UNIVERSITY OF PRIMORSKA

Faculty of Mathematics, Natural Sciences and Information Technologies

Selected Topics in Numerical Mathematics: Lecture notes

Amar Bapić and Assoc. prof. dr. Vito Vitrih

OTHER STUDY TEXTBOOK

Pages 80

Mathematical Sciences, 2nd Bologna Cycle

1st edition

Koper, 2019

Foreword

This textbook is based on lectures delivered for the course *Selected Topics in Numerical Mathematics*, which is a part of the study program Mathematical Sciences, 2nd Bologna Cycle, at the University of Primorska. Topics contained in this textbook are Approximation, Ordinary Differential Equations and Partial Differential Equations.

Contents

Approximation

1.1. Introduction

Idea: We want to approximate some function, curve, surface, solution of a differential equation ... denoted by f with some approximant, which we will denote with $\tilde{f}$. We need approximation since

- general functions may be too complicated (expensive) to compute;
- general functions may be known only implicitly.

We have to answer several questions first:

- where to choose the approximant?
- which properties should $\tilde{f}$ share with f?
- does $\tilde{f}$ even exist in the space where we are searching for it?
- is it uniquely defined?
- how to construct it?
- how good the approximation is?

We will restrict ourselves in this section to the case, when f is a function.

Notation: We will denote the general space by X ($f \in X$) and the subspace where $\tilde{f}$ lives will be denoted by S ($\tilde{f} \in S$). In the following examples we list some possible choices for spaces X and S.

Example 1.1

- $X = \mathscr{C}([a,b])$ - *space of all continuous functions on* $[a,b]$;
- $X = \mathscr{C}^k([a,b])$ - *space of all k-times continuously differentiable functions on* $[a,b]$;
- $X = L^2([a,b])$ - *space of all Lebesgue integrable functions of* 2^{nd} *order on* $[a,b]$, *i.e.,* $\int_a^b (f(t))^2 dt < \infty$.

Example 1.2

- $S = \mathbb{P}_n$ - *space of all polynomials of degree less or equal to* n;
- $S = T_n$ - *space of trigonometric polynomials of degree less or equal to* n;
- $S = R_{n,m}$ - *space of rational functions whose numerator is a polynomial of degree less or equal to* n *and denominator is a polynomial of degree less or equal to* m;
- $S = S_{k,\mathbf{x}}$ - *space of piecewise polynomial functions (splines) with knot vector* $\mathbf{x}$. *Each polynomial piece is of degree* k. *The smoothness of the spline is* $k-1$.

In order to measure the difference between f and $\tilde{f}$ we need some norm defined on the space X, that is, the space X needs to be normalised with a norm $||\cdot||$. Let us consider the most important norms in the following example.

Example 1.3

(i) *Continuous uniform approximation:*
We are looking for $\tilde{f}$ *which minimizes the infinity norm, that is,*

$$||f-\tilde{f}||_{\infty,[a,b]} := \max_{x\in[a,b]} |f(x)-\tilde{f}(x)|.$$

(ii) *Discrete uniform approximation:*
We want to minimize

$$\max_{\substack{x_i\in[a,b]\\ i=0,1,\ldots,N}} |f(x_i)-\tilde{f}(x_i)|.$$

(iii) *Continuous least squares approximation:*
We want to find $\tilde{f}$ *which minimizes*

$$||f-\tilde{f}||_{2,[a,b]} := \left(\int_a^b (f(x)-\tilde{f}(x))^2 dx\right)^{\frac{1}{2}}.$$

(iv) *Discrete least squares approximation:*
We want to minimize

$$\sum_{i=0}^{N} (f(x_i)-\tilde{f}(x_i))^2; \qquad x_i \in [a,b].$$

1.2. Weierstrass theorem

When increasing the dimension of the subspace S, we expect that S will become dense in X.

Remark 1.1

Bernstein basis polynomials are of the form

$$B_i^n(x) = \binom{n}{i} x^i (1-x)^{n-i}.$$

With these we can represent any polynomial curve as

$$p_n(x) = \sum_{i=0}^{n} \mathbf{b}_i \cdot B_i^n(x),$$

where $\mathbf{b}_i$ represent the so-called control points. Curve represented in the Bernstein basis is called Bézier curve.

Theorem 1.1: (Weierstrass)

Let $f \in \mathscr{C}([a,b])$ be an arbitrary continuous function. Then

$$dist_\infty(f, \mathbb{P}_n) := \inf_{p \in \mathbb{P}_n} ||f-p||_\infty \xrightarrow[n\to\infty]{} 0.$$

Proof. Let's prove this theorem for $[0,1]$ (the proof on $[a,b]$ follows directly, since we can find a linear, bijective mapping $[0,1] \to [a,b]$ and such a mapping does not affect the polynomial degree). Let's define an operator $B_n : \mathscr{C}([a,b]) \to \mathbb{P}_n$ which maps

$$f(x) \mapsto \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} f\left(\frac{i}{n}\right) := (B_n f)(x) =: (B_n(f))(x).$$

Let us denote $p_i(x) = x^i$, $i = 0,1,2$, and calculate $B_n p_0$, $B_n p_1$ and $B_n p_2$.

$$(B_n p_0)(x) = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot 1 = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} = (x+1-x)^n = 1^n = 1.$$

$$\begin{aligned}(B_n p_1)(x) &= \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot \frac{i}{n} = \sum_{i=1}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot \frac{i}{n} \\ &= \sum_{i=0}^{n-1} \binom{n}{i+1} x^{i+1} (1-x)^{n-1+i} \cdot \frac{i+1}{n} = x \cdot \underbrace{\sum_{i=0}^{n-1} \binom{n-1}{i} x^i (1-x)^{n-1+i}}_{=B_{n-1}(p_0)} \\ &= x \cdot B_{n-1}(p_0) = x \cdot 1 = x.\end{aligned}$$

$$
\begin{aligned}
(B_n p_2)(x) &= \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot \left(\frac{i}{n}\right)^2 = \sum_{i=1}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot \left(\frac{i}{n}\right)^2 \\
&= \sum_{i=1}^{n} \binom{n-1}{i-1} x^i (1-x)^{n-i} \cdot \frac{i}{n} = \sum_{i=0}^{n-1} \binom{n-1}{i} x^{i+1} (1-x)^{n-1+i} \cdot \frac{i+1}{n} \\
&= x \cdot \sum_{i=0}^{n-1} \binom{n-1}{i} x^i (1-x)^{n-1+i} \cdot \underbrace{\frac{i+1}{n}}_{=\frac{n-1}{n}\left(\frac{i}{n-1}+\frac{1}{n-1}\right)} \\
&= \frac{x}{n} \left((n-1) \cdot \sum_{i=0}^{n-1} \binom{n-1}{i} x^i (1-x)^{n-1+i} \cdot \frac{i}{n-1} + \sum_{i=0}^{n-1} \binom{n-1}{i} x^i (1-x)^{n-1+i} \right) \\
&= \frac{x}{n} \left((n-1) B_{n-1}(p_1)(x) + B_{n-1}(p_0)(x) \right) = \frac{x}{n} \left((n-1)x + 1 \right) \\
&= x^2 + \frac{1}{n} x(1-x) = x^2 + \mathscr{O}\left(\frac{1}{n}\right).
\end{aligned}
$$

We have shown that

$$
B_n(p_0) = p_0, \quad B_n(p_1) = p_1, \quad B_n(p_2) \xrightarrow[n\to\infty]{} p_2.
$$

Our theorem will be proved if we show that

$$
||f - B_n(f)||_{\infty,[0,1]} := \max_{x\in[0,1]} |f(x) - B_n(f)(x)| \xrightarrow[n\to\infty]{} 0.
$$

Let $x \in [0,1]$ be arbitrary and let's prove that

$$
|f(x) - B_n(f)(x)| \xrightarrow[n\to\infty]{} 0.
$$

We have

$$
\begin{aligned}
|f(x) - B_n(f)(x)| &= \left| f(x) \cdot 1 - \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot f\left(\frac{i}{n}\right) \right| \\
&= \left| f(x) \cdot \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} - \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot f\left(\frac{i}{n}\right) \right| \\
&= \left| \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \left(f(x) - f\left(\frac{i}{n}\right) \right) \right| \\
&\leq \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \left| f(x) - f\left(\frac{i}{n}\right) \right|.
\end{aligned}
$$

Let us split the set of indices $I = \{0, 1, \ldots, n\}$ into two disjoint subsets I_1 and I_2 such that

$$I_1 = \left\{ i \in I \ : \ \left| \frac{i}{n} - x \right| < \frac{1}{\sqrt[4]{n}} \right\}, \quad I_2 = I \setminus I_1.$$

Firstly, let us consider the set I_1 and after that the set I_2. With $\omega(f_i; h)$ let us denote the so-called modulus of continuity:

$$\omega(f_i; h) = \max_{|x-y| \leq h} |f(x) - f(y)|$$

If f is continuous and if $h \to 0$, then $\omega(f_i; h) \to 0$. We now have:

$$\sum_{i \in I_1} \binom{n}{i} x^i (1-x)^{n-i} \underbrace{\left| f(x) - f\left(\frac{i}{n}\right) \right|}_{\leq \omega\left(f_i; \frac{1}{\sqrt[4]{n}}\right)} \leq \omega\left(f_i; \frac{1}{\sqrt[4]{n}}\right) \sum_{i \in I_1} \binom{n}{i} x^i (1-x)^{n-i}$$
$$= \omega\left(f_i; \frac{1}{\sqrt[4]{n}}\right) \cdot 1 = \omega\left(f_i; \frac{1}{\sqrt[4]{n}}\right) \xrightarrow[n \to \infty]{} 0.$$

In I_2 we have that

$$\left| \frac{i}{n} - x \right| \geq \frac{1}{\sqrt[4]{n}} \quad \Rightarrow \left(\frac{i - nx}{n} \right)^2 \geq \frac{1}{\sqrt{n}} \quad \Rightarrow \frac{(i-nx)^2}{n\sqrt{n}} \geq 1.$$

For any two $x, y \in [0, 1]$ we have that

$$|f(x) - f(y)| \leq 2||f||_{\infty, [0,1]} \leq 2M < \infty,$$

because f is a continuous function on $[0, 1]$, and thus it always has a maximal value somewhere on $[0, 1]$. Therefore

$$\sum_{i \in I_2} \binom{n}{i} x^i (1-x)^{n-i} \left| f(x) - f\left(\frac{i}{n}\right) \right| \leq 2M \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot 1$$
$$\leq 2M \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \cdot \frac{(i-nx)^2}{n\sqrt{n}} = 2M\sqrt{n} \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \left(\left(\frac{i}{n}\right)^2 - 2x\frac{i}{n} + x^2 \right)$$
$$= 2M\sqrt{n} \left((B_n p_2)(x) - 2x(B_n p_1)(x) + x^2 \right) = 2M\sqrt{n}(1-x) \cdot \frac{x}{n} = \frac{2M}{\sqrt{n}} \underbrace{x(1-x)}_{\leq \frac{1}{4}} \leq \frac{M}{2\sqrt{n}} \xrightarrow[n \to \infty]{} 0.$$

Since both sums over the indices in I_1 and I_2 approach 0 as $n \to \infty$, the sum over the indices in I approaches 0 as $n \to \infty$. In other words,

$$|f(x) - B_n(f)(x)| \xrightarrow[n \to \infty]{} 0.$$

□

The Weierstrass theorem shows that we always have a convergence with Bernstein polynomials, but this convergence is slow. Usually it holds:

$$||f - B_n(f)||_\infty = \mathcal{O}\left(n^{-1}\right) \quad \text{or} \quad ||f - B_n(f)||_\infty = \mathcal{O}\left(n^{-\frac{1}{2}}\right)$$

How do we guess these exponents?

Let us compute the infinity norm in the discrete way as

$$||f - B_n(f)||_\infty = \max_{i=0,1,\ldots,N} |f(x_i) - (B_n f)(x_i)| := e_n = c \cdot n^\alpha + \ldots$$

We now compute e_n and e_m, and divide them:

$$\left.\begin{array}{rl} e_n &= c \cdot n^\alpha \\ e_m &= c \cdot m^\alpha \end{array}\right\} \Rightarrow \frac{e_n}{e_m} = \left(\frac{n}{m}\right)^\alpha \Rightarrow \alpha = \frac{\log\left(\frac{e_n}{e_m}\right)}{\log\left(\frac{n}{m}\right)}.$$

Generalization: Let us recall the main properties of the operator B_n:

- B_n is a linear operator:

$$B_n(\alpha f) = \alpha B_n(f); \quad B_n(f+g) = B_n(f) + B_n(g).$$

- B_n is a positive operator:

$$f \geq 0 \;\Rightarrow\; B_n(f) \geq 0.$$

- $$B_n(1) = 1, \;\; B_n(x) = x, \;\; B_n(x^2) \xrightarrow[n\to\infty]{} x^2.$$

Theorem 1.2: (Korovkin)

Let $(L_n)_n$ be a sequence of positive linear operators mapping from $\mathscr{C}([a,b])$ to $\mathscr{C}([a,b])$ and let

$$||f - L_n(f)||_\infty \xrightarrow[n\to\infty]{} 0, \;\; f \in \{1, x, x^2\}.$$

Then

$$||f - L_n(f)||_\infty \xrightarrow[n\to\infty]{} 0 \quad \forall f \in \mathscr{C}([a,b]).$$

Example 1.4

Instead of polynomials we will now show that the space $S_{1,\boldsymbol{x}}$ is also an appropriate approximation space.

Now we will increase the number of knots

$$\mathbf{x} = \{a = x_0, x_1, \ldots, x_{n-1}, x_n = b\}.$$

Let us denote $\Delta x_j := x_{j+1} - x_j$ and $\Delta\mathbf{x} := \max_{j=0,1,\ldots,n-1} \Delta x_j$. When increasing knots we will require that $\Delta\mathbf{x} \xrightarrow[n\to\infty]{} 0$. Let's define an operator

$$I_n : \mathscr{C}([a,b]) \to S_{1,\mathbf{x}},$$

such that

$$f(x) \mapsto \sum_{i=0}^{n} f(x_i) H_i(x),$$

where H_i represents the so called "hat" function. We define them as follows:

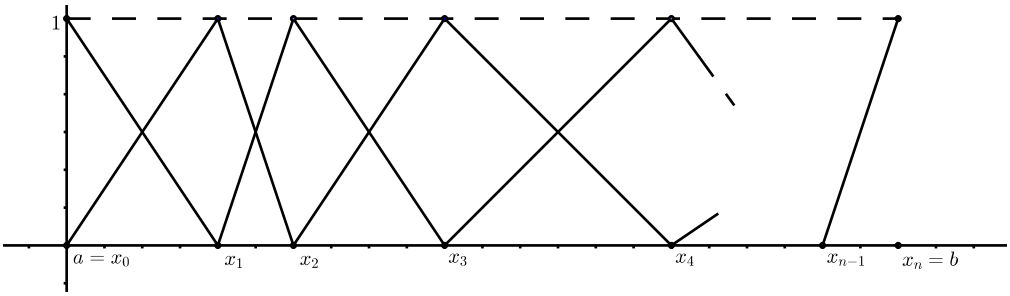


Figure 1.1: Graph of "hat" functions.

$$H_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - x_0}, & x \in [x_0, x_1] \\ 0, & \text{elsewhere} \end{cases}, \quad H_n(x) = \begin{cases} \frac{x - x_{n-1}}{x_n - x_{n-1}}, & x \in [x_{n-1}, x_n] \\ 0, & \text{elsewhere} \end{cases},$$

$$H_i(x) = \begin{cases} \frac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}] \\ \frac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i] \\ 0, & \text{elsewhere} \end{cases}, \quad i = 1, 2, \ldots, n-1.$$

Clearly, I_n is a linear and positive operator. Let us consider now $(I_n p_0)(x)$, $(I_n p_1)(x)$ and $(I_n p_2)(x)$ for some arbitrary $x \in [a,b]$. Since $x \in [a,b]$, we can always find an index $j \in \{0, 1, \ldots, n\}$ such that $x \in [x_{j-1}, x_j]$, implying $H_j(x) \neq 0$ and $H_{j-1}(x) \neq 0$, and for all $i \neq j, j-1$ we have $H_i(x) = 0$. Thus we have the following:

$$\begin{aligned} (I_n p_0)(x) &= \sum_{i=0}^{n} 1 \cdot H_i(x) = \sum_{i=0}^{n} H_i(x) = H_{j-1}(x) + H_j(x) \\ &= \frac{x_j - x}{x_j - x_{j-1}} + \frac{x - x_{j-1}}{x_j - x_{j-1}} = \frac{x_j - x_{j-1}}{x_j - x_{j-1}} = 1. \end{aligned}$$

$$
\begin{aligned}
(I_n p_1)(x) &= \sum_{i=0}^{n} x_i \cdot H_i(x) = x_{j-1} \cdot H_{j-1}(x) + x_j \cdot H_j(x) \\
&= x_{j-1} \cdot \frac{x_j - x}{x_j - x_{j-1}} + x_j \cdot \frac{x - x_{j-1}}{x_j - x_{j-1}} = \frac{x(x_j - x_{j-1})}{x_j - x_{j-1}} = x.
\end{aligned}
$$

$$
\begin{aligned}
\left|(I_n p_2)(x) - x^2\right| &= \left| \sum_{i=0}^{n} x_i^2 \cdot H_i(x) - x^2 \right| = \left| x_{j-1}^2 \cdot H_{j-1}(x) + x_j^2 \cdot H_j(x) - x^2 \right| \\
&= \left| x_{j-1}^2 \cdot \frac{x_j - x}{x_j - x_{j-1}} + x_j^2 \cdot \frac{x - x_{j-1}}{x_j - x_{j-1}} - x^2 \right| \\
&= \left| \frac{x_{j-1}^2 x_j - x_{j-1}^2 x + x_j^2 x - x_{j-1} x_j^2}{x_j - x_{j-1}} - x^2 \right| \\
&= \left| \frac{x_{j-1} x_j (-x_j + x_{j-1}) + x(x_j - x_{j-1})(x_j + x_{j-1}}{x_j - x_{j-1}} - x^2 \right| \\
&= \left| -x^2 + x(x_{j-1} + x_j) - x_{j-1} x_j \right| \\
&= \left| -\underbrace{(x - x_{j-1})}_{\geq 0} \underbrace{(x - x_j)}_{\leq 0} \right| = -(x - x_{j-1})(x - x_j) \\
&\leq -\left(\frac{x_{j-1} + x_j}{2} - x_{j-1} \right) \left(\frac{x_{j-1} + x_j}{2} - x_j \right) \\
&= \frac{(x_j - x_{j-1})^2}{4} \leq \frac{(\Delta \mathbf{x})^2}{4} \xrightarrow[n \to \infty]{} 0.
\end{aligned}
$$

Remark 1.2

In a similar way we can show the same for spaces $S_{k,\mathbf{x}}$ with $k > 1$.

1.3. Existence and uniqueness of approximant

Let X be a normalized space with $f \in X$, and let $\tilde{f} \in S$, $S \subseteq X$. We say that $\tilde{f}$ for which

$$||f - \tilde{f}|| \leq \inf_{s \in S} ||f - s||$$

an *element of best approximation*.

Existence: When S is of a finite dimension, then the existence of an approximant is guaranteed.

Theorem 1.3

Let X be a normalized space and $S \subseteq X$ such that $|S| < \infty$. Then for every $f \in X$ there exists an element of best approximation.

Proof. Let us choose an arbitrary $f \in X$ and let $\overline{K} = K(f, ||f||)$ be a closed ball with center in f and radius $||f||$. We observe that

$$\inf_{s \in S} ||f - s|| \overset{0 \in S}{\leq} ||f - 0|| = ||f|| \Rightarrow \tilde{f} \in \overline{K} \cap S := E.$$

Set E is closed and bounded, thus it is a compact set. We consider now the function from $S \to \mathbb{R}$, $s \mapsto ||f - s||$, which is continuous on a compact set, therefore it has it's minimal value somewhere on E. That is

$$\inf_{s \in S} ||f - s|| = \min_{s \in S} ||f - s|| = ||f - \tilde{f}||.$$

□

Remark 1.3

If $|S| = \infty$, such conclusion cannot be made.

Uniqueness: We always have it if X is strictly normalized.

Definition 1.1

A normalized vector space X is strictly normalized if for any $f, g \in X$, $g \neq 0$, for which we have

$$||f + g|| = ||f|| + ||g||,$$

it holds that

$$f = \lambda \cdot g,\ \lambda \in \mathbb{R}.$$

Theorem 1.4

Let X be strictly normalized and $S \subseteq X$, then there exists at most one best approximant.

Example 1.5

Let us take $X = \mathscr{C}([0,1])$ equipped with the infinity norm $||\cdot||_\infty$ and let $S = \mathbb{P}_n$. For $f(x) = x$ and $g(x) = x+1$ we have that

$$||f+g||_{\infty,[0,1]} = ||2x+1||_{\infty,[0,1]} = 3 = 1+2 = ||f||_{\infty,[0,1]} + ||g||_{\infty,[0,1]},$$

but

$$f \neq \lambda g,\ \lambda \in \mathbb{R}.$$

Therefore, X is not strictly normalized.

However, even if the space X is not strictly normalized, we can have uniqueness of the best approximant. But this now usually depends on space S. One such example will be presented in the next section.

