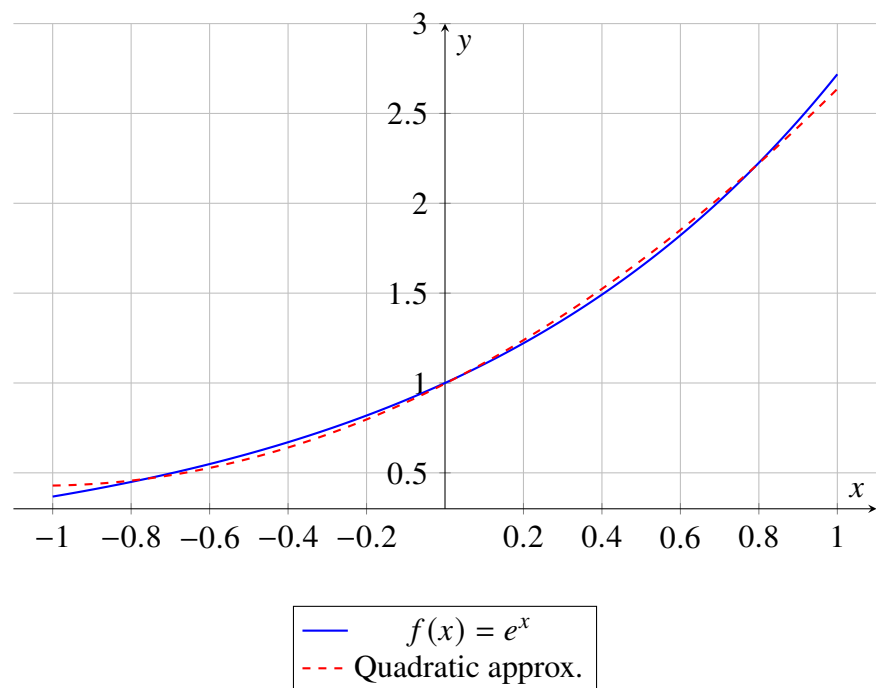


(Quadratic approximation) For a quadratic approximation, we now compute as well

$$a_2 = \frac{45}{8} \int_{-1}^1 e^x (x^2 - 1/3) dx = \frac{15(2e^2 - 14)}{8e}$$

So, the approximation is

$$P(x) = \frac{1}{2}(e - e^{-1}) + \frac{3}{e} \cdot x + \frac{15(2e^2 - 14)}{8e} \left( x^2 - \frac{1}{3} \right)$$



(10) Approximate  $f(x) = e^{-3x}$  for  $x \in (0, \infty)$  with a quadratic polynomial using least squares with weight function  $\omega(x) = e^{-x}$ , considering Laguerre polynomials defined as:

$$\phi_k(x) = \frac{e^x}{k!} \frac{d^k}{dx^k} (x^k e^{-x})$$

for all  $x \in [0, \infty)$ ,  $k = 0, 1, 2, \dots$ . Help: For each  $n \in \mathbb{N}$ , we know

$$\int_0^\infty x^n e^{-x} dx = n!$$

We will need  $\phi_0, \phi_1, \phi_2$  for a quadratic polynomial, so we might just as well begin by finding these polynomials. See the calculations below for their derivation.

**Derivation of  $\phi_1, \phi_2, \phi_3$ :** See that

$$\begin{aligned}\phi_0(x) &= e^x (x^0 e^{-x}) \\ &= e^{x-x} \\ &= 1\end{aligned}$$

Furthermore,

$$\begin{aligned}\phi_1(x) &= e^x \frac{d}{dx} (x e^{-x}) \\ &= e^x (e^{-x} - x e^{-x}) \\ &= 1 - x\end{aligned}$$

Lastly, since

$$\frac{d}{dx} x^2 e^{-x} = 2x e^{-x} - x^2 e^{-x}$$

then the second derivative of  $x^2 e^{-x}$  is:

$$2 \cdot \frac{d}{dx} (x e^{-x}) - \frac{d}{dx} x^2 e^{-x} = 2(e^{-x} - x e^{-x}) - (2x e^{-x} - x^2 e^{-x})$$

Por lo tanto,

$$\begin{aligned}
 \phi_2(x) &= \frac{e^x}{2!} \frac{d^2}{dx^2} (x^2 e^{-x}) \\
 &= \frac{e^x}{2} [2e^{-x} - 2xe^{-x} - 2xe^{-x} + x^2 e^{-x}] \\
 &= \frac{2 - 2x - 2x + x^2}{2} \\
 &= (x^2 - 4x + 2)1/2
 \end{aligned}$$

In summary,

$$\phi_0(x) = 1, \quad \phi_1(x) = -x + 1, \quad \phi_2(x) = (x^2 - 4x + 2)1/2$$

We know these polynomials are orthogonal with respect to  $\omega$ , which simply means that the Hilbert matrix with coefficients

$$a_{ij} = \int_0^\infty \omega(x) \phi_i(x) \phi_j(x) dx$$

satisfies

$$a_{ij} = \delta_{ij} \alpha_j, \quad j = 0, 1, 2$$

for some values  $\alpha_0, \alpha_1, \alpha_2$  which we must determine. Expressed differently, the coefficients  $a_0, a_1, a_2$  of the polynomial which approximates  $f$  are given by

$$a_j = \frac{1}{\alpha_j} \int_0^\infty \omega(x) f(x) \phi_j^2(x) dx$$

So let us now compute  $\alpha_0, \alpha_1, \alpha_2$ .

**Computation of the  $\alpha$  coefficients.** See that



$$\begin{aligned}\alpha_0 &= \int_0^\infty \omega(x) \phi_0^2(x) dx = \int_0^\infty x^0 e^{-x} dx \\ &= 1\end{aligned}$$

$$\begin{aligned}\alpha_1 &= \int_0^\infty \omega(x) \phi_1^2(x) dx = \int_0^\infty e^{-x} (1-x)^2 dx \\ &= \int_0^\infty e^{-x} (1 - 2x + x^2) dx \\ &= \int_0^\infty e^{-x} dx - 2 \int_0^\infty x e^{-x} dx + \int_0^\infty x^2 e^{-x} dx \\ &= 0! - 2 \times 1! + 2! \\ &= 1 - 2 + 2 \\ &= 1\end{aligned}$$

For the last problem, see that the numerator of  $\phi_2$  squared is:

$$\begin{aligned}(x^2 - 4x + 2)(x^2 - 4x + 2) &= (x^4 - 4x^3 + 2x^2) + (-4x^3 + 16x^2 - 8x) + (2x^2 - 8x + 4) \\ &= x^4 + (-4x^3 - 4x^3) + (2x^2 + 16x^2 + 2x^2) + (-8x - 8x) + 4 \\ &= x^4 - 8x^3 + 20x^2 - 16x + 4\end{aligned}$$

Then

$$\begin{aligned}\alpha_2 &= \int_0^\infty \omega(x) \phi_2^2(x) dx \\ &= \int_0^\infty e^{-x} (x^2 - 4x + 2)^2 (1/4) dx \\ &= \frac{1}{4} \int_0^\infty e^{-x} (x^4 - 8x^3 + 20x^2 - 16x + 4) dx \\ &= \frac{1}{4} [4! - 8 \times 3! + 20 \times 2! - 16 \times 1! + 4 \times 0!] \\ &= \frac{1}{4} [24 - 48 + 40 - 16 + 4] \\ &= \frac{1}{4} [68 - 64] \\ &= 1\end{aligned}$$

