Discrete maximum principle and the Ultraviolet Catastrophe of finite difference schemes on staggered Cartesian grids for heterogeneous and anisotropic diffusion equations

Roberta Dani, Chiara Simeoni

Research Report : discrete maximum principle and the *ultraviolet catastrophe* of finite difference schemes on staggered Cartesian grids for heterogeneous and anisotropic diffusion equations

Roberta Dani

Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica
Università degli Studi dell'Aquila, Italy
http://www.disim.univaq.it/

Chiara Simeoni

Laboratoire de Mathématiques J.A. Dieudonné
Université Nice Sophia Antipolis, France
http://math.unice.fr/

February 21st, 2014

Introduction

Anisotropic and heterogeneous diffusion equations have different fields of application such as image processing [38] and computer vision [1],[32], modeling of tumors growth [35],[14], flows in heterogeneous media [31],[9], plasma physics [27], option pricing in finance [18], biological processes [4],[13], medicine [2] and ecology [24]. Beside typical regularizing effects, the most important feature of such models is that inconstant diffusion coefficients could produce strikingly nontrivial patterns [15]. Therefore, the numerical solution often requires very long computational time, for the large amount of data to be traded in order to accurately capture the details of physical phenomena, and the numerical methods used for solving these models should be chosen with great care. Under specific hypotheses, those equations exhibit the notable properties of *maximum principle* and *comparison principle* [29],[11]. Indeed, the maximum principle, closely related to the non-negativity property, is one of the basic characteristics of classical solutions of second order PDEs of parabolic type. The preservation of this property for solutions to corresponding discretized problems is very important and is a natural requirement in reliable and meaningful numerical modeling of various real-life phenomena. In all the cases, it is indispensable for a physical meaning that the solution methods do not produce negative values, otherwise the results of the equations which correspond to density or concentration of a certain substance no longer make sense. Consequently, suitable numerical solvers for simulating anisotropic and heterogeneous diffusion equations must exhibit characteristics analogous to the theoretical properties of the mathematical models (issuing from physical considerations) : mass conservation, positivity preserving, discrete maximum/comparison principle. All such properties are, in fact, *stability properties* for the numerical methods.

In some literature (see [30],[10] for instance), the parabolic equations are typically treated with the *finite element method*, which provides a powerful tool to determine the validity of the maximum principle and the non-negativity property also in presence of non-structured grids, under certain conditions [23]. But, sometimes, it might be useful to simulate the parabolic equations with other simple numerical schemes, such as the *finite differences* or rather the *finite volume methods*.
These are far from being faster than finite element schems, but they have

the advantage of being easily *parallelizable*. Indeed, the ultimate purpose of future works is to put the schemes presented in this report into a parallel computing platform, for example *CUDA GPU* [17], to have several processes that perform at the same time to greatly decrease the computational costs. Even if the resolution of linear systems is not generally efficient in this type of architecture, for the type of computational domains we take into account is particularly suitable to employ them. In particular, we will refer to rectangular domains with *Cartesian meshes*, approximating the solution at mesh points, to have a numerical grid which is as consistent as possible with the GPU virtual mesh. In this sense, a *tread* in the CUDA machine corresponds to an iteration in the numerical scheme.

Therefore, we restrict to Finite Difference schemes on Cartesian grids for the ease of implementation in parallel computing systems, and especially CUDA GPUs, but we will always consider the Finite Volume formulation for recovering the numerical fluxes in more advanced applications (for the analogy with the integral formulation of conservation laws), thus leading to the relevant class of *finite differences/volumes on staggered grids*. The aim of this report is to give a uniform introduction to the finite difference and finite volume methods for approaching the anisotropic and heterogeneous diffusion equations, for which the validity of the discrete maximum principle, with the related non-negativity property, is satisfied.

The report is organized as follows. In Chapter 1, we give a mathematical statement of the problem : we introduce the anisotropic and heterogeneous diffusion equations, by situating them in the wider context of parabolic problems, we briefly discuss the physical derivation [6], and we present the main theoretical results on the maximum/minimum principle, also quoting few other related models. In Chapter 2, we describe Finite Difference and Finite Volume schemes, by trying to underline their analogies and recast them inside a common framework (this becomes especially relevant for future extensions to nonuniform meshes, for the question of consistency and super-convergence as only some formulations are really appropriate).
We present general techniques to approximate the partial derivatives, which are used to combine several numerical methods. In particular, the existence of the Nonnegative Scheme for two-dimensional problems is carefully proven, according to [38], where applications to image processing are considered.
In Chapter 3 and Chapter 4, we provide details about the stability analysis of numerical scheme for one-dimensional and two-dimensional problems, respectively. The issue of discrete maximum principle, and then the question of the L^∞-stability, is treated through the *algebraic theory of positive matrices* following [28], to determine the range of numerical parameters under which that important property is satisfied. On the other hand, the analysis of L^2-stability is confined to linear problems to apply the *Fourier analysis*, and we will discuss the peculiar phenomenon of the *Ultraviolet Catastrophe*.
Finally, an extensive series of numerical tests is proposed in Chapter 5.

Contents

List of Figures

Chapter 1

Linear Diffusion Equations and Qualitative Properties

The diffusion equations, that are generally stated in the form

$$\partial_t u = \nabla \cdot (A \cdot \nabla u)\,, \tag{1.1}$$

have multiple historical origins, each building upon a unique physical interpretation. For the most common applications, u is interpreted as a concentration or density and A is the *diffusion tensor*. In general, we will consider $u(t; x, y)$, where $(x, y) \in \mathbb{R}^2$, because in many practical situations the density spreads in a two-dimensional domain, for example the diffusion of diseases in a given geographic region or applications for population dynamics.

Doubtless, the maximum principle is one of the main properties of second order PDEs of parabolic type, as mentioned in most of the literature about this argument [29],[12]). It is a powerful feature because it is related to other properties, in particular the *non-negativity property*. Both of them are a natural requirement in reliable and meaningful real-life phenomena, so the importance of studying these properties in details.

1.1 Physical background

The diffusion is a physical process that equilibrates concentration differences without creating or destroying mass, thus resulting in a *conservative process*.

From a mathematical point of view, let $\Omega \subset \mathbb{R}^2$ be a specific region and u the density of a certain substance that varies over time $t > 0$. We want to study how the density moves (spreads) around this domain over the time. The law describing the diffusion process is the *Fick's law*, which postulates that the flux j goes from regions of high concentration to regions of low concentration, with a magnitude that is proportional to the concentration gradient ∇u, namely

$$j = -A \cdot \nabla u\,,$$

where the diffusion tensor A is a positive definite symmetric matrix. This last has dimensions $length^2/time$ and represents how the diffusion takes place. So, it is physically unreal for it to be negative. The sign minus is because the diffusion operates from higher to lower densities. The observation that diffusion is only a transportation phenomenon without destroying or creating mass is expressed by the *continuity equation*, that reads

$$\partial_t u = -\nabla \cdot j \,.$$

Replacing the flux in the last equation, we obtain the *diffusion equation* (1.1). In particular, the solution function u will be of the type

$$u \,=\, u(t;x,y) \,:\, \mathbb{R}^+ \times \Omega \longrightarrow \mathbb{R}^+.$$

We note that this *conservation law* holds for the special linear flux j when the entries of the diffusion matrix do not depend explicitly on the solution.

We focus our attention on the heterogeneous and anisotropic case, which includes some properties for the diffusion tensor A, that is :

- *heterogeneous* , i.e. the diffusion tensor A is not constant, indeed it is a matrix whose entries are functions of the space variables. In general, they can also be functions of the time and density itself, but we will limit to consider the time-independent linear case. So, A will be of the type
$$A = \begin{bmatrix} a(x,y) & c_1(x,y) \\ c_2(x,y) & b(x,y) \end{bmatrix}. \tag{1.2}$$

- *anisotropic* , i.e. the flux j and the gradient ∇u are not parallel, so that the diffusion will happen in all the directions of domain Ω in a different manner. The diffusion mechanism depends on the particular direction in space : in terms of the matrix A, we can have several cases, going from the simpler case in which $c_1 = c_2 = 0$, and so we have diffusion only along the directions of the Cartesian axes, to more complicate cases in which all the entries are different from zero.
In general, we focus on symmetric matrices, that is $c_1 = c_2$.

1.2 Parabolic Operators

The diffusion equation (1.1) lies in the class of the parabolic equations, according to the following definition.

Definition 1. *The partial differential operator*

$$L(u) = \sum_{i,j=1}^{n} a_{ij}(t,z)\frac{\partial^2 u}{\partial z_i \partial z_j} + \sum_{i=1}^{n} b_i(t,z)\frac{\partial u}{\partial z_i} - \frac{\partial u}{\partial t} \tag{1.3}$$

is said to be ***parabolic*** *at* $(t, z) = (t, z_1, z_2, \ldots, z_n) \in \mathbb{R}^+ \times \mathbb{R}^n$ *if for fixed* $t > 0$ *the second order operator consisting of the first sum is elliptic at* (t, z)*, that is if there exists a constant* $\mu > 0$ *such that*

$$\sum_{i,j=1}^{n} a_{ij}(t, z)\xi_i \xi j \geq \mu \sum_{i=1}^{n} \xi_i^2 \tag{1.4}$$

for all n*-tuples of real numbers* $(\xi_1, \xi_2, \ldots, \xi_n)$*. The operator* L *is* ***uniformly parabolic*** *in a region* $\Omega_T \subset \mathbb{R}^+ \times \mathbb{R}^n$ *if* (1.4) *holds with the same constant* $\mu > 0$ *for all* (t, z) *in* Ω_T*.*

In this report, we focus on parabolic equations in two space dimensions, for $n = 2$, that are computed as

$$\begin{aligned} L(u) = a_{11}(t; x, y)\frac{\partial^2 u}{\partial x^2} + a_{12}(t; x, y)\frac{\partial^2 u}{\partial x \partial y} + a_{21}(t; x, y)\frac{\partial^2 u}{\partial x \partial y} + \\ + a_{22}(t; x, y)\frac{\partial^2 u}{\partial y^2} + b_1(t; x, y)\frac{\partial u}{\partial x} + b_2(t; x, y)\frac{\partial u}{\partial y} - \frac{\partial u}{\partial t}\,. \end{aligned}$$

For comparing this parabolic operator to equation (1.1), we expand on the diffusion term using (1.2) and we rewrite it with a lighter notation,

$$\begin{aligned} \nabla \cdot (A \cdot \nabla u) &= \nabla \cdot \left[\begin{bmatrix} a & c_1 \\ c_2 & b \end{bmatrix} \cdot \begin{bmatrix} u_x \\ u_y \end{bmatrix} \right] = \nabla \cdot \begin{bmatrix} a\, u_x + c_1\, u_y \\ c_2\, u_x + b\, u_y \end{bmatrix} \\ &= \big(a\, u_x + c_1\, u_y\big)_x + \big(c_2\, u_x + b\, u_y\big)_y = \big(a\, u_x\big)_x + \big(c_1\, u_y\big)_x + \big(c_2\, u_x\big)_y + \big(b\, u_y\big)_y \\ &= a\, u_{xx} + (c_1 + c_2)\, u_{xy} + b\, u_{yy} + \big(a_x + (c_2)_y\big) u_x + \big((c_1)_x + b_y\big) u_y\,, \end{aligned} \tag{1.5}$$

also imposing the *Schwartz's theorem* (for regular solutions), so that the expression (1.3) is obtained by identifying the coefficients

$$\begin{aligned} a_{11} = a\,, \quad a_{12} = c_1\,, \quad a_{21} = c_2\,, \quad a_{22} = b\,, \\ b_1 = a_x + (c_2)_y\,, \quad b_2 = (c_1)_x + b_y\,. \end{aligned} \tag{1.6}$$

Remark 1. *The presence of first order terms in* (1.5) *is essentially due to the heterogeneity of the diffusion tensor, otherwise only second order terms would be included. In terms of numerical approximations, this fact implies the necessity of some* ***upwinding*** *in the construction of the numerical fluxes (refer to Section 4.2.2).*

We want to see under which conditions the inequality (1.4) holds, and we start by giving the definition of *positive definite* matrices.

1.2.1 Properties of the diffusion tensor

For modeling reasons coming from the physical interpretation, the real-valued diffusion matrix $A : \Omega \longrightarrow \mathbb{R}^{2\times 2}$ is typically chosen to be symmetric and positive definite, according to the following definition.

Definition 2. *Let A be a real-valued $n \times n$ symmetric matrix, then A is said to be* ***positive definite*** *if and only if, for all $\vec{\xi} \in \mathbb{R}^n/\{\vec{0}\}$, it holds $\vec{\xi}^T A \vec{\xi} > 0$, and* ***positive semidefinite*** *if and only if $\vec{\xi}^T A \vec{\xi} \geq 0$.*
This intuitively extends to the definition of ***negative definite****.*

We list some characteristic properties of the positive definite matrices [33],[16].

Proposition 1. *If A is a real-valued symmetric positive definite matrix, then*

- *its eigenvalues are all real and positive, i.e. $spec(A) \subset \mathbb{R}^+$;*
- *its leading principal minors are all positive (Sylvester's criterion);*
- *the associated symmetric bilinear form is an inner product in $\mathbb{R}^n$;*
- *its inverse, A^{-1}, exists and is also positive definite;*
- *the determinant of A is bounded by the product of its diagonal elements;*
- *there exists μ positive real number, such that $A > \mu I$, where I is the identity matrix.*

For the matrix (1.2) with $c_1 = c_2 = c$, the Definition 2 implies $a > 0, b > 0$ and $c < (a+b)/2$, and the eigenvalues are given by

$$\lambda_{1,2} = \frac{(a+b) \pm \sqrt{(a+b)^2 - 4(ab-c^2)}}{2},$$

so that, to satisfy the properties listed in Proposition 1, we have

$$a + b > 0 \quad \text{and} \quad 0 < \sqrt{(a+b)^2 - 4(ab-c^2)} < a + b,$$

and we finally obtain optimal conditions for the entries of the matrix,

$$a > 0, \quad b > 0, \quad c^2 < ab. \tag{1.7}$$

We remark that a more general (weaker) notion of positive-definiteness is given by the first statement of Proposition 1, which does not require the condition for a real-valued matrix of being symmetric. In that case, however, an equivalent requirement is that the corresponding symmetric part $\left(A + A^T\right)/2$ is positive definite in the narrower sense of Definition 2.

1.2.2 Relationship between *parabolic operators* and *positive definite matrices*

The definitions of parabolic equations and positive definite diffusion tensors are strictly related, as stated through the following result.

Lemma 1. *The equation* (1.1) *with A positive definite is parabolic.*

Proof. We consider the extended form (1.5) in $\Omega \subset \mathbb{R}^2$ and, recalling (1.6), we can rewrite the inequality (1.4) as

$$a\,\xi_1^2 + (c_1 + c_2)\,\xi_1\xi_2 + b\,\xi_2^2 > \mu(\xi_1^2 + \xi_2^2)\,,$$

for some $\mu > 0$, or rather in matrix-form as

$$\begin{bmatrix}\xi_1 & \xi_2\end{bmatrix}\begin{bmatrix}a & c_1\\ c_2 & b\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix} - \mu(\xi_1^2 + \xi_2^2) > 0\,,$$

that is

$$\vec{\xi}^{\,T}(A - \mu I)\,\vec{\xi} > 0\,, \qquad \vec{\xi} = (\xi_1, \xi_2) \neq \vec{0}\,,$$

where I is the identity matrix.
This last statement is certainly verified if A is a symmetric positive definite matrix, according to Proposition 1, thus L for (1.1) is a parabolic operator. Conversely, the same inequality directly implies that the matrix $A-\mu I$ is definite positive, therefore $a > \mu > 0$, $b > \mu > 0$, and $c^2 < (a-\mu)(b-\mu) < ab$, which guarantees that also A is positive definite (as sum of positive definite matrices). □

As a consequence of Lemma 1, positive definite matrices play an important role in *optimization problems* : as a matter of fact, any real quadratic form $Q : \mathbb{R}^n \longrightarrow \mathbb{R}$ can be written as $\vec{\xi}^{\,T}A\,\vec{\xi} + \vec{\xi}^{\,T}B + C$, where A is a symmetric $n \times n$ matrix, B is a $n-$vector and C is a scalar function; moreover, this functional is strictly convex (and hence it has a unique finite global minimum) if and only if A is positive definite. The analogy with (1.3) suggests that the parabolicity of the operator L is equivalent to the convexity of its associated *quadratic form*, that holds if and only if A is symmetric and positive definite.

Under broad regularity assumptions on the diffusion coefficients, linear parabolic PDEs in *conservation/divergence form* as given in (1.1) have solutions for all $t > 0$, which should be understood as *weak solutions* : in particular, we refer to [3][Chapter 10] for an existence theory in the case of L^∞-bounded coefficients, whereas stronger regularity assumptions are needed to treat the extended form (1.21) below. We remark that the symmetry of the diffusion tensor is not explicitly required to show most theoretical properties, but this becomes mandatory to establish a *variational formulation* of the solution as the global minimum of some convex entropy/energy

functional (as already seen above). Besides, the symmetry is justified for the L^2-contraction property to hold, or rather to provide a supplementary (weighted) inner product inside the specific functional framework [19],[5].

1.3 Maximum Principle for linear operators

We consider the parabolic operator (1.3) with all coefficients bounded inside Ω a bounded domain in $\mathbb{R}^n$ and $\partial\Omega$ its boundary. We introduce the sets

$$\Omega_T = \Omega \times (0,T)\,, \quad \Gamma_T = \big(\partial\Omega \times [0,T]\big) \cup \big(\Omega \times \{0\}\big), \tag{1.8}$$

the last one being the so-called *parabolic boundary*, for any arbitrary positive time T. We will prove that, under specific conditions, the maximum of the solution to the equation $L(u) = 0$ for (1.3) is attained at a point of the parabolic boundary Γ_T, and never in the interior of Ω_T.

The result holds immediately for the simplest *isotropic heat equation*, thanks to an explicit expression for the self-similar solutions, which exhibit the typical Gaussian shape (see [20][Chapter 8]), but clearly this technique cannot be extended to more general cases. On the other hand, the results reproduced in this report extend also to nonlinear operators (refer to [29] [Chapter 3, Section 7]), although we do not pursue explicitly that issue.

1.3.1 The one-dimensional case

We focus on the one-dimensional parabolic equation

$$L(u) = a(t,x)\frac{\partial^2 u}{\partial x^2} + b(t,x)\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} = 0\,, \qquad \forall\,(t,x) \in \Omega_T\,, \tag{1.9}$$

where Ω_T is a rectangular region of $\mathbb{R}^+ \times \mathbb{R}$, for example

$$\Omega_T = \{(t,x) \mid t \in (0,T)\,,\ x \in \Omega = (0,\Lambda)\}\,, \tag{1.10}$$

which is also appropriate for the discretization through finite difference schemes. We will see later on that we can extend the same results to the general case of parabolic equations in $\Omega \subset \mathbb{R}^n$.

We remark that the definition of parabolic operators leads to state that $a(t,x) > 0$ for the one-dimensional case.

The strategy for proving the results is the following : we will prove a sort of maximum principle for the strict inequality $L(u) > 0$ and, then, we will extend the result for the inequality $L(u) \geq 0$, which allows us to state the validity of the maximum principle for the equation $L(u) = 0$.

So, we proceed with the following results. For more details in the proof of these statements one can refer to [29].

Proposition 2. *If the parabolic operator $L(u)$ is strictly positive in Ω_T, namely $L(u) > 0$, then the maximum cannot occur neither in the interior of the domain Ω_T nor along the open segment forming the upper boundary of Ω_T, that is in a point (T, x), where $0 < x < \Lambda$.*

Proof. Suppose that there is an interior maximum point, so the following relations hold for regular solutions,

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} = 0\,, \qquad \frac{\partial^2 u}{\partial x^2} \leq 0\,.$$

In this case, $L(u) = a(x,t)\frac{\partial^2 u}{\partial x^2} \leq 0$, which violates the hypothesis.
Moreover, the maximum cannot be at $t = T$ because, if so, we would have

$$\frac{\partial u}{\partial t} \geq 0\,, \qquad \frac{\partial u}{\partial x} = 0\,, \qquad \frac{\partial^2 u}{\partial x^2} \leq 0\,,$$

and the positivity of the operator L is still contradicted. □

The maximum principle for the operator (1.3) is now extended to solutions to the differential inequality $L(u) \geq 0$. In [29], there is also a generalization for domains Ω_T of the (t, x)-plane which are not necessarily rectangular. From now on, to avoid mistakes with the notation, we also denote by $\Omega_t \subset \mathbb{R}^+ \times \mathbb{R}$ the generic domain where the index t simply indicates that we consider the time variable and the space variable together (see Figure 1.1).

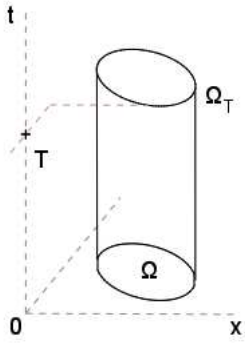
We proceed with three lemmas. Through these we will see that, if the maximum of the solution to the parabolic equation is attained at an interior point of the domain Ω_T, then the solution is constant and its value is just the maximum, otherwise this value is attained on the boundary of the domain.

Lemma 2. *Let u satisfy the differential inequality*

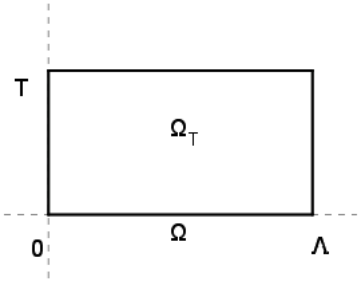
$$L(u) = a(t,x)\frac{\partial^2 u}{\partial x^2} + b(t,x)\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} \geq 0$$

in Ω_t, where a and b are bounded functions and L is uniformly parabolic. Let K be a disk such that it and its boundary ∂K are contained in Ω_t. Suppose the maximum of u in Ω_t is M, that $u < M$ in the interior of K, and that $u = M$ at the point P on the boundary of K. Then, the tangent to K at P is parallel to the x-axis. That is, P is either the point at the top or the point at the bottom of the disk K.

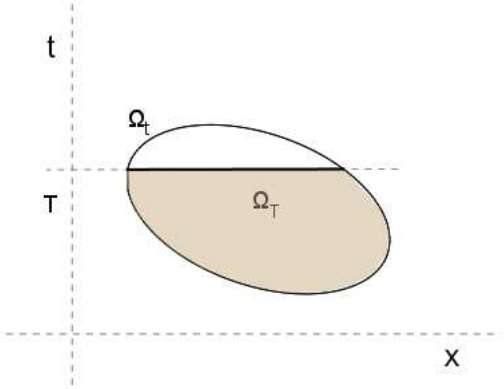
Proof. We consider $(\bar{t}, \bar{x})$ the center of the disk K, and R its radius. We suppose that the point P on ∂K is not at the top or bottom of the disk, and we prove that we reach a contradiction.
We also assume, without loss of generality, that P is the only boundary



(a) domain in $\mathbb{R}^+ \times \mathbb{R}^2$ as defined in (1.8)



(b) rectangular domain in $\mathbb{R}^+ \times \mathbb{R}$



(c) general domain in $\mathbb{R}^+ \times \mathbb{R}$

Figure 1.1: example of computational domains

point where $u = M$. We can always do this because, if there are other points on the boundary ∂K which attain the maximum, we can replace K by a slightly smaller disk K' whose boundary is interior to K except at the one point P where $\partial K'$ and ∂K are tangent.
Suppose P has coordinates (t_1, x_1) with $x_1 \neq \bar{x}$, and we construct a disk K_1 with center at P and radius R_1 so small that

$$R_1 < R = |x_1 - \bar{x}|\,,$$

and also such that K_1 lies completely in Ω_t (see Figure 1.2).
We note that ∂K_1 consists of two arcs,

$$\partial K_1 = C' \cup C'',$$

where C' is the intersection of ∂K_1 with the closed disk $\bar{K}$, so in this part there are also its endpoints, and C'' is the remaining part.

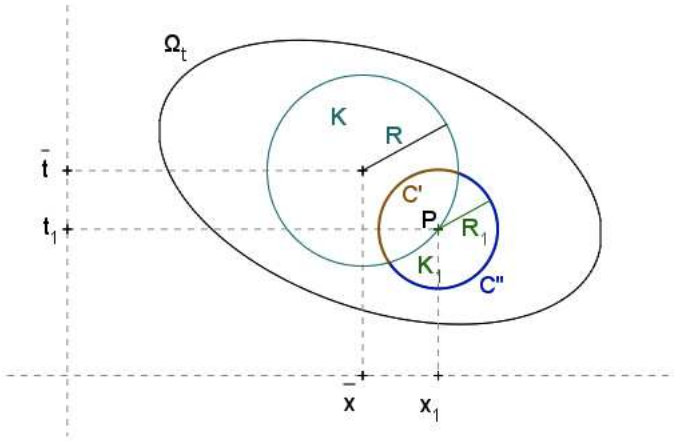


Figure 1.2: geometric construction for Lemma 2

Since $u < M$ on the closed arc C', we can find a constant η such that

$$u \leq M - \eta \quad \text{on } C'. \tag{1.11}$$

Moreover, since $u \leq M$ throughout Ω_t, it holds

$$u \leq M \quad \text{on } C''.$$

We define the following auxiliary function,

$$v(t,x) = e^{-\alpha\left[(t-\bar{t})^2+(x-\bar{x})^2\right]} - e^{-\alpha R^2},$$

and we choose $\alpha > 0$, so that the function v is positive in K, zero on ∂K and negative in the exterior of K.
Now, we apply the partial differential operator L to v and we obtain

$$L(v) = 2\alpha\, e^{-\alpha\left[(t-\bar{t})^2+(x-\bar{x})^2\right]}\left[2\alpha\, a\,(x-\bar{x})^2 - a - b\,(x-\bar{x}) + (t-\bar{t})\right].$$

We observe that

$$|x-\bar{x}| \geq |x_1-\bar{x}| - R_1 > 0 \quad \text{on } \bar{K_1}\,,$$

so we can choose α large enough such that

$$L(v) > 0 \quad \text{on } \bar{K_1}\,.$$

We also define a function w such that

$$w(t,x) = u(t,x) + \varepsilon v(t,x)\,, \tag{1.12}$$

with $\varepsilon > 0$, and we observe that

$$L(w) > 0 \quad \text{in } K_1 . \tag{1.13}$$

We can choose ε so small that

$$w = u + \varepsilon v < M \quad \text{on } C',$$
$$w = u + \varepsilon v < M \quad \text{on } C'',$$

because (1.11) holds and $v < 0$ on C'' and $u \leq M$. Thus $w < M$ on the entire boundary ∂K_1 but $w(t_1, x_1) = M$, because v vanishes on ∂K.
So, we have proven that the maximum of w is attained at the interior point of K_1. This fact contradicts the Proposition 2 because we have that (1.13) holds. □

Remark 2. *We note that the proof of Lemma 2 fails if P is located at the top or the bottom of K because, in this case, we have $x_1 = \bar{x}$ and we cannot choose $R_1 < |x_1 - \bar{x}|$.*

Lemma 3. *Suppose that u satisfies the inequality $L(u) \geq 0$ in Ω_t, with L as in Lemma 2. Suppose that $u < M$ at some interior point (t_0, x_0) of Ω_t and that $u \leq M$ throughout Ω_t. If $\mathcal{L}$ is any horizontal line in the interior of Ω_t which contains (t_0, x_0), then $u < M$ on $\mathcal{L}$.*

Proof. We proceed by contradiction : we suppose that $u = M$ at some interior point (t_0, x_1) on $\mathcal{L}$ and that $u < M$ at (t_0, x_0).
Without loss of generality, we suppose that $x_1 < x_0$, so we have

$$u(t_0, x) < M \quad \forall\, x \mid x_1 < x \leq x_0 .$$

We define the positive value d_0 as either $x_0 - x_1$ or the minimum of the distances from any point of the line segment $(t = t_0, x_1 \leq x \leq x_0)$ to $\partial\Omega_t$, whichever is smaller, namely

$$d_0 = min\big\{ |x_0 - x_1| , \min_{x \in \mathcal{L} \,|\, x_1 \leq x \leq x_0} dist(x, \partial\Omega_t) \big\}.$$

For $x_1 \leq x \leq x_1 + d_0$, let the positive function $d(x)$ be the distance from (t_0, x) to the nearest point in Ω_t where $u = M$. We have that

$$d(x) \leq x - x_1 , \tag{1.14}$$

because $u(t_0, x_1) = M$. By Lemma 2, this point is directly above or below (t_0, x), that is (see Figure 1.3)

$$u(t_0 + d(x), x) = M \quad \text{or} \quad u(t_0 - d(x), x) = M . \tag{1.15}$$

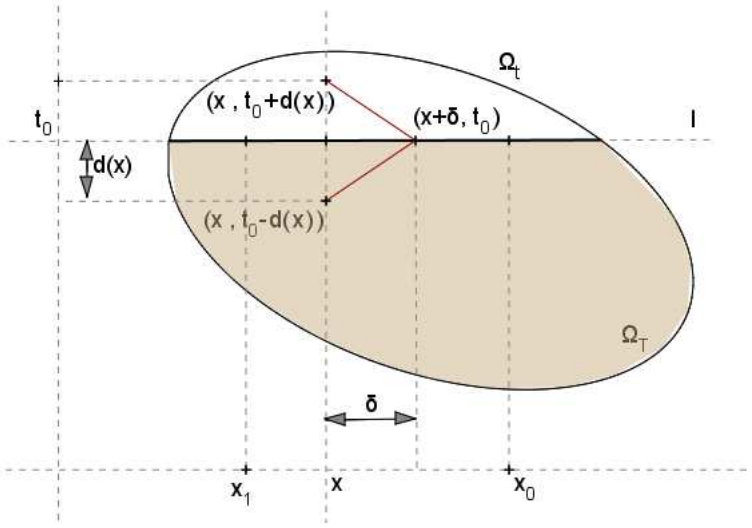


Figure 1.3: geometric construction for Lemma 3

We choose an arbitrary δ such that $0 < \delta < d(x)$, and we note that the distance from a point $(t_0, x+\delta)$ to $(t_0 \pm d(x), x)$ is $\sqrt{d(x)^2 + \delta^2}$. We compute a Taylor's expansion of that distance and we see that

$$d(x+\delta) < d(x) + \frac{\delta^2}{2d(x)}\,. \tag{1.16}$$

Replacing x by $x+\delta$ and δ by $-\delta$ we also observe that

$$d(x+\delta) > \sqrt{d(x)^2 - \delta^2}\,. \tag{1.17}$$

For the choice of δ, we have that $\sqrt{d(x)^2 - \delta^2} > 0$.
Now, we divide the interval $(x, x+\delta)$ into N equal parts,

$$x_j = x + \frac{j}{N}\delta\,, \quad j = 0, \ldots, N\,,$$

we apply the previous inequalities (1.16)-(1.17) and sum on j to find

$$d(x+\delta) - d(x) = \sum_{j=0}^{N-1} d(x_{j+1}) - d(x_j) \leq \frac{\delta^2}{2N^2 d(x + \frac{j}{N}\delta)} \leq \frac{\delta^2}{2N^2\sqrt{d(x)^2 - \delta^2}}\,.$$

For the arbitrary of N, we see that

$$d(x+\delta) \leq d(x) \quad \text{for } N \longrightarrow \infty\,.$$

Therefore $d(x)$ is a non-increasing function of x. Since (1.14) holds also for $x \to x_1$, we can deduce that

$$d(x) \equiv 0 \quad \text{for } x_1 < x < x_1 + d_0\,.$$

In other words, by (1.15) we have that

$$u(t_0, x) = M \quad \text{for } x_1 < x < x_1 + d_0\,,$$

and this statement clearly contradicts the hypothesis that $u < M$, for all points such that $x_1 < x \leq x_0$. □

Remark 3. *Lemma 3 states that, if there is a single interior point where $u = M$, then u remains constant, i.e. $u \equiv M$, along the largest horizontal segment containing this point whose interior lies in Ω_t.*

Lemma 4. *Suppose that in the lower half*

$$K_{t_1} = \left\{(t, x) \mid t \leq t_1\,,\ (t - t_1)^2 + (x - x_1)^2 < R^2\right\}$$

of a disk K centered at $P = (t_1, x_1)$, the solution u satisfies the differential inequality $L(u) \geq 0$, with L as in Lemma 2. Suppose that $u < M$ in the portion of K where $t < t_1$. Then $u(P) < M$.

Proof. We define the auxiliary function v as

$$v(x, t) = e^{-\left[\alpha(t-t_1)^2 + (x-x_1)^2\right]} - 1$$

and we apply the partial differential operator,

$$L(v) = e^{-\left[\alpha(t-t_1)^2 + (x-x_1)^2\right]}\left[4a\,(x - x_1)^2 - 2a - 2b\,(x - x_1) + \alpha\right],$$

with $\alpha > 0$ and large enough so that

$$L(v) \geq 0 \quad \text{in } K_{t_1}\,.$$

We introduce the parabola

$$\Pi = \alpha(t - t_1)^2 + (x - x_1)^2,$$

which is tangent to the line $t = t_1$ at the point P. We denote by C' the portion of ∂K which is below the parabola, with the endpoints, by C'' the portion of Π located within the disk K, and by D the region enclosed by C' and C'' (see Figure 1.4).
By hypothesis, there exists a value $\eta > 0$ such that

$$u \leq M - \eta \quad \text{on } C'.$$

We define the function w as in (1.12), such that

$$\begin{aligned} L(w) = L(u) + \varepsilon L(v) > 0 \quad &\text{in } D\,, \\ w = u + \varepsilon v < M \quad &\text{on } C', \\ w = u + \varepsilon v \leq M \quad &\text{on } C''. \end{aligned}$$

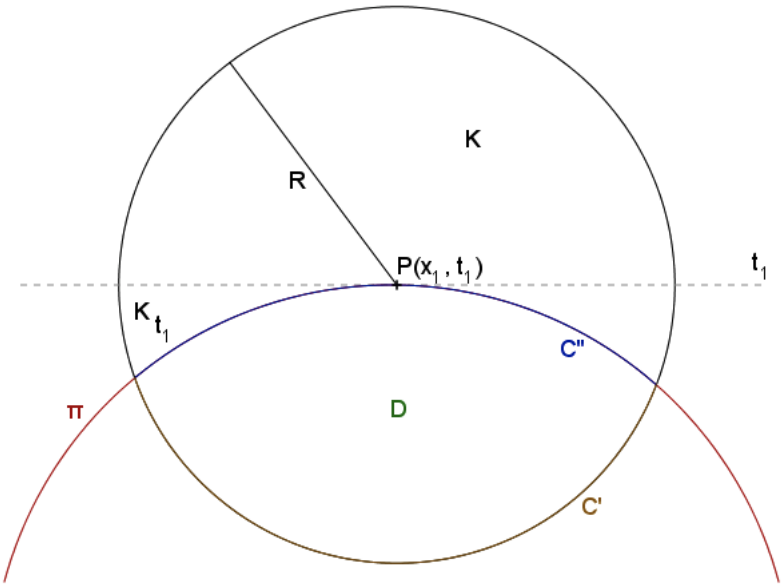


Figure 1.4: geometric construction for Lemma 4

The first condition above shows that w cannot attain its maximum in D, by the Proposition 2, so the maximum M occurs at the point P, and therefore

$$\frac{\partial w}{\partial t} \geq 0\,, \quad \frac{\partial v}{\partial t} = -\alpha < 0\,, \quad \frac{\partial u}{\partial x} = 0\,, \quad \frac{\partial^2 u}{\partial x^2} \leq 0\,,$$

implies that the hypothesis $L(u) \geq 0$ is contradicted, so the Lemma is demonstrated. □

In what follows, the region

$$\Omega_T = \{(t, x) \in \Omega_t \mid t \leq T\}$$

is a portion of the domain Ω_t, which can be defined for all $T > 0$.
Moreover, we assume that u is continuously differentiable in both of its variables, and twice differentiable in x throughout Ω_T, with $\frac{\partial u}{\partial t}|_{t=T}$ defined as a one-sided derivative. On the basis of the previous lemmas, we can now establish the following result.

Theorem 1. *Assume that the differential inequality*

$$L(u) = a(x,t)\frac{\partial^2 u}{\partial x^2} + b(x,t)\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} \geq 0$$

holds in Ω_T, with a and b bounded functions, and that L is uniformly parabolic. If $u \leq M$ in Ω_T, and $u(T, x_1) = M$, then $u \equiv M$ at every point $(t, x) \in \Omega_T$ which can be connected with (T, x_1) by a horizontal and vertical line segment, both of which lie in Ω_T (see Figure 1.5).

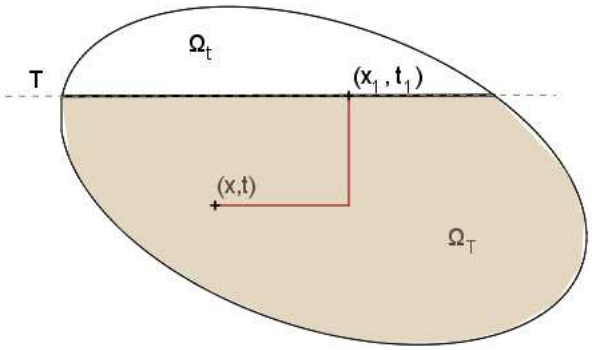


Figure 1.5: horizontal and vertical paths between two points

Proof. Let (t_0, x_1) be an arbitrary point in Ω_T such that $u(t_0, x_1) < M$ and the line segment $\mathcal{L} = \{(t, x) \mid t_0 \leq t \leq T, x = x_1\}$ lies in Ω_t.
Let τ be the least upper bound of values of u on $\mathcal{L}$ such that $u(t, x_1) < M$. By continuity, $u(\tau, x_1) = M$ while Lemma 3 shows that there is $R > 0$ such that $u < M$ for $t_0 \leq t < \tau$, $|x - x_1| < R$, as we see in Figure 1.6.
This leads to a contradiction of Lemma 4, because we have found a disk K where in its lower half K_τ, it holds $L(u) \geq 0$ and $u < M$ where $t < \tau$, but we also have that $u(\tau, x_1) = M$. □

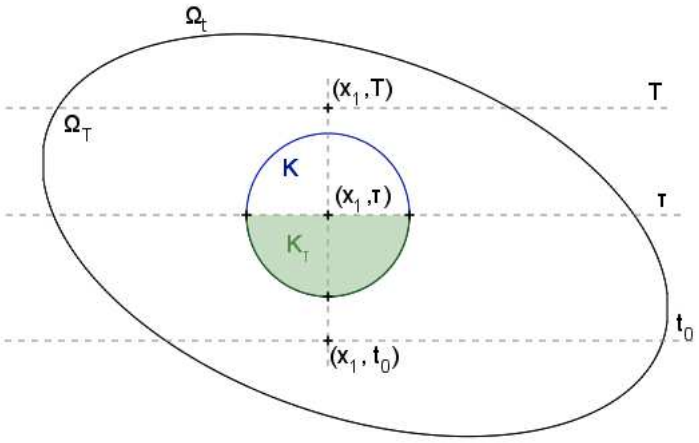


Figure 1.6: geometric construction for the theorem in dimension $n = 1$

Remark 4. *Theorem 1 can be further combined with Lemma 3 to identify the entire region where the solution is constant and its value is equal to the maximum, if this is attained at an interior point. Indeed, once obtained a point Q at which $u = M$, the maximum, so $u \equiv M$ on the largest horizontal segment in Ω_t containing Q. Moreover, if P is a point of Ω_t which can be connected with Q by a path in Ω_t consisting only of horizontal and vertical*

segments, then $u(P) = M$. So, in general, if we have a domain Ω_t connected and the maximum M is at an interior point, then $u \equiv M$ in all the domain. Obviously, this is no longer true if the computational domain is the union of connected components.

1.3.2 The n-dimensional case

We can extend the previous results to the n-dimensional case in a completely straightforward way for the general parabolic operatorgiven by (1.3).
In this section, we provide only the adaptation of the last theorem, by keeping the notations coherent with (1.8).

Theorem 2. *Let u satisfy the uniformly parabolic differential inequality*

$$L(u) = \sum_{i,j=1}^{n} a_{ij}(t,z)\frac{\partial^2 u}{\partial z_i \partial z_j} + \sum_{i=1}^{n} b_i(t,z)\frac{\partial u}{\partial z_i} - \frac{\partial u}{\partial t} \geq 0$$

in a region $\Omega_T \subset \mathbb{R}^+ \times \mathbb{R}^n$, and suppose that the coefficients of L are bounded functions. Let $\Omega_{\bar{t}}$ be a section of the domain, such that

$$\Omega_{\bar{t}} = \{(t,z) \mid 0 < t \leq \bar{t} < T,\ z \in \Omega\}.$$

Suppose that the maximum of u in $\Omega_{\bar{t}}$ is M and that it is attained at a point $P = (t,z)$ of $\Omega_{\bar{t}}$. Thus if Q is a point of Ω_T which can be connected to P through a path in Ω_T consisting only of horizontal segments and upward vertical segments, then $u(Q) = M$ (see Figure 1.7).

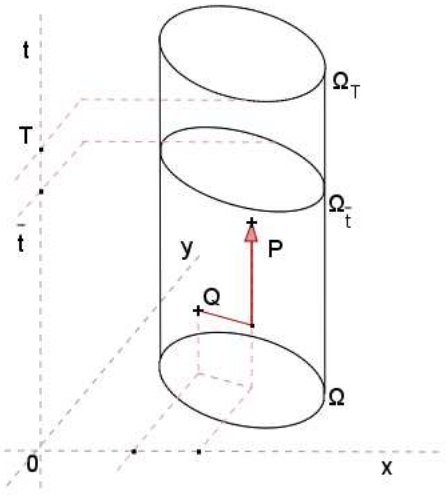


Figure 1.7: geometric construction for the theorem in dimension $n > 1$

Proof. The proof is derived in exactly the same way as for Theorem 1, by replacing the auxiliary function v in the proof of Lemma 2 by

$$v(x,t) = e^{-\alpha\left[(t-T)^2+\sum_{i=1}^{n}(x_i-\bar{x_i})^2\right]} - e^{-\alpha R^2},$$

while the auxiliary function corresponding to Lemma 4 is given by

$$v(x,t) = e^{-\left[\alpha(t-T)^2+\sum_{i=1}^{n}(x_i-\bar{x_i})^2\right]} - 1\,.$$

We replace disks by $(n+1)$- dimensional balls, and the parabola in the proof of Lemma 4 by the hyper-paraboloid

$$\alpha(t-T)^2 + \sum_{i=1}^{n}(x_i-\bar{x_i})^2) = 0\,.$$

The same arguments as in Section 1.3.1 allows to conclude the proof. □

We can recast the maximum principle in a more general framework, according to [11] and [10]. The following definition will be useful later on, in order to prove the non-negativity property, which in turn can be used to check the maximum principle in practical situations, because they are closely related : if the parabolic operator satisfies the maximum principle, then it satisfies the non-negativity property, and vice versa.

We denote by $\mathcal{D}_L$ the domain of the operator L in (1.3), defined as the space of functions u such that

$$\mathcal{D}_L := \{u \in C(\Omega_T \cup \Gamma_T) \mid \partial^{\zeta} u\,, u_t \text{ exist and are bounded}\},$$

where $\zeta = (\zeta_1, \ldots, \zeta_n)$ is a multi-index, with $0 < |\zeta| < n$ for $|\zeta| = \zeta_1 + \ldots + \zeta_n$. With this notation, the operator $L(u)$ is bounded on Ω_T for each $u \in \mathcal{D}_L$ and $0 < t < T$. Therefore, $inf_{\Omega_T} L(u)$ and $sup_{\Omega_T} L(u)$ are finite.
According to the previous results, if u attains its maximum at an interior point of the connected domain Ω_T, so the solution is necessarily constant in all the domain; thus, the maximum is attained at the parabolic boundary.

Definition 3 (maximum/minimum principle)**.** *We say that the parabolic operator defined by* (1.3) *satisfies the maximum/minimum principle if, for any function $u \in \mathcal{D}_L$, the inequality*

$$\min_{\Gamma_{\bar{t}}} u + \bar{t} \cdot \min\{0, \inf_{\Omega_{\bar{t}}} L(u)\} \leq u(\bar{t}, z) \leq \max_{\Gamma_{\bar{t}}} u + \bar{t} \cdot \max\{0, \sup_{\Omega_{\bar{t}}} L(u)\}$$

is satisfied for all $0 < \bar{t} < T$ and $z \in \Omega$.

We remark that, if the operator L is identically zero, i.e. $L(u) = 0$ as for the diffusion equations (1.1) treated in this report, the solution u must satisfy

$$\min_{\Gamma_{\bar{t}}} u \leq u(\bar{t}, z) \leq \max_{\Gamma_{\bar{t}}} u \qquad \text{for all } \; 0 < \bar{t} < T, \; z \in \Omega \,, \tag{1.18}$$

so, either the solution is constant or it has a maximum/minimum value at a point of the (local in time) parabolic boundary $\Gamma_{\bar{t}}, 0 < \bar{t} < T$. That is precisely equivalent to the results of the previous theorems and, therefore, the parabolic operator satisfies the maximum/minimum principle above. Moreover, the solution can be estimated from above and below.

1.4 Applications of the maximum principle

The importance of the maximum/minimum principle for parabolic operators lies in the fact that it implies other important properties, such as the *uniqueness* of the solution, the *comparison principle*, and the *non-negative property*. All these statements are also important for the numerical aspects.

1.4.1 Uniqueness

We show that it is possible to establish the uniqueness of a solution by means of the maximum principle alone. That is, there can be at most one solution to the equation $L(u) = 0$ which satisfies certain boundary conditions, that will be defined from time to time.

First, we study the one-dimensional case in a rectangular domain (1.10) with the operator L defined by (1.9). We also consider the following initial data and Dirichlet boundary conditions,

$$\begin{aligned} u(0, x) &= u_0(x) \quad \text{for } \; x \in [0, \Lambda] \\ u(t, 0) &= g_1(t) \quad \text{for } \; t \in [0, T) \\ u(t, \Lambda) &= g_2(t) \quad \text{for } \; t \in [0, T) \end{aligned} \tag{1.19}$$

where the functions g_1 and g_2 are bounded and continuous, and the initial data u_0 is positive and bounded.

Theorem 3. *Let u be a solution in Ω_T to the uniformly parabolic equation*

$$L(u) = a(t, x)\frac{\partial^2 u}{\partial x^2} + b(t, x)\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} = 0 \,, \tag{1.20}$$

satisfying the initial and boundary conditions (1.19)*. If v is another solution to* (1.20) *with the same initial and boundary conditions, then $u \equiv v$ in Ω_T.*

Proof. We define the function $w = u - v$ and observe that

$$L(w) = a(t, x)\frac{\partial^2 w}{\partial x^2} + b(t, x)\frac{\partial w}{\partial x} - \frac{\partial w}{\partial t} = 0 \,,$$

by the linearity of the operator L, and w satisfies the following conditions,

$$\begin{aligned} w(0,x) &= 0 \quad \text{for } x \in [0,\Lambda] \\ u(t,0) &= 0 \quad \text{for } t \in [0,T) \\ u(t,\Lambda) &= 0 \quad \text{for } t \in [0,T) \end{aligned}$$

According to the maximum/minimum principle, the function w cannot have a positive maximum in Ω_T, and so $w \leq 0$ everywhere. Applying the same reasoning to $-w$, we obtain the $w \geq 0$ in Ω_T. Hence the only possibility is $w \equiv 0$, so the two solutions u and v coincide. □

The result just established in the one-dimensional framework extends to solutions of parabolic operators in the $n-$dimensional case. We refer to the domains defined in (1.8) and the parabolic operator (1.3). In particular, we consider the Neumann boundary conditions but, obviously, the result holds also in the case of Dirichlet boundary conditions $u|_{\Gamma_T} = g(t,z)$, because the function w defined in the previous proof is such that $w|_{\Gamma_T} = 0$.

