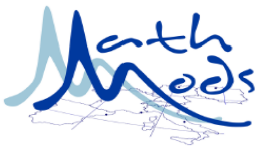


Application of the Sparse grid technique to Discontinuous Galerkin methods: some simple hyperbolic problems

Author:

Saverio Castelanelli

Advisor:

Prof. B. Ayuso de Dios

A thesis submitted in partial fulfillment
of the requirements for the degree of Master of Science
in Mathematical Modelling in Engineering: Theory, Numerics, Applications[1].

Universitat Autònoma de Barcelona
Barcelona, Catalunya, Spain

[1]Part of Erasmus Mundus Program sponsored by EU Commission scholarship and partially by CRM grant.

Ringraziamenti

Un ringraziamento particolare alla Professoressa Blanca Ayuso de Dios, con cui ho avuto la fortuna di poter lavorare a questa tesi, per il fatto di aver sempre potuto, dall'inizio alla fine di questa avventura, contare sul suo aiuto e sostegno nei momenti di difficoltà, per la pazienza e il tempo dedicatomi, e per il suo contagiante entusiasmo nel tema e nella matematica tutta.

Vorrei anche ringraziare Ilario Mazzieri del Politecnico di Milano (I) e Xiaozhe Hu dell'università di Penn State (US) per l'interesse mostrato in questo lavoro e per i loro utili commenti e suggerimenti per la parte riguardante l'implementazione.

Infine un grazie speciale ai miei genitori, Silvana e Sergio, per tutto ciò che hanno fatto per me in questi $27 + \varepsilon$ *anni.*

Contents

Chapter 1

Introduction

1.1 Motivation

High dimensional problems arise in several real life applications (of different nature) in science and technology. Many of them can be expressed by mathematical models; the valuation of Stock options in mathematical finance, data-mining problems related to medicine or the underlying problems in the controlled fusion in the area of plasma physics. The numerical approximation of such high dimensional models, by conventional or classical methods, often run into difficulties due in part to the huge number of unknowns that are required to provide a reasonable solution. In fact, at the present time classical numerical methods for approximating many of these problems do not supply a feasible option. In spite of the advance of big computers in recent times, classical approximation of high dimensional problems involves enormous storage requirements and extremely large computational complexity. Traditionally, probabilistic methods of Monte Carlo type have been frequently used, but depending on the specific application, the approximation is far from being satisfactory. This is indeed the case, in the numerical approximation of plasma physics problems, typically modeled by kinetic equations, for which deterministic or Eulerian solvers (based on a fixed grid) can produce more accurate descriptions. However unless they were specifically tailored, their cost increases exponentially with the dimension of the problem, which forbids their use in many of the real practical applications.

By using an eulerian classical scheme; say a finite element method, if we consider a uniform grid with piecewise d-polynomial functions over a bounded domain, the approximation error $||f - f^h||$ (in some norm) is of the order $O(h^\alpha)$ where h refers to the mesh size and α is a parameter depending on the smoothness of the exact solution f and the polynomial degree used for its approximation f^h. This complexity estimate translates into $O(N^d)$ grid points or degrees of freedom, where N is the number of grid points in one coordinate direction at the boundary. Thus, the computational cost and storage requirements grow exponentially with the dimensionality of the problem. We encounter the so-called *curse of dimensionality*.
The sparse grid approach, introduced by Zenger (1991), allow to partially overcome this problem, affecting only slightly the accuracy of the numerical solution. By using a higher-dimensional multiscale basis derived from a one-dimensional multiscale basis (by a tensor product construction) a discretization based on the sparse grid technique involve only $O(N(\log N)^{d-1})$ degrees of freedom, where d denotes the underlying dimension of the problem. The accuracy (in L^2-norm) obtained with, for instance piecewise linear basis functions is of order $O(N^{-2}(\log N)^{d-1})$ if the solution has bounded second mixed derivatives. This feature renders the technique particularly

suitable to tackle problems of high dimensionality.

In the last decades, this technique has been applied in various contexts. In the context of numerical analysis, the theory for linear *conforming finite element methods* has been widely studied and applied, in particular, to tackle problems whose underlying model is a partial differential equation of elliptic or parabolic type.
Nevertheless, in science, many high dimensional problems are of hyperbolic types. In kinetic theory, for example, system of equations such as the Vlasov-Poisson and the Vlasov-Maxwell belong to this category.
In this work we explore the application of the sparse grid technique to solve some hyperbolic problems by means of a discontinuous Galerkin (DG) approach.

Based on a totally discontinuous finite element spaces, DG methods are extremely versatile and have numerous attractive features: they conserve local properties, can easily handle irregularly refined meshes and vary the approximation degrees of the polynomials from element to element (hp-adaptivity). Moreover, DG mass matrices are block-diagonal, and, so, they can be inverted at a low computational cost, giving rise to very efficient time-stepping algorithms in the context of time-dependent problems. DG methods have undergone a rapid development in recent years, and nowdays new methods are designed to solve more complex linear and nonlinear problems. On the other hand, DG methods usually require more degrees of freedom than their conforming relatives. Furthermore, since they are finite element methods, they do also suffer the *curse of dimensionality* when approximating high dimensional problems.

The aim of this work is to tackle some simple hyperbolic problems by using a DG-method combined with the sparse grid technique. We would like to reduce the number of degrees of freedom, trying to take advantage, at the same time, of the nice features of the DG-methods.
Since in DG-methods discontinuities on interelement boundaries are allowed, the construction of a multilevel hierarchy of spaces imposes extra-difficulties as compared to the conforming case. We also note that, to the best of our knowledge, this issue has not been studied before in literature nor in the elliptic nor in the hyperbolic case. For these reasons, in this work, we have started by studying a simple hyperbolic equation and by considering the lowest order polynomials ($\mathbf{P}^0$) for the approximation.

The second chapter is devoted to the one-dimensional transport equation. Here the sparse grid technique reduces to use a multilevel hierarchy. What we observe in one dimension is just a change of basis, but this step is fundamental for understanding the construction of the multilevel hierarchy of spaces and their corresponding basis functions in higher dimensions.

In the third chapter, we consider the two-dimensional transport equation for both constant and variable coefficients. Here, the main issues concern the construction and the way of choosing the two-dimensional hierarchical subspaces when we apply the idea of the sparse grid. For the sparse grid we show the convergence of the method and the conservation of the total discrete mass. Unfortunately, with the new proposed schemes, we do not preserve the positivity. We have explored two different strategies to overcome this issue, although none of them gave a completely satisfactory answer. The former consist to add sequentially some of the subspaces that were deleted from the full space in the sparse construction. However, it turns out that to get a positive approximation all but one subspaces need to be included which ruins completely the sparse idea. The latter approach consist in using a modified projection at the initial time, which

ensures the positivity of the discrete initial data. Here the approximation preserves positivity, but the overall accuracy (and order) of sparse-DG method, degrades slightly.

In the fourth chapter, we use the method to approximate a more challenging problem: the Vlasov-Poisson system of equations with periodic boundary conditions. The nonlinearity of the problem, given by the coupling of the Poisson problem and the Vlasov-equation (transport equation), requires particular attention. We consider the LDG-method to approximate the Poisson equation and the sparse DG-method to solve the Vlasov equation. Here as well, we show convergence and total discrete mass conservation. We also state the result with the error analysis of the method.

In the last chapter the implementation of the different methods is given.

Finally we wish to emphasize that we demonstrate numerically that the new methods (resulting from the combination of sparse grid and DG) allow to significantly reduce the number of degrees of freedom in the discretization with essentially no loss of accuracy. This, however, does not guarantee that the computational time to solve the problem scales down at the same rate. The development and implementation efficient algorithms tailored to sparse DG finite elements will be subject of future research. The final aim would be to solve the VP system at overall computational costs that are proportional to the number of degrees of freedom in the sparse discretization.

In the following section, we present the function spaces and some notations which we will use throughout the work.

1.2 Function spaces and basic notations

Given a domain $\Omega \subset \mathbb{R}^2$. A typical point in $\mathbb{R}^2$ will be denoted by (x, y).
Let $\alpha = (\alpha_1, \alpha_2)$ with $\alpha_j \geq 0$ for $j = 1, 2$, we call α a *multi-index* and denote by

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x^{\alpha_1} \partial y^{\alpha_2}}$$

the *differential operator* of order $|\alpha| = \alpha_1 + \alpha_2$.
For a function f we will usually write its partial derivatives as

$$\frac{\partial^{|\alpha|}}{\partial x^{\alpha_1} \partial y^{\alpha_2}} f = f\underbrace{x \dots x}_{\alpha_1}\underbrace{y \dots y}_{\alpha_2}.$$

For any nonnegative integer m let $C^m(\Omega)$ be the vector space consisting of all functions φ which, together with all their partial derivatives $D^\alpha \varphi$ of order $|\alpha| \leq m$, are continuous on Ω.

Furthermore, we use the standard notation for Sobolev spaces and their norms [4], namely, we denote by

$$H^m(\Omega) = \{\varphi \in L^2(\Omega) \mid D^\alpha \varphi \in L^2(\Omega) \ \forall \ |\alpha| \leq m\}$$

the L^2-Sobolev space of order m. Here, the derivatives D^α are taken in the weak sense.
For $m = 0$, we write $L^2(\Omega)$ instead of $H^0(\Omega)$.

We denote by $|| \cdot ||_0$ the usual L^2-norm

$$||\varphi||_0^2 = \int_\Omega |\varphi(\mathbf{x})|^2 dxdy.$$

The norm and seminorm in $H^m(\Omega)$ are given by

$$||\varphi||_m^2 = \sum_{|\alpha| \leq m} ||D^\alpha \varphi||_0^2 \qquad \text{and}$$
$$|\varphi|_m^2 = \sum_{|\alpha| = m} ||D^\alpha \varphi||_0^2 \qquad \text{respectively.}$$

Finally, to denote *periodic* boundary conditions we use the following notation:
for any space $H^m(\Omega)$

$$H^m_{per}(\Omega) = \{f \in H^m(\Omega) \mid f \text{ periodic at the boundaries } \partial\Omega\}.$$

and we denote by L^p_{loc} the space of local L^p-functions.

Chapter 2

Transport equation - 1D case

Let $a \in \mathbb{R}$ be a given constant.
We consider the one dimensional transport equation

$$u_t(x,t) + au_x(x,t) = 0, \qquad \text{for } x \in \Omega_x \subset \mathbb{R}, t \in [0,\infty), a \in \mathbb{R}, \tag{2.1}$$

$$u(x,0) = u_0(x), \qquad \text{for } x \in \Omega_x. \tag{2.2}$$

In the case considered we set Ω_x to be the interval $[0,1]$.
We complement the equation by imposing periodic boundary conditions,

$$u(0,t) = u(1,t), \text{ for all } t \geq 0. \tag{2.3}$$

2.1 The weak formulation

In order to derive the weak formulation, we multiply the equation (2.1) with a test function $\psi \in C^\infty_{per}(\Omega_x)$ and then integrate over $[0,1]$, this yields

$$(u_t, \psi)_{\Omega_x} + (au_x, \psi)_{\Omega_x} = 0$$

in which the following notation has been used:

$$(u, \psi)_{\Omega_x} = \int_0^1 u(x,t)\psi(x)dx.$$

Integrating by parts the second term yields

$$(u_t, \psi)_{\Omega_x} - (au, \psi_x)_{\Omega_x} + a\left[u(1,t)\psi(1,t) - u(0,t)\psi(0,t)\right] = 0.$$

Using (2.3) the terms regarding the boundaries disappear and we obtain the weak formulation which reads: *Find u such that*

$$(u_t, \psi)_{\Omega_x} - (au, \psi_x)_{\Omega_x} = 0, \qquad \textit{for all } \psi \in C^\infty_{per}(\Omega_x). \tag{2.4}$$

Before deriving the DG-method, we define the partition, the finite element space and we introduce the notation we will use.

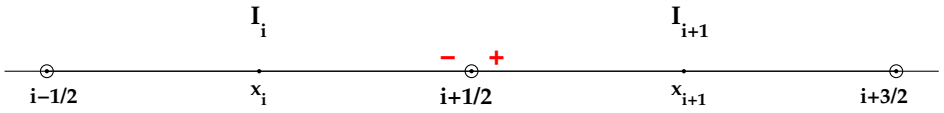
2.2 Notations for the discrete formulation

We fix $\Omega_x = [0,1]$ and consider a **uniform partition** $\mathcal{I}_h = \{I_i\}_{i=1}^N$ of N elements

$$I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \text{ for all } i = 1, \dots, N.$$

We denote by x_i the center of the $i^{.th}$ element I_i and since the partition is assumed to be uniform we have $h = x_{i+1} - x_i$, for all $i = 1, \dots, N-1$. Moreover, we define the following set of points: $\gamma_x = \{x_{i+\frac{1}{2}}\}_{i=0}^N$.

Figure 2.1: Notation at a node.



We define the **finite element space** V_h^0, in which we will search our approximate solution u^h, as

$$V_h^0(\Omega_x) = \{v \in L^2(\Omega_x) |\ v|_{I_i} \in \mathbb{P}^0(I_i) \text{ for all } i = 1, \dots, N\},$$

where $\mathbb{P}^0(I_i)$ is the space of constants functions on I_i.