1.4. Uniform approximation with polynomials

Let us consider the space $X = \mathscr{C}([a,b])$ normalized with the infinity norm $||\cdot||_\infty$ and the subspace $S = \mathbb{P}_n$. We are looking for $\tilde{f} \in S$ which minimizes the norm

$$||f-s||_{\infty, s\in S}.$$

We are asking ourselves three questions regarding $\tilde{f}$:

1. Does it exist?
2. Is it unique?
3. How do we construct it?

Definition 1.2

*Let $E \subseteq [a,b]$ and let f be a function defined on E with the norm being defined as $||f||_{\infty,E} = \max_{x\in E} |f(x)|$. We define the **minimax** for the function f on E and for $\mathbb{P}_n$ as*

$$M_n(E;f) := \min_{p\in\mathbb{P}_n} ||f-p||_{\infty,E} = \min_{p\in\mathbb{P}_n} \max_{x\in E} |f(x)-p(x)| =: dist_{\infty,E}(f,\mathbb{P}_n).$$

Definition 1.3

Let $\tilde{f} \in S$ be the polynomial of best uniform approximation for f in the given norm. The error of the approximation

$$f - \tilde{f}$$

will be called ***residual****.*

Since $S = \mathbb{P}_n$, we will usually write $\tilde{p}$ instead of $\tilde{f}$.

Idea: We want to construct $\tilde{p}$ with respect to $|E|$, the power of the set E.

(i) Let $|E| = n+1$. In this case $\tilde{p}$ is equal to the interpolating polynomial

$$\tilde{p}(x) = \sum_{i=0}^{n} f(x_i) \cdot L_{n,i}(x),$$

where

$$L_{n,i}(x) = \prod_{j=0,\ j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)}.$$

Let's prove that this representation is unique. To prove that let us assume that there are two different interpolating polynomials $\tilde{p}_1$ and $\tilde{p}_2$ for f. Since

$$\tilde{p}_1(x_i) = \tilde{p}_2(x_i) = f(x_i), \ \ i = 0, 1, \ldots, n,$$

that means that for $\tilde{p}_1 - \tilde{p}_2 \in \mathbb{P}_n$ we have

$$(\tilde{p}_1 - \tilde{p}_2)(x_i) = 0, \ \ i = 0, 1, \ldots, n.$$

Therefore $\tilde{p}_1 \equiv \tilde{p}_2$, which is a contradiction.

(ii) Let $|E| = n+2$, that is

$$E = \{x_i \mid a \leq x_0 < x_1 < \ldots < x_n < x_{n+1} \leq b\}.$$

We are looking for a polynomial p for which the value

$$m = \max_{x \in E} |f(x) - p(x)|$$

is the smallest among all $p \in \mathbb{P}_n$. We notice that

$$f(x_i) - p(x_i) = (-1)^i u_i m, \ \ |u_i| \leq 1, \ i = 0, 1, \ldots, n+1. \tag{1.4.1}$$

Let us write the polynomial p in the form

$$p(x) = \sum_{i=0}^{n} a_i x^i. \tag{1.4.2}$$

Together, (1.4.1) and (1.4.2) give us a system of linear equations which, written in the matrix form, is:

$$\begin{bmatrix} u_0 & 1 & x_0 & x_0^2 & \dots & x_0^n \\ -u_1 & 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ (-1)^{n+1}u_{n+1} & 1 & x_{n+1} & x_{n+1}^2 & \dots & x_{n+1}^n \end{bmatrix} \cdot \begin{bmatrix} m \\ a_0 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_{n+1}) \end{bmatrix}. \tag{1.4.3}$$

We have to choose u_i so that we get the best approximant, that is, we have to choose u_i so that m is as small as possible.

Notation: Let's denote the matrix of the system (1.4.3) with $A \in \mathbb{R}^{n+2,n+2}$ and with A_{ij} the matrix we obtain from A by deleting the i-th row and j-th column. We denote with D_i the determinant of $A_{i+1,1}$ where $i = 0,1,\dots,n+1$. Notice that D_i are Vandermonde matrices, hence

$$D_i = \det V(x_0,x_1,\dots,x_{i-1},x_{i+1},\dots,x_{n+1}) = \det V(y_0,y_1,\dots,y_n) = \prod_{i=0}^{n}\prod_{j=0}^{i-1}(y_i - y_j) > 0,$$

since for $i > j$ we have that $y_i > y_j$ for all $i = 0,1,\dots,n$.
By using the Cramer's rule we have that

$$m = |m| = \left| \frac{D_0 f(x_0) - D_1 f(x_1) + \ldots + (-1)^{n+1} D_{n+1} f(x_{n+1})}{u_0 D_0 + u_1 D_1 + \ldots + u_{n+1} D_{n+1}} \right| = \left| \frac{\sum_{i=0}^{n+1} (-1)^i f(x_i) D_i}{\sum_{i=0}^{n+1} u_i D_i} \right|$$

$$= \frac{\left|\sum_{i=0}^{n+1} (-1)^i f(x_i) D_i\right|}{\left|\sum_{i=0}^{n+1} u_i D_i\right|} \geq \frac{\left|\sum_{i=0}^{n+1} (-1)^i f(x_i) D_i\right|}{\sum_{i=0}^{n+1} |u_i| D_i} \geq \frac{\left|\sum_{i=0}^{n+1} (-1)^i f(x_i) D_i\right|}{\sum_{i=0}^{n+1} D_i}.$$

By looking at the right side of the inequality, m will reach his minimal value for

$$u_i = \varepsilon \ \ \forall i = 0,1,\dots,n+1, \text{ where } \varepsilon = 1 \text{ or } \varepsilon = -1.$$

We conclude that the best polynomial is the solution of the linear system

$$f(x_i) - \tilde{p}(x_i) = (-1)^i \varepsilon M_n(E;f),$$

where $\varepsilon = 1$ or $\varepsilon = -1$ is chosen such that $m \geq 0$.

Uniqueness: We are asking ourselves if $\det A \neq 0$. Let's say that

$$A = \begin{bmatrix} \pm 1 & 1 & x_0 & \dots & x_0^n \\ \mp 1 & 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ (\pm 1)^{n+1} & 1 & x_{n+1} & \dots & x_{n+1}^n \end{bmatrix}$$

then

$$\det A = \pm \sum_{i=0}^{n+1} D_i \neq 0,$$

since $D_i > 0$ for $i = 0, 1, \dots, n+1$.

Remark 1.4

We see that the residual $f - \tilde{p}$ alternately reaches the same value in all points of the set E.

Let us consider now two theorems without proofs for $|E| > n+2$.

(iii) Let $|E| < \infty$:

Theorem 1.5

Let $f \in \mathscr{C}([a,b])$ be an arbitrary function and $S = \mathbb{P}_n$ the subspace in which we are looking for the approximants. Let $G \subseteq [a,b]$ be a set of finite cardinality and let us denote with $E \subseteq G$ a set of cardinality $n+2$ such that

$$M_n(E; f) \geq M_n(E'; f), \ \ \forall E' \subseteq G, \ |E'| = n+2.$$

Then the polynomial of best uniform approximation for f on E is also the polynomial of best uniform approximation for f on G. It is defined uniquely.

(iv) Let $|E| = [a,b]$:

Theorem 1.6

Let $f \in \mathscr{C}([a,b])$. In $\mathbb{P}_n$ there is a uniquely defined polynomial of the best uniform approximation for f in $[a,b]$. It matches with the polynomial of best uniform approximation for f on $E \subseteq [a,b]$, where E is of cardinality $n+2$ for which

$$M_n(E;f) \geq M_n(E';f),$$

for all $E' \subseteq [a,b]$, $|E'| = n+2$.

Recall previous remark, which says that the residual $f - \tilde{p}$ alternately reaches the same value in all points of the set E. Also the reverse sentence is true in the following way.

Theorem 1.7: De la Vallée-Poussin

Let $f \in \mathscr{C}([a,b])$. If the polynomial $p \in \mathbb{P}_n$ is chosen such that the difference $f - p$ alternately reaches its norm $||f-p||_\infty$ in $n+2$ different points $x_i \in [a,b]$, then p is the polynomial of best uniform approximation for f on $[a,b]$.

Proof. Let $\tilde{p}$ be the polynomial of best uniform approximation for f on $[a,b]$. We have to show that $p = \tilde{p}$. Let us denote

$$m := ||f-p||_\infty \geq M_n([a,b];f) = ||f-\tilde{p}||_\infty \tag{1.4.4}$$

Let us suppose that

$$||f-\tilde{p}||_\infty < m = ||f-p||_\infty = \max_{0\leq i\leq n+1} |f(x_i) - p(x_i)|. \tag{1.4.5}$$

We have the following

$$\tilde{p}(x_i) - p(x_i) = (f(x_i) - p(x_i)) - (f(x_i) - \tilde{p}(x_i)).$$

Let us multiply both sides with $\operatorname{sgn}(f(x_i) - p(x_i))$. Then

$$\begin{aligned} \operatorname{sgn}(f(x_i) - p(x_i))(\tilde{p}(x_i) - p(x_i)) &= \operatorname{sgn}(f(x_i) - p(x_i)) \cdot (f(x_i) - p(x_i)) - (f(x_i) - \tilde{p}(x_i)) \\ &= |f(x_i) - p(x_i)| - \operatorname{sgn}(f(x_i) - p(x_i)) \cdot (f(x_i) - \tilde{p}(x_i)) \\ &\geq m - ||f-\tilde{p}||_\infty > 0, \end{aligned}$$

where $m = |f(x_i) - p(x_i)|$ because the residual reaches it's norm in the points x_i and $\operatorname{sgn}(f(x_i) - p(x_i)) \cdot (f(x_i) - \tilde{p}(x_i))$ is obviously less or equal than $||f-\tilde{p}||_\infty$. Also,

$$\tilde{p}(x_{i+1}) - p(x_{i+1}) = (f(x_{i+1}) - p(x_{i+1})) - (f(x_{i+1}) - \tilde{p}(x_{i+1}))$$

Let us multiply again both sides with $\operatorname{sgn}(f(x_i)-p(x_i))$. Then

$$\begin{aligned}
&\operatorname{sgn}(f(x_i)-p(x_i))(\tilde{p}(x_{i+1})-p(x_{i+1})) \\
&\qquad = \operatorname{sgn}(f(x_i)-p(x_i))\cdot(f(x_{i+1})-p(x_{i+1}))-(f(x_{i+1})-\tilde{p}(x_{i+1})) \\
&\qquad = -|f(x_{i+1})-p(x_{i+1})|+\operatorname{sgn}(f(x_{i+1})-p(x_{i+1}))\cdot(f(x_{i+1})-\tilde{p}(x_{i+1})) \\
&\qquad \leq -m+||f-\tilde{p}||_\infty < 0,
\end{aligned}$$

where $\operatorname{sgn}(f(x_i)-p(x_i)) = -\operatorname{sgn}(f(x_{i+1})-p(x_{i+1}))$. We see that the polynomial $\tilde{p}-p \in \mathbb{P}_n$ has strict opposite signs in the points x_i and x_{i+1}, $i = 0,1,\ldots,n$. Hence, $\tilde{p}-p \in \mathbb{P}_n$ has zeros on all intervals $(x_i,\ x_{i+1})$, $i = 0,\ldots,n$. We got that a n-degree polynomial has at least $n+1$ zeros on a given segment $[a,b]$, which is a contradiction. Hence

$$m = ||f-p||_\infty = ||f-\tilde{p}||_\infty \Rightarrow p \equiv \tilde{p}.$$

□

EXAMPLE: Let us consider the function $f(x) = x^n \in \mathscr{C}([a,b])$ and let $S = \mathbb{P}_{n-1}$. For $x = \cos\theta$, where $\theta \in [0,\pi]$, we have that $\theta = \arccos x$ (the mapping $[0,\pi] \to [-1,1]$ is bijective.) Let's define the Chebyshev polynomials of the first kind T_n as

$$T_n(x) := \cos(n\theta) = \cos(n\arccos x)$$

First, let us prove that $T_n \in \mathbb{P}_n$. From

$$\cos(n\theta) = \frac{1}{2}\left(e^{in\theta}+e^{-in\theta}\right) = \frac{1}{2}\left(\left(e^{i\theta}\right)^n+\left(e^{-i\theta}\right)^n\right)$$

and

$$e^{\pm i\theta} = \cos\theta \pm i\sin\theta = \cos(\arccos x) \pm i\sin(\arccos x) = x \pm i\sqrt{1-x^2} = x \pm \sqrt{x^2-1}$$

we get that

$$\begin{aligned}
\cos(n\theta) &= \frac{1}{2}\left(\left(x+\sqrt{x^2-1}\right)^n+\left(x-\sqrt{x^2-1}\right)^n\right) \\
&= \frac{1}{2}\left(\sum_{i=0}^{n}\binom{n}{i}x^i\left(x^2-1\right)^{\frac{n-i}{2}}+\sum_{i=0}^{n}\binom{n}{i}(-1)^{n-i}x^i\left(x^2-1\right)^{\frac{n-i}{2}}\right) \\
&= \frac{1}{2}\sum_{i=0}^{n}\binom{n}{i}x^i\left(x^2-1\right)^{\frac{n-i}{2}}\left(1+(-1)^{n-i}\right) \in \mathbb{P}_n.
\end{aligned}$$

The leading coefficient of T_n is equal to

$$\frac{1}{2}\sum_{i=0}^{n}\binom{n}{i}+\frac{1}{2}\sum_{i=0}^{n}\binom{n}{i}(-1)^{n-i}=\frac{1}{2}\left((1+1)^n+(1-1)^n\right)=2^{n-1}\neq 0.$$

We conclude, T_n is a polynomial of degree n with leading coefficient 2^{n-1}. Let us denote

$$x_k=\cos(\theta_k),\quad \theta_k=\frac{k\pi}{n},\ \ k=0,1,\ldots,n.$$

We have

$$T_n(x_k)=\cos(n\arccos x_k)=\cos(k\pi)=(-1)^k,\ \ k=0,1,\ldots,n.$$

Obviously, $||T_n||_\infty=1$. We are looking for a polynomial $p\in\mathbb{P}_{n-1}$ for which the residual alternately reaches its norm in $(n-1)+2=n+1$ points. We see that $2^{1-n}T_n$ is obviously the residual $f-p$. Hence,

$$2^{1-n}T_n=x^n-p(x)\Rightarrow p(x)=x^n-2^{1-n}T_n.$$

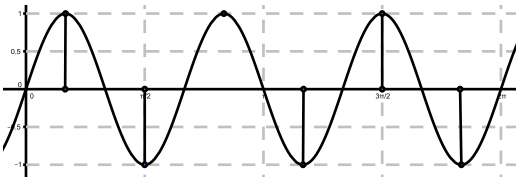
Remark 1.5

$$dist_\infty(x^n,\mathbb{P}_{n-1})=\min_{p\in\mathbb{P}_{n-1}}||f-p||_\infty=2^{1-n}=\frac{1}{2^{n-1}}.$$

This tells us that calculations with the monomial basis $(1,x,x^2,\ldots,x^n)$ can be numerically unstable.

Example 1.6

Let $f(x)=\sin(3x)$ be defined on $[0,2\pi]$. Find $\tilde{p}\in\mathbb{P}_4$.



If we take $p\equiv 0$ we have that $r=f-p=f$ alternately reaches its norm $(=1)$ in $n+2=6$ points.

```
Determine the polynomial of the best uniform approximation p̃ ∈ P_n for the given
    f ∈ C([a,b]) and ε > 0.

E_0 := {x_i : a ≤ x_0 ≤ x_1 ≤ ... ≤ x_{n+1} ≤ b};
k := 0;
repeat
```

```
6
7    find p̃_k as polynomial of the best uniform approximation for f on E_k;
8    (we solve the system f(x_i) − p̃_k(x_i) := e_k(x_i) = (−1)^i m_k, i = 0,...,n+1)
9
10   find y ∈ [a,b] such that |f(y) − p̃_k(y)| = ||f − p̃_k||_∞;
11
12   if |f(y) − p̃_k(y)| − m_k < ε
13    return;
14   else
15    E_{k+1} = (E_k ∪ {y}) \ {x_j},
16    where x_j is the nearest left or right point of y
17    such that sgn(e_k(y)) = sgn(e_k(x_j));
18   k++;
19 end
```

Listing 1.1: Remez algorithm

Theorem 1.8

Let $f \in \mathscr{C}([a,b])$ and $(\tilde{p}_k)_{k\geq 0}$ be a sequence of polynomials obtained by the Remez algorithm. Let $\tilde{p}$ be the polynomial of best uniform approximation for f on $[a,b]$. Then there exist $c > 0$ and $0 < \theta < 1$ such that

$$0 \leq ||f - \tilde{p}_k||_\infty - M_n([a,b]; f) \leq c \cdot \theta^k.$$

Therefore

$$||f - \tilde{p}_k||_\infty \xrightarrow[k\to\infty]{} ||f - \tilde{p}||_\infty \quad \text{and} \quad ||\tilde{p}_k - \tilde{p}||_\infty \xrightarrow[k\to\infty]{} 0.$$

Remark 1.6

The speed of convergence of the given algorithm is linear. If we would change in each step all of the points in the set E_k, then the speed of the convergence would be quadratic.

Example 1.7

Let $f(x) = e^x \in \mathscr{C}([0,1])$. Find $\tilde{p}_1 \in \mathbb{P}_1$ using Remez algorithm.

Solution: Let $k = 0$, $p_0(x) = k_0 x + n_0$ and $E_0 = \{0, \frac{1}{2}, 1\}$. Let us solve the system $f(x_i) - p_0(x_i) = (-1)^i m_0$, $i = 0,1,2$.

$$\begin{cases} 1 - n_0 & = m_0 \\ e^{\frac{1}{2}} - \frac{k_0}{2} - n_0 & = -m_0 \\ e - k_0 - n_0 & = m_0 \end{cases}$$

We obtain that

$$k_0 = e - 1 = 1.71828\ldots,\ n_0 = 0.89479\ldots,\ m_0 = 0.10521\ldots$$

Let us find the value y from the algorithm.

$$\begin{aligned} e_0'(x) = 0 &\Leftrightarrow e^x - k_0 = 0 \\ &\Leftrightarrow e^x = e - 1 \\ &\Leftrightarrow x = \ln(e-1) \\ &\Leftrightarrow x = 0.5413\ldots = y \end{aligned}$$

Obivously, we conclude that

$$E_1 = \{0, \ln(e-1), 1\}.$$

Now let $k = 1$ and $p_1(x) = k_1 x + n_1$. Solving the system

$$\begin{cases} 1 - n_1 &= m_1 \\ e - 1 - k_1 \cdot \ln(e-1) - n_1 &= -m_1 \\ e - k_1 - n_1 &= m_1 \end{cases}$$

gives us that

$$k_1 = e - 1,\ \ n_1 = 0.89407\ldots,\ \ m_1 = 0.10593\ldots$$

Again,

$$\begin{aligned} e_1'(x) = 0 &\Leftrightarrow e^x - k_1 = 0 \\ &\Leftrightarrow e^x = e - 1 \\ &\Leftrightarrow x = \ln(e-1) \\ &\Leftrightarrow x = 0.5413\ldots = y \end{aligned}$$

Hence,

$$E_2 = \{0, \ln(e-1), 1\} = E_1.$$

□

We can now generalize Remez algorithm to other approximation spaces.

Definition 1.4

The functions $S = \{f_0, f_1, \ldots, f_n\}$ form the so called Chebyshev system of functions on $[a,b]$ if the generalized Vandermonde determinant

$$\det V(x_0, x_1, \ldots, x_n; S) = \det\left(V\left(f_i(x_j)\right)_{i,j=0}^{n}\right)$$

is nonzero for $x_i \neq x_j$, $i \neq j$.

One can prove the following theorem:

Theorem 1.9

Let $S = \{f_0, f_1, \ldots, f_n\}$ be a Chebyshev system of functions on $[a,b]$. We can then use Remez algorithm for the space $\mathscr{L}in\{f_0, f_1, \ldots, f_n\}$.

Example 1.8

(a) *Is $S = \{1, x, x^2, \ldots, x^n\}$ a Chebyshev system of functions on $[a,b]$?*

(b) *Is $S = \{1, x^2\}$ a Chebyshev system of functions on $[-1,1]$?*

(c) *Is $S = \{1, x^2\}$ a Chebyshev system of functions on $[0,1]$?*

Solution:

(a) Since

$$D = \det\left(V\left(f_i(x_j)\right)_{i,j=0}^{n}\right) = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix} = \prod_{i=0}^{n} \prod_{j=0}^{i-1} (x_i - x_j),$$

we know that $D \neq 0$ for $x_i \neq x_j$, for all $i \neq j$. This is true on any interval $[a,b]$. Hence, S is a Chebyshev system of functions on $[a,b]$.

(b)

$$D = \begin{vmatrix} 1 & x_0^2 \\ 1 & x_1^2 \end{vmatrix} = x_1^2 - x_0^2 = (x_1 - x_0)(x_1 + x_0).$$

If we take $x_0 < x_1$, $x_0, x_1 \in [-1,1]$, then for example for $x_0 = -\frac{1}{2}$ and $x_1 = \frac{1}{2}$, we have $D = 0$ but $x_0 \neq x_1$. Hence, S is not a Chebyshev system of functions on $[-1,1]$.

(c) In this case S is a Chebyshev system of functions on $[0,1]$.

□

Example 1.9

We would like to approximate the function $f(x) = x$ on $[0,2]$ with elements of the space $\mathscr{L}in\{e^x, e^{2x}\}$ by the method of best uniform approximation. Show that we can use Remez algorithm and present one step of the algorithm.

Solution: We first compute

$$D = \begin{vmatrix} e^{x_0} & e^{2x_0} \\ e^{x_1} & e^{2x_1} \end{vmatrix} = e^{x_0+2x_1} - e^{x_1+2x_0}.$$

Let us take $x_0 < x_1$; $x_0, x_1 \in [0,2]$. If we suppose that $D = 0$ then we would have that $x_0 + 2x_1 = x_1 + 2x_0$, implying $x_0 = x_1$, which is a contradiction. Hence, $D \neq 0$.