Theorem 4. *Let u be a solution in Ω_T to the uniformly parabolic equation*

$$L(u) = \sum_{i,j=1}^{n} a_{ij}(t,z)\frac{\partial^2 u}{\partial z_i \partial z_j} + \sum_{i=1}^{n} b_i(t,z)\frac{\partial u}{\partial z_i} - \frac{\partial u}{\partial t} = 0\,, \tag{1.21}$$

satisfying the initial and Neumann boundary conditions

$$\begin{aligned} u(0,z) &= u_0(z) \quad \textit{for } z \in \Omega \\ \nabla u \cdot \vec{\nu}|_{\Gamma_T} &= g(t,z) \quad \textit{for } (t,z) \in \Gamma_T \end{aligned} \tag{1.22}$$

where $\vec{\nu}$ is the normal outward vector on Γ_T, and g is bounded and sufficiently regular in Γ_T. If v is another solution to (1.21) *with the same initial and boundary conditions, then $u \equiv v$ in Ω_T.*

Proof. The result is derived in exactly the same way as in the previous case. □

1.4.2 Monotonicity and comparison principle

Another important application of the maximum/minimum principle for parabolic equations is the *comparison principle*. The idea behind it is that if u and v are both solutions to the problem (1.21), and the initial data are such that $u_0(z) \geq v_0(z)$, then $u(t,z) \geq v(t,z)$ for all $t > 0$.

Theorem 5. *Suppose that u is a solution to the parabolic equation* (1.21) *in Ω_T, with the initial and Dirichlet boundary conditions*

$$\begin{aligned} u(0,z) &= u_0(z) \quad \textit{for } z \in \Omega \\ u(t,z) &= g(t,z) \quad \textit{for } (t,z) \in \Gamma_T \end{aligned} \tag{1.23}$$

where g is bounded and sufficiently regular in Γ_T. We assume that v_1 and v_2 are sub- and super-solutions, respectively, i.e. they satisfy

$$L(v_1) \leq L(u) \leq L(v_2) \quad in \ \ \Omega_T,$$

and, moreover, it holds that

$$\begin{aligned} v_1(0,z) \leq u_0(z) \leq v_2(0,z) \quad & for \ \ z \in \Omega \\ v_1(t,z) \leq g(t,z) \leq v_2(t,z) \quad & for \ \ (t,z) \in \Gamma_T \end{aligned}$$

Then

$$v_1(t,z) \leq u(t,z) \leq v_2(t,z) \quad for \ all \ \ (t,z) \in \Omega_T.$$

We remark that a similar result is still valid, under appropriate supplementary hypotheses, for the more general case $L(u) = f(t,z)$, with some external *source term* function, and also for nonlinear parabolic operators [29].

The previous theorem shows that the initial-boundary value problem for the parabolic operator (1.21) is stable under small perturbations, and this is an important concept for the well-posedness of the problem.

Definition 4. *A mathematical problem is said to be well-posed if it presents a unique solution that is stable under small perturbations. Otherwise, it is said to be ill-posed.*

The results stated above guarantee that, if we take a solution to (1.21), together with (1.22) or (1.23), for some initial data u_0, and we perturb this, that is we study the problem with an initial data like $\bar{u}_0 = u_0 + \varepsilon$, for a parameter $\varepsilon > 0$ small enough, we have that the solutions u and $\bar{u}$ remain close over the time. Moreover, according to the comparison principle, if we consider a special case in which the initial data $\bar{u}_0 = 0$, then the solution $\bar{u}(t,z)$ is identically equal to zero, and therefore any other solution u, with initial data $u_0 > 0$, is such that $u(t,z) \geq 0$. This last statement is better explained in the following section.

1.4.3 Non-negativity property

Now we can present the strict connection between the maximum principle (MP) and the non-negativity (NP) for the partial differential operator (1.21), namely

$$MP \iff NP \tag{1.24}$$

Before demonstrating this statement, we give the following definition from [11].

Definition 5 (Non-negativity preservation)**.** *The differential operator L is said to be non-negativity preserving if the following implication is satisfied,*

$$\min_{\Gamma_{\bar{t}}} u \geq 0 \ \ and \ \ L(u) \geq 0 \ \ on \ \ \Omega_{\bar{t}} \quad \Longrightarrow \quad u \geq 0 \ \ on \ \ \Omega_{\bar{t}} \ \ for \ all \ \ 0 < \bar{t} < T\,.$$

This means that, if the initial data $u_0 \geq 0$ and the operator $L(u)$ is nonnegative, so the solution u remains nonnegative over the all domain.

Theorem 6. *The operator L defined in* (1.3) *satisfies the maximum/minimum principle if and only if it preserves the non-negativity.*

Proof. The necessity of the condition is trivial because derived directly from the above inequality. Then, we show the sufficiency. We choose an arbitrary function $u \in \mathcal{D}_L$ and we define

$$\bar{u} = u - \min_{\Gamma_{\bar{t}}} u - \bar{t} \cdot \min\{0, \inf_{\Omega_{\bar{t}}} L(u)\}.$$

It follows immediately that $\bar{u} \geq 0$ on $\Gamma_{\bar{t}}$. Now we apply the operator L to the function $\bar{u}$, obtaining $L(\bar{u}) = L(u) - \min\{0, \inf_{\Omega_{\bar{t}}} L(u)\}$ by the definition of L, which implies that $L(\bar{u}) \geq 0$ on $\Omega_{\bar{t}}$. By virtue of the non-negativity preservation assumption, we have that $\bar{u} \geq 0$ on $\Omega_{\bar{t}}$. Thus, the lower estimation $\min_{\Gamma_{\bar{t}}} u + \bar{t} \cdot \min\{0, \inf_{\Omega_{\bar{t}}} L(u)\} \leq u(\bar{t}, z)$ is satisfied. For the upper estimation, we choose the function

$$\bar{u} = \max_{\Gamma_{\bar{t}}} u - u + \bar{t} \cdot \min\{0, \inf_{\Omega_{\bar{t}}} L(u)\}$$

and the steps of the proof are similar. □

Chapter 2

Numerical Schemes for Parabolic Conservation Laws

In this chapter we describe how to solve numerically the two-dimensional diffusion equation (1.1) with time variable $t > 0$ and $(x, y) \in \mathbb{R}^2$.
According to [30],[22],[21], the goal is to approximate the exact solution to (1.1) by finding some discrete function that satisfies a given relationship between various of its derivatives, on some region of space and time, along with boundary conditions.

Typically, to treat this kind of second order problems, one can use the *finite element method* because, especially in the case of anisotropic and heterogeneous diffusion, this approach allows to have a mesh that may follow the behaviour of the diffusion to catch the nature of the diffusion coefficients, thus as consistent as possible with the exact solution. Nevertheless, in some situations, we would have not only a good approximation but also to enable such schemes to properly perform on CUDA GPU applications [17]. So, we need particular schemes, that are easy to handle to make them *parallelizable*, and discretizations on fixed rectangular grids which are compatible with the pixel structure of the digital grids. We refer to such discrete domains as *Cartesian grids*, i.e. rectangular domains with a Cartesian meshing.
At a discrete time $t^n, n = 1, 2, ...$, and at any point (x_i, y_j), $i, j = 1, 2, ...$, we calculate the approximate solution $u(t^n; x_i, y_j)$ using the values at previous times. Most of the schemes we take into account are *time-explicit schemes*, but sometimes we operate also with *semi-implicit* or *fully-implicit* schemes. Explicit schemes are the simplest to code and, therefore, they are used in combination with GPU architectures almost exclusively. Due to their local structure, they are well-suited for parallel applications. Unfortunately, they suffer from the fact that fairly small time-step sizes are needed in order to ensure *stability* [30]. Semi-implicit schemes, or the purely implicit schemes, are considered in the case of the one-dimensional equations, as they possess better stability properties, but for other situations they exhibit too much computational costs.

2.1 Discretization of the spatial domain

For two-dimensional numerical simulations, we consider the computational domain $\Omega_T = \Omega \times [t_0, T]$, where Ω is a rectangular subset of $\mathbb{R}^2$ located from a_x to b_x along the x-direction and from a_y to b_y along the y-direction (see Figure 2.1). We apologize for the misleading notation, as the above symbols could be confused with the entries of the diffusion matrix, but we guarantee that no confusion is possible in the following.
For the time, we discretize the interval $[t_0, T]$ by means of

$$t^n = t_0 + n\,\Delta t\,, \qquad n = 0, 1, 2, \dots\,,$$

so that we have

$$t^0 = t_0\,, \quad t_{N_t} = T, \quad \Delta t = \frac{T - t_0}{N_t} = t^{n+1} - t^n, \quad n = 0, 1, ..., N_t - 1\,.$$

Typically, the parameter Δt will be chosen to satisfy the stability condition, according to the equation and the specific numerical scheme under consideration; usually, the time-step Δt must be less than some given value, this being determined through the so-called *CFL stability condition* [25],[34],[22].

For the set Ω, we can decide among two types of discrete domains.

1. The first one as in Figure 2.1, where the *nodes* of the spatial grid are the points (x_i, y_j) located at the boundaries of the grid cells.

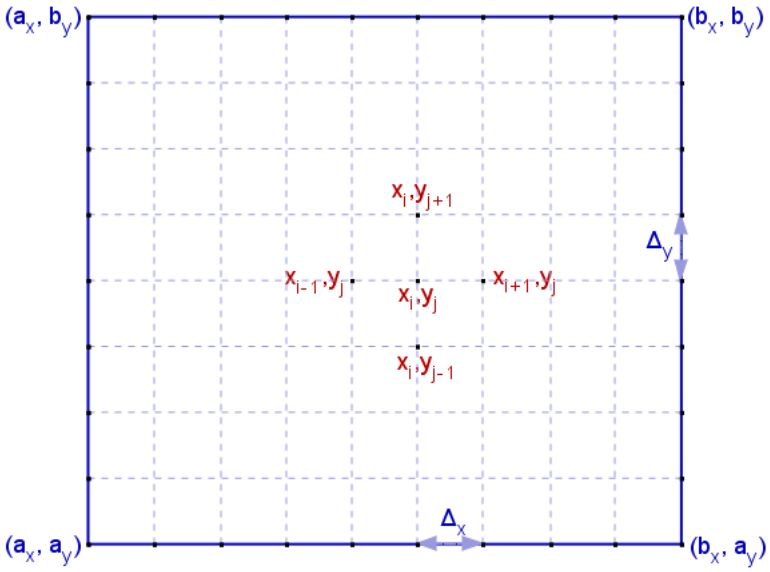


Figure 2.1: spatial grid for the Finite Difference method

We have divided the one-dimensional intervals $[a_x, b_x]$ and $[a_y, b_y]$ in N_x and N_y sub-intervals, respectively, such that

$$\Delta x = \frac{b_x - a_x}{N_x}\,, \qquad \Delta y = \frac{b_y - a_y}{N_y}\,,$$

and with this notation we have that

$$\begin{aligned} x_i &= a_x + i\Delta x\,, \quad i = 0, 1, \ldots, N_x \\ y_j &= a_y + j\Delta y\,, \quad j = 0, 1, \ldots, N_y \\ x_0 = a_x\,, \quad x_{N_x} &= b_x\,, \quad y_0 = a_y\,, \quad y_{N_y} = b_y\,. \end{aligned}$$

In this framework, we will use **finite difference schemes** based on Taylor's expansions at the grid points (x_i, y_j), in order to calculate the approximate solution at the generic points of the grid.

2. The second one as in Figure 2.2, where we have introduced the *interfacial points* given by $x_{i+\frac{1}{2}}$ and $y_{j+\frac{1}{2}}$ with

$$\begin{aligned} x_i &= \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2}\,, \quad i = 0, 1, \ldots, N_x\,, \\ y_j &= \frac{y_{j+\frac{1}{2}} + y_{j-\frac{1}{2}}}{2}\,, \quad j = 0, 1, \ldots, N_y\,. \end{aligned}$$

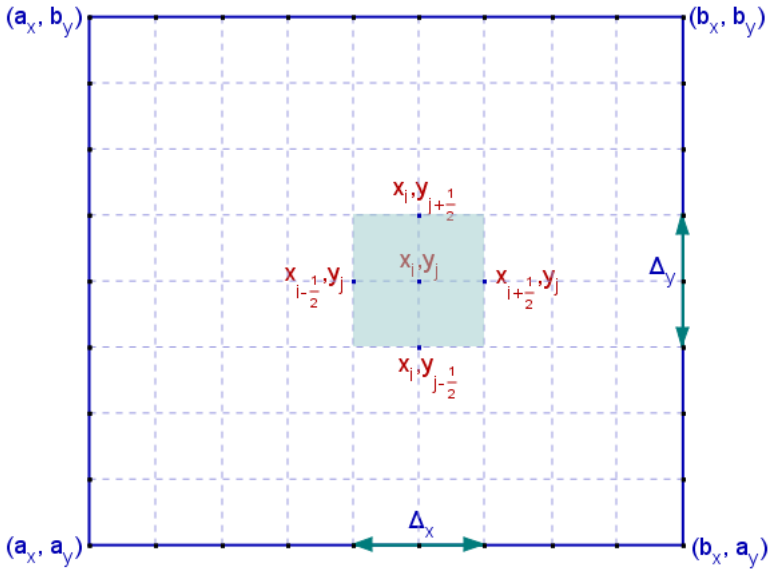


Figure 2.2: spatial grid for the Finite Volume method

The grid nodes are still the points (x_i, y_j) but now they are located at the center of the cells, the so-called *finite volumes* denoted by

$$C_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}] \tag{2.1}$$

and the length of the intervals is computed as

$$\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}\,, \qquad \Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}\,. \tag{2.2}$$

According to these definitions, the *boundaries* of the domain are

$$a_x = x_0 = x_{\frac{1}{2}} - \frac{\Delta x}{2}\,, \quad b_x = x_{N_x} = x_{N_x-\frac{1}{2}} + \frac{\Delta x}{2}\,,$$
$$a_y = y_0 = y_{\frac{1}{2}} - \frac{\Delta y}{2}\,, \quad b_y = y_{N_y} = y_{N_y-\frac{1}{2}} + \frac{\Delta y}{2}\,.$$

The numerical approach in this case is that of **finite volume schemes** : at fixed time t^n the approximate solution U^n is defined by the values calculated at points (x_i, y_j) and extended to the whole cell C_{ij}, namely

$$U^n(x,y) = \sum_{i=0}^{N_x}\sum_{j=0}^{N_y} u_{ij}^n\, \chi_{C_{ij}}(x,y)\,, \tag{2.3}$$

where $\chi_{C_{ij}}$ is the characteristic function of the cell C_{ij}. We make the assumption that we approximate the solution $u(t^n; x_i, y_j)$ through the *integral average* on the grid cells,

$$u_{ij}^n \simeq \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \left[\frac{1}{\Delta x \Delta y}\int_{C_{ij}} u(t;x,y)\,dx\,dy\right] dt\,, \tag{2.4}$$

where C_{ij} is defined in (2.1) with volume's size $mes(C_{ij}) = \Delta x \Delta y$.

In the definitions (2.2), we have considered that Δx and Δy have fixed length for all the grid cells, but we could also use a *nonuniform mesh*. For the purpose of this report, there is no stringent reason to use nonuniform grids, because when performing the simulations in CUDA GPU, an appropriately small Δx can be chosen, thanks to the computational power of GPU architectures. For a general analysis, we can simply impose

$$\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}\,, \qquad \Delta y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}\,,$$

where the length of the intervals can vary according to the cell in which we operate. This type of spatial discretization is useful if we want to adjust the grid of the domain to the initial data for a given problem : if the function $u_0(x,y)$ has rapid variations in the (x,y)-domain, we can choose to use a nonuniform grid to catch the exact behaviour of this function.

As we will see, there are some cases in which we will use the second type of grid also in the context of finite difference schemes, and therefore we refer to this approach as *staggered grid*, where the derivatives of the solution are calculated by Taylor's expansions at the grid points (x_i, y_j) through the values at the interfacial points.

2.2 The Finite Difference method

We focus on the first case presented above. The finite difference scheme proceeds by replacing the derivatives of functions by the *incremental ratios*, that is the finite difference between values at the grid points (we refer to [25] and [22], for instance).

For the mathematical derivation of the method, we consider a function u of one variable x in the given interval I, that is assumed to be smooth, meaning that we can differentiate the function several times and each derivative is a well-defined bounded function over the interval. As said before, let $u(x_i)$ be the value of the function at grid point x_i , and u_i its approximate value. We begin by approximating the first order derivatives $u'(x_i)$ by means of finite differences based on the given set of points. So, we start from

$$u'(x_i) = \lim_{\Delta x \to 0^+} \frac{u(x_i + \Delta x) - u(x_i)}{\Delta x}$$

and we approximate it through the incremental ratio

$$u'_i = \frac{u(x_{i+1}) - u(x_i)}{\Delta x}, \qquad 0 \le i \le N_x - 1\,. \tag{2.5}$$

The right-hand side of this expression is called *forward finite difference*. Instead of (2.5), we can choose other approximations, for example we could employ a centered incremental ratio and obtain the *centered finite difference*

$$u'_i = \frac{u(x_{i+1}) - u(x_{i-1})}{2\Delta x}, \qquad 1 \le i \le N_x - 1\,. \tag{2.6}$$

Finally, with a similar procedure, we can derive the *backward finite difference*

$$u'_i = \frac{u(x_i) - u(x_{i-1})}{\Delta x}, \qquad 1 \le i \le N_x\,.$$

Now, let us approximate the second order derivatives of u. We use Taylor's expansions again, with the space-step Δx. We suppose that $u \in C^4([a_x, b_x])$ and we sum the following results for $u(x_{i+1})$ and $u(x_{i-1})$, that is

$$\begin{aligned} u(x_{i+1}) &= u(x_i) + \Delta x\, u'(x_i) + \frac{\Delta x^2}{2} u''(x_i) + \frac{\Delta x^3}{6} u'''(x_i) + O(\Delta x^4), \\ u(x_{i-1}) &= u(x_i) - \Delta x\, u'(x_i) + \frac{\Delta x^2}{2} u''(x_i) - \frac{\Delta x^3}{6} u'''(x_i) + O(\Delta x^4), \end{aligned} \tag{2.7}$$

thus obtaining

$$u''_i = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{\Delta x^2}. \tag{2.8}$$

This is a centered approximation and it is said to be $O(\Delta x^2)$ according to the order of the incremental ratio (2.8) for the exact function u from (2.7).

We remark that one could have chosen to approximate second order derivatives using compositions of the centered first order discretization (2.6), i.e.

$$u_i'' = \frac{u_x(x_{i+1}) - u_x(x_{i-1})}{2\Delta x} = \frac{u(x_{i+2}) - 2u(x_i) + u_x(x_{i-2})}{4\Delta x^2}, \tag{2.9}$$

but, in addition to producing a larger *stencil* and, therefore, requiring to set more boundary data, the last approximation is more *diffusive* since its consistency error behaves like $\frac{\Delta x^2}{3} u_{xxxx}$ whereas we have $\frac{\Delta x^2}{12} u_{xxxx}$ for (2.8).

2.2.1 Approximation of parabolic equations

Within the framework introduced above, we can attempt at approximating the equation (1.1). For the sake of readability, we adopt a slight abuse of notation, by identifying numerical and exact values of the function u on the grid points, whereas the correct notation should be $u_{ij}^n \simeq u(t^n; x_i, x_j)$.

For the time variable, we use the forward finite difference scheme,

$$u_{ij}^{n+1} = u_{ij}^n + \Delta t \frac{\partial u_{ij}^n}{\partial t} + \frac{\Delta t^2}{2} \frac{\partial^2 u_{ij}^n}{\partial t^2} + O(\Delta t^3),$$

and we approximate the time derivative as

$$\frac{\partial u}{\partial t}(t^n; x_i, y_j) \simeq \frac{u_{ij}^{n+1} - u_{ij}^n}{\Delta t}, \tag{2.10}$$

with the following *truncation error*

$$\tau_{ij} = \frac{\Delta t}{2} \frac{\partial^2 u_{ij}^n}{\partial t^2} + O(\Delta t^2) = O(\Delta t).$$

If $||\tau_{ij}|| \to 0$ as $\Delta t \to 0$ for some appropriate norm to be made explicit later, we say that the approximation is *consistent*; moreover, if $||\tau_{ij}|| = O(\Delta t^p)$ for some integer $p > 0$, we also say that the scheme has p as *order of accuracy*. So, the scheme (2.10) has an order of accuracy of 1 and it is consistent.

For the space variable, we have to calculate second order derivatives. Referring to Figure 2.3, let us begin with the second order centered scheme (2.6) for first order derivatives on a staggered grid. We denote by $u_{i+\frac{1}{2},j}$ and $u_{i,j+\frac{1}{2}}$ the numerical solution evaluated at the half points $(x_{i+\frac{1}{2}}, y_j) = (x_i + \frac{\Delta x}{2}, y_j)$ and $(x_i, y_{j+\frac{1}{2}}) = (x_i, y_j + \frac{\Delta y}{2})$, respectively, so that

$$\begin{aligned} u_{i+\frac{1}{2},j} &\simeq u_{ij} + \frac{\Delta x}{2} \frac{\partial u_{ij}}{\partial x} + \frac{\Delta x^2}{4} \frac{\partial^2 u_{ij}}{\partial x^2} + O(\Delta x^3), \\ u_{i-\frac{1}{2},j} &\simeq u_{ij} - \frac{\Delta x}{2} \frac{\partial u_{ij}}{\partial x} + \frac{\Delta x^2}{4} \frac{\partial^2 u_{ij}}{\partial x^2} + O(\Delta x^3). \end{aligned} \tag{2.11}$$

Subtracting the second from the first above, we get

$$u_{i+\frac{1}{2},j} - u_{i-\frac{1}{2},j} \simeq \Delta x \frac{\partial u_{ij}}{\partial x} + O(\Delta x^3) \Longrightarrow \frac{\partial u_{ij}}{\partial x} \simeq \frac{u_{i+\frac{1}{2},j} - u_{i-\frac{1}{2},j}}{\Delta x} + O(\Delta x^2).$$

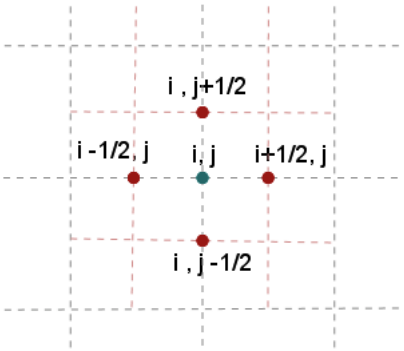


Figure 2.3: staggered grid for first order derivatives

Intuitively, with the same procedure for the y-direction, this leads to obtain

$$\begin{aligned} \frac{\partial u}{\partial x}(t^n; x_i, y_j) &\simeq \frac{u^n_{i+\frac{1}{2},j} - u^n_{i-\frac{1}{2},j}}{\Delta x}\,, \\ \frac{\partial u}{\partial y}(t^n; x_i, y_j) &\simeq \frac{u^n_{i,j+\frac{1}{2}} - u^n_{i,j-\frac{1}{2}}}{\Delta y}\,. \end{aligned} \tag{2.12}$$

As before, the schemes (2.12) are consistent and with order of accuracy 2 with respect to Δx and Δy. Alternatively, we can consider the schemes

$$\begin{aligned} \frac{\partial u}{\partial x}(t^n; x_i, y_j) &\simeq \frac{u^n_{i+1,j} - u^n_{i-1,j}}{2\Delta x}\,, \\ \frac{\partial u}{\partial y}(t^n; x_i, y_j) &\simeq \frac{u^n_{i,j+1} - u^n_{i,j-1}}{2\Delta y}\,, \end{aligned} \tag{2.13}$$

which are also consistent approximations, but are valid only for the case of uniform grids (refer to Figure 2.4).

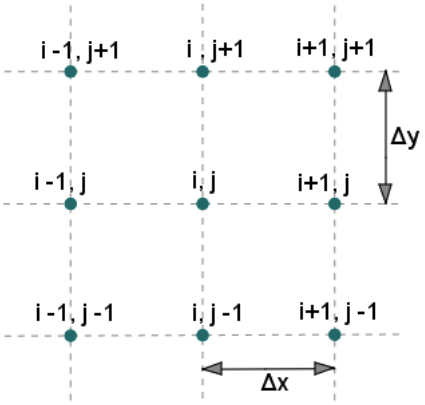


Figure 2.4: finite difference grid for first order derivatives

Remark 5. *An easy calculation shows that the truncation error for* (2.13) *is bigger with respect to the one derived for* (2.12) *from* (2.11)*, and this would be relevant for the diffusive character of the scheme.*

We proceed with the second order derivatives. We detail the computation only for the x variable because it will be the same also for the y variable. We begin to sum the Taylor's expansions for u_{i+1} and u_{i-1} as given in (2.7), with the same abuse of notation observed above,

$$u^n_{i+1,j} + u^n_{i-1,j} = 2\,u^n_{ij} + \Delta x^2 \frac{\partial^2 u^n_{ij}}{\partial x^2} + O(\Delta x^4)\,,$$

from which we obtain

$$\frac{\partial^2 u^n_{ij}}{\partial x^2} = \frac{u^n_{i+1,j} - 2u^n_{ij} + u^n_{i-1,j}}{\Delta x^2} + O(\Delta x^2).$$

Finally, the approximation for the second order derivatives are as follows,

$$\begin{aligned} \frac{\partial^2 u^n_{ij}}{\partial x^2}(t^n; x_i, y_j) &\simeq \frac{u^n_{i+1,j} - 2u^n_{ij} + u^n_{i-1,j}}{\Delta x^2}\,, \\ \frac{\partial^2 u^n_{ij}}{\partial y^2}(t^n; x_i, y_j) &\simeq \frac{u^n_{i,j+1} - 2u^n_{ij} + u^n_{i,j-1}}{\Delta y^2}\,, \end{aligned} \tag{2.14}$$

which are consistent with order of accuracy 2 with respect to both variables.

2.2.2 Mixed derivatives and the θ-scheme for time

In order to discretize the whole diffusion equation (1.1), we have to calculate the approximations for the mixed derivatives. We expand the right-hand side as in (1.5) and we note that mixed derivatives appear only with the coefficients $c_1 = c_2 = c$ (we consider the symmetric case).
First, we choose a composition of centered finite differences (2.13), that gives

$$\begin{aligned} \frac{\partial}{\partial y}\frac{\partial u}{\partial x}(t^n; x_i, y_j) &\simeq \frac{\left(\frac{\partial u}{\partial x}\right)^n_{i,j+1} - \left(\frac{\partial u}{\partial x}\right)^n_{i,j-1}}{2\Delta y} \\ &\simeq \frac{1}{2\Delta y}\left(\frac{u^n_{i+1,j+1} - u^n_{i-1,j+1}}{2\Delta x} - \frac{u^n_{i+1,j-1} - u^n_{i-1,j-1}}{2\Delta x}\right) \\ &= \frac{1}{4\Delta x \Delta y}\left(u^n_{i+1,j+1} - u^n_{i-1,j+1} - u^n_{i+1,j-1} + u^n_{i-1,j-1}\right). \end{aligned} \tag{2.15}$$

A different approach consists in performing finite difference approximations using the staggered grid, so that we have

$$\begin{aligned} \frac{\partial}{\partial y}\frac{\partial u}{\partial x}(t^n; x_i, y_j) &\simeq \frac{\left(\frac{\partial u}{\partial x}\right)^n_{i,j+\frac{1}{2}} - \left(\frac{\partial u}{\partial x}\right)^n_{i,j-\frac{1}{2}}}{\Delta y} \\ &\simeq \frac{1}{\Delta x \Delta y}\left(u^n_{i+\frac{1}{2},j+\frac{1}{2}} - u^n_{i-\frac{1}{2},j+\frac{1}{2}} - u^n_{i+\frac{1}{2},j-\frac{1}{2}} + u^n_{i-\frac{1}{2},j-\frac{1}{2}}\right). \end{aligned} \tag{2.16}$$

Recalling that the numerical solution of first order schemes is usually defined as a piecewise constant function on the spatial grid, for the staggered grid in Figure 2.2 we have the expression (2.3). Then, to derive a coherent formula for the derivatives at the interfacial points in (2.16), we can use the Dirac functions $\delta_{i+\frac{1}{2},j+\frac{1}{2}}(x,y)$ at the mesh interfaces as follows,

$$\begin{aligned}
\frac{\partial u}{\partial x}(t^n; x, y_{j+\frac{1}{2}}) &\simeq \sum_{i+\frac{1}{2}} \frac{1}{\Delta x}\big(u^n_{i+1,j+\frac{1}{2}} - u^n_{i,j+\frac{1}{2}}\big)\delta_{i+\frac{1}{2},j+\frac{1}{2}}(x,y) \\
\implies \quad \frac{\partial u}{\partial x}(t^n; x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}) &\simeq \frac{1}{\Delta x}\big(u^n_{i+1,j+\frac{1}{2}} - u^n_{i,j+\frac{1}{2}}\big), \\
\frac{\partial u}{\partial y}(t^n; x_{i+\frac{1}{2}}, y) &\simeq \sum_{j+\frac{1}{2}} \frac{1}{\Delta y}\big(u^n_{i+\frac{1}{2},j+1} - u^n_{i+\frac{1}{2},j}\big)\delta_{i+\frac{1}{2},j+\frac{1}{2}}(x,y) \\
\implies \quad \frac{\partial u}{\partial y}(t^n; x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}) &\simeq \frac{1}{\Delta y}\big(u^n_{i+\frac{1}{2},j+1} - u^n_{i+\frac{1}{2},j}\big),
\end{aligned}$$

which will be useful in constructing finite volume schemes through integral averages on the grid cells. To pass from interfacial to point derivatives in case of uniform meshes, we simply make an arithmetic average, for example

$$\begin{aligned}
\left(\frac{\partial u}{\partial x}\right)^n_{i,j+\frac{1}{2}} &\simeq \frac{1}{2}\frac{\partial u}{\partial x}(t^n; x_{i-\frac{1}{2}}, y_{j+\frac{1}{2}}) + \frac{1}{2}\frac{\partial u}{\partial x}(t^n; x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}) \\
&\simeq \frac{1}{2\Delta x}\big(u^n_{i+1,j+\frac{1}{2}} - u^n_{i-1,j+\frac{1}{2}}\big),
\end{aligned}$$

so that from (2.16) we will finally recover the centered discretization (2.15). We remark that the same expressions (2.15) and (2.16) are obtained starting from $\frac{\partial}{\partial x\partial y} = \frac{\partial}{\partial x}\frac{\partial u}{\partial y}$, so the requirement of symmetric diffusion tensor is coherent also within the discrete framework (at least for constant coefficients).

As an example for introducing a general class of discrete time-operators, we apply the above formulas to derive a discretization for the one-dimensional homogeneous *heat equation*, with the diffusion coefficient $a > 0$ and constant.

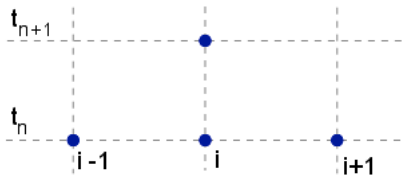


Figure 2.5: stencil of the explicit forward Euler scheme

Three schemes can be taken into account : the *explicit solver* leads to

$$\frac{u^{n+1}_i - u^n_i}{\Delta t} = a\frac{u^n_{i+1} - 2u^n_i + u^n_{i-1}}{\Delta x^2}, \tag{2.17}$$

which is the well-known *forward Euler* method, whose *stencil* is represented in Figure 2.5; the *implicit solver* leads to

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = a \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{\Delta x^2}, \tag{2.18}$$

where the time-discretization comes from a Taylor's expansion at point t^n instead of t^{n+1}, which is known as the *backward Euler* method, with the *stencil* represented in Figure 2.6; finally, the *semi-implicit solver* leads to

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = a(1-\theta) \frac{u_{i+1}^{n} - 2u_i^{n} + u_{i-1}^{n}}{\Delta x^2} + a\,\theta \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{\Delta x^2}, \tag{2.19}$$

where $\theta \in [0, 1]$, whose *stencil* is represented in Figure 2.7, and this type of operator is generally called a θ-method. If $\theta = 1/2$ we have the well-known *Crank-Nicolson scheme*.

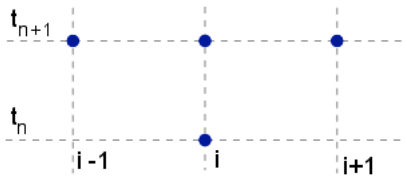


Figure 2.6: stencil of the implicit backward Euler scheme

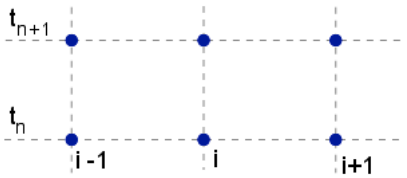


Figure 2.7: stencil of the semi-implicit θ-method

We will consider these three methods in combination with space-discretizations in the next chapters, and we will examine their properties in details depending on the specific applications.

2.3 The Finite Volume method

The finite volume schemes are an alternative approach for the approximation of PDEs [25],[30]. Similarly to the finite difference method, the values of the numerical solution are calculated at discrete locations on a meshed geometry, and *finite volume* refers to the small volume surrounding each

node on the grids (see Figure 2.2). In the finite volume method, volume integrals of partial differential operators which contain a divergence term are converted into surface integrals, through the extensive use of the *divergence theorem*. These terms are then evaluated as *fluxes* at the surfaces of each finite volume [21].

To give an example of derivation of these schemes, we focus on a simple case of heterogeneous one-dimensional diffusion equation

$$u_t - \big(a(x)u_x\big)_x = 0\,, \tag{2.20}$$

for $t > 0$ and $x \in [x_1, x_2] \subset \mathbb{R}$. According to the theory of *conservation laws*, we denote by $F(t,x) = a(x)u_x$ the *parabolic flux* (eventually depending also explicitly on time), so the above equation can be rewritten in conservative form as

$$u_t - \big(F(t,x)\big)_x = 0\,. \tag{2.21}$$

Typically, the conservation law (2.21) concerns the (local) dynamics of the *mass* of a physical quantity, that is related to the density/concentration by

$$m(t) = \int_{x_1}^{x_2} u(t,x)\,dx\,,$$

which provides a connection between global/macroscopical and local/microscopical quantities. We write the *integral form* of the equation (2.20) as

$$\frac{d}{dt}\int_{x_1}^{x_2} u(t,x)\,dx = \int_{x_1}^{x_2} \big(a(x)u_x\big)_x\,dx = F(t,x_2) - F(t,x_1)\,, \tag{2.22}$$

and this actually holds for any arbitrary interval.
The above equality states that the variation of mass over time inside the region $[x_1, x_2] \subset \mathbb{R}$ is equal to the difference of the flux at the initial and final points (the boundary of a one-dimensional domain). This represents a conservation law for the mass because the mass is conserved if the value of the flux at the boundary points is constant, namely the entering flux equals the outgoing flux. That property is satisfied for any physical situation which is not creating or destroying mass, like the diffusion processes considered in this report, and the only way to locally variate the mass is through the behavior of the flux (we refer to [6] for a remarkable presentation of that subject).

We aim at deriving a finite volume scheme for the equation (2.20).
We adapt to the one-dimensional framework the definition (2.4) and the discretization of the spatial domain as in Figure 2.2, then we compute the cell-averages of the equation as follows,

$$\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \left\{ \frac{1}{\Delta x}\int_{C_i} \Big[u_t - \big(a(x)u_x\big)_x\Big]dx \right\} dt = 0\,. \tag{2.23}$$

We assume that u has all the regularity properties necessary to exchange derivatives with integrals, and for the first term we obtain

$$\frac{1}{\Delta t}\left[\frac{1}{\Delta x}\int_{C_i} u(t^{n+1},x)\,dx - \frac{1}{\Delta x}\int_{C_i} u(t^n,x)\,dx\right] = \frac{\bar{u}_i^{n+1}-\bar{u}_i^n}{\Delta t}, \qquad (2.24)$$

where the *cell-averages* of the exact solution are defined as

$$\bar{u}_i^n = \frac{1}{\Delta x}\int_{C_i} u(t^n,x)\,dx = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(t^n,x)\,dx\,,$$

recalling that $\bar{u}_i^n \simeq u(t^n,x_i)$ to the order $O(\Delta x^2)$ for the choice to locate the nodes at the center of the finite volumes. We recover the integral form (2.22) with $C_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ for the second term in (2.23) and we have

$$\begin{aligned}&\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \frac{1}{\Delta x}\left[F(t,x_{i+\frac{1}{2}}) - F(t,x_{i-\frac{1}{2}})\right]dt \\ &= \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \frac{1}{\Delta x}\left[a(x_{i+\frac{1}{2}})\,u_x(t,x_{i+\frac{1}{2}}) - a(x_{i-\frac{1}{2}})\,u_x(t,x_{i-\frac{1}{2}})\right]dt\,.\end{aligned} \qquad (2.25)$$

Therefore, the time variation (2.24) is given by the difference of flux values at the interfacial points, which are the boundary of a grid cell, and this has a physical sense for the conservation law (2.22).
We still need to approximate the time-average in (2.25). In general, for a function $g \in C^1(\mathbb{R})$, we can write

$$\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} g(t)\,dt = g(t^n) + R_1 = g(t^{n+1}) + R_2\,,$$

with

$$|R_k| \le \max_{t^n \le t \le t^{n+1}} |g'(t)| \cdot \Delta t\,, \quad k = 1,2\,,$$

so that we can choose an explicit scheme (with t^n) or an implicit one (with t^{n+1}) or also a combination of both. For example, we substitute t^n in (2.25), committing an error of order Δt, and we consider the discrete fluxes

$$\frac{1}{\Delta x}\left[a(x_{i+\frac{1}{2}})\,u_x(t^n,x_{i+\frac{1}{2}}) - a(x_{i-\frac{1}{2}})\,u_x(t^n,x_{i-\frac{1}{2}})\right].$$

Finally, we have to approximate the diffusion coefficient a and u_x at the cell-interfaces by using the values at the nodes, because only these last data are available at the discrete level. As usual, we set $a_i = \frac{1}{\Delta x}\int_{C_i} a(x)\,dx \simeq a(x_i)$ and we can exploit the *arithmetic averages*, namely

$$a(x_{i+\frac{1}{2}}) \simeq \frac{a_i + a_{i+1}}{2}\,, \qquad a(x_{i-\frac{1}{2}}) \simeq \frac{a_{i-1}+a_i}{2}\,, \qquad (2.26)$$

while for the first order derivatives at the interfaces we have

$$u_x(t^n, x_{i+\frac{1}{2}}) \simeq \frac{u_{i+1}^n - u_i^n}{\Delta x}\,, \qquad u_x(t^n, x_{i-\frac{1}{2}}) \simeq \frac{u_i^n - u_{i-1}^n}{\Delta x}\,. \tag{2.27}$$

So, the fully-discrete scheme for the equation (2.20) reads

$$\begin{aligned}
\frac{u_i^{n+1} - u_i^n}{\Delta t} &= \frac{1}{\Delta x}\left(\frac{a_i + a_{i+1}}{2} \cdot \frac{u_{i+1}^n - u_i^n}{\Delta x} - \frac{a_{i-1} + a_i}{2} \cdot \frac{u_i^n - u_{i-1}^n}{\Delta x}\right) \\
&= \frac{a_i}{2} \cdot \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} + \frac{a_{i+1}}{2\Delta x} \cdot \frac{u_{i+1}^n - u_i^n}{\Delta x} - \frac{a_{i-1}}{2\Delta x} \cdot \frac{u_i^n - u_{i-1}^n}{\Delta x} \\
&= u_{i+1}^n \left(\frac{a_i + a_{i+1}}{2\Delta x^2}\right) + u_i^n \left(\frac{-a_{i+1} - 2a_i - a_{i-1}}{2\Delta x^2}\right) + u_{i-1}^n \left(\frac{a_{i-1} + a_i}{2\Delta x^2}\right),
\end{aligned}$$

or rather

$$\begin{aligned}
u_i^{n+1} = {} & \frac{\Delta t}{2\Delta x^2}(a_i + a_{i+1})\, u_{i+1}^n + \big(1 - \frac{\Delta t}{2\Delta x^2}(a_{i+1} + 2a_i + a_{i-1})\big) u_i^n \\
& + \frac{\Delta t}{2\Delta x^2}(a_{i-1} + a_i)\, u_{i-1}^n\,.
\end{aligned} \tag{2.28}$$

One easily recognizes in the second line above the two terms of the heterogeneous diffusion equation (2.20), after rearranging the terms as follows,

$$\left(\frac{a_i}{2} + \frac{a_{i-1}}{2}\right) \cdot \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} + \frac{a_{i+1} - a_{i-1}}{2\Delta x} \cdot \frac{u_{i+1}^n - u_i^n}{\Delta x}\,,$$

and, moreover, we recover the finite difference explicit scheme (2.17) for the one-dimensional heat equation if a is constant.

Remark 6. *In the case of nonuniform spatial grids, the approximations* (2.26) *and* (2.27) *must be computed taking into account the size of the cells explicitly into the formulation, thus introducing a lack of symmetry in the schemes. We will address that delicate issue in a forthcoming work.*

Despite, in some easy cases, finite difference and finite volume schemes are formally the same, the interpretation in the spirit of the finite volume method is very useful to catch the physical sense of the parabolic equations, as clearly expressed through the weak/integral formulation (2.22).

2.4 Finite difference schemes for two-dimensional heterogeneous equations

With the tools introduced in Section 2.2, we analyze three different schemes, following the presentation in [26].

2.4.1 The Chain Rule method

The simplest way of discretizing second order operators like (1.1) is by applying the *product rule* to expand the derivatives and, then, use centered finite difference schemes (2.13) and (2.14). From the second line of (1.5), it holds

$$\begin{aligned}\big(a\,u_x\big)_x &\simeq \frac{a_{i+1,j}-a_{i-1,j}}{2\Delta x}\cdot\frac{u_{i+1,j}-u_{i-1,j}}{2\Delta x}+a_{ij}\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{\Delta x^2}\\ &= u_{i+1,j}\Big(\frac{a_{i+1,j}-a_{i-1,j}}{4\Delta x^2}+\frac{a_{ij}}{\Delta x^2}\Big)+u_{ij}\Big(-2\frac{a_{ij}}{\Delta x^2}\Big)\\ &\quad+u_{i-1,j}\Big(-\frac{a_{i+1,j}-a_{i-1,j}}{4\Delta x^2}+\frac{a_{ij}}{\Delta x^2}\Big).\end{aligned} \tag{2.29}$$

where we have omitted the time-dependence, for simplicity, and also

$$\begin{aligned}\big(c\,u_y\big)_x &\simeq \frac{c_{i+1,j}-c_{i-1,j}}{2\Delta x}\cdot\frac{u_{i,j+1}-u_{i,j-1}}{2\Delta y}\\ &\quad+\frac{c_{ij}}{2\Delta x}\left(\frac{u^n_{i+1,j+1}-u^n_{i+1,j-1}}{2\Delta y}-\frac{u^n_{i-1,j+1}-u^n_{i-1,j-1}}{2\Delta y}\right)\end{aligned} \tag{2.30}$$

where we used the same approach as in (2.15). Analogous calculations hold for the other terms $\big(c\,u_x\big)_y$ and $\big(b\,u_y\big)_y$ in (1.5).

To make the notation more compact, we rewrite the stencil of the scheme as reported in Table 2.1, where the central entry is the coefficient corresponding to the term u_{ij}, the right one corresponds to the coefficient of $u_{i+1,j}$, the left one to the coefficient of $u_{i-1,j}$, and similarly for all the others.

$-\frac{c_{ij}}{2\Delta x\Delta y}$	$\frac{b_{i,j+1}-b_{i,j-1}}{4\Delta y^2}+\frac{b_{ij}}{\Delta y^2}$ $+\frac{c_{i+1,j}-c_{i-1,j}}{4\Delta x\Delta y}$	$\frac{c_{ij}}{2\Delta x\Delta y}$
$-\frac{a_{i+1,j}-a_{i-1,j}}{4\Delta x^2}+\frac{a_{ij}}{\Delta x^2}$ $-\frac{c_{i,j+1}-c_{i,j-1}}{4\Delta x\Delta y}$	$-2\frac{a_{ij}}{\Delta x^2}-2\frac{b_{ij}}{\Delta y^2}$	$\frac{a_{i+1,j}-a_{i-1,j}}{4\Delta x^2}+\frac{a_{ij}}{\Delta x^2}$ $+\frac{c_{i,j+1}-c_{i,j-1}}{4\Delta x\Delta y}$
$\frac{c_{ij}}{2\Delta x\Delta y}$	$-\frac{b_{i,j+1}-b_{i,j-1}}{4\Delta y^2}+\frac{b_{ij}}{\Delta y^2}$ $-\frac{c_{i+1,j}-c_{i-1,j}}{4\Delta x\Delta y}$	$-\frac{c_{ij}}{2\Delta x\Delta y}$

Table 2.1: stencil of the two-dimensional Chain Rule scheme

We recall that the above stencil refers only to the space-discretization. We will see later on that this method is inappropriate for the validity of the discrete maximum principle, and then useless for practical applications.

2.4.2 The Standard Discretization method

Another possible approach is based mainly on the approximation on staggered grids (2.12), so the double derivatives (2.29) are treated as follows,

$$\begin{aligned}
\big(a\,u_x\big)_x &\simeq \frac{(a\,u_x)_{i+\frac{1}{2},j} - (a\,u_x)_{i-\frac{1}{2},j}}{\Delta x} \\
&\simeq \frac{1}{\Delta x}\left(a_{i+\frac{1}{2},j}\frac{u_{i+1,j}-u_{ij}}{\Delta x} - a_{i-\frac{1}{2},j}\frac{u_{ij}-u_{i-1,j}}{\Delta x}\right) \\
&\simeq \frac{a_{ij}+a_{i+1,j}}{2}\cdot\frac{u_{i+1,j}-u_{ij}}{\Delta x^2} + \frac{a_{i-1,j}+a_{ij}}{2}\cdot\frac{u_{i-1,j}-u_{ij}}{\Delta x^2} \qquad (2.31)\\
&= u_{i+1,j}\frac{a_{ij}+a_{i+1,j}}{2\Delta x^2} + u_{ij}\frac{-a_{i+1,j}-2a_{ij}-a_{i-1,j}}{2\Delta x^2} \\
&\quad + u_{i-1,j}\frac{a_{i-1,j}+a_{ij}}{2\Delta x^2}\,,
\end{aligned}$$

where the unknown interfacial values $a_{i+\frac{1}{2},j}$ and $a_{i-\frac{1}{2},j}$ are calculated by the arithmetic averages, and we recover precisely the finite volume scheme (2.28). The formula (2.30) for the mixed derivatives is modified according to (2.16), and we have several choices for interpolating the interfacial values : by using standard arithmetic averages, for example, we end up with some weighted modification of (2.15) which is also more compact than (2.30), namely

$$\begin{aligned}
\big(c\,u_y\big)_x &\simeq \frac{(c\,u_y)_{i+\frac{1}{2},j} - (c\,u_y)_{i-\frac{1}{2},j}}{\Delta x} \\
&\simeq \frac{(c\,u_y)_{ij} + (c\,u_y)_{i+1,j} - (c\,u_y)_{i-1,j} - (c\,u_y)_{ij}}{2\Delta x} = \frac{(c\,u_y)_{i+1,j} - (c\,u_y)_{i-1,j}}{2\Delta x} \\
&\simeq \frac{1}{2\Delta x}\left(c_{i+1,j}\frac{u_{i+1,j+\frac{1}{2}}-u_{i+1,j-\frac{1}{2}}}{\Delta y} - c_{i-1,j}\frac{u_{i-1,j+\frac{1}{2}}-u_{i-1,j-\frac{1}{2}}}{\Delta y}\right) \\
&\simeq \frac{1}{2\Delta x}\left(c_{i+1,j}\frac{u_{i+1,j+1}-u_{i+1,j-1}}{2\Delta y} - c_{i-1,j}\frac{u_{i-1,j+1}-u_{i-1,j-1}}{2\Delta y}\right),
\end{aligned} \qquad (2.32)$$

therefore, in this case, the extra-diagonal coefficient c appears only with the terms at the vertices of Table 2.2, where we have reported the whole stencil.