Let $w \in L^2(\mathcal{I}_h)$, we denote by P^0 the standard L^2-projection of function onto the finite Element Space V_h^0 defined locally, i.e., for each $1 \le i \le N$

$$\int_{I_i} \left(P^0(w) - w\right) q^h \, dx = 0 \qquad \forall\ q^h \in \mathbb{P}^0(I_i). \tag{2.5}$$

For dealing with the discontinuities of the finite element functions at the interelement boundaries, we introduce the following notation

$$\varphi(x^{\pm}_{i+\frac{1}{2}}) = \varphi^{\pm}_{i+\frac{1}{2}} := \lim_{s \to 0^+} \varphi(x_{i+\frac{1}{2}} \pm s) \qquad \text{(see figure 2.1).} \tag{2.6}$$

In order to simplify the notation we will sometimes write u_k instead of $u(x_k)$.

2.3 The Discontinous Galerkin formulation

We consider the weak formulation of problem (2.1)-(2.3) restricted to an element I_i,

$$(u_t, \psi)_{I_i} + (au_x, \psi)_{I_i} = 0, \tag{2.7}$$

We start by considering the second term in (2.7). Integration by parts yields

$$\begin{aligned}(au_x, \psi)_{I_i} &= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} au_x(x,t)\psi(x)dx \\ &= -\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} au(x,t)\psi_x(x)dx + \left[au(x_{i+\frac{1}{2}},t)\psi(x_{i+\frac{1}{2}}) - au(x_{i-\frac{1}{2}},t)\psi(x_{i-\frac{1}{2}})\right]\end{aligned}$$

The next step is to substitute u and ψ by the approximations u^h, $\psi^h \in V_h^0$.
Since these functions are constants on each interval, the derivative ψ_x^h vanishes and as a consequence, in the second line only the second term is left.
Thus, we have

$$\begin{aligned}\left(a\left(u^h\right)_x, \psi^h\right)_{I_i} &= \left[au^h(x_{i+\frac{1}{2}},t)\psi^h(x_{i+\frac{1}{2}}) - au^h(x_{i-\frac{1}{2}},t)\psi^h(x_{i-\frac{1}{2}})\right] \\ &= \left(\widehat{au^h}\right)_{i+\frac{1}{2}} \psi^{h-}_{i+\frac{1}{2}} - \left(\widehat{au^h}\right)_{i-\frac{1}{2}} \psi^{h+}_{i-\frac{1}{2}},\end{aligned}$$

where $\left(\widehat{au^h}\right)_k$ for $k \in \{i-\frac{1}{2}, i+\frac{1}{2}\}$ denotes the *numerical flux*.

The numerical flux $\widehat{au^h}$ is a function defined on interelement boundaries. It depends on the value of a and on the values of the numerical approximation u^h from the elements that share the observed boundary. Its definition is what really characterize the method.

In this case we take the *upwind-flux*,

$$\left(\widehat{au^h}\right)_k = \begin{cases} a\left(u^h\right)^-_k & \text{if } a > 0, \\ a\left(u^h\right)^+_k & \text{if } a < 0. \end{cases} \tag{2.8}$$

Furthermore, since $\psi^h \in \mathbb{P}^0(I_i)$ and we consider the element I_i, we have

$$\psi^{h-}_{i+\frac{1}{2}} = \psi^{h+}_{i-\frac{1}{2}} = \psi^h_i.$$

Also, assuming $a > 0$, (2.8) implies

$$\left(u^h\right)^-_{i+\frac{1}{2}} = u^h_i \qquad \text{and} \qquad \left(u^h\right)^-_{i-\frac{1}{2}} = u^h_{i-1}.$$

On the other hand, if $a < 0$, we have

$$\left(u^h\right)^+_{i+\frac{1}{2}} = u^h_{i+1} \qquad \text{and} \qquad \left(u^h\right)^+_{i-\frac{1}{2}} = u^h_i.$$

So, considering both cases, the second term in (2.7) becomes

$$\left(a\left(u^h\right)_x, \psi^h\right)_{I_i} = \left[\min(a,0)(u^h_{i+1} - u^h_i) + \max(a,0)(u^h_i - u^h_{i-1})\right]\psi^h_i.$$

Finally, we set $u^h(0) = P^h(u_0)$ and then considering also the first term in (2.7), and fixing a positive T, we obtain the following formulation:
find $u^h : [0,T] \to V_h^0$ *such that for all* $i = 1, \ldots, N$

$$\int_{I_i} \left(u^h\right)_t \psi^h dx + \left[\min(a,0)(u_{i+1}^h - u_i^h) + \max(a,0)(u_i^h - u_{i-1}^h)\right] \psi^h = 0, \qquad \text{for all } \psi^h \in V_h^0. \tag{2.9}$$

2.4 Implementation - Basis functions

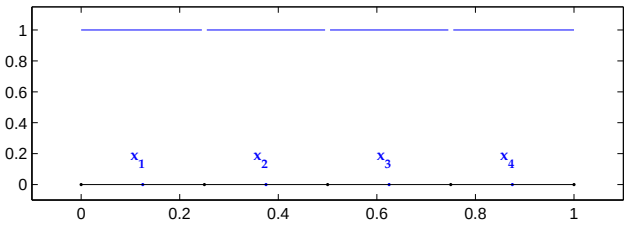
2.4.1 Standard basis

The standard constant basis functions are defined by (the characteristic function of the interval I_i)

$$\chi_i(x) = \begin{cases} 1, & \text{if } x \in I_i, \\ 0, & \text{otherwise,} \end{cases} \tag{2.10}$$

for all $i = 1, \ldots, N$. The case $N = 4$ is depicted in figure 2.2.

Figure 2.2: Standard basis for $N = 4$



Notice (and see also figure 2.2) that the functions χ_i have non-overlapping support.
The approximate solution u^h can be written as

$$u^h(x,t) = \sum_{i=1}^{N} \bar{u}_i(t)\chi_i(x),$$

where $\bar{u}_i$ is the time-dependent coefficient with respect to the basis (2.10), i.e.

$$\bar{u}_i(t) = \frac{1}{h}\left(u^h(x,t), \chi_i\right)_{L^2}.$$

Then (2.9) becomes

$$h\left(\bar{u}_i\right)_t + \min(a,0)(\bar{u}_{i+1} - \bar{u}_i) + \max(a,0)(\bar{u}_i - \bar{u}_{i-1}) = 0. \qquad \forall\ i.$$

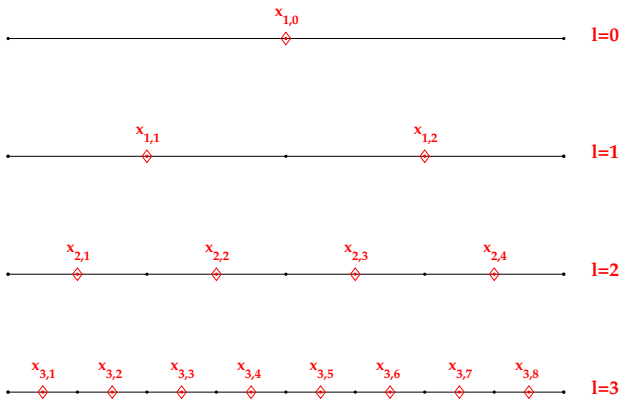
For the initial condition we set

$$u^h(0) = P^0(u_0).$$

2.4.2 Two sets of hierarchical basis

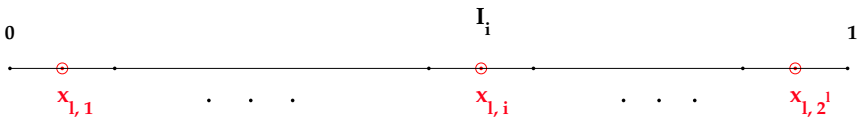
In the following we present another way to approximate a function using sets of functions based on a multilevel construction, a hierarchical basis.
To this end, we still consider the interval $\Omega_x = [0, 1]$ but we need to introduce some further notation:

We define l to be the *level* in hierarchy which indicates how many times we halve our domain.
In fact, when l is fixed, we consider a uniform partition of Ω_x in 2^l elements.
This implies that all elements have the same length h_l, in this case (for $\Omega_x = [0, 1]$), we have $h_l = 2^{-l}$.
The hierarchical meshes for $l \leq 3$ are shown in the following figure.



For a fixed level l, we identify the $i^{.th}$ element of the grid with its center, this means we are considering the following set of points

$$x_{l,i} = 2^{-l-1} + (i-1)\ h_l, \qquad \text{where } i = 1, \ldots, 2^l. \tag{2.11}$$



Next, we present the two sets of hierarchical basis function we consider throughout this work.

The hierarchical *One* basis

For a fixed level $l \geq 0$ we define the functions $\phi_{l,i}$ as

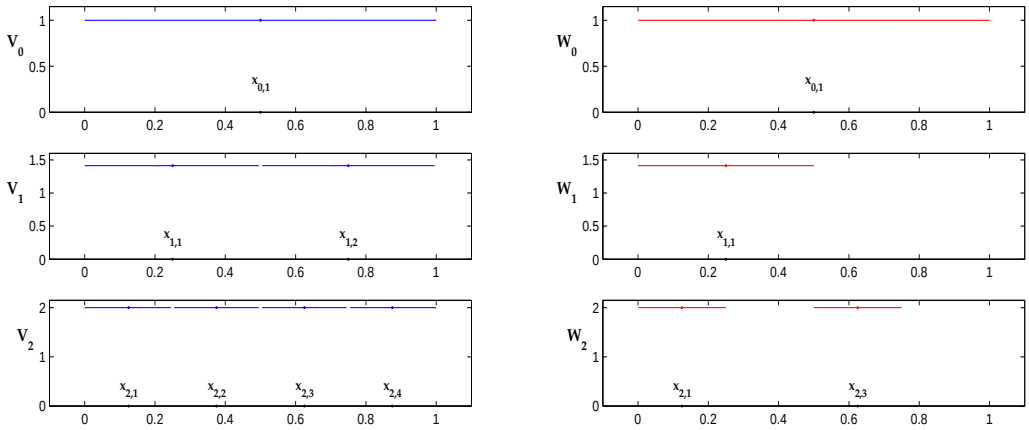
$$\phi_{l,i}(x) = \begin{cases} 2^{\frac{l}{2}} & \text{if } x \in [x_{l,i} - \frac{h_l}{2}, x_{l,i} + \frac{h_l}{2}], \\ 0 & \text{otherwise,} \end{cases} \tag{2.12}$$

for $i = 1, \ldots, 2^l$.
We can define the following spaces:

$$\begin{aligned} V_l &= \text{span}\{\phi_{l,i}(x) : \text{ for } i = 1, \ldots, 2^l\} \\ W_l &= \text{span}\{\phi_{l,i}(x) : \text{ for } i = 1, \ldots, 2^l - 1, i \text{ odd}\} \end{aligned} \tag{2.13}$$

Notice that W_l is a subspace of V_l. In fact, it consists on the space spanned by the functions with the odd indices in the level l [1]. In figure 2.3 the collection of spaces V_l and W_l until level 2 are shown.

Figure 2.3: V_l and W_l for $l \leq 2$



For fixed level l note that:

- ◇ The functions that span the space V_l are a scaled version of the standard basis for $\mathbb{P}^0$ elements given in (2.10):

  $$\phi_{l,i} = 2^{\frac{l}{2}} \chi_i, \qquad \text{for } i = 1, \ldots, 2^l.$$

  Hence $V_l = V_{h_l}^0$.

- ◇ Note that the basis functions (2.12) have disjoint support which implies

  $$\int_{\Omega_x} \phi_{l,k}(x)\phi_{l,j}(x)dx = \begin{cases} \int_{\text{supp}(\phi_{l,k})} \phi_{l,k}^2 dx = 2^{-l} 2^{2\frac{l}{2}} = 2^0 = 1 & \text{if } k = j, \\ \int_0^1 \phi_{l,k}\phi_{l,j} dx = 0 & \text{if } k \neq j. \end{cases}$$

In the following lemma an important relation between the spaces V_l and W_l is stated.
The important consequence of this lemma is the fact that it allows us to consider the direct sum of subspaces W_l for $l \leq n$ to describe the space V_n.

Lemma 2.4.1 *For $n \geq 0$ and for V_n and $\{W_l\}_{l \leq n}$ defined as in (2.13) we have that*

$$V_n = \bigoplus_{l \leq n} W_l. \tag{2.14}$$

Proof 1 There are two things that have to be shown:

(i) The functions that span the spaces are linearly independent.

(ii) The basis of the two spaces have the same dimension (number of elements).

We start by showing (ii):
The number of basis functions in V_n is

$$|V_n| = |\text{span}\{\phi_{n,i} : \text{ for } i = 1, \ldots, 2^n\}| = 2^n.$$

So one has to show that also $|\bigoplus_{l \leq n} W_l| = 2^n$.

From (2.12) and from the definition of the space W_l we have

$$|W_0| = 1 \qquad |W_l| = \frac{2^l}{2} = 2^{l-1} \qquad \text{for } l \geq 1.$$

Thus

$$|\bigoplus_{l \leq n} W_l| = 1 + |\bigoplus_{1 \leq l \leq n} W_l| = 1 + \sum_{i=1}^{n} 2^{i-1} = 1 + \underbrace{\sum_{i=0}^{n-1} 2^i}_{=2^n-1} = 2^n.$$

Hence $\dim(V_n) = \dim\left(\bigoplus_{l \leq n} W_l\right)$.