So we have  $\alpha_0 = \alpha_1 = \alpha_2 = 1$ . This means

$$a_j = \int_0^\infty \omega(x) \phi_j(x) f(x) dx$$

I don't know how to do the following fast without using the Gamma function:

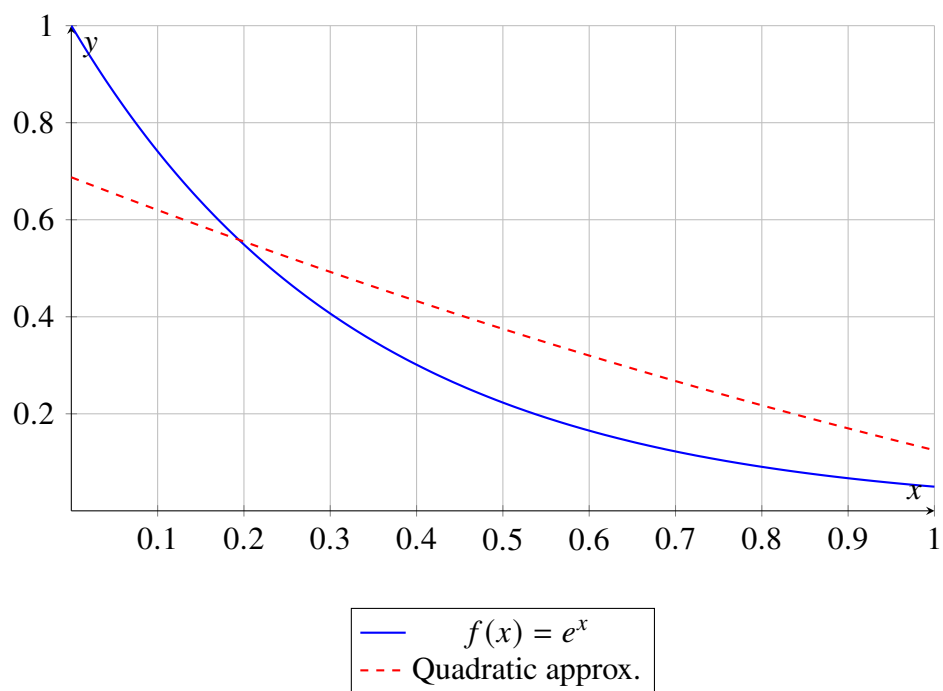
$$\int_0^\infty x^n e^{-kx} dx = \frac{n!}{k^{n+1}}$$

But we have:

$$\begin{aligned} a_0 &= \int_0^\infty e^{-x} e^{-3x} dx = \int_0^\infty e^{-4x} dx = \frac{1}{4} \\ a_1 &= \int_0^\infty e^{-x} (1-x) e^{-3x} dx \\ &= \int_0^\infty e^{-4x} (1-x) dx \\ &= \int_0^\infty e^{-4x} dx - \int_0^\infty e^{-4x} x dx \\ &= \frac{1}{4} - \frac{1}{4^2} \\ a_2 &= \frac{1}{2} \int_0^\infty e^{-x} (x^2 - 4x + 2) e^{-3x} dx \\ &= \frac{1}{2} \int_0^\infty e^{-4x} (x^2 - 4x + 2) dx \\ &= \frac{1}{2} \left[ \int_0^\infty x^2 e^{-4x} dx - 4 \int_0^\infty x e^{-4x} dx + 2 \int_0^\infty e^{-4x} dx \right] \\ &= \frac{1}{2} \left[ \frac{2!}{4^3} - 4 \frac{1!}{4^2} + 2 \frac{1}{4} \right] \\ &= \frac{1}{2} [1/4 - 1/4 + 1/2] \\ &= 1/4 \end{aligned}$$

So the approximating polynomial is

$$\begin{aligned}
 P(x) &= \sum_{i=0}^2 a_i \phi_i(x) \\
 &= \frac{1}{4} + \frac{3}{16}(-x+1) + \frac{1}{4} \frac{x^2 - 4x + 2}{2} \\
 &= \frac{1}{4} - \frac{3}{16}x + \frac{3}{16} + \frac{1}{8}x^2 - \frac{1}{2}x + \frac{1}{4} \\
 &= \frac{1}{8}x^2 - \frac{11}{16}x + \frac{11}{16}
 \end{aligned}$$



## 7 Numerical integration

We'll estimate  $\int_a^b f(x) dx$  via  $\sum_{i=0}^n a_i f(x_i)$ , where  $(x_0, y_0), \dots, (x_n, y_n)$  are known values of  $f$  (which itself might be unknown) in the interval  $[a, b]$ .

More generally, take  $\{x_i\}_{i=0}^n$  in  $[a, b]$  and  $P_n$  the interpolating polynomial of  $f$  in those nodes in Lagrange's form. Then we have

$$P_n(x) = \sum_{i=0}^n f(x_i) \ell_i(x), \quad e_n(x) = \frac{f^{(n+1)}(\zeta_x)}{(n+1)!} \eta_{n+1}(x)$$

for  $\zeta_x \in (x_0, x_n)$ . Using the fact that  $f(x) = P_n(x) + e_n(x)$ ,

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b P_n(x) dx + \int_a^b e_n(x) dx \\ &= \int_a^b \sum_{i=0}^n f(x_i) \ell_i(x) dx + \int_a^b \frac{f^{(n+1)}(\zeta_x)}{(n+1)!} \eta_{n+1}(x) dx \\ &= \sum_{i=0}^n a_i f(x_i) + \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\zeta_x) \eta_{n+1}(x) dx \end{aligned}$$

from  $\zeta_x \in (x_0, x_n)$  and  $a_i = \int_a^b \ell_i(x) dx, i = 0, \dots, n$ .

$$\therefore \int_a^b f(x) dx \approx \sum_{i=0}^n a_i f(x_i) \tag{1}$$

with an error

$$E_n(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\zeta_x) \eta_{n+1}(x) dx \tag{2}$$

### 7.1 Trapeze rule

**Summary.**

$$\int_a^b f(x) dx \approx \frac{h}{2} [f(a) + f(b)], \quad E(x) = -\frac{h^3}{12} f''(\zeta)$$

for some  $\zeta \in (a, b)$  and  $h = b - a$ .

**Observation.** If  $f'' \equiv 0$  then the rule is exact.

Trapeze's rule is the name for the case  $n = 1$  (linear integration). By convention,  $x_0 = a, x_1 = b, h = b - a$ . The interpolating polynomial is

$$P_1(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1)$$

with error

$$e_1(x) = \frac{f''(\zeta_x)}{2!} (x - x_0)(x - x_1)$$

Simplifying equation (1) for this case,

$$\begin{aligned} \int_a^b f(x) dx &\approx f(x_0) \int_a^b \frac{x - x_1}{x_0 - x_1} dx + f(x_1) \int_a^b \frac{x - x_0}{x_1 - x_0} dx \\ &= \frac{h}{2} (f(a) + f(b)) \end{aligned}$$

To simplify the expression of the error, the following theorem is used.