$-\frac{c_{i-1,j}+c_{i,j+1}}{4\Delta x\Delta y}$	$\frac{b_{ij}+b_{i,j+1}}{2\Delta y^2}$	$\frac{c_{i+1,j}+c_{i,j+1}}{4\Delta x\Delta y}$
$\frac{a_{i-1,j}+a_{ij}}{2\Delta x^2}$	$-\frac{a_{i+1,j}+2a_{ij}+a_{i-1,j}}{2\Delta x^2}$ $-\frac{b_{i,j+1}+2b_{ij}+b_{i,j-1}}{2\Delta y^2}$	$\frac{a_{ij}+a_{i+1,j}}{2\Delta x^2}$
$\frac{c_{i-1,j}+c_{i,j-1}}{4\Delta x\Delta y}$	$\frac{b_{i,j-1}+b_{ij}}{2\Delta y^2}$	$-\frac{c_{i+1,j}+c_{i,j-1}}{4\Delta x\Delta y}$

Table 2.2: stencil of the two-dimensional Standard Discretization

We remark that, in the case of homogeneous diffusion tensors, with constant diffusion coefficients, the scheme represented in Table 2.2 is the same

as the one in Table 2.1. Moreover, also according to [7], the above scheme is *conditionally stable* but does not satisfy the *discrete maximum principle*. In Section 4.3, we will derive another proof of stability, giving a less restrictive condition for $c = 0$, and under which also the discrete maximum principle holds. Nevertheless, in the general case of $c \neq 0$ or c is a space-dependent function, this method fails to satisfy the discrete maximum principle and for this reason we must resort to other schemes.

We conclude this section by introducing another approach for the mixed derivatives, which makes use of the approximation (2.16) on staggered grids,

$$\begin{aligned}
\big(c\,u_y\big)_x &\simeq \frac{(c\,u_y)_{i+\frac{1}{2},j} - (c\,u_y)_{i-\frac{1}{2},j}}{\Delta x} \\
&\simeq \frac{c_{i+\frac{1}{2},j}}{\Delta x \Delta y}\big(u_{i+\frac{1}{2},j+\frac{1}{2}} - u_{i+\frac{1}{2},j-\frac{1}{2}}\big) - \frac{c_{i-\frac{1}{2},j}}{\Delta x \Delta y}\big(u_{i-\frac{1}{2},j+\frac{1}{2}} - u_{i-\frac{1}{2},j-\frac{1}{2}}\big), \\
\big(c\,u_x\big)_y &\simeq \frac{(c\,u_x)_{i,j+\frac{1}{2}} - (c\,u_x)_{i,j-\frac{1}{2}}}{\Delta y} \\
&\simeq \frac{c_{i,j+\frac{1}{2}}}{\Delta x \Delta y}\big(u_{i+\frac{1}{2},j+\frac{1}{2}} - u_{i-\frac{1}{2},j+\frac{1}{2}}\big) - \frac{c_{i,j-\frac{1}{2}}}{\Delta x \Delta y}\big(u_{i+\frac{1}{2},j-\frac{1}{2}} - u_{i-\frac{1}{2},j-\frac{1}{2}}\big),
\end{aligned}$$

which is recovered through a finite volume approach like in Section 2.3, and coincide with the scheme (2.31)-(2.32) for c constant if using arithmetic averages to pass from interfacial values to grid points. But this is actually a more rigorous way of derivation, because we use the same approach for the second order derivatives and the mixed ones.

2.4.3 The Nonnegative method

As we will discuss later, the main property that efficient numerical methods for diffusion equations have to satisfy is the non-negativity of extra-diagonal stencil entries, in order to guarantee the *discrete maximum principle*. If we look at the two previous tables (2.1) and (2.2), we cannot be sure that these coefficients are nonnegative. The existence of a third method that solves this problem is demonstrated in [38], where applications to image processing are considered, leading to the so-called *nonnegative method* as its name suggests.

The proof in [38] is constructive : the strategy is based on calculating numerical derivatives in new directions, in addition to the (x, y)-directions, and the two-dimensional framework uses the diagonal directions of the spacial grid, as shown in Figure 2.8. Therefore, the diffusion terms are modified including an angle $\beta = \arctan(\frac{\Delta y}{\Delta x})$ and the mixed derivatives are replaced by the new directional ones. To obtain a nonnegative stencil, it is enough to find the conditions under which the stencil weights of the new directions are nonnegative. We reproduce the complete derivation of the two-dimensional numerical scheme as given in [38], through the following theorem.

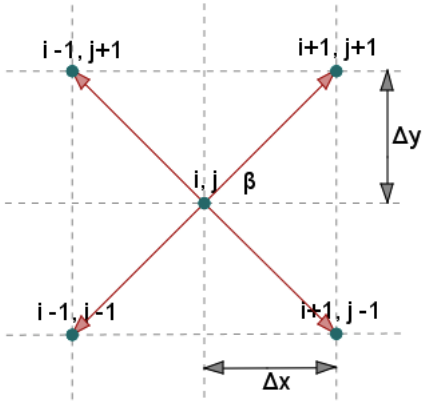


Figure 2.8: diagonal directions for numerical mixed derivatives

Theorem 7. *Let* $A = \begin{bmatrix} a & c \\ c & b \end{bmatrix} \in \mathbb{R}^{2\times 2}$ *be a symmetric positive definite matrix, with bounded spectral condition number* $\kappa = max|\lambda_l|/min|\lambda_l|$*, for* λ_l*,* $l = 1, 2$ *eigenvalues. Then, there exists some* $m(\kappa) \in \mathbb{N}$ *such that the parabolic differential operator* $\nabla \cdot (A \cdot \nabla u)$ *reveals a second-order nonnegative finite difference discretization with a* $(2m+1) \times (2m+1)$*-stencil.*

Proof. We start by fixing arbitrarily $m \in \mathbb{N}$ and consider the corresponding $(2m+1)\times(2m+1)$-stencil. We call *boundary pixels* the grid points at the boundary of the stencil, which induce $4m$ principal orientations $\beta_i \in (-\frac{\pi}{2}, \frac{\pi}{2}]$, $i = -2m+1, \ldots, 2m$, according to

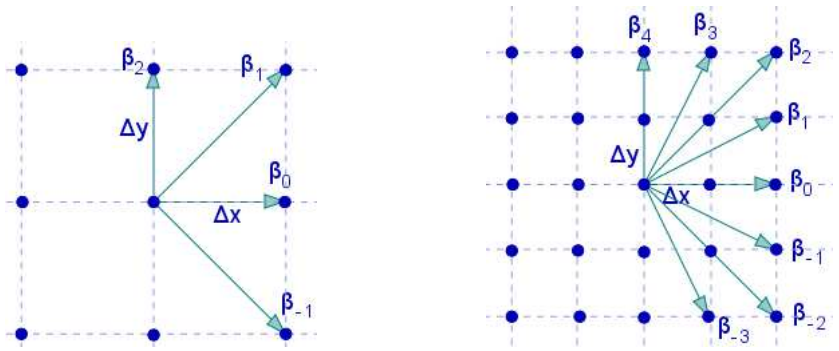
$$\beta_i = \begin{cases} \arctan\left(\frac{i\Delta y}{m\Delta x}\right) & |i| \leq m \\ \operatorname{arccot}\left(\frac{(2m-i)\Delta x}{m\Delta y}\right) & m < i \leq 2m \\ \operatorname{arccot}\left(\frac{(i-2m)\Delta x}{m\Delta y}\right) & -2m+1 \leq i < -m \end{cases} \tag{2.33}$$

We refer to Figure 2.9, for example. Let $J_m = \{1, \ldots, 2m-1\}$ and define a partition of $(-\frac{\pi}{2}, \frac{\pi}{2}]$ into $4m-2$ subintervals I_i, $|i| \in J_m$ as

$$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right] = \bigcup_{i=-2m+1}^{-1} I_i \cup \bigcup_{i=1}^{2m-1} I_i = \bigcup_{i=-2m+1}^{-1} (\theta_i, \theta_{i+1}] \cup \bigcup_{i=1}^{2m-1} (\theta_{i-1}, \theta_i],$$

where

$$\theta_i = \begin{cases} 0 & i = 0 \\ \frac{1}{2}\arctan\left(\frac{2}{\cot\beta_i - \tan\beta_{i+1}}\right) & i \in \{1, \ldots, 2m-2\} \text{ and } \beta_i + \beta_{i+1} < \frac{\pi}{2} \\ \frac{\pi}{4} & i \in \{1, \ldots, 2m-2\} \text{ and } \beta_i + \beta_{i+1} = \frac{\pi}{2} \\ \frac{\pi}{2} + \frac{1}{2}\arctan\left(\frac{2}{\cot\beta_i - \tan\beta_{i+1}}\right) & i \in \{1, \ldots, 2m-2\} \text{ and } \beta_i + \beta_{i+1} > \frac{\pi}{2} \\ \frac{\pi}{2} & i = 2m-1 \end{cases}$$



(a) $m = 1$ (b) $m = 2$

Figure 2.9: examples of two-dimensional stencils

By definition, $\theta_i = -\theta_{-i}$ for $i \in \{-2m+1, \ldots, -1\}$ and $\beta_i \in I_i$ for $|i| \in J_m$.

Let us focus on the eigenvalues of A : we know that they are real and positive, so let $\lambda_1 \geq \lambda_2 > 0$, with the corresponding normal eigenvectors $(\cos\psi, \sin\psi)^T$ and $(-\sin\psi, \cos\psi)^T$, where $\psi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$.

We want to show that there exists a stencil direction with respect to the first eigenvector, based on β_k with $|k| \in J_m$, such that the splitting

$$\nabla \cdot (A \cdot \nabla u) = \partial_{e_{\beta_0}}\big(\alpha_0\, \partial_{e_{\beta_0}} u\big) + \partial_{e_{\beta_k}}\big(\alpha_k\, \partial_{e_{\beta_k}} u\big) + \partial_{e_{\beta_{2m}}}\big(\alpha_{2m}\, \partial_{e_{\beta_{2m}}} u\big)\,, \quad (2.34)$$

where $e_{\beta_i} = (\cos\psi, \sin\psi)^T$, reveals nonnegative *directional diffusivity* coefficients α_0, α_k and α_{2m} along the stencil orientations β_0, β_k and β_{2m}.
To prove the statement (2.34), we demonstrate the following properties :

(a) with $\psi \in I_k$ and $A = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$, a nonnegative spitting is possible if

$$\min\big(a - c\,\mathrm{cotan}\,\beta_k\,, b - c\,\tan\beta_k\big) \geq 0\,; \quad (2.35)$$

(b) the previous inequality (2.35) is satisfied for

$$\frac{\lambda_1}{\lambda_2} \leq \min\big(\cot(\rho_k - \beta_k)\ \tan\rho_k\,, \cot(\beta_k - \eta_k)\ \cot\eta_k\big) = \kappa_{k,m} \quad (2.36)$$

with

$$\rho_k = \begin{cases} \theta_k & |k| \in \{1, \ldots, 2m-2\} \\ \frac{1}{2}(\theta_k + \beta_k) & |k| = 2m-1 \end{cases}$$

and

$$\eta_k = \begin{cases} \frac{1}{2}\beta_k & |k| = 1 \\ \theta_{k-1} & |k| \in \{1, \ldots, 2m-2\} \end{cases}$$

(c) $\lim_{m\to\infty}\big(\min_{|i|\in J_m} \kappa_{i,m}\big) = \infty\,.$

Once these assertions are proven, we will find an explicit formula for the coefficient α_k through which a nonnegative second-order discretization arises. We assume that the point (c) is proven (refer to [38] for details), because we want to concentrate on the two others, for which we have the tools to construct nonnegative discretizations.

(a) Let $\varphi_0 = 0$, $\varphi_1 = \beta_k$ where $\psi \in I_k$, and $\varphi_2 = \frac{\pi}{2}$. Moreover, let $\gamma_0 = \alpha_0$, $\gamma_1 = \alpha_k$ and $\gamma_{2m} = \alpha_{2m}$. With this notation we rewrite the equality (2.34) as a type of *change of representation*, that is

$$\begin{aligned}
\nabla \cdot \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} \cdot \nabla u \right) &= \sum_{i=0}^{2} \frac{\partial}{\partial e_{\varphi_i}} \Big(\gamma_i \frac{\partial u}{\partial e_{\varphi_i}} \Big) \\
&= \sum_{i=0}^{2} \left(\frac{\partial}{\partial x} \frac{\partial x}{\partial e_{\varphi_i}} + \frac{\partial}{\partial y} \frac{\partial y}{\partial e_{\varphi_i}} \right) \gamma_i \left(\frac{\partial}{\partial x} \frac{\partial x}{\partial e_{\varphi_i}} + \frac{\partial}{\partial y} \frac{\partial y}{\partial e_{\varphi_i}} \right) u \\
&= \frac{\partial}{\partial x} \sum_{i=0}^{2} \cos\varphi_i \big[\gamma_i (u_x \cos\varphi_i + u_y \sin\varphi_i) \big] \\
&\quad + \frac{\partial}{\partial y} \sum_{i=0}^{2} \sin\varphi_i \big[\gamma_i (u_x \cos\varphi_i + u_y \sin\varphi_i) \big] \\
&= \nabla \cdot \left(\begin{bmatrix} \sum_{i=0}^{2} \gamma_i \cos^2\varphi_i & \sum_{i=0}^{2} \gamma_i \sin\varphi_i \cos\varphi_i \\ \sum_{i=0}^{2} \gamma_i \sin\varphi_i \cos\varphi_i & \sum_{i=0}^{2} \gamma_i \sin^2\varphi_i \end{bmatrix} \cdot \nabla u \right)
\end{aligned}$$

Comparing the coefficients and developing the sums, we obtain

$$\begin{aligned}
a &= \gamma_0 \cos^2\varphi_0 + \gamma_1 \cos^2\varphi_1 = \gamma_0 + \gamma_1 \cos^2\beta_k \,, \\
c &= \gamma_1 \sin\varphi_1 \, \cos\varphi_1 = \gamma_1 \sin\beta_k \, \cos\beta_k \,, \\
b &= \gamma_1 \sin^2\varphi_1 + \gamma_2 \sin^2\varphi_2 = \gamma_1 \sin^2\beta_k + \gamma_2 \,.
\end{aligned}$$

We can solve directly the second equation above as $\gamma_1 = \frac{c}{\sin\beta_k \cos\beta_k}$, and then we can substitute into the two others. Recalling that $\tan\beta_k = \frac{\sin\beta_k}{\cos\beta_k}$ and $\cot\beta_k = \frac{\cos\beta_k}{\sin\beta_k}$, we easily deduce

$$\gamma_0 \, = \, a - c \, \cot\beta_k \,, \quad \gamma_1 \, = \, \frac{c}{\sin\beta_k \, \cos\beta_k} \,, \quad \gamma_2 \, = \, b - c \, \tan\beta_k \,. \tag{2.37}$$

We need to study the sign of these three coefficients, aiming at having them always nonnegative because we are searching a discretization with this property. By definition of eigenvalue/eigenvector, it holds

$$A \begin{bmatrix} \cos\psi \\ \sin\psi \end{bmatrix} = \lambda_1 \begin{bmatrix} \cos\psi \\ \sin\psi \end{bmatrix}, \qquad A \begin{bmatrix} -\sin\psi \\ \cos\psi \end{bmatrix} = \lambda_2 \begin{bmatrix} -\sin\psi \\ \cos\psi \end{bmatrix},$$

therefore, developing the products, we have

$$\begin{aligned}
a \, \cos\psi + c \, \sin\psi = \lambda_1 \cos\psi \,, \quad c \, \cos\psi + b \, \sin\psi = \lambda_1 \sin\psi \,, \\
-a \, \sin\psi + c \, \cos\psi = -\lambda_2 \sin\psi \,, \quad -c \, \sin\psi + b \, \cos\psi = \lambda_2 \cos\psi \,.
\end{aligned}$$

Now, multiplying the first by $\sin\psi$ and the third by $\cos\psi$, and summing these two equivalences, we obtain

$$c = (\lambda_1 - \lambda_2)\,\cos\psi\,\sin\psi\,,$$

so γ_1 in (2.37) is always nonnegative, since $\lambda_1 - \lambda_2 \geq 0$ by hypothesis and $\psi, \beta_k \in I_k$ belong to the same quadrant. In order to verify the non-negativity of γ_0 and γ_2, we must impose that

$$\min\left(a - c\,\cot\beta_k\,, b - c\,\tan\beta_k\right) \geq 0\,,$$

which is the inequality (2.35) we wanted to prove.

(b) To prove the second assertion, we suppose that the inequality (2.35) is satisfied under the assumption (2.36).
Let $\frac{\lambda_1}{\lambda_2} \leq \kappa_{k,m}$ and consider the case $0 < \beta_k < \frac{\pi}{2}$. Let us define

$$\begin{aligned} B(\varphi) &= \cos^2\varphi - \sin\varphi\,\cos\varphi\,\cot\beta_k\,, \\ C(\varphi) &= \sin^2\varphi + \sin\varphi\,\cos\varphi\,\cot\beta_k\,, \end{aligned}$$

with the properties

$$\begin{aligned} &B(\varphi) < 0 \quad \text{for } \varphi \in (\beta_k, \frac{\pi}{2})\,, \qquad B(\varphi) \geq 0 \quad \text{for } \varphi \in [-\frac{\pi}{2}, \beta_k]\,, \\ &C(\varphi) < 0 \quad \text{for } \varphi \in [-\frac{\pi}{2}, 0)\,, \qquad C(\varphi) \geq 0 \quad \text{for } \varphi \in [0, \frac{\pi}{2}]\,. \end{aligned} \tag{2.38}$$

By the definition of $\kappa_{k,m}$, we have that

$$\frac{\lambda_1}{\lambda_2} \leq \cot(\rho_k - \beta_k)\,\tan\rho_k = -\frac{C(\rho_k)}{B(\rho_k)} = \min_{\varphi\in(\beta_k,\theta_k)}\Big(-\frac{C(\varphi)}{B(\varphi)}\Big) \leq -\frac{C(\varphi)}{B(\varphi)}\,.$$

In particular, the equality $\cot(\rho_k - \beta_k)\,\tan\rho_k = -\frac{C(\rho_k)}{B(\rho_k)}$ comes from the following trigonometrical formula,

$$\cot(a-b) = \frac{\cot(a)\,\cot(b) + 1}{\cot(b) - \cot(a)}$$

when applied to the case above, namely

$$\begin{aligned} \cot(\rho_k - \beta_k)\,\tan\rho_k &= \frac{\cot\rho_k\,\cot\beta_k + 1}{\cot\beta_k - \cot\rho_k}\cdot\frac{\sin\rho_k}{\cos\rho_k} \\ &= \frac{1 + \frac{\cos\rho_k}{\sin\rho_k}\cot\beta_k}{\cot\beta_k - \frac{\cos\rho_k}{\sin\rho_k}}\cdot\frac{\sin\rho_k}{\cos\rho_k} \\ &= \frac{\sin\rho_k(\sin\rho_k + \cos\rho_k\,\cot\beta_k)}{\cos\rho_k(-\cos\rho_k + \sin\rho_k)} \\ &= -\frac{\sin^2\rho_k + \sin\rho_k\,\cos\rho_k\,\cot\beta_k}{\cos^2\rho_k - \sin\rho_k\,\cos\rho_k\,\cot\beta_k} = -\frac{C(\rho_k)}{B(\rho_k)}\,. \end{aligned}$$

Due to the properties in (2.38), we have that $\forall\, \varphi \in (\beta_k, \theta_k)$ it holds

$$\frac{\lambda_1}{\lambda_2} \leq -\frac{C(\varphi)}{B(\varphi)} \;\text{ and }\; B(\varphi) < 0\,, \quad \frac{\lambda_1}{\lambda_2} B(\varphi) \geq -C(\varphi) \;\text{ and }\; C(\varphi) > 0\,,$$

and also $\lambda_1 B(\varphi) + \lambda_2 C(\varphi) \geq 0$, and by the continuity of $B(\varphi)$ and $C(\varphi)$ we may extend this relation to the entire interval $I_k = (\theta_{k-1}, \theta_k]$. In particular, since $\psi \in I_k$, we have

$$0 \leq \lambda_1 B(\varphi) + \lambda_2 C(\varphi) = (\lambda_1 \cos^2\psi + \lambda_2 \sin^2\psi) - (\lambda_1 - \lambda_2)\, \sin\psi\, \cos\psi\, \cot\beta_k\,.$$

By the representation through eigenvalues and eigenvectors,

$$\begin{aligned} \begin{bmatrix} a & c \\ c & b \end{bmatrix} &= \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 \cos^2\psi + \lambda_2 \sin^2\psi & (\lambda_1 - \lambda_2)\sin\psi\, \cos\psi \\ (\lambda_1 - \lambda_2)\sin\psi\, \cos\psi & \lambda_1 \cos^2\psi + \lambda_2 \sin^2\psi \end{bmatrix}, \end{aligned}$$

we obtain that

$$a - c\, \cot\beta_k = (\lambda_1 \cos^2\psi + \lambda_2 \sin^2\psi) - (\lambda_1 - \lambda_2)\sin\psi\, \cos\psi\, \cot\beta_k \geq 0\,,$$

which is one of the two terms of inequality (2.35) we wanted to prove. For the second, with the same technique, we can conclude that

$$\frac{\lambda_1}{\lambda_2} \leq \cot(\beta_k - \eta_k)\, \cot\eta_k \implies b - c\, \tan\beta_k \geq 0\,,$$

which is the other term of inequality (2.35), so also the minimum between those terms is greater than zero, and the point (b) is proven.

This construction finally provides a nonnegative discretization. □

In order to better illustrate the ideas for the proof of Theorem 7, we proceed by applying these results to the case considered in this report.
We want to find a nonnegative spatial discretization of the differential operator $\nabla \cdot (A \cdot \nabla u)$ on a (3×3)-stencil, with $A = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$ for the simple case of constant entries satisfying the parabolicity constraint (1.7).
Since $m = 1$, we have 4 principal directions $\beta_i \in (-\frac{\pi}{2}, \frac{\pi}{2}]$, $i = -1, 0, 1, 2$ from (2.33), as shown in Figure 2.9(a), which are given by

$$\beta_{-1} = \arctan\Big(-\frac{\Delta y}{\Delta x}\Big)\,, \quad \beta_0 = 0\,, \quad \beta_1 = \arctan\Big(\frac{\Delta y}{\Delta x}\Big)\,, \quad \beta_2 = \frac{\pi}{2}\,.$$

Now $J_1 = \{1\}$ defines a partition of $(-\frac{\pi}{2}, \frac{\pi}{2}]$ into 2 subintervals I_i, $|i| \in J_m$ such that

$$\big(-\frac{\pi}{2}, \frac{\pi}{2}\big] = \big(-\frac{\pi}{2}, 0\big] \cup \big(0, \frac{\pi}{2}\big] = I_{-1} \cup I_1\,,$$

where I_{-1} and I_1 belongs to the grid angles β_{-1} and β_1. First, we focus on the case $\psi \in I_1$, with $k = 1$. We have that a nonnegative splitting is possible if (2.35) is satisfied, from point (a), and this is true if (2.36) is valid, from point (b). So, we have to calculate $\cot(\rho_1 - \beta_1)\, \tan\rho_1$ and $\cot(\beta_1 - \eta_1)\, \cot\eta_1$. With the notation of Theorem 7, we obtain

$$\begin{aligned}
\theta_1 &= \frac{\pi}{2} \qquad \text{because} \quad i = 2m-1\,,\\
\rho_1 &= \frac{\theta_1+\beta_1}{2} = \frac{\pi}{4} + \frac{\beta_1}{2} \qquad \text{because} \quad |k| = 1\,,\\
\eta_1 &= \frac{\beta_1}{2} \qquad \text{because} \quad |k| = 1\,,
\end{aligned}$$

and, therefore, we have that

$$\begin{aligned}
\cot(\rho_1-\beta_1)\,\tan\rho_1 &= \cot\big(\frac{\pi}{4}+\frac{\beta_1}{2}-\beta_1\big)\,\tan\big(\frac{\pi}{4}+\frac{\beta_1}{2}\big)\\
&= \frac{1+\tan\frac{\pi}{4}\tan\frac{\beta_1}{2}}{\tan\frac{\pi}{4}-\tan\frac{\beta_1}{2}} \cdot \frac{\tan\frac{\pi}{4}+\tan\frac{\beta_1}{2}}{1-\tan\frac{\pi}{4}\tan\frac{\beta_1}{2}}\\
&= \frac{1+\tan\frac{\beta_1}{2}}{1-\tan\frac{\beta_1}{2}} \cdot \frac{1+\tan\frac{\beta_1}{2}}{1-\tan\frac{\beta_1}{2}} = \frac{(\cos\frac{\beta_1}{2}+\sin\frac{\beta_1}{2})^2}{(\cos\frac{\beta_1}{2}-\sin\frac{\beta_1}{2})^2}\\
&= \frac{1+2\cos\frac{\beta_1}{2}\sin\frac{\beta_1}{2}}{1-2\cos\frac{\beta_1}{2}\sin\frac{\beta_1}{2}} = \frac{1+\sin\beta_1}{1-\sin\beta_1}
\end{aligned}$$

and also

$$\cot(\beta_1-\eta_1)\,\cot\eta_1 = \cot(\beta_1-\frac{\beta_1}{2})\,\cot\frac{\beta_1}{2} = \cot^2\frac{\beta_1}{2} = \frac{1}{\tan^2\frac{\beta_1}{2}} = \frac{1+\cos\beta_1}{1-\cos\beta_1}\,,$$

where we have used the following trigonometrical formulas,

$$\tan(a-b) = \frac{\tan(a)-\tan(b)}{1+\tan(a)\,\tan(b)}\,, \qquad \tan(a+b) = \frac{\tan(a)+\tan(b)}{1-\tan(a)\,\tan(b)}\,,$$

$$2\cos(a)\,\sin(a) = \sin(2a)\,, \qquad \tan\big(\frac{a}{2}\big) = \sqrt{\frac{1-\cos(a)}{1+\cos(a)}}\,.$$

Then, we have that $\kappa_{1,1} = \min\left(\frac{1+\sin\beta_1}{1-\sin\beta_1}, \frac{1+\cos\beta_1}{1-\cos\beta_1}\right)$ and, thanks to the symmetry, we obtain the same condition for $\kappa_{-1,1}$, with $\psi \in I_{-1}$. Therefore, if (2.36) is satisfied, with the value of κ given above, a nonnegative discretization occurs. In particular, the inequality (2.35) is true and so

$$a \geq c\,\cot\beta_k = c\frac{\Delta x}{\Delta y}\,, \qquad b \geq c\,\tan\beta_k = c\frac{\Delta y}{\Delta x}\,. \tag{2.39}$$

To find the nonnegative discretization, we have to calculate the coefficients α_i, $i = -1, 0, 1$ through γ_0, γ_1 and γ_2. Since $\sin\beta_k = \sin(\arctan(\frac{\Delta y}{\Delta x})) = \Delta y$

and $\cos\beta_k = \cos(\arctan(\frac{\Delta y}{\Delta x})) = \Delta x$, we can calculate γ_i through (2.37). For $k = 1$ we have

$$\gamma_0 = a - c\,\cot\beta_1 = a - c\frac{\Delta x}{\Delta y}\,,$$
$$\gamma_1 = \frac{c}{\sin\beta_1\,\cos\beta_1} = \frac{c}{\Delta x\Delta y}\,,$$
$$\gamma_2 = b - c\,\tan\beta_1 = b - c\frac{\Delta y}{\Delta x}\,,$$

and for $k = -1$ we have

$$\gamma_0 = a - c\,\cot\beta_{-1} = a + c\frac{\Delta x}{\Delta y}\,,$$
$$\gamma_1 = \frac{c}{\sin\beta_{-1}\,\cos\beta_{-1}} = -\frac{c}{\Delta x\Delta y}\,,$$
$$\gamma_2 = b + c\,\tan\beta_{-1} = b + c\frac{\Delta y}{\Delta x}\,.$$

We recall that the γ_i's are all positive values, and we obtain the corresponding values α_i as follows,

$$\alpha_{-1} = \frac{|c| - c}{2\Delta x\Delta y}\,, \quad \alpha_0 = a - |c|\frac{\Delta x}{\Delta y}\,, \quad \alpha_1 = \frac{|c| + c}{2\Delta x\Delta y}\,, \quad \alpha_2 = b - |c|\frac{\Delta y}{\Delta x}\,,$$

which induce a second-order spatial discretization through the splitting (2.34).

$\frac{\lvert c\rvert - c}{2\Delta x\Delta y}$	$\frac{b}{\Delta y^2} - \frac{\lvert c\rvert}{\Delta x\Delta y}$	$\frac{\lvert c\rvert + c}{2\Delta x\Delta y}$
$\frac{a}{\Delta x^2} - \frac{\lvert c\rvert}{\Delta x\Delta y}$	$-\frac{2a}{\Delta x^2} - \frac{2b}{\Delta y^2} + \frac{2\lvert c\rvert}{\Delta x\Delta y}$	$\frac{a}{\Delta x^2} - \frac{\lvert c\rvert}{\Delta x\Delta y}$
$\frac{\lvert c\rvert + c}{2\Delta x\Delta y}$	$\frac{b}{\Delta y^2} - \frac{\lvert c\rvert}{\Delta x\Delta y}$	$\frac{\lvert c\rvert - c}{2\Delta x\Delta y}$

Table 2.3: stencil of the homogeneous Nonnegative Discretization

We report the whole stencil in Table (2.3), where the central entry is

$$-2\frac{\alpha_0}{\Delta x^2} - 2\frac{\alpha_2}{\Delta y^2} - 2\,(\alpha_{-1} + \alpha_1) = -\frac{2a}{\Delta x^2} - \frac{2b}{\Delta y^2} + \frac{2|c|}{\Delta x\Delta y}\,,$$

while the upper and the right-hand entries are

$$\frac{\alpha_2}{\Delta y^2} = \frac{b}{\Delta y^2} - \frac{|c|}{\Delta x\Delta y}\,, \qquad \frac{\alpha_0}{\Delta x^2} = \frac{a}{\Delta x^2} - \frac{|c|}{\Delta x\Delta y}\,.$$

We refer to [38] for the expression of the stencil of the nonnegative scheme for general heterogeneous diffusion tensors (1.2), as given in Table 2.4.

We observe that only three directions are sufficient to guarantee a *nonnegative directional splitting*. Thus, unless m is very small as in the above

$\frac{\lvert c_{i-1,j+1}\rvert - c_{i-1,j+1}}{4\Delta x\Delta y}$ $+\frac{\lvert c_{ij}\rvert - c_{ij}}{4\Delta x\Delta y}$	$\frac{b_{ij}+b_{i,j+1}}{2\Delta y^2} - \frac{\lvert c_{ij}\rvert + \lvert c_{i,j+1}\rvert}{2\Delta x\Delta y}$	$\frac{\lvert c_{i+1,j+1}\rvert + c_{i+1,j+1}}{4\Delta x\Delta y}$ $+\frac{\lvert c_{ij}\rvert + c_{ij}}{4\Delta x\Delta y}$
$\frac{a_{i-1,j}+a_{ij}}{2\Delta x^2}$ $-\frac{\lvert c_{i-1,j}\rvert + \lvert c_{ij}\rvert}{2\Delta x\Delta y}$	$-\frac{a_{i-1,j}+2a_{ij}+a_{i+1,j}}{2\Delta x^2}$ $-\frac{\lvert c_{i-1,j+1}\rvert - c_{i-1,j+1} + \lvert c_{i+1,j+1}\rvert + c_{i+1,j+1}}{4\Delta x\Delta y}$ $-\frac{\lvert c_{i-1,j-1}\rvert + c_{i-1,j-1} + \lvert c_{i+1,j-1}\rvert - c_{i+1,j-1}}{4\Delta x\Delta y}$ $+\frac{\lvert c_{i-1,j}\rvert + \lvert c_{i+1,j}\rvert + 2\lvert c_{ij}\rvert + \lvert c_{i,j-1}\rvert + \lvert c_{i,j+1}\rvert}{2\Delta x\Delta y}$ $-\frac{b_{i,j-1}+2b_{ij}+b_{i,j+1}}{2\Delta y^2}$	$\frac{a_{ij}+a_{i+1,j}}{2\Delta x^2}$ $-\frac{\lvert c_{ij}\rvert + \lvert c_{i+1,j}\rvert}{2\Delta x\Delta y}$
$\frac{\lvert c_{i-1,j-1}\rvert + c_{i-1,j-1}}{4\Delta x\Delta y}$ $+\frac{\lvert c_{ij}\rvert + c_{ij}}{4\Delta x\Delta y}$	$\frac{b_{i,j-1}+b_{ij}}{2\Delta y^2} - \frac{\lvert c_{i,j-1}\rvert + \lvert c_{ij}\rvert}{2\Delta x\Delta y}$	$\frac{\lvert c_{i+1,j-1}\rvert - c_{i+1,j-1}}{4\Delta x\Delta y}$ $+\frac{\lvert c_{ij}\rvert - c_{ij}}{4\Delta x\Delta y}$

Table 2.4: stencil of the heterogeneous Nonnegative Discretization

application, most of the stencil coefficients can be set to zero. Especially for large m values, a $(2m+1)\times(2m+1)$-stencil reveals much more directions than those $4m$ that are induced by the boundary pixels (the cases shown in Figure (2.9) are optimal for the respective dimensions). Therefore, even if we use only 3 directions, we may expect to find stricter estimates than those given in the proof of Theorem 7, and these estimates might be improved further by admitting more than 3 directions.

2.5 Discrete Maximum Principle

In the previous sections, we have seen how to develop numerical schemes for the diffusion equation (1.1) based on finite difference/volume methods.
It is often desirable to have a genuinely discrete theory which guarantees that an algorithm exactly reproduces the qualitative properties of its continuous counterpart. For the discrete operators introduced in this report, the main property we ask to benefit from is some analogue of (1.18), and Table (2.5) summerizes the suitable properties needed for the well-posedness of continuous or discrete diffusion processes [38].
These criteria are easy to check for many discretizations. More particularly, for the numerical schemes presented in this report, if the above properties are valid for the semi-discrete formulation, they are also valid for the fully discrete one. Indeed, because of the choice of one-step time-discretizations, these schemes generally admit the representation

$$u_{ij}^{n+1} = u_{ij}^{n} + \Delta t\,\{\text{space-discretization at time } n \text{ or/and } n+1\}\,.$$

Consequently, the property of conservation becomes that the sum of columns is equal to 1, which simply comes from the supplementary diagonal term u_{ij}^n, and the non-negativity of the entries extends also to the diagonal elements

	continuous problem $u_t = \nabla \cdot (A \cdot \nabla u)$ + initial/boundary conditions	*semi-discrete formulation* $u_t = M_s\, u$	*fully discrete formulation* $u^{n+1} = M_d\, u^n$
symmetry	A symmetric	M_s symmetric	M_d symmetric
conservation	divergence form	columns sum to 0	column sums to 1
nonnegativity	A positive semidefinite	nonnegative off-diagonals	nonnegative entries
connectivity	A uniformly positive definite	M_s irreducible	M_d irreducible, positive diagonal

Table 2.5: requirements of well-posed problems for diffusion equations

(under a suitable CFL-condition to be calculated explicitly for each numerical scheme). As a matter of fact, we will always consider matrices whose diagonal entries are strictly positive, otherwise it could be pathological cases of singular matrices.

Remark 7. *For the requirement of irreducibility for numerical matrices, we refer to [28][Chapter 2, Theorem 1.3], and we admit that the set of nonnegative irreducible matrices includes the positive matrices (so, if we have a positive matrix, automatically we have that it is irreducible).*

2.5.1 Theoretical results for semi-discrete problems

To show why the properties listed in Table 2.5 are important for numerical discretizations, we focus on semi-discrete problems, where only the spacial approximation is considered. The class of semi-discrete problems we are concerned with is characterized by the following definition. We remark that the theory is well-established even for the case of nonlinear operators, whereas in this report we deal only with linear PDEs.

Definition 6. *Let $u_0 \in \mathbb{R}^n$ and $u \in C^1([0,+\infty);\mathbb{R}^n)$ a solution to the initial value problem for a first order differential operator, namely*

$$\frac{\partial u}{\partial t} = A(u)u\,, \qquad u(0) = u_0\,, \tag{2.40}$$

where $A = \big(a_{ij}(u)\big)$, $i,j = 1,2,...,n$, has the following properties, $\forall\, u \in \mathbb{R}^n$,

P1) $A \in C(\mathbb{R}^n;\mathbb{R}^{n\times n})$ for every bounded subset of $\mathbb{R}^n$ (Lipschitz-continuity);

P2) symmetry : $a_{ij}(u) = a_{ji}(u)\,,\ \forall\, i,j = 1,2,...,n$;

P3) vanishing row sums : $\sum_j a_{ij}(u) = 0\,,\ \forall\, i = 1,2,...,n$;

P4) nonnegative off-diagonal entries : $a_{ij} \geq 0\,,\ \forall\, i \neq j$;

P5) irreducibility for all $u \in \mathbb{R}^n$.

For linear applications, where the diffusion tensor A does not depend on the function u, the property of Lipschitz-continuity is always satisfied because A is constant with respect to u. The properties $(P2)$ and $(P3)$ correspond to the specific structure of parabolic conservation laws, usually expressed in the so-called *divergence form*, while properties $(P4)$ and $(P5)$ play a similar role as the non-negativity of the eigenvalues of A and its uniform positive definiteness, respectively. We observe an immediate consequence of $(P3)$ and $(P4)$, that is the diagonal entries are strictly non-positive, also in view of $(P5)$. The proof of the discrete maximum principle mainly involves properties $(P3)$ and $(P4)$, as we will see in the following theorem, and we refer to [38][Chapter 3, Theorem 4] for the other theoretical results (global existence, regularity, conservation, energy contraction, large time behaviour) invoking in particular properties $(P2)$ and $(P5)$.

Theorem 8. *For every* $T > 0$*, the problem* (2.40) *admits a unique solution* $u(t) \in C^1([0,T];\mathbb{R}^n)$*. This solution depends continuously on the initial data and it satisfies the maximum/minimum principle*

$$u_{min} \leq u_i(t) \leq u_{max}\,, \quad \forall\, i = 1,\dots,n\,, \quad \forall\, t \in [0,T]\,, \tag{2.41}$$

where $u_{min} = \min_{i=1,\dots,n} u_0^i$ *and* $u_{max} = \max_{i=1,\dots,n} u_0^i$*, with* $u_0^i = u_i(0)$.

Proof. We demonstrate only the upper bound of the above inequality, but the proof is analogous for the left-hand side of (2.41).
We assume that the problem (2.40) has a unique solution on $[0,T]$, $T > 0$.

First, we show that the derivative of the largest component of $u(t)$ is nonpositive for every t inside the interval. Let $u_k(\bar{t}) = \max_{i=1,\dots,n} u_i(\bar{t})$ for some $\bar{t} \in [0,T]$. We fix k and, for $t = \bar{t}$, we obtain

$$\frac{\partial u_k}{\partial t} = \sum_j a_{kj}\, u_j = a_{kk}\, u_k + \sum_{j \neq k} a_{kj}\, u_j \leq u_k \sum_j a_{kj} = 0\,, \tag{2.42}$$

the inequality being justified through the property $(P4)$, while the last step is derived from $(P3)$. From now on, the implications are purely analytical. Let $\varepsilon > 0$ and set

$$u^\varepsilon(t) = u(t) - \begin{bmatrix} \varepsilon t \\ \vdots \\ \varepsilon t \end{bmatrix} \in \mathbb{R}^n,$$

together with $P := \{p = 1,\dots,n \mid u_p^\varepsilon(0) = \max_{i=1,\dots,n} u_i^\varepsilon(0) = \max_{i=1,\dots,n} u_0^i\}$. Then, from 2.42 we have that

$$\frac{\partial u_p^\varepsilon}{\partial t}(0) = \frac{\partial u_p}{\partial t}(0) - \varepsilon < 0\,, \quad \forall\, p \in P\,. \tag{2.43}$$

Since $\max_{i\notin P} u_i^\varepsilon(0) < \max_{i=1,\dots,n} u_i^\varepsilon(0)$, for the continuity of u, there exists some $t_1 \in (0,T)$ such that

$$\max_{i\notin P} u_i^\varepsilon(t) < \max_{i=1,\dots,n} u_i^\varepsilon(t)\,, \quad \forall\, t \in [0,t_1)\,. \tag{2.44}$$

Let us consider some fixed $p \in P$. Due to (2.43) and the smoothness of the solution, we may find $\bar{t}_p \in (0,T)$ such that

$$\frac{\partial u_p^\varepsilon}{\partial t}(t) < 0\,, \quad \forall\, t \in [0,\bar{t}_p)\,.$$

Thus, we have $u_p^\varepsilon(t) < u_p^\varepsilon(0)\,,\ \forall\, t \in (0,\bar{t}_p)$. We define $t_2 = \min_{p\in P} \bar{t}_p$, and we obtain

$$\max_{p\in P} u_p^\varepsilon(t) < \max_{i=1,\dots,n} u_i^\varepsilon(0)\,, \quad \forall\, t \in (0,t_2)\,. \tag{2.45}$$

At this point, taking $t_0 = \min\{t_1,t_2\}$ and using (2.44) and (2.45), we have

$$\max_{i=1,\dots,n} u_i^\varepsilon(t) < \max_{i=1,\dots,n} u_i^\varepsilon(0)\,, \quad \forall\, t \in (0,t_0)\,. \tag{2.46}$$

Now, we extend the estimate (2.46) to the whole interval $(0,T)$. We proceed by contradiction, assuming that the opposite is true. By means of the *mean value theorem*, there exists t_3 the smallest time in the interval such that

$$\max_{i=1,\dots,n} u_i^\varepsilon(t_3) \;= \max_{i=1,\dots,n} u_i^\varepsilon(0)\,.$$

Let $u_k^\varepsilon = \max_{i=1,\dots,n} u_i^\varepsilon(t_3)$. For the minimality of t_3 we have that

$$u_k^\varepsilon(t) < u_k^\varepsilon(t_3)\,, \quad \forall\, t \in (0,t_3)\,. \tag{2.47}$$

By the inequality (2.42), we deduce that

$$\frac{\partial u_k^\varepsilon}{\partial t}(t_3) = \frac{\partial u_k}{\partial t}(t_3) - \varepsilon < 0\,,$$

and, by continuity of the derivative, there exists some $t_4 \in (0,t_3)$ with

$$\frac{\partial u_k^\varepsilon}{\partial t}(t) < 0\,, \quad \forall\, t \in (t_4,t_3]\,. \tag{2.48}$$

The *mean value theorem*, combined with (2.47), implies that we have found $t_5 \in (t_4,t_3)$ with

$$\frac{\partial u_k^\varepsilon}{\partial t}(t_5) > 0\,.$$

This clearly contradicts (2.48). Hence the inequality (2.46) has to be valid on the whole interval $(0,T)$. Finally, together with $u = \lim_{\varepsilon\to 0} u_\varepsilon$ and the continuity of the solution, we have that the maximum principle holds. □

2.5.2 Application to the finite difference schemes

For the sake of readability, we restrict to the one-dimensional equation (2.20), although the results of Theorem 8 apply to numerical schemes in any spatial dimension, as we will see in the next chapter, and we adopt the notation introduced in Section 2.1.

For the **Chain Rule method** described in Section 2.4.1, there is no way to impose the non-negativity of all stencil entries. Indeed, if we discretize the one-dimensional domain with points $\{x_i\}_{0\leq i\leq N_x}$, where $\Delta x = |x_{i+1} - x_i|$ for a uniform mesh, and we consider the explicit Euler method for the time-discretization, also setting $a_i = a(x_i) > 0\,, \forall\, i = 0, 1, ...$, we have

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{a_{i+1} - a_{i-1}}{2\Delta x} \cdot \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} + a_i \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2}\,, \qquad (2.49)$$

so that the scheme in Table 2.1 reduces to

$$\begin{aligned} u_i^{n+1} = {} & \frac{\Delta t}{\Delta x^2}\Big(\frac{a_{i+1} - a_{i-1}}{4} + a_i\Big)u_{i+1}^n + \Big(1 - 2\frac{\Delta t}{\Delta x^2}\, a_i\Big)u_i^n \\ & + \frac{\Delta t}{\Delta x^2}\Big(-\frac{a_{i+1} - a_{i-1}}{4} + a_i\Big)u_{i-1}^n\,. \end{aligned}$$

In order to satisfy the non-negativity of the coefficients, we have to impose

$$\frac{\Delta t}{\Delta x^2}\Big(\frac{a_{i+1} - a_{i-1}}{4} + a_i\Big) > 0\,, \quad \frac{\Delta t}{\Delta x^2}\Big(-\frac{a_{i+1} - a_{i-1}}{4} + a_i\Big) > 0\,, \qquad (2.50)$$

independently from the choice of the diffusion function $a(x)$, because we can always assume $1 - 2\frac{\Delta t}{\Delta x^2} a_i > 0$ by modifying appropriately the ratio between time-step and space-step to fulfill the so-called *parabolic CFL-condition*

$$\Delta t < \frac{\Delta x^2}{2\, a_{\max}}\,, \qquad a_{\max} := \max_{0\leq i\leq N_x} a_i\,. \qquad (2.51)$$

For the second condition in (2.50), we would have

$$a_i - \frac{a_{i+1} - a_{i-1}}{4} > 0\,, \quad \forall\, i = 0, 1, \ldots, N_x\,.$$

We deduce from a Taylor's expansion that $\frac{a_{i+1} - a_{i-1}}{4} = \frac{\Delta x}{2}\, a_i' + O(\Delta x^3)$, with abuse of notation if we denote by a_i' the value of the derivative at point x_i. Finally, the diffusion coefficients should satisfy the following constraint,

$$a_i > \frac{\Delta x}{2} a_i'\,, \quad \forall\, i = 0, 1, \ldots, N_x\,,$$

which alternatively gives an extra-restriction on the space-step Δx (that is practically viable only for linear problems, i.e. in case of diffusion functions solely depending on space, and eventually time). We remark that a similar

constraint is deduced from the first condition in (2.50) and this concerns all possible signs for the derivatives. Therefore, if we approximate the parabolic operator on a *coarse grid* (as motivated by computational cost) and the diffusion function a attains low values over the domain but with rapid growth, i.e. a_i is very small but its derivative is very high, the Chain Rule method fails to be nonnegative, and the maximum principle can be violated.

Remark 8. *For the one-dimensional parabolic equation in non-conservative form with constant coefficients, i.e.* $u_t = b\,u_x + a\,u_{xx}$, $b \in \mathbb{R}$, $a \in \mathbb{R}^+$, *we can consider the finite difference centered scheme*

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} \;=\; b\,\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} + a\,\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2}\,,$$

which is structurally similar to (2.49) *since based on the same principle of separating the terms of different order, and we can rewrite it as*

$$u_i^{n+1} = \frac{\Delta t}{\Delta x}\Big(\frac{b}{2} + \frac{a}{\Delta x}\Big)u_{i+1}^n + \Big(1 - 2\frac{\Delta t}{\Delta x^2}\,a\Big)u_i^n + \frac{\Delta t}{\Delta x}\Big(-\,\frac{b}{2} + \frac{a}{\Delta x}\Big)u_{i-1}^n\,.$$

Under the parabolic CFL-condition $\Delta t < \frac{\Delta x^2}{2\,a}$ *, also corresponding to* (2.51)*, the central coefficient above is positive and less than* 1*, moreover the sum of the coefficients is equal to* 1 *and they can be made all positive (independently from the sign of* b*) providing that the space-step* Δx *is sufficiently small. Nevertheless, there is no reason why those coefficients should be less than* 1 *, thus preventing an* L^∞*-stability of the scheme. Indeed, that scheme turns out to be a weakly-parabolic correction of the centered scheme for hyperbolic problems, which is well-known to exhibit oscillations [22], because the maximum principle is not fulfilled. The* L^2*-stability is guaranteed, however, as we can easily check by performing an analysis of its modified equation (obtained by Taylor's expansions and manipulating the exact equation)*

$$\begin{aligned} u_t - b\,u_x - a\,u_{xx} - \big(\frac{a}{12}\Delta x^2 - \frac{a^2}{2}\Delta t\big)u_{xxxx} \\ + \frac{b^2}{2}\Delta t\,u_x + \big(a\,b\,\Delta t - \frac{b}{6}\Delta x^2\big)u_{xxx} + O(\Delta t^2, \Delta x^5) = 0\,, \end{aligned}$$

under some quite restrictive CFL-type condition, namely $\Delta t < \frac{\Delta x^2}{6\,a}$*, which is the same predicted for the fully-parabolic problem in Section 3.1.4.*

The limits of the Chain Rule method become insurmountable when passing to two-dimensional problems : unless one deals with diagonal matrices, for which the issue discussed above is however relevant, the off-diagonal elements of the stencil in Table 2.1 cannot always be positive for fully anisotropic diffusion tensors, as the coefficients c may have any sign.