Let $E_0 = \{0,1,2\}$ and $p_0(x) = a \cdot e^x + b \cdot e^{2x}$. Solving the system $f(x_i) - p_0(x_i) = (-1)^i m, \;\; x_i \in E_0$, in other words,

$$\begin{cases} 0 - a - b &= m \\ 1 - ae - be^2 &= -m \\ 2 - ae^2 - be^4 &= m \end{cases}$$

gives us that

$$a = \frac{e^2 - 3}{e(e^2-1)} = 0.2527\ldots, \;\; b = \frac{3-e}{e(e-1)(e^2+1)} = 0.0072\ldots$$

The residual is $r(x) = f(x) - p_0(x) = x - ae^x - be^{2x}$. Hence, we have

$$r'(x) = 0 \Leftrightarrow 1 - ae^x - 2be^{2x} = 0.$$

Using the substitution $z = e^x$ and the fact that $z > 0$, we conclude that

$$z = \frac{a - \sqrt{a^2 + 8b}}{-4b}.$$

The new point is therefore $x = \ln z = 1.2021\ldots$ and thus

$$E_1 = \{0, 1.2021\ldots, 2\}.$$

1.5. Least squares approximation method

We will consider a space X which is normalized with the second norm $||\cdot||_2 = \sqrt{\langle \cdot, \cdot \rangle}$ and finite subset S.

From before we know that the existence and uniqueness of the best approximant is guaranteed. The problem we are facing now is how to construct such an approximant.

Example 1.10

1. *Continuous case:*
$$\langle f,g\rangle_\rho = \int_a^b f(x)g(x)\rho(x)dx,\ \ \rho \geq 0.$$
2. *Discrete case:*
$$\langle f,g\rangle_\rho = \sum_{i=0}^{n} f(x_i)g(x_i)\rho(x_i).$$

Theorem 1.10

Let $S \subseteq X$. The element $\tilde{f} \in S$ is the best approximant by the least squares method for $f \in X$ if and only if $f - \tilde{f} \perp S$.

Proof.

($\Leftarrow$) Let us suppose that $f - \tilde{f} \perp S$. For $\tilde{f}$ to be the best approximate it suffices to prove that

$$||f-s||_2^2 \geq ||f-\tilde{f}||_2^2,\ \ \forall s \in S.$$

Using the assumption we obtain

$$\begin{aligned} ||f-s||_2^2 &= ||\underbrace{(f-\tilde{f})}_{\in S^\perp} + \underbrace{(\tilde{f}-s)}_{\in S}||_2^2 = \langle (f-\tilde{f})+(\tilde{f}-s),(f-\tilde{f})+(\tilde{f}-s)\rangle \\ &= ||f-\tilde{f}||_2^2 + 2\underbrace{\langle \tilde{f}-s, f-\tilde{f}\rangle}_{=0} + \underbrace{||\tilde{f}-s||_2^2}_{\geq 0} \geq ||f-\tilde{f}||_2^2. \end{aligned}$$

($\Rightarrow$) Let $s \in S$, let $\lambda > 0$ be an arbitrary real number and let $\tilde{f}$ be the best approximant. We need to prove that $\langle f-\tilde{f}, s\rangle = 0$. We have

$$\begin{aligned} 0 \leq ||f-\tilde{f}+\lambda s||_2^2 - ||f-\tilde{f}||_2^2 &= ||f-\tilde{f}||_2^2 + 2\lambda\langle f-\tilde{f},s\rangle + \lambda^2||s||_2^2 - ||f-\tilde{f}||_2^2 \\ &= 2\lambda\langle f-\tilde{f},s\rangle + \lambda^2||s||_2^2. \end{aligned}$$

When $\lambda \to 0$, then $\lambda^2 \to 0$ faster than $\lambda \to 0$. Thus $\lambda^2||s||_2^2$ can be neglected and we have that

$$\langle f-\tilde{f}, s\rangle \geq 0. \tag{1.5.1}$$

Since s was chosen arbitrarily and we know that $-s \in S$, we have also that

$$\langle f - \tilde{f}, -s \rangle \geq 0.$$

Multiplying the last inequality with -1 we get that

$$\langle f - \tilde{f}, s \rangle \leq 0. \tag{1.5.2}$$

From (1.5.1) and (1.5.2) we conclude that

$$\langle f - \tilde{f}, s \rangle = 0.$$

□

Let $(s_i)_{i=1}^n$ be the basis of the space S. That means we can write

$$\tilde{f} = \sum_{i=1}^{n} \alpha_i s_i.$$

By Theorem 1.10 we have that

$$f - \sum_{i=1}^{n} \alpha_i s_i \perp S.$$

It suffices to consider the basis elements, that is

$$\left\langle f - \sum_{i=1}^{n} \alpha_i s_i, s_j \right\rangle = 0, \quad j = 1, 2, \ldots, n.$$

Equivalently,

$$\sum_{i=1}^{n} \alpha_i \langle s_i, s_j \rangle = \langle f, s_j \rangle$$

or in the matrix form

$$G \cdot \mathbf{a} = \mathbf{b},$$

where

$$G = \begin{bmatrix} \langle s_1, s_1 \rangle & \langle s_2, s_1 \rangle & \ldots & \langle s_n, s_1 \rangle \\ \langle s_1, s_2 \rangle & \langle s_2, s_2 \rangle & \ldots & \langle s_n, s_2 \rangle \\ \vdots & \vdots & \ldots & \vdots \\ \langle s_1, s_n \rangle & \langle s_2, s_n \rangle & \ldots & \langle s_n, s_n \rangle \end{bmatrix}$$

is the Gram matrix and

$$\mathbf{a} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \langle f, s_1 \rangle \\ \langle f, s_2 \rangle \\ \vdots \\ \langle f, s_n \rangle \end{bmatrix}.$$

Remark 1.7

The solution $\tilde{f}$ is obtained by solving the given linear system. Thus, the problem of finding the best approximant for $||\cdot||_2$ is much simpler than finding it for $||\cdot||_\infty$.

The linear system $G \cdot \mathbf{a} = \mathbf{b}$ is called a normal system. Since for $\mathbf{x} \neq 0$ and $(s_i)_{i=1}^n$ being a basis

$$\mathbf{x}^T G \mathbf{x} = \sum_{i,j=1}^{n} x_i g_{ij} x_j = \sum_{i,j=1}^{n} x_i \langle s_i, s_j \rangle x_j = \left\langle \sum_{i=1}^{n} x_i s_i, \sum_{i=1}^{n} x_j s_j \right\rangle = \langle \mathbf{y}, \mathbf{y} \rangle > 0.$$

Therefrom we conclude that the Gram matrix G is a symmetric positive definite matrix. Thus, for solving the system we can use the Cholesky decomposition for which we need $\frac{1}{3}n^3 + \mathscr{O}(n^2)$ operations.

We have two possibilities to find a solution:

1. We choose an arbitrary basis $(s_i)_i$ and solve the linear system $G \cdot \mathbf{a} = \mathbf{b}$.
2. From the given basis $(s_i)_i$, we compute the orthogonal basis $(\tilde{s}_i)_i$ and with it we find the solution without solving a linear system.

The following example will demonstrate why the first option is not always an appropriate one.
Example: Let $X = \mathscr{C}([0,1])$, $S = \mathbb{P}_n$ whose basis is $s_i = x^i$, $i = 0, 1, \ldots, n$, and the inner product is given with $\langle f, g \rangle = \int_0^1 f(x)g(x)dx$.

If we denote $G = (g_{ij})_{i,j=1}^n$, then

$$g_{ij} = \int_0^1 x^{i-1} x^{j-1} dx = \int_0^1 x^{i+j-2} dx = \frac{x^{i+j-1}}{i+j-1}\Big|_0^1 = \frac{1}{i+j-1}.$$

Thus,

$$G = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix},$$

which is the so called Hilbert matrix. Let us take for $\mathbf{b} = G \cdot (1,1,\ldots,1)^T$ and let $n = 10$. From

$$G \cdot \mathbf{x} = \mathbf{b} \Leftrightarrow G \cdot \mathbf{x} = G \cdot (1,1,\ldots,1)^T,$$

we can conclude that the exact solution is $\mathbf{x} = (1,1,\ldots,1)^T$, but using Octave we have that the approximate solution $\overline{\mathbf{x}}$ is

```
G=hilb(10);
b=G*ones(10,1);
xapr=G\b

xapr=
   1.00000
   1.00000
   1.00000
   1.00001
   0.99995
   1.00013
   0.99980
   1.00019
   0.99990
   1.00002
```

The error is

$$||\mathbf{x} - \overline{\mathbf{x}}||_2 = 3.286 \cdot 10^{-4},$$

which is not acceptable for such a "small" system.

Example 1.11

We are given points $(x_1,y_1)^T$, $(x_2,y_2)^T$, $\ldots$, $(x_m,y_m)^T$. Find a line $y = kx+n$ using the least squares method which best approximates the given data.

Solution:

First possibility:
We are trying to find

$$\min_{k,n} \sum_{i=1}^{m} (y_i - (kx_i + n))^2 = \min_{k,n} f(k,n).$$

Let us solve the system

$$\frac{\partial f}{\partial k} = 0, \quad \frac{\partial f}{\partial n} = 0.$$

We have

$$2\sum_{i=1}^{m} x_i(y_i - kx_i - n) = 0, \quad 2\sum_{i=1}^{m}(y_i - kx_i - n) = 0.$$

Further,

$$\sum_{i=1}^{m}(x_i y_i - kx_i^2 - nx_i) = 0, \quad \sum_{i=1}^{m}(y_i - kx_i - n) = 0.$$

Finally,

$$\sum_{i=1}^{m} x_i y_i - k\sum_{i=1}^{m} x_i^2 - n\sum_{i=1}^{m} x_i = 0, \quad \sum_{i=1}^{m} y_i - k\sum_{i=1}^{m} x_i - nm = 0,$$

which gives

$$\begin{bmatrix} \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \\ m & \sum_{i=1}^{m} x_i \end{bmatrix} \cdot \begin{bmatrix} n \\ k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} x_i y_i \\ \sum_{i=1}^{m} y_i \end{bmatrix}.$$

Second possibility:
Let $S = \mathscr{L}in\{1, x\}$. We have then

$$G = \begin{bmatrix} \langle \mathbf{1}, \mathbf{1} \rangle & \langle \mathbf{1}, \mathbf{x} \rangle \\ \langle \mathbf{x}, \mathbf{1} \rangle & \langle \mathbf{x}, \mathbf{x} \rangle \end{bmatrix} = \begin{bmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{bmatrix},$$

and

$$\mathbf{b} = \begin{bmatrix} \langle \mathbf{y}, \mathbf{1} \rangle \\ \langle \mathbf{y}, \mathbf{x} \rangle \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{bmatrix}.$$

Hence,

$$G\mathbf{a} = \mathbf{b} \Leftrightarrow \begin{bmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{bmatrix} \cdot \begin{bmatrix} n \\ k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{bmatrix}$$

which is the same system as we obtained above. Solving this system will give us the parameters k and n, which will determine the best approximant.

♣

Example 1.12

Let $f(x) = e^x$ be defined on $[-1, 1]$. Find $\tilde{p}_1 \in \mathbb{P}_1$ by the continuous least squares method.

Solution: We are looking for $\tilde{p}_1(x) = kx + n$, where $S = \mathscr{L}in\{1, x\}$ and $\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)dx$. Let us compute the inner products.

$$\langle 1, 1 \rangle = \int_{-1}^{1} dx = x\Big|_{-1}^{1} = 2,$$

$$\langle 1,x\rangle = \langle x,1\rangle = \int_{-1}^{1} x dx = 0,$$
$$\langle x,x\rangle = \int_{-1}^{1} x^2 dx = \frac{x^3}{3}\Big|_{-1}^{1} = \frac{2}{3},$$
$$\langle e^x,1\rangle = \int_{-1}^{1} e^x dx = e^x\Big|_{-1}^{1} = e - \frac{1}{e},$$
$$\langle e^x,x\rangle = \int_{-1}^{1} xe^x dx = xe^x\Big|_{-1}^{1} - \int_{-1}^{1} e^x dx = e + \frac{1}{e} - e + \frac{1}{e} = \frac{2}{e}.$$

Hence, we have the system

$$\begin{bmatrix} 2 & 0 \\ 0 & \frac{2}{3} \end{bmatrix} \cdot \begin{bmatrix} n \\ k \end{bmatrix} = \begin{bmatrix} e - \frac{1}{e} \\ \frac{2}{e} \end{bmatrix}$$

whose solutions are

$$n = \frac{e^2-1}{2e}, \quad k = \frac{3}{e}.$$

Thus,

$$\tilde{p}_1(x) = \frac{3}{e}x + \frac{e^2-1}{2e}.$$

♣

Example 1.13

Using the discrete least squares method approximate points $(-1,0)^T$, $(-1,1)^T$, $(0,1)^T$, $(1,2)^T$, $(1,3)^T$ *with a*

(a) polynomial from the space $\mathbb{P}_2$,

(b) polynomial from the space $\mathbb{P}_3$.

Solution:

(a) Let $p_2(x) = a_0 + a_1x + a_2x^2$ and $s_1 = 1, s_2 = x, s_3 = x^2$ be the basis of our space S. We have that

$$\mathbf{1} = (1,1,1,1,1), \quad \mathbf{x} = (-1,-1,0,1,1), \quad \mathbf{x^2} = (1,1,0,1,1), \quad \mathbf{y} = (0,1,1,2,3).$$

The inner products are

$$\langle \mathbf{1},\mathbf{1}\rangle = 5, \ \langle \mathbf{1},\mathbf{x}\rangle = 0, \ \left\langle \mathbf{1},\mathbf{x^2}\right\rangle = 4,$$
$$\langle \mathbf{x},\mathbf{x}\rangle = 4, \ \left\langle \mathbf{x},\mathbf{x^2}\right\rangle = 0, \ \left\langle \mathbf{x^2},\mathbf{x^2}\right\rangle = 4,$$
$$\langle \mathbf{y},\mathbf{1}\rangle = 7, \ \langle \mathbf{y},\mathbf{x}\rangle = 4, \ \left\langle \mathbf{y},\mathbf{x^2}\right\rangle = 6.$$

Hence, we have the system

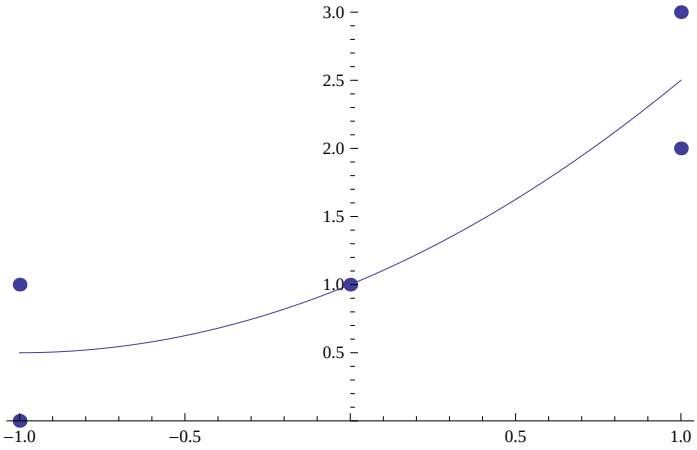
$$\begin{bmatrix} 5 & 0 & 4 \\ 0 & 4 & 0 \\ 4 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 4 \\ 6 \end{bmatrix},$$

whose solutions are

$$a_0 = 1, \quad a_1 = 1, \quad a_2 = \frac{1}{2}.$$

Thus,

$$p_2(x) = 1 + x + \frac{x^2}{2}.$$



(b) Let $p_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ and $s_1 = 1, s_2 = x, s_3 = x^2, s_4 = x^3$. We calculate the remaining inner products:

$$\left\langle \mathbf{1}, \mathbf{x^3} \right\rangle = 0, \ \left\langle \mathbf{x}, \mathbf{x^3} \right\rangle = 4, \ \left\langle \mathbf{x^2}, \mathbf{x^3} \right\rangle = 0, \ \left\langle \mathbf{x^3}, \mathbf{x^3} \right\rangle = 4, \ \left\langle \mathbf{y}, \mathbf{x^3} \right\rangle = 4.$$

By solving the system

$$\begin{bmatrix} 5 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 4 \\ 6 \\ 4 \end{bmatrix}$$

we obtain

$$a_0 = 1, \quad a_2 = \frac{1}{2}, \quad a_3 = \alpha \in \mathbb{R}, \quad a_1 = 1 - \alpha.$$

Hence,

$$p_3(x) = 1 + (1 - a)x + \frac{x^2}{2} + \alpha x^3, \ \ \alpha \in \mathbb{R}.$$

Remark 1.8

The solution in (b) is not unique. The reason for that is that the given data is not sampled from some function $f \in X$.

♣

Example 1.14

Approximate points $(1,0)^T, (2,2)^T, (1,2)^T, (3,2)^T, (1,1)^T, (2,1)^T, (2,0)^T, (3,0)^T, (3,1)^T$ with the least squares method in the space $S = \mathscr{L}in\{1, e^x\}$.

Solution: We want to find $y = a + be^x$. We have

$$\begin{aligned}\mathbf{1} &= (1,1,1,1,1,1,1,1,1)\\ \mathbf{e^x} &= (e,e^2,e,e^3,e,e^2,e^2,e^3,e^3)\\ \mathbf{y} &= (0,2,2,2,1,1,0,0,1)\end{aligned}$$

Let us compute the inner products

$$\begin{aligned}\langle \mathbf{1},\mathbf{1}\rangle = 9, \quad \langle \mathbf{1},\mathbf{e^x}\rangle &= 3e+3e^2+3e^3 = 3e(1+e+e^2)\\ \langle \mathbf{e^x},\mathbf{e^x}\rangle = 3e^2+3e^4+3e^6 &= 3e^2(1+e+e^2), \quad \langle \mathbf{y},\mathbf{1}\rangle = 9,\\ \langle \mathbf{y},\mathbf{e^x}\rangle &= 2e^2+2e+2e^3+e+e^2+e^3 = 3e(1+e+e^2).\end{aligned}$$

Thus, we have the system

$$\begin{bmatrix} 9 & 3e(1+e+e^2) \\ 3e(1+e+e^2) & 3e^2(1+e+e^2)\end{bmatrix} \cdot \begin{bmatrix} a \\ b\end{bmatrix} = \begin{bmatrix} 9 \\ 3e(1+e+e^2)\end{bmatrix},$$

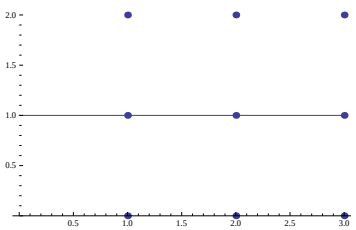
whose solution is

$$a = 1, \quad b = 0.$$

Hence,

$$y = 1$$

is the best approximant in the given space.



♣

Gram-Schmidt procedure is the standard procedure to compute an orthogonal basis for a given space S. However, for $S = \mathbb{P}_n$ we have a particular three-term recursive formula available to compute it.

Recursive formula:
Let $\{Q_0, Q_1, \ldots, Q_{n-1}\}$ be an orthogonal set of polynomials where $\deg(Q_i) = i$. Let us compute Q_n, where we want that $\langle Q_n, Q_i \rangle = 0$, $\forall i = 0, 1, \ldots, n-1$, and that the leading coefficient of Q_n is the same as the leading coefficient of Q_{n-1}. Thus,

$$Q_n(x) = (x + a_n)Q_{n-1}(x) + b_n Q_{n-2}(x) + \sum_{j=0}^{n-3} c_j Q_j(x).$$

We have, for $i = 0, 1, \ldots, n-3$,

$$0 = \langle Q_n, Q_i \rangle = \langle (x + a_n)Q_{n-1}, Q_i \rangle + \langle b_n Q_{n-2}, Q_i \rangle + \sum_{j=0}^{n-3} c_j \langle Q_j, Q_i \rangle = c_i ||Q_i||^2.$$

Thus,

$$c_i = 0, \quad i = 0, 1, \ldots, n-3.$$

Similarly,

$$\begin{aligned}
0 &= \langle Q_n, Q_{n-2} \rangle = \langle (x + a_n)Q_{n-1}, Q_{n-2} \rangle + \langle b_n Q_{n-2}, Q_{n-2} \rangle \\
&= \langle xQ_{n-1}, Q_{n-2} \rangle + \langle a_n Q_{n-1}, Q_{n-2} \rangle + b_n ||Q_{n-2}||^2 \\
\Rightarrow b_n &= -\frac{\langle Q_{n-1}, xQ_{n-2} \rangle}{||Q_{n-2}||^2}. \\
0 &= \langle Q_n, Q_{n-1} \rangle = \langle (x + a_n)Q_{n-1}, Q_{n-1} \rangle + \langle b_n Q_{n-2}, Q_{n-1} \rangle \\
&= \langle xQ_{n-1}, Q_{n-1} \rangle + a_n ||Q_{n-1}||^2 + \langle b_n Q_{n-2}, Q_{n-1} \rangle \\
\Rightarrow a_n &= -\frac{\langle xQ_{n-1}, Q_{n-1} \rangle}{||Q_{n-1}||^2}.
\end{aligned}$$

```
begin
    Q_0(x) = 1 ;
    Q_1(x) = x - <x,1> / ||Q_0||^2 ;
    k = 2,3,...,n
        a_k = - <xQ_{k-1}, Q_{k-1}> / ||Q_{k-1}||^2 ;
        b_k = - <Q_{k-1}, xQ_{k-2}> / ||Q_{k-2}||^2 ;
        Q_k(x) = (x + a_k)Q_{k-1}(x) + b_k Q_{k-2}(x) ;
end
```

Listing 1.2: Algorithm for computing an orthogonal set of polynomials

Example 1.15

Let $\langle f,g\rangle = \sum_{i=1}^{5} f(i)g(i)$.

(i) Compute the first three orthogonal polynomials.