**Theorem 16** (Pretty theorem). Assume  $f \in C[a, b]$  and  $g$  integrable and either always positive or always negative in  $[a, b]$ . Then there is some  $c \in (a, b)$  such that

$$\int_a^b f(x)g(x) dx = f(c) \int_a^b g(x) dx$$

If  $g(x) \equiv 1$  then  $\int_a^b f(x) dx = f(c)(b - a)$ , entailing that

$$f(c) = \frac{1}{b - a} \int_a^b f(x) dx$$

It is easy to see that  $(x - x_0)(x - x_1)$  satisfies the conditions of the theorem, meaning that there is some  $\zeta$  (independent of  $x$ ) such that

$$\begin{aligned}
\int_a^b (\zeta_x)(x-a)(x-b) dx &= f''(\zeta) \int_a^b (x-a)(x-b) dx \\
&= f''(\zeta) \left[ \frac{x^3}{3} - \frac{a+b}{2}x^2 + abx \right]_a^b \\
&= f''(\zeta) \left( -\frac{h^3}{6} \right)
\end{aligned}$$

In summary, the error is:

$$E(x) = \frac{1}{2} f''(\zeta) \left( -\frac{h^3}{6} \right) = -\frac{h^3}{12} f''(\zeta)$$

with  $h = b - a$  and for some  $\zeta \in (a, b)$  independent of  $x$ .

## 7.2 Simpson's rule

Simpson's rule is the name for the case  $n = 2$  with three equidistant nodes  $a = x_0, x_1 = \frac{a+b}{2}, x_2 = b$ . We let

$$h = \frac{b-a}{2}, \text{ which entails the nodes are } a, a+h, a+2h$$

Three nodes allow us to construct an interpolating polynomial of degree 2. The rule is:

$$\int_a^b f(x) dx \approx \frac{h}{3} [f(a) + 4f(a+h) + f(a+2h)]$$

with

$$E = -\frac{h^5}{90} f^{(4)}(\zeta)$$

for some  $\zeta \in (a, b)$ .

## 7.3 Rectangle's rule

Stupid:

$$\int_a^b f(x) dx \approx f(a)(b-a)$$

which is the area in the rectangle from  $a, b$  with height  $f(a)$ , giving

$$E = \frac{(b-a)^2}{2} f'(\zeta), \quad \zeta \in (a, b)$$

One could take  $x_0 = b$  and estimate via the rectangle from  $a, b$  with height  $f(b)$ .

## 7.4 Midpoint rule

Stupid: Exactly the same as the rectangle rule, but the rectangle has height  $f(m)$ , where  $m = (a+b)/2$  is the midpoint of  $[a, b]$ :

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

with error

$$E = \frac{(b-a)^3}{24} f''(\zeta), \quad \zeta \in (a, b)$$

Since the error has a derivative of order 2, this method will integrate exactly all polynomials of degree 1.

## 7.5 Precision of estimation rules

The **precision** of a rule is the greatest non-negative integer  $n$  s.t. the formula is exact for  $x^k$ , for all  $k = 0, \dots, n$ .

## 7.6 Composite rules

**Simpson's composite rule.** Let  $f \in C^4[a, b]$ ,  $n$  a positive integer,  $h = (b-a)/2n$  and  $x_j = a + jH$  for  $j = 0, \dots, 2n$ . There is some  $\mu \in (a, b)$  such that the composite rule of Simpson for  $n$  sub-intervals is given by

Regla	Puntos	Fórmula	Error	Precisión
<b>Rectángulo</b>	1	$f(a)(b-a)$	$\frac{(b-a)^2}{2} f'(\xi)$	0
<b>Punto medio</b>	1	$f\left(\frac{a+b}{2}\right)(b-a)$	$\frac{(b-a)^3}{24} f''(\xi)$	1
<b>Trapecio</b>	2	$\frac{(b-a)}{2} [f(a) + f(b)]$	$-\frac{(b-a)^3}{12} f''(\xi)$	1
<b>Simpson</b>	3	$\frac{(b-a)/2}{3} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$	$-\frac{((b-a)/2)^5}{90} f^{(4)}(\xi)$	3

Table 1: Fórmulas de integración numérica y sus errores

$$\int_a^b f(x) dx = \frac{h}{3} \left( f(x_0) + 2 \sum_{j=1}^n f(x_{2j}) + 4 \sum_{j=1}^n f(x_{2j-1}) + f(x_n) \right) - \frac{b-a}{180} h^4 f^{(4)}(\mu)$$

To estimate large intervals, the interval is segmented into  $k$  segments of equal length of relatively small size and the estimations within segment are summed.

In general, if  $[a, b]$  is segmented in  $n$  sub-intervals of equal length, we obtain  $(2n + 1)$  equally spaced points  $x_j = a + jh$ , with  $j = 0, \dots, 2n$  and  $h = (b - a)/2n$ . Then,

$$\int_a^b f(x) dx = \sum_{j=1}^n \int_{x_{2j-2}}^{x_{2j}} f(x) dx$$

For Simpson's rule, this gives

$$\int_a^b f(x) dx \approx \sum_{j=1}^n \left( \frac{h}{3} [f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})] - \frac{h^5}{90} f^{(4)}(\zeta_j) \right)$$

with  $\zeta_j \in (x_{2j-2}, x_{2j})$ ,  $j = 1, \dots, n$ ,  $f \in C^4[a, b]$ . For  $j = 1, \dots, n$ , the term  $f(x_{2j})$  appears both in  $[x_{2j-2}, x_{2j}]$  and  $[x_{2j}, x_{2j+2}]$ , which allows us to simplify:

$$\int_a^b f(x) dx \approx \frac{h}{3} \left( f(x_0) + 2 \sum_{j=1}^n f(x_{2j}) + 4 \sum_{j=1}^n f(x_{2j-1}) + f(x_{2n}) \right) - \frac{h^5}{90} \sum_{j=1}^n f^{(4)}(\zeta_j)$$

Since  $f^{(4)}$  is assumed to be continuous in  $[a, b]$ , the extreme value theorem for continuous functions ensures:

- $f^{(4)}(\zeta_j)$  is between the minimum and the maximum of  $f^{(4)}$  in  $[a, b]$ .



- $\sum_{i=1}^n f^{(4)}(\zeta_j)$  is between  $n$  times the minimum and  $n$  times the maximum of  $f^{(4)}$  in  $[a, b]$ .
- Dividing in (2) by  $n$ , we obtain

$$\min_{x \in [a, b]} f^{(4)}(x) \leq \frac{1}{n} f^{(4)}(\zeta_j) \leq \max_{x \in [a, b]} f^{(4)}(x)$$

The intermediate value theorem ensures then that there is some  $\mu \in (a, b)$  such that

$$f^{(4)}(\mu) = \frac{1}{n} f^{(4)}(\zeta_j) \Rightarrow \sum_{j=1}^n f^{(4)}(\zeta_j) = n f^{(4)}(\mu)$$

Using this, the error can be expressed independently of  $\zeta_j$  as:

$$E(f) = -\frac{h^5}{90} \sum_{j=1}^n f^{(4)}(\zeta_j) = -\frac{h^5}{90} n f^{(4)}(\mu)$$

## 7.7 Other composite rules

No time. Memorize the formulas from apunte.