The discussion becomes more interesting for the **Standard Discretization** introduced in Section 2.4.2, which equals the finite volume scheme (2.28) in the one-dimensional setting, as given by (2.31), thus resulting in a positive approximation. For two-dimensional problems, we can see in Table 2.2 that horizontal and vertical off-diagonal elements of the stencil are nonnegative, since a and b must be nonnegative from (1.7), but the non-negativity of the whole stencil cannot be guaranteed since c has undefined sign.
Nevertheless, the scheme in Table 2.2 enjoys certain *stability* properties, as reported in [7], because the approximate solution at any arbitrary fixed time remains bounded for some suitable norm defined on Ω. Typically, we consider the *maximum norm* and we say that the scheme is stable in $(\Omega, ||\cdot||_\infty)$ if there exists a constant $C_T > 0$ such that $||u^n||_\infty \leq C_T||u_0||_\infty$, $\forall\ n > 0$. Explicit methods are often conditionally stable, i.e. if the time-step Δt is chosen under certain *CFL-conditions*, as we will discuss in the next chapter.

The **Nonnegative method** derived in Section 2.4.3 reduces to the standard discretization, i.e. the finite difference scheme on staggered grids, in the one-dimensional case : this comes from the fact that the only modification needed to achieve non-negativity concerns the mixed derivatives, whose effect cannot be appreciated when no dimensional interaction is present.
Then, it seems that the failure of the non-negativity property can actually occur whenever the scheme involves terms with mixed derivatives, and a nonnegative correction has to be introduced to control the coefficients at the vertices of Table 2.2, which are the only ones produced by the mixed derivatives. We will see in the next chapter that the nonnegative method in Table 2.4 satisfies all the properties listed in Table (2.5).

Chapter 3

Stability Analysis of one-dimensional methods

In this chapter, we revisit details of classical methods in the one-dimensional setting, focusing on several finite difference/volume schemes, essentially in order to illustrate the main tools and strategies we will later apply to the study of two-dimensional problems.
We remark that most schemes introduced in the previous chapter coincide for one-dimensional equations, because the finite difference method on staggered grids can be reinterpreted as the finite volume approach, and for those schemes we check the validity of the properties listed in Table 2.5.

3.1 The simplest homogeneous case

We start by considering the homogeneous one-dimensional heat equation

$$u_t - a\, u_{xx} = 0\,, \qquad (t,x) \in \Omega_T = [a_x, b_x] \times [0,T]\,, \tag{3.1}$$

with scalar constant diffusion $a > 0$, so that the density u feels the effect of diffusion uniformly and independently from its position, together with initial data and homogenous Dirichlet boundary conditions

$$\begin{aligned} u(0,x) &= u_0(x) && \text{for } x \in [a_x, b_x]\,, \\ u(t,a_x) &= g_1(t) = 0 && \text{for } t \in (0,T)\,, \\ u(t,b_x) &= g_2(t) = 0 && \text{for } t \in (0,T)\,. \end{aligned} \tag{3.2}$$

We divide the computational domain in N_t and N_x intervals, for $[0,T]$ and $[a_x, b_x]$ respectively, according to the following notation

$$\begin{aligned} t_n &= n\,\Delta t\,, && n = 0, 1, \ldots, N_t\,, \\ x_i &= a_x + i\,\Delta x\,, && i = 0, 1, \ldots, N_x\,, \end{aligned} \tag{3.3}$$

where $t_0 = 0$ and $\Delta x = \frac{b_x - a_x}{N_x}$, so that $x_0 = a_x$ and $x_{N_x} = b_x$, while Δt will be determined by means of the CFL-condition introduced later.
We approximate the initial data in (3.2) as a vector $u_0 \in \mathbb{R}^{N_x+1}$, with

$$u_0^i \simeq u_0(x_i) = u(0, x_i)\,, \qquad i = 0, 1, \ldots, N_x\,. \tag{3.4}$$

We denote by $x_{i+\frac{1}{2}}$, $i = 1, \ldots, N_x - 1$, the interfacial points which characterize *cell-centered nodes* through $x_i = \frac{x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}}{2}$, and we observe that this definition holds also for nonuniform meshes, although we do not treat that case explicitly in this report. Therefore, since $x_{i+\frac{1}{2}} - x_{i+\frac{1}{2}} = \Delta x$, we have the following property for the cell-averages,

$$\frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_0(x)\, dx = u_0(x_i) + O(\Delta x^2)\,, \tag{3.5}$$

which allows using indifferently (3.4) and (3.5) as discretization for the initial data at least for second order schemes (the first approximation being easier to compute in some practical applications).

3.1.1 Numerical schemes for one-dimensional heat equation

For the derivation of the numerical schemes reviewed in this section, we refer to the formulas in Section 2.2. For the general study of the heat equation in the one-dimensional case, one can refer to [20], for instance.

We recall the formulation of the **explicit scheme** from (2.17), that reads

$$u_i^{n+1} = u_i^n + a\frac{\Delta t}{\Delta x^2}\big(u_{i+1}^n - 2u_i^n + u_{i-1}^n\big), \quad i = 1, 2, \ldots, N_x - 1\,, \tag{3.6}$$

for which the numerical solution at time t_{n+1} is computed entirely in terms of the values at previous time t_n. We introduce the parameter $\lambda = \frac{\Delta t}{\Delta x^2}$ and we rewrite (3.6) in compact matrix-form, for any fixed $n = 0, 1, \ldots$,

$$U^{n+1} = M_{exp} \cdot U^n\,,$$

where $U^n = (u_1^n, u_2^n, \ldots, u_{N_x-1}^n)$ is the $(N_x - 1)$-vector of discrete unknowns and the matrix of the numerical scheme $M_{exp} \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ is given by

$$M_{exp} = \begin{bmatrix} 1 - 2a\lambda & a\lambda & 0 & \cdots \\ a\lambda & 1 - 2a\lambda & a\lambda & \cdots \\ 0 & a\lambda & 1 - 2a\lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.7}$$

It is worthwhile observing that (3.7) is meaningful only for homogeneous Dirichlet boundary conditions (3.2), namely for $u_0^n = u_{N_x}^n = 0\,, \forall\, n \geq 0\,,$

otherwise the computation of the first and last component of U^n should be adapted, also modifying consequently the expression (3.7).

More generally, the matrix-form of a numerical scheme looks like

$$U^{n+1} = M \cdot U^n + g^n, \qquad U^n = \left(u_0^n, u_1^n, \dots, u_{N_x}^n\right) \in \mathbb{R}^{N_x+1},$$

where $M \in \mathbb{R}^{(N_x+1)\times(N_x+1)}$ and $g^n \in \mathbb{R}^{N_x+1}$ is the vector including initial data and boundary conditions in some appropriate way.

The **implicit scheme** given by (2.18) is based on the approximation of the second order derivatives at time t^{n+1}, and it directly results in

$$u_i^{n+1} = u_i^n + a\frac{\Delta t}{\Delta x^2}\left(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}\right), \quad i = 1, 2, \dots, N_x - 1\,. \tag{3.8}$$

The matrix-form is the following system of linear algebraic equations,

$$M_{imp} \cdot U^{n+1} = U^n\,,$$

where

$$M_{imp} = \begin{bmatrix} 1+2a\lambda & -a\lambda & 0 & \cdots \\ -a\lambda & 1+2a\lambda & -a\lambda & \cdots \\ 0 & -a\lambda & 1+2a\lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.9}$$

so that it can be rewritten as

$$U^{n+1} = M_{imp}^{-1} \cdot U^n\,, \tag{3.10}$$

under the hypothesis that M_{imp} is not singular, and with the same remark for the numerical boundary conditions.

The **Crank-Nicolson scheme** is a combination of the explicit and the implicit schemes, whose formulation for $i = 1, 2, \dots, N_x - 1$ is given by

$$u_i^{n+1} = u_i^n + \frac{a}{2}\frac{\Delta t}{\Delta x^2}\left(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}\right) + \frac{a}{2}\frac{\Delta t}{\Delta x^2}\left(u_{i+1}^{n} - 2u_i^{n} + u_{i-1}^{n}\right), \tag{3.11}$$

and the corresponding matrix-form reads

$$M_1 \cdot U^{n+1} = M_2 \cdot U^n\,, \tag{3.12}$$

with

$$M_1 = \begin{bmatrix} 1+a\lambda & -\frac{a}{2}\lambda & 0 & \cdots \\ -\frac{a}{2}\lambda & 1+a\lambda & -\frac{a}{2}\lambda & \cdots \\ 0 & -\frac{a}{2}\lambda & 1+a\lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

and

$$M_2 = \begin{bmatrix} 1-a\lambda & \frac{a}{2}\lambda & 0 & \cdots \\ \frac{a}{2}\lambda & 1-a\lambda & \frac{a}{2}\lambda & \cdots \\ 0 & \frac{a}{2}\lambda & 1-a\lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The three schemes presented before are particular cases of **θ-methods**, for $0 \leq \theta \leq 1$, introduced in (2.19), with the alternative representation

$$\begin{aligned} u_i^{n+1} = u_i^n + a(1-\theta)\lambda\big(u_{i+1}^n - 2u_i^n + u_{i-1}^n\big) \\ + a\theta\lambda\big(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}\big), \end{aligned} \tag{3.13}$$

so that for $\theta = 0$ we have the explicit method, for $\theta = 1$ we have the implicit method, and for $\theta = \frac{1}{2}$ we have the Crank-Nicolson method.
This compact formulation has the advantage to allow the theory of stability to be established uniquely for the general class, as well as the discussion of the validity of the maximum principle.

In the next sections, we will first address the issue of stability of the θ-methods through the *Von Neumann analysis* and, then, we will enunciate and demonstrate the discrete maximum principle, also providing precise comments about the necessary conditions on the numerical parameters.

3.1.2 L^2-stability of the θ-methods

The Von Neumann stability analysis, also known as *Fourier stability analysis*, is a standard procedure to check the stability of finite difference schemes applied to linear PDEs, which is based on the study of the propagation of *numerical errors*. A numerical scheme is stable if the error made after one time-step of the calculation is limited by a constant which does not depend on time : if the errors decay and eventually damp out, the numerical scheme is said to be stable; if, on the contrary, the errors grow with time, the numerical scheme is said to be unstable. The stability of numerical schemes for a large class of PDEs of hyperbolic and parabolic type can be successfully investigated by performing the Von Neumann analysis [25],[34],[12],[22],[20].

The Von Neumann analysis refers to the stability of the L^2-norm of the solution, which is motivated by the functional properties of the continuous model (3.1)-(3.2). Indeed, assuming that u enjoys all the regularity needed to work with derivatives and integrals, we multiply equation (3.1) by its

solution and we integrate over the spatial domain Ω, thus obtaining

$$\begin{aligned}\int_\Omega u\, u_t\, dx &= \frac{1}{2}\frac{\partial}{\partial t}\int_\Omega |u|^2 dx = \frac{1}{2}\frac{\partial}{\partial t}||u||^2_{L^2} \\ &= a\int_\Omega u\, u_{xx}\, dx = a\int_\Omega (u\, u_x)_x\, dx - a\int_\Omega |u_x|^2 dx \\ &= a\int_{\partial\Omega} u\, u_x\cdot\vec{\nu}\, ds - a\,||u_x||^2_{L^2}\,,\end{aligned}$$

where we have applied the *divergence theorem* to $\Omega = [a_x, b_x]$ and, according to the boundary conditions (3.2) we have $\int_{\partial\Omega} u\, u_x\cdot\vec{\nu}\, ds = 0$. Therefore,

$$\frac{1}{2}\frac{\partial}{\partial t}||u||^2_{L^2} = -a\,||u_x||^2_{L^2} \le 0\,,$$

from which we infer that the L^2-norm of the analytical solution decreases as time evolves. In particular, we would recover similar results for the numerical solution, so we have to designate some *discrete L^2-norm*, denoted by $||\cdot||_{l_2}$, that is appropriate for functions defined only at discrete points (t^n, x_i) of the computational grid. Referring to the staggered grid 2.3, for example,

$$\begin{aligned}&\int_{a_x}^{b_x} |u(t^n,x)|^2\, dx \\ &= \int_{a_x}^{x_{\frac{1}{2}}} |u(t^n,x)|^2\, dx + \int_{x_{\frac{1}{2}}}^{x_{N_x-\frac{1}{2}}} |u(t^n,x)|^2\, dx + \int_{x_{N_x-\frac{1}{2}}}^{b_x} |u(t^n,x)|^2\, dx \\ &\simeq \frac{\Delta x_0}{2}|u_0^n|^2 + \sum_{i=1}^{N_x-1}\Delta x_i|u_i^n|^2 + \frac{\Delta x_{N_x}}{2}|u_{N_x}^n|^2 = \sum_{i=0}^{N_x}\delta x_i|u_i^n|^2 =: ||u^n||^2_{l^2}\,,\end{aligned}$$

where in the last equality we have set the length of the cells as following,

$$\delta x_0 = \frac{\Delta x_0}{2}\,,\quad \delta x_{N_x} = \frac{\Delta x_{N_x}}{2}\,,\qquad \delta x_i = \Delta x_i\,,\quad i = 1,2,\ldots,N_x-1\,,$$

and we have $\Delta x_i = \Delta x\,,\ \forall\, i = 0,1,\ldots,N_x$, if the mesh is uniform.
The above definition of discrete L^2-norm is coherent with the assumption that the numerical solution is a piecewise constant function reconstructed through the values u_i^n on the cells, i.e. $U^n = \sum_{i=0}^{N_x}\chi_{C_i}u_i^n$, with $C_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, as the one-dimensional counterpart of (2.1).

As an example, for simplicity, we refer to the explicit scheme (3.6). Let v_i^n be the truncated solution u at some order $p > 0$ in time and $q > 0$ in space, such that

$$\frac{v_i^{n+1}-v_i^n}{\Delta t} - a\frac{v_{i+1}^n - 2v_i^n + v_{i-1}^n}{\Delta x^2} = R(\Delta t^p, \Delta x^q)\,.$$

We note that this is equivalent to imposing that the scheme in *consistent*. We define the *local truncation error* as $E_i^n := v_i^n - u_i^n$ and, since the numerical scheme is linear, we can deduce an equation of the same type, i.e.

$$\frac{E_i^{n+1} - E_i^n}{\Delta t} - a\frac{E_{i+1}^n - 2E_i^n + E_{i-1}^n}{\Delta x^2} = R(\Delta t^p, \Delta x^q)\,,$$

or rather

$$E_i^{n+1} = a\frac{\Delta t}{\Delta x^2}\,E_{i+1}^n + \big(1 - 2a\frac{\Delta t}{\Delta x^2}\big)E_i^n + a\frac{\Delta t}{\Delta x^2}\,E_{i-1}^n + R(\Delta t^p, \Delta x^q)\,.$$

The right-hand side of the above equation is actually a *convex combination* of the values E_j^n, $j = i-1, i, i+1$, with coefficients that are positive and less than 1 under the CFL-condition $\big(1 - 2a\frac{\Delta t}{\Delta x^2}\big) > 0$. Finally, for the l_2-norm defined before, we obtain

$$||E^{n+1}||_{l_2} \le C\,||E^n||_{l_2}\,, \qquad \forall\, n = 0, 1, \dots, \left[\frac{T}{\Delta t}\right]\,,$$

with $C < 1$, so that the numerical scheme is stable.

Remark 9. *The basic property allowing to perform the above stability analysis, and the Von Neumann analysis developed in this section, is the* ***linearity*** *of the numerical scheme, otherwise stronger assumptions must be considered. To see this, we look at the nonlinear equation*

$$u_t - \big(\Phi(u)\big)_{xx} = 0\,, \quad t > 0\,,\ x \in \mathbb{R}\,, \tag{3.14}$$

where $\Phi(u)$ is some nonlinear function (usually increasing with respect to u). We discretize the domain through a staggered grid and we make the assumption that the approximate solution is piecewise constant on the mesh, so that we can choose of treating the nonlinear term like $\Phi(u)_i^n = \Phi(u_i^n)$. Thus, we compute the numerical scheme as in Section (2.3) and we obtain

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{\Phi(u_{i+1}^n) - 2\Phi(u_i^n) + \Phi(u_{i-1}^n)}{\Delta x^2}\,. \tag{3.15}$$

For the truncated solution v_i^n, thanks to Taylor's expansions, it holds

$$\frac{v_i^{n+1} - v_i^n}{\Delta t} - \frac{\Phi(v_{i+1}^n) - 2\Phi(v_i^n) + \Phi(v_{i-1}^n)}{\Delta x^2} = R(\Delta t^p, \Delta x^q)\,,$$

and for the error $E_i^n = v_i^n - u_i^n$ we have the equation

$$\frac{E_i^{n+1} - E_i^n}{\Delta t} - \frac{\Delta\Phi_{i+1} - 2\Delta\Phi_i + \Delta\Phi_{i-1}}{\Delta x^2} = R(\Delta t^p, \Delta x^q)\,,$$

where $\Delta\Phi_i := \Phi(v_i^n) - \Phi(u_i^n)$. Now, it is clear that only under structural hypotheses on the model, for instance Φ is Lipschitz-contractive, namely

$$|\Phi(v_i^n) - \Phi(u_i^n)| \le C_\Phi |v_i^n - u_i^n|\,, \tag{3.16}$$

for some constant $C_\Phi < 1$, the scheme is stable according to the previous analysis. Otherwise, specific terms must be added to stabilize the method.

There is another approach to the L^2-stability, using the *Fourier analysis*, which is possible in case of linear, autonomous and homogeneous equations like (3.1). This consists in studying the evolution of the L^2-norm in the *space of frequencies*, due to the *Parseval's equality* $||u||_{L^2}^2 = ||\hat{u}||_{L^2}^2$, where $\hat{u}$ represents the Fourier transform of the function u (through this, the properties of regularity of the solution are transformed into properties of decreasing at infinity of its Fourier transform). We fix the time t^n at which we want to calculate the solution, so we have $u^n(x) = u(t^n, x)$. We denote by $\hat{u}^n(\xi)$ the Fourier transform in the one-dimensional space variable,

$$\hat{u}^n(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\iota x\xi}\, u^n(x)\, dx \simeq \frac{1}{\sqrt{2\pi}} \sum_{i=-\infty}^{+\infty} e^{-\iota(i\Delta x)\xi}\, u_i^n \Delta x\,,$$

for $\xi \in [-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}]$. Then, the inverse Fourier transform is given by

$$u^n(x_i) = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{\iota(i\Delta x)\xi}\, \hat{u}^n(\xi)\, d\xi\,. \tag{3.17}$$

Because of the assumption $u_i^n \simeq u^n(x_i)$, we can substitute the representation (3.17) into the finite difference schemes, and for the general θ-method (3.13) we obtain

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{\iota(i\Delta x)\xi} \left[\frac{\hat{u}^{n+1}(\xi) - \hat{u}^n(\xi)}{\Delta t}\right] d\xi \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} \frac{a(1-\theta)}{\Delta x^2} \left[e^{\iota(i+1)\Delta x\,\xi}\, \hat{u}^n(\xi) - 2e^{\iota(i\Delta x)\xi}\, \hat{u}^n(\xi) + e^{\iota(i-1)\Delta x\,\xi}\, \hat{u}^n(\xi)\right] d\xi \\
&\quad + \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} \frac{a\theta}{\Delta x^2} \left[e^{\iota(i+1)\Delta x\,\xi}\, \hat{u}^{n+1}(\xi) - 2e^{\iota(i\Delta x)\xi}\, \hat{u}^{n+1}(\xi) + e^{\iota(i-1)\Delta x\,\xi}\, \hat{u}^{n+1}(\xi)\right] d\xi\,.
\end{aligned}$$

Neglecting the common multiplication constant and integrals, also collecting the exponential term $e^{\iota(i\Delta x)\xi}$, we deduce the following *sufficient condition* :

$$\begin{aligned}
\frac{\hat{u}^{n+1}(\xi) - \hat{u}^n(\xi)}{\Delta t} &= \frac{a(1-\theta)}{\Delta x^2} \left(e^{\iota\Delta x\,\xi} - 2 + e^{-\iota\Delta x\,\xi}\right) \hat{u}^n(\xi) \\
&\quad + \frac{a\theta}{\Delta x^2} \left(e^{\iota\Delta x\,\xi} - 2 + e^{-\iota\Delta x\,\xi}\right) \hat{u}^{n+1}(\xi)\,,
\end{aligned}$$

which can be further rewritten as

$$\begin{aligned}
&\left[1 - a\theta\frac{\Delta t}{\Delta x^2} \left(e^{\iota\Delta x\,\xi} - 2 + e^{-\iota\Delta x\,\xi}\right)\right] \hat{u}^{n+1} \\
&= \left[1 + a(1-\theta)\frac{\Delta t}{\Delta x^2} \left(e^{\iota\Delta x\,\xi} - 2 + e^{-\iota\Delta x\,\xi}\right)\right] \hat{u}^n\,,
\end{aligned}$$

or rather in a more compact form, with the usual notation $\lambda = \frac{\Delta t}{\Delta x^2}$,

$$\hat{u}^{n+1} = G\,\hat{u}^n\,, \qquad G := \frac{1 + a(1-\theta)\lambda\left(e^{\iota\Delta x\,\xi} - 2 + e^{-\iota\Delta x\,\xi}\right)}{1 - a\theta\lambda\left(e^{\iota\Delta x\,\xi} - 2 + e^{-\iota\Delta x\,\xi}\right)}, \tag{3.18}$$

where $G = G(a, \theta, \lambda)$ is the so-called *amplification factor* of the scheme. Through the analysis of G we have the stability according to the following theorem (referring to [30],[25],[22], for the classical proof).

Theorem 9. *A one-step finite difference scheme for the equation* (3.1) *is stable iff there exists a constant* K *and fixed values* Δt_0 , Δx_0 *such that*

$$|G| \leq 1 + K\Delta t\,,$$

for all $\Delta x\,\xi \in \mathbb{C}$ *and* $0 < \Delta t < \Delta t_0$, $0 < \Delta x < \Delta x_0$.
Moreover, if G *is independent from* Δt *and* Δx*, one can choose* $K = 0$.

We aim at verifying the statement of Theorem 9 for the amplification factor in (3.18) associated to the θ-methods. First, we focus on the term

$$e^{\iota\Delta x\,\xi} - 2 + e^{-\iota\Delta x\,\xi} = 2\,\cos(\Delta x\,\xi) - 2 = -4\,\sin^2\big(\frac{\Delta x\,\xi}{2}\big)\,. \tag{3.19}$$

We put the last expression into (3.18) and we obtain

$$G = \frac{1 - 4\,a(1-\theta)\lambda\,\sin^2(\frac{\Delta x\,\xi}{2})}{1 + 4\,a\theta\lambda\,\sin^2(\frac{\Delta x\,\xi}{2})}. \tag{3.20}$$

We need to check the condition $|G| \leq 1$. Since $\lambda > 0$, we always have that

$$1 - 4\,a(1-\theta)\lambda\,\sin^2\big(\frac{\Delta x\,\xi}{2}\big) \leq 1 + 4\,a\theta\lambda\,\sin^2\big(\frac{\Delta x\,\xi}{2}\big)\,,$$

therefore $G \leq 1$ and it remains to analyze the case when $G \geq -1$ to have stability, namely from (3.20) this occurs (otherwise we have instability) for

$$4\,a(1-2\theta)\lambda\,\sin^2\big(\frac{\Delta x\,\xi}{2}\big) \leq 2\,. \tag{3.21}$$

Thanks to $0 \leq \sin^2(\frac{\Delta x\,\xi}{2}) \leq 1$, the above condition is fulfilled if it holds

$$4\,a(1-2\theta)\lambda \leq 2\,.$$

In conclusion, for $0 \leq \theta < \frac{1}{2}$ the method is stable if and only if

$$\lambda \leq \frac{1}{2a(1-2\theta)}\,, \tag{3.22}$$

and for $\frac{1}{2} \leq \theta \leq 1$ the method is stable for all λ .
In particular, the explicit scheme is stable if $\lambda \leq \frac{1}{2a}$, which is the well-known *parabolic CFL condition*, while the implicit scheme and the Crank-Nicolson scheme are *unconditionally stable*.

Remark 10. *The parabolic CFL condition* $\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2a}$ *is actually quite restrictive, because if we want to halve the space-step we have to take a quarter of the time-step, with a relevant cost in terms of computational time. The semi-implicit methods resolve this problem. Nevertheless, explicit schemes are much simpler to be implemented, and they produce less numerical viscosity with respect to the implicit ones, thus rehabilitating their use especially in the context of CUDA GPU applications.*

3.1.3 The question of the Ultraviolet Catastrophe

The Fourier analysis gives also informations about the qualitative behaviour of the numerical solution with respect to the underlying physical problem.

For simplicity, we consider $x \in \Omega = [0,1]$ and we calculate an analytical solution to (3.1) by *separation of variables*, i.e. with the special form

$$\bar{u}(t,x) = e^{-ak^2 t} \sin(kx)\,,$$

where k is a real constant, which can be determined imposing the boundary conditions (3.2), and we have $k = m\,\pi\,,\ m \in \mathbb{N}\,.$ Any linear combination of such solutions will satisfy the differential equation (3.1), so that we write

$$u(t,x) = \sum_{m \geq 1} f_m\, e^{-a(m\pi)^2 t} \sin(m\,\pi x)\,,$$

with coefficients f_m to be chosen in order to satisfy the initial data at $t = 0\,,$

$$u_0(x) = \sum_{m \geq 1} f_m\, \sin(m\,\pi x)\,.$$

This shows that $f_m\,,\ m = 1, 2, \ldots,$ are just the coefficients of the expansion in *Fourier series* of the given function, that is

$$f_m = 2 \int_0^1 u_0(x) \sin(m\,\pi x)\, dx\,.$$

Therefore, we have expressed the exact solution to the partial differential equation as a Fourier series, and this expression is based on the observation that a particular set of *Fourier modes* are exact solutions to (3.1).

Following [25], for the numerical solution we can use the expression (3.18), thus obtaining $\hat{u}^n = [G(\Delta x\,\xi)]^n \hat{u}_0$ by iteration, and putting this back into (3.17) we get

$$u_i^n = \sum_{-\infty}^{\infty} A_m e^{\iota a(m\pi)(i\Delta x)} [g(m\pi)]^n. \tag{3.23}$$

The low frequency terms in the above expansion give a good approximation to the exact solution, but for large values of ξ the modes of the exact solution are rapidly damped by the exponential factor $e^{-\xi^2 \Delta t}$, whilst in the numerical solution the damping factor G will became greater than 1 when $\lambda > \frac{1}{2a(1-2\theta)}$, for $0 \leq \theta < \frac{1}{2}$, thus bringing the scheme into instability regimes.
In particular, from (3.21) we observe that the sharpest condition to be satisfied for stability is when $\Delta x\,\xi \simeq \pi\,,$ for that $\sin^2(\frac{\Delta x\,\xi}{2}) \simeq 1\,,$ otherwise also bigger values of λ would be possible as the factor $\sin^2(\frac{\Delta x\,\xi}{2}) << 1$ for small frequencies $\xi\,.$ Moreover, this explains why the instabilities cannot be resolved by simply reducing the space-step and, of course, these unstable Fourier modes grow unboundedly as n increases (see Section 5.2.2).

This phenomenon is known as the *Ultraviolet Catastrophe*, referring to problems typically occurring at high frequencies, and it takes place in a large class of physical models : for the one-dimensional heat equation, for example, all frequencies are dissipated in the analytical model, whereas only low frequencies are correctly treated by the numerical schemes, whilst high frequencies are responsible for spurious oscillations if the CFL-constraint is violated. We will see its consequences on the numerical results in Section 5.2.2, in the sense that the solution is forced to highly oscillate if the condition of stability is violated.

3.1.4 Modified Equation and Consistency

The stability and accuracy of finite difference/volume approximations to simple linear PDEs can also be analyzed by studying the so-called *modified equation* [37],[25]. Aside from round-off errors, the modified equation represents the actual partial differential equation solved when a numerical solution is computed using finite difference/volume methods. The modified equation is derived by first expanding each term of a difference scheme in Taylor series, and then eliminating derivatives higher than a certain order by the algebraic manipulations of the model under study. In addition to the determination of necessary and sufficient conditions for computational stability, a truncated version of the modified equation can be used to gain insight into the nature of both dissipative and dispersive errors.

As we will see in this section, one has to be rather careful in using the modified equation for the numerical issues of second order equations.
For this purpose, we refer to the finite difference explicit scheme (3.6) for the one-dimensional homogeneous heat equation, and we substitute into the discrete equation the truncated Taylor's expansions of the exact solution at the grid points (t^n, x_i). By identifying $u_i^n = u(t^n, x_i)$, with some abuse of notation for the sake of readability, we obtain

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{u_i^n + \Delta t \big(u_t\big)_i^n + \frac{\Delta t^2}{2}\big(u_{tt}\big)_i^n + O(\Delta t^3) - u_i^n}{\Delta t}\,,$$

together with

$$\begin{aligned}
&\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} \\
&= \frac{u_i^n + \Delta x \big(u_x\big)_i^n + \frac{\Delta x^2}{2}\big(u_{xx}\big)_i^n + \frac{\Delta x^3}{3!}\big(u_{xxx}\big)_i^n + \frac{\Delta x^4}{4!}\big(u_{xxxx}\big)_i^n + O(\Delta x^5)}{\Delta x^2} \\
&- \frac{2u_i^n}{\Delta x^2} + \frac{u_i^n - \Delta x \big(u_x\big)_i^n + \frac{\Delta x^2}{2}\big(u_{xx}\big)_i^n - \frac{\Delta x^3}{3!}\big(u_{xxx}\big)_i^n + \frac{\Delta x^4}{4!}\big(u_{xxxx}\big)_i^n + O(\Delta x^5)}{\Delta x^2}\,,
\end{aligned}$$

so that we deduce, dropping the sub/super-scripts for simplicity,

$$u_t + \frac{\Delta t}{2} u_{tt} + O(\Delta t^2) - a\Big[u_{xx} + \frac{\Delta x^2}{12} u_{xxxx} + O(\Delta x^3)\Big] = 0\,. \qquad (3.24)$$

We derive the equation (3.1) once for time and twice for space, to obtain

$$u_{tt} = a\,u_{xxt}\,, \quad u_{txx} = a\,u_{xxxx} \quad \Longrightarrow \quad u_{tt} = a^2 u_{xxxx}\,,$$

through which substituting in (3.24) we have

$$u_t - a\,u_{xx} + \Big(a^2\frac{\Delta t}{2} - a\frac{\Delta x^2}{12}\Big)u_{xxxx} + O(\Delta t^2, \Delta x^3) = 0\,. \tag{3.25}$$

It seems reasonable to expect instabilities when the coefficient of the fourth order derivative becomes positive, so the condition for stability coming from the analysis of the modified equation would be $\Delta t \leq \frac{\Delta x^2}{6a}$, which is by far more restrictive than the condition $\Delta t \leq \frac{\Delta x^2}{2a}$ obtained by the Von Neumann analysis. Then, the modified equation analysis may have no use for parabolic problems, although it remains relevant for hyperbolic problems [22].
However, it is necessary to perform the above computations in order to characterize the (local) *pointwise consistency* of the method, therefore from (3.24) and (3.25) we conclude that the explicit scheme is consistent to the order 1 in time and order 2 in space.

As we have seen in (3.20), the amplification factor G is real and instability occurs first for the most oscillatory mode $\Delta x\,\xi \simeq \pi$, when the amplification factor becomes less than -1. In [25], the *dilemma* of the modified equation is justified by taking $\Delta x\,\xi = \pi - \Delta x\,\xi'$ for the Fourier modes in the range $|\Delta x\,\xi| \leq \pi$, where $|\Delta x\,\xi'|$ is small for the most oscillatory modes (the *power expansions* of this quantity correspond to expansions in spatial derivatives of the amplitudes of the oscillatory modes). Under this hypothesis, we split the solution u into a smooth part u^s and an oscillatory part u^o, and then $u_i^n = (u^s)_i^n + (-1)^{i+n}(u^o)_i^n$. We focus only on the oscillatory modes and we take out the common factor $(-1)^{i+n}$, obtaining from (3.6)

$$-(u^o)_i^{n+1} = (u^o)_i^n + a\lambda\Big[-(u^o)_{i+1}^n - 2(u^o)_i^n - (u^o)_{i-1}^n\Big].$$

We compute the modified equation for the right-hand side,

$$-(u^o)_i^{n+1} = (1-2a\lambda)(u^o)_i^n - 2a\lambda\Big[(u^o)_i^n + \frac{\Delta x^2}{2}(u^o_{xx})_i^n + \frac{\Delta x^4}{4!}(u^o_{xxxx})_i^n + O(\Delta x^5)\Big],$$

and we estimate the time derivative of the oscillatory term at the grid points using the last equation,

$$\begin{aligned}(u^o)_t \simeq \frac{\big[(u^o)_i^{n+1} - (u^o)_i^n\big]}{\Delta t} &= \frac{1}{\Delta t}(4a\lambda - 2)(u^o)_i^n \\ &+ \frac{2a\lambda}{\Delta t}\Big[\frac{\Delta x^2}{2}(u^o_{xx})_i^n + \frac{\Delta x^4}{4!}(u^o_{xxxx})_i^n + O(\Delta x^5)\Big].\end{aligned}$$

Looking at the coefficient of $(u^o)_i^n$, we can observe exponential growth when $\lambda > \frac{1}{2a}$, otherwise the oscillations are damped out exponentially in time, which is the same condition obtained by the Von Neumann analysis for the explicit scheme performed in the previous section.

3.1.5 Discrete Maximum Principle

We enunciate in the following theorem the discrete maximum principle of θ-methods for the one-dimensional homogeneous heat equation, to determine conditions on the numerical parameters for which its validity is guaranteed.

Theorem 10. *Let u_i^n be the approximate solution generated by the θ-method under the condition $(1-\theta)\lambda < \frac{1}{2a}$. Let $u_{min} = \min\{u_0^i\,, i = 0, 1, \ldots, N_x; 0\}$ and $u_{max} = \max\{u_0^i\,, i = 0, 1, \ldots, N_x; 0\}$, with $u_0^i = u_0(x_i)$ the approximate initial data. Then, it holds*

$$u_{min} \,\leq\, u_i^n \,\leq\, u_{max}\,, \quad \forall\, n \geq 0\,,\; i = 0, 1, \ldots, N_x\,. \tag{3.26}$$

Proof. We rewrite 3.13 as

$$\begin{aligned}(1 + 2a\theta\lambda)u_i^{n+1} = \big[1 - 2a(1-\theta)\lambda\big]u_i^n \\ + a(1-\theta)\lambda\big(u_{i+1}^n + u_{i-1}^n\big) + a\,\theta\lambda\big(u_{i+1}^{n+1} + u_{i-1}^{n+1}\big)\,.\end{aligned}$$

Under the hypothesis $(1-\theta)\lambda < \frac{1}{2a}$, all the coefficients of the right-hand side of this equation are positive and their sum is equal to $(1 + 2a\,\theta\lambda)$.
We suppose that the maximum is attained at the interior point (x_i, t^{n+1}), and we denote by $u^* = \max\{u_{i-1}^n, u_{i+1}^n, u_i^n, u_{i-1}^{n+1}, u_{i+1}^{n+1}\}$, then $u_{max} \leq u^*$. But, by definition of the maximum, we have that $u_{max} = u^*$, so the maximum must be attained at all the six points (since the coefficients are strictly positive). We can reiterate this argument until reaching the boundaries at $i = 0$ or $i = N_x$, and coming back in time until the initial data. So, we have demonstrated the right-hand side of (3.26). For the left-hand inequality, the procedure is the same. □

We can extend the above result to more general cases, for example with $u(t, a_x) = g_1(t)$ and $u(t, b_x) = g_2(t)$, $\forall\, t \geq 0$. Thus, we would have that

$$\begin{aligned}u_{max} &= \max\big\{u_0^i\,, i = 0, 1, \ldots, N_x\,;\, g_1^n,\, g_2^n,\, n = 1, 2, \ldots, N_t\big\}\,,\\ u_{min} &= \min\big\{u_0^i\,, i = 0, 1, \ldots, N_x\,;\, g_1^n,\, g_2^n,\, n = 1, 2, \ldots, N_t\big\}\,,\end{aligned}$$

where g_1^n and g_2^n are numerical approximations of the boundary conditions, $g_1^n = g_1(t^n)$ and $g_2^n = g_2(t^n)$, respectively.

The condition $(1-\theta)\lambda < \frac{1}{2a}$ is more restrictive than that for the stability, except for the explicit scheme ($\theta = 0$) where $\lambda \leq \frac{1}{2a}$ corresponds also to the condition under which the discrete maximum principle holds. The implicit scheme ($\theta = 1$) satisfies the discrete maximum principle for all λ, while the Crank-Nicolson scheme holds if $\lambda \leq \frac{1}{a}$. Obviously, another important concept is the non-singularity of the matrix, because this guarantees the existence of a unique solution to the algebraic system defining all linear schemes : for the explicit scheme (3.7) it holds for $\lambda \neq \frac{1}{2a}$, for the implicit

	L^2-stability	maximum/minimum principle	nonsingular matrix
explicit	$\Delta t \leq \frac{\Delta x^2}{2a}$	$\Delta t \leq \frac{\Delta x^2}{2a}$	$\frac{\Delta t}{\Delta x^2} \neq \frac{1}{2a}$
Crank-Nicolson	$\forall\ \Delta t, \Delta x$	$\Delta t \leq \frac{\Delta x^2}{a}$	$\frac{\Delta t}{\Delta x^2} \neq \frac{1}{a}$
implicit	$\forall\ \Delta t, \Delta x$	$\forall\ \Delta t, \Delta x$	$\forall\ \Delta t, \Delta x$

Table 3.1: well-posedness requirements for the discrete heat equation

scheme the matrix (3.9) is always nonsingular, and for the Crank-Nicolson scheme (3.12) it holds for $\lambda \neq \frac{1}{a}$. We summarize these results in Table (3.1).

Now, we analyze the matrix-form $U^{n+1} = M \cdot U^n$, $M \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$, of the numerical schemes introduced in Section 3.1.1, to recover the results previously stated by checking the properties in Table 2.5.

For the explicit three-points scheme, we observe that the matrix (3.7) is quadratic and tri-diagonal, symmetric, nonsingular if $\lambda \neq \frac{1}{2a}$ and positive if $\lambda < \frac{1}{2a}$ (all its entries are strictly positive), which implies that M_{exp} is irreducible [28]. Moreover, the sum of elements on the rows is equal to 1, except for the rows including the boundary conditions, i.e. the first and the last one. So, all the properties listed in Table 2.5 are satisfied and, according to the analysis of Theorem 8, this method satisfies the maximum principle under the CFL-condition $\lambda < \frac{1}{2a}$ (because the method is also L^2-stable), as reported in Table 3.1.

For the implicit scheme, we consider the matrix of the formulation (3.10), which is the inverse of the nonsingular matrix (3.9). Therefore, we have that M_{imp}^{-1} is tri-diagonal, symmetric and nonnegative for all λ, thus verifying the sufficient conditions required in Table 3.1. We observe that the matrix M_{impl} in (3.9) verifies the *conservation property*, but this is not necessarily true for its inverse. As a matter of fact, for properties of the inverse matrix M_{imp}^{-1}, we have to resort to the theory of *M-matrices* stated in [28].
Indeed, from (3.9) we obtain

$$M_{imp} = (1 + 2a\lambda)\mathbb{I} - a\lambda\,\mathbb{B} \ \in\ \mathbb{R}^{(N_x-1)\times(N_x-1)},$$

where $\mathbb{I}$ denotes the identity matrix and $\mathbb{B} = diag(\mathrm{d}, -1) + diag(\mathrm{d}, 1)$, with $\mathrm{d} = ones(N_x - 2)$, which has spectral radius strictly less than 1, independently from its size. Therefore, M_{imp} is a nonsingular M-matrix, symmetric and positive definite, which implies that its inverse M_{imp}^{-1} exists and it is a positive matrix (so M_{imp} is also said to be *monotone*), i.e. all the entries of M_{imp}^{-1} are positive, thus ensuring the positivity of the numerical scheme.
According to [23], since the matrix M_{imp} has rows with nonnegative sum, the implicit scheme satisfies the discrete maximum principle.

For the Crank-Nicolson scheme (3.12), the situation is more complicated, as we have to compute the matrix $M_1^{-1} \cdot M_2$, but it is guaranteed that M_1 is

invertible (and, moreover, the scheme is contractive). Nevertheless, the analysis of the matrix does not allow neither checking the results obtained with the Von Neumann stability analysis nor adding significant observations...

3.2 The heterogeneous linear case

The simplest model of diffusion in heterogeneous media has been introduced in Section 2.3, which can be effectively generalized to anisotropic operators in higher dimensions, and it is given by

$$u_t = \big(a(x)u_x\big)_x = a'(x)\,u_x + a(x)\,u_{xx}\,, \tag{3.27}$$

with a nonnegative function $a(x)$ for the correct definition of one-dimensional parabolic equations. The diffusion coefficient depending on x, the density u feels it in a different way according to the physical position.

Remark 11. *The right-hand side of equation* (3.27) *is meaningful only for enough regularity of the solution, whereas the conservation form on the left-hand side has the advantage to be analytically relevant also for discontinuous data (both the solution and the diffusion coefficient). Indeed, the theory developed in [3] generally holds for bounded data, in the integral/weak sense.*

We consider the following initial data and Neumann boundary conditions,

$$\begin{aligned} &u(0,x) = u_0(x) \qquad \text{for } \; x \in \Omega = [a_x, b_x]\,, \\ &\frac{\partial u}{\partial n}\Big|_{\partial\Omega} = 0 \qquad\quad \text{for } \; t \in (0,T)\,. \end{aligned}$$

We discretize the computational domain as in (3.3), and the Neumann boundary conditions in the one-dimensional case can be rewritten as

$$\begin{aligned} 0 = u_x(t,x_0) \simeq \frac{u(t,x_0+\Delta x) - u(t,x_0)}{\Delta x} \quad &\Longrightarrow \quad u(t,x_0) \simeq u(t,x_1)\,, \\ 0 = u_x(t,x_{N_x}) \simeq \frac{u(t,x_{N_x}) - u(t,x_{N_x}-\Delta x)}{\Delta x} \quad &\Longrightarrow \quad u(t,x_{N_x}) \simeq u(t,x_{N_x-1})\,, \end{aligned}$$

where $x_0 = a_x$ and $x_{N_x} = b_x$ are the initial and final points of the grid. From the numerical point of view, the previous formulas translate into

$$u(t^n,x_0) = u(t^n,x_1)\,, \quad u(t^n,x_{N_x}) = u(t^n,x_{N_x-1})\,, \tag{3.28}$$

for all $n = 0, 1, \ldots, N_t$, also known as *no-flux boundary conditions*.

To construct an approximation of the solution u to (3.27), we compute the values at the grid points $u_i^n \simeq u(t^n,x_i)$ and we consider a discretization of the diffusion function through $a_i = a(x_i)$. We have already seen in Section 2.5.2 that the simple finite difference Chain Rule method generally fails to satisfy the discrete maximum principle, so it is not useful in practice.

According to (2.31) in Section 2.4.2, the finite difference Standard Discretization method built on a staggered grid, and readapted to the one-dimensional framework, becomes

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}=\frac{1}{\Delta x}\left(\frac{a_i+a_{i+1}}{2}\cdot\frac{u_{i+1}^n-u_i^n}{\Delta x}-\frac{a_{i-1}+a_i}{2}\cdot\frac{u_i^n-u_{i-1}^n}{\Delta x}\right),$$

or rather

$$u_i^{n+1}=\alpha_i^1\frac{\Delta t}{2\Delta x^2}\,u_{i-1}^n+\Big(1-\alpha_i^2\frac{\Delta t}{2\Delta x^2}\Big)u_i^n+\alpha_i^3\frac{\Delta t}{2\Delta x^2}\,u_{i+1}^n\,, \tag{3.29}$$

where, for $i=1,2,\dots,N_x-1$, we have set

$$\alpha_i^1=a_{i-1}+a_i\,,\quad \alpha_i^2=a_{i+1}+2a_i+a_{i-1}\,,\quad \alpha_i^3=a_i+a_{i+1}\,, \tag{3.30}$$

and the boundary values are computed as $u_0^n=u_1^n$ and $u_{N_x}^n=u_{N_x-1}^n$, for all $n=1,2,\dots,N_t$, according to (3.28).

Because of the typical conservation/divergence form of equation (3.27), we can propose an alternative derivation of the above scheme by applying the finite volume method, following the arguments in Section 2.3.
Finally, we get precisely (3.29)-(3.30) also with the Nonnegative Method derived in Section 2.4.3, since the nonnegative modification eventually concerns only the mixed derivatives, which are not present in this case.

Now, we rewrite the numerical scheme (3.29) in compact matrix-form,

$$U^{n+1}=M\cdot U^n, \tag{3.31}$$

with $U^n=(u_1^n,u_2^n,\dots,u_{N_x-1}^n)\in\mathbb{R}^{N_x-1}$ and M the $(N_x-1)\times(N_x-1)$ tri-diagonal matrix with entries (3.30), namely

$$M=\left[\begin{array}{c} diag\Big(\dfrac{\Delta t}{2\Delta x^2}\alpha_i^3\,,+1\Big)_{1\le i\le N_x-2}\,; \\ diag\Big(1-\dfrac{\Delta t}{2\Delta x^2}\alpha_i^2\Big)_{1\le i\le N_x-1}\,; \\ diag\Big(\dfrac{\Delta t}{2\Delta x^2}\alpha_i^1\,,-1\Big)_{2\le i\le N_x-1} \end{array}\right], \tag{3.32}$$

where $diag(\cdot)$ denotes the principal diagonal of M, while $diag(\cdot,-1)$ and $diag(\cdot,+1)$ indicate its lower and upper first diagonals, respectively, so that

$$M=\begin{bmatrix} 1-\frac{\Delta t}{\Delta x^2}\frac{a_0+2a_1+a_2}{2} & \frac{\Delta t}{\Delta x^2}\frac{a_1+a_2}{2} & 0 & \cdots \\ \frac{\Delta t}{\Delta x^2}\frac{a_1+a_2}{2} & 1-\frac{\Delta t}{\Delta x^2}\frac{a_1+2a_2+a_3}{2} & \frac{\Delta t}{\Delta x^2}\frac{a_2+a_3}{2} & \cdots \\ 0 & \frac{\Delta t}{\Delta x^2}\frac{a_2+a_3}{2} & 1-\frac{\Delta t}{\Delta x^2}\frac{a_2+2a_3+a_4}{2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Let us verify the properties listed in Table (2.5) for the fully discrete problem (3.31) with matrix M given above : *symmetry* is obvious from the definition, and also *conservation* as the sum of the elements on each row is equal to 1 (except for the first and the last component); the *non-negativity* of the off-diagonal entries is ensured from $a_i > 0$ for all $0 \le i \le N_x$, so that $\frac{\Delta t}{2\Delta x^2}\alpha_i^1 > 0$, $2 \le i \le N_x - 1$, and $\frac{\Delta t}{2\Delta x^2}\alpha_i^3 > 0$, $1 \le i \le N_x - 2$, while for the diagonal elements we ask for $(1 - \frac{\Delta t}{2\Delta x^2}\alpha_i^2) \ge 0$, $1 \le i \le N_x - 1$, namely

$$\frac{\Delta t}{\Delta x^2} \cdot \frac{a_{i-1} + 2a_i + a_{i+1}}{2} \le 1 \,, \quad \forall\, i = 1, 2, \ldots, N_x - 1 \,. \tag{3.33}$$

We define

$$a_{\max} := \max_{0 \le i \le N_x} a_i$$

and we substitute into (3.33) to obtain the following *sufficient condition* :

$$\frac{\Delta t}{\Delta x^2} \frac{a_{i-1} + 2a_i + a_{i+1}}{2} \le \frac{\Delta t}{2\Delta x^2} a_{\max} \le 1 \quad \Longrightarrow \quad \Delta t \le \frac{\Delta x^2}{2\, a_{\max}} \,, \tag{3.34}$$

which generalizes the CFL-condition of the explicit scheme for homogeneous heat equation stated in Table 3.1, and it has been already derived in (2.51). Therefore, all the theoretical hypotheses of Theorem 8 being verified, the validity of the discrete maximum principle is proven.

