Chapter 2

Finite differences for boundary value problems

2.1 Principle of the method

Amongst numerical methods for the approximate solution to boundary value problems, the finite difference method most closely resembles the numerical schemes used to approximate the solutions of ordinary differential equations. The basic idea is the same: replace the derivatives which appear in the equation by appropriate difference quotients, whence the name of the method, *finite difference method*, as opposed to "infinitesimal" differences, which would correspond to the derivatives themselves. In this method, we do not actually compute a function defined on the full interval of consideration, but rather an approximation of the values of the solution to the boundary value problem at a *finite* number of points on this interval (this is necessary in order to be able to implement the corresponding algorithm). One is of course free then to interpolate the values thus obtained to construct a function (one may for example construct a continuous piecewise-linear function or a more regular function using cubic splines) and draw, for instance, an appealing graph, but this is not part of the method itself, at most a post-processing of the result of the method.

We therefore begin by introducing a *uniform discretisation mesh* by taking an integer $N \geq 1$ and letting $h = \frac{1}{N+1}$ and $x_i = ih$ for $i = 0, 1, \dots, N+1$, such that the points x_i are uniformly spaced with *spacing h*, *i.e.*, $x_{i+1} - x_i = h$, with $x_0 = 0$ and $x_{N+1} = 1$. There are thus N grid points (also known as mesh points) in the interior of the interval, corresponding to indices $i = 1, \dots, N$, cf. figure 2.1 below. In the following, we will let N tend to infinity, which is equivalent to letting h tend towards 0. We will compute some numerical values denoted by u_i, $i = 1, \dots, N$, which are approximations to the exact values $u(x_i)$. The quality of approximation will improve for larger N (or, equivalently, smaller h), as we will prove later assuming that $u_0 = \alpha, u_{N+1} = \beta$ (in figure 2.1 we illustrated the case $\alpha = \beta = 0$). The heuristic idea is that, the derivatives are by definition the limit of the difference quotients, thus we expect not to commit too large an error by replacing these with quotients which are traditionally called *finite differences*.

If φ is a sufficiently regular function on $[0, 1]$, we can approximate the derivative of φ with respect to x_i, supposing that $x_{i\pm1} - x_i = \pm h$ is "sufficiently small", by the *right-sided finite*

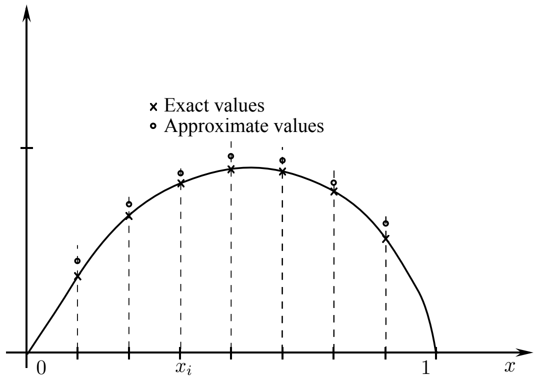


Figure 2.1: Idea of the finite difference method

difference

$$\varphi'(x_i) \approx \frac{\varphi(x_{i+1}) - \varphi(x_i)}{x_{i+1} - x_i} = \frac{\varphi(x_{i+1}) - \varphi(x_i)}{h},$$

or, alternatively, by the left-sided finite difference

$$\varphi'(x_i) \approx \frac{\varphi(x_i) - \varphi(x_{i-1})}{x_i - x_{i-1}} = \frac{\varphi(x_i) - \varphi(x_{i-1})}{h}.$$

Combining these two approximations, we evidently see for the second derivative

$$\varphi''(x_i) \approx \frac{\varphi'(x_{i+1}) - \varphi'(x_i)}{x_{i+1} - x_i} \approx \frac{\frac{\varphi(x_{i+1})-\varphi(x_i)}{h} - \frac{\varphi(x_i)-\varphi(x_{i-1})}{h}}{h} = \frac{\varphi(x_{i+1}) - 2\varphi(x_i) + \varphi(x_{i-1})}{h^2}.$$

Naturally, here the symbol $\approx$ does not have a precise meaning. It simply indicates an *a priori* reasonable fashion of approximating the second derivative of a function at a grid point if we know the function values, or approximations of those, on the neighboring points of this same grid. Let us make this slightly more precise.

Theorem 2.1.1 *Let $\varphi \in C^4([0,1])$. For every $i \in \{1,\dots,N\}$, there is a θ_i, with $|\theta_i| < 1$ such that*

$$-\varphi''(x_i) = \frac{-\varphi(x_{i+1}) + 2\varphi(x_i) - \varphi(x_{i-1})}{h^2} + \frac{h^2}{12}\varphi^{(4)}(x_i + \theta_i h).$$

Proof. As is always the case for results of this type, the proof is based on Taylor's theorem. Since φ is assumed to be of class C^4 on $[0,1]$, we can use Taylor's theorem up to order 4 on every grid point. In particular, for every $i \in \{1,\dots,N\}$, there exists $\theta_i^+ \in]0,1[$ such that

$$\varphi(x_{i+1}) = \varphi(x_i) + h\varphi'(x_i) + \frac{h^2}{2}\varphi''(x_i) + \frac{h^3}{6}\varphi'''(x_i) + \frac{h^4}{24}\varphi^{(4)}(x_i + \theta_i^+ h).$$

Similarly, there is $\theta_i^- \in]0,1[$ such that

$$\varphi(x_{i-1}) = \varphi(x_i) - h\varphi'(x_i) + \frac{h^2}{2}\varphi''(x_i) - \frac{h^3}{6}\varphi'''(x_i) + \frac{h^4}{24}\varphi^{(4)}(x_i - \theta_i^- h).$$

Adding those two relations it follows that

$$\varphi(x_{i+1})+\varphi(x_{i-1})=2\varphi(x_i)+h^2\varphi''(x_i)+\frac{h^4}{24}\big(\varphi^{(4)}(x_i+\theta_i^+h)+\varphi^{(4)}(x_i-\theta_i^-h)\big).$$

Since $\varphi^{(4)}$ is continuous by assumption, we have by the intermediate value theorem that there are constants $y_i\in[x_i-\theta_i^-h,x_i+\theta_i^+h]$ such that

$$\frac{1}{2}\big(\varphi^{(4)}(x_i+\theta_i^+h)+\varphi^{(4)}(x_i-\theta_i^-h)\big)=\varphi^{(4)}(y_i).$$

The term on the left hand side is indeed the mean of the values taken by $\varphi^{(4)}$ at the end points of the interval $[x_i-\theta_i^-h,x_i+\theta_i^+h]$. Thus, we see that

$$-\varphi''(x_i)=\frac{-\varphi(x_{i+1})+2\varphi(x_i)-\varphi(x_{i-1})}{h^2}+\frac{h^2}{12}\varphi^{(4)}(y_i).$$

To conclude, we note that $y_i\in[x_i-\theta_i^-h,x_i+\theta_i^+h]\subset\,]x_{i-1},x_{i+1}[$, thus $\theta_i=\frac{y_i-x_i}{h}$ is such that $|\theta_i|<1$ and, trivially, $y_i=x_i+\theta_ih$. □

Corollary 2.1.2 *Under the assumptions of theorem 2.1.1, we have*

$$\max_{1\le i\le N}\left|-\varphi''(x_i)-\frac{-\varphi(x_{i+1})+2\varphi(x_i)-\varphi(x_{i-1})}{h^2}\right|\le\frac{h^2}{12}\|\varphi^{(4)}\|_{L^\infty}. \tag{2.1}$$

Proof. This is an immediate consequence of theorem 2.1.1. □

Remark 2.1.1 *i) The quantity on the left hand side of (2.1) is called the consistency error of the numerical method (a vector in $\mathbb{R}^N$, measured in the norm $\|.\|_\infty$), similarly to the terminology for approximation schemes for ODEs.*

ii) If φ is a polynomial of degree less than or equal to 3 the consistency error is zero.

iii) If we only assume φ to be of class C^3, we can only conclude that the consistency error is of $O(h)$ because in this case Taylor's theorem can only be applied up to order three. Similarly, if φ is only C^2, we can only say that the consistency error tends to 0 as h tends to zero, but a priori nothing more than that.

We can now apply these results to the boundary value problem (P). Let us denote the vector $(u(x_1),u(x_2),\dots,u(x_N))^T\in\mathbb{R}^N$ by $\overline{U}_h$ (be careful with this traditional notation which lacks a bit of coherence; h and N are linked by the relation $(N+1)h=1$ thus, in particular, the dimension of the vector $\overline{U}_h$ depends on h).