## 7.8 Gaussian rules

**Theorem 17.** Let  $w$  a positive weight function in  $[a, b]$  and  $q \neq 0$  a polynomial of degree exactly  $n + 1$  ortogonal to all polynomials of degree  $\leq n$  (with respect to  $n$ ). In other words, assume

$$\int_a^b q(x)p(x)w(x) \, dx = 0$$

Then if  $x_0, x_1, \dots, x_n$  are the  $n + 1$  roots of  $q$ , then

$$\int_a^b f(x)w(x) \, dx \approx \sum_{i=0}^n a_i f(x_i)$$

with  $a_i = \int_a^b w(x)\ell_i(x) \, dx$  is exact for all polynomials  $f$  of degree  $\leq 2n + 1$ .

## 7.9 Exercises

(2) Define

$$f(x) = \begin{cases} x & 0 \leq x \leq 1/2 \\ 1-x & 1/2 \leq x \leq 1 \end{cases}$$

in  $[0, 1]$ . Estimate  $\int_0^1 f(x) dx$  using (a) the trapeze rule in  $[0, 1]$ , the trapez rule on a half-segmentation of  $[0, 1]$ , (c) Simpson's rule on  $[0, 1]$ . (d) What can you conclude?

Let us first observe that

$$\begin{aligned} \int_0^1 f(x) dx &= \int_0^{1/2} x dx + \int_{1/2}^1 1-x dx \\ &= \frac{1}{8} + \frac{1}{2} - \left(\frac{1}{2} - \frac{1}{8}\right) \\ &= 1/4 \end{aligned}$$

(a) The trapeze rule was defined as

$$\int_a^b f(x) dx = \frac{h}{2} [f(a) + f(b)] - \frac{h^3}{12} f''(\zeta)$$

for some  $\zeta \in (a, b)$ . Here,  $h = 1 - 0 = 1$ ,  $f(0) = 0$ ,  $f(1) = 0$ . So its estimation gives  $1/2(0) = 0$ , with an error of 0.25.

(b) When using the rule on the two equal-split segments of the interval, we obtain

$$\begin{aligned} \int_0^1 f(x) dx &\approx \frac{1/2}{2} [f(0) + f(1/2)] + \frac{1/2}{2} [f(1/2) + f(1)] \\ &= \frac{1}{4} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} \\ &= 1/4 \end{aligned}$$

which is an exact approximation.

(c) Simpson's rule is defined for  $n = 2$  (three points), so we select  $1/2$  as midpoint having  $h = 1/2$ .

$$\begin{aligned}
 \int_0^1 f(x) \, dx &\approx \frac{h}{3} [f(0) + 4f(1/2) + f(1)] \\
 &= \frac{1}{6} [0 + 4/2 + 0] \\
 &= \frac{1}{3}
 \end{aligned}$$

(4) (a) Build a rule of the form

$$\int_{-1}^1 f(x) dx \approx \alpha f\left(-\frac{1}{2}\right) + \beta f(0) + \gamma f\left(\frac{1}{2}\right)$$

that is exact for all polynomials  $f$  degree  $\leq 2$ .

(b) Determine the precision of the formula for

$$\int_{-1}^1 f(x) dx \approx \frac{4}{3} f\left(-\frac{1}{2}\right) - \frac{2}{3} f(0) + \frac{4}{3} f\left(\frac{1}{2}\right)$$

(a) Let  $p_0(x) = 1, p_1(x) = x, p_2(x) = x^2$ . For the formula to be exact for polynomials of degree  $\leq 2$ , it must satisfy:

$$\begin{bmatrix} p_0\left(-\frac{1}{2}\right) & p_0(0) & p_0\left(\frac{1}{2}\right) \\ p_1\left(-\frac{1}{2}\right) & p_1(0) & p_1\left(\frac{1}{2}\right) \\ p_2\left(-\frac{1}{2}\right) & p_2(0) & p_2\left(\frac{1}{2}\right) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \int_{-1}^1 p_0(x) dx \\ \int_{-1}^1 p_1(x) dx \\ \int_{-1}^1 p_2(x) dx \end{bmatrix}$$

Let  $\varphi_i = \int_{-1}^1 p_i(x) dx$ . Note that  $\varphi_0 = 2, \varphi_1 = 0, \varphi_2 = \frac{2}{3}$ . Then, we can express the problem more succinctly.  $\alpha, \beta, \gamma$  are solutions to the following system of equations:

$$\left[ \begin{array}{ccc|c} 1 & 1 & 1 & 2 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{2}{3} \end{array} \right]$$

From the second row, we readily have  $\alpha = \gamma$ . Then, the third row expresses that

$$\frac{1}{2}\alpha = \frac{2}{3} \Rightarrow \alpha = \frac{4}{3}$$

Then, row one gives

$$\beta = 2 - \frac{8}{3} = -\frac{2}{3}$$

The formula then is

$$\int_{-1}^1 f(x) \, dx \approx \frac{4}{3}f\left(-\frac{1}{2}\right) - \frac{2}{3}f(0) + \frac{4}{3}f\left(\frac{1}{2}\right)$$

**(5)** (a) Determine the number  $n$  of sub-intervals needed in the composite trapeze rule to approximate  $\int_0^1 e^{-x^2} dx$  with an error  $\leq \frac{1}{2}10^{-6}$ , assuming  $e^{-x^2}$  can be computed precisely. (b) Repeat for Simpson's composite rule.

(a) If  $n$  sub-intervals are taken in  $[0, 1]$ , then each sub-interval is of length  $h(n) = \frac{1}{n}$ . Note as well that if  $n$  sub-intervals are taken, then there are  $n+1$  points forming intervals  $[x_1, x_2], \dots, [x_n, x_{n+1}]$ . The composite approximation is

$$\int_0^1 f(x) dx \approx \sum_{i=1}^n \left( \frac{h(n)}{2} [f(x_i) + f(x_{i+1})] - \frac{h^3}{12} f''(\zeta_i) \right)$$

where  $\zeta_i \in (x_i, x_{i+1})$ . We only care about the error here, which we can clear out, since

$$\begin{aligned} & \sum_{i=1}^n \left( \frac{h(n)}{2} [f(x_i) + f(x_{i+1})] - \frac{h^3(n)}{12} f''(\zeta_i) \right) \\ &= \sum_{i=1}^n \left( \frac{1}{2n} [f(x_i) + f(x_{i+1})] \right) - \frac{1}{12n^3} \sum_{i=1}^n f''(\zeta_i) \end{aligned}$$

Now, observe that

$$\frac{d}{dx} e^{-x^2} = \frac{d}{du} e^u \frac{d}{dx} (-x^2) = -2x e^{-x^2}$$

from which follows that the second derivative of  $f$  is

$$\frac{d}{dx} -2x e^{-x^2} = -2 \left( e^{-x^2} + x(-2x e^{-x^2}) \right) = -2e^{-x^2} + 4x^2 e^{-x^2} = f''(x)$$

The derivative of this function is

$$-2(4e^{-x^2} x^3 - 6e^{-x^2} x)$$

which is always negative in  $(0, 1)$  (since  $e^{-x^2}$  is always positive in  $(0, 1)$ ). Therefore,  $f''$  is a decreasing function and its maximum (in  $(0, 1)$ ) is bounded by  $f''(0) = -2$ . From this readily follows that the error term satisfies

$$-\frac{1}{12n^3} \sum_{i=1}^n f''(\xi_i) \leq -\frac{1}{12n^3} \sum_{i=1}^n (-2) = \frac{2n}{12n^3} = \frac{1}{6n^2}$$

Using the fact that the error term is less than or equal to  $1/6n^2$ , we must simply bound

$$\frac{1}{6n^2} \leq \frac{1}{2} 10^{-6}$$



(7) Estimate  $\int_{-1}^1 f(x) dx$  through a rule of the form

$$\int_{-1}^1 f(x) dx = \alpha f(x_1) + \beta f(x_2) + \gamma f(x_3)$$

that is exact for polynomials of degree  $\leq 5$ . Use it to approximate  $f(x) = \cos x$ .