For (i) we show first that, fixing the level $l = n$, the functions $\phi_{n,i} \in V_n$ for $i = 1, \ldots, 2^n$ are a basis of V_n.
We have said that the $\phi_{n,i}$ have non-overlapping support, hence they are linearly independent. Then, since the number of functions is exactly the same as the number of intervals in which we divide our domain, it follows that $\{\phi_{n,i}\}_{1 \leq i \leq 2^n}$ is a basis of V_n.

Next, we have to show that also the functions $\phi_{l,i}$ that appear on the right part of (2.14) are linearly independent and, consequentely, a basis.
This is different from above because here we consider all levels from 0 to n (not only n as before) but let us also point out that here we only consider the odd indeces i.

We recall that, in order to show that a finite set of functions $\{\varphi_k\}_{1 \leq k \leq N}$ is linearly independent on Ω, the following equivalence has to hold:

$$\sum_{k=1}^{N} a_k \varphi_k(x) = 0, \text{ for all } x \in \Omega \Leftrightarrow a_k = 0, \text{ for all } k = 1, \ldots, N.$$

The implication "$\Leftarrow$" holds trivially, so, next, we show "$\Rightarrow$" for our case by induction.

◇ For $n = 1$ we have $\phi_{0,1}$ and $\phi_{1,1}$ and we need to show that

$$\text{if} \quad \alpha_{0,1}\phi_{0,1}(x) + \alpha_{1,1}\phi_{1,1}(x) = 0, \quad \text{for any } x \in [0,1] \quad \Rightarrow \quad \alpha_{0,1} = \alpha_{1,1} = 0,$$

To this end we choose particular values of x. Recalling (2.11), we take

$$z = x_{1,1} \qquad \text{and} \qquad w = x_{1,1} + h_1.$$

This choice implies

$$\phi_{0,1}(z) = \phi_{0,1}(w) = 1, \qquad \phi_{1,1}(z) = \sqrt{2} \qquad \text{and} \qquad \phi_{1,1}(w) = 0,$$

see also figure 2.3.
So, we have

$$\begin{cases} 0 &= \alpha_{0,1}\phi_{0,1}(z) + \alpha_{1,1}\phi_{1,1}(z) = \alpha_{0,1} + \alpha_{1,1}\sqrt{2} \\ 0 &= \alpha_{0,1}\phi_{0,1}(w) + \alpha_{1,1}\phi_{1,1}(w) = \alpha_{0,1}, \end{cases}$$

and this yields $\alpha_{0,1} = \alpha_{1,1} = 0$.

◇ The induction step $n - 1 \to n$, goes as follows:
assuming

$$V_{n-1} = \bigoplus_{l \le n-1} W_l \tag{2.15}$$

we want to show that

$$\bigoplus_{l \le n-1} W_l \oplus W_n = V_n.$$

We know that $|W_n| = 2^{n-1}$, so we need to show that

$$\sum_l \sum_k \alpha_{l,k}\phi_{l,k}(x) = 0 \ \forall \ x \Leftrightarrow \alpha_{l,k} = 0 \ \forall \ l, k. \tag{2.16}$$

We recall that $l = 0, \ldots, n$ and $k = 1, \ldots, 2^l - 1$ odd.

To this end, similarly to above, for every function of the n.th level we define the following points:

$$z_k = x_{n,k} \qquad \text{and} \qquad w_k = x_{n,k} + h_n.$$

This definition implies

$$\phi_{n,k}(z_j) = \delta_{kj} 2^{\frac{n}{2}} \qquad \text{and} \qquad \phi_{n,k}(w_j) = 0 \quad \forall k,$$

where

$$\delta_{kj} = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{otherwise.} \end{cases}$$

From this we have that the left hand side of the equivalence (2.16) is the following system of equations:
for $l = 0, \ldots, n-1$ and $k = 1, \ldots, 2^l - 1$ odd,

$$\sum_l \sum_k \alpha_{l,k}\phi_{l,k}(z_i) + \alpha_{n,i}2^{\frac{n}{2}} = 0 \tag{2.17}$$

$$\sum_l \sum_k \alpha_{l,k}\phi_{l,k}(w_i) = 0. \tag{2.18}$$

Observe that the special choice of z_k an w_k implies

$$\phi_{l,k}(z_i) = \phi_{l,k}(w_i) \qquad \forall\ l \leq n \quad \text{and} \quad \forall\ k, i.$$

This implies that in (2.17) we remain with just

$$\alpha_{n,i}2^{\frac{n}{2}} = 0, \qquad \text{thus} \qquad \alpha_{n,i} = 0, \quad \forall\ i = 1, \ldots, 2^n - 1 \text{ odd}.$$

Note that we have $|\underset{l\leq n-1}{\oplus} W_l| = 2^{n-1}$ and also that we still have 2^{n-1} equations involving the points w_i that can be used.

Let us recall our asumption (2.15) which says that $V_{n-1} = \underset{l\leq n-1}{\oplus} W_l$.
This implies that the set $\{\phi_{l,k}\}_{l=0}^{n-1}$ is a basis of $\underset{l\leq n-1}{\oplus} W_l$, and so the $\phi_{l,i}$'s are linearly independent.

Now, we consider the 2^{n-1} equations concerning the point w_i.
These equations have the following form

$$\sum_l \sum_k \alpha_{l,k}\phi_{l,k}(w_i) = 0,$$

and because of the linear independence of the $\phi_{l,i}$'s we have that all the coefficients $\alpha_{l,k}$, for $l = 0, \ldots, n-1$ and $k = 1, \ldots, 2^{n-1} - 1$ odd, are equal to zero.

■

The hierarchical *Haar* basis

Another set of basis function that we consider is the one based on the **hierarchical *Haar* basis.**
The functions $\theta_{l,i}$ for $i = 1, \ldots, 2^l$ are defined in the following way:

$$\text{for } l = 0, \qquad \theta_{0,1}(x) = \begin{cases} 1 & \text{if } x \in \Omega_x, \\ 0 & \text{otherwise,} \end{cases}$$

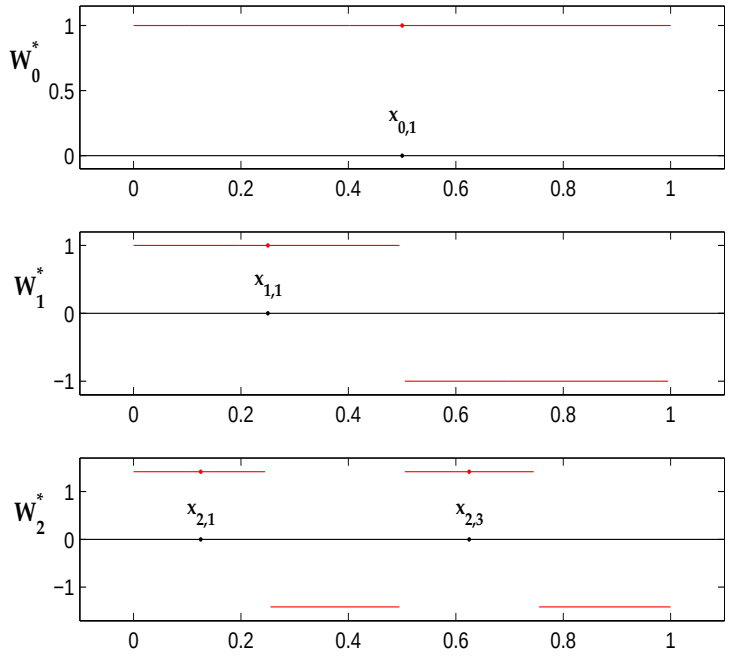
$$\text{for } l \geq 1, \qquad \theta_{l,i}(x) = \begin{cases} 2^{\frac{l-1}{2}} & \text{if } x \in [x_{l,i} - \frac{h_l}{2}, x_{l,i} + \frac{h_l}{2}], \\ -2^{\frac{l-1}{2}} & \text{if } x \in [x_{l,i+1} - \frac{h_l}{2}, x_{l,i+1} + \frac{h_l}{2}], \\ 0 & \text{otherwise.} \end{cases} \tag{2.19}$$

We define W_l^* in the same way as before (2.13), i.e.,

$$W_l^* = \text{span}\{\theta_{l,i}(x) : \text{ for } i = 1, \ldots, 2^l - 1, i \text{ odd}\}$$

The subspaces W_l^* up to level 2 are shown in figure 2.4.

Figure 2.4: W_l^* for $l \leq 2$ using hierarchical Haar basis functions



Notice that hierarchical Haar basis is orthogonal in L^2, i.e., for any choice of (l, m, k, i) we have

$$\int_{\Omega_x} \theta_{l,k}(x)\theta_{m,i}(x)dx = \begin{cases} 1 & \text{if } l = m,\ k = i \\ \\ 0 & \text{otherwise.} \end{cases}$$

2.5 Time integration

We now describe how we perform the integration in time.

Before presenting the method we introduce the notation needed:

For a $T_{end} > 0$ we consider the time interval $[0, T_{end}]$, so we take a uniform partition

$$0 = t^0 < t^1 < \ldots < t^M = T_{end},$$

and for a fixed time-step $dt = \left[\frac{T_{end}-t^0}{M}\right]$, the discretization in time is given recursively by

$$t^1 = t^0 + dt, \qquad \text{and, for } m \in \mathbb{N}, \qquad t^{m+1} = t^m + dt.$$

We set $(u^h)^0 := P^h(u_0)$ and for any future t^m the numerical approximation u^h will be denoted by $u^h(t^m) = (u^h)^m$.

Now, using that V_h^0 is a space of piecewise constants, we rearrange the DG-formulation (2.9) as follows

$$\int_{I_i} \left(u^h\right)_t \psi^h dx = -\left[\min(a,0)(u_{i+1}^h - u_i^h) + \max(a,0)(u_i^h - u_{i-1}^h)\right] \psi^h.$$

Then we use the explicit Euler method in which $(u^h)_t$ at time m is approximated through a *forward* difference which reads

$$(u^h)_t = \frac{(u^h)^{m+1} - (u^h)^m}{dt}.$$

This is a one-step explicit method, because the solution at time $m+1$ is computed directly in terms of the value u^m at the previous time-step:
for $m = 0$, given $(u^h)^0 := P^h(u_0)$

$$\begin{aligned}\int_{I_i} \left(u^h\right)^1 \psi^h dx = &\int_{I_i} \left(u^h\right)^0 \psi^h dx \\ &- dt\left[\min(a,0)(\ \left(u^h\right)_{i+1}^0 - \left(u^h\right)_i^0) + \max(a,0)(\ \left(u^h\right)_i^0 - \left(u^h\right)_{i-1}^0)\right] \psi^h.\end{aligned}$$

and, for any $m > 0$,

$$\begin{aligned}\int_{I_i} \left(u^h\right)^{m+1} \psi^h dx = &\int_{I_i} \left(u^h\right)^m \psi^h dx \\ &- dt\left[\min(a,0)(\ \left(u^h\right)_{i+1}^m - \left(u^h\right)_i^m) + \max(a,0)(\ \left(u^h\right)_i^m - \left(u^h\right)_{i-1}^m)\right] \psi^h.\end{aligned}$$

In order to give a stability result we consider the time discretization in one element I_i when the standard basis function are used, this reads

$$u_i^{m+1} = u_i^m - \frac{dt}{h}\left[\min(a,0)\left(u_{i+1}^m - u_i^m\right) + \max(a,0)\left(u_i^m - u_{i-1}^m\right)\right].$$

This is the usual formulation that we derive using *finite difference* (with forward difference) or *finite volume* (with upwind fluxes) methods, on a uniform mesh.

For this case the method is stable provided

$$dt < \frac{h}{|a|}. \tag{2.20}$$

which is the, so-called, CFL condition (Courant, Friedrichs, Lewy). A prove can be found in [10].

Lemma 2.5.1 *Using the standard basis functions* (2.10)*, the scheme* (2.9) *for the one-dimensional transport equation is total variation diminishing (TVD), which means that the approximate solution u^h of the discretized problem satisfies*

$$TV((u^h)^{m+1}) \leq TV((u^h)^m)$$

where the total variation seminorm is defined by

$$TV(u) = \sum_i |u_{i+1} - u_i|.$$

To this end let us state the following lemma due to Harten.

Lemma 2.5.2 ***(Harten's lemma)***
If a scheme can be written in the form

$$u_i^{m+1} = u_i^m + C_{i+\frac{1}{2}} \Delta u_{i+1}^m - D_{i-\frac{1}{2}} \Delta u_i^m,$$

with periodic or compactly supported boundary conditions, where $\Delta u_i^m = u_i^m - u_{i-1}^m$ and $C_{i+\frac{1}{2}}$ and $D_{i-\frac{1}{2}}$ may be nonlinear functions of the grid values u_j^m for $j = i-p, \ldots, i+q$ with some $p, q \leq 0$ satisfying

$$C_{i+\frac{1}{2}} \geq 0, \qquad D_{i+\frac{1}{2}} \geq 0, \qquad C_{i+\frac{1}{2}} + D_{i+\frac{1}{2}} \leq 1, \qquad \forall\ i \tag{2.21}$$

then the scheme is total variation diminishing.

Proof of Harten's lemma: see [11].