Corollary 2.1.3 *Suppose that the solution u to the boundary value problem is of class C^4 on $[0,1]$. Then there are points $y_i\in\,]x_{i-1},x_{i+1}[$ such that the vector $\overline{U}_h$ is a solution to the following linear system:*

$$A_h\overline{U}_h=F_h+K_h, \tag{2.2}$$

with

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2+c(x_1)h^2 & -1 & 0 & \dots & \dots & & \dots & 0 \\ -1 & 2+c(x_2)h^2 & -1 & 0 & & & & \\ 0 & -1 & \ddots & \ddots & & & & \\ \vdots & & \ddots & \ddots & \ddots & & & \\ \vdots & & 0 & -1 & 2+c(x_i)h^2 & -1 & 0 & \\ \vdots & & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & 0 & -1 & & 2+c(x_{N-1})h^2 & -1 \\ 0 & & \dots & & 0 & & -1 & 2+c(x_N)h^2 \end{pmatrix} \tag{2.3}$$

where $c_i = c(x_i)$*,*

$$F_h = \begin{pmatrix} f(x_1)+\frac{\alpha}{h^2} \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N)+\frac{\beta}{h^2} \end{pmatrix} \tag{2.4}$$

where $f_i = f(x_i)$ *and*

$$K_h = K_h(u) = -\frac{h^2}{12}\begin{pmatrix} u^{(4)}(y_1) \\ u^{(4)}(y_2) \\ \vdots \\ u^{(4)}(y_N) \end{pmatrix}. \tag{2.5}$$

Proof. The proof is almost immediate. Indeed, for every point x_i, $1 \le i \le N$, we have, by virtue of the differential equation,

$$-u''(x_i)+c(x_i)u(x_i) = f(x_i).$$

It suffices to replace $-u''(x_i)$ by the expressions derived in theorem 2.1.1. We then distinguish three cases, when $i = 1$, $2 \le i \le N-1$ and when $i = N$.

- The case $i = 1$. In this case it follows that

$$f(x_1) = -u''(x_1)+c(x_1)u(x_1) = \frac{-u(x_2)+2u(x_1)-u(x_0)}{h^2}+c(x_1)u(x_1)+\frac{h^2}{12}u^{(4)}(y_1).$$

Since $u(x_0) = u(0) = \alpha$ is determined by the boundary condition, we can move it to the right hand side and thus obtain

$$\frac{-u(x_2)+2u(x_1)}{h^2}+c(x_1)u(x_1) = f(x_1)+\frac{u(x_0)}{h^2}-\frac{h^2}{12}u^{(4)}(y_1) = f(x_1)+\frac{\alpha}{h^2}-\frac{h^2}{12}u^{(4)}(y_1),$$

or, alternatively,

$$\frac{1}{h^2}[(2+c(x_1)h^2)u(x_1)-u(x_2)] = f(x_1)+\frac{\alpha}{h^2}-\frac{h^2}{12}u^{(4)}(y_1).$$

- The case $i = N$. Analogously we have

$$f(x_N) = \frac{-u(x_{N+1}) + 2u(x_N) - u(x_{N-1})}{h^2} + c(x_N)u(x_N) + \frac{h^2}{12}u^{(4)}(y_N).$$

Since $u(x_{N+1}) = u(1) = \beta$, we thus obtain

$$\frac{1}{h^2}[-u(x_{N-1}) + (2 + c(x_N)h^2)u(x_N)] = f(x_N) + \frac{\beta}{h^2} - \frac{h^2}{12}u^{(4)}(y_N).$$

- The case $2 \leq i \leq N-1$. There is no special treatment required, we have

$$f(x_i) = \frac{-u(x_{i+1}) + 2u(x_i) - u(x_{i-1})}{h^2} + c(x_i)u(x_i) + \frac{h^2}{12}u^{(4)}(y_i),$$

which is exactly the i-th row of the desired linear system. □

Remark 2.1.2 *i) The grid point values of the solution satisfy the linear system (2.2) exactly: there is no approximation here. Of course, there is no magic bullet: we cannot solve this linear system because the right hand side contains the unknown term $K_h = K_h(u)$! Nevertheless, this expression is useful for the convergence analysis of the method.*

ii) On the other hand, the matrix A_h, the dimension $N \times N$ and the vector b_h are all known.

iii) We denote by $||\cdot||_\infty$ the ℓ^∞-norm on $\mathbb{R}^N$ (defined in proposition 1.5.3). The quantity $||K_h||_\infty$ is the consistency error computed over u. By applying corollary 2.1.2 we find that it satisfies

$$||K_h||_\infty \leq \frac{h^2}{12}||u^{(4)}||_{L^\infty} \to 0 \text{ as } N \to +\infty, \tag{2.6}$$

because $h = \frac{1}{N+1}$. We note that the space $\mathbb{R}^N$ changes as we vary N and if we wanted to be precise, we could have written $||\cdot||_{\infty,N}$ in order to specify that we took the ℓ^∞-norm on the space $\mathbb{R}^N$. In the interest of notational simplicity we shall not use this more complicated notation here.

Idea of the finite difference method: Since $K_h(u)$ is small when h is small (assuming, of course, that u is sufficiently regular), we decide to omit it from the right hand side and consider the following *discrete* problem

$$\boxed{(S_h) \quad \begin{cases} \text{Find } U_h \in \mathbb{R}^N \text{ such that} \\ \\ A_h U_h = F_h. \end{cases}}$$

Thus it remains to solve a system of N linear equations with N unknowns given by the components $(u_1, \ldots, u_n)$ of the vector U_h, where the matrix A_h and the right hand side F_h are known. Several questions arise:

1) Is the matrix A_h invertible? If this is not the case, there is no chance of calculating this new vector U_h.

2) Assuming that this is the case, in general we will have $U_h \neq \overline{U}_h$, *i.e.* $u_i \neq u(x_i)$. We thus commit an *error* which we must be able to estimate. Do we then have $U_h - \overline{U}_h \to 0$ in a reasonable sense, and at what rate in terms of $h \to 0$ (or of $N \to +\infty$)? In other words, have we indeed constructed approximations to the solution values at the grid points and how good is the quality of these approximations?

Definition 2.1.1 *We say that the method is*
i) convergent if

$$\max_{1\le i\le N} |u_i - u(x_i)| \to 0 \text{ as } N \to +\infty.$$

ii) of order p if

$$\max_{1\le i\le N} |u_i - u(x_i)| \le C(u)h^p,$$

where $C(u)$ is a constant independent of u.

Thus, if the method is convergent, then we have indeed obtained approximations to the exact values (here uniform ones, but we could use norms other than $\|\cdot\|_\infty$ to measure this convergence) and these approximations converge faster if the order of the method is higher.

In summary, it is now a question of carrying out what is called the *numerical analysis* of the method: Is it well-defined, is it convergent and of what order?

To begin with, let us deal with the first question, namely if it is reasonable to want to compute the vector U_h.

Theorem 2.1.4 *If $c(x) \ge 0$ on $[0,1]$, then the matrix A_h is symmetric, positive definite and thus invertible.*

Proof. Clearly A_h is symmetric tridiagonal, regardless of the sign sign of c. Suppose now that $c \ge 0$. Let $V \in \mathbb{R}^N \setminus \{0\}$. We need to evaluate the scalar product $V^T A_h V = \langle A_h V, V\rangle$. We have

$$h^2 V^T A_h V = h^2 \sum_{i,j} a_{ij} v_i v_j = \sum_{i=1}^{N} (2 + c(x_i)h^2) v_i^2 - 2\sum_{i=1}^{N-1} v_i v_{i+1} \ge 2\sum_{i=1}^{N} v_i^2 - 2\sum_{i=1}^{N-1} v_i v_{i+1},$$

because $c_i = c(x_i) \ge 0$. Consequently, by rearranging the above two sums, it follows that

$$\begin{aligned} h^2 V^T A_h V &\ge v_1^2 + (v_1^2 - 2v_1v_2 + v_2^2) + (v_2^2 - 2v_2v_3 + v_3^2) \\ &\qquad + \cdots + (v_{N-1}^2 - 2v_{N-1}v_N + v_N^2) + v_N^2 \\ &= v_1^2 + (v_1 - v_2)^2 + (v_2 - v_3)^2 + \cdots + (v_{N-1} - v_N)^2 + v_N^2 \ge 0. \end{aligned}$$