We know from previous problems that  $\varphi_0 = 2, \varphi_1 = 0, \varphi_2 = \frac{2}{3}$ , where  $\varphi_i = \int_{-1}^1 x^i dx$ .

$$\int_{-1}^1 x^3 dx = 0, \quad \int_{-1}^1 x^4 dx = \frac{2}{5}, \quad \int_{-1}^1 x^5 dx = 0$$

We will take  $x_1 = -1, x_2 = 0, x_3 = 1$ . Then,  $\alpha, \beta, \gamma$  are the solutions of the following system of equations:

$$\begin{bmatrix} 1 & 1 & 1 & 2 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 1 & \frac{2}{3} \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 1 & \frac{2}{5} \\ -1 & 0 & 1 & 0 \end{bmatrix} \quad (1)$$

## 8 Systems of equations

The problem is finding the solution to  $Ax = b$  with  $A \in \mathbb{R}^{n \times n}$  and  $b \in \mathbb{R}^n$ .

**Diagonal case.** Obviously, if  $A$  is diagonal and its determinant is non-zero (i.e.  $a_{ii} \neq 0$  for all  $i = 1, \dots, n$ ) then  $Ax = b$  (with  $x, b \in \mathbb{R}^n$ ) has the single solution  $x_i = \frac{b_i}{a_{ii}}$ . So an algorithm is simply:

```

input  $n, A, b$ 
for  $i = 1, \dots, n$  do
     $x_i \leftarrow b_i / a_{ii}$ 
od

```

with a computational complexity  $O(n)$  (exactly  $n$  operations).

**Triangular case.** A matrix is upper-triangular iff  $a_{ij} = 0$  for  $i < j$  (lower-triangular if this holds for  $i > j$ ). Assume  $A \in \mathbb{R}^n$  is lower-triangular.

Assume  $A$  is lower-triangular and non-singular (i.e. invertible, i.e. non-zero determinant, ect.):

$$A = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \dots & 0 \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

Then  $Ax = b$  requires

$$\begin{cases} a_{11}x_1 & = b_1 \\ a_{21}x_1 + a_{22}x_2 & = b_2 \\ \vdots & \\ a_{n1}x_1 + \dots + a_{nn}x_n & = b_n \end{cases} \Rightarrow \begin{cases} x_1 & = b_1 / a_{11} \\ x_2 & = \frac{1}{a_{22}}(b_2 - a_{21}x_1) \\ \vdots & \\ x_i & = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j \right) \\ \vdots & \\ x_n & = \frac{1}{a_{nn}} \left( b_n - \sum_{j=1}^{n-1} a_{nj}x_j \right) \end{cases}$$

Therefore, an algorithm is simply setting for  $i = 1, \dots, n$  the value

$$x_i \leftarrow \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j \right) / a_{ii}$$

Esto hace  $O(n^2)$  operaciones.

## 8.1 Gaussian elimination

Gaussian elimination consists of solving  $Ax = b$  by finding  $(n - 1)$  transformations of  $A$  and  $b$ ,

$$A^{(1)}, \dots, A^{(n-1)}, \quad b^{(1)}, \dots, b^{(n-1)}$$

satisfying that  $U = A^{(n-1)}$  is upper-triangular and therefore contains the solution to the system. Let us show the transitions in part before generalizing.

(Transition  $A \rightarrow A^{(1)}$ ) Take  $a_{11}$  as pivot. We want to convert all values under  $a_{11}$  (first column except  $a_{11}$ ) into zeros. So for each row, we compute the multiplicative factor

$$m_i^{(1)} = \frac{a_{i1}}{a_{11}}$$

Observe that this factor is such that  $a_{i1} - m_i^{(1)}a_{11} = 0$ , in other words it is the number such that the row operation

$$i\text{th row} - \text{first row} \times m_i \tag{1}$$

ensures  $a_{i1} = 0$ . We then perform precisely the operation specified in (1), which gives  $A^{(1)}$  with coefficients

$$a_{ij}^{(1)} = \begin{cases} a_{11} & i = j = 1 \\ 0 & 1 < i \leq n, j = 1 \\ a_{ij} - m_i^{(1)}a_{1j} & c.c. \end{cases}, \quad b_i^{(1)} = \begin{cases} b_1 & i = 1 \\ b_i - m_i b_1 & 1 < i \leq n \end{cases}$$

How many operations are involved? Well, we can ignore the direct assignments and consider only the updating of non-zero values. In the matrix, there are  $(n - 1) \times (n - 1) = (n - 1)^2$  such computations to be made, each of which consists of subtracting from an existing coefficient a product. Then  $(n - 1)$  such computations must be performed in vector  $b$  as well. So the passing  $A \rightarrow A^{(1)}$  requires  $(n - 1)^2 + n - 1$  products,  $(n - 1)^2 + (n - 1)$  sums.

However, before the passing, we must also compute  $m_2, \dots, m_n$ , which involves  $n - 1$  quotients. So we have in total:

$$\text{\#Products} = (n - 1)^2 + 2(n - 1) = n^2 - 1, \quad \text{\#Sums} = (n - 1)^2 + (n - 1) = n^2 - n \quad (2)$$

( $A^{(1)} \rightarrow A^{(2)}$ ) We have all zeros in the first column (except for  $a_{11}$ ), so now we want all zeros in the second column (except  $a_{12}, a_{22}$ ). This involves in doing exactly the same as we did before, but in the  $(n - 1)^2$  sub-matrix that has  $a_{22}$  as pivot. So the same reasoning applies.

(General case:  $A^{(k)} \rightarrow A^{(k+1)}$ ) The pivot is  $a_{(k+1)(k+1)}$ , the multiplicative factor is  $m_i^{(k+1)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ , and the operations are:

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & 1 \leq i \leq k, 1 \leq j \leq n \\ 0 & k + 1 \leq i \leq n, j = k \\ a_{ij}^{(k)} - m_i^{(k+1)} a_{kj}^{(k)} & k \leq i \leq n, k + 1 \leq j \leq n \end{cases}$$

The number of operations are the same, but instead of dealing with matrix  $A$  of  $n^2$  dimensions, we are dealing with matrix  $A^k$  of  $(n - k)$  dimensions.

## 8.2 LU factorization

It is an exercise to show that the computational cost of Gaussian elimination is  $\mathcal{O}\left(\frac{2}{3}n^3\right)$ .