(ii) Approximate the function $f(x) = \sin\frac{\pi x}{2}$ *with a parabola by the least squares method.*

Solution:

(i) We have $\mathbf{x} = (1,2,3,4,5)$. Let's compute Q_0, Q_1 and Q_2.

$$Q_0(x) = 1, \quad \mathbf{Q}_0 = (1,1,1,1,1), \quad ||\mathbf{Q}_0||^2 = 5.$$

$$Q_1(x) = x - \frac{15}{5} = x-3, \quad \mathbf{Q}_1 = (-2,-1,0,1,2), \quad ||\mathbf{Q}_1||^2 = 10.$$

$$a_2 = -\frac{\langle xQ_1, Q_1\rangle}{10} = -\frac{30}{10} = -3, \quad b_2 = -\frac{\langle x-3, x\rangle}{5} = -\frac{10}{5} = -2.$$

$$Q_2(x) = (x-3)Q_1(x) - 2Q_0(x) = (x-3)^2 - 2 = x^2 - 6x + 7,$$
$$\mathbf{Q}_2 = (2,-1,-2,-1,2), \quad ||\mathbf{Q}_2||^2 = 14.$$

(ii) By (i) we have

$$G = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 14 \end{bmatrix}.$$

We have that $\mathbf{f} = (1,0,-1,0,1)$ and thus

$$\langle \mathbf{f}, \mathbf{Q}_0\rangle = 1, \quad \langle \mathbf{f}, \mathbf{Q}_1\rangle = 0, \quad \langle \mathbf{f}, \mathbf{Q}_2\rangle = 6.$$

By solving the system $G\cdot\mathbf{a} = \mathbf{b}$ we get that

$$\alpha_0 = \frac{1}{5}, \ \alpha_1 = 0, \ \alpha_2 = \frac{3}{7}.$$

Hence,

$$p_2(x) = \frac{1}{5} + 0 + \frac{3}{7}(x^2 - 6x + 7) = \frac{3}{7}x^2 - \frac{18}{7}x + \frac{16}{5}.$$

♣

Example 1.16

We are given points $(-1,12)^T$, $(0,7)^T$, $(1,6)^T$, $(2,9)^T$. *Approximate these points with a parabola by the least squares method using orthogonal polynomials.*

Solution: We have $\mathbf{x} = (-1,0,1,2)$. Let us compute Q_0, Q_1 and Q_2.

$$Q_0(x) = 1, \quad \mathbf{Q}_0 = (1,1,1,1), \quad ||\mathbf{Q}_0||^2 = 4.$$

$$Q_1(x) = x - \frac{\langle x,1\rangle}{||\mathbf{Q}_0||^2} = x - \frac{1}{2}, \quad ||\mathbf{Q}_1|| = \frac{1}{2}(-3,-1,1,3), \quad ||\mathbf{Q}_1||^2 = 5.$$

$$a_2 = -\frac{\langle xQ_1, Q_1\rangle}{||\mathbf{Q}_1||^2} = -\frac{1}{2}, \quad b_2 = -\frac{\langle Q_1, xQ_0\rangle}{||\mathbf{Q}_0||^2} = -\frac{5}{4}.$$

$$Q_2(x) = \left(x - \frac{1}{2}\right)^2 - \frac{5}{4} = x^2 - x - 1, \quad \mathbf{Q}_2 = (1,-1,-1,1), \ ||\mathbf{Q}_2||^2 = 4.$$

We denote $\mathbf{f} = (12,7,6,9)$. Hence,

$$\langle \mathbf{f}, \mathbf{Q}_0\rangle = 34, \quad \langle \mathbf{f}, \mathbf{Q}_1\rangle = -5, \quad \langle \mathbf{f}, \mathbf{Q}_2\rangle = 8.$$

By solving the system

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 34 \\ -5 \\ 8 \end{bmatrix}$$

we obtain

$$\alpha_0 = \frac{17}{2}, \quad \alpha_1 = -1, \quad \alpha_2 = 2.$$

Thus,

$$p_2(x) = \frac{17}{2} - x + \frac{1}{2} + 2x^2 - 2x - 2 = 2x^2 - 3x + 7.$$

♣

Ordinary differential equations

2.1. Introduction

An ordinary differential equation (ODE) represents an equation which connects the independent variable x, dependent variable $y = y(x)$ and it's derivatives. We can write it as

$$F\left(x, y, y', \ldots, y^{(n)}\right) = 0.$$

Sometimes we have an explicit expression for the highest derivative as

$$y^{(n)} = f\left(x, y, \ldots, y^{(n-1)}\right) = 0.$$

The solution of such a DE is a function g for which

$$F\left(x, g(x), g'(x), g^{(n)}(x)\right) = 0.$$

The **order** of an ODE represents the highest derivative of the variable y which appears in the equation. The **solution** of an ODE is a n-parametric family of solutions $h(x; c_1, c_2, \ldots, c_n)$ where fixing the constants $c_1, c_2, \ldots, c_n$ gives us the so-called *particular solution*.

Example 2.1: DE with separable variables

If we can write the given DE in the form

$$g(x)dx = h(y)dy,$$

we say for that equation that it is a differential equation with separable variables. The idea of finding a solution is to integrate both sides of the equation, that is

$$\int g(x)dx = \int h(y)dy \;\Leftrightarrow\; G(x) = H(y) + C.$$

Sometimes we can explicitly express solution $y(x)$.

Example 2.2

Solve ODE $y'(x) = -2xy$.

Solution:

$$\begin{aligned} y' = -2xy &\Leftrightarrow \frac{dy}{dx} = -2xy \\ &\Leftrightarrow \frac{dy}{y} = -2xdx \\ &\Leftrightarrow \int \frac{dy}{y} = -2\int xdx \\ &\Leftrightarrow \ln|y| = -x^2 + \ln|C| \\ &\Leftrightarrow y(x) = De^{-x^2} \end{aligned}$$

♣

Let us consider now the situation with more than one DE. Suppose we have a system of DE of first order

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \ldots, y_n), \\ y_2' &= f_2(x, y_1, y_2, \ldots, y_n), \\ &\vdots \\ y_n' &= f_n(x, y_1, y_2, \ldots, y_n), \end{aligned}$$

where $f_1, f_2, \ldots, f_n$ are given and $y_1(x), y_2(x), \ldots, y_n(x)$ are the unknown functions. If we are given initial values

$$y_1(x_0) = y_{1,0},\ y_2(x_0) = y_{2,0},\ \ldots y_n(x_0) = y_{n,0},$$

then the initial value problem can be written as

$$Y'(x) = F(x, Y(x)), \quad Y(x_0) = Y_0,$$

where $Y = (y_1, y_2, \ldots, y_n)^T$ and $F = (f_1, f_2, \ldots, f_n)^T$.

2.2. Some simple numerical methods

Methods for solving DE can be partitioned with respect to several parameters:

1. Discrete vs. Continuous

   (a) *Discrete methods:* The interval $[a, b]$ on which we are solving the DE is separated into $a = x_0 < x_1 < \ldots < x_n = b$. We are looking for approximate values of $y(x_1), y(x_2), \ldots, y(x_n)$.

(b) *Continuous methods*: Now the solution lies in some approximation space S, e.g., space of (piecewise) polynomial functions. Some methods belonging to this class of methods are the *finite element method*, *collocation*, *isogeometric analysis* ...

2. Approximation of derivatives vs. Approximation of integrals

   (a) *Approximation of derivatives:* We use some approximation method to approximate

   $$y'(x) = \sum_{i=0}^{n-1} \alpha_i y(x_i).$$

   (b) *Approximation of integrals:* We integrate DE e.g. on the segment $[x_{n-1}, x_n]$, that is

   $$\int_{x_{n-1}}^{x_n} y'(x)dx = \int_{x_{n-1}}^{x_n} f(x, y(x))dx \;\Leftrightarrow\; y(x_n) - y(x_{n-1}) = \int_{x_{n-1}}^{x_n} f(x, y(x))dx.$$

   We use now some numerical method to compute the integral, e.g. we use one of the *Newton-Cotes methods* or *Gauss methods*.

3. One-step vs. Multi-step

   (a) *One-step methods:* For computing $y_n \approx y(x_n)$ we only use y_{n-1}.

   (b) *Multi-step methods:* For computing y_n we use $y_{n-1}, y_{n-2}, \ldots, y_{n-k}$, where $k \geq 1$.

4. Explicit vs. Implicit

   (a) *Explicit methods:* We can express y_n as $y_n = g(y_0, y_1, \ldots, y_{n-1})$.

   (b) *Implicit methods:* we only have the relation $y_n = h(y_0, y_1, \ldots, y_{n-1}, y_n)$.

Global and local error of a numerical method

The global error in the point x_n is the difference $||y_n - y(x_n)||$. Considering the local error is an approach which enables us to compute the global error. The local error in the point x_n is the difference $||y_n - y(x_n)||$ with the additional assumption that $y_{n-1} = y(x_{n-1})$. The global error is not the sum of local errors, but the approach considering local errors makes it easier for us to compute the global error.

Convergence

We want the numerical approximants to converge to the exact values as $\max_i \Delta x_i$ decreases, that is

$$\max_{0 \leq m \leq n} ||y(x_m) - y_m|| \xrightarrow[h \to 0]{} 0, \quad h := \max_i \Delta x_i.$$

If that holds for the solutions of all DE, then we say that this method is a convergent method. We say that a method is of order r if

$$\max_{0\leq m\leq n} ||y(x_m) - y_m|| = \mathcal{O}\left(h^r\right).$$

Remark 2.1

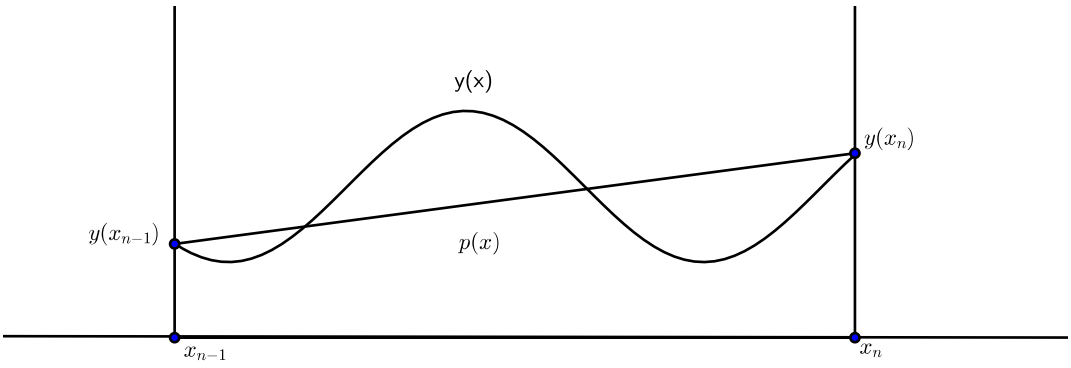
1. *The method is convergent if it is of order $r \geq 1$.*
2. *Considering initial value problems we usually require that $r \geq 4$.*
3. *Considering boundary value problems we are usually satisfied with $r \geq 4$.*

2.2.1 Euler methods

We have the initial value problem

$$y' = f(x,y), \quad y(x_0) = y_0. \tag{2.2.1}$$

For the approximation of the derivative of a function $y(x)$ we will use the derivative of a linear function in the point x. How to approximate $y'(x_{n-1})$ and $y'(x_n)$?



We have :

$$y'(x_{n-1}) \doteq p'(x_{n-1}) = \frac{y(x_n) - y(x_{n-1})}{x_n - x_{n-1}},$$
$$y'(x_n) \doteq p'(x_n) = \frac{y(x_n) - y(x_{n-1})}{x_n - x_{n-1}}.$$

Let us suppose that $\Delta x_i = h$, $\forall i$. Let us write the DE (2.2.1) in the point x_{n-1}:

$$\begin{aligned} & y'(x_{n-1}) = f(x_{n-1}, y(x_{n-1})) \\ \Rightarrow\ & y(x_n) - y(x_{n-1}) = hf(x_{n-1}, y(x_{n-1})) \\ \overset{y(x_i)\approx y_i}{\Rightarrow}\ & y_n = y_{n-1} + hf(x_{n-1}, y_{n-1}). \end{aligned}$$

The last expression represents the **Explicit Euler Method (EEM)**. Let us write the DE (2.2.1) in the point x_n:

$$\begin{aligned} & y'(x_n) = f(x_n, y(x_n)) \\ \Rightarrow\ & y(x_n) - y(x_{n-1}) = hf(x_n, y(x_n)) \\ \overset{y(x_i)\approx y_i}{\Rightarrow}\ & y_n = y_{n-1} + hf(x_n, y_n). \end{aligned}$$

The last expression represents the **Implicit Euler Method (IEM)**. Now let us consider the order of the EEM.

EEM: We compute $y(x_n) - y_n$ with the assumption that $y(x_{n-1}) = y_{n-1}$. The order of the local error is now computed as the following.

$$\begin{aligned} y_n &= y_{n-1} + hf(x_{n-1}, y_{n-1}) \\ y(x_n) &= y(x_{n-1}) + hy'(x_{n-1}) + \frac{h^2}{2} y''(x_{n-1}) + \dots \\ &= y_{n-1} + hf(x_{n-1}, y_{n-1}) + \mathcal{O}(h^2). \end{aligned}$$

Therefore

$$y(x_n) - y_n = \mathcal{O}(h^2).$$

We conclude, that the order of the local error is 2.

Remark 2.2

For initial value problems, the order of the global error of the method is one less than the order of the local error.

Hence, the order of EEM is equal to one. It is straightforward to verify that the IEM has also

order one.

The Taylor expansion of a function of two variables is given as

$$f(x+h,y+k) = f(x,y) + hf_x + kf_y + \frac{h^2}{2}f_{xx} + \frac{k^2}{2}f_{yy} + hkf_{xy} + \dots$$

Now, let us approach the point (x_n, y_n) from (x_{n-1}, y_{n-1}) in the direction given in the point at parameter $\frac{x_{n-1}+x_n}{2}$. Let us denote $x_{n-\frac{1}{2}} := x_{n-1} + \frac{h}{2}$ and $y_{n-\frac{1}{2}} := y_{n-1} + \frac{h}{2}f(x_{n-1}, y_{n-1})$. Now we have,

$$y_n = y_{n-1} + hf\left(x_{n-\frac{1}{2}}, y_{n-\frac{1}{2}}\right) \Leftrightarrow y_n = y_{n-1} + hf\left(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2}f\left(x_{n-1}, y_{n-1}\right)\right).$$

The last equation represents the **Improved Euler Method**. Let us consider the local order of this method. We assume that $y(x_{n-1}) = y_{n-1}$. Then

$$\begin{aligned}
y(x_n) &= y(x_{n-1}) + hy'(x_{n-1}) + \frac{h^2}{2}y''(x_{n-1}) + \mathcal{O}(h^3) \\
&= y(x_{n-1}) + hf(x_{n-1}, y_{n-1}) + \frac{h^2}{2}\left(f_x(x_{n-1}, y_{n-1}) + f(x_{n-1}, y_{n-1})f_y(x_{n-1}, y_{n-1})\right). \\
y_n &= y_{n-1} + h\left(f(x_{n-1}, y_{n-1}) + \frac{h}{2}f_x(x_{n-1}, y_{n-1}) + \frac{h}{2}f(x_{n-1}, y_{n-1})f_y(x_{n-1}, y_{n-1})\right) + \mathcal{O}(h^3). \\
\Rightarrow y(x_n) - y_n &= \mathcal{O}(h^3).
\end{aligned}$$

Example 2.3

We are given the following equation: $y'(x) = -50y + 100,\ y(0) = y_0$.

(a) Find the exact solution of the ODE;

(b) Solve the ODE using the explicit Euler method;

(c) Solve the ODE using the implicit Euler method.

In (b) *and* (c) *find the closed-form formula for* y_n*. Show for which* h *the approximation* y_n *converges to the exact solution when* $n \to \infty$.

Solution:

(a)

$$\begin{aligned}
& y'(x) = -50y + 100 \\
\Leftrightarrow\ & \frac{dy}{dx} = -50y + 100 \\
\Leftrightarrow\ & \frac{dy}{-50y + 100} = dx
\end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow \frac{1}{50}\cdot\frac{dy}{(-y+2)} = dx \quad \Big|\int \\
&\Leftrightarrow -\ln|2-y| = 50x+\ln|c| \\
&\Leftrightarrow \ln|2-y| = -50x-\ln|c| \\
&\Leftrightarrow 2-y = \mp\ln|c|e^{-50x} \\
&\Leftrightarrow y = 2+ke^{-50x}.
\end{aligned}$$

Further we have

$$y_0 = y(0) = 2+k \;\Rightarrow\; k = y_0-2.$$

Therefore

$$y(x) = (y_0-2)e^{-50x}+2 \xrightarrow[x\to\infty]{} 2.$$

(b) We have $f(x,y) = -50y+100$.

$$\begin{aligned}
y_n &= y_{n-1}+hf(x_{n-1},y_{n-1}) \\
&= y_{n-1}+h(-50y_{n-1}+100) \\
&= y_{n-1}(1-50h)+100h \\
&= [y_{n-2}(1-50h)+100h](1-50h)+100h \\
&= y_{n-2}(1-50h)^2+100h((1-50h)+1) \\
&= y_{n-3}(1-50h)^3+100h\left((1-50h)^2+(1-50h)+1\right) = \ldots \\
&= y_0(1-50h)^n+100h\sum_{k=0}^{n-1}(1-50h)^k \\
&= y_0(1-50h)^n+100h\cdot\frac{1-(1-50h)^n}{1-1+50h} \\
&= y_0(1-50h)^n+2\left(1-(1-50h)^n\right).
\end{aligned}$$

Therefore

$$y_n = (1-50h)^n(y_0-2)+2.$$

For which h will we have the convergence when $x\to\infty$? We observe that $x\to\infty \Leftrightarrow n\to\infty$, so to achieve that $y_n \xrightarrow[n\to\infty]{} 2$ we must have that $|1-50h|<1$. This gives $h\in\left(0,\frac{1}{25}\right)$.

(c) Like in (b) we have $f(x,y) = -50y+100$. Then

$$\begin{aligned}
y_n &= y_{n-1}+hf(x_n,y_n) \\
&= y_{n-1}+h(-50y_n+100).
\end{aligned}$$

We have

$$y_n(1+50h) = y_{n-1}+100h$$

and further

$$\begin{aligned} y_n &= \frac{1}{1+50h}\,(y_{n-1}+100h) = \\ &= \frac{1}{1+50h}\left(\frac{1}{1+50h}\,(y_{n-2}+100h)+100h\right) = \dots \\ &= \frac{1}{(1+50h)^n}(y_0-2)+2. \end{aligned}$$

When $n \to \infty$ we need that $\frac{1}{(1+50h)^n} \to 0$. Thus, $|1+50h| > 1$ which holds for all $h > 0$.

♣

2.3. Trapezoidal method

Instead of approximating the derivative like in the case of Euler methods, the idea here is to integrate equation $y'(x) = f(x,y)$ on $[x_{n-1}, x_n]$. With the assumption that $y_n \approx y(x_n)$ and $y_{n-1} \approx y(x_{n-1})$ we have that

$$y_n - y_{n-1} \approx \int_{x_{n-1}}^{x_n} f(x, y(x))dx.$$

We have to use now a numerical method for computing the integral. We will use the so-called trapezoidal quadrature rule, that is

$$\int_{x_{n-1}}^{x_n} g(x)dx \approx \frac{h}{2}(g(x_{n-1}) + g(x_n)).$$

Using this, we get

$$y_n = y_{n-1} + \frac{h}{2}(f(x_{n-1}, y_{n-1}) + f(x_n, y_n)).$$

The last formula represents the **trapezoidal method**. With the assumption that $y_{n-1} = y(x_{n-1})$, let us compute the order of this method.

$$\begin{aligned} y(x_n) &= y(x_{n-1}) + hy'(x_{n-1}) + \frac{h^2}{2}y''(x_{n-1}) + \mathcal{O}(h^3) \\ &= y(x_{n-1}) + hf(x_{n-1}, y(x_{n-1})) + \frac{h^2}{2}(f_x + f_y f) + \mathcal{O}(h^3) \\ &= y_{n-1} + hf(x_{n-1}, y_{n-1}) + \frac{h^2}{2}(f_x + f_y f) + \mathcal{O}(h^3). \end{aligned}$$

Further

$$y_n = y_{n-1} + \frac{h}{2}f(x_{n-1}, y_{n-1}) + \frac{h}{2}\left(f(x_{n-1}, y_{n-1}) + hf_x + f_y\left(\frac{h}{2}(2f + \mathcal{O}(h))\right)\right) + \mathcal{O}(h^3).$$

Therefore

$$y(x_n) - y_n = \mathcal{O}(h^3).$$

Example 2.4

Compute the approximation y_1 for equation $y' = x^2 - y^2$, $y(x_0) = y_0$, using the trapezoidal method.

Solution:

$$\begin{aligned} y_1 &= y_0 + \frac{h}{2}\left(f(x_0,y_0) + f(x_1,y_1)\right) \\ &= y_0 + \frac{h}{2}\left(x_0^2 - y_0^2 + x_1^2 - y_1^2\right). \end{aligned}$$

Then

$$hy_1^2 + 2y_1 - \left(2y_0 + hx_0^2 - hy_0^2 + hx_1^2\right) = 0.$$

It follows

$$y_1^{\pm} = \frac{-1 \pm \sqrt{1 + 2hy_0 + h^2x_0^2 - h^2y_0^2 + h^2x_1^2}}{h}.$$

Since $\sqrt{1+x} = 1 + \frac{x}{2} + \mathcal{O}(x^2)$, we have that

$$y_1^{\pm} = \frac{-1 \pm \left(1 + hy_0 - \frac{h^2}{2}\left(x_0^2 + y_0^2 + x_1^2\right)\right)}{h} = \frac{-1 \pm \left(1 + hy_0 + \mathcal{O}(h^2)\right)}{h}.$$

Then

$$\begin{aligned} y_1^+ &= \frac{hy_0 + \mathcal{O}(h^2)}{h} = y_0 + \mathcal{O}(h) \xrightarrow{h\to 0} y_0, \\ y_1^- &= \frac{-2 - hy_0 + \mathcal{O}(h^2)}{h} \xrightarrow{h\to 0} -\infty. \end{aligned}$$

Hence, the solution is $y_1 = y_1^+$.