We conclude this section by establishing the stability of the scheme presented above. We recall the matrix-form (3.31) and we want that

$$||U^{n+1}||_* \le ||M||\, ||U^n||_* \,, \qquad \forall\, n = 1, 2, \ldots \,,$$

with $||M|| \le 1$, for some *operator norm* defined by

$$||M|| = \sup_{U \in \mathbb{R}^m} \frac{||M \cdot U||_*}{||U||_*} \,, \quad m = N_x - 1 \,,$$

where $||\cdot||_*$ is the underlying vectorial norm.
In particular, when $||\cdot||_* = ||\cdot||_\infty$ the following identity holds

$$||M||_\infty = \max_{1 \le j \le m} \sum_{i=1}^{m} |m_{ij}| \,,$$

where i and j are indicators for the rows and the columns of M, respectively, and M is a $m \times m$ matrix (refer to [30]).

Since M in (3.32) is symmetric, with the sum of rows equal to 1 (leading to a convex combination of the vector elements), and moreover nonnegative if (3.34) holds, we have that $||M||_\infty \le 1$, and so $||U^{n+1}||_\infty \le ||U^n||_\infty$.
This is perfectly consistent with the requirement of the discrete maximum principle, because at every fixed time t^{n+1} the numerical solution remains

bounded in l^∞-norm with respect to time t^n. Moreover, we have stability also in the l^2-norm, due to the equivalence of the norms in $\mathbb{R}^m$. Indeed, requiring stability in l^∞-norm is a stronger statement than in l^2-norm, and so the validity of the maximum principle implies also the l^2-stability. We thus conclude that the scheme is stable if the condition (3.34) is satisfied.

Remark 12. *The type of analysis performed in this section can be extended to other cases, according to a general protocol : if we have a matrix-form in whose diagonal elements are less than* 1*, and it holds the convex combination of its row elements, this assures the overall stability of the method.*

3.3 Other two scalar heterogeneous models

For a positive function varying on the space, we consider the diffusion model

$$u_t = a(x)u_{xx} = \big(a(x)u_x\big)_x - a'(x)\,u_x\,, \tag{3.35}$$

which is not a conservation law (there is no way to rewrite equation (3.35) in conservation form). However, it can be interpreted as an approximation of a diffusion equation (3.27) where $a(x)$ has very weak slope $a'(x) \simeq 0$, i.e. the function a is almost constant.

For the discretization of the time derivative, we use the same explicit scheme calculated in the previous cases, while for the second order term the finite difference and the finite volume approach gives exactly the same discretization : indeed, using (2.8) we have

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = a_i \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2}\,,$$

which can be rewritten as

$$u_i^{n+1} = \big(\frac{\Delta t}{\Delta x^2}a_i\big)u_{i+1}^n + \big(1 - 2\frac{\Delta t}{\Delta x^2}a_i\big)u_i^n + \big(\frac{\Delta t}{\Delta x^2}a_i\big)u_{i-1}^n\,,$$

and the same expression is deduced from the discrete integral form

$$\frac{1}{\Delta t}\int_{t^n}^{t_{n+1}} \left(\frac{1}{\Delta x}\int_{C_i} a(x)u_{xx}\,dx\right) dt \simeq \frac{a_i}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_{xx}(t^n, x)\,dx$$

$$= \frac{a_i}{\Delta x}\Big[u_x(t^n, x_{i+\frac{1}{2}}) - u_x(t^n, x_{i-\frac{1}{2}})\Big] \simeq \frac{a_i}{\Delta x}\left(\frac{u_{i+1}^n - u_i^n}{\Delta x} - \frac{u_i^n - u_{i-1}^n}{\Delta x}\right).$$

We compute the matrix-form of the scheme, with the tri-diagonal matrix

$$M = \Bigg[\qquad\qquad diag\Big(\frac{\Delta t}{\Delta x^2}a_i\,, +1\Big)_{1\le i\le N_x-2}\,;$$

$$diag\Big(1 - 2\frac{\Delta t}{\Delta x^2}a_i\Big)_{1\le i\le N_x-1}\,;$$

$$diag\Big(\frac{\Delta t}{\Delta x^2}a_i\,, -1\Big)_{2\le i\le N_x-1} \qquad\qquad \Bigg],$$

so that

$$M = \begin{bmatrix} 1-2\frac{\Delta t}{\Delta x^2}a_1 & \frac{\Delta t}{\Delta x^2}a_1 & 0 & \cdots \\ \frac{\Delta t}{\Delta x^2}a_2 & 1-2\frac{\Delta t}{\Delta x^2}a_2 & \frac{\Delta t}{\Delta x^2}a_2 & \cdots \\ 0 & \frac{\Delta t}{\Delta x^2}a_3 & 1-2\frac{\Delta t}{\Delta x^2}a_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Concerning the properties in Table 2.5, we have the conservation (the sum of the rows is equal to 1) and the non-negativity of all entries if

$$2\frac{\Delta t}{\Delta x^2}a_i < 1\,, \quad \forall\, i = 1, 2, \ldots, N_x - 1 \quad \Longrightarrow \quad \Delta t < \frac{\Delta x^2}{2\, a_{\max}}\,, \tag{3.36}$$

with $a_{\max} = \max_{1\le i\le N_x-1} a_i$, which is the same condition as in (3.34). Therefore, the validity of the discrete maximum principle is satisfied. The only missed property is the *symmetry* and this can explained as we are not dealing with a proper conservation form.

Finally, we consider the one-dimensional parabolic equation

$$\begin{aligned} u_t = \big(a(x)\, u\big)_{xx} &= \big(a'(x)u + a(x)u_x\big)_x \\ &= a''(x)u + 2a'(x)u_x + a(x)u_{xx}\,, \end{aligned} \tag{3.37}$$

which is a conservation law (2.21) with flux $F(t,x) = \big(a(x)\, u\big)_x$ belonging to the class of porous media equations [8],[36], then it is the heterogenous linear version of (3.14) with $\Phi(x\,;u) = a(x)\, u$, so that (3.16) is satisfied.

Many references in the literature are devoted to the two-dimensional case, also adding a reaction term for describing more general phenomena,

$$\partial_t u - \Delta\Phi = f(u)\,, \quad \Phi = \Phi(x, y\,; u)\,, \qquad (x, y) \in \Omega \subset \mathbb{R}^2,$$

that is common for modeling the diffusion in porous media, under the structural assumption that $\Phi : \Omega \times \mathbb{R}^+ \to \mathbb{R}$, so that the above equation can be rewritten as

$$\partial_t u - \nabla \cdot \big(\nabla\Phi + \Phi'\nabla u\big) = f(u)\,,$$

with the first term in parenthesis corresponding to the contribution of *transport* and the second term reproducing a nonlinear heterogeneous but *purely isotropic* diffusion for $\Phi' \in \mathbb{R}^+$. Therefore, there is no way to rewrite the anisotropic model (1.1) in such form : in comparison with isotropic nonlinear diffusion, anisotropic tensors dare to reproduce the (averaged) choice of motion of the agents composing the density. Certainly, we could also include a nonlinearity in the model (1.1) by assuming that $A = A(x, y\,; u)$ with $A : \Omega \times \mathbb{R}^+ \to \mathbb{R}^{2\times 2}$, for example to describe the effects of crowding on the spatial domain (higher density values tends to reduce the diffusivity coefficients).

Despite the fact that models like (3.14) are not of interest in this report, we briefly derive finite difference/volumes schemes for (3.37), using the staggered grid in Figure 2.2 mainly because of the conservation form. Therefore, from (3.15) we have

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{a_{i+1}u_{i+1}^n - 2a_i u_i^n + a_{i-1}u_{i-1}^n}{\Delta x^2}\,, \tag{3.38}$$

or rather

$$u_i^{n+1} = \big(\frac{\Delta t}{\Delta x^2}\,a_{i+1}\big)u_{i+1}^n + \big(1 - 2\frac{\Delta t}{\Delta x^2}\,a_i\big)u_i^n + \big(\frac{\Delta t}{\Delta x^2}\,a_{i-1}\big)u_{i-1}^n\,,$$

which again can be reformulated in terms of the finite volume approach,

$$\begin{aligned}
&\frac{1}{\Delta t}\int_{t^n}^{t_{n+1}}\Big(\frac{1}{\Delta x}\int_{C_i}\big(F(t,x)\big)_x\,dx\Big)\,dt \simeq \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\big(F(t^n,x)\big)_x\,dx\\
&= \frac{1}{\Delta x}\left[F(t^n,x)\Big|_{x_{i+\frac{1}{2}}} - F(t^n,x)\Big|_{x_{i-\frac{1}{2}}}\right]\\
&\simeq \frac{1}{\Delta x}\left[\frac{a_{i+1}\,u_{i+1}^n - a_i\,u_i^n}{\Delta x} - \frac{a_i\,u_i^n - a_{i-1}\,u_{i-1}^n}{\Delta x}\right]
\end{aligned}$$

where we have set $F(t^n,x) = \big(a(x)\,u(t^n,x)\big)_x$.
As usual, we compute the matrix-form of the scheme with

$$\begin{aligned}
M = \Bigg[& \qquad\qquad diag\Big(\frac{\Delta t}{\Delta x^2}\,a_{i+1}\,,+1\Big)_{1\le i\le N_x-2};\\
& diag\Big(1 - 2\frac{\Delta t}{\Delta x^2}\,a_i\Big)_{1\le i\le N_x-1};\\
& diag\Big(\frac{\Delta t}{\Delta x^2}\,a_{i-1}\,,-1\Big)_{2\le i\le N_x-1} \qquad\qquad \Bigg],
\end{aligned}$$

so that

$$M = \begin{bmatrix} 1-2\frac{\Delta t}{\Delta x^2}a_1 & \frac{\Delta t}{\Delta x^2}a_2 & 0 & \cdots\\ \frac{\Delta t}{\Delta x^2}a_1 & 1-2\frac{\Delta t}{\Delta x^2}a_2 & \frac{\Delta t}{\Delta x^2}a_3 & \cdots\\ 0 & \frac{\Delta t}{\Delta x^2}a_2 & 1-2\frac{\Delta t}{\Delta x^2}a_3 & \cdots\\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

We note that this matrix has the same diagonal of the previous one, so we find the same condition for the non-negativity of the entries,

$$\Delta t \le \frac{\Delta x^2}{2\,a_{\max}}\,, \qquad a_{\max} := \max_{1\le i\le N_x-1} a_i\,.$$

Moreover, it is easy to check that the other properties in Table 2.5 are not satisfied, and thus an l^∞-stability cannot be guaranteed.

Chapter 4

The two-dimensional Anisotropic and Heterogeneous case

The main target of this report is to present the problems arising when dealing with higher order anisotropic operators, because the interaction of diffusions spreading differently along different directions may actually lead to complex and interesting phenomena (see [32],[35],[14],[31],[27],[15], for instance).
We restrict to the two-dimensional case, that is motivated by several applications to real systems, and we consider the equation (1.1) with an anisotropic and heterogeneous diffusion tensor (1.2), which is also symmetric, namely

$$\partial_t u = \nabla \cdot \left(\begin{bmatrix} a(x,y) & c(x,y) \\ c(x,y) & b(x,y) \end{bmatrix} \nabla u \right), \qquad t \in \mathbb{R}^+,\ (x,y) \in \mathbb{R}^2, \tag{4.1}$$

for $u(t;x,y) \in \mathbb{R}^+$, under the uniform *parabolicity condition* given by (1.7).

4.1 Diagonal diffusion tensors

In the simplest case, the directions where the material spreads faster are parallel to the x and y axes, i.e. $c = 0$, although most of the times this is not the case. The model of diagonal diffusion is however a good representation for several physical situations, for example when the main diffusion directions are locally orthogonal and rotated by some known angle ϑ from the x-axis, as Figure 4.1 shows.
Let us denote this direction by ξ and its normal by η, and we express them with respect to x, y and through ϑ as

$$\xi = x\,\cos\vartheta + y\,\sin\vartheta\,, \qquad \eta = -x\,\sin\vartheta + y\,\cos\vartheta\,,$$

so the variables x and y can be derived from

$$x = \xi\,\cos\vartheta - \eta\,\sin\vartheta\,, \qquad y = \xi\,\sin\vartheta + \eta\,\cos\vartheta\,.$$

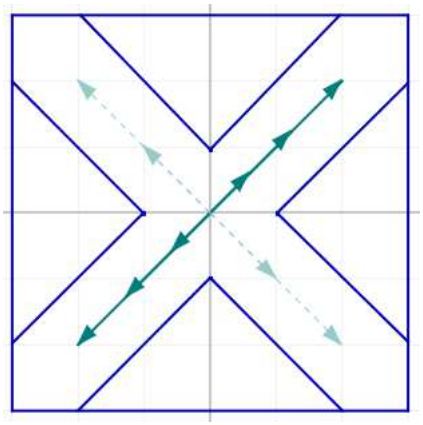


Figure 4.1: example of orthogonal non-Cartesian diffusion directions

Within the (ξ, η)-referential frame, we can write the diffusion term as

$$\nabla_{(\xi,\eta)} \cdot \Big[A(\xi,\eta) \cdot \nabla_{(\xi,\eta)} u(t;\xi,\eta) \Big], \tag{4.2}$$

where A is a 2×2 diagonal matrix of the form

$$A = \begin{bmatrix} a(\xi,\eta) & 0 \\ 0 & b(\xi,\eta) \end{bmatrix}. \tag{4.3}$$

In order reproduce the diffusion operator (4.2)-(4.3) in Cartesian coordinates, we expand and we substitute the derivatives calculated in terms of x and y , that are

$$\begin{aligned}
u_\xi &= \frac{\partial u}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial \xi} = u_x \cos\vartheta + u_y \sin\vartheta\,, \\
u_\eta &= \frac{\partial u}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial \eta} = -u_x \sin\vartheta + u_y \cos\vartheta\,, \\
u_{\xi\xi} &= \frac{\partial}{\partial \xi}(u_x \cos\vartheta + u_y \sin\vartheta) \\
&= \big(\frac{\partial u_x}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial u_x}{\partial y}\frac{\partial y}{\partial \xi}\big)\cos\vartheta + \big(\frac{\partial u_y}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial u_y}{\partial y}\frac{\partial y}{\partial \xi}\big)\sin\vartheta \\
&= \big(u_{xx}\cos\vartheta + u_{xy}\sin\vartheta\big)\cos\vartheta + \big(u_{xy}\cos\vartheta + u_{yy}\sin\vartheta\big)\sin\vartheta \\
&= u_{xx}\cos^2\vartheta + 2\,u_{xy}\sin\vartheta\cos\vartheta + u_{yy}\sin^2\vartheta \\
&= u_{xx}\cos^2\vartheta + u_{xy}\sin(2\vartheta) + u_{yy}\sin^2\theta\,, \\
u_{\eta\eta} &= \frac{\partial}{\partial \eta}(-u_x \sin\vartheta + u_y \cos\vartheta) \\
&= -\big(-u_{xx}\sin\vartheta + u_{xy}\cos\vartheta\big)\sin\vartheta + \big(-u_{xy}\sin\vartheta + u_{yy}\cos\vartheta\big)\cos\vartheta \\
&= u_{xx}\sin^2\vartheta - u_{xy}\sin(2\vartheta) + u_{yy}\cos^2\vartheta\,,
\end{aligned}$$

where we used the classical equality $2\sin\vartheta\cos\vartheta = \sin(2\vartheta)$, so that we obtain

$$\begin{aligned}
&\big(a(\xi,\eta)\,u_\xi\big)_\xi + \big(b(\xi,\eta)\,u_\eta\big)_\eta = a\,u_{\xi\xi} + a_\xi\,u_\xi + b_\eta\,u_\eta + b\,u_{\eta\eta} \\
&= \big(a_x\cos^2\vartheta + b_x\sin^2\vartheta\big)u_x + \frac{a_y - b_y}{2}\sin(2\vartheta)u_x \\
&\quad + \big(a\cos^2\vartheta + b\sin^2\vartheta\big)u_{xx} + (a-b)\sin(2\vartheta)u_{xy} \\
&\quad + \frac{a_x - b_x}{2}\sin(2\vartheta)u_y + \big(a_y\sin^2\vartheta + b_y\cos^2\vartheta\big)u_y \\
&\quad + \big(b\cos^2\vartheta + a\sin^2\vartheta\big)u_{yy} \\
&= \nabla_{(x,y)}\cdot\left(\begin{bmatrix}\alpha(x,y) & \gamma(x,y)\\ \gamma(x,y) & \beta(x,y)\end{bmatrix}\cdot\nabla_{(x,y)}u\right),
\end{aligned} \tag{4.4}$$

where identify the entries of the new matrix as following,

$$\alpha = a\cos^2\vartheta + b\sin^2\vartheta\,, \quad \beta = a\sin^2\vartheta + b\cos^2\vartheta\,, \quad \gamma = \frac{a-b}{2}\sin(2\vartheta)\,.$$

We note that the matrix in the (x,y)-coordinates is symmetric and positive definite (it is actually the counter-diagonalization of the original matrix, so with the same eigenvalues) : indeed, we have $\alpha,\beta > 0$ and $\gamma^2 < \alpha\beta$ since

$$\begin{aligned}
\gamma^2 - \alpha\beta &= (a-b)^2\sin^2\vartheta\cos^2\vartheta \\
&\quad - (a^2+b^2)\sin^2\vartheta\cos^2\vartheta - a\,b\,(\cos^4\vartheta + \sin^4\vartheta) \\
&= -2\,a\,b\,\cos^2\vartheta\sin^2\vartheta - a\,b\,(\cos^4\vartheta + \sin^4\vartheta) < 0\,,
\end{aligned}$$

$\forall\, a,b > 0$, $\forall\,\vartheta$, so it is a good candidate for being a diffusion tensor, the rotation of coordinates preserving the positive definiteness.
For the isotropic case $a = b$, we obtain a *circular diffusion* with no preferred direction, so the rotation of coordinates does not make sense in that context since the transformed matrix must coincide with the original one.

As an example, suppose that we have a preferential diffusion rotated with an angle $\vartheta = 45°$, and another crossing perpendicularly as in Figure 4.1. To model such a situation we have to correctly choose the diffusion functions a, b and c. If we fix a constant $d > 0$ and we choose the data

$$a(x,y) = \begin{cases} d & \text{for } y = x \\ 0 & \text{otherwise}\end{cases}, \quad b(x,y) = \begin{cases} d & \text{for } y = 1-x \\ 0 & \text{otherwise}\end{cases}, \quad c(x,y) = 0\,,$$

as a matter of fact, we state that the direction $y = x$ has faster propagation along the x-axis, and not along the angle ϑ, and the same holds for $y = 1-x$. So we are modelling the situation in Figure 4.2, which is different from the situation we want to take into account. To solve this problem, the only consistent approach is that of computing a rotation of coordinates, thus

getting the new diffusion tensor (4.4) depending on ϑ. In this way, we recover the case in which the direction of propagation is a combination along the x-axis and the y-axis. In [39], the authors propose some geometrical remarks about the case described in this section, and they also explore the issue of the interaction between Cartesian grids and non-Cartesian diffusion directions through numerical experiments.

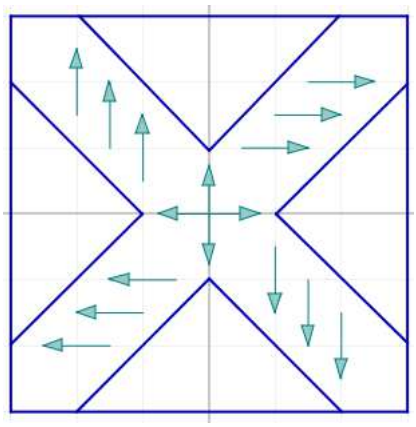


Figure 4.2: wrong diffusion along orthogonal non-Cartesian directions

4.2 Discrete maximum principle for two-dimensional problems

We will analyze in details three different cases of the model (4.1), according to the specific form of the diffusion tensor, namely

- the diagonal anisotropic and homogeneous case, involving different constant diffusions only along the Cartesian axes, with $a(x,y) = a$, $b(x,y) = b$ and $c(x,y) = 0$, so that the equation becomes

$$u_t \;=\; \nabla \cdot \big(a\,u_x\,, b\,u_y\big)^T \;=\; a\,u_{xx} + b\,u_{yy}\,. \tag{4.5}$$

- the diagonal anisotropic and heterogeneous case, describing different diffusions along the Cartesian axes which vary also depending on the space position, by means of the parabolic equation

$$u_t \;=\; \big(a(x,y)\,u_x\big)_x + \big(b(x,y)\,u_y\big)_y\,. \tag{4.6}$$

- the fully anisotropic but homogeneous case, with $c(x,y) \neq 0$, given by

$$u_t \;=\; \big(a\,u_x + c\,u_y\big)_x + \big(c\,u_x + b\,u_y\big)_y\,, \tag{4.7}$$

  which involves the mixed derivatives with all different constant diffusion coefficients.

For simplicity, we restrict to rectangular domains as in Figure 2.1 or rather as in Figure 2.2, and we consider homogeneous Dirichlet boundary conditions, which translate into the numerical framework as $u_{ij}^n = 0$ for $i \in \{0, N_x\}$ and $j \in \{0, N_y\}$, for all finite difference/volume schemes introduced in the next sections.

4.2.1 Diagonal anisotropic homogeneous diffusion

We use the notations of Section 2.2 and Section 2.3.

To construct finite difference discretizations, we consider the equation (4.5) at the grid points $(t^n; x_i, y_j)$ and we apply (2.10) and (2.14), so that we can assemble the fully discrete scheme as

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\Delta t} = a\,\frac{u_{i+1,j}^n - 2u_{ij}^n + u_{i-1,j}^n}{\Delta x^2} + b\,\frac{u_{i,j+1}^n - 2u_{ij}^n + u_{i,j-1}^n}{\Delta y^2}\,, \tag{4.8}$$

which finally looks like

$$\begin{aligned} u_{ij}^{n+1} = {} & a\frac{\Delta t}{\Delta x^2}\big(u_{i+1,j}^n + u_{i-1,j}^n\big) + \Big(1 - 2a\frac{\Delta t}{\Delta x^2} - 2b\frac{\Delta t}{\Delta y^2}\Big)u_{ij}^n \\ & + b\frac{\Delta t}{\Delta y^2}\big(u_{i,j+1}^n + u_{i,j-1}^n\big)\,, \end{aligned} \tag{4.9}$$

for $i = 1, 2, \ldots, N_x - 1$ and $j = 1, 2, \ldots, N_y - 1$.

We recover the same discretization through the finite volume method, starting from the integral averages of the equation (4.5) on the grid cells (2.1), according to (2.4). Indeed, we perform the two-dimensional analogue of (2.23) and we use (2.24) for the time derivative,

$$\begin{aligned} & \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \left(\frac{1}{\Delta x \Delta y}\int_{C_{ij}} u_t\,dx\,dy\right) dt \\ & = \frac{1}{\Delta t}\cdot\frac{1}{\Delta x\Delta y}\int_{C_{ij}} \Big[u(t^{n+1}; x, y) - u(t^n; x, y)\Big]dx\,dy \;\simeq\; \frac{u_{ij}^{n+1} - u_{ij}^n}{\Delta t}\,, \end{aligned} \tag{4.10}$$

while for the term in divergence form, we approximate the time-average with the values at t^n (forward Euler scheme) and then we apply the *divergence theorem* to obtain

$$\begin{aligned} & \frac{1}{\Delta x\Delta y}\int_{C_{ij}} \nabla\cdot\left(\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}\cdot\begin{bmatrix} u_x(t^n; x, y) \\ u_y(t^n; x, y) \end{bmatrix}\right) dx\,dy \\ & = \frac{1}{\Delta x\Delta y}\int_{\partial C_{ij}} \Big[a\,u_x(t^n, s)\,, b\,u_y(t^n, s)\Big]\cdot\vec{\nu}\,ds\,, \end{aligned} \tag{4.11}$$

with the functions expressed in local variables $s \in \partial C_{ij}$, where ∂C_{ij} denotes the boundary of the cell C_{ij}, referring to Figure 4.3,

$$\partial C_{ij} = [A, B]\big|_{y_{j-\frac{1}{2}}} \cup\ [B, C]\big|_{x_{i+\frac{1}{2}}} \cup\ [C, D]\big|_{y_{j+\frac{1}{2}}} \cup\ [A, D]\big|_{x_{i-\frac{1}{2}}} \tag{4.12}$$

and $\vec{\nu} = (\nu_k)$, $k = 1, 2, 3, 4$, collects the normal vectors at the intervals composing ∂C_{ij}. The notation $[A, B]\big|_{y_{j-\frac{1}{2}}}$ indicates, for example, all the points (x, y) where y is fixed at $y_{j-\frac{1}{2}}$ and x varies in the interval $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$.

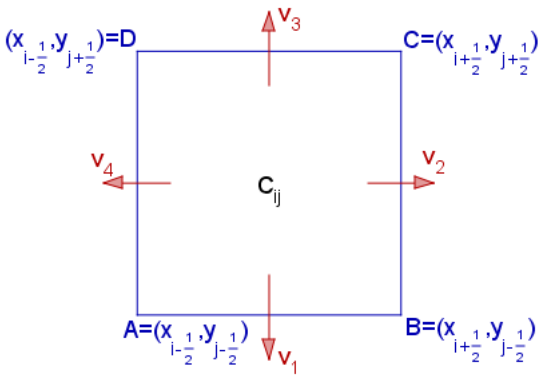


Figure 4.3: boundaries of the grid cell C_{ij}

We proceed with explicit calculations from (4.11), where the boundary integral can be decomposed into four parts according to the geometrical setting in Figure 4.3, and we further simplify the notation using $f^n(\cdot) = f(t^n, \cdot)$,

$$
\begin{aligned}
F_1 &= \frac{1}{\Delta x \Delta y} \int_A^B \left[a\, u_x^n(x, y_{j-\frac{1}{2}}), b\, u_y^n(x, y_{j-\frac{1}{2}}) \right] \cdot \begin{bmatrix} 0 \\ -1 \end{bmatrix} dx \\
&= -\frac{b}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_y^n(x, y_{j-\frac{1}{2}})\, dx \simeq -\frac{b}{\Delta y} u_y^n(x_i, y_{j-\frac{1}{2}}) \simeq -\frac{b}{\Delta y} \cdot \frac{u_{ij}^n - u_{i,j-1}^n}{\Delta y},
\end{aligned}
$$

$$
\begin{aligned}
F_2 &= \frac{1}{\Delta x \Delta y} \int_B^C \left[a\, u_x^n(x_{i+\frac{1}{2}}, y), b\, u_y^n(x_{i+\frac{1}{2}}, y) \right] \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} dy \\
&= \frac{a}{\Delta x \Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u_x^n(x_{i+\frac{1}{2}}, y)\, dy \simeq \frac{a}{\Delta x} u_x^n(x_{i+\frac{1}{2}}, y_j) \simeq \frac{a}{\Delta x} \cdot \frac{u_{i+1,j}^n - u_{ij}^n}{\Delta x},
\end{aligned}
$$

$$
\begin{aligned}
F_3 &= \frac{1}{\Delta x \Delta y} \int_C^D \left[a\, u_x^n(x, y_{j+\frac{1}{2}}), b\, u_y^n(x, y_{j+\frac{1}{2}}) \right] \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} dx \\
&= \frac{b}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_y^n(x, y_{j+\frac{1}{2}})\, dx \simeq \frac{b}{\Delta y} u_y^n(x_i, y_{j+\frac{1}{2}}) \simeq \frac{b}{\Delta y} \cdot \frac{u_{i,j+1}^n - u_{ij}^n}{\Delta y},
\end{aligned}
$$

$$
\begin{aligned}
F_4 &= \frac{1}{\Delta x \Delta y} \int_A^D \left[a\, u_x^n(x_{i-\frac{1}{2}}, y), b\, u_y^n(x_{i-\frac{1}{2}}, y) \right] \cdot \begin{bmatrix} -1 \\ 0 \end{bmatrix} dy \\
&= -\frac{a}{\Delta x \Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u_x^n(x_{i-\frac{1}{2}}, y)\, dy \simeq -\frac{a}{\Delta x} u_x^n(x_{i-\frac{1}{2}}, y_j) \simeq -\frac{a}{\Delta x} \cdot \frac{u_{ij}^n - u_{i-1,j}^n}{\Delta x},
\end{aligned}
$$

where we made use of the approximation $\frac{1}{\Delta x \Delta y}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x,y_j)\,dx \simeq \frac{1}{\Delta y} f(x_i,y_j)\,,$ for example. We sum these terms and we equalize to the time-discretization (4.10), thus obtaining the same finite difference scheme (4.9).

We rewrite (4.9) in matrix-form $U^{n+1} = M \cdot U^n$, where U^n is now is a column vector in $\mathbb{R}^{(N_x-1)\times(N_y-1)}$ given by

$$U^n = (U^n_{i1}, U^n_{i2}, \dots, U^n_{i,N_y-1})^T, \qquad i = 1,2,\dots,N_x-1\,,$$

with entries

$$U^n_{ij} = (u^n_{1j}, u^n_{2j}, \dots, u^n_{N_x-1,j})^T, \qquad j = 1,2,\dots,N_y-1\,,$$

and the matrix $M \in \mathbb{R}^{((N_x-1)\times(N_y-1))^2}$ is a block-matrix in the form

$$M = \big[diag(D_2,-1)\,; diag(D_1)\,; diag(D_2,1)\big],$$

where the lower-block and the upper-block matrices are diagonal as

$$D_2 = \begin{bmatrix} b\frac{\Delta t}{\Delta y^2} & 0 & 0 & \dots \\ 0 & b\frac{\Delta t}{\Delta y^2} & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \in \mathbb{R}^{(N_x-1)\times(N_y-1)},$$

and the matrices D_1 are tri-diagonal in $\mathbb{R}^{(N_x-1)\times(N_y-1)}$ as

$$D_1 = \begin{bmatrix} 1-2a\frac{\Delta t}{\Delta x^2}-2b\frac{\Delta t}{\Delta y^2} & a\frac{\Delta t}{\Delta y^2} & 0 & \dots \\ a\frac{\Delta t}{\Delta y^2} & 1-2a\frac{\Delta t}{\Delta x^2}-2b\frac{\Delta t}{\Delta y^2} & a\frac{\Delta t}{\Delta y^2} & \dots \\ 0 & a\frac{\Delta t}{\Delta y^2} & 1-2a\frac{\Delta t}{\Delta x^2}-2b\frac{\Delta t}{\Delta y^2} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Let us verify the properties listed in Table 2.5 for the fully discrete problem in this case. Obviously, we do not consider the conservation for the first and the last rows of the matrix above, although the homogeneous Dirichlet boundary conditions are compatible with its expression.

We can easily see that M is symmetric. Moreover, the sum of the rows is equal to 1, since the off-diagonal elements of matrix D_1 cancel the term $-2a\frac{\Delta t}{\Delta x^2}$ inside the diagonal, while the term $-2b\frac{\Delta t}{\Delta x^2}$ is canceled by the two terms $b\frac{\Delta t}{\Delta x^2}$ of the corresponding upper-block and lower-block matrices D_2.

For the non-negativity property, it is immediately valid for off-diagonal elements, indeed we have $a, b > 0$, while for the diagonal entries we have as *sufficient condition* that

$$2\max\{a,b\}\Delta t\Big(\frac{1}{\Delta x^2}+\frac{1}{\Delta y^2}\Big) < 1 \quad \Longrightarrow \quad 2a\frac{\Delta t}{\Delta x^2}+2b\frac{\Delta t}{\Delta y^2} < 1\,. \tag{4.13}$$

Now, if we suppose that $\Delta x^2 > \Delta y^2$, the inequality (4.13) becomes

$$2\max\{a,b\}\Delta t\,\frac{2}{\Delta y^2} \leq 1 \quad \Longrightarrow \quad \Delta t < \frac{\Delta y^2}{4\max\{a,b\}},$$

conversely, if we suppose that $\Delta y^2 > \Delta x^2$, it becomes

$$\Delta t < \frac{\Delta x^2}{4\max\{a,b\}},$$

so finally we consider the sufficient condition for stability given by

$$\Delta t < \frac{\min\{\Delta x^2, \Delta y^2\}}{4\max\{a,b\}}. \tag{4.14}$$

This is the theoretical condition under which the discrete maximum principle holds, also implying the l^∞-stability of the numerical scheme (4.9) since all the coefficients of its stencil are less than 1.
We will further verify the experimental validity of (4.14) in Section 5.4.1.

Remark 13. *The classical two-dimensional heat equation $u_t - a\,\Delta u = 0$ corresponds to isotropic homogeneous diffusion tensors, with $a = b$ and $c = 0$, i.e. for $A = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$, and then the scheme above has its natural adaptation.*

4.2.2 Diagonal anisotropic heterogeneous diffusion

Now we focus on the case of equation (4.6), where the entries of the diffusion tensor A are generic functions of the space variables. This represents the situation in which the material feels the diffusion in a different way depending on its position, through the spatial dependence of a and b, and moreover the diffusion is different along the Cartesian axes. We calculate finite difference and finite volume discretizations and, again, we will see that are equal.

We start from (4.6) and we apply directly the approximation of derivatives at the grid points $(t^n; x_i, y_j)$ using the staggered grid in Figure 2.3,

$$\begin{aligned}
\big(a(x,y)u_x\big)_x &\simeq \frac{1}{\Delta x}\left(a(x,y)\,u_x^n\Big|_{x_{i+\frac{1}{2}}} - \; a(x,y)\,u_x^n\Big|_{x_{i-\frac{1}{2}}} \right) \\
&\simeq a_{i+\frac{1}{2},j}\,\frac{u_{i+1,j}^n - u_{ij}^n}{\Delta x^2} - a_{i-\frac{1}{2},j}\,\frac{u_{i,j}^n - u_{i-1,j}^n}{\Delta x^2} \\
&\simeq \frac{a_{ij} + a_{i+1,j}}{2}\cdot\frac{u_{i+1,j}^n - u_{ij}^n}{\Delta x^2} - \frac{a_{i-1,j} + a_{ij}}{2}\cdot\frac{u_{i,j}^n - u_{i-1,j}^n}{\Delta x^2},
\end{aligned}$$

together with

$$\begin{aligned}\big(b(x,y)u_y\big)_y &\simeq \frac{1}{\Delta y}\left(b(x,y)\,u_y^n\Big|_{y_{j+\frac12}} - b(x,y)\,u_y^n\Big|_{y_{j-\frac12}}\right)\\ &\simeq b_{i,j+\frac12}\,\frac{u^n_{i,j+1}-u^n_{ij}}{\Delta y^2} - b_{i,j-\frac12}\,\frac{u^n_{i,j}-u^n_{i,j-1}}{\Delta y^2}\\ &\simeq \frac{b_{ij}+b_{i,j+1}}{2}\cdot\frac{u^n_{i,j+1}-u^n_{ij}}{\Delta y^2} - \frac{b_{i,j-1}+b_{ij}}{2}\cdot\frac{u^n_{i,j}-u^n_{i,j-1}}{\Delta y^2}\,.\end{aligned}$$

For the time-discretization, we have the same formula (4.10) derived before, thus the fully discrete scheme reads

$$\begin{aligned}u_{ij}^{n+1} &= \frac{\Delta t}{\Delta x^2}\cdot\frac{a_{ij}+a_{i+1,j}}{2}\,u^n_{i+1,j} + \frac{\Delta t}{\Delta x^2}\cdot\frac{a_{i-1,j}+a_{ij}}{2}\,u^n_{i-1,j}\\ &\quad + \Big(1-\frac{\Delta t}{\Delta x^2}\cdot\frac{a_{i+1,j}+2a_{ij}+a_{i-1,j}}{2} - \frac{\Delta t}{\Delta y^2}\cdot\frac{b_{i,j+1}+2b_{ij}+b_{i,j-1}}{2}\Big)u^n_{ij}\\ &\quad + \frac{\Delta t}{\Delta x^2}\cdot\frac{b_{ij}+b_{i,j+1}}{2}\,u^n_{i,j+1} + \frac{\Delta t}{\Delta x^2}\cdot\frac{b_{i,j-1}+b_{ij}}{2}\,u^n_{i,j-1}\,.\end{aligned} \tag{4.15}$$

We remark that this scheme coincides with the Standard Discretization introduced in Section 2.4.2, which is moreover the nonnegative method described in Table 2.4 for $c_{ij}=0$, for all $0\le i\le N_x$, $0\le j\le N_y$, as mixed derivatives do not play any role in case of diagonal diffusion tensors.

We do not even take the Chain Rule method into account, because we already know from Section 2.5.2 that it may fail to satisfy the discrete maximum principle, neither other centered schemes generalizing (2.9) on finite difference grids as in Figure 2.1 that will exhibit a too large stencil unsuitable for practical applications.

For recovering the above discretization from the finite volume approach, we start from the divergence form of equation (4.6) and we integrate through the *divergence theorem* as done in (4.11), so that

$$\begin{aligned}&\frac{1}{\Delta x\Delta y}\int_{C_{ij}}\nabla\cdot\left(\begin{bmatrix}a(x,y) & 0\\ 0 & b(x,y)\end{bmatrix}\cdot\begin{bmatrix}u_x(t^n;x,y)\\ u_y(t^n;x,y)\end{bmatrix}\right)dx\,dy\\ &= \frac{1}{\Delta x\Delta y}\int_{\partial C_{ij}}\Big[a(s)\,u_x(t^n,s)\,,b(s)\,u_y(t^n,s)\Big]\cdot\vec{\nu}\,ds\,,\end{aligned} \tag{4.16}$$

with the same notation in (4.12) and Figure 4.3. Then, we reproduce the arguments developed in the previous section, with the only difference that now the diffusion coefficients also depend on the space variables. We treat

explicitly the first two terms (the same calculation holds for the others)

$$
\begin{aligned}
F_1 &= \frac{1}{\Delta x \Delta y} \int_A^B \Big[a(x, y_{j-\frac{1}{2}})\, u_x^n(x, y_{j-\frac{1}{2}})\,, b(x, y_{j-\frac{1}{2}})\, u_y^n(x, y_{j-\frac{1}{2}}) \Big] \cdot \begin{bmatrix} 0 \\ -1 \end{bmatrix} dx \\
&= -\frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} b(x, y_{j-\frac{1}{2}})\, u_y^n(x, y_{j-\frac{1}{2}})\, dx \\
&\simeq -\frac{1}{\Delta y} b(x_i, y_{j-\frac{1}{2}})\, u_y^n(x_i, y_{j-\frac{1}{2}}) \simeq -\frac{b_{i,j-1} + b_{ij}}{2} \cdot \frac{u_{ij}^n - u_{i,j-1}^n}{\Delta y^2}\,, \\
F_2 &= \frac{1}{\Delta x \Delta y} \int_B^C \Big[a(x_{i+\frac{1}{2}}, y)\, u_x^n(x_{i+\frac{1}{2}}, y)\,, b(x_{i+\frac{1}{2}}, y)\, u_y^n(x_{i+\frac{1}{2}}, y) \Big] \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} dy \\
&= \frac{1}{\Delta x \Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} a(x_{i+\frac{1}{2}}, y)\, u_x^n(x_{i+\frac{1}{2}}, y)\, dy \\
&\simeq \frac{1}{\Delta x} a(x_{i+\frac{1}{2}}, y_j)\, u_x^n(x_{i+\frac{1}{2}}, y_j) \simeq \frac{a_{ij} + a_{i+1,j}}{2} \cdot \frac{u_{i+1,j}^n - u_{ij}^n}{\Delta x^2}\,.
\end{aligned}
$$

These two integrals are analogous of the terms in the finite difference scheme (4.15), that we will recover entirely summing all other terms and equalizing to the time-discretization (4.10). We introduce the notations

$$
\begin{aligned}
a_{i+\frac{1}{2},j} &= \frac{a_{ij} + a_{i+1,j}}{2}\,, \quad a_{i-\frac{1}{2},j} = \frac{a_{i-1,j} + a_{ij}}{2}\,, \\
b_{i,j+\frac{1}{2}} &= \frac{b_{ij} + b_{i,j+1}}{2}\,, \quad b_{i,j-\frac{1}{2}} = \frac{b_{i,j-1} + b_{ij}}{2}\,,
\end{aligned}
$$

and we write the matrix-form of the scheme (4.15) using the block-matrix

$$
M = \big[diag(D_2^-, -1)\,; diag(D_1)\,; diag(D_2^+, 1) \big],
$$

where the lower-block matrix D_2^- and the upper-block matrix D_2^+ are diagonal in $\mathbb{R}^{(N_x-1)\times(N_y-1)}$ for all $i = 1, 2, \ldots, N_x - 1$, such that

$$
D_2^- = diag\big(\frac{\Delta t}{\Delta y^2}\, b_{i,j-\frac{1}{2}}\big)_{1\le j\le N_y-1} = \begin{bmatrix} \frac{\Delta t}{\Delta y^2}\, b_{i,1-\frac{1}{2}} & 0 & 0 & \ldots \\ 0 & \frac{\Delta t}{\Delta y^2}\, b_{i,2-\frac{1}{2}} & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},
$$

and

$$
D_2^+ = diag\big(\frac{\Delta t}{\Delta y^2}\, b_{i,j+\frac{1}{2}}\big)_{1\le j\le N_y-1} = \begin{bmatrix} \frac{\Delta t}{\Delta y^2}\, b_{i,1+\frac{1}{2}} & 0 & 0 & \ldots \\ 0 & \frac{\Delta t}{\Delta y^2}\, b_{i,2+\frac{1}{2}} & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},
$$

while the matrices D_1 are tri-diagonal in $\mathbb{R}^{(N_x-1)\times(N_y-1)}$ as

$$D_1 = \Big[diag(A_{i-\frac{1}{2},j}, -1)\,; diag(1 - A_{ij} - B_{ij})\,; diag(A_{i+\frac{1}{2},j}, 1)\Big]_{1\le i\le N_x-1}$$
$$= \begin{bmatrix} 1 - A_{1j} - B_{1j} & A_{1+\frac{1}{2},j} & 0 & 0 & \dots \\ A_{2-\frac{1}{2},j} & 1 - A_{2j} - B_{2j} & A_{2+\frac{1}{2},j} & 0 & \dots \\ 0 & A_{3-\frac{1}{2},j} & 1 - A_{3j} - B_{3j} & A_{3+\frac{1}{2},j} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

for all $j = 1, 2, \dots, N_y - 1$, and we fixed

$$A_{i-\frac{1}{2},j} = \frac{\Delta t}{\Delta x^2}\, a_{i-\frac{1}{2},j}\,, \qquad A_{ij} = \frac{\Delta t}{\Delta x^2} \cdot \frac{a_{i+1,j} + 2a_{ij} + a_{i-1,j}}{2}\,,$$
$$A_{i+\frac{1}{2},j} = \frac{\Delta t}{\Delta x^2}\, a_{i+\frac{1}{2},j}\,, \qquad B_{ij} = \frac{\Delta t}{\Delta y^2} \cdot \frac{b_{i,j+1} + 2b_{ij} + b_{i,j-1}}{2}\,.$$

Concerning the properties in Table 2.5, we immediately observe that the symmetry and the conservation are satisfied, since the off-diagonal elements $A_{i-\frac{1}{2},j}$ and $A_{i+\frac{1}{2},j}$ of the matrix D_1 cancel the term A_{ij} inside the diagonal, while the term B_{ij} is canceled by the two terms $\frac{\Delta t}{\Delta y^2}\, b_{i,j+\frac{1}{2}}$ and $\frac{\Delta t}{\Delta y^2}\, b_{i,j-\frac{1}{2}}$ of the corresponding rows of the upper-block matrix D_2^+ and the lower-block matrix D_2^-, respectively.

For the non-negativity property, it is valid for the off-diagonal elements of the whole matrix M, indeed we have that $a_{ij}, b_{ij} > 0$ for all i, j, while for the diagonal entries we have to impose that

$$\frac{\Delta t}{\Delta x^2} \cdot \frac{a_{i+1,j} + 2a_{ij} + a_{i-1,j}}{2} + \frac{\Delta t}{\Delta y^2} \cdot \frac{b_{i,j+1} + 2b_{ij} + b_{i,j-1}}{2} < 1\,.$$

We consider the maximum for all values of the diffusion coefficients over the grid points, and we deduce the following sufficient condition

$$2\,\Delta t \max\{a_{ij}, b_{ij}\}\Big(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\Big) < 1\,,$$

so finally we can assert a candidate for the CFL-condition given by

$$\Delta t < \frac{\min\{\Delta x^2, \Delta y^2\}}{4 \max\{a_{ij}, b_{ij}\}}\,, \tag{4.17}$$

which turns out to be the analogue of (4.14) found in the previous case. Therefore, the numerical scheme (4.15) is also l^∞-stable, since the non-negativity combined with the property of elements to be always less than 1 ensures stability.

4.2.3 Fully anisotropic homogeneous diffusion

We consider the model (4.7), which describes a diffusion phenomenon spreading in all directions according to the values of the entries of diffusion tensor.