Therefore the A_h is positive. Moreover, if $V^T A_h V = 0$, we see that we necessarily have

$$v_1 = 0, v_1 - v_2 = 0, v_2 - v_3 = 0, \dots, v_{N-1} - v_N = 0 \text{ and } v_N = 0,$$

meaning that in fact $V = 0$. The matrix is thus positive definite. We thus immediately deduce that it is invertible, since

$$V \in \ker A_h \Leftrightarrow A_h V = 0 \Rightarrow V^T A_h V = 0 \Leftrightarrow V = 0,$$

implies $\ker A_h = \{0\}$. □

Corollary 2.1.5 *If $c(x) \ge 0$ on $[0,1]$, then for all f and N, there is a unique vector $U_h \in \mathbb{R}^N$ which solves the finite difference problem (S_h).*

Remark 2.1.3 *It is interesting to note that the same assumption on the sign of c, which guarantees existence of the solution to the discrete problem, also ensures the existence of the solution to the boundary value problem itself.*

2.2 Convergence analysis

In the following, we always assume that the solution u is of class $\mathcal{C}^4$ on $[0,1]$. We will measure the difference between the discrete solution U_h and the values of the exact continuous solution at the gridpoints $\overline{U}_h$ using the uniform norm $\|U_h - \overline{U}_h\|_\infty$ which we have used in the definition of convergence (definition 2.1.1). All norms of $\mathbb{R}^N$ are equivalent but not all of them are suitable for the study of convergence because, in the limit, N tends to infinity. We note that the norm $\|\overline{U}_h\|_\infty$ remains a sensible quantity in this limit whereas, for example, $\|\overline{U}_h\|_1 \to \infty$ as $N \to \infty$. Moreover, as we will see, it is possible to evaluate the norm $\|A_h^{-1}\|_\infty$ and thus this norm will allow us to estimate the error $U_h - \overline{U}_h$.

Proposition 2.2.1 *We have*

$$\|U_h - \overline{U}_h\|_\infty \leq \|A_h^{-1}\|_\infty \left(\frac{h^2}{12} \|u^{(4)}\|_{L^\infty} \right).$$

Proof. Let us write out the linear systems satisfied respectively by U_h and $\overline{U}_h$: by (2.2)

$$\begin{aligned} A_h U_h &= F_h, \\ A_h \overline{U}_h &= F_h + K_h. \end{aligned}$$

Subtracting these two relations, we find

$$A_h(U_h - \overline{U}_h) = -K_h \Longleftrightarrow U_h - \overline{U}_h = -A_h^{-1} K_h$$

because A_h is invertible. Consider the norm $\|\cdot\|_\infty$ of these vectors, we find by the definition of subordinate matrix norms

$$\|U_h - \overline{U}_h\|_\infty \leq \|A_h^{-1}\|_\infty \|K_h\|_\infty.$$

The result then immediately follows from (2.6). □

The convergence study thus has been reduced to an understanding of the behavior of the quantity $\|A_h^{-1}\|_\infty$ (which no longer depends on u) in terms of N and h.

We note that, similarly to our treatment of numerical schemes for ordinary differential equations, the error is made up of two components: the consistency error $\|K_h\|_\infty$, which we have already estimated, and the quantity $\|A_h^{-1}\|_\infty$ which does not depend on the solution and which we can call the *stability constant*.

Our estimate of $\|A_h^{-1}\|_\infty$ is related to the properties of the matrix A_h which in turn arises from the modeling of a physical problem. The proof requires several steps. We begin with defining the notion of a *monotone* matrix.

Definition 2.2.1 *i) We introduce a partial ordering on* $\mathbb{R}^N$ *by writing*

$$V \geq W \text{ if and only if } \forall i,\, v_i \geq w_i.$$

ii) Similarly, for matrices, we say that

$$A \geq B \text{ if and only if } \forall i, j,\; a_{ij} \geq b_{i,j}.$$

iii) We say that a matrix A is monotone if it is invertible and if $A^{-1} \geq 0$.

Beware that the ordering on matrices thus defined has nothing to do with the one defined on symmetric matrices using quadratic forms: there are matrices that are positive in the sense of quadratic forms but not positive in the present sense and vice versa (exercise: find examples).

Let us provide an alternative characterization of monotonicity which is slightly more useful than the definition.

Lemma 1 *Let A be an $N \times N$-matrix. A is monotone if and only if for every vector V we have $AV \geq 0$ implies $V \geq 0$.*

Proof. We prove the condition is necessary and sufficient separately.

• Necessary condition. Let A be a monotone matrix. Consider a vector $V \in \mathbb{R}^N$ such that $AV \geq 0$, i.e. $(AV)_i \geq 0$ for every index $1 \leq i \leq N$. Naturally, $V = A^{-1}(AV)$, which means in terms of components

$$v_i = \sum_{j=1}^{N} (A^{-1})_{ij}(AV)_j.$$

Now we have $(A^{-1})_{ij} \geq 0$ because A is monotone, and thus it follows that $v_i \geq 0$, i.e. $V \geq 0$.

• Sufficient condition. Let A be a matrix such that $Av \geq 0$ implies $V \geq 0$. We firstly show that it is invertible. For this, let W be an element of the kernel of A, i.e. so that $AW = 0$. Since, of course, $AW = 0 \geq 0$, we thus deduce that $W \geq 0$. Similarly, $-W$ belongs to the kernel and thus $-W \geq 0$. Therefore, $w_i = 0$ for every i and the kernel consists of only the zero vector $\ker A = \{0\}$.

Denote now by b_j the j-th column vector of the matrix A^{-1}. This means that $A^{-1}e_j = b_j$ where e_j is the j-th vector in the standard basis. In other words $Ab_j = e_j$, where clearly $e_j \geq 0$. Consequently, we deduce that $b_j \geq 0$ for every j, which immediately implies that $A^{-1} \geq 0$. □

We continue with a property that will be useful in a later proof.

Lemma 2 *Let A and B be two monotone matrices, with $B \geq A$. Then*

$$A^{-1} \geq B^{-1}$$

and

$$\|A^{-1}\|_\infty \geq \|B^{-1}\|_\infty.$$

Proof. We note that for every pair of invertible matrices A and B, we have the identity:

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1},$$

which follows immediately from $A^{-1} - B^{-1} = A^{-1}(BB^{-1}) - (A^{-1}A)B^{-1}$. If the matrices are monotone, and if in addition $B - A \geq 0$ this implies that $A^{-1} - B^{-1} \geq 0$ because a product of positive matrices is clearly positive (an immediate consequence of the definition). We have thus shown that

$$A^{-1} \geq B^{-1}.$$

Let us consider now two matrices such that $B \geq A \geq 0$. This means simply that $b_{ij} \geq a_{ij} \geq 0$ for all i, j. Consequently,

$$\|B\|_\infty = \max_i \sum_j |b_{ij}| = \max_i \sum_j b_{ij} \geq \max_i \sum_j a_{ij} = \max_i \sum_j |a_{ij}| = \|A\|_\infty.$$

We can therefore conclude the proof by applying this result to $A^{-1} \geq B^{-1} \geq 0$. □

Remark 2.2.1 *If A is a positive matrix, then its ℓ^∞-norm is given by*

$$\|A\|_\infty = \max_i \sum_j |a_{ij}| = \max_i \sum_j a_{ij} = \|Ae\|_\infty,$$

where $e = (1,\dots,1)^T$ is the constant vector with all entries equal to 1. Similarly, if A is monotone, because A^{-1} is positive, then its ℓ^∞-norm is given by $\|A^{-1}e\|_\infty$, i.e.

$$\|A^{-1}\|_\infty = \|V\|_\infty,$$

where $V \in \mathbb{R}^N$ is the unique solution to $AV = e$.

An important class of monotone matrices is given by the following definition.

Definition 2.2.2 *A matrix A is called an M-matrix if and only if it satisfies the following three properties:*
i) Positive values on the principle diagonal: $a_{i,i} > 0, \quad i = 1,\dots,N$.
ii) Non-positive remaining coefficients: $a_{i,j} \leq 0, \quad i \neq j$.
iii) Strictly diagonally dominant: there is a $\mu > 0$ such that $\sum_j a_{i,j} \geq \mu, \quad i = 1,\dots,N$.