♣

2.4. One-step methods

All the methods considered above belong to a more general class of methods. The idea now is to find the values of the function f (directions) in m different points and to determine the final direction as a particular linear combination of the obtained directions.

So we first compute m coefficients

$$k_i = hf\left(x_{n-1} + h\alpha_i, y_{n-1} + \sum_{j=1}^{m} \beta_{ij} k_j\right), \quad i = 1, 2, \ldots, m,$$

and build the next approximation y_n using the linear combination of coefficients k_i:

$$y_n = y_{n-1} + \sum_{i=1}^{m} \gamma_i k_i.$$

We have to determine parameters $\alpha_i,\ \gamma_i$ and β_{ij} in such a way, that the order of the method is as high as possible. The obtained methods are called *Runge-Kutta methods*. Parameter m represents the degree of the method (note that degree is not the same as the order). It can be proved that if we want the highest possible order, then we must have

$$\alpha_i = \sum_{j=1}^{m} \beta_{ij}, \quad \sum_{i=1}^{m} \gamma_i = 1.$$

We have the following special cases:

α_1	β_{11}	β_{12}	$\ldots$	β_{1m}
α_2	β_{21}	β_{22}	$\ldots$	β_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
α_m	β_{m1}	β_{m2}	$\ldots$	β_{mm}
	γ_1	γ_2	$\ldots$	γ_m

Table 2.1: Butcher tableau

- Explicit Runge-Kutta method (ERK): $\beta_{ij} = 0,\ j \geq i$
- Diagonally implicit Runge-Kutta method (DIRK): $\beta_{ij} = 0,\ j > i$
- Implicit Runge-Kutta method (IRK): arbitrary β_{ij}

Examples:

(1) $m = 1$: $\begin{array}{c|c} \alpha_1 & \beta_{11} \\ \hline & \gamma_1 \end{array}$

(a) $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \Rightarrow \left.\begin{array}{ll} k_1 &= hf(x_{n-1}, y_{n-1}) \\ y_n &= y_{n-1} + k_1 \end{array}\right\}$ EEM

(b) $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \Rightarrow \left.\begin{array}{ll} k_1 &= hf(x_n, y_{n-1} + k_1) \\ y_n &= y_{n-1} + k_1 \end{array}\right\}$ IEM

(2) $m=2$:

$$\begin{array}{c|cc} \alpha_1 & \beta_{11} & \beta_{12} \\ \alpha_2 & \beta_{21} & \beta_{22} \\ \hline & \gamma_1 & \gamma_2 \end{array}$$

(a)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \Rightarrow \left.\begin{aligned} k_1 &= hf(x_{n-1},y_{n-1}) \\ k_2 &= hf\left(x_{n-1}+\tfrac{h}{2}, y_{n-1}+\tfrac{k1}{2}\right) \\ y_n &= y_{n-1}+k_2 \end{aligned}\right\} \text{Improved EM}$$

(b)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \Rightarrow \left.\begin{aligned} k_1 &= hf(x_{n-1},y_{n-1}) \\ k_2 &= hf\left(x_n, y_{n-1}+\tfrac{k1}{2}+\tfrac{k_2}{2}\right) \\ y_n &= y_{n-1}+\tfrac{1}{2}(k_1+k_2) \end{aligned}\right\} \text{Trapezoidal method}$$

(3) Arbitrary explicit Runge-Kutta method of degree 2:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \beta & 0 \\ \hline & \gamma_1 & \gamma_2 \end{array} \Rightarrow \left.\begin{aligned} k_1 &= hf(x_{n-1},y_{n-1}) \\ k_2 &= hf\left(x_{n-1}+\alpha h, y_{n-1}+\beta k_1\right) \\ y_n &= y_{n-1}+\gamma_1 k_1+\gamma_2 k_2 \end{aligned}\right\}$$

The goal is to have as high as possible order of the method. We want to compute $y(x_n)-y_n = \mathcal{O}(h^{\square})$ with the assumption that $y(x_{n-1})=y_{n-1}$. Using Taylor's expansion we have

$$y(x_n) = y_{n-1}+hf+\frac{h^2}{2}(f_x+ff_y)+\frac{h^3}{6}(f_{xx}+2f_{xy}f+f_{yy}f^2+f_y(f_x+f_y))+\mathcal{O}(h^4).$$

Let us compute k_2 and y_n. We will use the notation $f=f(x_{n-1},y_{n-1})$.

$$k_2 = h\left(f+\alpha h f_x+\beta h f f_y+\frac{1}{2}\left(\alpha^2h^2f_{xx}+\beta^2h^2f^2f_{yy}+2f_{xy}\alpha\beta h^2 f\right)\right)+\mathcal{O}(h^4),$$

$$y_n = y_{n-1}+h(\gamma_1 f+\gamma_2 f)+h^2(\alpha f_x+\beta f f_y)\gamma_2+\frac{h^3}{2}\left(\alpha^2 f_{xx}+\beta^2 f^2 f_{yy}+2f_{xy}\alpha\beta f\right)\gamma_2+\mathcal{O}(h^4).$$

Let us look at $y(x_n)-y_n$ and try to eliminate as many terms in the expression as possible.

$$\begin{aligned} &1: \checkmark \\ &h: \gamma_1+\gamma_2=1 \Rightarrow \gamma_1 = 1-\gamma_2 \\ &h^2: \alpha = \frac{1}{2\gamma_2},\ \beta=\frac{1}{2\gamma_2} \\ &h^3: \text{cannot be eliminated} \end{aligned}$$

We conclude,

$$\gamma_2 \in \mathbb{R},\ \gamma_1 = 1-\gamma_2,\ \alpha=\beta=\frac{1}{2\gamma_2}.$$

For example, when $\gamma_2=1$ we have the improved Euler method and for $\gamma_2=\frac{1}{2}$ we have the so-called Heun's method.

(4) Example of an implicit method of degree 2 and order 4 (Hammer-Hollingsworth's method)

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

(5) The most often used Runge-Kutta method is one of the explicit methods of order and degree 4

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array} \Rightarrow \begin{cases} k_1 &= hf(x_{n-1}, y_{n-1}) \\ k_2 &= hf\left(x_{n-1}+\frac{1}{2}h, y_{n-1}+\frac{1}{2}k_1\right) \\ k_3 &= hf\left(x_{n-1}+\frac{1}{2}h, y_{n-1}+\frac{1}{2}k_2\right) \\ k_4 &= hf\left(x_{n-1}+h, y_{n-1}+k_3\right) \\ y_n &= y_{n-1}+\frac{1}{6}k_1+\frac{2}{6}k_2+\frac{2}{6}k_3+\frac{1}{6}k_4 \end{cases}$$

Example 2.5

Compute $y_1 \approx y(0.1)$ for the equation $y' = -y + 5e^x \sin x,\ y(0) = 1,\ h = 0.1$.

Solution:

$$k_1 = 0.1 \cdot (-1-0) = -0.1$$
$$k_2 = 0.1 \cdot f(0.05, 0.95) = -0.12127$$
$$k_3 = 0.1 \cdot f(0.05, 0.93936) = -0.12021$$
$$k_4 = 0.1 \cdot f(0.1, 0.87979) = -0.14315$$
$$y_1 = 1 - \frac{0.1}{6} - \frac{0.12127}{3} - \frac{0.12021}{3} - \frac{0.14315}{6} = 0.87898$$

♣

Example 2.6

With the method given in the Butcher tableau bellow solve the differential equation $y' = \lambda y,\ y(0) = y_0$.

1. *First compute y_n explicitly.*
2. *For which stepsize h will y_n $(n \to \infty)$ behave as the exact solution for $\lambda = -4$.*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array}$$

Solution: The method is given with the following

$$\begin{aligned}
k_1 &= hf(x_{n-1}, y_{n-1}) \\
k_2 &= hf\left(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{k_1}{2}\right) \\
k_3 &= hf(x_{n-1} + h, y_{n-1} - k_1 + 2k_2) \\
y_n &= y_{n-1} + \frac{k_1}{6} + \frac{4k_2}{6} + \frac{k_3}{6}.
\end{aligned}$$

For $f(x,y) = \lambda y$, we get:

$$\begin{aligned}
k_1 &= h\lambda y_{n-1} \\
k_2 &= h\lambda\left(y_{n-1} + \frac{h}{2}\lambda y_{n-1}\right) = h\lambda y_{n-1}\left(1 + \frac{1}{2}h\lambda\right) \\
k_3 &= h\lambda\left(y_{n-1} - h\lambda y_{n-1} + 2h\lambda y_{n-1}\left(1 + \frac{1}{2}h\lambda\right)\right) \\
&= h\lambda y_{n-1}\left(1 - h\lambda + 2h\lambda\left(1 + \frac{1}{2}h\lambda\right)\right) \\
&= h\lambda y_{n-1}(1 + h\lambda + (h\lambda)^2)
\end{aligned}$$

$$\begin{aligned}
y_n &= y_{n-1} + \frac{1}{6}h\lambda y_{n-1}\left(1 + 4\left(1 + \frac{1}{2}h\lambda\right) + \left(1 + h\lambda + (h\lambda)^2\right)\right) \\
&= y_{n-1} + \frac{1}{6}h\lambda y_{n-1}\left(1 + 4 + 2h\lambda + 1 + h\lambda + h^2\lambda^2\right) \\
&= y_{n-1}\underbrace{\left(1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{6}\right)}_{\Phi(h,\lambda)} \\
&= \dots \\
&= y_0 \cdot [\Phi(h,\lambda)]^n.
\end{aligned}$$

We can easily compute the exact solution of the differential equation by writing it as $\frac{dy}{y} = \lambda dx$, where, after integration, we get that

$$y(x) = ce^{\lambda x}.$$

Since $y(0) = y_0$, we have that $c = y_0$, so

$$y(x) = y_0 e^{\lambda x}$$

is the exact solution of our given equation. For $\lambda = -4$ we have that $y(x) = y_0 e^{-4x}$. We have that $y(x) \to 0$ when $x \to \infty$. Thus the question is for which h does $y_n \to 0$ when $n \to \infty$. Since,

$y_n = y_0\,[\Phi(h,-4)]^n$, we conclude that the condition is satisfied when

$$|\Phi(h,-4)| < 1 \;\Leftrightarrow\; \left|1-4h+8h^2-\frac{32}{3}h^3\right| < 1 \;\Leftrightarrow\; h \in (0,0.6281)\,.$$

♣

Example 2.7

We are given the equation $y''-y'y^2+y=0, y(0)=1, y'(0)=0$, and the stepsize $h=0.2$. Transform this DE of second order to a system of two equations of first order. Use the RK method of order 4 and compute the approximation $y_1 \approx y(0.2)$.

Solution: Let us use a substitution $z=y'$, then $z'=y''=zy^2-y$. With this we get that our *DE* can be represented as

$$\begin{cases} y' &= z, & y(0)=1 \\ z' &= zy^2-y, & z(0)=0 \end{cases}$$

We can write this in the matrix form

$$Y=\begin{pmatrix} y \\ z \end{pmatrix},\; Y'=F(x,Y)=\begin{pmatrix} z \\ zy^2-y \end{pmatrix},\; Y(0)=\begin{pmatrix} 1 \\ 0 \end{pmatrix}=Y_0$$

We have

$$\begin{pmatrix} k_1 \\ \ell_1 \end{pmatrix} = hF(x_0,Y_0) = 0.2\cdot\begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.2 \end{pmatrix}$$
$$\begin{pmatrix} k_2 \\ \ell_2 \end{pmatrix} = hF\left(x_0+\frac{h}{2}, Y_0+\frac{1}{2}\begin{pmatrix} k_1 \\ \ell_1 \end{pmatrix}\right) = 0.2F\left(0.1,\begin{pmatrix} 1 \\ -0.1 \end{pmatrix}\right) = 0.2\cdot\begin{pmatrix} -0.1 \\ -1.1 \end{pmatrix} = \begin{pmatrix} -0.02 \\ -0.22 \end{pmatrix}$$
$$\begin{pmatrix} k_3 \\ \ell_3 \end{pmatrix} = \begin{pmatrix} -0.022 \\ -0.219562 \end{pmatrix}$$
$$\begin{pmatrix} k_4 \\ \ell_4 \end{pmatrix} = \begin{pmatrix} -0.04391 \\ 0.2376 \end{pmatrix}.$$

Therefore

$$Y_1=\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} y_0+\frac{1}{6}k_1+\frac{2}{6}k_2+\frac{2}{6}k_3+\frac{1}{6}k_4 \\ z_0+\frac{1}{6}\ell_1+\frac{2}{6}\ell_2+\frac{2}{6}\ell_3+\frac{1}{6}\ell_4 \end{pmatrix} = \begin{pmatrix} 0.97868\ldots \\ -0.219454\ldots \end{pmatrix}.$$

♣

2.4.1 Nested methods

One of the advantages of one-step methods is the fact that we can relatively easy adapt the stepsize h. Methods, which allow such generalization are called nested Runge-Kutta methods. The idea of such methods is to add on each step the value for the so-called *estimator*.

$$\begin{array}{c|cccc} \alpha_1 & \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \alpha_2 & \beta_{21} & \beta_{22} & \dots & \beta_{2m} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \alpha_m & \beta_{m1} & \beta_{m2} & \dots & \beta_{mm} \\ \hline & \gamma_1 & \gamma_2 & \dots & \gamma_m \\ & \tilde{\gamma}_1 & \tilde{\gamma}_2 & \dots & \tilde{\gamma}_m \end{array}$$

Remark 2.3

Explicit methods of degree m can be of order $r \leq m$. For $m \geq 5$ we can have $r \leq m-1$.

We usually use that the original Runge-Kutta method is of order r and the estimator is of order $r' = r \pm 1$. The method is then labelled as the pair (r, r').

Known methods:

- Runge-Kutta-Fehlberg's method (4,5) is a method of degree $m = 6$. The fundamental method is of order 4 and the estimator is of order 5

$$y_n = y_{n-1} + \sum_{i=1}^{6} \gamma_i k_i, \quad \tilde{y_n} = y_{n-1} + \sum_{i=1}^{6} \tilde{\gamma}_i k_i.$$

  Error is then

$$y_n - \tilde{y_n} = \sum_{i=1}^{6} (\gamma_i - \tilde{\gamma}_i) k_i.$$

  Usually we choose

$$h_{new} = qh, \quad q = \left(\frac{\varepsilon h}{2|y_n - \tilde{y_n}|} \right)^{\frac{1}{5}}.$$

- Dormand-Prince methods. The best known is the (5,4)-method.

2.5. Multi-step methods

We compute y_n with help of $y_{n-1}, y_{n-2}, \dots, y_{n-k}$.

Advantages: Usually less function evaluations are needed to achieve the same accuracy.
Disadvantages: It's difficult to adjust the stepsize h. We need a special way to compute the initial k approximants. They are less stable.

The idea is similar as the one used for one-step methods. We again distinguish methods depending on whether the numerical approximation for the derivative (BDF methods) or for the integral of the function (Adams methods, Milne's methods) is used.

2.5.1 Adams methods

We integrate the DE $y'(x) = f(x, y(x))$ from x_{n-1} to x_n:

$$\int_{x_{n-1}}^{x_n} y'(x)dx = \int_{x_{n-1}}^{x_n} f(x, y(x)) \Leftrightarrow y_n - y_{n-1} = \int_{x_{n-1}}^{x_n} f(x, y(x))$$

Now we have to calculate the approximation for $\int_{x_{n-1}}^{x_n} f(x, y(x))$ and we will use an interpolating polynomial $p(x)$ which interpolates the function f in the points $x_{n-k}, x_{n-k+1}, \ldots, x_{n-1}$. The points are distributet equidistantly:

$$p(x) = p(x_{n-1} + ht) = \sum_{i=0}^{k-1} (-1)^i \binom{-t}{i} \nabla^i f_{n-1}$$

Remark 2.4

- $x = x_{n-1} + ht \longrightarrow t = \frac{1}{h}(x - x_{n-1})$
- $\nabla^i f_k = \nabla^{i-1} f - \nabla^{i-1} f_{k-1},\ \nabla^0 f_k = f_k,\ f_k = f(x_k, y(x_k))$
- $\binom{-t}{i} = \frac{1}{i!}(-t-0)(-t-1)(-t-2)\ldots(-t-(i-1))$

From

$$\int_{x_{n-1}}^{x_n} f(x, y(x)) \approx \int_{x_{n-1}}^{x_n} p(x)dx = \int_0^1 \sum_{i=0}^{k-1} (-1)^i \binom{-t}{i} \nabla^i f_{n-1} h dt$$

we can derive the explicit Adams method (Adams-Bashforth method):

$$y_n = y_{n-1} + h \sum_{i=0}^{k-1} \gamma_i \nabla^i f_{n-1},\ \gamma_i = (-1)^i \int_0^1 \binom{-t}{i} dt.$$

Example 2.8

Determine the methods for $k \leq 3$.

Solution:

- $k = 1$:

$$\gamma_0 = (-1)^0 \int_0^1 \binom{-t}{0} dt = 1$$
$$y_n = y_{n-1} + h\gamma_0 \nabla^0 f_{n-1}$$
$$= y_{n-1} + h f_{n-1}$$

$$y_n = y_{n-1} + h f(x_{n-1}, y(x_{n-1}))\ \text{EEM}$$

- $k=2$:

$$\gamma_1 = (-1)^1 \int_0^1 \binom{-t}{1} dt = \frac{1}{2}$$
$$y_n = y_{n-1} + h\left(\gamma_0 \nabla^0 f_{n-1} + \gamma_1 \nabla^1 f_{n-1}\right)$$
$$= y_{n-1} + h\left(f_{n-1} + \frac{1}{2}\left(f_{n-1} - f_{n-2}\right)\right)$$

$$y_n = y_{n-1} + h\left(\frac{3}{2} f_{n-1} - \frac{1}{2} f_{n-2}\right)$$

- $k=3$:

$$\gamma_2 = (-1)^2 \int_0^1 \binom{-t}{2} dt = \frac{5}{12}$$
$$y_n = y_{n-1} + h\left(\gamma_0 \nabla^0 f_{n-1} + \gamma_1 \nabla^1 f_{n-1} + \gamma_2 \nabla^2 f_{n-1}\right)$$
$$= y_{n-1} + h\left(f_{n-1} + \frac{1}{2}\left(f_{n-1} - f_{n-2}\right) + \frac{5}{12}\left(f_{n-1} - 2f_{n-2} + f_{n-3}\right)\right)$$

$$y_n = y_{n-1} + h\left(\frac{17}{12} f_{n-1} - \frac{4}{3} f_{n-2} + \frac{5}{12} f_{n-3}\right)$$

♣

Theorem 2.1

For $m = 1, 2, \ldots$ we have that

$$\gamma_0 = 1,\ \gamma_m = 1 - \frac{1}{2}\gamma_{m-1} - \frac{1}{3}\gamma_{m-2} - \ldots - \frac{1}{m+1}\gamma_0.$$

There are also the implicit methods, so-called Adams-Mouton methods, for which holds

$$p \text{ additionally interpolates the function in } x_n \Rightarrow p(x) = p(x_n + ht) = \sum_{i=0}^{k} (-1)^i \binom{-t}{i} \nabla^i f_n$$

The method is given with

$$y_n = y_{n+1} + h \sum_{i=0}^{k} \gamma_i^* \nabla^i f_n,\ \gamma_i^* = (-1)^i \int_0^1 \binom{-t+1}{i} dt.$$

Also,

$$\gamma_i^* = \gamma_i - \gamma_{i-1},\ i \geq 1,\ \ \gamma_0^* = 1.$$

Example 2.9

Using the above relation between γ_i and γ_i^*, compute the methods for $k = 1,2,3$. We omit the solution of this problem.

2.6. General linear multi-step methods

The general form of these methods is

$$\sum_{i=0}^{k} \alpha_i y_{n-i} + h \sum_{i=0}^{k} \beta_i f_{n-i} = 0.$$

Since the equation is homogeneous we can assume that $\alpha_0 = -1$. The method is explicit if $\beta_0 = 0$. Let us define two **generating polynomials**:

$$\rho(\xi) = \sum_{i=0}^{k} \alpha_i \xi^{k-i}, \quad \sigma(\xi) = \sum_{i=0}^{k} \beta_i \xi^{k-i}.$$

Theorem 2.2

The linear k-step method is of order r if and only if

$$\rho(1+\xi) + \ln(1+\xi)\sigma(1+\xi) = c_{r+1}\xi^{r+1} + \mathcal{O}(\xi^{r+2}), \quad c_{r+1} \neq 0. \tag{2.6.1}$$

Remark 2.5

For $r \geq 0$ we evaluate (2.6.1) at $\xi = 0$ and obtain

$$\rho(1) = 0. \tag{2.6.2}$$

Moreover, for $r \geq 1$ we can derive (2.6.1) first and then evaluate it at $\xi = 0$. Then

$$\rho'(1) + \sigma(0) = 0. \tag{2.6.3}$$

Conditions (2.6.2) and (2.6.3) guaranty the consistency of the method.

The equation (2.6.1) enables us to compute the polynomial σ if we know ρ and vice versa. Let's recall:

$$\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} \pm \dots$$
$$\frac{1}{\ln(1+z)} = \frac{1}{2} + \frac{1}{z} - \frac{z}{12} + \frac{z^2}{24} - \frac{19z^3}{720} \pm \dots$$

Example 2.10

Compute the generic polynomials for the Adams method.