To compute the finite volume discretization, we use the explicit Euler scheme (4.10) for the time derivative, together with the usual integral average for the divergence term,

$$\begin{aligned}
&\frac{1}{\Delta x\Delta y}\int_{C_{ij}} \nabla\cdot\left(\begin{bmatrix} a & c\\ c & b\end{bmatrix}\cdot\begin{bmatrix} u_x(t^n;x,y)\\ u_y(t^n;x,y)\end{bmatrix}\right) dx\,dy\\
&= \frac{1}{\Delta x\Delta y}\int_{\partial C_{ij}} \Big[a\,u_x(t^n,s)+c\,u_y(t^n,s)\,,\; c\,u_x(t^n,s)+b\,u_y(t^n,s)\Big]\cdot\vec{\nu}\,ds\,,
\end{aligned} \tag{4.18}$$

with the same cell boundaries (4.12) and its normal vectors as in Figure 4.3. We focus only on the terms including the parameter c for mixed derivatives, because for the others the calculation already performed in the case (4.11) still holds. Therefore, we have

$$\begin{aligned}
-\frac{c}{\Delta x\Delta y}\int_{x_{i-\frac12}}^{x_{i+\frac12}} u_x^n(x,y_{j-\frac12})\,dx &\simeq -\frac{c}{\Delta x\Delta y}\Big(u^n_{i+\frac12,j-\frac12}-u^n_{i-\frac12,j-\frac12}\Big),\\
\frac{c}{\Delta x\Delta y}\int_{y_{j-\frac12}}^{y_{j+\frac12}} u_y^n(x_{i+\frac12},y)\,dy &\simeq \frac{c}{\Delta x\Delta y}\Big(u^n_{i+\frac12,j+\frac12}-u^n_{i+\frac12,j-\frac12}\Big),\\
\frac{c}{\Delta x\Delta y}\int_{x_{i-\frac12}}^{x_{i+\frac12}} u_x^n(x,y_{j+\frac12})\,dx &\simeq \frac{c}{\Delta x\Delta y}\Big(u^n_{i+\frac12,j+\frac12}-u^n_{i-\frac12,j+\frac12}\Big),\\
-\frac{c}{\Delta x\Delta y}\int_{y_{j-\frac12}}^{y_{j+\frac12}} u_y^n(x_{i-\frac12},y)\,dy &\simeq -\frac{c}{\Delta x\Delta y}\Big(u^n_{i-\frac12,j+\frac12}-u^n_{i-\frac12,j-\frac12}\Big).
\end{aligned} \tag{4.19}$$

which corresponds to the finite difference discretization (2.16) on a staggered grid. We recall that we do not apply the centered formula (2.15) because this would lead to the Chain Rule method derived in Section 2.4.1 for which the discrete maximum principle could actually fail (refer to Section 2.5.2). At this point, if we proceed from (4.19) by substituting with combinations of arithmetic averages, for example

$$u^n_{i+\frac12,j-\frac12}\simeq\frac{u^n_{i,j-\frac12}+u^n_{i+1,j-\frac12}}{2}\simeq\frac{u^n_{i,j-1}+u^n_{ij}+u^n_{i+1,j-1}+u^n_{i+1,j}}{4}\,,$$

summing over all the terms in (4.19), we obtain

$$\begin{aligned}
&2\,\frac{c}{\Delta x\Delta y}\Big(u^n_{i+\frac12,j+\frac12}-u^n_{i-\frac12,j+\frac12}-u^n_{i+\frac12,j-\frac12}+u^n_{i-\frac12,j-\frac12}\Big)\\
&\simeq\frac{c}{2\Delta x\Delta y}\big(u^n_{i+1,j+1}-u^n_{i+1,j-1}-u^n_{i-1,j+1}+u^n_{i-1,j-1}\big)\,,
\end{aligned}$$

thus coming back to the centered approximation (2.15) which is a special case of both the Chain Rule scheme from (2.30) and the Standard Discretization from (2.32) in the case of constant diffusion coefficients, namely

$$\frac{u_{ij}^{n+1}-u_{ij}^{n}}{\Delta t}=a\,\frac{u_{i+1,j}^{n}-2u_{ij}^{n}+u_{i-1,j}^{n}}{\Delta x^2}+b\,\frac{u_{i,j+1}^{n}-2u_{ij}^{n}+u_{i,j-1}^{n}}{\Delta y^2}$$
$$+\frac{c}{2\Delta x\Delta y}\big(u_{i+1,j+1}^{n}-u_{i+1,j-1}^{n}-u_{i-1,j+1}^{n}+u_{i-1,j-1}^{n}\big),$$

or rather

$$\begin{aligned}u_{ij}^{n+1}&=\big(1-2a\frac{\Delta t}{\Delta x^2}-2b\frac{\Delta t}{\Delta y^2}\big)\,u_{ij}^{n}\\&+a\frac{\Delta t}{\Delta x^2}\big(u_{i+1,j}^{n}+u_{i-1,j}^{n}\big)+b\frac{\Delta t}{\Delta y^2}\big(u_{i,j+1}^{n}+u_{i,j-1}^{n}\big)\\&+\frac{c\,\Delta t}{2\Delta x\Delta y}\big(u_{i+1,j+1}^{n}-u_{i+1,j-1}^{n}-u_{i-1,j+1}^{n}+u_{i-1,j-1}^{n}\big).\end{aligned}\tag{4.20}$$

The matrix-form the scheme (4.20) uses the block-matrix

$$M=\big[diag(D_2^-,-1)\,;diag(D_1)\,;diag(D_2^+,1)\big],$$

where the matrix D_1 is the same as the one calculated in Section 4.2.1, because it does not involve the coefficient c, while for the lower-block and the upper-block matrices we have the tri-diagonal modifications

$$D_2^-=\begin{bmatrix} b\frac{\Delta t}{\Delta y^2} & -\frac{c}{2\Delta x\Delta y} & 0 & \dots\\ \frac{c}{2\Delta x\Delta y} & b\frac{\Delta t}{\Delta y^2} & -\frac{c}{2\Delta x\Delta y} & \dots\\ 0 & \frac{c}{2\Delta x\Delta y} & b\frac{\Delta t}{\Delta y^2} & \dots\\ \vdots & \vdots & \vdots & \ddots\end{bmatrix}$$

and

$$D_2^+=\begin{bmatrix} b\frac{\Delta t}{\Delta y^2} & \frac{c}{2\Delta x\Delta y} & 0 & \dots\\ -\frac{c}{2\Delta x\Delta y} & b\frac{\Delta t}{\Delta y^2} & \frac{c}{2\Delta x\Delta y} & \dots\\ 0 & -\frac{c}{2\Delta x\Delta y} & b\frac{\Delta t}{\Delta y^2} & \dots\\ \vdots & \vdots & \vdots & \ddots\end{bmatrix}.$$

From these, despite the symmetry and the conservation property, it is clear that the matrix M cannot be nonnegative, due to the terms $\pm\frac{c}{2\Delta x\Delta y}$ in the upper and lower matrices D_2^+ and D_2^-. Indeed, we have no information about the sign of $c\in\mathbb{R}$, according to the positive definiteness of the matrix A which only imposes $c^2<ab$, and so if one of these two terms is positive its opposite is certainly negative. Therefore, the discrete maximum principle is never satisfied for the approximation (4.20) of the anisotropic homogeneous parabolic equation (4.7).

For that fundamental reason, we have to consider the Nonnegative Discretization developed in Section 2.4.3, with the stencil reported in Table (2.3) for the particular case of constant diffusion coefficients, which induces the following numerical method,

$$\begin{aligned} u_{ij}^{n+1} = &\Big(1 - 2a\frac{\Delta t}{\Delta x^2} - 2b\frac{\Delta t}{\Delta y^2} + 2|c|\frac{\Delta t}{\Delta x \Delta y}\Big)u_{ij}^n \\ &+ \Big(a\frac{\Delta t}{\Delta x^2} - |c|\frac{\Delta t}{\Delta x \Delta y}\Big)\big(u_{i+1,j}^n + u_{i-1,j}^n\big) \\ &+ \Big(b\frac{\Delta t}{\Delta y^2} - |c|\frac{\Delta t}{\Delta x \Delta y}\Big)\big(u_{i,j+1}^n + u_{i,j-1}^n\big) \\ &+ \frac{|c| - c}{2} \cdot \frac{\Delta t}{\Delta x \Delta y}\big(u_{i+1,j-1}^n + u_{i-1,j+1}^n\big) \\ &+ \frac{|c| + c}{2} \cdot \frac{\Delta t}{\Delta x \Delta y}\big(u_{i-1,j-1}^n + u_{i+1,j+1}^n\big), \end{aligned} \tag{4.21}$$

whose matrix-form is given through the block-matrix

$$M = \big[diag(D_2^-, -1)\,; diag(D_1)\,; diag(D_2^+, 1)\big],$$

where D_1 is a symmetric tri-diagonal matrix with entries

$$\begin{aligned} d_{kk} &= 1 - 2a\frac{\Delta t}{\Delta x^2} - 2b\frac{\Delta t}{\Delta y^2} + 2|c|\frac{\Delta t}{\Delta x \Delta y}\,, \quad \forall\, k = 1, 2, \ldots, N_x - 1\,, \\ d_{k,k+1} &= d_{k+1,k} = a\frac{\Delta t}{\Delta x^2} - |c|\frac{\Delta t}{\Delta x \Delta y}\,, \quad \forall\, k = 1, 2, \ldots, N_x - 2\,, \end{aligned}$$

and

$$D_2^- = \begin{bmatrix} b\frac{\Delta t}{\Delta y^2} - |c|\frac{\Delta t}{\Delta x \Delta y} & \frac{|c|-c}{2} \cdot \frac{\Delta t}{\Delta x \Delta y} & 0 & 0 & \ldots \\ \frac{|c|+c}{2} \cdot \frac{\Delta t}{\Delta x \Delta y} & b\frac{\Delta t}{\Delta y^2} - |c|\frac{\Delta t}{\Delta x \Delta y} & \frac{|c|-c}{2} \cdot \frac{\Delta t}{\Delta x \Delta y} & 0 & \ldots \\ 0 & \frac{|c|+c}{2} \cdot \frac{\Delta t}{\Delta x \Delta y} & b\frac{\Delta t}{\Delta y^2} - |c|\frac{\Delta t}{\Delta x \Delta y} & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

together with $D_2^+ = \big(D_2^-\big)^T$. Obviously, the symmetry and the conservation property are satisfied. Also the non-negativity of the coefficients should be guaranteed, by construction. Indeed, referring to (4.21) as represented in Table 2.3, the entries in the outer angles are either zero or always positive, since $|c| \geq c$. For the other elements, because of the relations (2.39) imposed in the proof of Theorem 7, we have that

$$\frac{a}{\Delta x} \geq \frac{c}{\Delta y}\,, \qquad \frac{b}{\Delta y} \geq \frac{c}{\Delta x}. \tag{4.22}$$

for all values of the coefficient $c \in \mathbb{R}$, which imply

$$\frac{a}{\Delta x^2} \geq \frac{|c|}{\Delta x \Delta y}\,, \qquad \frac{b}{\Delta y^2} \geq \frac{|c|}{\Delta x \Delta y}\,,$$

so we have the non-negativity of the off-diagonal elements inside the stencil as long as the constraints (4.22) hold. For the central entries, we must verify

$$2\Delta t\Big(\frac{a}{\Delta x^2}+\frac{b}{\Delta y^2}-\frac{|c|}{\Delta x\Delta y}\Big)<1\,, \tag{4.23}$$

and a sufficient condition is obtained by developing into the brackets as

$$2\Delta t\frac{\max\{a,b\}}{\min\{\Delta x^2,\Delta y^2\}}S(a,b,c)<1\,,$$

where $S(a,b,c):=\left[\frac{a}{\Delta x^2}\cdot\frac{\min\{\Delta x^2,\Delta y^2\}}{\max\{a,b\}}+\frac{b}{\Delta y^2}\cdot\frac{\min\{\Delta x^2,\Delta y^2\}}{\max\{a,b\}}-\frac{|c|\ min(\Delta x^2,\Delta y^2)}{\Delta x\Delta y\ max(a,b)}\right]$ and, therefore, the condition for the non-negativity of the central entries is

$$\Delta t<\frac{\min\{\Delta x^2,\Delta y^2\}}{2\ \max\{a,b\}}\,S^{-1}(a,b,c)\,. \tag{4.24}$$

Moreover, the matrix M is symmetric and the sum its rows (except the first and the last one) is equal to 1, as easily seen from the stencil in Table 2.3, since all the horizontal and vertical entries around the central one exactly cancel it, and adding the time-discretization the total sum is equal to 1.
Additionally, we have proven that all coefficients are less than 1, thanks again to (4.22), and this leads to the l^∞-stability of the scheme through the discrete maximum principle.

We conclude this section by remarking that the same properties above are also satisfied for the Nonnegative Discretization of fully anisotropic and heterogeneous diffusion equations, as we can deduce from a careful analysis of the stencil given in Table 2.4.

4.3 L^2-stability analysis of numerical schemes

In particular, for the case analyzed in Section (4.2.1), because the model (4.5) is linear, autonomous and homogeneous, we can apply the Fourier analysis as we have done for the one-dimensional case in Section 3.1.1.
We denote by $\hat{u}$ the Fourier transform of the function u, with $\hat{u}^n(\cdot)=\hat{u}(t^n;\cdot)$ for t^n fixed, such that

$$\hat{u}^n(\xi_x,\xi_y)=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}^2}e^{-\iota x\xi_x}e^{-\iota y\xi_y}u(t^n;x,y)\,dx\,dy\,,$$

and the inverse Fourier transform reads

$$u(t^n;x,y)=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}^2}e^{\iota x\xi_x}e^{\iota y\xi_y}\hat{u}^n(\xi_x,\xi_y)\,d\xi_x\,d\xi_y\,.$$

Let us calculate at the grid points $x_i=i\Delta x$, $i=0,1,\dots,N_x$, and $y_j=j\Delta y$, $j=0,1,\dots,N_y$, making the hypothesis that the computational domain is

$\Omega = [0,1] \times [0,1]$, so for the values of the numerical solution we have

$$u_{ij}^n \simeq u(t^n; x_i, y_j) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^2} e^{\iota(i\Delta x)\xi_x} e^{\iota(j\Delta y)\xi_y} \hat{u}^n(\xi_x, \xi_y)\, d\xi_x\, d\xi_y\,.$$

We substitute into the scheme in form (4.8) and, after reducing the integrals and common exponential factors (also neglecting the dependence upon the variables, for simplicity), we obtain

$$\frac{\hat{u}^{n+1} - \hat{u}^n}{\Delta t} = \frac{a}{\Delta x^2}\big(e^{\iota\Delta x\xi_x} - 2 + e^{-\iota\Delta x\xi_x}\big)\hat{u}^n + \frac{b}{\Delta y^2}\big(e^{\iota\Delta y\xi_y} - 2 + e^{-\iota\Delta y\xi_y}\big)\hat{u}^n\,,$$

which gives directly the formula $\hat{u}^{n+1} = G\,\hat{u}^n$, where we can recognize the amplification factor

$$G := 1 + a\frac{\Delta t}{\Delta x^2}\big(e^{\iota\Delta x\xi_x} - 2 + e^{-\iota\Delta x\xi_x}\big) + b\frac{\Delta t}{\Delta y^2}\big(e^{\iota\Delta y\xi_y} - 2 + e^{-\iota\Delta y\xi_y}\big)\,.$$

We use the standard relation (3.19) in order to rewrite

$$G = 1 - 4\,a\frac{\Delta t}{\Delta x^2}\sin^2\big(\frac{\Delta x\,\xi_x}{2}\big) - 4\,b\frac{\Delta t}{\Delta y^2}\sin^2\big(\frac{\Delta y\,\xi_y}{2}\big)\,. \tag{4.25}$$

To ensure the L^2-stability we must have $|G| \leq 1$. Since all the coefficients in front of $\sin^2(\cdot)$ are positive, the inequality $G \leq 1$ always holds, so let us analyze the case $G \geq -1$, namely

$$\Delta t\left[\frac{a}{\Delta x^2}\sin^2\big(\frac{\Delta x\,\xi_x}{2}\big) + \frac{b}{\Delta y^2}\sin^2\big(\frac{\Delta y\,\xi_y}{2}\big)\right] \leq \frac{1}{2}\,.$$

As usual, we consider a sufficient condition given by

$$\Delta t\frac{\max\{a,b\}}{\min\{\Delta x^2, \Delta y^2\}}\left[\sin^2\big(\frac{\Delta x\,\xi_x}{2}\big) + \sin^2\big(\frac{\Delta y\,\xi_y}{2}\big)\right] \leq \frac{1}{2}\,,$$

and so the worst case is satisfied when we have that

$$\Delta t\frac{\max\{a,b\}}{\min\{\Delta x^2, \Delta y^2\}} \leq \frac{1}{4}\,,$$

which is the same condition found in (4.14) for the validity of the discrete maximum principle, and thus ensuring the L^∞-stability of the method (obtained through the analysis of the matrix-form of the scheme).

We remark that a more restrictive stability condition is derived in [7], at least when applied to the case of diagonal anisotropic homogeneous diffusion tensors like (4.5), which reads $\Delta t \leq \frac{1}{8}\frac{\min\{\Delta x^2, \Delta y^2\}}{\max\{a,b\}}$. The result established in [7][Section 5, Theorem 5.1] shows that the maximum principle can be violated for the Standard Discretization in Table 2.2, but we still have stability :

at each time-iteration, the numerical solution is bounded by the maximum value of the initial data multiplied by a positive factor, and this guarantees it does not blow up in finite time and ensures its stability, but the maximum principle can be violated by the positivity of the multiplicative factor.
The reason of discrepancy between continuous and numerical solution is that the second order Standard Discretization is not a nonnegative approximation, as we already discussed in the previous section.

To improve the result in [7], we rewrite the scheme (4.8) using its original structure (2.31),

$$\begin{aligned} u_{ij}^{n+1} = u_{ij}^n + a\frac{\Delta t}{\Delta x^2}\Big[\big(u_{i+1,j}^n - u_{ij}^n\big) - \big(u_{ij}^n - u_{i-1,j}^n\big)\Big] \\ + b\frac{\Delta t}{\Delta y^2}\Big[\big(u_{i,j+1}^n - u_{ij}^n\big) - \big(u_{ij}^n - u_{i,j-1}^n\big)\Big] \end{aligned}$$

and we split the first term into two parts $u_{ij}^n = \frac{1}{2}u_{ij}^n + \frac{1}{2}u_{ij}^n$, then reorganizing the previous equation as

$$\begin{aligned} u_{ij}^{n+1} = \big(1 - 2a\frac{\Delta t}{\Delta x^2} - 2b\frac{\Delta t}{\Delta y^2}\big)u_{ij}^n \\ + a\frac{\Delta t}{\Delta x^2}\big(u_{i+1,j}^n + u_{i-1,j}^n\big) + b\frac{\Delta t}{\Delta y^2}\big(u_{i,j+1}^n + u_{i,j-1}^n\big)\,. \end{aligned}$$

As usual, we have to impose $0 < 1 - 2a\frac{\Delta t}{\Delta x^2} - 2b\frac{\Delta t}{\Delta y^2} < 1$, the upper bound being always satisfied, and for the lower bound we manipulate like

$$2\Delta t\frac{\max\{a,b\}}{\min\{\Delta x^2, \Delta y^2\}}\left[\frac{a}{\Delta x^2}\cdot\frac{\min\{\Delta x^2,\Delta y^2\}}{\max\{a,b\}} + \frac{b}{\Delta y^2}\cdot\frac{\min\{\Delta x^2,\Delta y^2\}}{\max\{a,b\}}\right],$$

thus finding a better CFL-condition, i.e.

$$\Delta t < \frac{1}{2}\frac{\min\{\Delta x^2, \Delta y^2\}}{\max\{a,b\}}S^{-1}, \tag{4.26}$$

where $S := \frac{a}{\Delta x^2}\cdot\frac{\min\{\Delta x^2,\Delta y^2\}}{\max\{a,b\}} + \frac{b}{\Delta y^2}\cdot\frac{\min\{\Delta x^2,\Delta y^2\}}{\max\{a,b\}}$, and $0 < S \leq 2$ as desired.

4.4 Other two isotropic heterogeneous models

For completeness, we take into account other two cases of parabolic equation, that are the two-dimensional version of those illustrated in Section 3.3.

First, we consider the heterogeneous two-dimensional heat equation, i.e.

$$u_t \;=\; a(x,y)\Delta u = a(x,y)\big(u_{xx} + u_{yy}\big)\,. \tag{4.27}$$

The numerical scheme computed through finite difference and finite volume methods is the same, just repeating the calculations performed for the previous cases, and it looks like

$$\begin{aligned} u_{ij}^{n+1} &= u_{ij}^n + a_{ij}\left[\frac{\Delta t}{\Delta x^2}\big(u_{i+1,j}^n - 2u_{ij}^n + u_{i-1,j}^n\big) + \frac{\Delta t}{\Delta y^2}\big(u_{i,j+1}^n - 2u_{ij}^n + u_{i,j-1}^n\big)\right] \\ &= a_{ij}\frac{\Delta t}{\Delta x^2}\big(u_{i+1,j}^n + u_{i-1,j}^n\big) + \left[1 - 2\,a_{ij}\Big(\frac{\Delta t}{\Delta x^2} + \frac{\Delta t}{\Delta y^2}\Big)\right]u_{ij}^n \\ &\quad + a_{ij}\frac{\Delta t}{\Delta y^2}\big(u_{i,j+1}^n + u_{i,j-1}^n\big), \end{aligned} \tag{4.28}$$

which can be rewritten in matrix-form with a block-matrix whose entries are now index-dependent matrices, namely

$$M = \big[diag(E_j^-, -1)_{1\le j\le N_y-2}\,; diag(D_j)_{1\le j\le N_y-1}\,; diag(E_j^+, 1)_{2\le j\le N_y-1}\big],$$

where the lower-block and the upper-block matrices are diagonal as

$$\begin{aligned} E_j^- &= \Big[diag\big(a_{ij}\frac{\Delta t}{\Delta y^2}\big)\Big]_{1\le i\le N_x-2} \quad \text{for all } j = 1, 2, \dots, N_y - 2\,, \\ E_j^+ &= \Big[diag\big(a_{ij}\frac{\Delta t}{\Delta y^2}\big)\Big]_{2\le i\le N_x-1} \quad \text{for all } j = 2, 3, \dots, N_y - 1\,, \end{aligned}$$

and the central blocks are tri-diagonal matrices

$$D_j = \Big[diag\big(a_{ij}\frac{\Delta t}{\Delta x^2}, -1\big)\,; diag\big(1-2a_{ij}\big(\frac{\Delta t}{\Delta x^2}+\frac{\Delta t}{\Delta y^2}\big)\big)\,; diag\big(a_{ij}\frac{\Delta t}{\Delta x^2}, 1\big)\Big]_{1\le i\le N_x-1}$$

for all $j = 1, 2, \dots, N_y - 1$.
About the properties listed in Table 2.5, although it may be harder to check using the matrices above, the *conservation* is actually valid because the sum of the coefficients in (4.28) is clearly equal to 1. This fact, together with the positiveness property below, guarantees the L^∞-stability of the method.

Remark 14. *There existes a direct correspondence between the rows of the matrix of a numerical scheme and its stencil, as one can easily see for non-homogeneous three points schemes like* (4.28)*, for instance.*

Nevertheless, the *symmetry* is not satisfied, that is coherent with the one-dimensional case in Section 3.3, since we are dealing with a parabolic equation (4.27) which is not in conservation form.
For the *non-negativity* of all entries, it is satisfied if

$$2\,a_{ij}\,\Delta t\Big(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\Big) < 1\,,$$

from which a sufficient condition for the stability is given by

$$\Delta t < \frac{\min\{\Delta x^2, \Delta y^2\}}{4\, a_{\max}}\,, \qquad a_{\max} = \max_{0\le i\le N_x,\, 0\le j\le N_y} a_{ij}\,, \tag{4.29}$$

that is the analogous condition to (3.34).
We remark that, under the above constraint, the numerical scheme (4.28) exhibit all coefficients positive and less than 1, thus providing a convex combination for the stability.

The second case is the two-dimensional version of equation (3.37), i.e.

$$u_t \,=\, \Delta\big(a(x,y)u\big) = \big(a(x,y)u\big)_{xx} + \big(a(x,y)u\big)_{yy}\,,$$

which is a *conservation law* with flux $F(t;x,y) = \nabla_{(x,y)}\big(a(x,y)u\big)$.
Therefore, we can directly apply the finite difference/volume scheme (3.38) to both components of the Laplacian operator, to obtain

$$\begin{aligned}\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\Delta t} \;=\; & \frac{a_{i+1,j}\, u_{i+1,j}^{n} - 2a_{ij}\, u_{ij}^{n} + a_{i-1,j}\, u_{i-1,j}^{n}}{\Delta x^2} \\ & + \frac{a_{i,j+1}\, u_{i,j+1}^{n} - 2\, a_{ij}\, u_{ij}^{n} + a_{i,j-1}\, u_{i,j-1}^{n}}{\Delta y^2}\,,\end{aligned}$$

that can be rewritten as

$$\begin{aligned}u_{ij}^{n+1} = {} & a_{i+1,j}\frac{\Delta t}{\Delta x^2}\, u_{i+1,j}^{n} + a_{i-1,j}\frac{\Delta t}{\Delta x^2}\, u_{i-1,j}^{n} \\ & + \left[1 - 2\, a_{ij}\Big(\frac{\Delta t}{\Delta x^2} + \frac{\Delta t}{\Delta y^2}\Big)\right] u_{ij}^{n} \\ & + a_{i,j+1}\frac{\Delta t}{\Delta y^2}\, u_{i,j+1}^{n} + a_{i,j-1}\frac{\Delta t}{\Delta y^2}\, u_{i,j-1}^{n}\,,\end{aligned} \tag{4.30}$$

with matrix-form given by means of the block-matrix

$$M = \big[diag(E_j^-, -1)_{1\le j\le N_y-2}\,; diag(D_j)_{1\le j\le N_y-1}\,; diag(E_j^+, 1)_{2\le j\le N_y-1}\big]\,,$$

where the lower-block and the upper-block matrices are diagonal as

$$\begin{aligned}E_j^- &= \Big[diag\big(a_{i,j+1}\frac{\Delta t}{\Delta y^2}\big)\Big]_{1\le i\le N_x-2} \quad \text{for all } j = 1, 2, \ldots, N_y - 2\,, \\ E_j^+ &= \Big[diag\big(a_{i,j-1}\frac{\Delta t}{\Delta y^2}\big)\Big]_{2\le i\le N_x-1} \quad \text{for all } j = 2, 3, \ldots, N_y - 1\,,\end{aligned}$$

and the central blocks are tri-diagonal matrices

$$D_j = \Big[diag\big(a_{i-1,j}\frac{\Delta t}{\Delta x^2}, -1\big)\,; diag\big(1 - 2\, a_{ij}\big(\frac{\Delta t}{\Delta x^2} + \frac{\Delta t}{\Delta y^2}\big)\big)\,; diag\big(a_{i+1,j}\frac{\Delta t}{\Delta x^2}, 1\big)\Big]_{1\le i\le N_x-1}\,,$$

for all with $j = 1, 2, \ldots, N_y - 1$.
Concerning the properties listed in Table 2.5, the *conservation* is not valid because if we fix any row of the above matrices, the value of the entries varies according to the column's index and the sum of the elements is typically different from 1 , that can also be easily seen from the coefficients in (4.30). For the same reason, the *symmetry* is not satisfied, that is also coherent with the one-dimensional case in Section 3.3. Solely, the *non-negativity* property holds if the inequality (4.29) is satisfied.

Chapter 5

Experimental validation and numerical results

In this chapter, we present several cases of parabolic/diffusion equation (4.1) that have been previously treated theoretically, to verify through coherent numerical results the conclusions formulated above. All the codes have been implemented in the framework of Scilab – http://www.scilab.org
We start with the simplest one-dimensional heat equation, and we progressively introduce complications like heterogenous coefficients and anisotropic diffusion tensors, to finally show the necessity of nonnegative discretizations to ensure the fundamental property of discrete maximum principle.

5.1 Definition of initial data, boundary conditions and numerical parameters

For all experimental tests, we consider the region where we would study the effects of the diffusion as the computational domain $\Omega_T = \Omega \times [0, T]$, where the two-dimensional spatial domain is given by

$$\Omega = [-10, 10] \times [-10, 10] \,, \tag{5.1}$$

which is discretized as in Figure 2.1, with the grid points (x_i, y_j) such that $i = 0, 1, \ldots, N_x$ and $j = 0, 1, \ldots, N_y$, while for the time-discretization we have

$$t^n = n\,\Delta t \,, \qquad n = 0, 1, \ldots, N_t \,,$$

and Δt is typically calculated according to some CFL-condition, depending on the analytical model and the numerical scheme under consideration.
The parameter $T > 0$ will be modified following the specific solution to be observed, but generally it is chosen small enough to avoid the solution to flatten too much, as this phenomenon would not allow to clearly see the behaviour of the numerical solution.

Since the well-posedness of time-dependent PDEs depends on both the characteristics of the equation and the type of supplementary conditions for the initial and boundary values problem, for second-order parabolic equations we have to carefully fix boundary conditions with respect to space and initial conditions with respect to time. For the *initial data*, we consider

$$u(0;x,y) = u_0(x,y)\,, \qquad \forall\,(x,y)\in\Omega\,,$$

which defines the solution at time $t=0$. From the physical point of view, we choose u_0 a positive function, because negative densities do not have sense. Then, we will take into account two types of *boundary conditions*, as evoked in the previous chapters,

- Dirichlet conditions : $u|_{\partial\Omega} = g_1(t;x,y)\,,\ \forall\, t\in[0,T]\,,\ (x,y)\in\partial\Omega\,,$
- Neumann conditions : $\frac{\partial u}{\partial n}\big|_{\partial\Omega} = g_2(t;x,y)\,,\ \forall\, t\in[0,T]\,,\ (x,y)\in\partial\Omega\,.$

In the first case, we simply assert that the density at the boundary of the domain has to be equal to a given function g_1, which may vary on time. For the numerical tests, if we put $g_1 = 0$, we force the solution to be equal to zero at the boundary, and physically we are describing a situation in which the density vanishes approaching the boundary.
The second case states that the gradient of u at the boundary of Ω is equal to a given function g_2, which may vary on time. The particular case in which the function $g_2 = 0$ suggests that inner and outer flows of the substance through $\partial\Omega$ are not allowed, and in this way we have that the domain Ω represents the whole physical space where the substance can diffuse. For any rectangular domain like (5.1), this translates into the two-dimensional analogue of (3.28), namely

$$\begin{aligned}
0 = u_x(t;x_0,y) &\simeq \frac{u(t;x_0+\Delta x,y) - u(t;x_0,y)}{\Delta x}\,,\\
0 = u_x(t;x_{N_x},y) &\simeq \frac{u(t;x_{N_x},y) - u(t;x_{N_x}-\Delta x,y)}{\Delta x}\,,\\
0 = u_y(t;x,y_0) &\simeq \frac{u(t;x,y_0+\Delta y) - u(t;x,y_0)}{\Delta y}\,,\\
0 = u_y(t;x,y_{N_y}) &\simeq \frac{u(t;x,y_{N_y}) - u(t;x,y_{N_y}-\Delta y)}{\Delta y}\,,
\end{aligned}$$

and for the numerical issues we represent that situation as

$$\begin{aligned}
u(t^n;x_0,y_j) &= u(t^n;x_1,y_j) && \forall\, j=0,1,\dots,N_y\,,\ \forall\, n\geq 0\\
u(t^n;x_{N_x},y_j) &= u(t^n;x_{N_x-1},y_j) && \forall\, j=0,1,\dots,N_y\,,\ \forall\, n\geq 0\\
u(t^n;x_i,y_0) &= u(t^n;x_i,y_1) && \forall\, i=0,1,\dots,N_x\,,\ \forall\, n\geq 0\\
u(t^n;x_i,y_{N_y}) &= u(t^n;x_i,y_{N_y-1}) && \forall\, i=0,1,\dots,N_x\,,\ \forall\, n\geq 0
\end{aligned} \tag{5.2}$$

Some considerations about the violation of the Discrete Maximum Principle are in order. The theoretical results on the maximum/minimum principle established in the first chapter of this report assert that the solution $u(t; x, y)$ has to be bounded by the maximum and the minimum values of the initial data $u_0(x, y)$, at least in the case of homogeneous Dirichlet boundary conditions. Due to the characteristics of the problem under analysis, especially the fact that (4.1) is autonomous, this also means that the maximum of the solution at fixed time t has to be less than the maximum of the solution at the previous time $t - \Delta t$ (thus suggesting a *monotonicity* property already mentioned in Section 1.4, since the solution at time t can be interpreted as the initial data for future times and, therefore, the maximum/minimum bounds should hold step by step).
Considering the dependence of the problem from the initial data, inside the areas of the physical domain Ω where there is an initial high concentration, the solution must decrease in time at any point, otherwise the solution must rise where the density is low at the initial time. This is indeed the description of the *diffusivity phenomenon*, when the density spreads through the domain from areas with higher concentrations to areas with lower ones.
Figure 5.1 shows a solution correctly decreasing from the top of the initial data towards the bottom, whilst Figure 5.2 shows a case of violation of the discrete maximum principle.

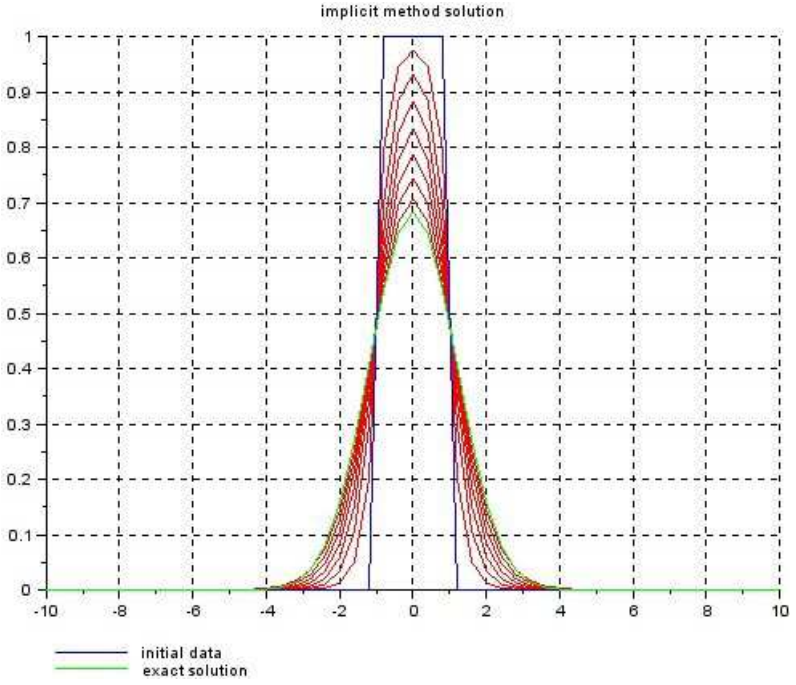


Figure 5.1: example of fulfillment of the discrete maximum principle

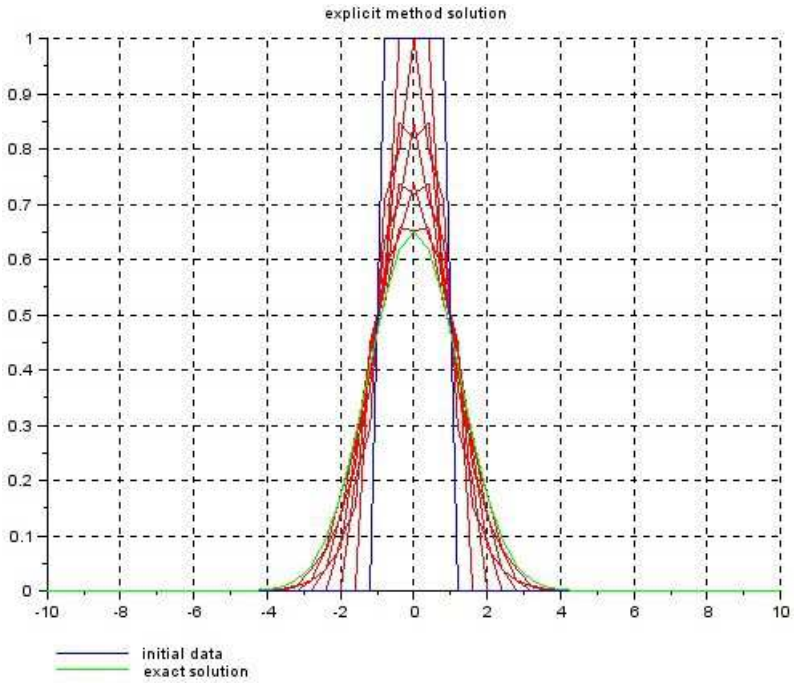


Figure 5.2: example of violation of the discrete maximum principle

5.2 Numerical tests for the one-dimensional heat equation

We refer to Section 3.1, and we consider the following initial data

$$u_0(x) = \begin{cases} 1 & |x| \leq 1 \\ 0 & otherwise \end{cases} \tag{5.3}$$

and homogeneous Dirichlet boundary conditions

$$u(t^n, x_0) = 0\,, \quad u(t^n, x_{N_x}) = 0\,, \qquad \forall\, n \geq 0\,.$$

This is the so-called *double Riemann problem* and we can calculate an exact solution, for comparison with the numerical solution at the final time T, according to [34][Chapter 6] given by

$$u(t,x) = \frac{1}{2}\left[\mathrm{erf}\Big(\frac{1-x}{\sqrt{4at}}\Big) + \mathrm{erf}\Big(\frac{1+x}{\sqrt{4at}}\Big)\right], \tag{5.4}$$

where erf denotes the *error function* defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2}\, ds\,.$$

The analytical solution becomes successively more spread out as the time increases. When t gets very large, the argument of the error functions in (5.4) are progressively smaller, for any x fixed. Thus, the corresponding value of the solution $u(t,x)$ tends to zero asymptotically in time.

Let us analyze the three numerical schemes proposed in Section 3.1, with the parameters varying as follows. The *CFL*-condition defines the length of the time-intervals as

$$\Delta t = CFL\,\frac{\Delta x^2}{2\,a}\,, \tag{5.5}$$

which not only describes the regime of stability by the Von Neumann analysis, but also the range of fulfillment of the discrete maximum principle. We recall that, theoretically from (3.22), for the explicit method the condition $CLF < 1$ provides both the l^2-stability and the maximum/minimum principle, for the implicit method these two properties are satisfied for all values of CFL, while for the Crank-Nicolson method we have that the stability is always satisfied but the maximum/minimum principle holds if $CFL < 2$, so we can introduce another parameter such that

$$CFL' = \frac{CFL}{2} < 1\,. \tag{5.6}$$

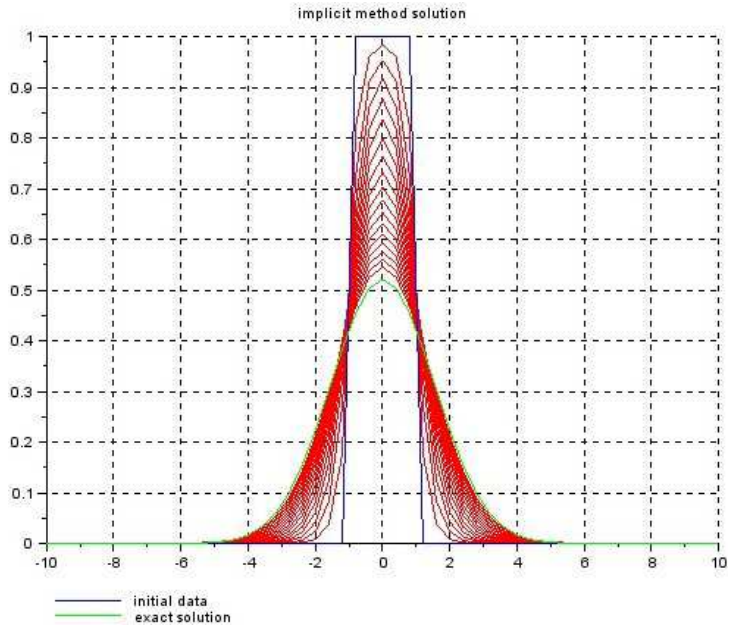
Obviously, the space-step Δx being computed as $\Delta x = \frac{x_{N_x} - x_0}{N_x}$, increasing the parameter N_x leads to refine the grid points and the time intervals through (5.5). We will see in the numerical tests that this parameter also characterizes the *numerical instability*.

5.2.1 The time-implicit method

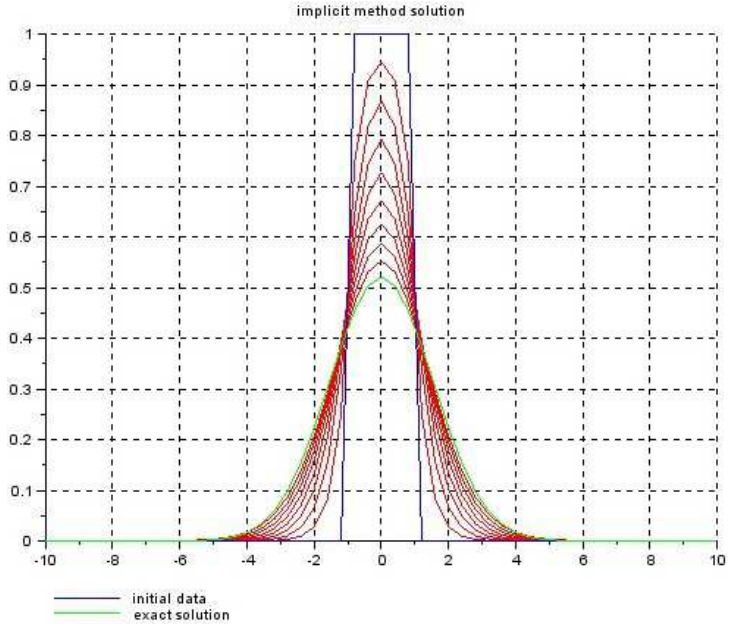
Several numerical tests have been performed, which confirm the theoretical results that this method is unconditionally stable and satisfies the discrete maximum principle. However, an important practical observation is the decreasing of the numerical diffusion as the *CFL* number decreases : this is due to the coefficient accompanying the term u_{xxxx} in the *modified equation*, that is derived by calculating the Taylor's expansion of $u(x_i, t^{n+1})$ to get

$$\begin{aligned}
(u_t)_i^{n+1} - a\,(u_{xx})_i^{n+1} &= \Big(a^2\frac{\Delta t}{2} + a\frac{\Delta x^2}{12}\Big)(u_{xxxx})_i^{n+1} + H.O.T. \\
&= a\Big(CFL\frac{\Delta x^2}{4} + \frac{\Delta x^2}{12}\Big)(u_{xxxx})_i^{n+1} + H.O.T. \\
&= a\frac{\Delta x^2}{4}\Big(CFL + \frac{1}{3}\Big)(u_{xxxx})_i^{n+1} + H.O.T.\,,
\end{aligned}$$

where $H.O.T. = O(\Delta t^2, \Delta x^4)$ for smooth analytical solutions, and therefore the *numerical viscosity* grows up with the value of CFL. We can see this phenomenon of artificially enlarging the diffusion in Figure 5.3, which makes



(a) $CFL = 0.7$



(b) $CFL = 1.5$

Figure 5.3: effects of artificial viscosity for the implicit method

the choice of time-step not really arbitrary even for the implicit scheme to have an accurate numerical solution.

Consequently, for numerical reasons, we can decide to use explicit schemes

which require less computational costs than the implicit ones to give good approximations of the exact solution. However, the implicit method remains very important for the particular case of *stiff problems*.
A typical example in this class is the ordinary differential equation, for $\epsilon > 0$,

$$y' = -\frac{1}{\epsilon} y \,, \qquad y(0) = y_0 \,,$$

for which the classical *forward Euler method* produces the explicit scheme

$$\frac{y^{n+1} - y^n}{\Delta t} = -\frac{1}{\epsilon} y^n \quad \Longrightarrow \quad y^{n+1} = \Big(1 - \frac{\Delta t}{\epsilon}\Big) y^n .$$

By iterating over n , and then choosing $T = 1$, $\Delta t = \frac{T}{n}$, we have

$$y^n = \Big(1 - \frac{\Delta t}{\epsilon}\Big)^n y_0 = \Big(1 - \frac{1}{n\,\epsilon}\Big)^n y_0 \quad \longrightarrow_{n\to\infty} \quad y(1) = y_0\, e^{-\frac{1}{\epsilon}} .$$

On the other hand, the *backward Euler method* gives the implicit scheme

$$\frac{y^{n+1} - y^n}{\Delta t} = -\frac{1}{\epsilon} y^{n+1} \quad \Longrightarrow \quad \Big(1 + \frac{\Delta t}{\epsilon}\Big) y^{n+1} = y^n ,$$

which can be rewritten as

$$y^n = \left[\Big(1 + \frac{\Delta t}{\epsilon}\Big)^n\right]^{-1} y_0 \,,$$

and again this converges to the exact solution above, at time $T = 1$, iterating over $n \to \infty$. Now, if we consider two different initial data y_0 and w_0, by linearity the error between the corresponding numerical solutions y and w satisfies the same equation, so that

$$E^n_{exp} = \Big(1 - \frac{1}{n\,\epsilon}\Big)^n E_0 \,, \qquad E^n_{imp} = \left[\Big(1 + \frac{1}{n\,\epsilon}\Big)^n\right]^{-1} E_0 \,.$$

Finally, the standard stability analysis for discrete equations ensures that the implicit scheme is unconditionally stable, because

$$0 < \left[\Big(1 + \frac{1}{n\,\epsilon}\Big)^n\right]^{-1} < 1 \,, \quad \forall\, n \geq 0 \,,$$

whilst the explicit scheme needs some constraint on the numerical parameters, the so-called condition for *error contraction* , i.e.

$$0 < \Big(1 - \frac{1}{n\,\epsilon}\Big)^n < 1 \quad \Longrightarrow \quad n > \frac{1}{\epsilon} \,,$$

So, if the physical problem describes a situation with ϵ becoming very small, a lot of time steps are necessary to make the scheme stable and catch the

exact behaviour of the solution, and in this sense an explicit scheme fails the approximation of stiff problems, making the implicit schemes preferable because they are unconditionally stable.

In the simplest case of one-dimensional heat equation, we have performed numerical tests with all the three schemes presented in Section 3.1, but we will consider only the explicit method when dealing with other general cases.

5.2.2 The time-explicit method

We focus on four experimental cases :

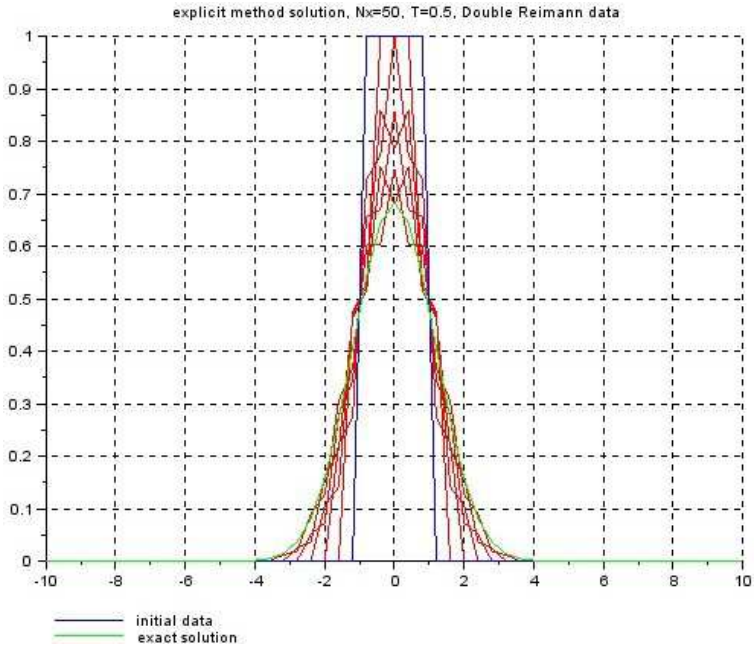
1. *CFL* number fixed by the stability condition, and different values for the grid parameter N_x ;
2. different values of *CFL* under the stability condition $CFL < 1$;
3. $CFL \simeq 1$;
4. violation of the theoretical stability condition, namely $CFL > 1$.

First case. We fix $CFL = 0.95 < 1$ in (5.5), so that increasing N_x induces Δx, and then Δt, to decrease. We observe an improvement of the numerical solution in Figure 5.4, which is due to the *numerical accuracy*. This property of the algorithm reflects into the accuracy of the results, it depends only on the numerical parameters and not on the problem itself, and typically the refinement of grid points through Δx leads the spurious modes to vanish.