Proof 2 In order to show that our numerical scheme is TVD, we start by recalling the scheme

$$h_x \frac{\bar{u}_i^{m+1} - \bar{u}_i^m}{dt} + \min(a,0)(\bar{u}_{i+1}^m - \bar{u}_i^m) + \max(a,0)(\bar{u}_i^m - \bar{u}_{i-1}^m) = 0,$$

where $\bar{u}_i^m$ is the value taken in the interval I_i (or associated at the grid point x_i).

Rewriting this expression by using the differential operator introduced above and moving some terms, we obtain

$$\bar{u}_i^{m+1} = \bar{u}_i^m - \frac{dt \min(a,0)}{h_x} \Delta \bar{u}_{i+1}^m - \frac{dt \max(a,0)}{h_x} \Delta \bar{u}_i^m.$$

Since a is a constant different from zero, we have that either $a > 0$ or $a < 0$, so

- $\diamond$ $\mathbf{a > 0}$:
  this implies $C_{i+\frac{1}{2}} = 0$ and $D_{i+\frac{1}{2}} = \frac{dt\ a}{h_x} > 0$. In order to satisfy also the third condition in (2.21) we need to have

$$\frac{dt\ a}{h_x} \leq 1 \qquad \Rightarrow \qquad dt \leq \frac{h_x}{a}.$$

- ◇ $\mathbf{a < 0}$:
  this implies $C_{i+\frac{1}{2}} = -\frac{dt\ a}{h_x} > 0$ and $D_{i+\frac{1}{2}} = 0$. The third condition in (2.21) is satisfied as long as dt is chosen in such a way that the following holds:

$$-\frac{dt\ a}{h_x} \leq 1 \qquad \Rightarrow \qquad dt \leq -\frac{h_x}{a}.$$

Considering both cases together, we observe that, if we choose a time stepping dt such that

$$dt \leq \frac{h_x}{|a|},$$

all conditions (2.21) of Harten's lemma are satisfied. This requirement is exactly the CFL condition we have seen before (2.20), hence our numerical scheme is TVD.

2.6 Numerical experiments

One-dimensional transport equation with constant coefficients

We check the convergence of the numerical method using the standard basis functions and the two different sets of hierarchical basis functions to solve the following one-dimensional transport equation with periodic boundary condition

$$\begin{aligned} u_t(x,t) + u_x(x,t) &= 0, && \text{for } (x,t) \in [0,1] \times [0, T_{end}], \\ u(x,0) &= \sin(2\pi x), && \text{for all } x \in [0,1], \\ u(0,t) &= u(1,t), && \text{for all } t \in [0, T_{end}]. \end{aligned}$$

The exact solution of this problem is given by

$$u_{ex}(x,t) = \sin(2\pi(x-t)), \qquad \text{for } (x,t) \in [0,1] \times [0, T_{end}].$$

2.6.1 Standard basis

We performed the experiment using first the standard basis dividing the domain $[0,1]$ in N =16, 32, 64, 128 and 256 uniform subintervals and fixing $T_{end} = 1$.
In this experiment we looked at the absolute errors computed using the L^2 norm

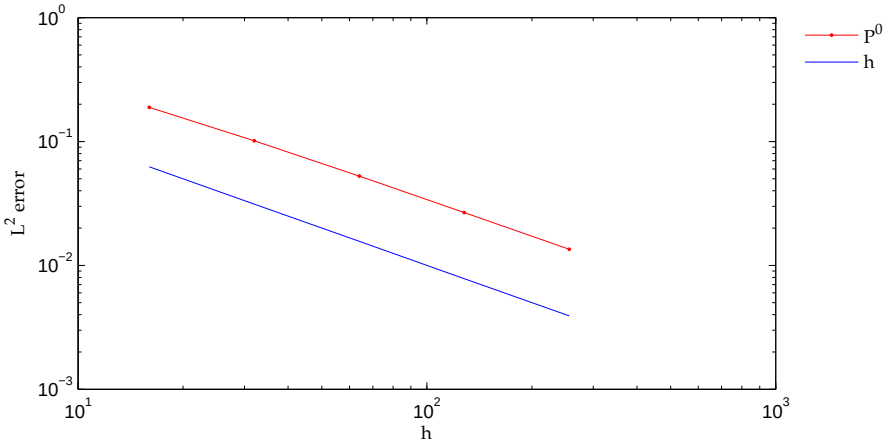
$$||u_{ex}(\cdot,1) - u^h(\cdot,1)||_0.$$

In the following table the errors in the L^2 norm and the rates of convergence are given.

N	16	32	64	128	256
L^2-*error*	0.1887	0.1012	0.0525	0.0267	0.0135
Rates	0.8989	0.9468	0.9755	0.9839	

The convergence is shown in figure 2.5.

Figure 2.5: Convergence using the standard basis



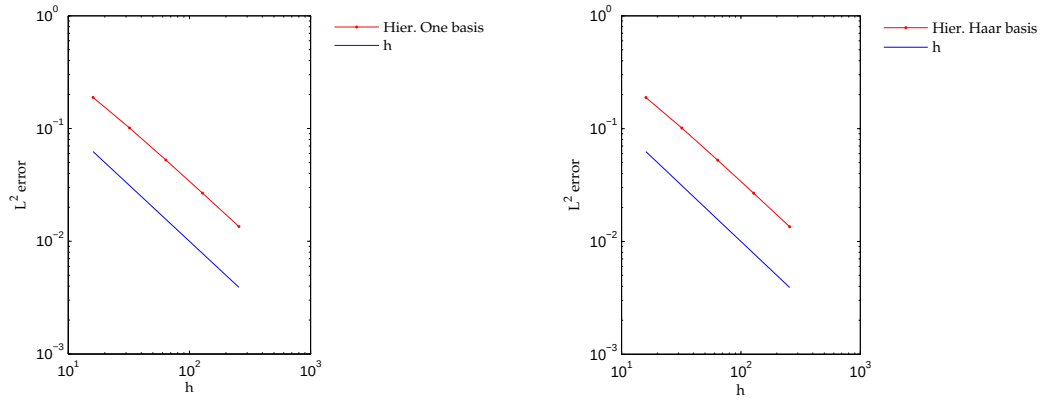
2.6.2 Hierarchical basis

With both sets of hierarchical basis functions we performed the same experiment fixing $n = 4$, 5, 6, 7 and 8. These values corresponds to $N = 16$, 32, 64, 128, 256 respectively.
In the following table, again, the absolute errors (in L^2 norm) and the rates of convergence are given.

n		4	5	6	7	8
Hier. *One* b.	L^2-*error*	0.1887	0.1012	0.0525	0.0267	0.0135
	Rates	0.8989	0.9468	0.9755	0.9839	
Hier. *Haar* b.	L^2-*error*	0.1887	0.1012	0.0525	0.0267	0.0135
	Rates	0.8989	0.9468	0.9755	0.9839	

In figure 2.6 we have represented the diagrams for the DG method using hierarchical basis. On the left it is represented the one which refers to the hierarchical *One* basis and on the right the one obtained by using the hierarchical *Haar* basis. Since we only perform a change of basis the results are not surprisingly exactly the same as the ones obtained by using the standard basis.

Figure 2.6: Convergence diagrams using hierarchical basis functions



Chapter 3

Transport equation - 2D case

3.1 Continuous problem

Let $\Omega = \Omega_x \times \Omega_y \subset \mathbb{R}^2$, $a \in L^\infty(\Omega_y)$ and $b \in L^\infty(\Omega_x)$.
We consider the **transport equation**:

$$\begin{aligned} u_t + a(y)u_x + b(x)u_y &= 0, && \text{for } x \in \Omega_x, y \in \Omega_y, t \in [0,\infty), \\ u(x,y,0) &= u_0(x,y) && \text{for } x \in \Omega_x, y \in \Omega_y. \end{aligned} \tag{3.1}$$

The problem we consider is obtained by completing the equation (3.1) with periodic boundary conditions both in x and y:

$$\partial\Omega_x = [x_{\text{left}}, x_{\text{right}}] \quad \text{and} \quad \partial\Omega_y = [y_{\text{down}}, y_{\text{up}}],$$

we have

$$\begin{aligned} u(x, y_{\text{down}}, t) &= u(x, y_{\text{up}}, t) && \forall\ x \in \Omega_x, \\ u(x_{\text{left}}, y, t) &= u(x_{\text{right}}, y, t) && \forall\ y \in \Omega_y. \end{aligned} \tag{3.2}$$

3.1.1 The weak formulation

In order to get the weak formulation, we proceed as in the one-dimensional case. We multiply the equation (3.1) by a test function $\psi \in C^\infty_{per}(\Omega)$ and then integrate over Ω.
We rewrite the equation as

$$(u_t, \psi)_\Omega + (au_x, \psi)_\Omega + (bu_y, \psi)_\Omega = 0$$

where

$$(u, \psi)_\Omega = \int_\Omega u(x,y,t)\psi(x,y)\ dxdy.$$

Integrating by parts the second and the third term we obtain

$$(u_t, \psi)_\Omega - (au, \psi_x)_\Omega + (au, \psi)_{\partial\Omega_y} - (bu, \psi_y)_\Omega + (bu, \psi)_{\partial\Omega_x} = 0,$$

where

$$(au,\psi)_{\partial\Omega_y} = \int_{\Omega_y} [au(x,y,t)\psi(x,y)]_{x\in\partial\Omega_x} dy,$$
$$(bu,\psi)_{\partial\Omega_x} = \int_{\Omega_x} [bu(x,y,t)\psi(x,y)]_{y\in\partial\Omega_y} dx.$$

Again, because of (3.2) the integrals defined on the boundaries simplify and the weak formulation reads: *Find u such that*

$$(u_t,\psi)_\Omega - (au,\psi_x)_\Omega - (bu,\psi_y)_\Omega = 0, \qquad \textit{for all } \psi \in C^\infty_{per}(\Omega). \tag{3.3}$$

3.1.2 Properties of the system

We now state some properties of the continuous solution of (3.1)-(3.2).
In deriving numerical methods, we will try to ensure that the resulting schemes will produce approximate solutions, which will be able to mimic some of these properties.

Mass conservation

The function u, continuous solution of the problem (3.1)-(3.2), conserves the total mass, i.e.

$$\frac{d}{dt}\int_\Omega u(x,y,t)\ dxdy = 0, \qquad \forall t > 0.$$

Proof 3 In (3.3) we take $\psi = 1 \in C^\infty_{per}(\Omega)$ and so we have

$$\int_\Omega u_t 1\ dxdy - \underbrace{\int_\Omega au\psi_x\ dxdy}_{=0} - \underbrace{\int_\Omega bu\psi_y\ dxdy}_{=0} = 0$$

Thus

$$\frac{d}{dt}\int_\Omega u(x,y,t)\ dxdy = 0, \qquad \forall t > 0.$$

■

L^2-conservation

For the continuous solution of problem (3.1)-(3.2) we also have L^2-conservation, i.e.

$$\frac{d}{dt}||u(\cdot,\cdot,t)||_0 = 0 \qquad \forall t > 0.$$

Proof 4 By taking $\psi = u$ in (3.3) we have

$$\begin{aligned}
0 &= \int_\Omega u_t u\ dxdy - \int_\Omega auu_x\ dxdy - \int_\Omega buu_y\ dxdy \\
&= \int_\Omega \left(\frac{u^2}{2}\right)_t dxdy - \int_{\Omega_y} a(y)\left(\int_{\Omega_x}\left(\frac{u^2}{2}\right)_x dx\right)dy - \int_{\Omega_x} b(x)\left(\int_{\Omega_y}\left(\frac{u^2}{2}\right)_y dy\right)dx \\
&= \frac{1}{2}\frac{d}{dt}\int_\Omega u^2\ dxdy - \frac{1}{2}\int_{\Omega_y} a\underbrace{u^2|_{\partial\Omega_x}}_{=0}dy - \frac{1}{2}\int_{\Omega_x} b\underbrace{u^2|_{\partial\Omega_y}}_{=0}dx
\end{aligned}$$

where we have used the periodicity of u in x and y.
So we end up with

$$\frac{d}{dt}\int_\Omega u^2(x,y,t)\ dxdy = 0,$$

which is

$$\frac{d}{dt}||u(\cdot,\cdot,t)||_0^2 = 0.$$

■

3.2 Notations for the discrete formulation

From now on, we consider our domain to be $\Omega = [0,1]^2$.
Before deriving the DG-method we define the partition, the finite element space and we present some notations.
Similarly to the one-dimensional case, we consider a **uniform partition**

$$\mathcal{T}_h = \mathcal{I}_h \times \mathcal{J}_h = \{K_{ij} = I_i \times J_j\}_{1\le i,j\le N},$$

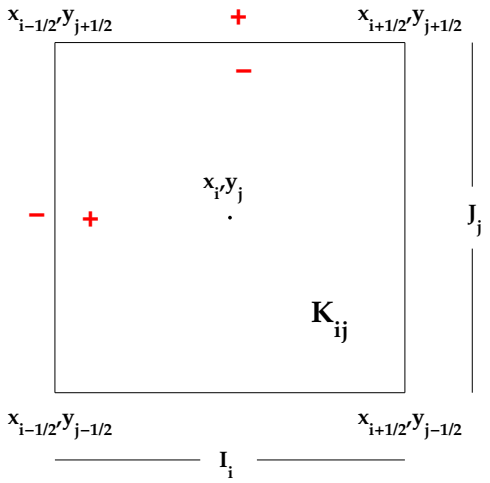
of N^2 elements (K_{ij}). Here we have

$$I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \qquad J_i = [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}],$$
$$h_x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \qquad h_y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}} \quad \text{and} \quad h = \min\{h_x, h_y\} \quad \text{for all } 1 \le i,j \le N.$$