Solution: We have $y_n = y_{n-1} + h\sum_{i=0}^{k} \beta_i^{(k)} f_{n-i}$. If we move all the variables to one side of the equation we get $-y_n + y_{n-1} + h\sum_{i=0}^{k} \beta_i^{(k)} f_{n-i} = 0$. So, we conclude

$$\rho(\xi) = -\xi^k + \xi^{k-1} = \xi^{k-1}(1-\xi).$$

Let's compute the polynomials for $k = 2$ when we have an explicit method:

$$\sigma(1+z) = -\frac{\rho(1+z)}{\ln(1+z)} + \mathcal{O}(z^2).$$

We have

$$\begin{aligned}\sigma(1+z) &= -\left(\frac{1}{2} + \frac{1}{z} - \frac{z}{12} + \frac{z^2}{24} - \frac{19z^3}{720} \pm \dots\right) \cdot (1+z) \cdot (-z) + \mathcal{O}(z^2) \\ &= (z+z^2)\left(\frac{1}{2} + \frac{1}{z} - \frac{z}{12} + \frac{z^2}{24} - \frac{19z^3}{720} \pm \dots\right) + \mathcal{O}(z^2) \\ &= 1 + z + \frac{1}{2}z + \mathcal{O}(z^2) = 1 + \frac{3}{2}z\end{aligned}$$

thus

$$\sigma(\xi) = 1 + \frac{3}{2}(\xi - 1) = -\frac{1}{2} + \frac{3}{2}\xi$$

The method is:

$$y_n = y_{n-1} + h\left(-\frac{1}{2}f_{n-2} + \frac{3}{2}f_{n-1}\right)$$

Now let's determine the implicit method for $k = 2$. In this case σ is a quadratic polynomial.

$$\sigma(1+z) = -\frac{\rho(1+z)}{\ln(1+z)} + \mathcal{O}(z^3)$$

Similarly as before we compute $\sigma(1+z)$ till the quadratic degree, that is

$$\sigma(1+z) = 1 + z + \frac{1}{2}z^2 - \frac{z^2}{12} + \mathcal{O}(z^3) = 1 + \frac{3}{2}z + \frac{5}{12}z^2$$

Hence,

$$\sigma(\xi) = 1 + \frac{3}{2}\xi - \frac{3}{2} + \frac{5}{12}\xi^2 - \frac{5}{6}\xi + \frac{5}{12} = -\frac{1}{12} + \frac{2}{3}\xi + \frac{5}{12}\xi^2.$$

The method is given with

$$y_n = y_{n-1} + h\left(-\frac{1}{12}f_{n-2} + \frac{2}{3}f_{n-1} + \frac{5}{12}f_n\right).$$

♣

If we want that the method is stable, it has to be at least stable for the simplest equation $y' = 0$. The numerical method

$$\sum_{i=0}^{k} \alpha_i \gamma_{n-i} + h \sum_{i=0}^{n} \beta_i f_{n-i} = 0$$

is simplified to a difference equation

$$\sum_{i=0}^{k} \alpha_i y_{n-i} = 0.$$

We use the substitution $y_k = \lambda^k$ and get

$$\begin{aligned} \sum_{i=0}^{k} \alpha_i \lambda^{n-i} = 0 \Rightarrow \alpha_0 \lambda^n + \alpha_1 \lambda^{n-1} + \ldots + \alpha_k \lambda^{n-k} = 0 \quad \Big| : \lambda^{n-k} \neq 0 \\ \Rightarrow \alpha_0 \lambda^k + \alpha_1 \lambda^{k-1} + \ldots + \alpha_k = 0 \\ \Rightarrow \rho(\lambda) = 0 \end{aligned}$$

The general solution is of the form

$$y_n = \sum_{i=1}^{m} p_i(n) \lambda_i^n, \; deg(p_i) = m_i - 1, \; \sum_{i=1}^{m} m_i = k,$$

where m_i represents the multiplicity of the zeros λ_i, $i = 1, 2, \ldots, m$.
We would like that $y_n \to c < \infty$ when $n \to \infty$. What has to hold for the zeros λ_i? If $|\lambda_i| < 1$ then the requirement is fulfilled. If $|\lambda_i| = 1$, then λ_i has to be a simple zero.

Definition 2.1

The method is zero-stable if

(i) $\rho(z) = 0 \Rightarrow |z| \leq 1$,

(ii) $\rho(z) = \rho'(z) = 0 \Rightarrow |z| < 1$.

Theorem 2.3

The linear multi-step method converges if it is zero-stable and consistent, i.e., $\rho(1) = 0$, $\rho'(1) + \sigma(1) = 0$.

Example 2.11

We have given the following method

$$y_n = \frac{1}{2}\left(y_{n-1} + 2y_{n-2} - y_{n-3}\right) + \frac{h}{6}\left(13f_{n-1} - 8f_{n-2} + f_{n-3}\right).$$

(a) Compute the generating polynomials.

(b) Compute the leading coefficient of the error.

(c) What is the order of the method?

(d) Is the method zero-stable?

(e) Is it convergent?

Solution:

(a)

$$\rho(\xi) = -\xi^3 + \frac{1}{2}\xi^2 + \xi - \frac{1}{2}, \quad \sigma(\xi) = \frac{13}{6}\xi^2 - \frac{4}{3}\xi + \frac{1}{6}.$$

(b)

$$\rho(1+z) = -(1+z)^3 + \frac{1}{2}(1+z)^2 + (1+z) - \frac{1}{2} = -z^3 - \frac{5}{2}z^2 - z = -z\left(z^2 - \frac{5}{2}z + 1\right),$$
$$\sigma(1+z) = \frac{13}{6}(1+z)^2 - \frac{4}{3}(1+z) + \frac{1}{6} = \frac{13}{6}z^2 + 3z + 1.$$

We have

$$\rho(1+z) + \ln(1+z)\sigma(1+z) = -z^3 - \frac{5}{2}z^2 - z + \left(z - \frac{z^2}{2} + \frac{z^3}{3} \pm \ldots\right) \cdot \left(\frac{13}{6}z^2 + 3z + 1\right).$$

If we consider all the elements in the sum, the coefficients at z, z^2 and z^3 are equal to 0. Further, the coefficient at z^4 equals $-\frac{1}{3} = c_4 = c_{r+1}$. So, the leading coefficient of the error is $-\frac{1}{3}$.

(c) From the above we conclude that the order of the method is $r = 3$. Since $r \geq 1$, we have consistency.

(d) The roots of the polynomial ρ:

$$\rho(\xi) = 0 \Leftrightarrow -\xi(\xi^2 - 1) + \frac{1}{2}(\xi^2 - 1) = 0 \Leftrightarrow \left(\frac{1}{2} - \xi\right)(\xi - 1)(\xi + 1) = 0 \Leftrightarrow \xi_1 = \frac{1}{2}, \xi_{2,3} = \pm 1.$$

Since $|\xi_i| \leq 1$, $i = 1, 2, 3$, and all zeros are simple, we conclude that the method is zero-stable.

(e) Since it is consistent and zero-stable, the method is convergent.

♣

Example 2.12

Let $\rho(\xi) = \xi^3 - \xi^2 + \frac{1}{4}\xi - \frac{1}{4}$. Compute σ so that the order of the local error is at least 4 and that the method is explicit. Is the method zero-stable and/or convergent?

Solution: We want to determine $\sigma(1+z) = Az^2 + Bz + C$.

$$\rho(1+z) = (1+z)^3 - (1+z)^2 + \frac{1+z}{4} - \frac{1}{4} = z^3 + 2z^2 + \frac{5}{4}z$$

Since the method is of order 4, the coefficients with z, z^2 and z^3 in $\rho(1+z) + \ln(1+z)\sigma(1+z)$ have to be 0. So,

$$\begin{aligned}\rho(1+z) + \ln(1+z)\sigma(1+z) &= z^3 + 2z^2 + \frac{5}{4}z + \left(z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} \pm \dots\right)\left(Az^2 + Bz + C\right) \\ &= \left(\frac{5}{4} + C\right)z + \left(2 + B - \frac{1}{2}C\right)z^2 + \left(1 + A - \frac{1}{2}B + \frac{1}{3}C\right)z^3 + \dots\end{aligned}$$

Therefore

$$\frac{5}{4} + C = 0, \quad 2 + B - \frac{1}{2}C = 0, \quad 1 + A - \frac{1}{2}B + \frac{1}{3}C = 0.$$

By solving this system we get

$$A = -\frac{91}{48}, \quad B = -\frac{21}{8}, \quad C = -\frac{5}{4}.$$

Hence,

$$\sigma(1+z) = -\frac{91}{48}z^2 - \frac{21}{8}z - \frac{5}{4} \iff \sigma(\xi) = -\frac{91}{48}\xi^2 + \frac{7}{6}\xi - \frac{25}{48},$$

and our method can be written as

$$0 = y_n - y_{n-1} + \frac{1}{4}y_{n-2} - \frac{1}{4}y_{n-3} + h\left(-\frac{91}{48}f_{n-1} + \frac{7}{6}f_{n-2} - \frac{25}{48}f_{n-3}\right).$$

Since $r \geq 1$, we have consistency. The zeros of $\rho(\xi)$ are $\xi_1 = 1, \xi_{2,3} = \pm\frac{1}{2}i$. Since $|\xi| \leq 1$, $i = 1,2,3$, and all zeros are simple, the method is zero-stable. Thus it is also convergent.

♣

2.7. Boundary problems

We are solving differential equations of order 2 with two boundary conditions.

$$y'' = f(x, y, y') + \text{ boundary conditions}$$

2.7.1 Linear boundary problem

$$\underbrace{y''(x) + p(x)y'(x) + q(x)y(x)}_{=Ly(x)} = r(x), \quad y(a) = \alpha,\ y(b) = \beta.$$

The numerical methods for solving the linear boundary problems are:

1. *Combination of the initial problems.* We choose different γ_1 and γ_2 and solve two initial problems to compute $y_1(b)$ and $y_2(b)$. That is, we have the following equations

$$\begin{aligned} y_1''(x) + p(x)y_1'(x) + q(x)y_1(x) &= r(x), \quad y_1(a) = \alpha,\ y_1'(a) = \gamma_1, \\ y_2''(x) + p(x)y_2'(x) + q(x)y_2(x) &= r(x), \quad y_2(a) = \alpha,\ y_2'(a) = \gamma_2. \end{aligned}$$

   By combining y_1 and y_2 we get the real solution as

$$y(x) = Ay_1(x) + (1 - A)y_2(x).$$

   Clearly,

$$Ly(x) = r(x), \quad y(a) = \alpha.$$

   From

$$y(b) = Ay_1(b) + (1 - A)y_2(b) = \beta$$

   we compute

$$A = \frac{\beta - y_2(b)}{y_1(b) - y_2(b)}.$$

2. *Difference method.* The most commonly used method for solving boundary problems is based on the approximation of the derivative

$$y''(x) + p(x)y'(x) + q(x)y(x) = r(x). \tag{2.7.1}$$

   Often, we are already satisfied with a method of order 2, where the first and the second derivatives are approximated with differences of order ≥ 2. The interval $[a, b]$ is divided into n equidistant intervals with points $a = x_0 < x_1 < \ldots < x_n = b$, where $x_i = x_0 + ih$ and $h = \Delta x_i$.

We use the following approximations

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_{i-1})}{2h} + \mathcal{O}(h^2),$$
$$y''(x_i) = \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} + \mathcal{O}(h^2).$$

We solve equation (2.7.1) in all the points x_i, $i = 1,2,\ldots,n-1$. Let us use the notation

$$p_i := p(x_i), \quad q_i := q(x_i), \quad r_i := r(x_i), \quad y_i := y(x_i).$$

We get the following system of equations

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p_i \cdot \frac{y_{i+1} - y_{i-1}}{2h} + q_i \cdot y_i = r_i, \quad i = 1,2,\ldots,n-1, \tag{2.7.2}$$

where $y_0 = \alpha$ and $y_{n+1} = \beta$. With this we have that the variables in our system are $y_1, y_2, \ldots, y_{n-1}$. Since the system is tridiagonal, it can be solved very efficiently.
We can also use the difference method if our boundary conditions contain derivatives, that is, if we have boundary conditions of the form:

$$\alpha_1 y(a) + \beta_1 y'(a) = \gamma_1, \quad \alpha_2 y(b) + \beta_2 y'(b) = \gamma_2.$$

The above derivatives are approximated with

$$y'(x_0) \approx \frac{y_1 - y_{-1}}{2h}, \quad y'(x_{n+1}) \approx \frac{y_{n+2} - y_n}{2h},$$

where y_{-1} and y_{n+2} are the so-called virtual variables. By applying these relations in our boundary conditions, we can express

$$y_{-1} = q_1(y_0, y_1), \quad y_{n+1} = q_2(y_{n-1}, y_n),$$

which we put into the equation (2.7.2) for $i = 0$ and $i = n$.

Example 2.13

Solve equation

$$y'' + 4y' - y = x, \quad 3y(0) - y'(0) = 1,\ y'(3) = 1$$

for $h = 1$ using the difference method.

Solution: We have to solve the system

$$y_{i+1} - 2y_i + y_{i-1} + 4 \cdot \frac{y_{i+1} - y_{i-1}}{2} - y_i = x_i, \quad i = 0,1,2,3,$$

or equivalently

$$3y_{i+1} - 3y_i - y_{i-1} = x_i, \quad i = 0,1,2,3.$$

For $i = 0$ and $i = 3$ we have to introduce variables y_{-1} and y_4 which can be expressed as

$$y_{-1} = 2 - 6y_0 + y_1, \quad y_4 = 2 + y_2.$$

Hence, we have

$$\begin{cases} 3y_1 - 3y_0 - (2 - 6y_0 + y_1) & = 0 \\ 3y_2 - 3y_1 - y_0 & = 1 \\ 3y_3 - 3y_2 - y_1 & = 2 \\ 3(2 + y_2) - 3y_3 - y_2 & = 3 \end{cases}$$

$$\Leftrightarrow \begin{cases} 2y_1 + 3y_0 & = 2 \\ 3y_2 - 3y_1 - y_0 & = 1 \\ 3y_3 - 3y_2 - y_1 & = 2 \\ -3y_3 + 2y_2 & = -3. \end{cases}$$

In the matrix form we obtain

$$\begin{bmatrix} 3 & 2 & 0 & 0 \\ -1 & -3 & 3 & 0 \\ 0 & -1 & -3 & 3 \\ 0 & 0 & 2 & -3 \end{bmatrix} \cdot \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 2 \\ -3 \end{bmatrix}.$$

The solution is then

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{3}{4} \\ \frac{3}{2} \end{bmatrix}.$$

♣

2.7.2 Non-linear boundary problems. Shooting method

We are solving the DE

$$y'' = f(x, y, y')$$

where f is a non-linear in y and y'. The boundary conditions are $y(a) = \alpha$ and $y(b) = \beta$. Now, we can not transform the problem to appropriate combination of initial value problems. Furthermore, using the difference method, we get a non-linear system with lots of unknown variables, so the Newton's method would not work. However, we can use the so-called *shooting method*, which we

describe now. We transform first our problem to the initial value problem

$$y'' = f(x,y,y'), \quad y(a) = \alpha,\ y'(a) = v, \tag{2.7.3}$$

where we denote the solution with $y(x;v)$. We are looking for v^* such that $y(b;v^*) = b$. Let us define the function $E(v) = y(b;v) - \beta$. We are looking for the roots of the function E. We can use an arbitrary method to compute the zeros of a non-linear function: bisection, tangent method, secant method, etc.
Let us consider more precisely the tangent method. We perform the following iterative procedure:

$$v_0 \text{ is chosen arbitrary}, \quad v_{r+1} = v_r - \frac{E(v_r)}{E'(v_r)}.$$

For $v = v_r$, we have $E(v_r) = y(b;v_r) - \beta$, where $y(b;v_r)$ is the solution of the initial value problem (2.7.3) for $v = v_r$. How do we compute $E'(v_r)$?

$$E'(v_r) = \frac{\partial}{\partial v}E(v)\Big|_{v=v_r} = \frac{\partial}{\partial v}\left(y(x;v) - \beta\right)\Big|_{x=b,v=v_r} = \frac{\partial}{\partial v}y(x;v)\Big|_{x=b,v=v_r}.$$

Let us define the function $z(x;v) := \frac{\partial}{\partial v}y(x;v)$. We are looking for $z(b;v_r)$. Let us differentiate our equation $y'' = f(x,y,y')$ with respect to variable v:

$$\frac{\partial}{\partial v}y'' = \frac{\partial}{\partial v}\left(\frac{\partial^2}{\partial x^2}y\right) = \frac{\partial^2}{\partial x^2}\left(\frac{\partial}{\partial v}y\right) = \frac{\partial^2}{\partial x^2}z = z''$$
$$\frac{\partial}{\partial v}f(x,y,y') = f_x \cdot \frac{\partial x}{\partial v} + f_y \cdot z + f_{y'} \cdot \frac{\partial}{\partial v}\left(\frac{\partial}{\partial x}y\right) = f_y z + f_{y'} z'$$

Hence,

$$z'' = f_y z + f_{y'} z'$$

and the boundary conditions are

$$z(a;v) = 0,\ z'(a;v) = 1.$$

We solve this initial value problem and z_n is the value of $E'(v_r)$.

Partial differential equations

We come across them in numerical problems in physics, finances, economics, etc. We will consider three model equations and how to solve them: *Poisson's equation, Wave equation* and *Heat equation*. We will only consider functions of two variables x and y, where with $u(x,y)$ we will denote the solution of the PDE. The simplest type are the equations of second order. Let us look at the so-called quasi-linear equations of the form

$$a\,u_{xx} + b\,u_{xy} + c\,u_{yy} = f(x,y,u,u_x,u_y). \tag{3.0.1}$$

We distinguish three types:

- elliptic type (Poisson's equation): $b^2 - 4ac < 0$.
- parabolic type (Heat equation): $b^2 - 4ac = 0$.
- hyperbolic type (Wave equation): $b^2 - 4ac > 0$.

We say that the PDE has a canonical form if it does not contain the mixed derivative u_{xy}. If we have a linear PDE of second order

$$a\,u_{xx} + b\,u_{xy} + c\,u_{yy} + e\,u_x + g\,u_y + r\,u = f(x,y),$$

where a,b,c,e,g,r are constants and $b \neq 0$, we have to determine new variables $p := p(x,y)$ and $q := q(x,y)$ such that the equation with variables p and q is of the form

$$A\,u_{pp} + C\,u_{qq} + E\,u_p + G\,u_q + R\,u = F(p,q).$$

Let us rotate the xOy coordinate system for the angle α which is determined with the equation

$$\tan(2\alpha) = \frac{b}{a-c}. \tag{3.0.2}$$

The new variables are of the form

$$p := x\cos\alpha + y\sin\alpha, \quad q := -x\sin\alpha + y\cos\alpha.$$

Then

$$u_{xx} = (u_x)_x = (u_p \cdot p_x + u_q \cdot q_x)_x = (u_p\cos\alpha - u_q\sin\alpha)_x = u_{pp}\cos^2\alpha - 2u_{pq}\sin\alpha\cos\alpha + u_{qq}\sin^2\alpha,$$
$$u_{xy} = (u_x)_y = (u_p\cos\alpha - u_q\sin\alpha)_y = u_{pp}\cos\alpha\sin\alpha + u_{pq}(\cos^2\alpha - \sin^2\alpha) - u_{qq}\cos\alpha\sin\alpha,$$
$$u_{yy} = \ldots = u_{pp}\sin^2\alpha + 2u_{pq}\sin\alpha\cos\alpha + u_{qq}\cos^2\alpha.$$

We have that

$$au_{xx} + bu_{xy} + cu_{yy} = Au_{pp} + C_{uqq} + Bu_{pq},$$

where

$$A = a\cos^2\alpha + b\sin\alpha\cos\alpha + c\sin^2\alpha \neq 0,$$
$$C = a\cos^2\alpha - b\sin\alpha\cos\alpha + c\sin^2\alpha \neq 0,$$

and we claim that $B = 0$. Indeed

$$\begin{aligned} B &= -2a\sin\alpha\cos\alpha + b(\cos^2\alpha - \sin^2\alpha) + 2c\sin\alpha\cos\alpha \\ &= -a\sin 2\alpha + b\cos 2\alpha + c\sin 2\alpha \\ &= (c-a)\sin 2\alpha + b\cos 2\alpha = 0. \end{aligned}$$

The last equality holds since (3.0.2). Some PDEs in canonical form are:

- parabolic type - heat equation
$$u_t = c^2 u_{xx} + f(x,t)$$
- elliptic type - Poisson's equation
$$\Delta u = -f(x,y),$$
and for $f = 0$ we have the Laplace equation.
- hyperbolic type - wave equation
$$u_{tt} = k^2 u_{xx} + f(x,t).$$

As in the case of ODE to have uniqueness of the solutions, we need boundary and initial conditions (when the time variable is involved).
If we are looking for the solution on a domain $\Omega \subseteq \mathbb{R}^2$, then the boundary conditions can be:

- *Dirichlet's condition.* We are given the value u at the boundary, that is $u(x,y) = g(x,y)$ for $(x,y) \in \partial\Omega$.
- *Neumann's condition.* We are given the value of the first derivative of u in the direction of the normal to the boundary, that is $\frac{\partial u}{\partial n}(x,y) = g(x,y)$ for $(x,y) \in \partial\Omega$.
- *Robbinson's condition.* Linear combination of the Dirichlet's and Neumann's condition.

If one of the variables is the time t, then additionally the initial conditions for u and it's derivative are given at $t = 0$. We have good numerical methods for solving such PDEs:

- difference method
- finite elements method
- collocation
- boundary elements method
- isogeometric analysis

Let us consider more precisely the difference method. Let Ω denote the computational domain and let us split the $x-$ and $y-$axis equidistantly as

$$x_i = x_0 + i\delta x, \quad i = 0,1,\ldots,$$
$$y_j = y_0 + j\delta y, \quad j = 0,1,\ldots$$

Let us denote with Ω_δ all the intersections of two lines inside the domain and with $\partial\Omega_\delta$ the intersection of one line and the boundary of the domain (see Fig. 3.1). In all the points of Ω_δ or $\partial\Omega_\delta$ we solve the PDE so that we approximate the derivatives in the differential equations.