Proposition 2.2.2 *An M-matrix A is monotone, and further satisfies $\|A^{-1}\|_\infty \leq \mu^{-1}$.*

Proof. We shall use the characterisation provided in lemma 1. Let $V \in \mathbb{R}^N$ be such that $AV \geq 0$. Consider i^* such that $v_{i^*} = \min v_i$. We may thus write

$$0 \leq (AV)_{i^*} = \sum_{j=1}^N a_{i^*,j} v_j \leq (\sum_{j=1}^N a_{i^*,j}) v_i^*,$$

because, by properties i) and ii) we have

$$\sum_{j=1}^n a_{i^*,j}(v_j - v_{i^*}) \leq 0.$$

By property iii), it therefore follows that, for some $\mu > 0$,

$$\mu v_i^* \geq 0,$$

which shows that $V \geq 0$. In order to control the norm $\|A^{-1}\|_\infty$, we can use remark 2.2.1: if V is such that $AV = e$, and if i_0 is such that $v_{i_0} = \max v_i = \|V\|_\infty$, then we have, in particular, that

$$1 = (AV)_{i_0} = \sum_{j=1}^N a_{i_0,j} v_j \geq (\sum_{j=1}^N a_{i_0,j}) v_{i_0},$$

because, by properties i) and ii), we have

$$\sum_{j=1}^n a_{i_0,j}(v_j - v_{i_0}) \geq 0,$$

and by property iii) it follows that $v_{i_0} \leq \mu^{-1}$, thus

$$\|A^{-1}\|_\infty = \|V\|_\infty \leq \mu^{-1}.$$

□

If we return to the matrix A_h, we see that it is "almost" an M-matrix, in the sense that the sign properties i) and ii) are true but the diagonal dominance iii) is not strict except for the first and final line $i = 1$ and $i = N$. For the remaining values of i we have

$$\sum_j (A_h)_{i,j} = \frac{1}{h^2}(-1 - 1 + 2 + c(x_i)h^2) = c(x_i) \geq 0$$

because $c \geq 0$ but we do not necessarily have $c(x_i) \geq \mu > 0$ for every i. Nevertheless, we may prove that A_h is monotone using the following proposition.

Proposition 2.2.3 *If A is invertible and satisfies properties i), ii), as well as*

$$\sum_j a_{i,j} \geq 0, \quad i = 1, \dots, N,$$

then A is monotone.

Proof. We note that for every $\varepsilon > 0$ the matrix

$$A_\varepsilon = A + \varepsilon I$$

is an M-matrix for $\mu = \varepsilon$. Consequently A_ε is invertible and $A_\varepsilon^{-1} \geq 0$. Since $A_\varepsilon \to A$ as $\varepsilon \to 0$ and since A is invertible, we have that $A_\varepsilon^{-1} \to A^{-1}$. Because the space of matrices $M_N(\mathbb{R})$ is a vector space of finite dimension $N \times N$, this convergence can be expressed in any norm and, in particular, means that every coefficient $(A_\varepsilon^{-1})_{i,j} \geq 0$ of A_ε converges to the coefficient $(A^{-1})_{i,j}$ of A^{-1} which is thus positive. This shows that, indeed, A is monotone. □

This observation can thus be applied to the matrix A_h which yields the following result.

Proposition 2.2.4 *If $c \geq 0$ then, for every $h > 0$, the matrix A_h is monotone.*

Remark 2.2.2 *The monotonicity of the matrix A_h is the discrete analogue of the monotonicity in the boundary value problem (cf. theorem 1.4.4). Indeed, by definition (2.4) of F_h, if the assumptions of theorem 1.4.4 are satisfied, then $F_h \geq 0$. If $F_h \geq 0$ and $A_h U_h = F_h$, we deduce that $U_h \geq 0$. We thus call this the discrete maximum principle.*

We may now establish an estimate on the ℓ^∞-norm of A_h^{-1}.

Proposition 2.2.5 *For every $h > 0$, the inverse of the matrix A_h satisfies the estimate*

$$\|A_h^{-1}\|_\infty \leq \frac{1}{8}.$$

Proof. Let us introduce the matrix

$$A_{0h} = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & & \\ 0 & -1 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & -1 & 2 & -1 \\ 0 & \dots & & 0 & -1 & 2 \end{pmatrix} \tag{2.7}$$

which corresponds to the case when $c = 0$. We already know that $A_h^{-1} \geq 0$ and that $A_{0h}^{-1} \geq 0$. Moreover,

$$A_h - A_{0h} = \begin{pmatrix} c(x_1) & 0 & \dots & 0 \\ 0 & c(x_2) & \dots & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & c(x_N) \end{pmatrix} \geq 0,$$

since $c(x_i) \geq 0$. Applying lemma 2, we thus find that

$$A_{0h}^{-1} \geq A_h^{-1} \geq 0.$$

and also that

$$\|A_h^{-1}\|_\infty \leq \|A_{0h}^{-1}\|_\infty,$$

which leaves us only to deal with the case $c = 0$.

Thus we are left to estimate the ℓ^∞-norm of the positive matrix A_{0h}^{-1}. Let hence be

$$U_{0h} = A_{0h}^{-1}e \Longleftrightarrow A_{0h}U_{0h} = e.$$

The vector U_{0h} is therefore none other than the solution to the finite difference equations associated with the particular boundary value problem

$$\begin{cases} -u_0''(x) = 1 \text{ in }]0,1[, \\ u_0(0) = u_0(1) = 0. \end{cases}$$

At this point it is easy to compute the exact solution to this problem: it is given by the function $u_0(x) = \frac{1}{2}x(1-x)$. Hence it turns out that it is a second degree polynomial and therefore $u_0^{(4)} = 0$, i.e. its consistency error is zero, $K_h(u_0) = 0$. We thus deduce, by corollary 2.1.3 that the vector $\overline{U}_{0h}$ with entries $u_0(x_i)$ is a solution to the *same* linear system as the vector U_{0h}. Hence we can express the latter as follows :

$$(U_{0h})_i = u_0(x_i) = \frac{1}{2}ih(1-ih).$$

Therefore we have

$$\|A_{0h}^{-1}\|_\infty = \|U_{0h}\|_\infty \leq \max_{x\in[0,1]}\left\{\frac{1}{2}x(1-x)\right\} = \frac{1}{8},$$

which immediately implies the desired result. □

Combining this with the error estimate from proposition 2.2.1, we thus obtain the following convergence result.

Theorem 2.2.1 *Assume that c and f are of class $\mathcal{C}^2$. Then we have the following upper bound on the error:*

$$||U_h - \overline{U}_h||_\infty = \max_{1 \le i \le N} |u_i - u(x_i)| \le \frac{h^2}{96} ||u^{(4)}||_{L^\infty}.$$

Remark 2.2.3 *i) We have shown that the finite difference method is convergent at second order (in the ℓ^∞-norm). The computed values are thus located in a neighborhood of width $O(h^2)$ around the graph of the exact solution.*

ii) It can be shown that, in general, the convergence is no faster than that, i.e., there exists data $f \in \mathcal{C}^2$ such that the exact solution and its corresponding approximations satisfy the estimate $||U_h - \overline{U}_h||_\infty = Ch^2(1 + \delta(h))$ with $\delta(h) \to 0$ as $h \to 0$ for some $C > 0$.

iii) The error estimate depends on the fourth derivative of the unknown u. It is therefore not an explicit function of the data of the problem, and does not provide any quantitative control on the error that is committed. We call this an a priori estimate.

We derived an estimate in the $||.||_\infty$-norm. We may also obtain a similar result in other norms. Recall that all norms on a finite dimensional vector space are equivalent, however, the constants involved in the inequalities between the norms may depend on the dimension N of the space. For example, the Euclidean norm on $\mathbb{R}^N$ is not interesting for our purposes because, as $h \to 0$, then $N \to \infty$ and thus the number of terms in the sum increases which typically makes this norm tend to $+\infty$ if we apply it to $\overline{U}_h$. In order to obtain a convergence result which we can interpret in terms of the L^2-norm, we firstly need to define a *discrete L^2-norm* which makes sense as $h = \frac{1}{N+1} \to 0$.

Definition 2.2.3 *For every $V \in \mathbb{R}^{N+2}$, $V = (v_0, ..., v_{N+1})^T$, we define the discrete L^2-norm $||.||_{2,\Delta}$ by*

$$||V||_{2,\Delta}^2 = h\Big(\frac{1}{2}v_0^2 + \sum_{i=1}^{N} v_i^2 + \frac{1}{2}v_{N+1}^2\Big).$$

It is easy to see that this indeed defines a norm on $\mathbb{R}^{N+2}$. It is the norm in $L^2(0,1)$ of the piecewise constant function v_Δ which takes the value v_i on the interval $](i-1/2)h, (i+1/2)h]$, $i = 1, ..N$, and v_0 on $[0, h/2]$, v_N on $]1 - h/2, 1]$. The notation using the symbol Δ simply emphasizes the *discrete* nature of this norm (we could have alternatively used the index h). Another interpretation is that $||V||_{2,\Delta}^2$ is the integral of the piecewise affine function which takes values v_i^2 at the points x_i. In other words, if the v_i are the values of some function φ at the points x_i then $||V||_{2,\Delta}^2$ is the approximation of $\int_0^1 |\varphi|^2$ using the trapezoidal rule.