The element K_{ij} (figure 3.1) is centered in (x_i, y_j) and its vertices are given by

$$(x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}}),\ (x_{i+\frac{1}{2}}, y_{j-\frac{1}{2}}),\ (x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}) \text{ and } (x_{i-\frac{1}{2}}, y_{j+\frac{1}{2}}).$$

Figure 3.1: The element K_{ij} and the notation at the boundaries



We introduce some trace operators that allow for dealing with the discontinuities (of the FE functions) at interelement boundaries:
First, we define the space

$$H^1(\mathcal{T}_h) = \{v \in L^2(\Omega) : v \in H^1(K_{ij}) \text{ for all } K_{ij} \in \mathcal{T}_h\},$$

then for any function $\varphi \in H^1(\mathcal{T}_h)$ we define its **jump** $[\![\cdot]\!]$ and its **average** $\{\cdot\}$ at $(x_{i+\frac{1}{2}}, y)$ for all $y \in J_j$ as

$$[\![\varphi]\!]_{i+\frac{1}{2},y} := \varphi^+_{i+\frac{1}{2},y} - \varphi^-_{i+\frac{1}{2},y}, \qquad \{\varphi\}_{i+\frac{1}{2},y} := \frac{1}{2}\left[\varphi^+_{i+\frac{1}{2},y} + \varphi^-_{i+\frac{1}{2},y}\right], \tag{3.4}$$

where $\varphi^+_{i+\frac{1}{2},y}$ and $\varphi^-_{i+\frac{1}{2},y}$ denote the values of φ evaluated at the point $(x_{i+\frac{1}{2}}, y)$ from the right element $K_{i+1,j}$ and from the left element K_{ij} respectively:

$$\varphi^\pm_{i+\frac{1}{2},y} = \varphi(x^\pm_{i+\frac{1}{2}}, y) = \lim_{s\to 0^+} \varphi(x_{i+\frac{1}{2}} \pm s, y). \tag{3.5}$$

In the same way, but in the other coordinate, we have that $\varphi^+_{x,j+\frac{1}{2}}$ and $\varphi^-_{x,j+\frac{1}{2}}$ denote the values of φ evaluated at the point $(x, y_{j+\frac{1}{2}})$ from the upper element $K_{i,j+1}$ and from the buttom element K_{ij} respectively:

$$\varphi^\pm_{x,j+\frac{1}{2}} = \varphi(x, y^\pm_{j+\frac{1}{2}}) = \lim_{s\to 0^+} \varphi(x, y_{i+\frac{1}{2}} \pm s), \tag{3.6}$$

see again figure 3.1 for the spatial representation.

The **finite element space** we consider is

$$Z^0_n = \{z \in L^2(\Omega) \ : \ z \in \mathbb{Q}^0(K_{ij}) = \mathbb{P}^0(I_i) \otimes \mathbb{P}^0(J_j), \ 1 \le i, j \le N = 2^n\}, \tag{3.7}$$

where we construct $\mathbb{Q}^0(K_{ij})$ by doing tensor product of the spaces of constant functions in both I_i and J_j.
Notice that, for reasons that will become clear soon (the use of hierarchical basis) we take $N = 2^n$ where $n \in \mathbb{N}$.

In the numerical experiments we will measure errors in the **discrete $\mathbf{L^2}$-norm** which is defined by

$$||u||^2_{0,\mathcal{T}_h} = \sum_{i,j} \int_{K_{ij}} |u|^2 \, dxdy.$$

In the two-dimensional case we define the standard L^2-projection $\mathcal{P}_h : L^2(\mathcal{T}_h) \to Z^0_n$ by

$$\mathcal{P}_h(w) = \left(P^0_x \otimes P^0_y\right)(w) \qquad \forall \ w \in L^2(\mathcal{T}_h),$$

where P^0_x and P^0_y are the one-dimensional projections (2.5) in x and y respectively,
i.e., for all $1 \le i, j \le 2^n$,

$$\int_{K_{ij}} (\mathcal{P}_h(w) - w)\, \psi^h \, dxdy = 0 \qquad \forall \ \psi^h \in \mathbb{P}^0(I_i) \otimes \mathbb{P}^0(J_j). \tag{3.8}$$

3.3 The Discontinuous Galerkin formulation

We now derive the DG-method for approximating (3.1)-(3.2). As in one dimension, we observe the weak formulation on an arbitrary element $K_{ij} \in \mathcal{T}_h$:

$$(u_t, \psi)_{K_{ij}} + (au_x, \psi)_{K_{ij}} + (bu_y, \psi)_{K_{ij}} = 0 \qquad \text{for all } \psi \in Z_n^0. \tag{3.9}$$

We replace u and ψ by their approximations u^h and ψ^h and we integrate by parts.
For the second term we have

$$\begin{aligned}(a\left(u^h\right)_x, \psi^h)_{K_{ij}} &= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} a\left(u^h\right)_x \psi^h \, dxdy \\ &= -\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u^h (a\psi^h)_x \, dxdy \\ &\quad + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left[au^h\psi^h\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} dy \qquad \forall\, \psi^h \in Z_n^0.\end{aligned}$$

Observe that, since $\psi^h \in \mathbb{Q}^0(K_{ij})$ and a depends only on the variable y, the derivatives in x-direction are equal to zero and, as a consequence, the first integral of the second line disappears. Precisely

$$\begin{aligned}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u^h (a\psi^h)_x dxdy &= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u^h [a\psi^h]_x dxdy \\ &= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u^h \underbrace{\left(a_x\psi^h + a\left(\psi^h\right)_x\right)}_{=0} dxdy = 0.\end{aligned}$$

Rewriting the term left, by using the numerical fluxes we obtain

$$(a\left(u^h\right)_x, \psi^h)_{K_{ij}} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left[\widehat{au^h_{i+\frac{1}{2},y}} \left(\psi^h\right)^-_{i+\frac{1}{2},y} - \widehat{au^h_{i-\frac{1}{2},y}} \left(\psi^h\right)^+_{i-\frac{1}{2},y}\right] dy.$$

We repeat the same procedure for the third term in (3.9)

$$\begin{aligned}(b\left(u^h\right)_y, \psi^h)_{K_{ij}} &= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} b\left(u^h\right)_y \psi \, dxdy \\ &= -\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,y,t) \left(b\psi^h\right)_y dxdy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} b\left[u^h\psi^h\right]_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} dx.\end{aligned}$$

Arguing as before, we have $\left(b\psi^h\right)_y = 0$, so

$$(b\left(u^h\right)_y, \psi^h)_{K_{ij}} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left[\widehat{bu^h_{x,j+\frac{1}{2}}} \left(\psi^h\right)^-_{x,j+\frac{1}{2}} - \widehat{au^h_{x,j-\frac{1}{2}}} \left(\psi^h\right)^+_{x,j-\frac{1}{2}}\right] dx.$$

At the end, we have the following element-based formulation:
find $u^h : [0, T_{end}] \to Z_n^0$ *such that, for all* i, j,

$$\left(\left(u^h\right)_t, \psi^h\right)_{K_{ij}} + \left\langle \widehat{au^h}, \psi^h \right\rangle_{J_j} + \left\langle \widehat{bu^h}, \psi^h \right\rangle_{I_i} = 0, \qquad \forall \psi^h \in Z_n^0, \tag{3.10}$$

where the following notation has been used

$$\begin{aligned}
\left\langle \widehat{au^h}, \psi \right\rangle_{J_j} &= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left[\widehat{au^h}(x_{i+\frac{1}{2}}, y)\psi(x_{i+\frac{1}{2}}, y) - \widehat{au^h}(x_{i-\frac{1}{2}}, y)\psi(x_{i-\frac{1}{2}}, y) \right] dy \\
&= \int_{J_j} \left[\left(\widehat{au^h} \left(\psi^h\right)^- \right)_{i+\frac{1}{2},y} - \left(\widehat{au^h} \left(\psi^h\right)^+ \right)_{i-\frac{1}{2},y} \right] \, dy, \\
\left\langle \widehat{bu^h}, \psi \right\rangle_{I_i} &= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left[\widehat{bu^h}(x, y_{j+\frac{1}{2}})\psi(x, y_{j+\frac{1}{2}}) - \widehat{bu^h}(x, y_{j-\frac{1}{2}})\psi(x, y_{j-\frac{1}{2}}) \right] dx \\
&= \int_{I_i} \left[\left(\widehat{bu^h} \left(\psi^h\right)^- \right)_{x,j+\frac{1}{2}} - \left(\widehat{bu^h} \left(\psi^h\right)^+ \right)_{x,j-\frac{1}{2}} \right] \, dx
\end{aligned}$$

and $u^h(0) := \mathcal{P}_h(u_0)$ is the approximation to the initial data. By summing up over all elements of the partition in (3.10) we will get the DG-method.
As in one dimension, $\widehat{au^h}$ and $\widehat{bu^h}$ denote the **numerical fluxes**, and, also here, we consider *upwind* fluxes:

$$\left(\widehat{au^h}\right)_{i+\frac{1}{2}} = \begin{cases} au^h(x^-_{i+\frac{1}{2}}, y) & \text{if } a(y) \geq 0 \\ au^h(x^+_{i+\frac{1}{2}}, y) & \text{if } a(y) < 0 \end{cases} \quad \left(\widehat{bu^h}\right)_{j+\frac{1}{2}} = \begin{cases} bu^h(x, y^-_{j+\frac{1}{2}}) & \text{if } b(x) \geq 0 \\ bu^h(x, y^+_{j+\frac{1}{2}}) & \text{if } b(x) < 0. \end{cases} \tag{3.11}$$

Here we used the notation defined in (3.5)-(3.6).

3.3.1 L^2-stability

The definition of the numerical fluxes is the key point in the construction of a DG-scheme, in particular they have to be defined in such a way to ensure the stability of the method.
Next, we show that our numerical method is L^2-stable when *upwind* fluxes (3.11) are used.

Lemma 3.3.1 *Consider the transport equation* (3.1) *and its DG-formulation* (3.10). *If upwind fluxes* (3.11) *are used, then the numerical solution* u^h *is* L^2*-stable, i.e.*

$$||u^h(\cdot,\cdot,t)||_{0,\mathcal{T}_h} \leq ||u^h(\cdot,\cdot,0)||_{0,\mathcal{T}_h} \qquad \forall\ t > 0. \tag{3.12}$$

Proof 5 We have to show (3.12).
We start by replacing ψ^h by u^h in (3.10).
For the first integral we have

$$\int_{K_{ij}} (u^h)_t u^h \, dxdy \; = \int_{K_{ij}} \left(\frac{(u^h)^2}{2} \right)_t \, dxdy = \frac{1}{2}\frac{d}{dt} \int_{K_{ij}} (u^h)^2 \, dxdy$$

and when we sum up over all elements we obtain

$$\frac{1}{2}\frac{d}{dt} \sum_{i,j} \int_{K_{ij}} (u^h)^2 \, dxdy = \frac{1}{2}\frac{d}{dt} ||u^h||^2_{0,\mathcal{T}_h}.$$

Now, observe that if the sums over all elements of the second and third term in (3.10) are non-negative we would have

$$\frac{1}{2}\frac{d}{dt}||u^h||^2_{0,\mathcal{T}_h} + \Xi = 0.$$

in which $\Xi \geq 0$. In this case, rearrenging the terms and integrating in time we would get

$$||u^h(\cdot,\cdot,t)||^2_{0,\mathcal{T}_h} - ||u^h(\cdot,\cdot,0)||^2_{0,\mathcal{T}_h} = -\Xi \leq 0$$

and the lemma would be proved.

So, next we show the non-negativity for the second term in (3.10).
We consider the sum over all i, j of the following term

$$\int_{J_j} \left(\widehat{au^h}_{i+\frac{1}{2},j}(u^h)^-_{i+\frac{1}{2},j} - \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^+_{i-\frac{1}{2},j} \right) dy.$$

Notice that, since $u^h \in \mathbb{Q}^0(K_{ij})$ and we deal with a uniform partition, this expression can be rewritten as follows

$$h_y \left(\widehat{au^h}_{i+\frac{1}{2},j}(u^h)^-_{i+\frac{1}{2},j} - \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^+_{i-\frac{1}{2},j} \right).$$

In order to try to keep the notation as clear as possible, let us divide the term by $h_y > 0$.
In fact, one notices that h_y is a positive quantity that can be easily simplified by multiplying all terms in (3.10) with $1/h_y$.

We have then

$$\begin{aligned}
\widehat{au^h}_{i+\frac{1}{2},j}(u^h)^-_{i+\frac{1}{2},j} - \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^+_{i-\frac{1}{2},j} &= \widehat{au^h}_{i+\frac{1}{2},j}(u^h)^-_{i+\frac{1}{2},j} \underbrace{\pm \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^-_{i-\frac{1}{2},j}}_{=0} - \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^+_{i-\frac{1}{2},j} \\
&= \underbrace{\widehat{au^h}_{i+\frac{1}{2},j}(u^h)^-_{i+\frac{1}{2},j}}_{:=F_{i+\frac{1}{2},j}} - \underbrace{\widehat{au^h}_{i-\frac{1}{2},j}(u^h)^-_{i-\frac{1}{2},j}}_{:=F_{i-\frac{1}{2},j}} \\
&+ \underbrace{\widehat{au^h}_{i-\frac{1}{2},j}(u^h)^-_{i-\frac{1}{2},j} - \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^+_{i-\frac{1}{2},j}}_{:=\Theta_{i-\frac{1}{2},j}} \\
&= F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j} + \Theta_{i-\frac{1}{2},j}.
\end{aligned}$$

Now, observe that, by summing up over all elements (sum over all i and j), the first two terms in the last line form a telescopic sum and, since we deal with periodic boundary conditions, the boundary terms cancel and so the whole sum.