Assume several equations  $Ax = b_1, Ax = b_2$ , etc. wish to be solved. Same matrix, different  $b$  vectors. It'd be stupid to recompute the transformation of  $A$ .

Idea: Factor  $A = LU$  where  $L$  is lower-triangular with  $l_{ii} = 1$  and  $U$  is upper-triangular. Thus,

$$Ax = b \iff (LU)x = b \iff Ux = y \text{ and } Ly = b$$

Thus, once we know  $LU$  which factorizes  $A$ , instead of solving  $Ax = b$  we solve the two triangular systems  $Ux = y, Ly = b$ .  $L, U$  are obtained via construction by posing:

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}.$$

### 8.3 Iterative methods

[Vector norm] A vector norm in  $\mathbb{R}^n$  is a function  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}^{>0}$  satisfying:

1.  $\|\mathbf{x}\| > 0$  if  $\mathbf{x} \neq \vec{0}$ , and  $\|\vec{0}\| = 0$ .
2.  $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
3.  $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

[Distance] The distance between two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  is  $d(x, y) = \|\mathbf{x} - \mathbf{y}\|$ .

[Matrix norm] A matrix norm in  $\mathbb{R}^{n \times n}$  is a function mapping each matrix in the space to a non-negative real denoted  $\|A\|$ , such that for all  $A, B \in \mathbb{R}^{n \times n}$  and all  $\alpha \in \mathbb{R}$ :

1. Same 3 points as in vector norm.
2.  $\|AB\| \leq \|A\|\|B\|$ .

If we have a vector norm  $\|\cdot\|$  in  $\mathbb{R}^n$ , then we say

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

is the matrix norm induced by  $\|\cdot\|$ . In general, if a matrix norm is induced by a vector norm,

- $\|Ax\| \leq \|A\|\|x\|$  for all  $x \in \mathbb{R}^n$
- There is some  $\tilde{x}$  with  $\|\tilde{x}\| = 1$  such that  $\|A\tilde{x}\| = \|A\|$ .

The point is to generate a sequence of vectors  $\{\mathbf{x}^{(k)}\}$  that converge to a solution of  $A\mathbf{x} = \mathbf{b}$  under certain hypothesis.

Assume we expressed  $A$  as  $M - N$ . Then

$$\begin{aligned} A\mathbf{x} = \mathbf{b} &\iff (M - N)\mathbf{x} = \mathbf{b} \\ &\iff M\mathbf{x} = N\mathbf{x} + \mathbf{b} \\ &\iff \mathbf{x} = M^{-1}(N\mathbf{x} + \mathbf{b}) \\ &\iff \mathbf{x} = \left(M^{-1}N\right)\mathbf{x} + M^{-1}\mathbf{b} \end{aligned}$$

which is recursive because  $\mathbf{x}$  is defined in terms of  $\mathbf{x}$ . Given  $\mathbf{x}^{(0)}$  an approximation, we can construct

$$\mathbf{x}^{(k+1)} = (M^{-1}N)\mathbf{x}^{(k)} + M^{-1}b, \quad k \geq 0 \quad (1)$$

**Theorem 18** (Convergence of iterative methods). Let  $\mathbf{b} \in \mathbb{R}^n$ ,  $A = M - N \in \mathbb{R}^{n \times n}$ , with  $A, M$  non-singular. If  $\|M^{-1}N\| < 1$  for some induced matrix norm, then the sequence generated in equation (1) converges to the solution of  $A\mathbf{x} = b$  for any initial vector  $\mathbf{x}^{(0)}$ .

**Proof.** Take

$$\mathbf{x}^{(k+1)} = (M^{-1}N)\mathbf{x}^{(k)} + M^{-1}b, \quad k \geq 0 \quad (2)$$

Subtracting this from  $x_* = (M^{-1}N)x_* + M^{-1}b$ , we obtain

$$\mathbf{x}^{(k+1)} - x_* = (M^{-1}N)(\mathbf{x}^{(k)} - x_*), \quad k \geq 0 \quad (3)$$

Now, given an induced matrix norm, we obtain

$$\|\mathbf{x}^{(k+1)} - x_*\| = \|M^{-1}N\| \|\mathbf{x}^{(k)} - x_*\|, \quad k \geq 0 \quad (4)$$

Repeating the last step,

$$\|\mathbf{x}^{(k+1)} - x_*\| = \|M^{-1}N\|^{k+1} \|\mathbf{x}^{(0)} - x_*\|, \quad k \geq 0 \quad (5)$$

Using the fact that  $\|M^{-1}N\| < 1$ , we then have

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - x_*\| = 0 \quad (6)$$

## 8.4 Jacobi

In Jacobi we use  $M = D$  and therefore

$$N = M - A = D - (L + D + U) = -(L + U)$$

which means  $M$  is the diagonal matrix with the diagonal of  $A$ , and is a copy of  $A$  but has zeros in the diagonal.

A matrix  $A$  is diagonally dominant if

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|$$

**Theorem 19** (Jacobi's convergence). If  $A$  is diagonally dominant, the sequence generated by Jacobi's method converges to the solution of  $Ax = b$  for any initial vector  $x^{(0)} \in \mathbb{R}^n$ .

**Proof.** In Jacobi's method, the matrix  $M$  is the diagonal of  $A$  and it must be invertible for the method to be well-defined, meaning that  $a_{ii} \neq 0$  for all  $i = 1, \dots, n$ . Then the iteration matrix is given by

$$\begin{aligned} M^{-1}N &= - \begin{bmatrix} 1/a_{11} & 0 & \dots & 0 \\ 0 & 1/a_{22} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix} \\ &= - \begin{bmatrix} 0 & a_{12}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & \dots & a_{2n}/a_{22} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & 0 \end{bmatrix} \end{aligned}$$

It follows that

$$\|M^{-1}N\|_{\infty} = \max_{1 \leq i \leq n} \sum_{j=1}^n \frac{|a_{ij}|}{|a_{ii}|} < 1$$

since  $A$  is diagonally dominant. Then the method converges by virtue of **Theorem 1**.

## 8.5 Gauss-Seidel method

Here, we take  $N = M - a = L + D - (L + D + U) = -U$ . In other words, we obtain  $M$  to be the lower-triangular conversion of  $A$ , including the diagonal, and  $N$  to be the upper-triangular version of  $A$ , with zeros on the diagonal. It can be shown that

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right), \quad i = 1, \dots, n$$

**Theorem 20.** If  $A$  is diagonally dominant, then the sequence generated by the Gauss-Seidel method converges to the solution of  $Ax = b$ , for any initial vector  $x^{(0)} \in \mathbb{R}^n$ .

Define with  $\rho(A)$  the spectral radius of  $A$ , defined as

$$\max \{ |\lambda| : \lambda \text{ is eigenvalue of } A \}$$

**Theorem 21.** For any matrix  $A \in \mathbb{R}^{n \times n}$ , it holds that  $\rho(A) = \inf \{\|A\|\}$  on all induced matrix norms,

**Theorem 22.** A necessary and sufficient condition for the convergence of an iterative method

$$x^{(k+1)} = (M^{-1}N)x^{(k)} + M^{-1}b, \quad k \geq 0$$

for any initial vector  $x^{(0)}$  is that  $\rho(M^{-1}N) < 1$ .