An immediate observation is that we have

$$||V||_{2,\Delta}^2 \le h(N+1) \max_{i=0,...,N+1} |v_i|^2 = ||V||_\infty^2,$$

or, alternatively,

$$||V||_{2,\Delta} \le ||V||_\infty,$$

which is consistent with the classical property $||\varphi||_{L^2(0,1)} \le ||\varphi||_{L^\infty(0,1)}$ that holds for functions defined on $]0,1[$. Until now, $\overline{U}_h$ and U_h have been vectors in $\mathbb{R}^N$, but we can also extend them to vectors in $\mathbb{R}^{N+2}$ by adding the boundary values imposed at $x_0 = 0$ et $x_{N+1} = 1$, and writing with abuse of notation

$$\overline{U}_h := (\alpha, u(x_1), .., u(x_N), \beta)^T \quad \text{and} \quad U_h := (\alpha, u_1, .., u_N, \beta)^T.$$

We thus deduce from theorem 2.2.1 a result concerning the convergence in the discrete L^2-norm.

Theorem 2.2.2 *Assume that c and f are of class $\mathcal{C}^2$, $c \geq 0$, and let u be the solution to the boundary value problem (P), and $U_h \in \mathbb{R}^{N+2}$ the solution to the discrete problem (S_h) extended by the values at x_0 and x_{N+1}. Then we have the following upper bound for the error:*

$$||U_h - \overline{U}_h||_{2,\Delta} \leq \max_{1\leq i\leq N} |u_i - u(x_i)| \leq \frac{h^2}{96} ||u^{(4)}||_{L^\infty}.$$

The conclusion of theorem 2.2.2 implies then that $\bar{u}_\Delta$ (piecewise constant function constructed from the values u_i) converges to u in $L^2(0,1)$. Indeed, if we denote by u_Δ the piecewise constant function constructed from the values $u(x_i)$, we can express

$$||u - \bar{u}_\Delta||_{L^2} \leq ||u - u_\Delta||_{L^2} + ||u_\Delta - \bar{u}_\Delta||_{L^2}.$$

the final term on the right hand side is exactly $||\overline{U}_h - U_h||_{2,\Delta}$ and it converges to 0 when $h \to 0$ by the aforementioned result. The first term also tends to zero (since we approximate a continuous function by a sequence of step functions on $[0,1]$). Of course, all of this can be generalised to discrete L^p-norms which we define by

$$||V||^p_{p,\Delta} = h(\frac{1}{2}v_0^p + \sum_{i=1}^{N} v_i^p + \frac{1}{2}v_{N+1}^p).$$

2.3 An excursion to dimension two

At this point of the course, we cannot say much about boundary value problems in dimension greater than one. Nevertheless, in particular cases, a certain number of properties can be proved in a similar manner to our above discussions.

Let us thus consider the open square $\Omega =]0,1[\times]0,1[$. We denote the coordinates by x and y. Let $f\colon \Omega \to \mathbb{R}$ and $g\colon \partial\Omega \to \mathbb{R}$ be two continuous functions. The boundary value problem will consist of finding a function $u\colon \bar{\Omega} \to \mathbb{R}$ belonging to $\mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ such that:

$$\begin{cases} -\Delta u = f \ \text{ in } \Omega, \\ u = g \ \text{ on } \partial\Omega. \end{cases} \tag{2.8}$$

Theorem 2.3.1 *(Maximum principle). Let $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ such that $\Delta u \leq 0$. Then u attains its minimum on the boundary $\partial\Omega$. Similarly, if $\Delta u \geq 0$, then u attains its maximum on the boundary $\partial\Omega$.*

Proof. Let $v \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ be a function which has a relative minimum at an *interior* point $(x_0, y_0) \in \Omega$. The map $v_{y_0}\colon\,]0,1[\to \mathbb{R}$, $t \mapsto v(t, y_0)$ thus admits, in particular, a relative minimum at $t = x_0$, which is an interior point of an interval over which this function is of class $\mathcal{C}^2$. By Taylor's theorem we thus deduce that $\frac{d^2 v_{y_0}}{dt^2}(x_0) \geq 0$ (proof by contradiction). By definition of partial derivatives, this means exactly that $\frac{\partial^2 v}{\partial x^2}(x_0, y_0) \geq 0$. Similarly, we can show that $\frac{\partial^2 v}{\partial y^2}(x_0, y_0) \geq 0$. By adding those two inequalities, we find

$$\Delta v(x_0, y_0) \geq 0$$

at every interior relative minimum of v.

Let now $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ such that $-\Delta u \geq 0$ in Ω. We shall reason by contradiction and suppose thus that u does not attain its minimum on $\partial\Omega$. Hence, the function attains its minimum at an interior point $(x_0, y_0) \in \Omega$, and we have

$$M = u(x_0, y_0) < N = \min_{(x,y) \in \partial\Omega} u(x, y).$$

Let us now introduce an auxiliary function

$$u_\varepsilon(x, y) = u(x, y) - \varepsilon(x^2 + y^2) \quad \text{with} \quad 0 < \varepsilon < \frac{(N - M)}{2}.$$

We see then that

$$u_\varepsilon(x_0, y_0) \leq M,$$

and that for every $(x, y) \in \partial\Omega$ we have

$$u_\varepsilon(x, y) \geq N - 2\varepsilon > N - (N - M) = M.$$

This shows that u_ε attains its minimum at an interior point $(x_1, y_1) \in \Omega$, and therefore $\Delta u_\varepsilon(x_1, y_1) \geq 0$. On the other hand, as $-\Delta u \geq 0$ in Ω, we have

$$-\Delta u_\varepsilon(x, y) = -\Delta u(x, y) + \varepsilon\Delta(x^2 + y^2) = -\Delta u(x, y) + 4\varepsilon > 0.$$

In particular, at $(x_1, y_1) \in \Omega$, we obtain

$$-\Delta u_\varepsilon(x_1, y_1) > 0,$$

which is a contradiction. We can show in a similar manner that if $\Delta u \geq 0$, then u attains its maximum on the boundary $\partial\Omega$. □

Remark 2.3.1 *We cannot perform the proof by contradiction directly on u. Indeed, this would result in $-\Delta u(x_0, y_0) \leq 0$ and $-\Delta u(x_0, y_0) \geq 0$, which is not a contradiction and instead just implies that $-\Delta u(x_0, y_0) = 0$. That is as far as this line of reasoning would go.*

From the maximum principle we can immediately deduce the following properties.

Corollary 2.3.2 *Let $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ such that $-\Delta u \geq 0$ in Ω and $u \geq 0$ on $\partial\Omega$. Then $u \geq 0$ on the whole of $\bar{\Omega}$. Similarly, if $-\Delta u \leq 0$ in Ω and $u \leq 0$ on $\partial\Omega$, then $u \leq 0$ on the whole of $\bar{\Omega}$. In particular, if $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ is such that $-\Delta u = 0$ in Ω and $u = 0$ on $\partial\Omega$, then $u = 0$ on $\bar{\Omega}$.*

From this corollary we can immediately deduce a *uniqueness* result.

Theorem 2.3.3 *The boundary value problem (2.8) admits at most one solution.*

Proof. Let u_1 and u_2 be two solutions of the boundary value problem (2.8). Take $v = u_1 - u_2$. It follows thus that on the one hand we have $\Delta v = \Delta u_1 - \Delta u_2 = f - f = 0$ in Ω and on the other hand $v = g - g = 0$ on $\partial\Omega$. By corollary 2.3.2, we therefore have $v = 0$. □

Remark 2.3.2 *The above arguments remain valid when Ω is a bounded open subset of $\mathbb{R}^n$, for any n, and not just for the two-dimensional case (exercise: prove this).*

The question of *existence* of a solution to problem (2.8) is much more delicate. As a result of the simple geometry of the domain, in the present case this question can be understood with the help of Fourier series. We note that any function φ defined on $[0,1]$ can be extended to an odd function on the interval $[-1,1]$, and then to the whole of $\mathbb{R}$ by 2-periodicity. The resulting function $\tilde{\varphi}$ can be expanded in a Fourier series of the form

$$\varphi(x) = \sum_{k=1}^{\infty} b_k \sin(k\pi x).$$

Indeed, the coefficients a_k of the functions $\cos(k\pi x)$ which appear in a general Fourier series are zero in this case because φ is odd.

It is important to pay attention to the sense in which this series converges. We note that all of the functions $\sin(k\pi x)$ are zero at 0 and 1 which renders uniform convergence impossible if φ does not vanish at these points. Using the Dirichlet conditions (a.k.a. Dirichlet theorem) we can prove (exercise) pointwise convergence of this series on $]0,1[$ if φ is of class $\mathcal{C}^1$ on $]0,1[$. By studying the coefficients b_k more carefully, we can also prove uniform convergence if φ is of class $\mathcal{C}^1$ and vanishes at 0 and 1.