For example, let us look at the Poisson's equation

$$u_{xx} + u_{yy} = -f.$$

We know what the approximations for y' and y'' are from the boundary problems for ODE, i.e.,

$$y'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h} + \mathcal{O}(h^2), \quad y''(x_i) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + \mathcal{O}(h^2).$$

Hence,

$$u_{xx}(x_i,y_j) \approx \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\delta x^2}, \quad u_{yy}(x_i,y_j) \approx \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\delta y^2}.$$

Similarly we can compute u_x and u_y.

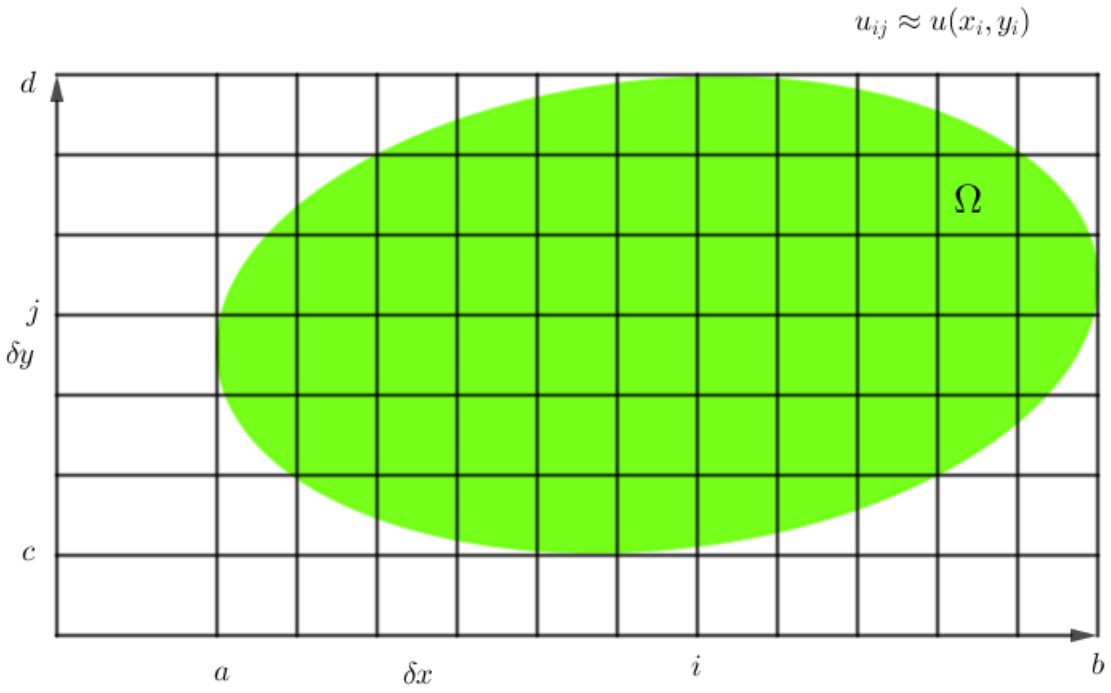


Figure 3.1: Discretized domain Ω.

3.1. Parabolic PDE

The model equation is the Heat equation

$$u_t = u_{xx},$$

where the domain is $\Omega = [0,1] \times [0,T]$. The boundary conditions are

$$u(0,t) = g(t), \quad u(1,t) = h(t).$$

Note that in several variables the equation would be of the form $u_t = \Delta u$. In time direction we denote the distance by δt. We equidistantly divide the interval $[0,1]$ with $n+2$ points $0 = x_0 < x_1 < \ldots < x_{n+1} = 1$. If we use symmetric differences in all of the points (see Fig. 3.2) we get

$$\frac{u_{i,j+1} - u_{i,j-1}}{2\delta t} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\delta x^2}.$$

Let us define the so-called Courant's number

$$\lambda := \frac{\delta t}{\delta x^2}.$$

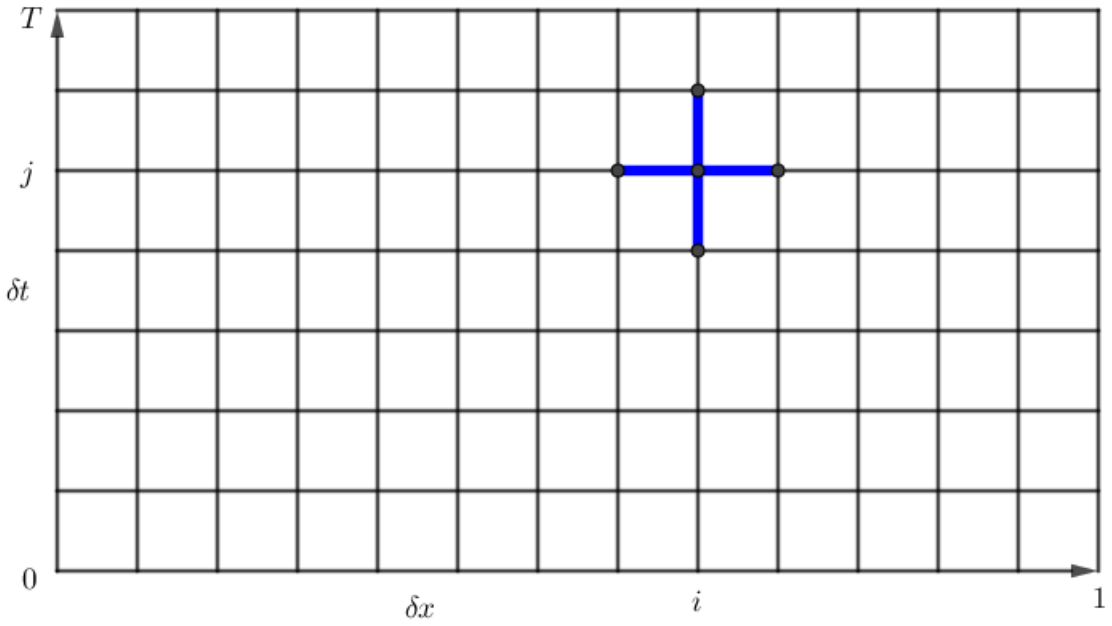


Figure 3.2: Discretisation of domain Ω and the "star" used when applying symmetric differences.

Then our equation becomes

$$u_{i,j+1} = u_{i,j-1} + 2\lambda(u_{i+1,j} - 2u_{ij} + u_{i-1,j}).$$

By defining $u_j := (u_{i,j})_{i=0}^{n+1}$ we get

$$\begin{pmatrix} u_{j+1} \\ u_j \end{pmatrix} = \begin{pmatrix} A & I \\ I & 0 \end{pmatrix} \cdot \begin{pmatrix} u_j \\ u_{j-1} \end{pmatrix}, \quad A = \begin{pmatrix} -4\lambda & 2\lambda & & & \\ 2\lambda & -4\lambda & 2\lambda & & \\ & \ddots & \ddots & \ddots & \\ & & & 2\lambda & -4\lambda \end{pmatrix}.$$

This notation reminds us to the power method. That method converges iff

$$\rho(R) = \max_i |\lambda_i| < 1, \quad \lambda_i \text{ eigenvalues of } R,$$

where R is the iteration matrix. In our case, $R = \begin{pmatrix} A & I \\ I & 0 \end{pmatrix}$. But, since

$$\det R = \prod_i \lambda_i = \pm 1,$$

we have that $\rho(R) \leq 1$ if all the λ_i are ± 1, but this does not have to be true always. This means that the symmetric difference method in time and space leads to numerical problems. Next idea is to choose forward one-step method or backward one-step method in time direction, instead of using

the symmetric difference. That is

$$u(x_i,t_j)' = \frac{u_{i,j+1}-u_{ij}}{\delta t} + \mathscr{O}(\delta t) \qquad \text{or} \qquad u(x_i,t_j)' = \frac{u_{ij}-u_{i,j-1}}{\delta t} + \mathscr{O}(\delta t).$$

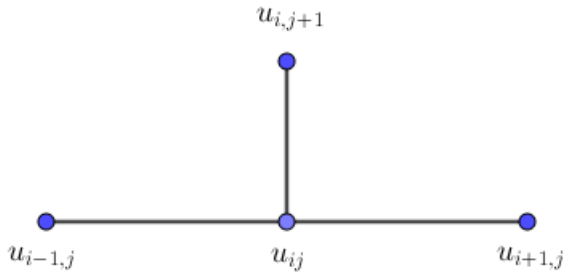
With this our equation becomes

$$\frac{u_{i,j+1}-u_{ij}}{\delta t} = \frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{\delta x^2}.$$

By multiplying with δt and using λ we get

$$u_{i,j+1} = \lambda u_{i-1,j} + \lambda u_{i+1,j} + (1-2\lambda)u_{ij},$$

which represents the so called *explicit difference method*.



Remark 3.1

1. *The method converges for all* $\lambda \leq \frac{1}{2}$.
2. *The best method is obtained for* $\lambda = \frac{1}{6}$.

Let us consider the local error $(D_\delta - D)u_{ij}$, $u_{ij} \doteq u(x_i,t_j)$, where

$$D_\delta u_{ij} := \frac{u_{i,j+1}-u_{ij}}{\delta t} - \frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{\delta x^2}, \quad D := \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}.$$

Let us expand $u_{i,j+1}$ and $u_{i\pm1,j}$ into the Taylor series around u_{ij}. We get:

$$u_{i,j+1} = u_{ij} + \delta t(u_t)_{ij} + \frac{\delta t^2}{2}(u_{tt})_{ij} + \dots$$
$$u_{i\pm1,j} = u_{ij} \pm \delta x(u_x)_{ij} + \frac{\delta x^2}{2}(u_{xx})_{ij} + \dots \pm \frac{\delta x^3}{6}(u_{xxx})_{ij} + \frac{\delta x^4}{24}(u_{xxxx})_{ij} \pm \dots$$

Therefore

$$(D_\delta - D)u_{ij} = \frac{\delta t}{2}(u_{tt})_{ij} - \frac{\delta x^2}{12}(u_{xxxx})_{ij} + \mathscr{O}\left(\delta t^2 + \delta x^4\right)$$

$$= \delta x^2 \left(\frac{\lambda}{2}(u_{tt})_{ij} - \frac{1}{12}(u_{xxxx})_{ij} \right) + \mathscr{O}\left(\delta t^2 + \delta x^4\right).$$

Since $u_t = u_{xx}$, it follows

$$u_{tt} = \frac{\partial}{\partial t}\left(\frac{\partial^2}{\partial x^2}u\right) = \frac{\partial^2}{\partial x^2}\left(\frac{\partial}{\partial t}u\right) = \frac{\partial^2}{\partial x^2}u_t = u_{xxxx}.$$

If we choose $\lambda = \frac{1}{6}$ we get

$$(D_\delta - D)u_{ij} = \mathscr{O}\left(\delta t^2 + \delta x^4\right).$$

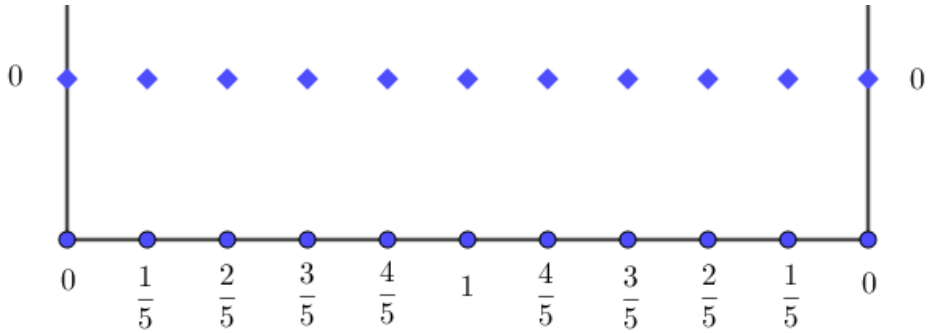
Example 3.1

Solve $u_t = u_{xx}$ with the conditions

$$u(0,t) = u(1,t) = 0,\ \forall t; \quad u(x,0) = \begin{cases} 2x, & 0 \le x \le \frac{1}{2}, \\ 2(1-x), & \frac{1}{2} \le x \le 1. \end{cases}$$

Take $\delta x = \frac{1}{10}$ and $\delta t = \frac{1}{600}$. Calculate the values for $t = \delta t$.

Solution:



The method for $\lambda = \frac{1}{6}$ equals

$$u_{i,j+1} = \frac{1}{6}(u_{i-1,j} + u_{i+1,j} + 4u_{ij}).$$

Straightforward computation gives the solutions

$$u_5 = \frac{14}{15}, \quad u_i = \frac{i}{5},\ u_{10-i} = u_i,\ i = 1,2,3,4.$$

♣

Let us also consider the implicit method. Let

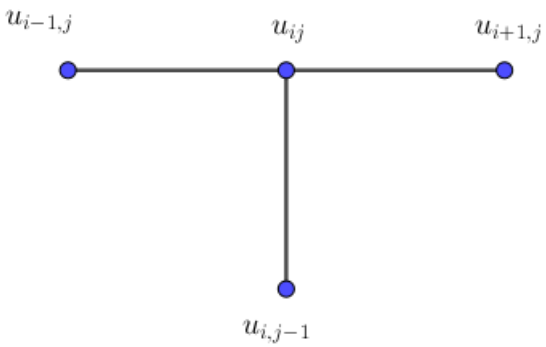
$$(u_t)_{ij} \approx \frac{u_{ij} - u_{i,j-1}}{\delta t}.$$

We get

$$\frac{u_{ij}-u_{i,j-1}}{\delta t}=\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{\delta x^2}$$

or equivalently

$$(1+2\lambda)u_{ij}-\lambda u_{i+1,j}-\lambda u_{i-1,j}=u_{i,j-1}.$$



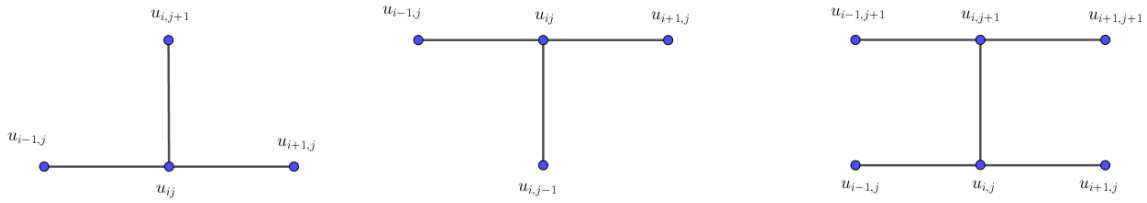
We obtain a tridiagonal linear system with matrix

$$\begin{pmatrix} 1+2\lambda & -\lambda & & & \\ -\lambda & 1+2\lambda & -\lambda & & \\ & \ddots & \ddots & \ddots & \\ & & & -\lambda & 1+2\lambda \end{pmatrix}.$$

Remark 3.2

This method converges for any $\lambda > 0$.

There also exists a method which combines the implicit and the explicit method, thus a method of higher order in time. This method is called the *Crank-Nicolson method*.



We take the average of both methods, that is

$$\frac{u_{i,j+1}-u_{ij}}{\delta t}=\frac{1}{2}\left(\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{\delta x^2}+\frac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{\delta x^2}\right).$$

Again, we get a tridiagonal system. It converges for all $\lambda > 0$ and the error is $\mathcal{O}(\delta t^2+\delta x^2)$.

When considering heat equation in higher dimension, i.e.,

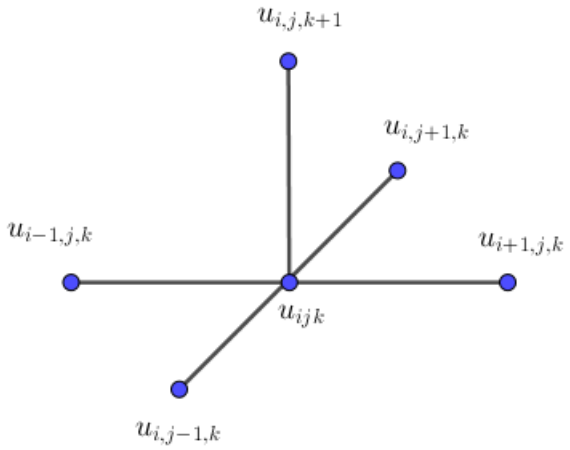
$$u_t = \Delta u, \quad (x,y) \in [0,1]^2 =: \Omega,$$

with the conditions

$$u(x,y,0) = f(x,y); \;\; u(x,y,t) = g(x,y,t), \;\; (x,y) \in \partial\Omega,$$

then, for example, the explicit method is of the form

$$\frac{u_{i,j,k+1} - u_{ijk}}{\partial t} = \frac{u_{i+1,j,k} - 2u_{ijk} + u_{i-1,j,k}}{\delta x^2} + \frac{u_{i,j+1,k} - 2u_{ijk} + u_{i,j-1,k}}{\delta y^2}.$$



3.2. Elliptic PDE

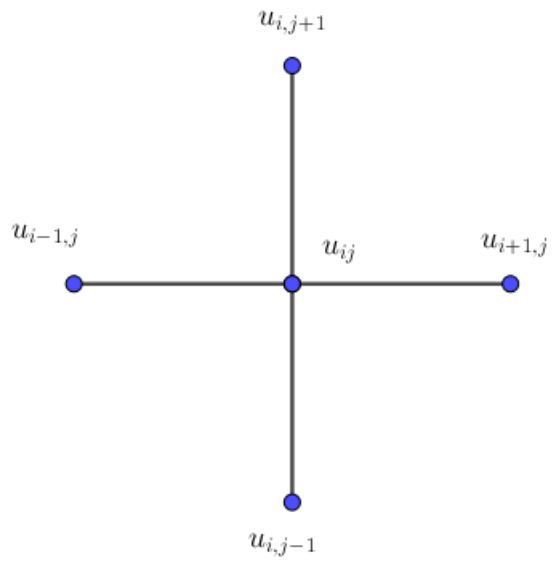
We will consider the Poisson's equation in 2D

$$-u_{xx} - u_{yy} = f$$

on the domain $\Omega = [a,b] \times [c,d]$. The boundary conditions are

$$u(x,y) = g(x,y), \quad (x,y) \in \partial\Omega.$$

Again, let us consider the difference method. We cover the domain Ω with a net determined by the equidistant division of $[a,b]$ with points $a = x_0, x_1, \ldots, x_n, x_{n+1} = b$ and $[c,d]$ with $c = y_0, y_1, \ldots, y_m, y_{m+1} = d$. The second partial derivatives with respect to x and y are approximated with symmetric differences, thus we have a 5-point scheme:



with the equation

$$-\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{\delta x^2}-\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{\delta y^2}=f_{ij}.$$

Let's define

$$\frac{1}{\delta^2}:=\frac{2}{\delta x^2}+\frac{2}{\delta y^2}=\frac{2(\delta x^2+\delta y^2)}{\delta x^2\delta y}.$$

By multiplying the above equation with δ^2 we get

$$u_{ij}-\Theta_x(u_{i+1,j}+u_{i-1,j})-\Theta_y(u_{i,j+1}+u_{i,j-1})=f_{ij}\delta^2,$$

where

$$\Theta_x=\frac{\delta y^2}{2(\delta x^2+\delta y^2)},\quad \Theta_y=\frac{\delta x^2}{2(\delta x^2+\delta y^2)}.$$

We introduce the notations

$$u_j:=(u_{ij})_{i=1}^n,\ \ u:=(u_j)_{j=1}^m,\ \ f_j:=(f_{ij})_{i=1}^n,\ \ I_n=id\in\mathbb{R}^{n\times n},\ \ L_n=\begin{bmatrix}0 & & & \\ 1 & \ddots & & \\ & \ddots & & \\ & & 1 & 0\end{bmatrix}\in\mathbb{R}^{n\times n}.$$

Now our equation is of the form

$$(I_n - \Theta_x(L_n + L_n^T))u_j - \Theta_y I_n(u_{j-1} + u_{j+1}) = \delta^2 f_j + \begin{bmatrix} \Theta_x g_{0,j} \\ 0 \\ \vdots \\ 0 \\ \Theta_x g_{n+1,j} \end{bmatrix}, \quad \forall j \in \{1,2,\ldots,m\}$$

We construct the following tridiagonal block matrix:

$$A = \begin{bmatrix} \begin{bmatrix} 1 & -\Theta_x & & \\ -\Theta_x & 1-\Theta_x & & \\ & \ddots & \ddots & \ddots \\ & & -\Theta_x & 1 \end{bmatrix} & \begin{bmatrix} -\Theta_y & & & \\ & -\Theta_y & & \\ & & \ddots & \\ & & & -\Theta_y \end{bmatrix} & & \\ \begin{bmatrix} -\Theta_y & & & \\ & -\Theta_y & & \\ & & \ddots & \\ & & & -\Theta_y \end{bmatrix} & & & \\ & \ddots & & \\ & & & \begin{bmatrix} -\Theta_y & & & \\ & -\Theta_y & & \\ & & \ddots & \\ & & & -\Theta_y \end{bmatrix} \\ & & \begin{bmatrix} -\Theta_y & & & \\ & -\Theta_y & & \\ & & \ddots & \\ & & & -\Theta_y \end{bmatrix} & \begin{bmatrix} 1 & -\Theta_x & & \\ -\Theta_x & 1-\Theta_x & & \\ & \ddots & \ddots & \ddots \\ & & -\Theta_x & 1 \end{bmatrix} \end{bmatrix}$$

We get the linear system:

$$Au = \delta^2 f + \begin{bmatrix} \Theta_y u_0 \\ 0 \\ \vdots \\ 0 \\ \Theta_y u_{m+1} \end{bmatrix}, \quad \delta^2 f = \left(f_j + \begin{bmatrix} \Theta_x g_{0,j} \\ 0 \\ \vdots \\ 0 \\ \Theta_x g_{n+1,j} \end{bmatrix} \right)_{j=1}^{m}.$$

Remark 3.3

We got a block tridiagonal system. In each row we have at most 5 non-zero elements. Hence, for solving it we will use some iterative method: Jacobi's method, Gauss-Seidel's method, SOR method, ADI method, ...