## 8.6 Eigenvalues, eigenvectors

In the iteration

$$\vec{x}^{(k+1)} = M^{-1}N\vec{x}^{(k)} + M^{-1}\vec{b},$$

we define the iteration matrix  $T = M^{-1}N$ . To analyze convergence of the method, we consider the behavior of powers of  $T$  applied to the initial vector. If  $\rho(T) < 1$ , convergence is guaranteed by Theorem 22. When  $\rho(T) \geq 1$ , however, a more refined analysis is needed.

Assume  $\vec{b} = 0$ . Assume  $T$  is diagonalizable, so there exists a basis of eigenvectors  $\mathcal{E} = \{\vec{v}_1, \dots, \vec{v}_n\}$  with corresponding eigenvalues  $\lambda_1, \dots, \lambda_n$ . Then any vector  $\vec{x}^{(k)} \in \mathbb{R}^n$  can be expressed as a linear combination:

$$\vec{x}^{(k)} = \sum_{i=1}^n a_i^{(k)} \vec{v}_i.$$

Applying the iteration:

$$\vec{x}^{(k+1)} = T\vec{x}^{(k)} = \sum_{i=1}^n \lambda_i a_i^{(k)} \vec{v}_i,$$

which implies:

$$a_i^{(k+1)} = \lambda_i a_i^{(k)} \quad \Rightarrow \quad a_i^{(k)} = \lambda_i^k a_i^{(0)}$$

Thus,

$$\vec{x}^{(k)} = \sum_{i=1}^n \lambda_i^k a_i^{(0)} \vec{v}_i = (\lambda_1^k a_1^{(0)}, \dots, \lambda_n^k a_n^{(0)})_{\mathcal{E}} \quad (1)$$

This means that finding the expression of  $\vec{x}^{(0)}$  in the basis  $\mathcal{E}$ , with coefficients  $a_1^{(0)}, \dots, a_n^{(0)}$ , is sufficient to compute  $\vec{x}^{(k)}$  just in terms of these coefficients and powers of the eigenvalues.

Incidentally, this is the reason why we need all eigenvalues to be  $< 1$  in absolute value to ensure convergence. Otherwise, the values  $x^{(1)}, x^{(2)}, \dots$  diverge as seen from equation (1).



## 8.7 Exercises

(3) Show that the operational cost of Gaussian elimination for solving  $Ax = b$ ,  $A \in \mathbb{R}^{n \times n}$ , is  $O\left(\frac{2}{3}n^3\right)$  flops.

**Hint.** Recall that

$$\sum_{k=1}^n k = \frac{n(n+1)}{2}, \quad \sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$$

The Gaussian elimination method performs the transitions

$$A \rightarrow A^{(1)} \rightarrow A^{(2)} \rightarrow \dots \rightarrow A^{(n-1)} = U$$

for a total of  $n$  transitions. We know

$$P(k) = (n-k)^2 - 1, \quad S(k) = (n-k)^2 - (n-k)$$

are the number of products ( $P$ ) and sums ( $S$ ) on the  $k$ th to  $k+1$  transition. Obviously, as  $k$  increases across  $0, \dots, n-1$ , the coefficient  $(n-k)$  ranges from  $n, n-1, \dots, 1$ . So ultimately,

$$\begin{aligned} \sum_{k=0}^{n-1} P(k) + S(k) &= \sum_{k=0}^{n-1} (n-k)^2 - 1 + (n-k)^2 - (n-k) \\ &= \sum_{k=1}^n k^2 - 1 + k^2 - k \\ &= 2 \times \frac{n(n+1)(2n+1)}{6} - n - \frac{n(n+1)}{2} \\ &= \frac{n(n+1)(2n+1)}{3} - n - \frac{n^2+n}{2} \\ &= \frac{2n^3 + 3n^2 + n}{3} - n - \frac{n^2}{2} + \frac{n}{2} \\ &= \frac{2}{3}n^3 + n^2 + \frac{1}{3}n - \frac{1}{2}n^2 - \frac{1}{2}n \\ &= \frac{2}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{3}n - \frac{1}{2}n \end{aligned}$$

Clearly, the dominant term in the number of flops is  $\frac{2}{3}n^3$ , which means the number of operations satisfies the complexity bound  $\frac{2}{3}n^3 + O(n^2)$ . This suffices to show the number of flops is in the order  $O\left(\frac{2}{3}n^3\right)$ .

(4) Solve  $Ax = b$  with

$$A = \begin{bmatrix} 2 & -2 & 1 \\ 1 & 1 & 3 \\ 0 & 4 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ 6 \\ 9 \end{bmatrix}$$

(a) First, solve via Gaussian elimination. (b) Then via LU decomposition.

(a) Observe that

$$\begin{aligned} A = \begin{bmatrix} 2 & -2 & 1 & -1 \\ 1 & 1 & 3 & 6 \\ 0 & 4 & 1 & 9 \end{bmatrix} &\rightarrow \begin{bmatrix} 2 & -2 & 1 & -1 \\ 0 & 2 & 2.5 & 6.5 \\ 0 & 4 & 1 & 9 \end{bmatrix} & (m = 1/2) \\ &\rightarrow \begin{bmatrix} 2 & -2 & 1 & -1 \\ 0 & 2 & 2.5 & 6.5 \\ 0 & 4 & 1 & 9 \end{bmatrix} & (m = 0) \\ &\rightarrow \begin{bmatrix} 2 & -2 & 1 & -1 \\ 0 & 2 & 2.5 & 6.5 \\ 0 & 0 & -4 & -4 \end{bmatrix} & (m = 4/2 = 2) \end{aligned}$$

This concludes Gaussian elimination, and it readily provides the  $U$  and  $L$  matrix for LU method:

$$U = \begin{bmatrix} 2 & -2 & 1 \\ 0 & 2 & 2.5 \\ 0 & 0 & -4 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix}$$

Solving via Gaussian elimination entails applying the upper-triangular algorithm to the system obtained via Gaussian elimination, which gives:

$$x_3 = 1, \quad x_2 = (6.5 - 2.5)/2 = 2, \quad x_1 = (-1 + 2 \times 2 - 1 \times 1)/2 = 1$$

(b) To solve using LU decomposition, we note that  $Ax = b$  if and only if  $Ux = y$  and  $Ly = b$ . So, we begin solving  $Ly = b$ , which gives (according to matrix  $L$ )

$$\begin{bmatrix} 1 & 0 & 0 & -1 \\ \frac{1}{2} & 1 & 0 & 6 \\ 0 & 2 & 1 & 9 \end{bmatrix}$$

$$y_1 = -1, \quad y_2 = 6.5, \quad y_3 = -4$$

Then we solve  $Ux = y$ , i.e. the system

$$\begin{bmatrix} 2 & -2 & 1 & -1 \\ 0 & 2 & 2.5 & 6.5 \\ 0 & 0 & -4 & -4 \end{bmatrix}$$

which is exactly the same system solved in (a) and gives the same solutions.