For a more general function φ, convergence occurs in $L^2(0,1)$ so long as φ belongs to this space and the coefficients $(b_k)_{k\geq 1}$ belong to the space ℓ^2 of square summable sequences. This reflects the fact that the family of functions $x \mapsto \sin(k\pi x)$ is an orthogonal basis of $L^2(0,1)$. We refer the reader (if part of course 4MA006) to chapter 6 for more details on these fundamental concepts.

Similarly, in dimension 2 we can expand any function $\varphi \in L^2(\Omega)$ with $\Omega =]0,1[^2$ in the following way

$$\varphi(x,y) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} b_{k,l} \sin(k\pi x) \sin(l\pi y),$$

where the series converges in the L^2-norm. We can also prove the pointwise convergence on $]0,1[^2$ under certain regularity assumptions on φ, for example under the assumption that $\varphi \in \mathcal{C}^1(\bar{\Omega})$. The functions

$$s_{k,l} = \sin(k\pi x) \sin(l\pi y),$$

form an orthogonal basis, vanish on the boundary of Ω, and are eigenfunctions of the operator $-\Delta$ because they satisfy

$$-\Delta s_{k,l} = \pi^2(k^2+l^2)s_{k,l}.$$

By expanding the right hand side of equation (2.8) as follows that

$$f = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} f_{k,l}s_{k,l},$$

we see that it is natural to consider a solution to the form

$$u = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} u_{k,l}s_{k,l}, \tag{2.9}$$

where the coefficients are given by

$$u_{k,l} = \frac{1}{\pi^2(k^2+l^2)} f_{k,l}.$$

This is because, by differentiating each term, we have that formally $-\Delta u = f$ and if the series converges uniformly then $u = 0$ on $\partial\Omega$. In order to establish a theorem it is necessary to decide on assumptions which allow us to prove this result. For example, we obtain the following conclusion.

Proposition 2.3.1 *Suppose that $f \in C^0(\bar{\Omega})$ is such that the functions $(f_{k,l})_{k,l\geq 1}$ form a sequence in ℓ^1, meaning that*

$$\sum_{k,l\geq 1} |f_{k,l}| < \infty.$$

Then the function u given in (2.9) *belongs to $C^2(\bar{\Omega})$ and is a solution to equation* (2.8).

Remark 2.3.3 *Of course, regularity assumptions on f are hidden in the assumption of absolute summability of the coefficients $f_{k,l}$ because the continuity of f alone is not sufficient for this to hold. In particular it is not enough for f to belong to $C^0(\bar{\Omega})$ in order for its Fourier series to be normally convergent. We see also that the assumption of this convergence implies that f vanishes on the boundary, which we do not necessarily want to assume (as opposed to u vanishing on the boundary which of course is a perfectly valid assumption). We cannot do without the proof of the convergence of the series and of the series of derivatives. Indeed, one can construct continuous functions which are nowhere differentiable using a series of functions which are arbitrarily regular. Thus the differentiability of a series of functions is nothing obvious, nor is the term-by-term differentiability of this series.*

2.4 The finite difference method in dimension two

Let us provide an overview of what can be done with finite differences on the previous problem. Let $N \geq 1$ be an integer, take $h = \frac{1}{N+1}$ and construct the discretisation grid composed of points or *nodes* $z_{ij} = (ih, jh)$ where $0 \leq i \leq N+1$ and $0 \leq j \leq N+1$. There is thus a total of $(N+2)^2$ points on this grid, of which N^2 are located in the interior ($1 \leq i \leq N$ et $1 \leq j \leq N$) and $4N+4$

are located on the boundary ($i = 0$ or $i = N+1$ or $j = 0$ or $j = N+1$). If φ is a function defined on $\bar{\Omega}$, we will write $\varphi(x,y) = \varphi(z)$ with $z = (x,y)$. In a manner analogous to dimension one, we introduce the discretisation of the Laplacian.

Definition 2.4.1 *Let $\varphi \in C^0(\bar{\Omega})$. We call $\Delta_h\varphi$ the 5-point discrete Laplacian of φ defined by*

$$\Delta_h\varphi(z_{ij}) = \frac{1}{h^2}[-4\varphi(z_{ij}) + \varphi(z_{i-1,j}) + \varphi(z_{i+1,j}) + \varphi(z_{i,j-1}) + \varphi(z_{i,j+1})]$$

for every interior grid point z_{ij} $(1 \le i \le N$ and $1 \le j \le N)$.

Remark 2.4.1 *The terminology is clear, the 5-point discrete Laplacian uses the four nearest neighbors of the point considered on the grid. There are variants with 9 points etc. Note the distinction between $\Delta\varphi$ which is a function defined on Ω and $\Delta_h\varphi$ which is defined only on the grid.*

Let us now estimate the consistency error.

Theorem 2.4.1 *Let $\varphi \in C^4(\bar{\Omega})$. Then for every $i,j \in \{1,\dots,N\}$,*

$$|\Delta\varphi(z_{ij}) - \Delta_h\varphi(z_{ij})| \le \frac{h^2}{12}\Big(\|\frac{\partial^4\varphi}{\partial x^4}\|_{L^\infty} + \|\frac{\partial^4\varphi}{\partial y^4}\|_{L^\infty}\Big),$$

where for every continuous function v on $\bar{\Omega}$, we denoted $\|v\|_{L^\infty} = \|v\|_{L^\infty(\bar{\Omega})} = \max_{x\in\bar{\Omega}}|v(x)|$.

Proof. Let us introduce the function $\theta_1^{ij}(t) = \varphi(ih+t, jh)$ which is defined and of class C^4 in a neighborhood of zero in t. This neighborhood contains, at the very least, the interval $]-h,h[$, for every i and j. Then we have, by the definition of partial derivatives, that $\frac{\partial^k\varphi}{\partial x^k}(ih+t, jh) = \frac{d^k\theta_1^{ij}}{dt^k}(t)$ for $0 \le k \le 4$. The same is true for $\theta_2^{ij}(s) = \varphi(ih, jh+s)$, $\frac{\partial^k\varphi}{\partial y^k}(ih, jh+s) = \frac{d^k\theta_2^{ij}}{ds^k}(s)$. In particular,

$$\Delta\varphi(z_{ij}) = (\theta_1^{ij})''(0) + (\theta_2^{ij})''(0).$$

Now the same computation of Taylor series expansions as in dimension 1, shows that

$$h^2(\theta_1^{ij})''(0) = \theta_1^{ij}(-h) - 2\theta_1^{ij}(0) + \theta_1^{ij}(h) + \frac{h^4}{12}(\theta_1^{ij})^{(4)}(t_{ij})$$

for a certain value of $t_{ij} \in\,]-h,h[$ and that

$$h^2(\theta_2^{ij})''(0) = \theta_2^{ij}(-h) - 2\theta_2^{ij}(0) + \theta_2^{ij}(h) + \frac{h^4}{12}(\theta^{ij})^{(4)}(s_{ij})$$

for a certain value of $s_{ij} \in\,]-h,h[$. By the definition of the functions θ_1^{ij} and θ_2^{ij}, it is easy to see that $\theta_1^{ij}(0) = \varphi(z_{ij})$, $\theta_1^{ij}(-h) = \varphi(z_{i-1,j})$, $\theta_1^{ij}(h) = \varphi(z_{i+1,j})$, $\theta_2^{ij}(0) = \varphi(z_{ij})$, $\theta_2^{ij}(-h) = \varphi(z_{i,j-1})$ and $\theta_1^{ij}(h) = \varphi(z_{i,j+1})$. By substituting and summing we thus find:

$$\Delta\varphi(z_{ij}) = \Delta_h\varphi(z_{ij}) + \frac{h^2}{12}\Big(\frac{\partial^4\varphi}{\partial x^4}(ih+t_{ij}, jh) + \frac{\partial^4\varphi}{\partial y^4}(ih, jh+s_{ij})\Big),$$

and the result immediately follows by taking the absolute value. □

Thus we see a consistency error of size $O(h^2)$. In the same way as for dimension 1 we now introduce the finite difference method in the following way. Let $\Omega_h = \{z_{ij}, 1 \le i \le N, 1 \le j \le N\}$ denote the set of interior grid points and $\bar{\Omega}_h = \{z_{ij}, 0 \le i \le N+1, 0 \le j \le N+1\}$ be the set of all grid points, including those on the boundary. Finally, we denote by $\partial\Omega_h := \bar{\Omega}_h \setminus \Omega_h$ the set of nodes on the boundary.