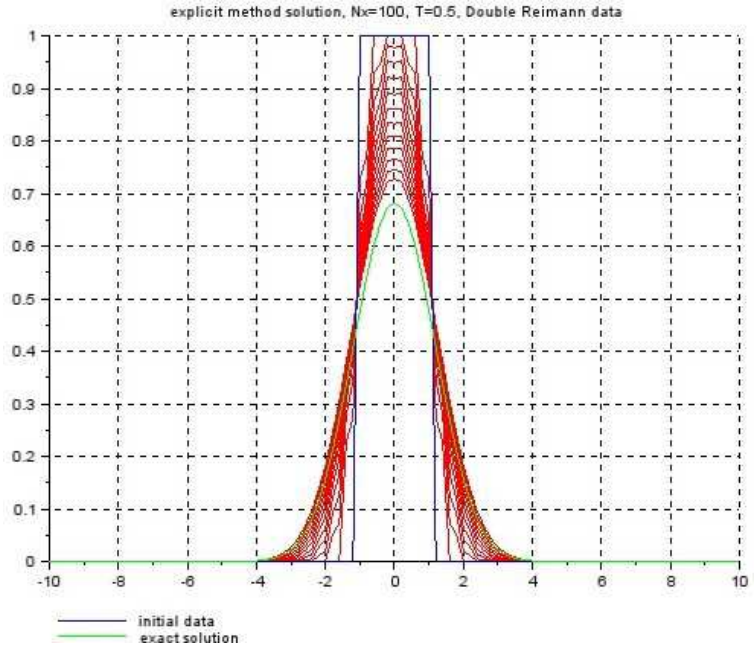
Second case. We want to check the theoretical condition $CFL < 1$ for the discrete maximum principle through coherent numerical results.
The main question reads : is the maximum principle satisfied whenever the condition $CFL < 1$ holds? Unfortunately, the answer is not. The range for which the discrete maximum principle is actually satisfied looks like

$$CFL \leq 0.85\,, \tag{5.7}$$

that is considerably lower than the theoretical one.
Referring to Figure 5.5(a), we can see in all subintervals of the domain where the solution decreases with time, it exhibits a maximum value below that at previous times, which is a sign that the maximum principle is satisfied.
In Figure 5.5(b), for a bigger CFL number, the solution changes its convexity, and clearly there is violation of the maximum principle. To emphasize this phenomenon, we consider $CFL = 0.98 >> 0.85$ and $N_x = 50$, and we perform the numerical tests reproduced in Figure 5.6. For $T = 0.3$, three iterations are achieved; for $T = 0.3 + \Delta t$, the method executes another iteration, but the maximum of the numerical solution at the last iteration is



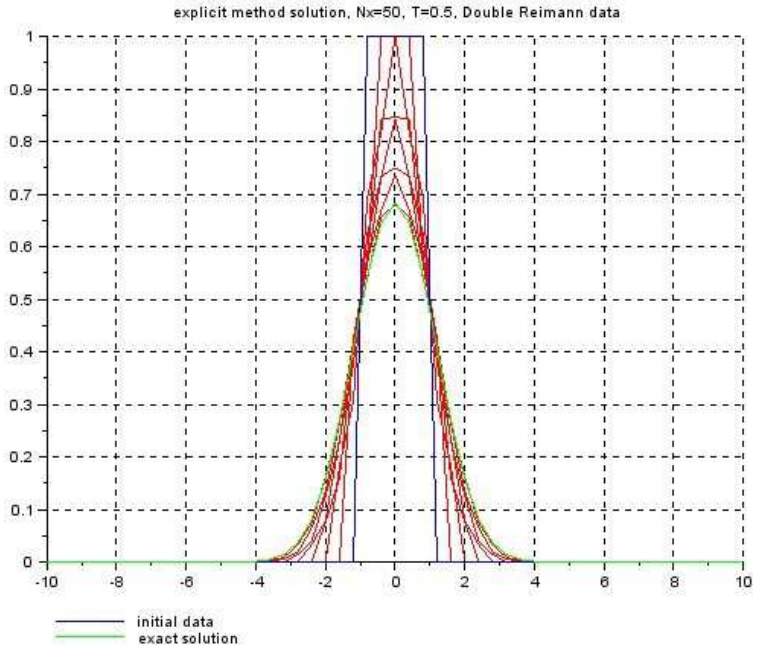
(a) $\Delta x = 0.4$



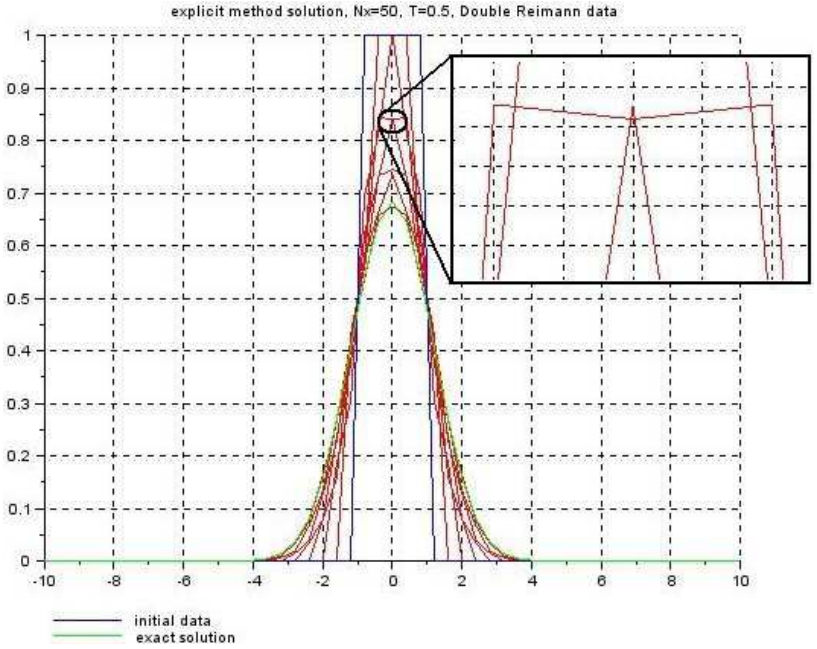
(b) $\Delta x = 0.2$

Figure 5.4: reducing the numerical instability with grid refinement

higher than the previous one, thus violating the maximum principle.

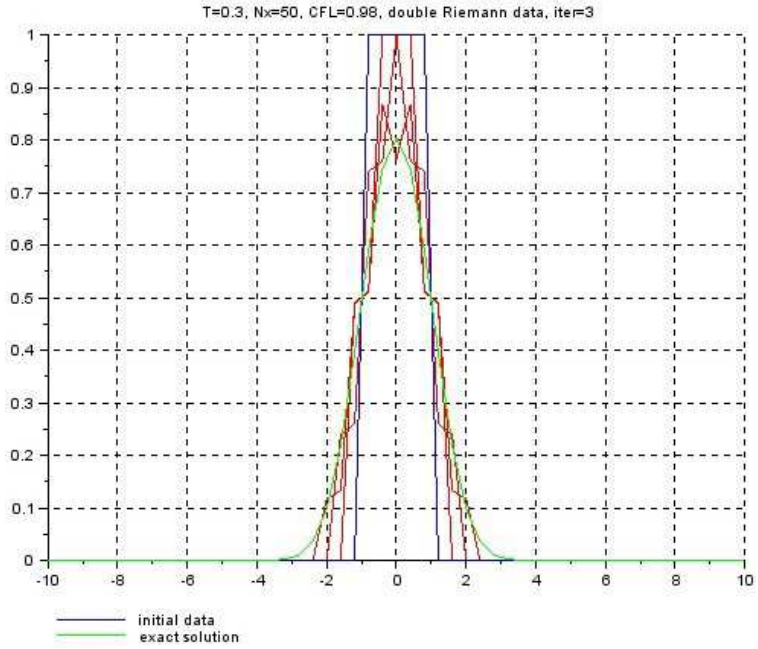


(a) $CFL = 0.85$

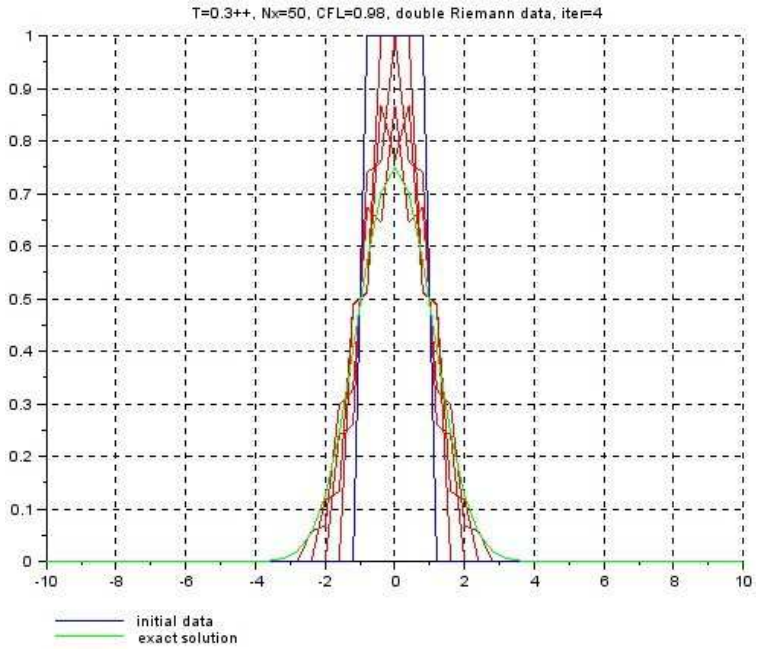


(b) $CFL = 0.86$

Figure 5.5: reducing the numerical instability with CFL constraints



(a) numerical solution at time $T = 0.3$



(b) numerical solution at time $T = 0.3 + \Delta t$

Figure 5.6: violation of discrete maximum principle for the explicit scheme

The fact that the experimental limit $CFL = 0.85$ for the discrete maximum/minimum principle is less than the theoretical prediction $CFL = 1$ can be explained through the analysis of the computational errors carried out by the computer's software when approximating arithmetical operations, and this question comes together with the *floating point representation* of the real numbers [30]. Whenever we use computer programs, we must take into account that any real number x actually has its machine representation

$$fl(x) = (1 \pm \epsilon_M)\, x\,, \tag{5.8}$$

where ϵ_M depends on the computer characteristics, and it measures the relative error made in replacing x with its floating point representation,

$$\frac{\big|x - fl(x)\big|}{|x|} = \epsilon_M\,.$$

We remark that (5.8) is better than the alternative definition $fl(x) = x \pm \epsilon_M$, because the latter expresses the *truncation error* in percent.
We return to the explicit scheme (3.6), and we assume that solely the values of the numerical solution are affected by representation errors, so that

$$\begin{aligned}(1 \pm \epsilon_1)u_i^{n+1} = \Big[1 - 2\,a\, fl\Big(\frac{\Delta t}{\Delta x^2}\Big)\Big](1 \pm \epsilon_2)u_i^n + a\, fl\Big(\frac{\Delta t}{\Delta x^2}\Big)(1 \pm \epsilon_3)u_{i+1}^n \\ + a\, fl\Big(\frac{\Delta t}{\Delta x^2}\Big)(1 \pm \epsilon_4)u_{i-1}^n\,,\end{aligned}$$

with the hypothesis that integer numbers satisfy $2 = fl(2)$ and $a = fl(a)$. For simplicity, we impose the same error for all the values u_i^n, because we can always estimate with the maximum of the ϵ_k, $k = 1, 2, 3, 4$. In order to have a convex combination of coefficients, the theoretical stability condition reads $0 \le 2a\frac{\Delta t}{\Delta x^2} \le 1$, and the floating point representation implies

$$2\,a\, fl\Big(\frac{\Delta t}{\Delta x^2}\Big) = 2\,a\left[\frac{(1 \pm \epsilon_{\Delta t})\Delta t}{(1 \pm \epsilon_{\Delta x})^2\,\Delta x^2}\right] = 2\,a\,\frac{\Delta t}{\Delta x^2}\,Err_M\,, \tag{5.9}$$

where $Err_M = \frac{(1\pm\epsilon_{\Delta t})}{(1\pm\epsilon_{\Delta x})^2}$ and, typically, it holds $Err_M >> 1$. The positivity is still satisfied because the floating point representation preserves the sign, instead the *contractivity* condition for (5.9) is fulfilled if

$$2\,a\,\frac{\Delta t}{\Delta x^2}\,Err_M \le 1 \quad \Longrightarrow \quad 2\,a\,\frac{\Delta t}{\Delta x^2} \le \frac{1}{Err_M}\,,$$

therefore the numerical errors influence the theoretical results by straitening the CFL-condition and, obviously, the experimental requirement (5.7) turns out to be more restrictive.

Third case. Now, we study more carefully the behaviour of the numerical solution when $CFL \simeq 1$. We start by remarking another interesting phenomenon in Figure 5.6 : we focus on the monotone branches and we see that, although exhibiting small artificial steps, the solution does not change its convexity. If we increase the CFL number, the convexity is unaltered, as shown in Figure 5.7(a), until $CFL = 1$ is reached and some parts of the vertical branches clearly flatten in Figure 5.7(b). If we further increase the value of CFL, the convexity changes drastically and a stronger instability occurs, resulting in uncontrolled oscillations (refer to Figure 5.8), thus showing that the l^2-stability persists more that the l^∞-stability encoded into the discrete maximum/minimum principle.

For $CFL = 1$, the explicit scheme (3.6) reduces to the arithmetic average between the values u_{i-1}^n and u_{i+1}^n , namely

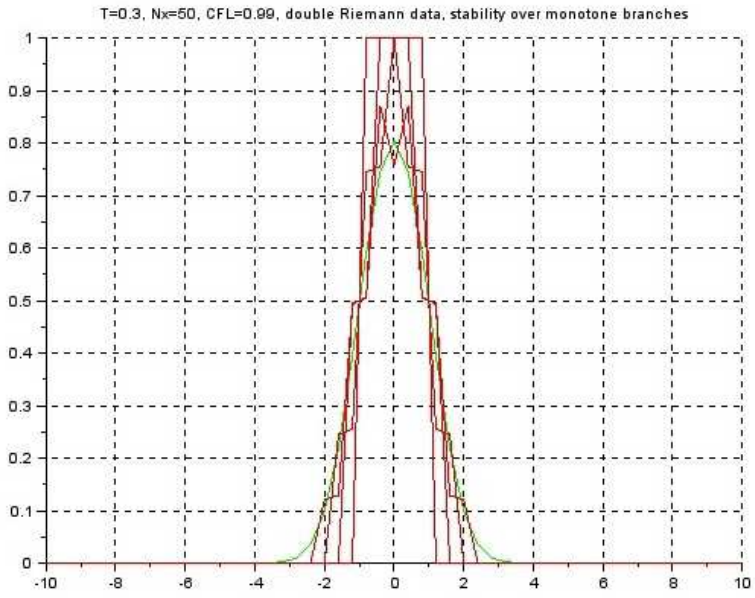
$$u_i^{n+1} = \frac{1}{2}\, u_{i-1}^n + \frac{1}{2}\, u_{i+1}^n \,,$$

so that, to calculate the numerical solution at the same point x_i at time t^{n+1}, we take the values at previous time t^n and we make the average of the left and right neighborhoods. In this way, for the components u_i^n on the monotone branches, the convexity does not change because the neighboring values are one above and one below u_i^n, respectively. But, if u_i^n is the maximum value at time t^n, very likely u_{i-1}^n and u_{i+1}^n are equal, and the convexity changes, as we have already seen in Figure 5.6. More precisely, the maximum value $u(t^3, 0)$ at the third iteration is below the maximum value $u(t^2, 0)$ at the previous one, but for the fourth iteration the value $u(t^4, 0)$ is greater than $u(t^3, 0)$ as it is just equal to

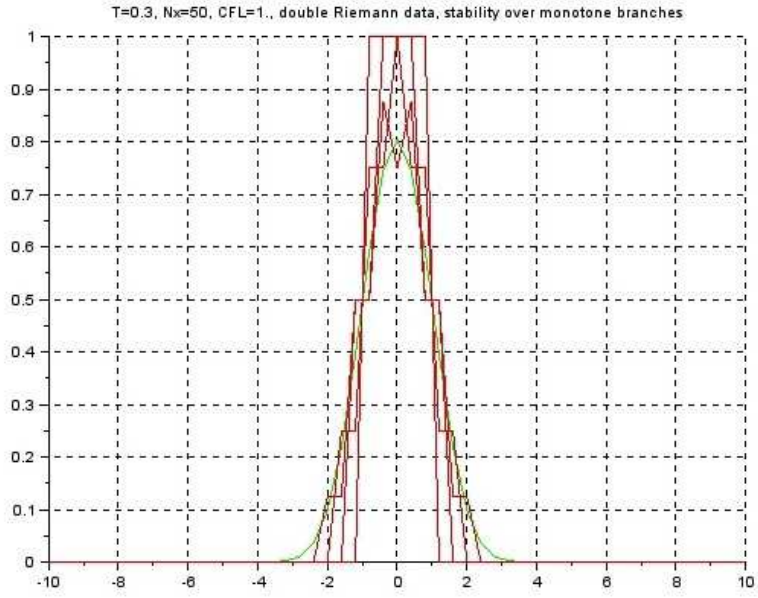
$$u(t^4, 0) = \frac{u(t^3, -\Delta x) + u(t^3, \Delta x)}{2} = u(t^3, -\Delta x) = u(t^3, \Delta x) > u(t^3, 0)$$

and this is clearly a violation of the maximum principle.

In conclusion, in case of change of convexity in the numerical solution, the maximum/minimum principle is no longer satisfied. Moreover, we have seen in the second case how the validity of the maximum principle deteriorates if $CFL \geq 0.85$ with Δx fixed, whereas it is satisfied if we fix the CFL number small enough and increase N_x in the first case. These two complementary aspects find an explanation in the sensitivity of the problem with respect to the discrete maximum/minimum principle, in comparison to the standard numerical instability (which is further analyzed in the next case) : the stability through the Von Neumann analysis considers the L^2-norm, i.e. the integral norm, which is lighter than the L^∞-norm used for the maximum/minimum principle, so the last one is more sensitive to the change of



(a) $CFL = 0.99$



(b) $CFL = 1$

Figure 5.7: change of convexity due to imminent instability regimes

parameters inside the numerical scheme.

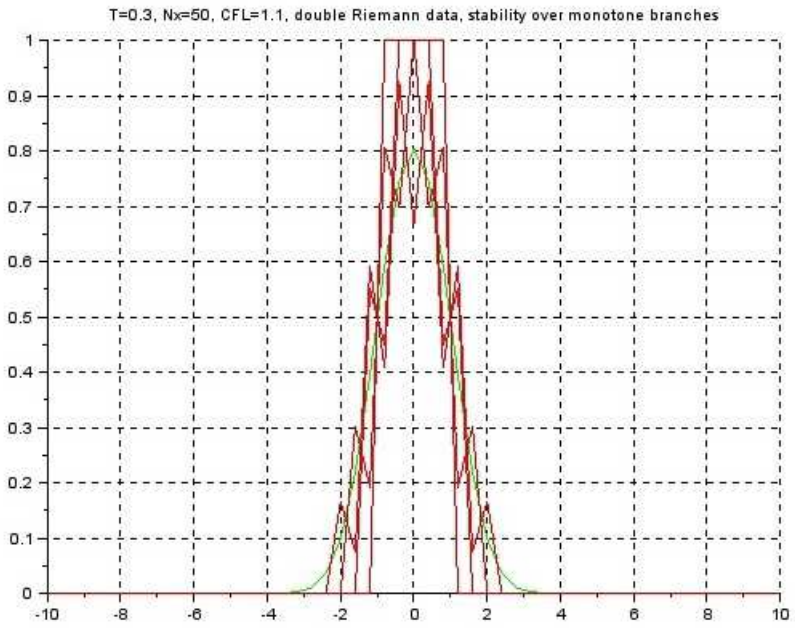


Figure 5.8: appearance of instabilities for $CFL > 1$

Fourth case. When the CFL-condition exceeds the theoretically predicted values for stability, the numerical solution starts oscillating around the exact solution, as we have seen in Figure 5.8. This phenomenon in known as the *Ultraviolet Catastrophe* (borrowing a term from quantum mechanics).
This usually happens in those situations in which one wants to approximate a model (the one-dimensional heat equation, for example) with another (the numerical algorithm) which has a similar behaviour on low frequencies, and we can apply the Fourier analysis, but it is very different on high frequencies, i.e. short wavelengths and, therefore, ultraviolet electromagnetic spectrum.

Indeed, recalling the arguments developed in Section 3.1.2, the *amplification factor* for the explicit scheme (3.6) is given by (3.20) fixing $\theta = 0$, so that using (5.5) and the definition $\lambda = \frac{\Delta t}{\Delta x^2}$, we can rewrite it as

$$G = 1 - 4\,a\,\lambda \sin^2\big(\frac{\Delta x\,\xi}{2}\big) = 1 - 2\,CFL \cdot \sin^2\big(\frac{\Delta x\,\xi}{2}\big).$$

For the stability condition $|G| \leq 1$, the upper bound is always satisfied since $CFL > 0$, while the lower bound holds if

$$-1 \leq 1 - 2\,CFL \cdot \sin^2\big(\frac{\Delta x\,\xi}{2}\big) \quad \Longrightarrow \quad CFL \cdot \sin^2\big(\frac{\Delta x\,\xi}{2}\big) \leq 1\,.$$

It is clear that, if we put $CFL \leq 1$, the above inequality is always satisfied independently from the value of ξ, whereas if we fix $CFL > 1$, the worst

value of $\sin^2(\frac{\Delta x\,\xi}{2})$ obviously being 1, the stability condition can be violated. In this sense, we have a stable numerical approximation for low frequencies because, as ξ becomes small, the factor $\sin^2(\frac{\Delta x\,\xi}{2}) < 1$ makes the inequality even stronger; moreover, for moderate frequencies, reducing Δx would always improve the stability, regardless the value of CFL, and then the dissipation mechanism of the numerical scheme appears similar to that of the continuous model. On the other hand, for the high frequencies, when ξ becomes large, the amplification factor could actually becomes greater than 1, and we can assist to oscillations inside the numerical solution, although the continuous model is able to dump out possible oscillations. Therefore, it is required to control the stability with the CFL number, to limit the amplification factor, otherwise we witness fluctuations that are not typical of the model to approximate. That is the essence of the *Ultraviolet Catastrophe*.

5.2.3 The semi-implicit Crank-Nicolson method

We recall (see Table 3.1) that this scheme is stable for all values of $\lambda = \frac{\Delta t}{\Delta x^2}$, as established by the Von Neumann analysis, but the theoretical condition for the validity of the maximum/minimum principle is given by (5.5)-(5.6). We see in the numerical tests that, actually, the Crank-Nicolson scheme satisfies that property if $CFL \leq 1.7$ and, when this experimental value is overpassed, the discrete maximum principle fails, as shown in Figure 5.9.
This behaviour is coherent with the condition (5.7) observed for the explicit scheme, indeed

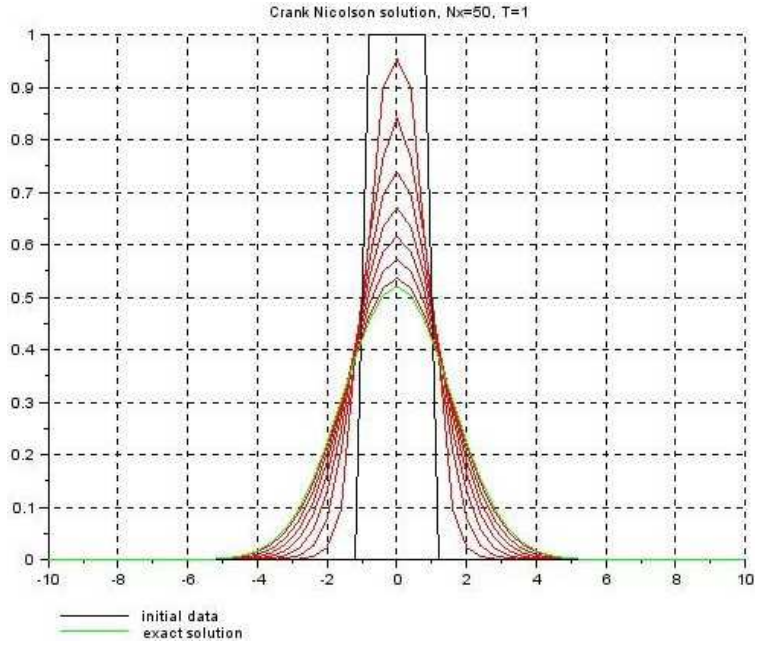
$$CFL' = \frac{CFL}{2} \leq \frac{1.7}{2} = 0.85\,,$$

and the explanation is the same described above concerning the characteristics of computations with floating point real numbers. That does not surprise since the Crank-Nicolson scheme (3.11) is a linear combination of the explicit and the implicit schemes, thus inheriting (half of) the stability constraints corresponding to the explicit solver.
Now, if we fix $CFL >> 1.7$, we observe in Figure 5.10 spurious oscillations starting in the graph of the numerical solution, which are not completely resolved even when reducing the space-step, and this clearly depends on the numerical instability.

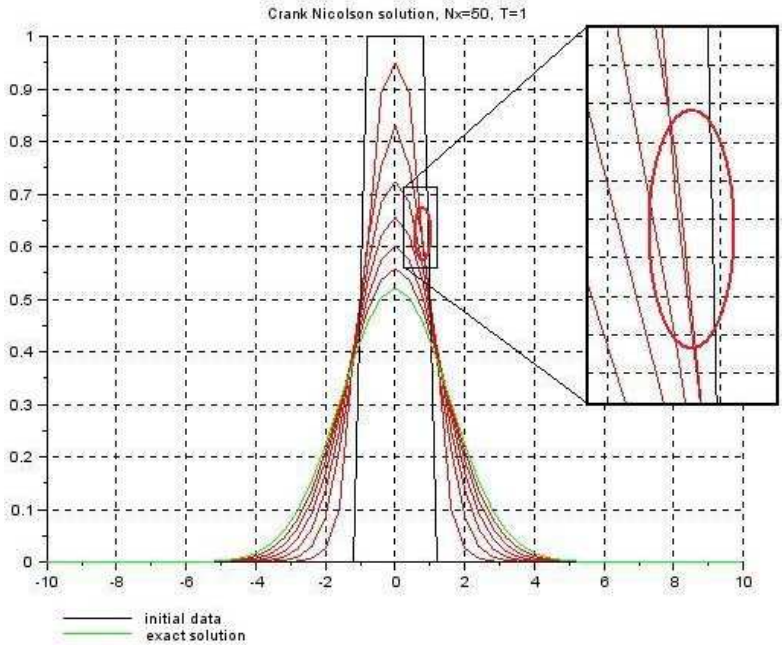
5.3 Numerical tests for one-dimensional heterogeneous diffusion

For the numerical tests of this section, we introduce a heterogeneous diffusion coefficient $a(x) > 0$ to treat the parabolic equation (3.27), together with the initial data (5.3) and Neumann boundary conditions (3.28), so that

$$u_0^n = u_1^n\,, \quad u_{N_x}^n = u_{N_x-1}^n\,, \qquad \forall\, n \geq 0\,.$$

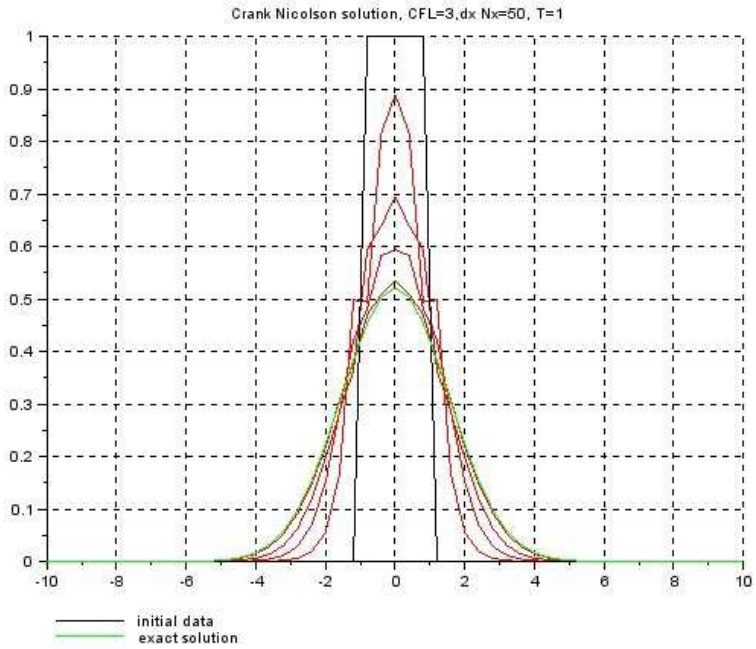


(a) $CFL = 1.7$

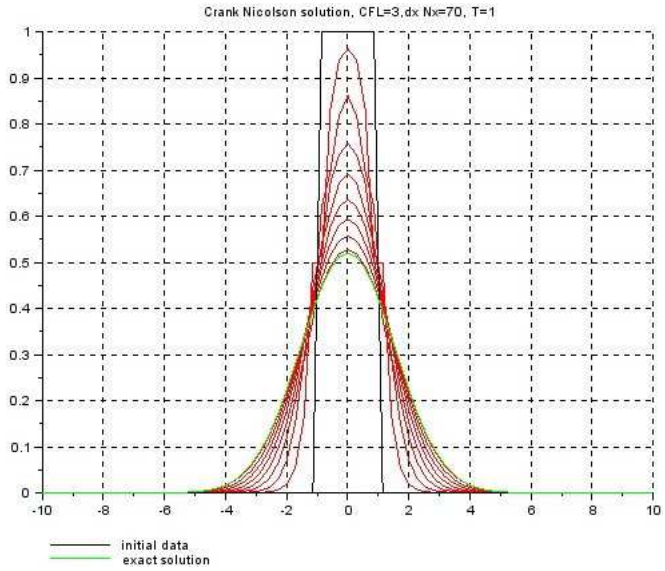


(b) $CFL = 1.8$

Figure 5.9: experimental validity/failure of discrete maximum principle for the Crank-Nicolson scheme



(a) $\Delta x = 0.4$



(b) $\Delta x \simeq 0.286$

Figure 5.10: numerical instability of the Crank-Nicolson scheme

We consider a positive diffusion function, which is symmetric with respect to the vertical axis $x = x_c$ and with its maximum equal to τ reached also at

the point x_c , namely

$$a(x) = \frac{\tau}{1 + \alpha(x - x_c)^2} , \tag{5.10}$$

where the parameter α measures the function's width (see Figure 5.11).

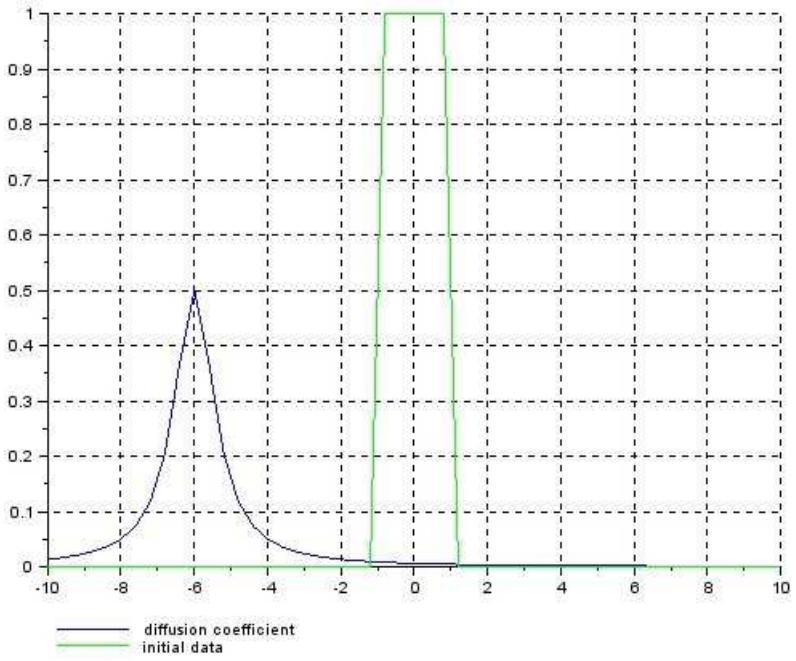


Figure 5.11: initial data and example of heterogeneous diffusion function

We recall the numerical scheme (3.29)-(3.30) derived in Section 3.2 and we analyze that method by varying the following *numerical parameters* :

- the final time of observation T ;
- the CFL number, which regulates the time-step according to (3.34), i.e.

$$\Delta t = CFL \, \frac{\Delta x^2}{2 \, a_{\max}} , \tag{5.11}$$

  where $a_{\max} = \max_{i=0,1,\ldots,N_x} a(x_i)$. Theoretically, we have proven that the condition $CLF \leq 1$ should cover both the l^2-stability and also the maximum/minimum principle, so we want to confirm this requirement with the numerical simulations.

- the amount of grid points N_x, to increase or decrease the space-step;

and the following *modelling variables* :

- the aspect ratio α , to evaluate how the solution behaves under shallow (small α) or sharp (big α) diffusion coefficients, respectively;

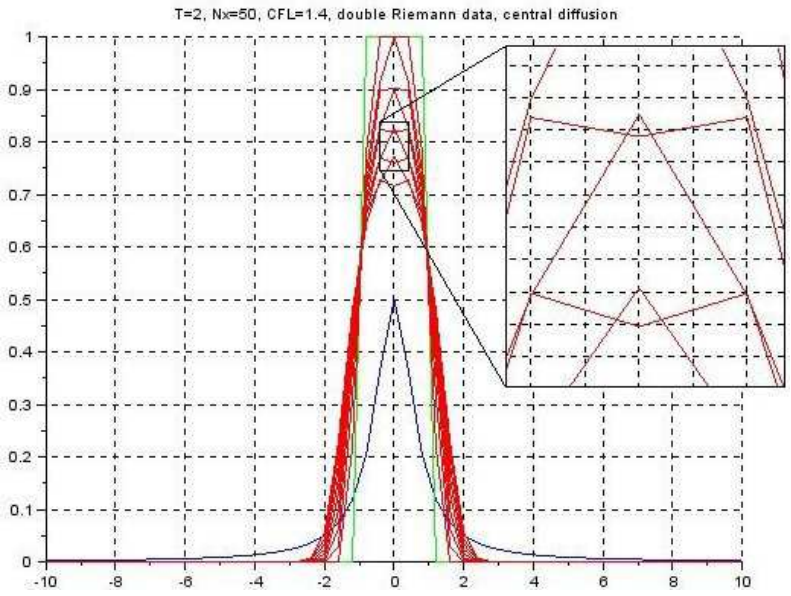
- the maximal position x_c , to estimate the connection between centered/ non-centered diffusion and the shape of the numerical solution.
  We set $x_c = -10 + 20/\beta$, so that *centered diffusion* occurs when $\beta = 2$ and then $x_c = 0$, otherwise we have a diffusion function concentrated on the left ($x_c < 0$) or on the right ($x_c > 0$) of the graph in Figure 5.11.

We remark that modifications of the last two parameters is possible only for theoretical observations because, when working with mathematical models which describe real situations, these data are given and not adjustable.

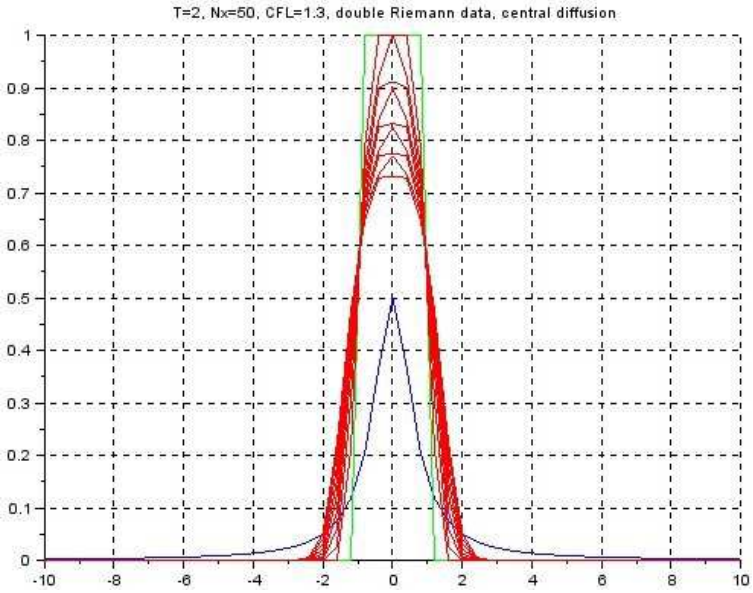
We have observed that symmetric configurations may accidentally confer more stability to the numerical solution, in the sense that the CFL number can be greater than 1 and still the discrete maximum principle be satisfied. Let us consider the case in which $x_c = 0$, so the diffusion function is centered inside the computational domain. We fix $N_x = 50$, $\alpha = 1.5$ and we start with $CFL = 1.4$, for which the maximum principle is violated as expected, that is shown in Figure 5.12(a). Nevertheless, the maximum principle is satisfied with $CFL = 1.3 >> 1$, as we can see in Figure 5.12(b).
So, the CFL parameter has a weaker upper bound than the one predicted by the theoretical results. This peculiar fact is due to the particular symmetry of the problem, where the value $a_{\max}$ is attained by the diffusion function precisely at the center of the computational interval $[-10, 10]$. From the formulas of the scheme (3.29)-(3.30), the coefficient $\frac{a_{i+1}+2a_i+a_{i-1}}{2}$ corresponding to the value u_i^n is clearly bigger than the two others, $\frac{a_{i-1}+a_i}{2}$ for u_{i-1}^n and $\frac{a_i+a_{i+1}}{2}$ for u_{i+1}^n . So, if $a_i = a_{\max}$ the symmetry implies that $a_{i-1} = a_{i+1}$ and, therefore, the value u_i^n contributes less to the computation of u_i^{n+1} , which is finally decreased by the average of smaller values.
Thus, the maximum principle is not violated, since the other factors inside the CFL-condition (5.5) compensate for $CFL > 1$ (but moderate) and the stability requirement is still fulfilled.

Now, if we relocate the diffusion function with $x_c \neq 0$, we recover the same experimental constraint $CFL \leq 0.85$ for the numerical stability as in the previous section, which translates the theoretical condition $CFL < 1$ in the presence of systematic truncation errors. The maximum principle begins to fail in Figure 5.13(a) for $CFL = 0.85$, whilst for $CFL = 0.8$ the scheme perfectly satisfies it as seen in Figure 5.13(b). Moreover, the density diffuses to the left, because we have fixed $x_c < 0$ and the effects of the diffusion are felt more on the left-hand side of the computational domain.

For $CFL = 1.3$, in the present configuration where the diffusion function and the initial data are not anymore symmetric with respect to the same axis, the maximum principle is not satisfied, as reported in Figure (5.14)(a). This produces a real instability, because the numerical scheme gives worse results if we increase the grid points N_x, and so Δx decreases, thus leading to the *Ultraviolet Catastrophe* shown in Figure (5.14)(b), and already dis-



(a) $CFL = 1.4$

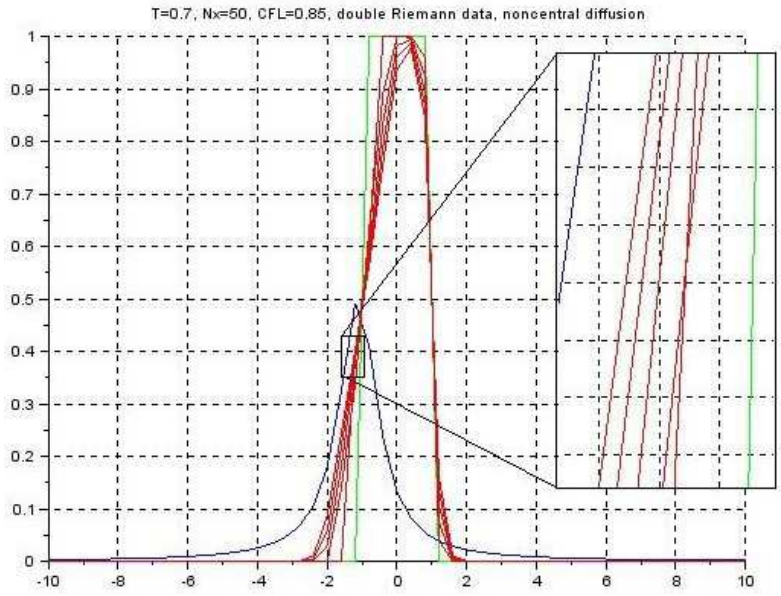


(b) $CFL = 1.3$

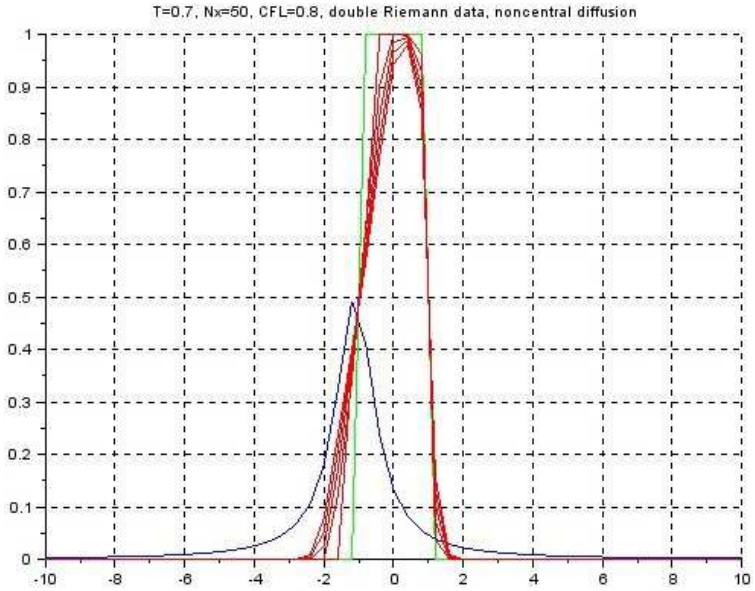
Figure 5.12: persistence of the maximum principle for centered diffusion

cussed in the case of the homogeneous heat equation (refer to Figure (5.8)).

We conclude this section by taking into account the *centered diffusion* in order to analyze the effect of decreasing/increasing the parameter α in (5.10).



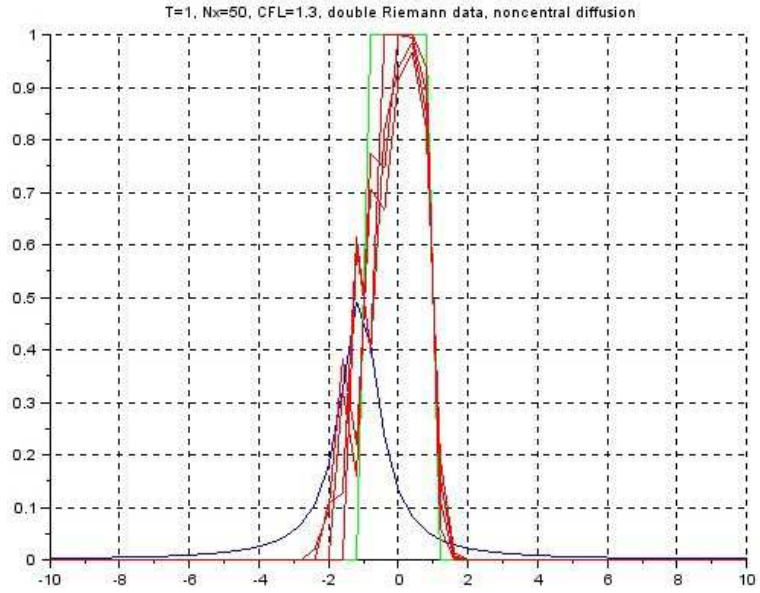
(a) $CFL = 0.85$



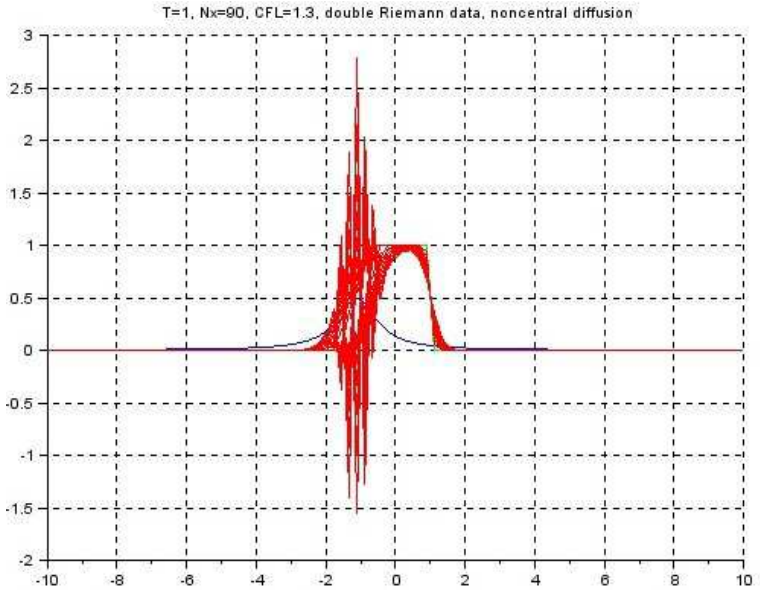
(b) $CFL = 0.80$

Figure 5.13: stable numerical solutions with non-centered diffusion function

We have just seen in Figure 5.12(b) that the discrete maximum principle is satisfied for $\alpha = 1.5$ and $CFL > 1$. We fix $CFL = 1.1$ and everything works as expected in Figure 5.15(a), whereas we obtain the Figure (5.15)(b) when we decrease the width of the diffusion function toward $\alpha = 0.5$.



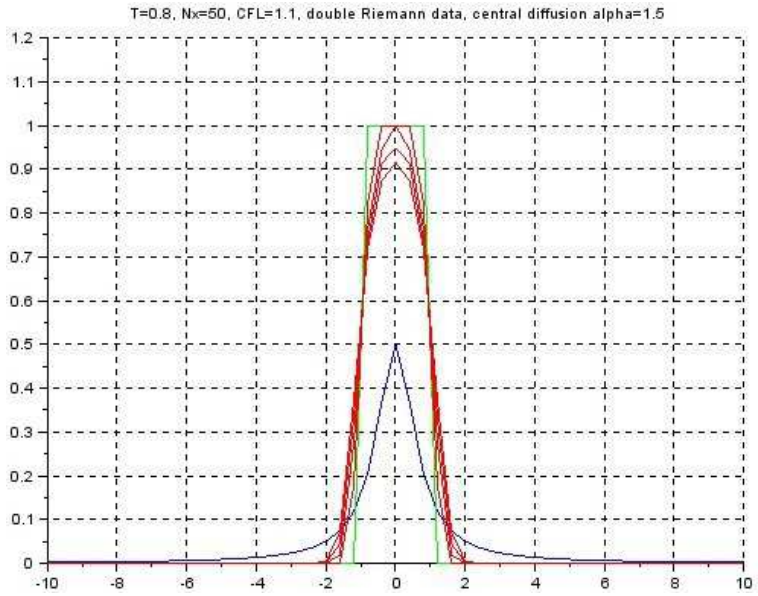
(a) $CFL = 1.3\,,\ \Delta x = 0.4$



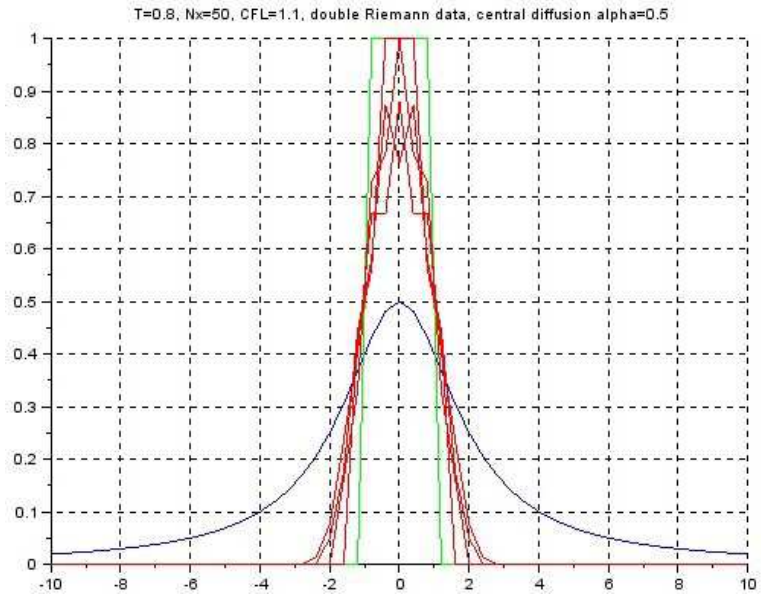
(b) $CFL = 1.3\,,\ \Delta x \simeq 0.22$

Figure 5.14: numerical instabilities with non-centered diffusion function

Therefore, more the diffusion is flatten and stronger the maximum principle is violated, so that this phenomenon depends on the concavity of the diffusion function $a(x)$. Indeed, considering the numerical scheme (3.29)-(3.30), the coefficient of u_i^n has to be less than 1 to satisfy the CFL-condition.