So, it remains to be shown the non-negativity of the sum of the Θ-terms.
We decompose the Θ-term as follows

$$\Theta_{i-\frac{1}{2},j} = \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^-_{i-\frac{1}{2},j} - \widehat{au^h}_{i-\frac{1}{2},j}(u^h)^+_{i-\frac{1}{2},j} = -\widehat{au^h}_{i-\frac{1}{2},j}[\![u^h]\!]_{i-\frac{1}{2},j}.$$

Notice that from the definition of the fluxes (3.11) and the definition of the trace operators (3.4), we have

$$\widehat{au^h}_{i-\frac{1}{2},j} = a\{u^h\}_{i-\frac{1}{2},j} - \frac{|a|}{2}\left[\!\left[u^h\right]\!\right]_{i-\frac{1}{2},j},$$
$$\left[\!\left[(u^h)^2\right]\!\right]_{i-\frac{1}{2},j} = 2\{u^h\}_{i-\frac{1}{2},j}\left[\!\left[u^h\right]\!\right]_{i-\frac{1}{2},j}.$$

Then each Θ-term can be rewritten as

$$\begin{aligned}\Theta_{i-\frac{1}{2},j} &= -a\{u^h\}_{i-\frac{1}{2},j}\left[\!\left[u^h\right]\!\right]_{i-\frac{1}{2},j} + \frac{|a|}{2}\left[\!\left[u^h\right]\!\right]^2_{i-\frac{1}{2},j} \\ &= -\frac{a}{2}\left[\!\left[(u^h)^2\right]\!\right]_{i-\frac{1}{2},j} + \frac{|a|}{2}\left[\!\left[u^h\right]\!\right]^2_{i-\frac{1}{2},j}\end{aligned} \tag{3.13}$$

Since $u^h \in \mathbb{Q}^0(K_{ij})$ we have that

$$(u^h)^+_{i-\frac{1}{2}} = (u^h)^-_{i+\frac{1}{2}} = u^h_i,$$

so, in particular, when we sum up over all i, the first term in 3.13 simplifies

$$\sum_i \left[\!\left[(u^h)^2\right]\!\right]_{i-\frac{1}{2},j} = \sum_i \left[(u^h)^+\right]^2_{i-\frac{1}{2},j} - \left[(u^h)^-\right]^2_{i-\frac{1}{2},j} = 0$$

and so

$$\sum_i \Theta_{i-\frac{1}{2}} = \sum_i \frac{|a|}{2}\left[\!\left[u^h\right]\!\right]^2_{i-\frac{1}{2}} \geq 0.$$

With the same procedure one can prove that with this choice of numerical fluxes also the third term in (3.10) is non-negative, which proves the L^2-stability of the numerical method.

■

3.3.2 Mass conservation

Let $u^h : [0, T_{end}] \to Z_n^0$ be the numerical solution of (3.10) for all $1 \leq i, j \leq 2^n$, then the following equalities hold

$$\sum_{i,j}\int_{K_{ij}} u^h(t)\ dxdy = \sum_{i,j}\int_{K_{ij}} u^h(0)\ dxdy = \sum_{i,j}\int_{K_{ij}} u_0\ dxdy, \qquad \forall t \in [0, T_{end}]. \tag{3.14}$$

Proof 6 Observe that the definition of the standard L^2-projection (3.8) and the fact that $u^h(0) = \mathcal{P}_h(u_0)$ imply

$$\sum_{i,j}\int_{K_{ij}} u^h(0)\ dxdy = \sum_{i,j}\int_{K_{ij}} \mathcal{P}_h(u_0)\ dxdy = \sum_{i,j}\int_{K_{ij}} u_0\ dxdy. \tag{3.15}$$

By taking

$$\psi^h(x,y) = \begin{cases} 1, & \text{if } (x,y) \in K_{ij} \\ 0, & \text{otherwise,} \end{cases}$$

in (3.10) we obtain

$$\int_{K_{ij}} \left(u^h\right)_t \, dxdy + \int_{J_j} \left[\left(\widehat{au^h}\right)_{i+\frac{1}{2},y} - \left(\widehat{au^h}\right)_{i-\frac{1}{2},y}\right] dy + \int_{I_i} \left[\left(\widehat{bu^h}\right)_{x,j+\frac{1}{2}} - \left(\widehat{bu^h}\right)_{x,j-\frac{1}{2}}\right] dx = 0.$$

Since this latter holds true for all i, j, summing up over all elements the boundary terms disappear because of periodic boundary conditions (in i and j). We get

$$\sum_{i,j} \int_{K_{ij}} \left(u^h\right)_t \, dxdy = 0,$$

integrating in time we obtain

$$\sum_{i,j} \int_{K_{ij}} u^h(t) \, dxdy - \sum_{i,j} \int_{K_{ij}} u^h(0) \, dxdy = 0$$

which, together with (3.15), yields (3.14).

■

3.3.3 Implementation - Basis functions

Standard basis

We start by using the usual standard basis functions χ_{ij} defined by

$$\chi_{ij}(x,y) = \chi_i(x)\chi_j(y) = \begin{cases} 1, \text{ if } (x,y) \in K_{ij}, \\ 0, \text{ otherwise.} \end{cases} \tag{3.16}$$

Considering (3.10), we observe that, by taking $\psi^h = \chi_{ij}$, we have

$$\psi^h|_{K_{ij}} = \chi_{ij}|_{K_{ij}} = 1, \quad \text{and} \quad u^h|_{K_{ij}} = \frac{1}{h_x h_y}(u^h, \chi_{ij})_{L^2} =: \bar{u}_{ij}.$$

Consequently the numerical solution can be written as

$$u^h(x,y,t) = \sum_{i,j=1}^{N} \bar{u}_{ij}(t)\chi_{ij}(x,y),$$

and the fluxes become

$$\left(\widehat{au^h}\right)_{i+\frac{1}{2}} = \begin{cases} au^h(x^-_{i+\frac{1}{2}}, y) = a\bar{u}_{ij} & \text{if } a(y) \geq 0 \\ au^h(x^+_{i+\frac{1}{2}}, y) = a\bar{u}_{i+1,j} & \text{if } a(y) < 0 \end{cases}$$

and

$$\left(\widehat{bu^h}\right)_{j+\frac{1}{2}} = \begin{cases} bu^h(x, y^-_{j+\frac{1}{2}}) = b\bar{u}_{ij} & \text{if } b(x) \geq 0 \\ bu^h(x, y^+_{j+\frac{1}{2}}) = b\bar{u}_{i,j+1} & \text{if } b(x) < 0. \end{cases}$$

Then (3.10) can be rewritten as

$$h_x h_y \left(\bar{u}_{ij}\right)_t + \int_{J_j} \left[\max(a(y),0)(\bar{u}_{ij} - \bar{u}_{i-1,j}) + \min(a(y),0)(\bar{u}_{i+1,j} - \bar{u}_{ij})\right] dy + \int_{I_i} \left[\max(b(x),0)(\bar{u}_{ij} - \bar{u}_{i,j-1}) + \min(b(x),0)(\bar{u}_{i,j+1} - \bar{u}_{i,j})\right] dx = 0,$$

which is similar to what one would obtain by using *finite differences* (with forward differences) and *finite volumes* (with upwind-fluxes) schemes.

Hierarchical basis

The construction of the hierarchical basis is done here using a **tensor product** construction of one dimensional hierarchical basis in each coordinates.
Before entering into details, we need to introduce the notation we will use. Basically, it's very similar to the one-dimensional case, in two dimension however, we use **bold** letters.

- $\diamond$ $\mathbf{l} = (l_1, l_2)$ is the **two-dimensional level** and it indicates how many times we halve each coordinate: l_1 refers to the one-dimensional level in the x-coordinate and l_2 to the one in the y-coordinate.
  Moreover, for $\mathbf{l} = (l_1, l_2)$ we define the following operators:

$$|\mathbf{l}|_\infty = \max(l_1, l_2) \quad \text{and} \quad |\mathbf{l}|_1 = l_1 + l_2.$$

- $\diamond$ $h_\mathbf{l} = (h_{l_1}, h_{l_2})$ indicates the grid size for each coordinate: $h_{l_1} = 2^{-l_1}$ refers to the x-coordinate and $h_{l_2} = 2^{-l_2}$ to the y-coordinate.

- $\diamond$ The two-dimensional grid points at level $\mathbf{l}$ are the ones given by

$$(x,y)_{\mathbf{l},\mathbf{i}} = (x_{l_1,i_1}, y_{l_2,i_2}), \qquad \text{for } i_1 = 1,\dots,2^{l_1} \text{ and } i_2 = 1,\dots,2^{l_2}$$

  in which the one-dimensional points x_{l_1,i_1} and y_{l_2,i_2} are defined as before (2.11).

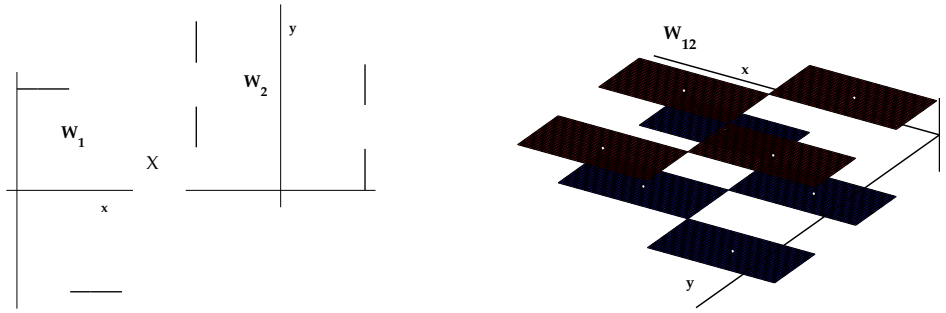
The hierarchical basis functions

We define the two-dimensional hierarchical basis functions by considering the tensor product of the corresponding one-dimensional ones:

$$\begin{aligned} \phi_{\mathbf{l},\mathbf{i}}(x,y) &= \phi_{(l_1,l_2),(i_1,i_2)}(x,y) = \phi_{l_1,i_1}(x)\phi_{l_2,i_2}(y), \\ \theta_{\mathbf{l},\mathbf{i}}(x,y) &= \theta_{(l_1,l_2),(i_1,i_2)}(x,y) = \theta_{l_1,i_1}(x)\theta_{l_2,i_2}(y) \end{aligned} \tag{3.17}$$

An example of this tensor product approach is shown in figure 3.2 : on the left, the one-dimensional hierarchical *Haar* basis functions in the different coordinates are given ($\theta_{1,1}(x)$, $\theta_{2,1}(y)$ and $\theta_{2,3}(y)$), on the right, the result of the tensor product among them is shown.

Figure 3.2: Tensor product approach for two-dimensional hierarchical *Haar* basis functions

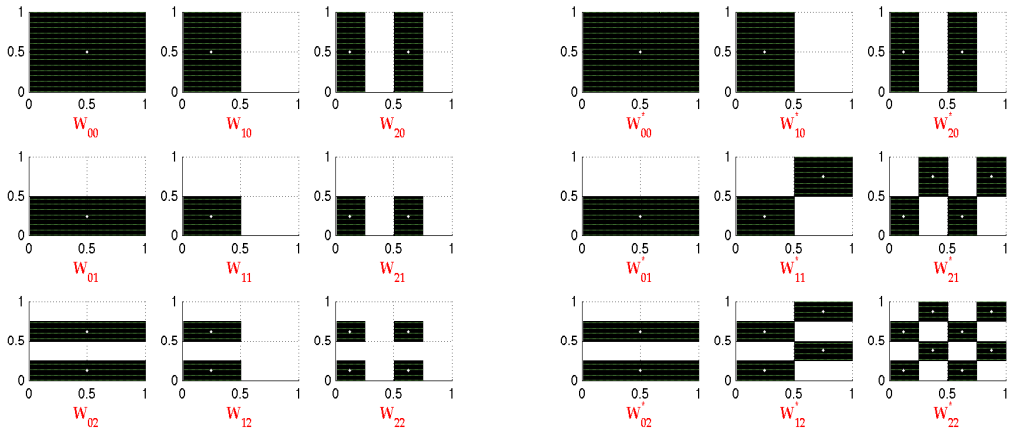


The definition of the two-dimensional hierarchical basis functions yields the following definitions for the subspaces $W_{\mathbf{l}}$ and $W_{\mathbf{l}}^*$ in two dimension:

$$W_{\mathbf{l}} = \{\phi_{\mathbf{l},\mathbf{i}} \mid i_1 = 1, \ldots, 2^{l_1} - 1 \text{ and } i_2 = 1, \ldots, 2^{l_2} - 1, \text{ with } i_1, i_2 \text{ odd}\}$$
$$W_{\mathbf{l}}^* = \{\theta_{\mathbf{l},\mathbf{i}} \mid i_1 = 1, \ldots, 2^{l_1} - 1 \text{ and } i_2 = 1, \ldots, 2^{l_2} - 1, \text{ with } i_1, i_2 \text{ odd}\}$$

In figure 3.3-3.4 the collection of subspaces $W_{\mathbf{l}}$ and $W_{\mathbf{l}}^*$ for $|\mathbf{l}|_\infty \leq 2$ are shown. On the left, the ones spanned by the set of hierarchial *One* basis functions, on the right, the ones spanned by the hierarchical *Haar* basis functions.