We will thus seek to compute values $u_{i,j}$ which converge towards the exact values $u(x_{i,j})$ by solving the following linear equations

$$\begin{cases} \frac{1}{h^2}(4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1})) = f(z_{ij}), & z_{i,j} \in \Omega_h, \\ u_{ij} = g(z_{ij}), & z_{i,j} \in \partial\Omega_h. \end{cases} \tag{2.10}$$

As for the method in dimension 1, we denote by $\overline{U}_h$ the vector of solution values on the grid and by U_h the vector of the approximation $u_{i,j}$, for $i, j = 1, \ldots, N$ and we study the error of the method by comparing those two quantities.

We can also identify with U_h a function u_h *defined on the grid* so that the previous equations can equivalently be written in the form

$$\begin{cases} -\Delta_h u_h(z_{ij}) = f(z_{ij}) \text{ on } \Omega_h, \\ u_h(z_{ij}) = g(z_{ij}) \text{ on } \bar{\Omega}_h \setminus \Omega_h. \end{cases} \tag{2.11}$$

There is of course no difficulty in applying the discrete Laplacian, defined initially for a continuous function on $\bar{\Omega}$, to a *discrete* function defined only on $\bar{\Omega}_h$. We can easily check by looking at the indices that all of the values which we need in order to compute the discrete Laplacian are available to us in the present case. In fact, we see that only the values of U_h at interior points are unknown, and that the values at boundary points are given by the boundary conditions (similar to dimension 1). Thus this is a linear problem with N^2 unknowns (the values of u_h on Ω_h) and N^2 linear equations (the values of $-\Delta_h u_h$ on Ω_h).

However, unlike in dimension 1, these unknowns and equations are not naturally arranged as the components of a vector – in fact, they are naturally arranged as the components of a matrix with two indices. The matrix form of the finite difference problem does not immediately appear similar to what we know from the one dimensional case treated above.

In order to arrive at this matrix form, we need to abstractly rearrange the entries of U_h, i.e. the grid points, into a single column. In other words, we must *number* the nodes. In dimension 1, this problem does not arise because we had a natural numbering imposed by the indices of the grid points. If we had desired of course, we could have chosen another numbering, i.e. effectively a permutation of the indices, which corresponds to a change of basis that permutes the basis vectors. The resulting matrix would then be obtained by performing a change of basis using a permutation matrix on the initial matrix. This however, would mean that it would have lost its favourable properties namely that it was tridiagonal and symmetric (exercise: what can you say about the monotonicity of the matrix?), which are incredibly useful when it comes to finding effective methods for the solution to these linear systems. Here, however, we have no choice but to design a suitable numbering because there is no obvious natural numbering. We will pull a suitable ordering *out of the hat* which will provide us with a symmetric matrix whose non-zero entries are relatively concentrated around the diagonal. These are two properties which

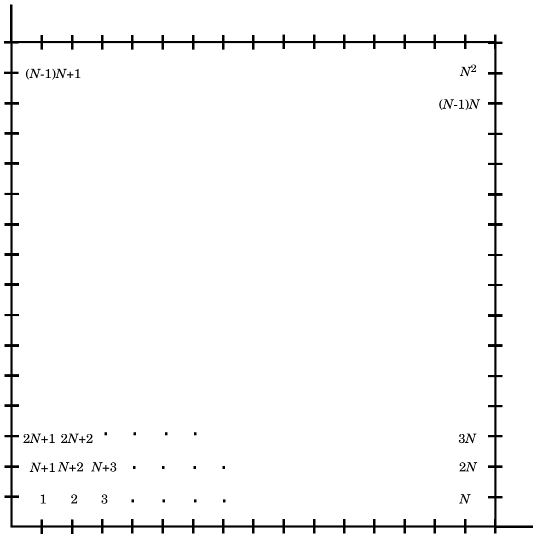


Figure 2.2: Numbering of the nodes

are very useful, in later step, for the application of efficient algorithms for the solution to the linear system.

We choose thus to order the nodes from left to right and from bottom to top, as shown in figure 2.2. We can easily convince ourselves that the number of the point z_{ij} is given by $(j-1)N+i$. So we define the vector U_h by its components: We can easily check that the number of the point z_{ij} is given by $(j-1)N+i$. Thus we can define the vector U_h through its components:

$$u_k = (U_h)_k = u_h(z_{ij}) = u_{i,j} \quad \text{for } k = (j-1)N+i.$$

Be careful with this abuse of notation: here we identified the function u_h defined on the grid with the vector U_h, even though their identification passes through a largely arbitrary numbering of the nodes.

The indices k thus defined vary between 1 and N^2, so $U_h \in \mathbb{R}^{N^2}$. The bijection between $\{1,\ldots,N\}\times\{1,\ldots,N\}$ and $\{1,\ldots,N^2\}$ that we have just introduced has the inverse map $k \mapsto \big(k - \big[\frac{k-1}{N}\big]N, \big[\frac{k-1}{N}\big]+1\big)$ where $[t]$ denotes the integer part of t (exercise: verify that this is indeed the inverse).

We can now write out the matrix form of the method associated with the chosen numbering. In order to do so, we introduce the $N\times N$ matrix

$$T_4 = \begin{pmatrix} 4 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 4 & -1 & 0 & & \vdots \\ 0 & -1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -1 & 4 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 4 \end{pmatrix}$$

and we denote the $N\times N$ identity matrix by I.

We will now define the matrix A_h of this linear system. The matrix has size $N^2 \times N^2$ and can be expressed in *block matrix form* with blocks of size $N \times N$, using the blocks T_4 and I, as well as the block 0 which corresponds to the $N \times N$ zero matrix. Similarly, we can write the right hand side F_h in block matrix form, we show below the first two blocks (out of a total of N blocks) and the beginning of the third one.

Proposition 2.4.1 *The vector $U_h \in \mathbb{R}^{N^2}$ is given as the solution to the following linear system:*

$$A_h U_h = F_h,$$

with

$$A_h = \frac{1}{h^2}\begin{pmatrix} T_4 & -I & 0 & \dots & \dots & 0 \\ -I & T_4 & -I & 0 & & \vdots \\ 0 & -I & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -I & T_4 & -I \\ 0 & \dots & \dots & 0 & -I & T_4 \end{pmatrix}$$

and

$$F_h = \begin{pmatrix} f_1 + \frac{1}{h^2}(g(z_{1,0}) + g(z_{0,1})) \\ f_2 + \frac{1}{h^2} g(z_{2,0}) \\ f_3 + \frac{1}{h^2} g(z_{3,0}) \\ \vdots \\ f_N + \frac{1}{h^2}(g(z_{N,0}) + g(z_{N+1,1})) \\ f_{N+1} + \frac{1}{h^2} g(z_{0,2}) \\ f_{N+2} \\ \vdots \\ f_{2N} + \frac{1}{h^2} g(z_{N+1,2}) \\ f_{2N+1} + \frac{1}{h^2} g(z_{0,3}) \\ f_{2N+2} \\ \vdots \end{pmatrix}$$

where $f_k = f(z_{ij})$ for $k = (j-1)N + i$.

Proof. We must distinguish several cases, depending on whether the row index k under consideration corresponds to a point whose four neighbors are inside Ω_h, a point where only three neighbors are inside Ω_h, or a point where only two neighbors are inside Ω_h.

♣ Four neighbors inside Ω_h. These are the points z_{ij} with $2 \le i \le N-2$ and $2 \le j \le N-2$. They correspond to the indices $k \in \cup_{l=1}^{N-2}\{lN+2, lN+3, \dots, (l+1)N-1\}$. (For example, for $l = 1$, this yields $k = N+2, N+3, \dots, 2N-1$. Then we jump directly to $2N+2$ and so on and so forth.)