(a) $CFL = 1.1\,,\ \alpha = 1.5$



(b) $CFL = 1.1\,,\ \alpha = 0.5$

Figure 5.15: appearance of instabilities for wider diffusion functions

By performing a Taylor's expansion at the point x_i, we obtain

$$1 - \frac{\Delta t}{2\Delta x^2}\big(a_{i+1} + 2\,a_i + a_{i-1}\big) = 1 - \frac{\Delta t}{2\Delta x^2}\big(4\,a_i + a_i''\,\Delta x^2 + O(\Delta x^4)\big)$$
$$\simeq 1 - \frac{\Delta t}{\Delta x^2}\big(2\,a_i + a_i''\frac{\Delta x^2}{2}\big),$$

with $a_i > 0$ and $a_i'' < 0$ by hypothesis, and the maximum/minimum principle is satisfied if $0 < \frac{\Delta t}{\Delta x^2}(2\,a_i + a_i''\frac{\Delta x^2}{2}) < 1$, for Δx small enough. Consequently, if $2\,a_i + a_i''\frac{\Delta x^2}{2} \leq 0$, the stability condition would be no longer satisfied, and this corresponds to the case of very high values for $|a_i''|$, although the space-step Δx moderates that behaviour. On the other hand, while the expression $2\,a_i + a_i''\frac{\Delta x^2}{2}$ remains positive, higher values of $|a_i''|$ contribute to weaken the stability constraint, thus compensating also for $CFL > 1$.
Indeed, the actual condition used for the simulations is (5.11), so we have

$$1 - \frac{\Delta t}{\Delta x^2}\big(2\,a_i + a_i''\frac{\Delta x^2}{2}\big) = 1 - \frac{CFL}{2\,a_{\max}} \cdot \frac{\Delta t}{\Delta x^2}\big(2\,a_i + a_i''\frac{\Delta x^2}{2}\big),$$

and the scheme needs

$$0 < \frac{CFL}{2\,a_{\max}} \cdot \frac{\Delta t}{\Delta x^2}\big(2\,a_i + a_i''\frac{\Delta x^2}{2}\big) < 1\,,$$

which becomes harder to satisfy when the diffusion function flattens, namely $|a_i''| \to 0$. Moreover, once the instability regime has occurred for $CFL >> 1$, to decrease Δx (by increasing N_x) does not lead to better results because the upper bound in the above inequality is clearly worse. Therefore, since we cannot modify the parameters inside the parenthesis, the right approach is to decrease the CFL number, bringing it back to the stability condition, as shown in Figure 5.16.

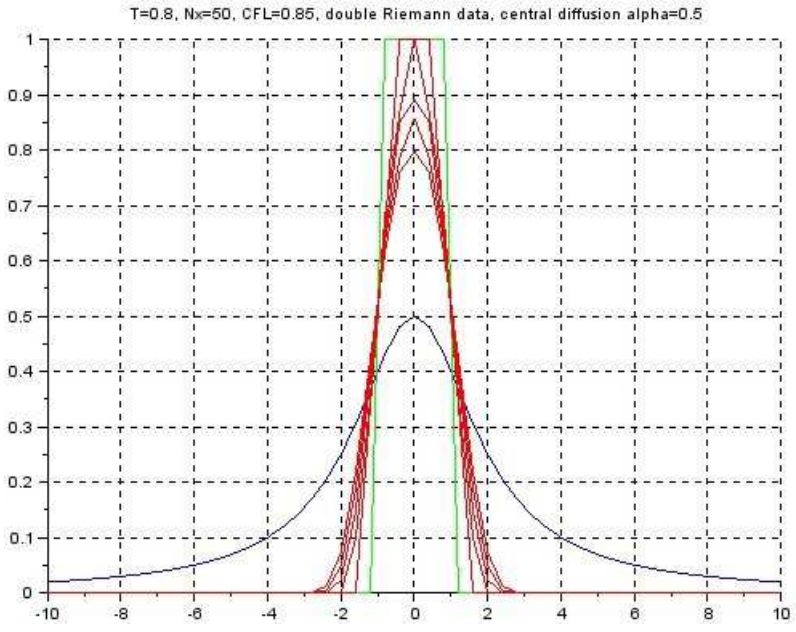


Figure 5.16: fulfillment of the discrete maximum principle

In conclusion, one could accidentally observe weaker stability conditions in case of specific symmetries of the problem, for example centered diffusion

with moderately picked shape, otherwise the maximum/minimum principle of numerical scheme (3.29)-(3.30) is always guaranteed only for $CFL < 1$.

5.4 Numerical tests for two-dimensional anisotropic diffusion

Let us focus on the parabolic equation (4.1) in the case of constant diffusion coefficients, under the condition (1.7) for the positive definiteness.
We consider the initial data shown in Figure 5.17, i.e.

$$u_0(x,y) = \begin{cases} 1 & |x^2+y^2| \leq 1 \\ 0 & otherwise \end{cases}$$

and (homogeneous) Neumann boundary conditions (5.2).

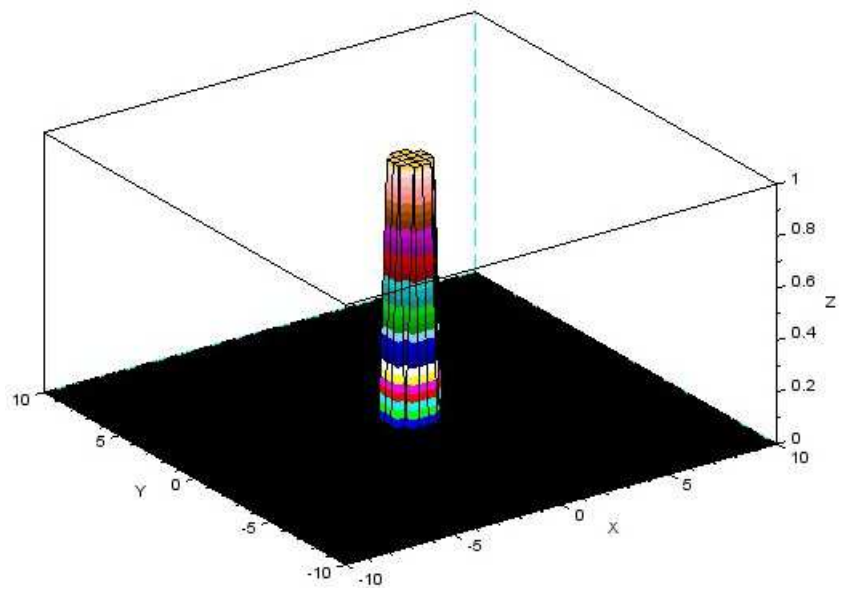


Figure 5.17: initial data for two-dimensional anisotropic diffusion equation

We will treat separately the simpler case $c = 0$, for which we verify numerically the theoretical conditions for the fulfillment of the discrete maximum principle established in Section 4.2.1. Then, we activate the term involving mixed derivatives, to check that effectively the Standard Discretization fails the maximum/minimum principle, and we need the Nonnegative Discretization to solve such a problem.

5.4.1 The purely diagonal case ($c = 0$)

We set $a = 3$, $b = 1$, and the computational domain is given by (5.1).
The grid points are specified by $N_x = N_y = 50$, so that $\Delta x = \Delta y = 0.4$.
The numerical scheme (4.9) has time-step computed according to (4.14) as

$$\Delta t = CFL \cdot \frac{\min\{\Delta x^2, \Delta y^2\}}{4 \max\{a, b\}}, \tag{5.12}$$

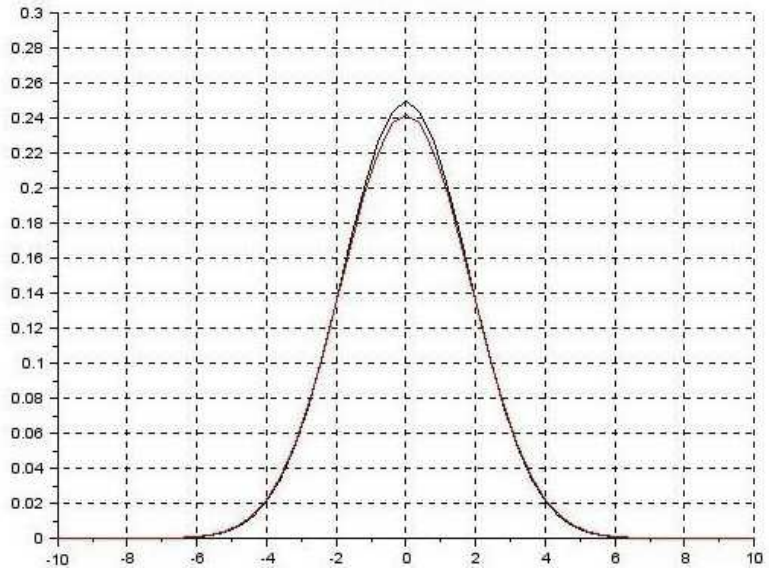
which corresponds to the stability constraint for $CFL < 1$.
During the numerical simulations, we have observed that the discrete maximum principle is satisfied even if the CFL number becomes greater than 1.
We refer to Figure (5.18), where the numerical solution is calculated with $CFL = 1.4$, and plotted at two different times, the black graph at $T = 0.5$ and the red graph at $T = 0.5 + \Delta t$, respectively. For both components along the x and y axes, the maximal/minimal value of the solution at time $T = 0.5 + \Delta t$ is always below/above that at time $T = 0.5$, which represents exactly the validity of the maximum/minimum principle.
However, this behaviour is not preserved in Figure (5.19), where the solution is calculated with $CFL = 1.45$, and a violation of the maximum/minimum principle suddenly occurs.
These facts suggest that other terms are participating in the characterization of the CFL-condition, which we are possibly neglecting, and an explanation comes directly from the improved stability condition (4.26). Indeed, if we replace (5.12) in the numerical codes with the following difinition

$$\Delta t = CFL' \cdot \frac{\min\{\Delta x^2, \Delta y^2\}}{2 \max\{a, b\}} S^{-1}, \tag{5.13}$$

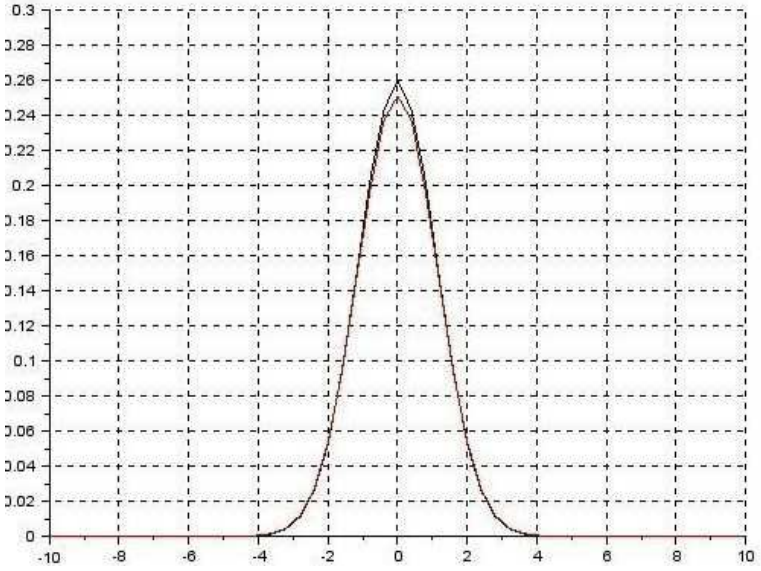
where $S = \frac{a}{\Delta x^2} \cdot \frac{\min\{\Delta x^2, \Delta y^2\}}{\max\{a,b\}} + \frac{b}{\Delta y^2} \cdot \frac{\min\{\Delta x^2, \Delta y^2\}}{\max\{a,b\}}$, the two parameters CFL and CFL' are correlated through

$$CFL' = CFL \cdot \frac{S}{2} \leq CFL. \tag{5.14}$$

Now, in the present case, we have $S = 1 + \frac{1}{3} \simeq 1.33$, and therefore $CFL \leq 1.4$ implies $CFL' \leq 0.931$ for the experimental setting.
The constraint $CFL' < 1$ is totally coherent with the theoretical results : the discrete maximum principle is not satisfied when performing simulations with $CFL' = 1.1$, as shown in Figure 5.20, and moreover numerical instabilities clearly pollute the experiments; on the other hand, accurate results are recovered with $CFL' = 0.95$, and we observe in Figure 5.21 that the maximal/minimal value of the numerical solution calculated at $T = 0.5$ (black graph) is always below/above that calculated at $T = 0.5 + \Delta t$ (red graph), that corresponds to the fulfillment of the maximum/minimum principle.



(a) plot of numerical solution with respect to the x-axis



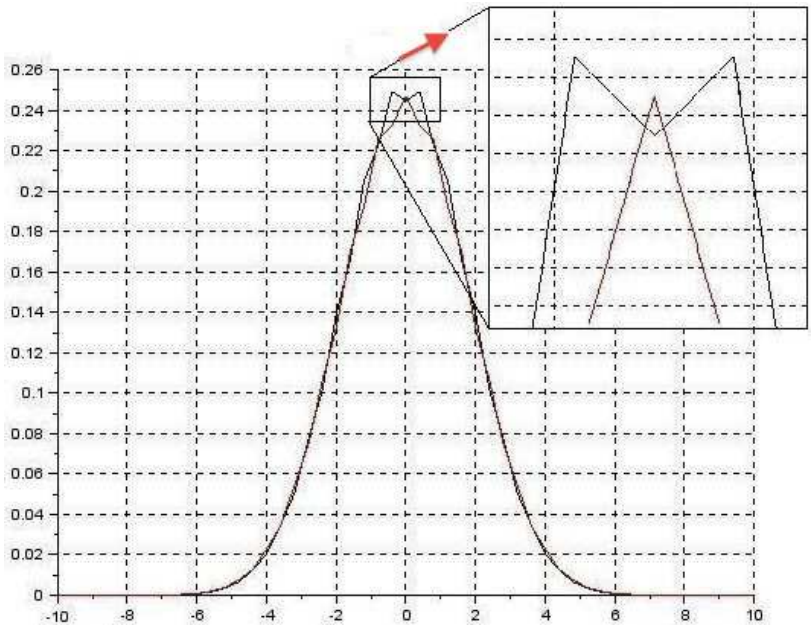
(b) plot of numerical solution with respect to the y-axis

Figure 5.18: $CFL = 1.4\,,\ T = 0.5$ (black line) versus $T = 0.5 + \Delta t$ (red line)

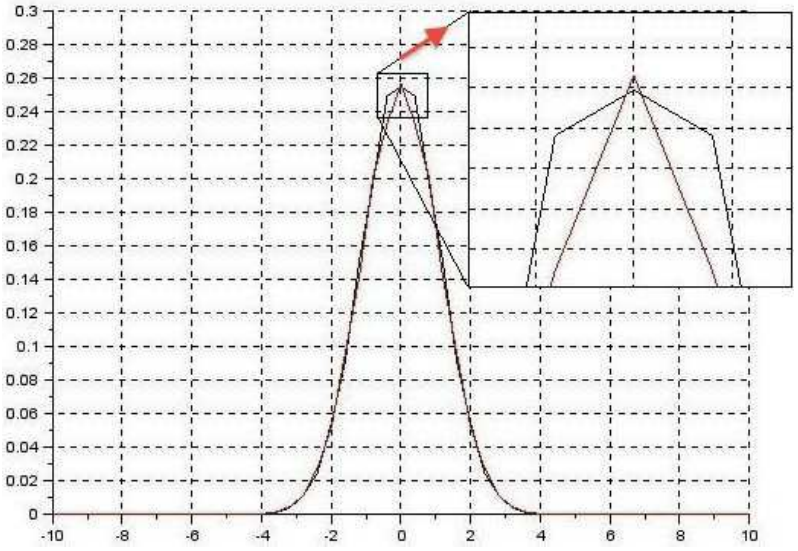
5.4.2 Taking into account the mixed derivatives ($c \neq 0$)

Finally, we focus on the fully anisotropic diffusion equation (4.7), with constant coefficients $a = 3\,,\ b = 1\,,\ c = 1.45\,,$ so that $c^2 < a\,b\,,$ which guarantees the positive definiteness property (1.7).

We have derived in Section 4.2.3 the Standard Discretization method,



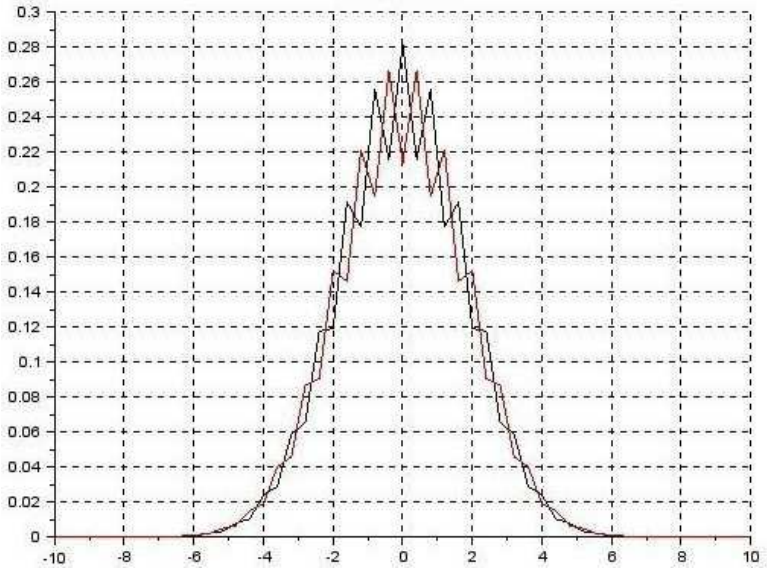
(a) plot of numerical solution with respect to the x-axis



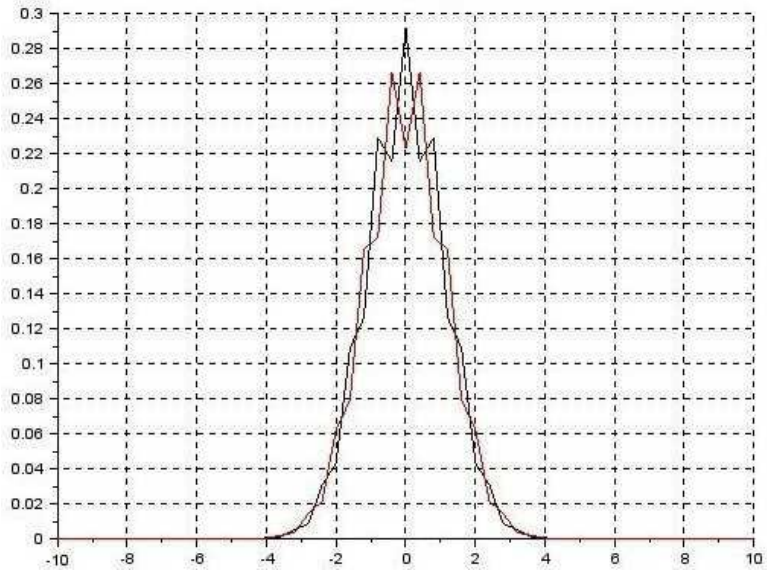
(b) plot of numerical solution with respect to the y-axis

Figure 5.19: $CFL = 1.45$, $T = 0.5$ (black line) versus $T = 0.5 + \Delta t$ (red line)

that produces the numerical scheme (4.20), but unfortunately the discrete maximum principle may fail for such an approximation, due to the possible negativity of the elements inside the stencil (refer also to Table 2.2).
We set $N_x = N_y = 70$ and $T = 0.5$, together with the CFL-condition (5.12)



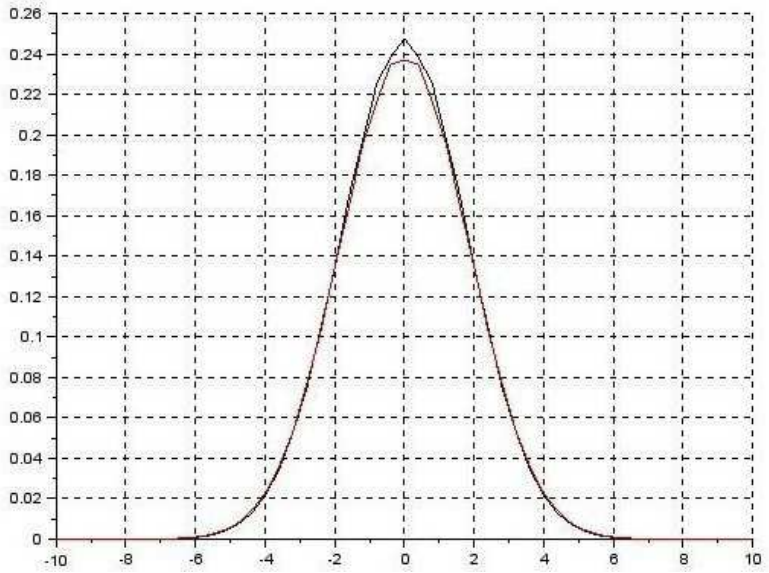
(a) plot of numerical solution with respect to the x-axis



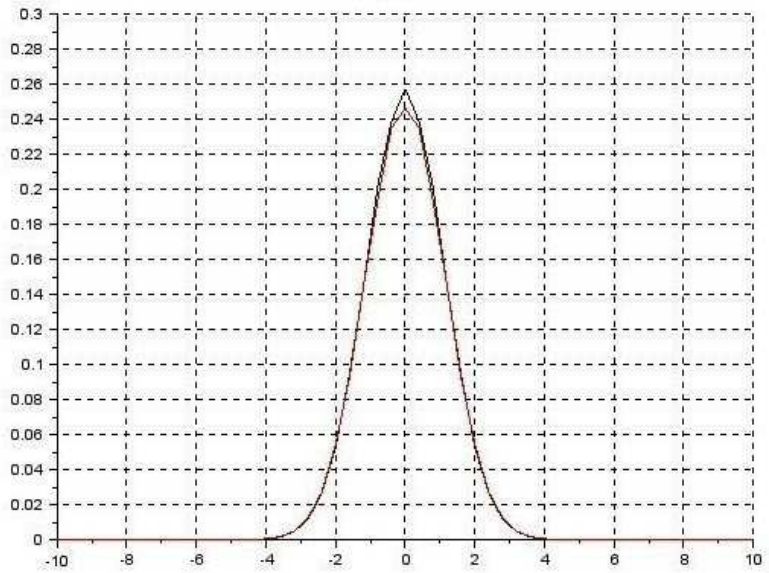
(b) plot of numerical solution with respect to the y-axis

Figure 5.20: $CFL' = 1.1\,,\ T = 0.5$ (black line) versus $T = 0.5 + \Delta t$ (red line)

already used for the diagonal case above, which should ensures at least the l^2-stability of the present method. We fix $CFL = 0.5$ and, indeed, we observe in Figure 5.22 a failure of the maximum/minimum principle, despite $CFL < 1$ and the numerical solution does not exhibit spurious oscillations. We have reported only the graph of the numerical solution with respect



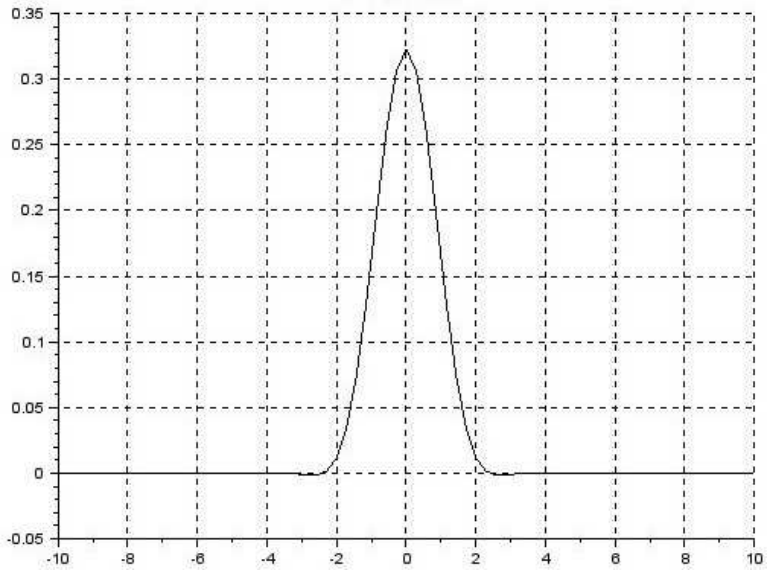
(a) plot of numerical solution with respect to the x-axis



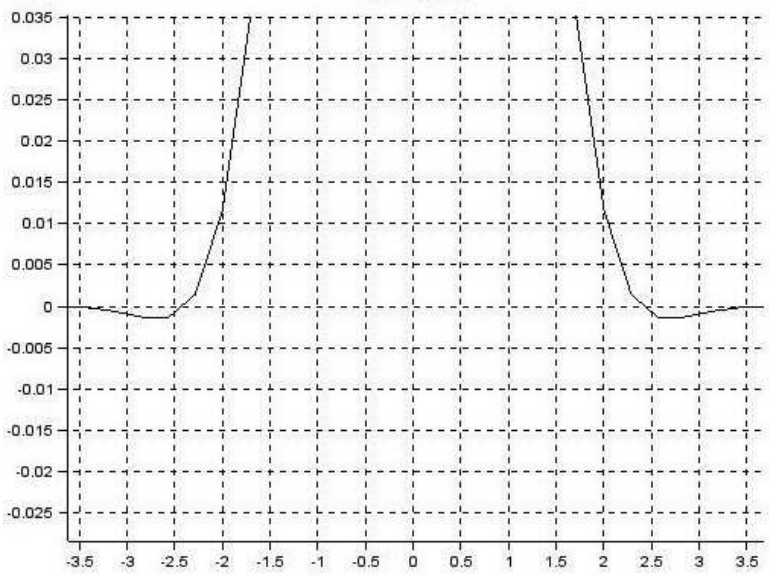
(b) plot of numerical solution with respect to the y-axis

Figure 5.21: $CFL' = 0.95\,,\ T = 0.5$ (black line) versus $T = 0.5 + \Delta t$ (red line)

to the y-axis, fixing x at the center of the interval $[-10, 10]$, to clearly see how the density becomes artificially negative. This violation of the discrete maximum principle is manifested by the loss of the *non-negativity property*, which is fundamental for the physical meaning of the problem, as discussed in Section 1.4 through the implication (1.24).



(a) plot of numerical solution with respect to the y-axis

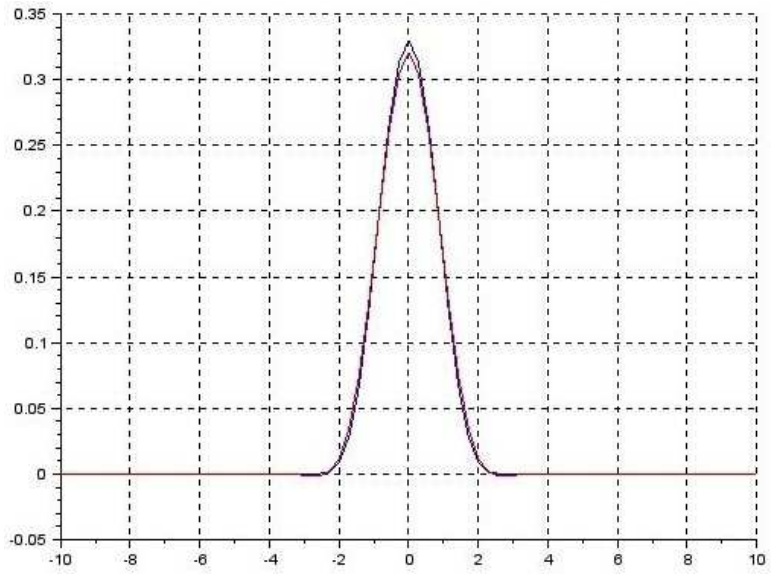


(b) zoom in the areas where the numerical solution becomes negative

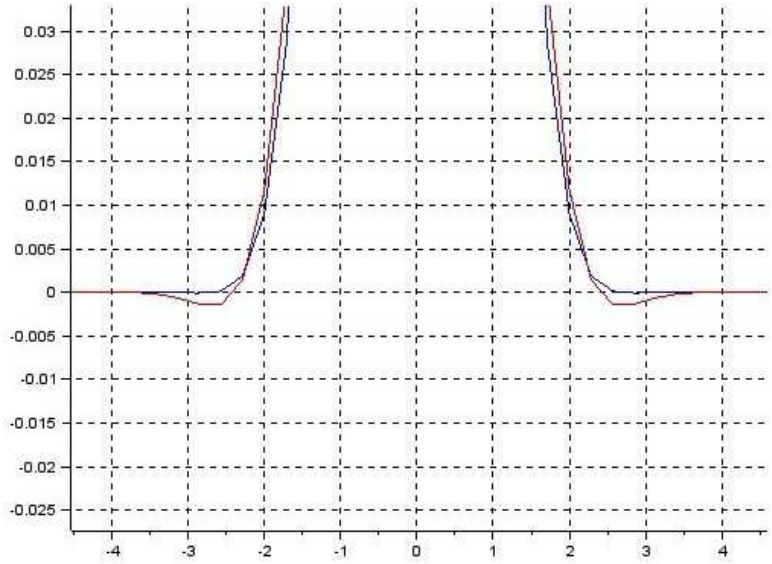
Figure 5.22: failure of non-negativity for the Standard Discretization

According to the theoretical results stated in Section 4.2.3, in order to secure overall stability properties, we must apply the Nonnegative Discretization method (4.21), with the same numerical parameters as before.
We compare the two approaches in Figure 5.23, for $CFL = 0.99$, and we have plotted the numerical solution issued from the Standard Discretization

(red graph) against the one calculated with the Nonnegative Discretization (blue graph), which remains always greater than 0. Therefore, non-physical negative values are eliminated when employing a nonnegative approximation, and the numerical solution satisfies the maximum/minimum principle.



(a) plot of numerical solutions with respect to the y-axis



(b) zoom in the areas where the numerical solution may become negative

Figure 5.23: Standard (red line) versus Nonnegative (blue line) method

Let us concentrate on the CFL-condition for the Nonnegative Discretization. In the previous simulations, we put the constraint (5.12) with $CFL<1$,

which is actually more restrictive than the theoretical requirement (4.23) for the case $c \neq 0$. This suggests that, again, we are considering larger positive terms on the parameter CFL, that could brings to numerical solutions which are stable also for CFL numbers greater than 1. Effectively, we have observed that the discrete maximum principle is satisfied until $CFL < 1.98$, as shown in Figure 5.24.

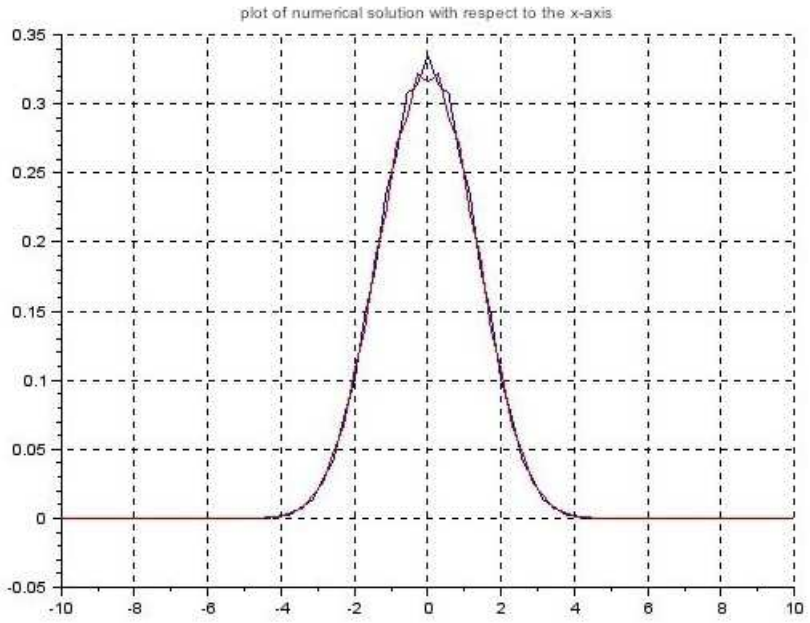


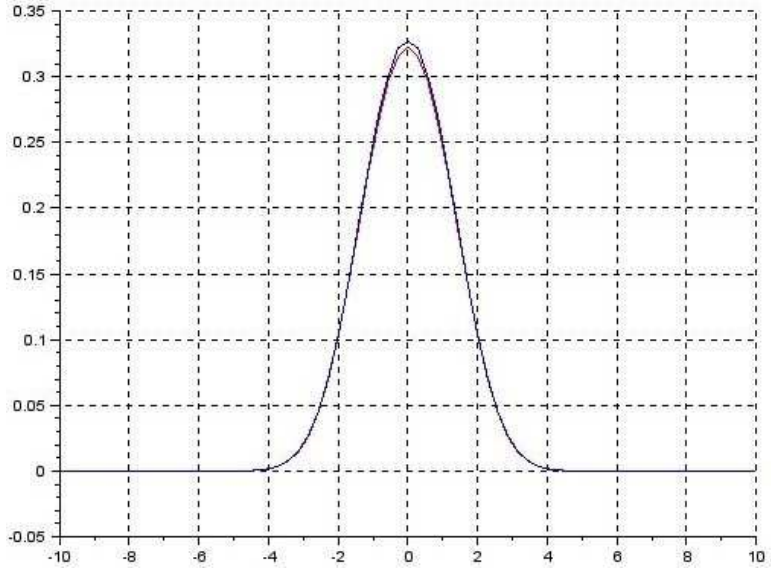
Figure 5.24: appearance of instabilities for $CFL \geq 1.98$

These numerical results are coherent with the theoretical statements in Section 4.2.3, since a sharper CFL-condition for the Nonnegative Discretization is actually given by (4.24), which is connected to (5.12) through a relationship between the CFL parameters similar to (5.14). From (4.24) we have that $0 < S < 2$, and for the present case $S = 1 + \frac{1}{3} - \frac{1.45}{3} \simeq 0.85$, so the stability constraint gains a factor 0.425 which allows to almost double the original CFL number, as reported in Figure 5.24.

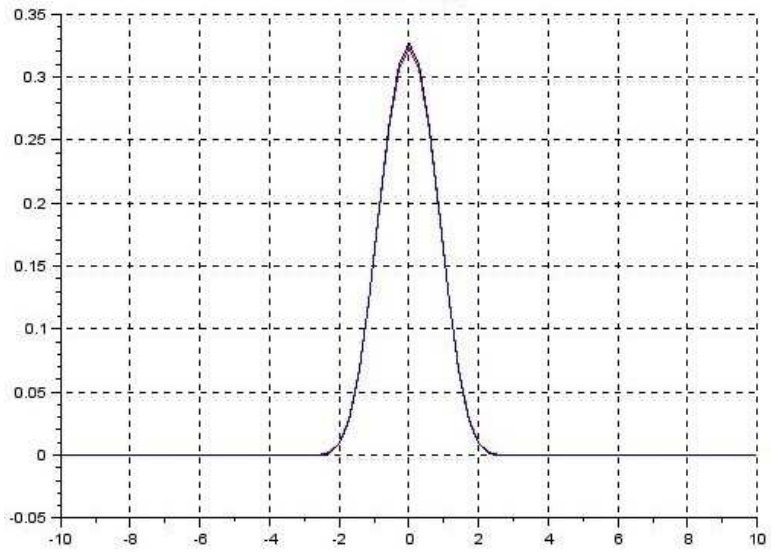
Remark 15. *It is worthwhile stressing that the actual gain in the stability constraint* (4.24) *depends very much on the values of the diffusion coefficients* a, b, c*, and the space-steps* $\Delta x, \Delta y$*, or rather on the ratio between those values, showing that strong anisotropy stabilizes the nonnegative scheme, as well as different space-steps (a type of non-uniformity of the spatial grid).*

Setting $\Delta t = CFL' \frac{\min\{\Delta x^2, \Delta y^2\}}{2 \max\{a,b\}} S^{-1}(a,b,c)$, with $S(a,b,c)$ defined as in (4.24), we report in Figure 5.25 the numerical solution evaluated for $CFL' = 0.82$ at time $T = 0.5$ and then at time $T = 0.5 + \Delta t$, which satisfies the discrete maximum principle. This property starts failing for $CFL' = 0.83$, as observed

in Figure 5.26, where the violation of the maximum/minimum principle is imperceptible but however it occurs as the red graph overcomes the blue one.



(a) plot of numerical solutions with respect to the x-axis



(b) plot of numerical solutions with respect to the y-axis

Figure 5.25: nonnegative solution at $T = 0.5$ (blue line) with its maximum above that at $T = 0.5 + \Delta t$ (red line)

We conclude this section about two-dimensional anisotropic diffusion and the discrete maximum principle by showing the three-dimensional plots of

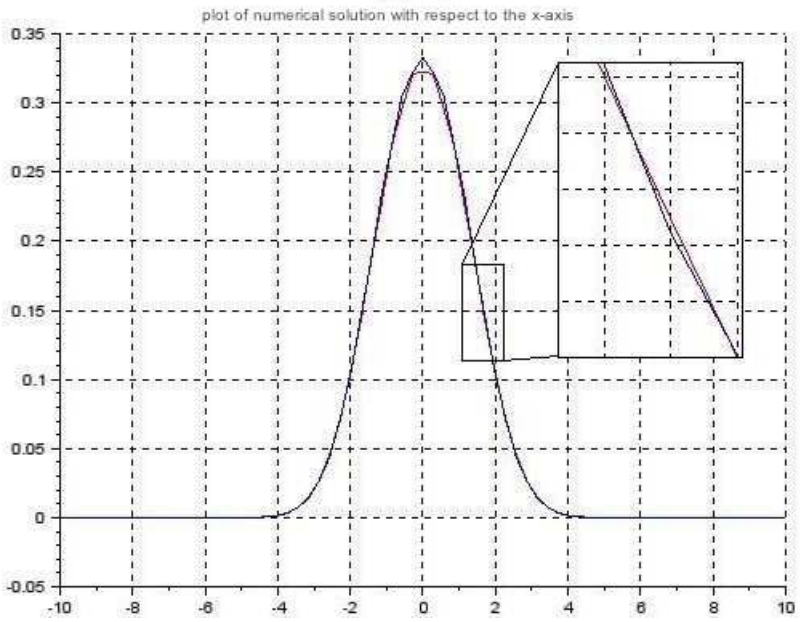


Figure 5.26: failure of the discrete maximum principle for $CFL \geq 0.83$

the numerical solution at time $T = 0.5$, for the diffusion tensor with $c = 0$ in Figure 5.27, and with $c \neq 0$ in Figure 5.27. Obviously, the introduction of the mixed derivatives makes the diffusion front to move also diagonally toward the axis $y = x$, unlike the case of purely diagonal diffusion where the density spreads only along the x and y axes.

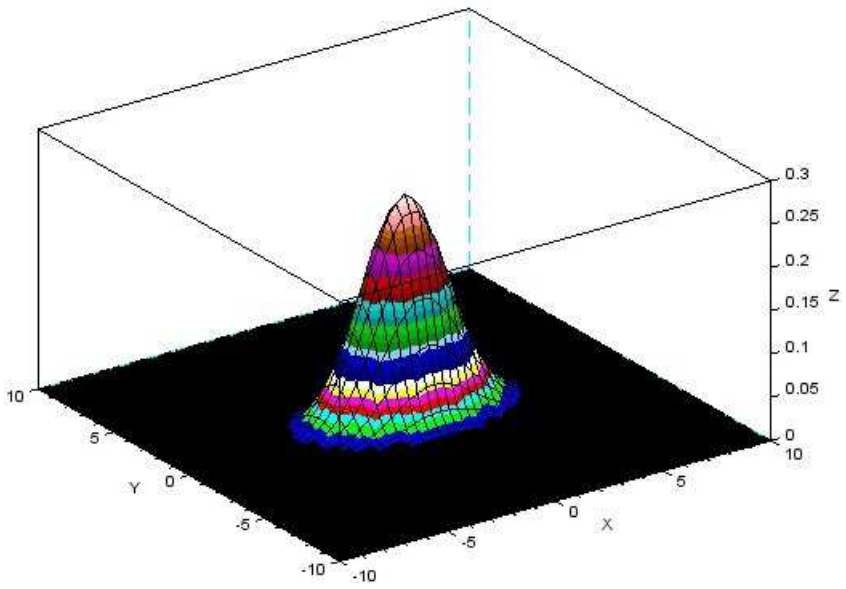


Figure 5.27: numerical solution in the case $c = 0$ with $a > b$

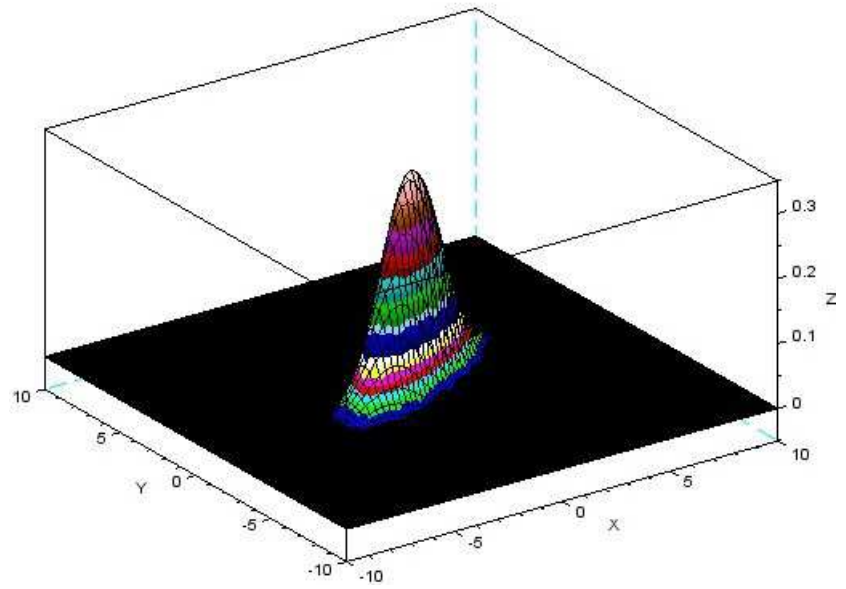


Figure 5.28: numerical solution in the case $c \neq 0$ with $a > b$

Bibliography

[1] P.J. Basser, D.K. Jones, Diffusion-tensor MRI: theory, experimental design, and data analysis – a technical review, NMR Biomed. 15 (2002) 456-467.

[2] C. Beaulieu, The basis of anisotropic water diffusion in the nervous system – a technical review, NMR Biomed. 15 (2002) 435-455.

[3] H. Brezis, Analyse fonctionnelle. Théorie et applications. 2nd Edition, Mathématiques Appliquées pour la Maîtrise, Dunod, Paris, 1999.

[4] A. Chauviere, T. Hillen, L. Preziosi, Modeling cell movement in anisotropic and heterogeneous network tissues, Netw. Heterog. Media 2 (2007) 333-357.

[5] R.F. Curtain, A.J. Pritchard, Functional analysis in modern applied mathematics, Mathematics in Science and Engineering 132, Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York, 1977.

[6] C.M. Dafermos, Hyperbolic conservation laws in continuum physics, Fundamental Principles of Mathematical Sciences 325, Springer-Verlag, Berlin, 2000.

[7] L. Dascal, N.A. Sochen, A maximum principle for Beltrami color flow, SIAM J. Appl. Math. 65 (2005) 1615-1632.

[8] R. de Boer, Trends in continuum mechanics of porous media, Theory and Applications of Transport in Porous Media 18, Springer, Dordrecht, 2005.

[9] Z. Dian-lin, C. Shao-chun, W. Yun-ping, L. Li, W. Xue-mei, X.L. Ma, K.H. Kuo, Anisotropic thermal conductivity of the 2D single quasicrystals: Al65Ni20Co15 and Al62Si3Cu20Co15, Phys. Rev. Lett. 66 (1991) 2778-2781.

[10] M. A.T. Elshebli, R. Horváth, Discrete maximum principle for the finite element solution of linear non-stationary diffusion-reaction problems, Applied Mathematical Modelling 32 (2008) 1530-1541.

[11] I. Faragó, R. Horváth, Discrete maximum principle and adequate discretizations of linear parabolic problems, SIAM J. Sci. Comput. 28 (2006), no. 6, 2313-2336.

[12] C. Grossmann, H.-G. Roos, Numerical Treatment of Partial Differential Equations. Translated and revised by Martin Stynes. Springer-Verlag, Berlin Heidelberg, 2007.

[13] A. Häcker, A mathematical model for mesenchymal and chemosensitive cell dynamics, J. Math. Biol. 64 (2012) 361-401.

[14] H.L.P. Harpold, E.C. Alvord, Jr., K.R. Swanson, The evolution of mathematical modeling of glioma proliferation and invasion, J. Neuropathol. Exp. Neurol. 66 (2007) 1-9.

[15] T. Hillen, K.J. Painter, M. Winkler, Anisotropic diffusion in oriented environments can lead to singularity formation, European J. Applied Math. (2013) in press.

[16] R.A. Horn, C.R. Johnson, Matrix analysis, 2nd Edition, Cambridge University Press, Cambridge, 2013.

[17] W.W. Hwu Editor, GPU computing gems, Emerald and Jade Editions, Applications of GPU Computing Series, Morgan Kaufmann Publishers, Elsevier, 2011.

[18] Y.-K. Kwok, Mathematical models of financial derivatives, second ed., Springer Finance, Springer, Berlin, 2008.

[19] O.A. Ladyženskaja, V.A. Solonnikov, N.N. Ural'ceva, Linear and quasilinear equations of parabolic type. (Russian) Translated from the Russian by S. Smith, Translations of Mathematical Monographs 23, American Mathematical Society, Providence, 1968.

[20] S. Larsson, V. Thomée, Partial Differential Equations with Numerical Methods,Texts in Applied Mathematics 45, Springer-Verlag Berlin Heidelberg, 2009.

[21] R.J. Leveque, Finite-Volume Methods for Hyperbolic Problems, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2004.

[22] R.J. LeVeque, Finite Difference Methods for Ordinary and Partial Differential Equations. Steady-State and Time-Dependent Problems, SIAM, Philadelphia, 2007.

[23] X. Li, W. Huang, An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems, J. Comput. Phys. 229 (2010), 8072-8094.

[24] H.W. McKenzie, E.H. Merrill, R.J. Spiteri, M.A. Lewis, How linear features alter predator movement and the functional response, Interface Focus 2 (2012) 205-216.

[25] K.W. Morton, D.F. Mayers, Numerical Solution of Partial Differential Equations. An Introduction, 2nd Edition, Cambridge University Press, New York, 2005.

[26] P. Mosayebi, D. Cobzas, M. Jagersand, A. Murtha, Stability effects of finite difference methods on mathematical tumor growth model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, San Francisco, 2010, 1-8.

[27] I.J. Parrish, J.M. Stone, Nonlinear evolution of the magnetothermal instability in two dimensions, Astrophys. J. 633 (2005) 334-347.

[28] A. Berman, R.J. Plemmons, Nonnegative Matrices in the Mathematical Sciences, Classics in Applied Mathematics 9, SIAM, Philadelphia, 1994.

[29] M.H. Protter, H.F. Weinberger, Maximum Principles in Differential Equations, Springer-Verlag, New York, 1984.

[30] A. Quarteroni, R. Sacco and F. Saleri, Numerical Mathematics, Texts in Applied Mathematics 37, Springer-Verlag, New york, 2000.

[31] M. Saadatfar, M. Sahimi, Diffusion in disordered media with long-range correlations: anomalous, Fickian, and superdiffusive transport and log-periodic oscillations, Phys. Rev. E 65 (2002) 036116.

[32] A.R. Sanderson, C.R. Johnson, R.M. Kirby, L. Yang, Advanced reaction-diffusion models for texture synthesis, J. Graph. Tools 11 (2006) 47-71.

[33] G. Strang, Linear Algebra and Its Applications, 4th Edition, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 2006.

[34] J.C. Strikwerda, Finite Difference Schemes and Partial Differential Equations, 2nd Edition, SIAM, Philadelphia, 2004.

[35] K.R. Swanson, C. Bridge, J.D. Murray, E.C. Alvord Jr., Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion, J. Neurol. Sci. 216 (2003) 1-10.

[36] J.L. Vázquez, The porous medium equation. Mathematical theory, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, Oxford, 2007.

[37] R.F. Warming, B.J. Hyett, The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods, Journal of Computational Physics 14 (1974), no. 2, 159-179.

[38] J. Weickert, Anisotropic diffusion in image processing, ECMI Series, Teubner-Vertlag, B.G. Teubner Stuttgart, 1998.

[39] V. Yépez, P. Sapun, Modeling of the spread of populations through reaction systems with anisotropic diffusion using Finite Difference and Finite Element methods, MasterÕs Thesis in Mathematical Engineering, Erasmus Mundus M.Sc. MathMods, University of L'Aquila, 2012.