Figure 3.3: $W_{\mathbf{l}}$ (left) and $W_{\mathbf{l}}^*$ (right) for $|\mathbf{l}|_\infty \leq 2$



Because of the definition of the hierarchical basis functions the black and white parts in the different basis have different meaning.
For the hierarchical One basis (left) the black parts represent the non-zero value taken by the functions in the corresponding subspace while the white parts corresponds to the regions in which

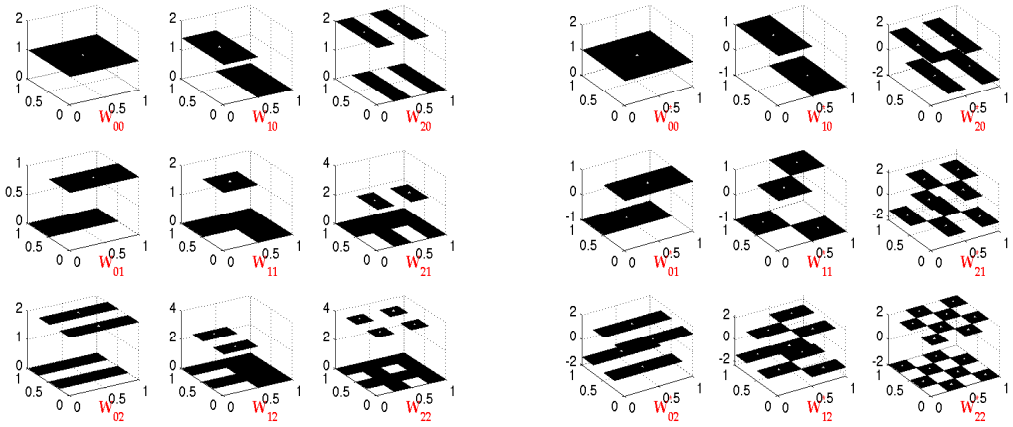
the basis functions are zero. For example the hierarchical One basis function in W_{11} is given by

$$\phi_{(1,1),(1,1)}(x,y) = \phi_{1,1}(x)\phi_{1,1}(y) = \begin{cases} 2, & \text{if } 0 \le x, y \le 0.5, \\ 0, & \text{otherwise.} \end{cases}$$

On the other hand, for hierarchical Haar basis the black parts correspond to the regions where the corresponding function takes its positive value while in the white parts it takes the negative one. Again, the example of the function of the subspace W_{11}^*

$$\theta_{(1,1),(1,1)}(x,y) = \theta_{1,1}(x)\theta_{1,1}(y) = \begin{cases} 1, & \text{if } 0 \le x, y \le 0.5 \text{ or } 0.5 \le x, y \le 1, \\ -1, & \text{otherwise.} \end{cases}$$

Figure 3.4: $W_{\mathbf{l}}$ (left) and $W_{\mathbf{l}}^*$ (right) for $|\mathbf{l}|_\infty \le 2$



Next, we state the two-dimensional version of the important property (2.14). Here, it involves the space Z_n^0 (3.7) and the subspaces $W_{\mathbf{l}}$ or $W_{\mathbf{l}}^*$ for the levels that satisfy $|\mathbf{l}|_\infty \le n$.

Lemma 3.3.2 *For a fixed $n \in \mathbb{N}$ we have that*

$$Z_n^0 = \bigoplus_{|\mathbf{l}|_\infty \le n} W_{\mathbf{l}} = \bigoplus_{|\mathbf{l}|_\infty \le n} W_{\mathbf{l}}^*.$$

Proof: the proof is done similarly as in the one-dimensional case, we omit the details for sake of conciseness.

3.3.4 Sparse grid

In the following we present the ingredient behind the sparse grid technique we use. The main idea of the sparse grid is to choose wisely the subspaces $W_{\mathbf{l}}$ or $W_{\mathbf{l}}^*$ in which we want to find our approximate solution in order to reduce the complexity of the method without losing too much accuracy.
We also have to point out that, in what follows, we give the presentation of the idea using the hierarchical Haar basis and the corresponding notation.
For an index set $\mathbf{I}$ that has to be identified, we define the sparse grid space Z_n^S as

$$Z_n^S = \bigoplus_{\mathbf{l}\in\mathbf{I}} W_{\mathbf{l}}^*,$$

in which the index S stands for *sparse*.
The definition of the index set $\mathbf{I}$ is obtained by solving a discrete optimization problem [1]. The question we ask is how to construct discrete approximation spaces that are better than Z_n^0 in the sense that the same relative error is obtained by a lower number of degrees of freedom.
In the following we look for an optimal space $Z_n^{(\mathrm{opt})}$ which will consist in a collection of finite sets $W_{\mathbf{l}}^*$, i.e.,

$$Z_n^{(\mathrm{opt})} = \bigoplus_{\mathbf{l}\in\mathbf{I}^{(\mathrm{opt})}} W_{\mathbf{l}}^*,$$

where $\mathbf{I}^{(\mathrm{opt})}$ is a finite index set for which the indices belong to $\mathbb{N}^2$.

We follow the discrete optimization presented by Bungartz and Griebel [1].

We search for an optimal grid $\mathbf{I}^{(\mathrm{opt})}$ among all possible grids $\mathbf{I} \subset \mathbf{I}^{(\mathrm{max})} := \{(0,0),\ldots,(n,n)\}$.
Notice that for any level $\mathbf{l}$ and some norm $||\cdot||$ we can bound the difference between the interpolant $u_{\mathbf{l}} \in W_{\mathbf{l}}^*$ ($u_{\mathbf{l}} = \sum_{\mathbf{i}} \alpha_{\mathbf{l},\mathbf{i}}\theta_{\mathbf{l},\mathbf{i}}$) and the solution u in the following way:

$$\left|\left| u - \sum_{\mathbf{l}\in\mathbf{I}} u_{\mathbf{l}} \right|\right|^2 \simeq \left|\left| \sum_{\mathbf{l}\in\mathbf{I}^{(\mathrm{max})}} u_{\mathbf{l}} - \sum_{\mathbf{l}\in\mathbf{I}} u_{\mathbf{l}} \right|\right|^2 \leq \sum_{\mathbf{l}\in\mathbf{I}^{(\mathrm{max})}\setminus\mathbf{I}} ||u_{\mathbf{l}}||^2$$

We start by presenting the *local cost* $c(\mathbf{l})$ and the *benefit* $b(\mathbf{l})$ functions for any two-dimensional level $\mathbf{l}$:

$$c(\mathbf{l}) := \text{number of degrees of freedom in the level } \mathbf{l}.$$

$$b(\mathbf{l}) := \text{upper bound for } ||u_{\mathbf{l}}||, \text{ where } ||\cdot|| \text{ is some norm.}$$

In order to state our optimization problem we need to define, for an arbitrary $\mathbf{I}$ its *global cost* $C(\mathbf{I})$ and *global benefit* $B(\mathbf{I})$.
For the global cost function one takes the sum of the local cost functions

$$C(\mathbf{I}) = \sum_{\mathbf{l}\in\mathbf{I}} c(\mathbf{l}) = \sum_{\mathbf{l}\in\mathbf{I}^{(\mathrm{max})}} \eta(\mathbf{l})\ c(\mathbf{l}),$$

where

$$\eta(\mathbf{l}) = \begin{cases} 1, & \text{if } \mathbf{l} \in \mathbf{I} \\ 0, & \text{if } \mathbf{l} \notin \mathbf{I}, \end{cases}$$

while the global benefit is derived considering the interpolant to u on the grid defined by $\mathbf{I}$ as follows

$$\begin{aligned}\left|\left|u-\sum_{\mathbf{l}\in\mathbf{I}}u_{\mathbf{l}}\right|\right|^2 &\simeq \left|\left|\sum_{\mathbf{l}\in\mathbf{I}^{(\max)}}u_{\mathbf{l}}-\sum_{\mathbf{l}\in\mathbf{I}}u_{\mathbf{l}}\right|\right|^2\\ &\leq \sum_{\mathbf{l}\in\mathbf{I}^{(\max)}\setminus\mathbf{I}}||u_{\mathbf{l}}||^2 \leq \sum_{\mathbf{l}\in\mathbf{I}^{(\max)}}(1-\eta(\mathbf{l}))\;b(\mathbf{l})\\ &= \sum_{\mathbf{l}\in\mathbf{I}^{(\max)}}b(\mathbf{l}) - \underbrace{\sum_{\mathbf{l}\in\mathbf{I}^{(\max)}}\eta(\mathbf{l})\;b(\mathbf{l})}_{=:B(\mathbf{I})}.\end{aligned}$$

Now, for some prescribed cost or work count w, the optimization problem is to find the grid $\mathbf{I}\subset\mathbf{I}^{(\max)}$ which maximizes the global benefit, i.e.,

$$\max_{\mathbf{I}\subset\mathbf{I}^{(\max)}}\sum_{\mathbf{l}\in\mathbf{I}^{(\max)}}\eta(\mathbf{l})\;b(\mathbf{l}) \tag{3.18}$$

for the fixed cost w, so it has to satisfy also

$$\sum_{\mathbf{l}\in\mathbf{I}^{(\max)}}\eta(\mathbf{l})\;c(\mathbf{l}) = w. \tag{3.19}$$

Arranging the $\mathbf{l}\in\mathbf{I}^{(\max)}$ in some linear order $i=1,\ldots,(n+1)^2=M$ with local cost $\mathbf{c}=(c_1,\ldots,c_M)^T$ and local benefit $\mathbf{b}=(b_1,\ldots,b_M)^T$, the optimization problem reads

$$\max_{\mathbf{x}}\;\mathbf{b}^T\mathbf{x} \qquad \text{with} \qquad \mathbf{c}^T\mathbf{x}=w,$$

where $w\in\mathbb{N}$ and $\mathbf{x}\in\{0,1\}^M$ is the vector of length M for which x_i is either 0 or 1 for all $i=1,\ldots,M$.

A problem of this kind is called *binary knapsack problem* which is known to be difficult to solve (NP-hard). However, by allowing $\mathbf{x}\in([0,1]\cup\mathbb{Q})^M$, a simple algorithm provide an optimal solution. The idea of the algorithm is to sort the *local* cost-benefit ratios b_i/c_i in decreasing order and then put a 1 to the corresponding position of $\mathbf{x}$ till the work count w is reached and set the rest to 0. In this way only one component of $\mathbf{x}$ would be rational.

1. rearrange the order in such a way that
$$\frac{b_1}{c_1}\geq\frac{b_2}{c_2}\geq\ldots\geq\frac{b_M}{c_M}.$$
2. set the limit of the knapsack to r, i.e.
$$r:=\max\{j:\sum_{i=1}^{j}c_i\leq w\}.$$
3. set the solution vector $\mathbf{x}=(x_1,\ldots,x_M)$ as follows:
$$\begin{aligned}x_i &:= 1, \qquad \text{for all } 1\leq i\leq r,\\ x_{r+1} &:= \left(w-\sum_{i=1}^{r}c_i\right)/c_{r+1},\\ x_k &:= 0, \qquad \text{for all } r+2\leq k\leq M.\end{aligned}$$

Because the work count w is an arbitrarily chosen natural number, our knapsack problem is of variable size, therefore it is possible to force the solution of the *rational* problem to be a *binary* one. Consequently, the optimization problem (3.18)-(3.19) can be reduced to the discussion of the *local* cost-benefit ratios $b(\mathbf{l})/c(\mathbf{l})$ of the underlying subspaces $W_{\mathbf{l}}^*$. Those subspaces with the best cost–benefit ratios are taken into account first, and the smaller these ratios become, the more negligible the underlying subspaces turn out to be.
In order to define these ratios, we first need to have the local cost and benefit functions.
The local cost function $c(\mathbf{l})$ depends on the degrees of freedom of the underlying subspace $W_{\mathbf{l}}^*$. We have

$$c(\mathbf{l}) = |W_{\mathbf{l}}^*| = |W_{l_1}^*|\ |W_{l_2}^*|.$$

We have seen previousely that $|W_0^*| = 1$ and for all $l \geq 1$ $|W_l^*| = 2^{l-1}$, these imply

$$c(\mathbf{l}) = \begin{cases} 1, & \text{if } l_1 = l_2 = 0, \\ 2^{l_1-1}, & \text{if } l_1 \neq 0,\ l_2 = 0, \\ 2^{l_2-1}, & \text{if } l_1 = 0,\ l_2 \neq 0, \\ 2^{|\mathbf{l}|_1-2}, & \text{if } l_1 \neq 0,\ l_2 \neq 0. \end{cases}$$

For the local benefit function we take a bound of $||u_{\mathbf{l}}||^2$ in some norm. By taking the L^2-norm we have

$$||u_{\mathbf{l}}||_0^2 \ = \ \left|\left|\sum_{\mathbf{i}} \alpha_{\mathbf{l},\mathbf{i}}\theta_{\mathbf{l},\mathbf{i}}\right|\right|_0^2 \ \leq \ \sum_{\mathbf{i}} |\alpha_{\mathbf{l},\mathbf{i}}|^2\ ||\theta_{\mathbf{l},\mathbf{i}}||_0^2. \tag{3.20}$$