Remark 3.4

For $\delta x = \delta y$, we have that $\Theta_x = \Theta_y = \frac{1}{4}$ and we get

$$A = \begin{bmatrix} T & -I & & \\ -I & T & -I & \\ & \ddots & \ddots & \ddots \\ & & -I & T \end{bmatrix},$$

where

$$T = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 4 \end{bmatrix}.$$

The local error of the 5-point approximation is $\tau(x,y) = (\Delta u - \Delta_\delta u)(x,y)$. We have that

$$\Delta_\delta u(x,y) = \frac{u(x+\delta_x,y) - 2u(x,y) + u(x-\delta_x,y)}{\delta x^2} + \frac{u(x,y+\delta_y) - 2u(x,y) + u(x,y-\delta_y)}{\delta_y^2}$$

$$u(x \pm \delta_x, y) = u(x,y) \pm \delta_x u_x(x,y) + \frac{\delta_x^2}{2} u_{xx}(x,y) \pm \dots$$

$$u(x, y \pm \delta_y) = u(x,y) \pm \delta_y u_y(x,y) + \frac{\delta_y^2}{2} u_{yy}(x,y) \pm \dots$$

Hence,

$$\tau(x,y) = \frac{1}{12}\left(\delta_x^2 u_{xxxx}(x,y) + \delta_y^2 u_{yyyy}(x,y)\right) + \mathcal{O}(\delta_x^4 + \delta_y^4).$$

If $\delta_x = \delta_y = h$, then

$$\tau(x,y) = \frac{h^2}{12}\left(u_{xxxx}(x,y) + u_{yyyy}(x,y)\right) + \mathcal{O}(h^4).$$

Example 3.2

Solve the PDE

$$u_{xx}+u_{yy}+2=0 \quad \text{on } \Omega=[-1,1]^2$$

with the boundary condition $u\Big|_{\partial\Omega}=0$. *Write the system, assuming we have a symmetry, for the step sizes:*

(a) $\delta x=\delta y=h=1$.

(b) $\delta x=\delta y=h=\frac{1}{2}$.

(c) Assuming that the error can be expressed as $u(x_i,y_j)=u_{ij}+ch^2+\mathcal{O}(h^4)$, *can we get in a simple way a better approximation for* $u(0,0)$?

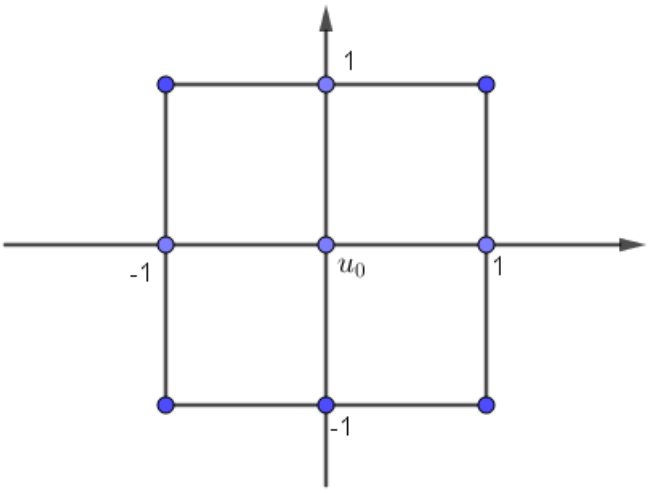
Solution:

(a) For $\delta x=\delta y=h=1$ we have the equation

$$\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{1^2}+\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{1^2}=-2.$$

Therefore

$$u_{i,j+1}+u_{i,j-1}+u_{i+1,j}+u_{i-1,j}-4u_{ij}=-2.$$



In other words, since $h=1$, we have only one equation

$$-4u_0=-2 \quad \Rightarrow \quad u_0=\frac{1}{2}.$$

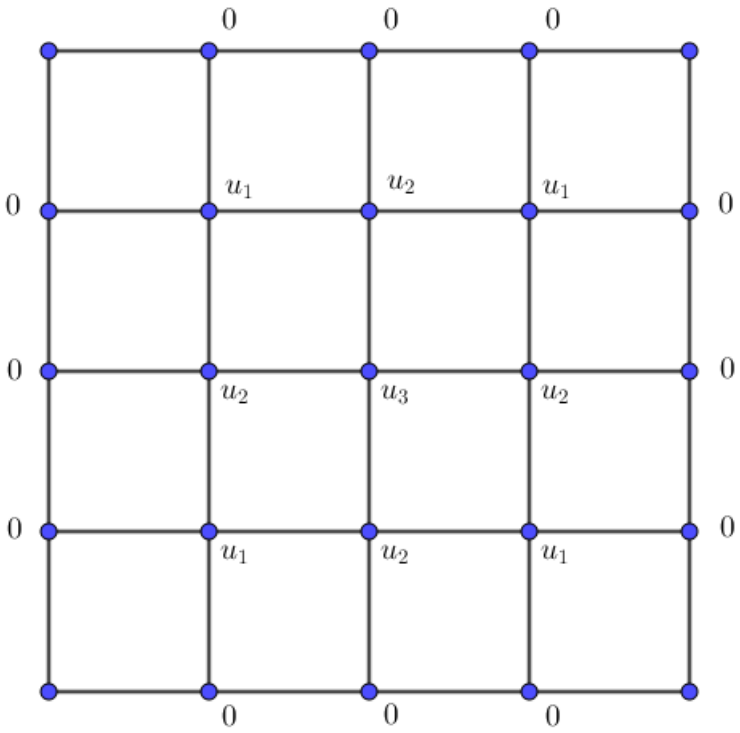
(b) For $h = \frac{1}{2}$ we get the equation

$$u_{i,j+1} + u_{i,j-1} + u_{i+1,j} + u_{i-1,j} - 4u_{ij} = -\frac{1}{2}.$$

Assuming symmetry

$$u(x,y) = u(-x,y) = u(x,-y) = u(y,x)$$

we get the scheme:



We get the following system

$$\begin{cases} 4u_1 - 2u_2 & = \frac{1}{2} \\ -2u_1 + 4u_2 - u_3 & = \frac{1}{2} \\ 4u_2 - 4u_3 & = -\frac{1}{2} \end{cases}$$

whose solution is

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} \frac{11}{32} \\ \frac{7}{16} \\ \frac{9}{16} \end{bmatrix} \begin{bmatrix} 0.34375 \\ 0.4375 \\ 0.5625 \end{bmatrix}.$$

(c) By the assumption for the step h we have that

$$u(0,0) = \frac{1}{2} + ch^2 + \mathcal{O}(h^4).$$

Replacing h with $\frac{h}{2}$, we get

$$u(0,0) = \frac{9}{16} + c \cdot \frac{h^2}{4} + \mathcal{O}(h^4).$$

By multiplying the last equation with (-4) and adding it to the first equation we get

$$3\,u(0,0) = \frac{7}{12} + \mathcal{O}(h^4) \doteq 0.5833 + \mathcal{O}(h^4).$$

The exact solution is

$$u(0,0) \approx 0.589\ldots$$

♣

3.2.1 Solving elliptic PDE on areas with curved boundaries

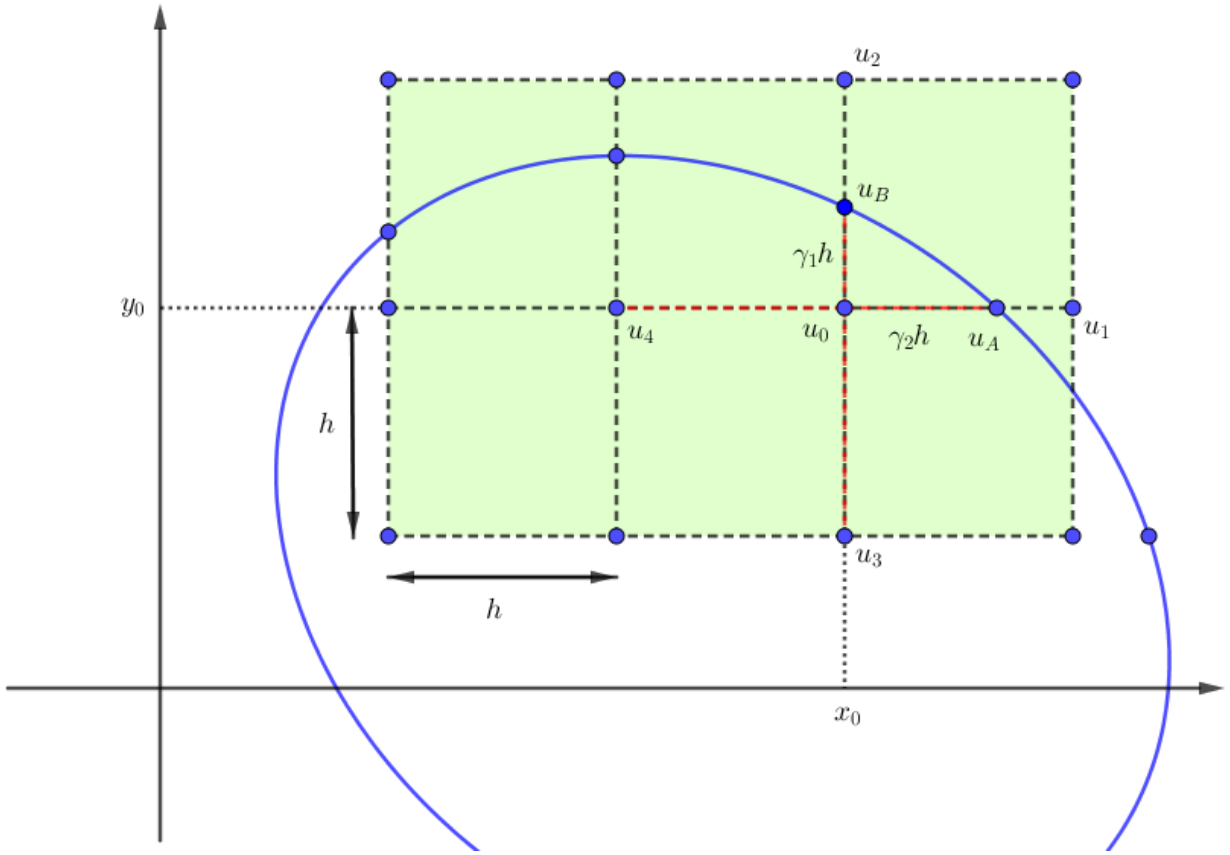


Figure 3.3: Discretization of domain Ω with curved boundaries.

Consider Fig. 3.3. We are given two points, u_A and u_B, on the boundary of the domain. The question

is how to express $(u_0)_{xx}$ and $(u_0)_{yy}$. Let us write the Taylor series for u_A around the point u_0:

$$u_A = u(x_0 + \gamma_2 h, y_0) = u_0 + \gamma_2 h \cdot (u_0)_x + \frac{(\gamma_2 h)^2}{2}(u_0)_{xx} + \dots$$

Similarly,

$$u_4 = u(x_0 - h, y_0) = u_0 - h(u_0)_x + \frac{h^2}{2}(u_0)_{xx} + \dots$$

We have

$$\gamma_2 u_4 + u_A = (1+\gamma_2)u_0 + \frac{h^2}{2}\gamma_2(1+\gamma_2)(u_0)_{xx} + \dots,$$

or equivalently

$$(u_0)_{xx} = \frac{2u_A}{\gamma_2(1+\gamma_2)h^2} + \frac{2u_4}{(1+\gamma_2)h^2} - \frac{2u_0}{\gamma_2 h^2}.$$

Analogously, we get that

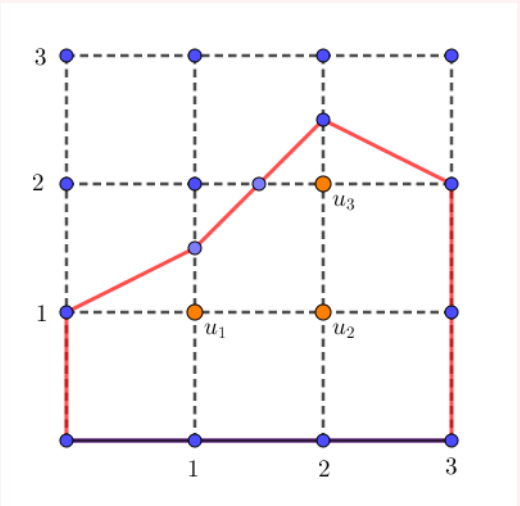
$$(u_0)_{yy} = \frac{2u_B}{\gamma_1(1+\gamma_1)h^2} + \frac{2u_3}{(1+\gamma_1)h^2} - \frac{2u_0}{\gamma_1 h^2}.$$

If we want to solve the equation $\Delta u = f$ in the point (x_0, y_0) we obtain the equation

$$\frac{u_A}{\gamma_2(1+\gamma_2)} + \frac{u_4}{(1+\gamma_2)} + \frac{u_B}{\gamma_1(1+\gamma_1)} + \frac{u_3}{(1+\gamma_1)} - \frac{u_0(\gamma_1+\gamma_2)}{\gamma_1\gamma_2} = \frac{1}{2}h^2 f.$$

Example 3.3

Solve the equation $\Delta u = 0$ over domain Ω, shown bellow:



The boundary conditions are $u = 0$ over the bottom line and $u = 1$ over all of the other lines of $\partial\Omega$. Find the approximants for $u(1,1), u(2,1)$ and $u(2,2)$.

Solution: Point u_2 is a "standard" point, hence, we have the equation

$$\frac{1-2u_2+u_1}{1}+\frac{u_3-2u_2+0}{1}=0 \quad \Rightarrow \quad u_1-4u_2+u_3=-1.$$

For the point u_1 we have $\gamma_x=1$ and $\gamma_y=\frac{1}{2}$. Then

$$\frac{1}{2}+\frac{u_2}{2}+\frac{1}{\frac{1}{2}\cdot\frac{3}{2}}+0-\frac{u_1\cdot\frac{3}{2}}{\frac{1}{2}}=0 \quad \Rightarrow \quad -6u_1+u_2=-\frac{11}{3}.$$

For u_3 we have $\gamma_x=\frac{1}{2}$ and $\gamma_y=\frac{1}{2}$. Therefore

$$\frac{4}{3}+\frac{2}{3}+\frac{4}{3}+\frac{2u_2}{3}-4u_3=0 \quad \Rightarrow \quad u_2-6u_3=-5.$$

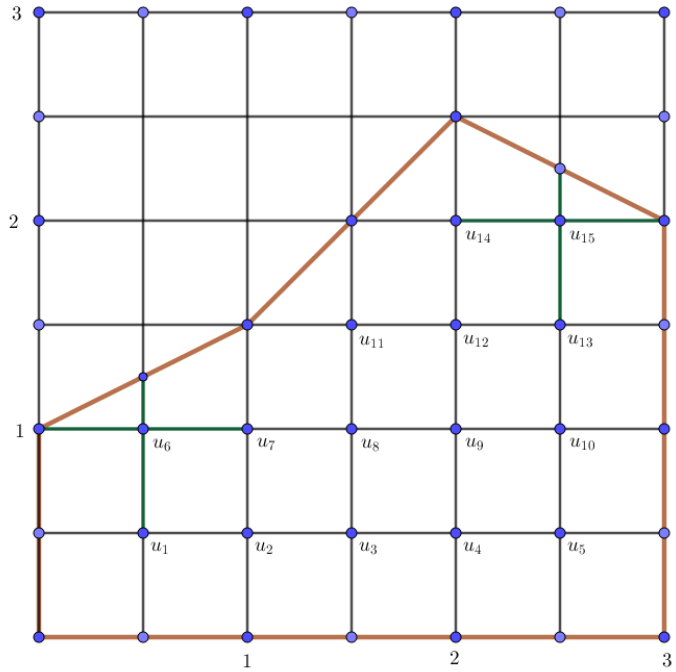
We get the following system

$$\begin{cases} u_1-4u_2+u_3 &= -1 \\ -6u_1+u_2 &= -\frac{11}{3} \\ u_2-6u_3 &= -5 \end{cases} \quad \Rightarrow \quad \begin{bmatrix} 1 & -4 & 1 \\ -6 & 1 & 0 \\ 0 & 1 & -6 \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} -1 \\ -\frac{11}{3} \\ -5 \end{bmatrix}.$$

The solution of the system is

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0.7222\ldots \\ 0.6666\ldots \\ 0.9444\ldots \end{bmatrix}.$$

For $h=\frac{1}{2}$ we get following scheme



The only "non-standard" points are u_6 and u_{15}. For u_6 we have that $\gamma_x = \frac{1}{2}$ and $\gamma_y = \frac{1}{4}$. Thus, we have

$$\frac{4}{3} + \frac{2u_7}{3} + \frac{16}{5} + \frac{4u_1}{5} - 6u_6 = 0 \quad \Rightarrow \quad 6u_1 - 45u_6 + 5u_7 = -34.$$

For u_{15} we also have $\gamma_x = \frac{1}{2}$ and $\gamma_y = \frac{1}{4}$. Hence,

$$6u_{13} - 45u_{15} + 5u_{14} = -34.$$

After some computations, we get the following system

$$\left[\begin{array}{ccccc|ccccc|ccccc}
-4 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -4 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -4 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -4 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\hline
6 & 0 & 0 & 0 & 0 & -45 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & -4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & -4 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & -4 & 0 & 0 & 1 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -4 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -4 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 5 & -45
\end{array}\right]
\left[\begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ \hline u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \\ \hline u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ u_{15} \end{array}\right]
=
\left[\begin{array}{c} -1 \\ 0 \\ 0 \\ 0 \\ -1 \\ \hline -34 \\ -1 \\ 0 \\ 0 \\ -1 \\ \hline -2 \\ 0 \\ -1 \\ -2 \\ -34 \end{array}\right].$$

♣

3.3. Hyperbolic PDE

The model equation is the wave equation

$$u_{tt} = \alpha^2 u_{xx} \quad \text{on } [0,1] \times [0,T].$$

The boundary conditions are

$$u(0,t) = u(1,t) = 0,$$

and the initial conditions are

$$u(x,0) = f(x), \;\; u_t(x,0) = g(x).$$

One possible method to solve this equation is the following. Let $u_{ij} \approx u(x_i, t_j)$. By using symmetric differences we get

$$\frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\delta t^2} = \alpha^2 \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\delta x^2}.$$

Let us denote the Courant's number $\lambda^2 = \frac{\alpha^2 \delta t^2}{\delta x^2}$. We get the equation

$$u_{i,j+1} = u_{ij}(2 - 2\lambda^2) + \lambda^2 u_{i+1,j} + \lambda^2 u_{i-1,j} - u_{i,j-1},$$

which represents the explicit method for solving the wave equation.

We know $u(x_i, 0) = u_{i,0} = f(x_i)$. To get $u_{i,1}$ we express the point $u_{i,1} = u(x_i, t_1)$ in the Taylor series around $u(x_i, 0)$,

$$u_{i,1} = u(x_i, 0) + \delta t u_t(x_i, 0) + \frac{\delta t^2}{2} u_{tt}(x_i, 0).$$

Since $u_{tt} = \alpha^2 u_{xx}(x_i, 0)$ and

$$u_{xx}(x_i, 0) = \frac{u(x_{i+1}, 0) - 2u(x_i, 0) + u(x_{i-1}, 0)}{\delta x^2}$$

we obtain

$$u_{i,1} = f(x_i) + \delta t g(x_i) + \frac{\lambda^2}{2}(f(x_{i+1} - 2f(x_i) + f(x_{i-1})),$$

or equivalently

$$u_{i,1} = \delta t g(x_i) + (1 - \lambda^2) f(x_i) + \frac{\lambda^2}{2}(f(x_{i+1}) - f(x_{i-1})).$$

Remark 3.5

The method converges for all $\lambda \leq 1$. There exists also an implicit method which converges for all λ.

Remark 3.6

The method can be expressed in a matrix form $u_{j+1} = Au_j - u_{j-1}$, where $u_j = (u_{ij})_{i=0}^n$ and

$$A = \begin{bmatrix} 2(1-\lambda^2) & \lambda^2 & & & \\ \lambda^2 & 2(1-\lambda^2) & \lambda^2 & & \\ & \ddots & \ddots & \ddots & \\ & & & \lambda^2 & 2(1-\lambda^2) \end{bmatrix}.$$

Example 3.4

Solve the wave equation $u_{tt} = 4u_{xx}$ *with boundary and initial conditions*

$$u(0,t) = u(1,t) = 0, \;\; u(x,0) = \sin(\pi x), \;\; u_t(x,0) = 0, \; x \in [0,1]$$

and $\delta t = \frac{1}{10}$, $\lambda = 1$. *Write approximations for levels* $t \in \{0, \delta t, 2\delta t\}$.

Solution: From $\alpha = 2, \lambda = 1$ and $\delta x = \frac{1}{10}$, we compute

$$\delta t = \frac{\delta x \cdot \lambda}{\alpha} = \frac{1}{20}.$$

Let $x_i = \frac{i}{10}$, $i = 0, 1, \ldots, 10$. Then

$t = 0$:

$$u_{i,0} = \sin(\pi x_i),$$

$t = \frac{1}{20}$:

$$u_{i,1} = \frac{1}{2}\left(\sin(\pi x_{i-1}) + \sin(\pi x_{i+1})\right),$$

$t = \frac{1}{10}$:

$$\begin{aligned} u_{i,2} &= 2(1-4)u_{i,1} + 4(u_{i+1,1} - u_{i-1,1}) - u_{i,0} \\ &= -\sin(\pi x_i) - 3\left(\sin(\pi x_{i-1}) + \sin(\pi x_{i+1})\right) + 2\left(\sin(\pi x_{i-2}) + \sin(\pi x_{i+2})\right). \end{aligned}$$

♣