(8) Consider the system

$$\begin{cases} x - y = 0 \\ x + y = 0 \end{cases}$$

Obtain the eigenvalues and eigenvectors of the associated iteration matrix (Gauss-Seidel) to decide if the method converges independently of the initial vector  $\vec{x}_0$ . Predict the behavior of the sequence with  $\vec{x}_0 \in \{(2, 0), (-0.03, 0.03), (0, 1)\}$ . Decide if Jacobi's method converges for these vectors.

The matrix associated to the system is simply

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

Recall that Gauss-Seidel's method writes  $A = (M - N)\vec{x} = \vec{b}$  with  $M$  the lower-triangular of  $A$  (including diagonal), and  $N$  the (negative) upper-triangular of  $A$  (with zeros in the diagonal). So we take

$$M = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Recall that the iteration is given by  $\vec{x}^{(k+1)} = (M^{-1}N)\vec{x}^{(k)} + M^{-1}\vec{b}$ , but since  $\vec{b} = \vec{0}$  the second term disappears. Thus, all that is needed is to compute the iteration matrix  $T = M^{-1}N$ . It is straightforward to see that the inverse of  $M$  is

$$M^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

which entails

$$T = M^{-1}N = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}$$

The eigenvalues are the solutions to the characteristic equation  $\det(T - \lambda I) = 0$ . Clearly,

$$\det(T - \lambda I) = \det \begin{pmatrix} -\lambda & 1 \\ 0 & -1 - \lambda \end{pmatrix} = -\lambda(-1 - \lambda)$$

with roots  $\lambda = 0, \lambda = -1$ . We know a necessary and sufficient condition for an iterative method to converge independently of the initial vector is  $\rho(T) = \rho(M^{-1}T) < 1$ , where  $\rho$  is the spectral radius. However,

$$\rho(T) = \max \{|\lambda| : \lambda \text{ is eigenvalue of } T\} = 1 \not< 1$$

$\therefore$  The matrix does not converge independently of the initial vector.

(b) To determine if the matrix converges with the given initial vectors, we need to compute the eigenvectors of  $T$ . To do this, we solve  $(T - \lambda I)\vec{v} = 0$  for each eigenvalue  $\lambda$ . Now, clearly

$$\begin{aligned} (T - 0 \times I)\vec{v} &= 0 \\ \iff T\vec{v} &= 0 \\ \iff v_1(0, 0) + v_2(1, -1) &= 0 \\ \iff (v_2, -v_2) &= 0 \\ \iff v_2 &= 0 \end{aligned}$$

Then any multiple of the eigenvector  $\vec{e}_1 = (1, 0)$  is an eigenvector. Similarly,

$$\begin{aligned} (T + I)\vec{v} &= 0 \\ \iff \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \vec{v} &= 0 \\ \iff \vec{v} &= (1, -1)^\top \end{aligned}$$

So the eigenvectors are  $e_1 = (1, 0), e_2 = (1, -1)$ . Then these vectors are a base and any  $\vec{x}^{(0)}$  is a linear combination of them. In other words,

$$\vec{x}^{(0)} = (x_1, x_2) = (a, 0) + (b, -b) = (a + b, -b)$$

for some  $a, b \in \mathbb{R}$ . We can then infer that  $b = -x_2$  and  $a = x_1 - b = x_1 + x_2$ . In other words, given any vector  $(x_1, x_2)$ , we can write it in terms of the eigen-basis using coefficients  $x_1 + x_2$  and  $-x_2$ . Then

$$\begin{aligned}
T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= T \left( (x_1 + x_2) \begin{bmatrix} 1 \\ 0 \end{bmatrix} - x_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) \\
&= (x_1 + x_2) T \begin{bmatrix} 1 \\ 0 \end{bmatrix} - x_2 T \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
&= (x_1 + x_2) \cdot 0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} - x_2 \cdot (-1) \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
&= (x_2, -x_2)
\end{aligned}$$

Therefore, if  $\vec{x}^{(k)} = (x, 0)$  for some  $x \in \mathbb{R}$ , immediately  $\vec{x}^{(k+1)} = T\vec{x}^{(k)} = (0, 0)$ , which means the method converges in one step. If  $\vec{x}^{(0)} = (x, y)$  with  $y \neq 0$ , then

$$\vec{x}^{(1)} = (y, -y), \quad \vec{x}^{(2)} = (-y, y), \quad \vec{x}^{(3)} = (y, -y), \dots$$

and the method does not converge.

## 9 Linear programming

A problem suited for linear programming consists in finding the solution to a system of equations

$$A\mathbf{x} = \mathbf{b}$$

which minimizes a cost function  $f(\mathbf{x})$  and which operates under certain constraints, among which we must impose  $\mathbf{x} \geq \vec{0}$ . The cost can itself be written as a system in function of  $\mathbf{x}$ :

$$\mathbf{c}^T \mathbf{x}$$

where  $c_i$  is the cost associated to each  $x_i$ . In general, it is very simple to translate a problem that does not fit the previous form into a problem that does. For instance, if one of the constraints is  $2x_1 + 7x_2 - 3x_3 \leq 10$ , we could write  $2x_1 + 7x_2 - 3x_3 + s_1 = 10$  with  $s_1 \geq 0$ .

A few definitions.

[Hyperplane, semi-space] A hyperplane of  $\mathbb{R}^n$  is the set of solutions to an arbitrary  $n$ -dimensional system:

$$\{\mathbf{x} \in \mathbb{R}^n \mid a_1x_1 + \dots + a_nx_n = b\}$$

A semi-space of  $\mathbb{R}^n$  is

$$\{\mathbf{x} \in \mathbb{R}^n \mid a_1x_1 + \dots + a_nx_n \leq b\}$$

[Convex set] A set  $S$  of  $\mathbb{R}^n$  is convex if for any  $\mathbf{x}, \mathbf{y} \in S$ , the segment joining these points is also within  $S$ . Formally, if

$$\alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \in S$$

for  $\alpha \in [0, 1]$ .

Two facts: a finite intersection of convex sets is convex, and every semi-space is convex. A finite intersection of closed semi-spaces in  $\mathbb{R}^n$  is called a **closed polyhedral region** of  $\mathbb{R}^n$ .

It follows that every closed polyhedral region is a convex set. Furthermore, the set of restrictions in a linear problem given by



$$\Omega = \{x \in \mathbb{R}^n \mid Ax = b \wedge Rx \leq s\}$$

is a closed polyhedral region. To characterize the polyhedral regions such as  $\Omega$ , we find the vertices, i.e. the intersection among the semi-spaces. Take the following linear problem:

$$\begin{cases} 3x + 2y & \geq 3 \\ x + 3y & \geq 1.5 \\ 8x + 2y & \geq 4 \\ x & \geq 0 \\ y & \geq 0 \end{cases}$$

The vertices among the lines are  $P_1 = (0, 2)$ ,  $P_2 = \left(\frac{1}{5}, \frac{5}{6}\right)$ ,  $P_3 = \left(\frac{6}{7}, \frac{3}{14}\right)$ ,  $P_4 = \left\{\frac{3}{2}, 0\right\}$ . For instance,  $P_2$  is obtained from solving

$$3x + 2y = 3, \quad 8x + 2y = 4$$

which are the hyperplanes (lines) in the first and third restrictions. Note that  $\left(0, \frac{3}{2}\right)$  is a solution to  $3x + 2y = 3, x = 0$  but it is not a vertex of *om*.