At every such point z_{ij}, of number $k=(j-1)N+i$, the neighboring points are all numbered with the following indices respectively

$$\begin{aligned}
\#z_{i-1,j} &= (j-1)N+(i-1)=k-1,\\
\#z_{i+1,j} &= (j-1)N+(i+1)=k+1,\\
\#z_{i,j-1} &= ((j-1)-1)N+i=k-N,\\
\#z_{i,j+1} &= ((j+1)-1)N+i=k+N.
\end{aligned}$$

We thus obtain the following row for the discrete problem

$$\frac{1}{h^2}(-u_{k-N}-u_{k-1}+4u_k-u_{k+1}-u_{k+N})=f(z_{ij})=f_k.$$

♠ Three neighbors inside Ω_h. This case arises in four different ways: $j=1, i=2,3,\dots,N-1$; $j=N, i=2,3,\dots,N-1$; $i=1, j=2,3,\dots,N-1$ or $i=N, j=2,3,\dots,N-1$. These four ways correspond, respectively, to $k=2,3,\dots,N-1$; $k=(N-1)N+2,(N-1)N+3,\dots,N^2-1$; $k=N+1,2N+1,\dots,(N-2)N+1$ and $k=2N,3N,\dots,(N-1)N$. Consider the first possibility (the others are left as an exercise to the reader). The point $z_{i,j-1}=z_{i,0}$ is located on the boundary. We must thus move the corresponding value of U_h, which is given by the boundary condition, to the right hand side, which yields

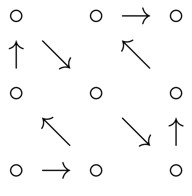
$$\frac{1}{h^2}(-u_{k-1}+4u_k-u_{k+1}-u_{k+N})=f_k+\frac{1}{h^2}g(z_{k,0}).$$

♢ Two neighbors inside Ω_h. This arises in four cases: $z_{1,1}$, $z_{N,1}$, $z_{1,N}$ and $z_{N,N}$, so for $k=1$, $k=N$, $k=(N-1)N+1$ and $k=N^2$. Consider the case $k=1$. It follows that

$$\frac{1}{h^2}(4u_1-u_2-u_{1+N})=f_1+\frac{1}{h^2}(g(z_{1,0})+g(z_{0,1})).$$

Finally, we need to join together all of the above observations. This is a bit tedious, but using large piece of paper to check the calculations, we ultimately arrive at the desired matrix form. □

Remark 2.4.2 *To convince ourselves of the importance of the numbering, we can for our entertainment try to write the matrix in a low-dimensional case, for example when $N=3$ (i.e. a matrix of size 9×9), firstly with the proposed numbering and then secondly with another numbering, which a priori appears just as natural, namely the one we use to prove that $\mathbb{N}^2$ is bijective to $\mathbb{N}$:*



Following exactly the same procedures as for the A_h matrix in dimension 1, we obtain the following result for the present two-dimensional case.

Proposition 2.4.2 *The matrix A_h is symmetric, positive definite and monotone.*

An immediate consequence is that the finite difference method is still well-posed. It remains to check the norm of its inverse. For this we can use a discrete version of the maximum principle established in theorem 2.3.1.

Theorem 2.4.2 *(Discrete maximum principle). Let v_h be a function defined on the grid $\bar{\Omega}_h$ such that $\Delta_h v_h \leq 0$ in Ω_h. Then v_h attains its maximum on $\partial\Omega_h$. Similarly, if v_h is such that $\Delta_h v_h \geq 0$ in Ω_h, then v_h attains its maximum on $\partial\Omega_h$.*

Proof. The proof is quite similar to that of theorem 2.3.1 so we only indicate the main steps. We begin by showing that if a function w_h defined on the grid admits a minimum at an interior point $z_{i,j}$, then we have

$$\Delta_h w_h(z_{i,j}) \geq 0.$$

Then we reason by contradiction, by assuming that v_h takes a minimum at an interior grid point $z_{i_0,j_0} \in \Omega_h$, strictly smaller than the minimum value on the boundary $\partial\Omega_h$. We introduce the following auxiliary discrete function

$$v_{h,\varepsilon}(z_{i,j}) = v_h(z_{i,j}) - \varepsilon(i^2 + j^2),$$

with $\varepsilon > 0$ sufficiently small such that $v_{h,\varepsilon}$ attains a minimum at an interior grid point $z_{i_1,j_1} \in \Omega_h$ which means we arrived at a contradiction because $\Delta_h v_{h,\varepsilon} < 0$. □

Theorem 2.4.3 *We have $\|A_h^{-1}\|_\infty \leq \frac{1}{2}$.*

Proof. Let $F_h \in \mathbb{R}^{N^2}$ and $V_h \in \mathbb{R}^{N^2}$ non-zero such that $A_h V_h = F_h$. Identifying V_h with a function v_h defined on $\bar{\Omega}_h$ and zero at the boundary and F_h with a function f_h defined in Ω_h, this is equivalent to $-\Delta_h v_h = f_h$ in Ω_h.

Let us introduce the discrete function $w_h(z_{ij}) = \frac{h^2}{4}(i^2 + j^2)$. We can check through a simple calculation that $-\Delta_h w_h = -1$ in Ω_h. Let us thus take

$$w_h^+ = \|-\Delta_h v_h\|_\infty w_h - v_h,$$

where we kept the sign in order to remind us that $-\Delta_h v_h = f_h$. Taking the discrete Laplacian of w_h^+, it follows, for every $z_{i,j} \in \Omega_h$, that

$$-\Delta_h w_h^+(z_{i,j}) = -\|-\Delta_h v_h\|_\infty \Delta_h w_h(z_{i,j}) + \Delta_h v_h(z_{i,j}) = \Delta_h v_h(z_{i,j}) - \|-\Delta_h v_h\|_\infty \leq 0.$$

The discrete maximum principle applies and shows that w_h^+ attains its maximum on $\partial\Omega_h$. However, we have $v_h = 0$ on $\partial\Omega_h$, thus for every $z_{i,j} \in \Omega_h$

$$w_h^+(z_{ij}) \leq \|-\Delta_h v_h\|_\infty \max_{\partial\Omega_h} w_h = \frac{1}{2}\|-\Delta_h v_h\|_\infty.$$

Consequently,

$$v_h(z_{i,j}) = \|-\Delta_h v_h\|_\infty w_h(z_{i,j}) - w_h^+(z_{i,j}) \geq -w_h^+(z_{i,j}) \geq -\frac{1}{2}\|-\Delta_h v_h\|_\infty.$$

Using in a similar manner the function $w_h^- = \|-\Delta_h v_h\|_\infty w_h + v_h$, we can show that

$$v_h(z_{i,j}) \le \frac{1}{2}\|-\Delta_h v_h\|_\infty.$$

Therefore,

$$\|v_h\|_\infty = \max_{z_{i,j}\in\Omega_h} |v_h(z_{i,j})| \le \frac{1}{2}\|-\Delta_h v_h\|_\infty.$$

In matrix terms, this can be written again as

$$\|A_h^{-1}F_h\|_\infty \le \frac{1}{2}\|F_h\|_\infty,$$

whence, by dividing this inequality by $\|F_h\|_\infty$, the result follows. □

In the above, we have studied the consistency using theorem 2.4.1 and the stability by estimating $\|A_h^{-1}\|_\infty$. We can therefore now announce the convergence theorem.

Theorem 2.4.4 *Assume that $u \in \mathcal{C}^4(\bar\Omega)$, then*

$$\max_{i,j} |u(z_{ij}) - u_{ij}| \le \frac{2h^2}{24}\Big(\|\frac{\partial^4 u}{\partial x^4}\|_{L^\infty} + \|\frac{\partial^4 u}{\partial y^4}\|_{L^\infty}\Big).$$

Proof. We denote by $\overline{U}_h$ (resp. F_h) the vector with entries $u(z_{ij}), 1 \le i,j \le N$ (resp. $f(z_{ij})$), and $K_h = A_h\overline{U}_h - F_h$ the consistency error. By associating them with the functions $\bar u_h$, f_h and κ_h defined on Ω_h, this can also be written as

$$\kappa_h = -\Delta_h \bar u_h - f_h.$$

We extend $\bar u_h$ and u_h to the boundary $\partial\Omega_h$ by the discrete boundary conditions $g(z_{i,j})$. We already know that

$$\|\kappa_h\|_\infty \le \frac{h^2}{12}\Big(\|\frac{\partial^4 u}{\partial x^4}\|_{L^\infty} + \|\frac{\partial^4 u}{\partial y^4}\|_{L^\infty}\Big).$$

However

$$\begin{aligned} -\Delta_h u_h &= f_h, \\ -\Delta_h \bar u_h &= f_h + \kappa_h, \end{aligned}$$

when

$$\begin{cases} -\Delta_h(\bar u_h - u_h) = \kappa_h \text{ in } \Omega_h, \\ \bar u_h - u_h = 0 \text{ on } \partial\Omega_h. \end{cases}$$

Because $\bar u_h - u_h$ vanishes on the boundary, we may thus write this in the form

$$A_h(\overline{U}_h - U_h) = K_h \Longrightarrow \|\overline{U}_h - U_h\|_\infty \le \|A_h^{-1}\|_\infty \|K_h\|_\infty.$$

This completes the proof. □

Remark 2.4.3 *i) In principle, we can extend these ideas to dimensions 3, 4 or higher without difficulty. However, the matrices involved become more complicated and, in particular, their size grows like N^3, N^4 etc. This exponential increase in size with respect to the dimension quickly becomes prohibitively expensive from the practical point of view.*

ii) Beware of the regularity assumption $u \in \mathcal{C}^4(\bar\Omega)$. As we have already noted in passing, this is not self-evident in dimension 2.