It follows that we need a bound for both coefficients and basis functions.
We notice that for a function $\theta_{\mathbf{l},\mathbf{i}}$ the measure of its support is $|\mathrm{supp}(\theta_{\mathbf{l},\mathbf{i}})| = 2^{-|\mathbf{l}|_1+2}$,
then we have

$$||\theta_{\mathbf{l},\mathbf{i}}||_0^2 = \int_{\mathrm{supp}(\theta_{\mathbf{l},\mathbf{i}})} \left(2^{\frac{l_1-1}{2}} 2^{\frac{l_2-1}{2}}\right)^2 dx = 2^{-|\mathbf{l}|_1+2}\ 2^{2\frac{|\mathbf{l}|_1-2}{2}} = 2^0 = 1,$$

for the basis functions.
For the coefficients the bound is obtained in a more complicated way.
First, let $\mathbf{C} \in \mathbb{R}^{2^{2n}\times 2^{2n}}$ be the matrix which is needed to change basis:

$$\bar{u} = \mathbf{C}\alpha.$$

Observe that a linear ordering for the two-dimensional coefficients has been used: $\bar{u} = (\bar{u}_{11}, \bar{u}_{12}, \ldots, \bar{u}_{2^n 2^n})^T$ and $\alpha = (\alpha_{(0,0),(1,1)}, \ldots, \alpha_{(n,n),(2^n,2^n)})^T$.
Notice that because we are using hierarchical *Haar* basis functions (L^2-orthogonal) the matrix $\mathbf{C}$ results to satisfy

$$\mathbf{C^{-1}} = 2^{-2n}\mathbf{C^T}.$$

Let $p \in \mathbb{N}$ to represent any couple $(\mathbf{l}, \mathbf{i})$ and $k \in \mathbb{N}$ an index of the linear ordering of the couples (i, j), then the bound on the coefficient α_p is

$$|\alpha_p| = \left|\sum_k \left(\mathbf{C^{-1}}\right)_{pk} \bar{u}_k\right| \ = 2^{-2n}\left|\sum_k \left(\mathbf{C^T}\right)_{pk} \bar{u}_k\right| \ \leq 2^{-2n}\sum_k \left|(\mathbf{C})_{kp}\right| \max_k(\bar{u}_k).$$

Observe that $|\text{supp}(\chi_{ij})| = 2^{-2n}$ and recalling that $\theta_{\mathbf{l},\mathbf{i}}$ is either $\pm 2^{\frac{|\mathbf{l}|_1-1}{2}}$ or 0, we have

$$\mathbf{C}_{kp} = (\chi_k, \theta_p)_{L^2} = \begin{cases} \int_{2^{-2n}} \pm 2^{\frac{|\mathbf{l}|_1-1}{2}}, & \text{if } \text{supp}(\chi_k \cap \theta_p) = \text{supp}(\chi_k) \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore we have,

$$\sum_k \zeta(\chi_k, \theta_p) = 2^{2n-|\mathbf{l}|_1+2},$$

where

$$\zeta(\chi_k, \theta_p) = \begin{cases} 1, & \text{if } |\text{supp}(\chi_k \cap \theta_p)| = |\text{supp}(\chi_k)| \\ 0, & \text{otherwise.} \end{cases}$$

With this preliminaries we can now focus on the following

$$\sum_k |\mathbf{C}_{kp}| = \sum_k \zeta(\chi_k, \theta_p)| \pm 2^{\frac{|\mathbf{l}|_1-1}{2}} 2^{-2n}| = 2^{\frac{-|\mathbf{l}|_1-1}{2}} 2^{-2n} 2^{2n-|\mathbf{l}|_1+2} = 2^{-\frac{|\mathbf{l}|_1-3}{2}}.$$

So for the bound of the coefficient α_p we have:

$$|\alpha_p| \leq 2^{-2n} 2^{-\frac{|\mathbf{l}|_1-3}{2}} \max_k(\bar{u}_k)$$

Then (3.20) becomes

$$||u_{\mathbf{l}}||_0^2 \leq \sum_{\mathbf{i}} |\alpha_{\mathbf{l},\mathbf{i}}|^2 \leq 2^{2n} 2^{-4n-|\mathbf{l}|_1+3} \max_k \bar{u}_k$$

and with this latter result the *local cost-benefit* ratios, denoted by cbr, turns out to be

$$\mathsf{cbr}(\mathbf{l}) := \frac{b(\mathbf{l})}{c(\mathbf{l})} = \frac{2^{-2n-|\mathbf{l}|_1+3} \max_k(\bar{u}_k)}{2^{|\mathbf{l}|_1-2}} = 2^{-2n-2|\mathbf{l}|_1+5} \max_k(\bar{u}_k)$$

An optimal grid $\mathbf{I}^{(\text{opt})}$ consists of all levels $\mathbf{l}$ where $\mathsf{cbr}(\mathbf{l})$ is bigger than some prescribed threshold $\varkappa(n)$ that we choose to be of the order of $\mathsf{cbr}(\bar{\mathbf{l}})$ with $\bar{\mathbf{l}} := (n, 0)$:

$$\varkappa(n) := \mathsf{cbr}(\bar{\mathbf{l}}) = 2^{-4n+5} \max_k(\bar{u}_k).$$

That is, we search for all $W_{\mathbf{l}}^*$ whose cost–benefit ratio is equal or better than the one of the subspace $W_{(n,0)}^*$. Thus, applying the criterion $\mathsf{cbr}(\mathbf{l}) \geq \varkappa(n)$ we have the following

$$2^{-2n-2|\mathbf{l}|_1+5} \max_k(\bar{u}_k) \geq 2^{-4n+5} \max_k(\bar{u}_k) \quad \Rightarrow \quad -2n - 2|\mathbf{l}|_1 + 5 \geq -4n + 5 \quad \Rightarrow \quad |\mathbf{l}|_1 \leq n.$$

Thus, the relation

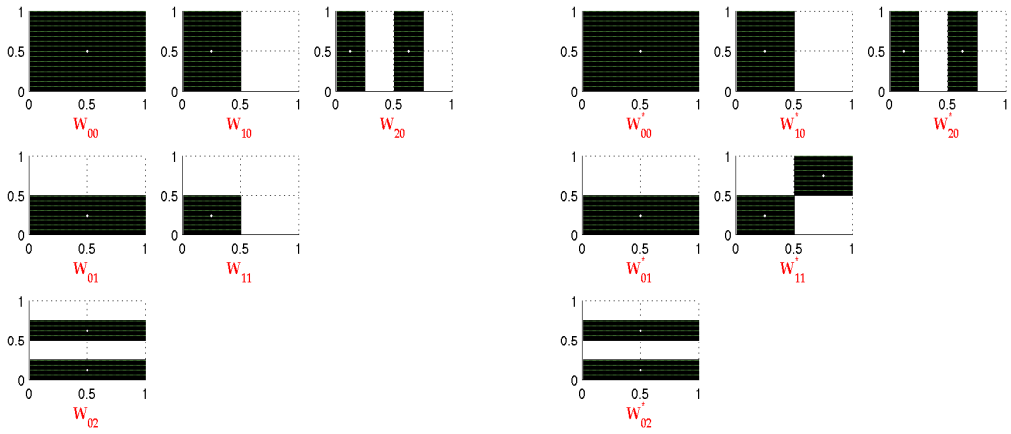
$$|\mathbf{l}|_1 \leq n$$

qualifies a subspace $W_{\mathbf{l}}^*$ to be taken into account. This result leads us to the definition of the desired approximation space

$$Z_n^S = \bigoplus_{|\mathbf{l}|_1 \leq n} W_{\mathbf{l}}^*,$$

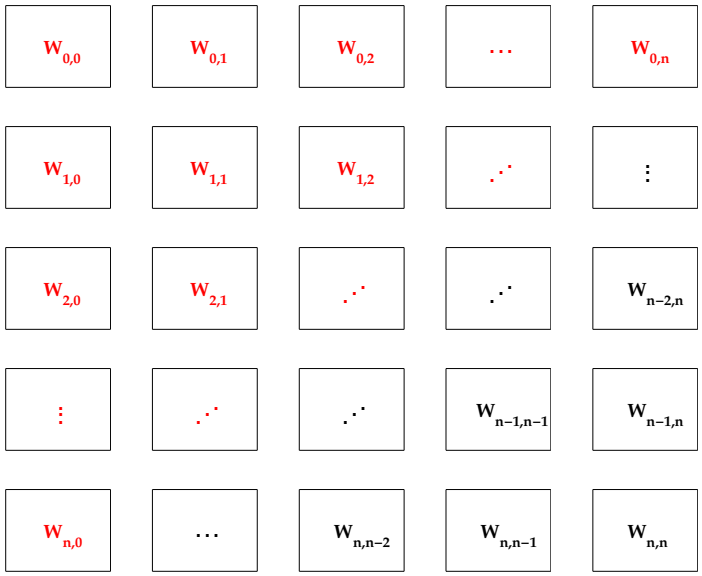
which is optimal with respect to our cost–benefit setting. The selected subspaces $W_{\mathbf{l}}^*$ and $W_{\mathbf{l}}$, for the case with $n = 2$ are shown in figure 3.5.

Figure 3.5: Sparse grid for hierarchical *One* (left) and *Haar* (right) basis functions



For general $n \in \mathbb{N}$, the sparse grid Z_n^S is the collection of subspaces which lie in the upper-left triangle in our representation's diagram of the subspaces (see figure 3.6).

Figure 3.6: Selected subspaces in the two-dimensional case



Mass conservation

Also in the Sparse grid space, the approximation $u^h : [0, T_{end}] \to Z_n^S$ that solves (3.10) for all $1 \leq i, j \leq 2^n$ conserves the total mass, i.e., we have

$$\sum_{i,j} \int_{K_{ij}} u^h(t)\ dxdy = \sum_{i,j} \int_{K_{ij}} u^h(0)\ dxdy \quad \forall t \in [0, T_{end}]. \tag{3.21}$$

Proof 7 Since the hierarchical basis functions $\phi_{\mathbf{l},\mathbf{i}}$ have global support we have to consider the DG-formulation on the whole domain

$$\sum_{i,j} \int_{K_{ij}} \left(u^h\right)_t \psi^h\ dxdy + \sum_{i,j} \int_{J_j} \left[\left(\widehat{au^h} \left(\psi^h\right)^- \right)_{i+\frac{1}{2},y} - \left(\widehat{au^h} \left(\psi^h\right)^+ \right)_{i-\frac{1}{2},y} \right] dy \tag{3.22}$$

$$+ \sum_{i,j} \int_{I_i} \left[\left(\widehat{bu^h} \left(\psi^h\right)^- \right)_{x,j+\frac{1}{2}} - \left(\widehat{bu^h} \left(\psi^h\right)^+ \right)_{x,j-\frac{1}{2}} \right] dx = 0. \tag{3.23}$$

By taking $\psi^h = 1\ \forall\ (x, y)$ this latter equation becomes

$$\sum_{i,j} \int_{K_{ij}} \left(u^h\right)_t\ dxdy + \sum_{i,j} \int_{J_j} \left[\left(\widehat{au^h} \right)_{i+\frac{1}{2},y} - \left(\widehat{au^h} \right)_{i-\frac{1}{2},y} \right] dy$$

$$+ \sum_{i,j} \int_{I_i} \left[\left(\widehat{bu^h} \right)_{x,j+\frac{1}{2}} - \left(\widehat{bu^h} \right)_{x,j-\frac{1}{2}} \right] dx = 0.$$

Observe that we have $u^h = \sum_{\mathbf{l},\mathbf{i}} \alpha_{\mathbf{l},\mathbf{i}} \phi_{\mathbf{l},\mathbf{i}}$, so for all $(\mathbf{l}, \mathbf{i})$ let us define

$$J_{\mathbf{l},\mathbf{i}} := \text{all vertical edges in which the function } \phi_{\mathbf{l},\mathbf{i}} \text{ has non-zero jump,}$$
$$I_{\mathbf{l},\mathbf{i}} := \text{all horizontal edges in which the function } \phi_{\mathbf{l},\mathbf{i}} \text{ has non-zero jump.}$$

Now, with this new notation we rewrite the boundary terms in (3.22) obtaining

$$\sum_{i,j} \int_{J_j} \left[\left(\widehat{au^h} \right)_{i+\frac{1}{2},y} - \left(\widehat{au^h} \right)_{i-\frac{1}{2},y} \right] dy = \sum_{\mathbf{l},\mathbf{i}} \int_{J_{\mathbf{l},\mathbf{i}}} \sum_{i,j} \left[\left(\widehat{a\alpha_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}}} \right)_{i+\frac{1}{2},y} - \left(\widehat{a\alpha_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}}} \right)_{i-\frac{1}{2},y} \right] dy = 0$$

$$\sum_{i,j} \int_{I_i} \left[\left(\widehat{bu^h} \right)_{x,j+\frac{1}{2}} - \left(\widehat{bu^h} \right)_{x,j-\frac{1}{2}} \right] dx = \sum_{\mathbf{l},\mathbf{i}} \int_{I_{\mathbf{l},\mathbf{i}}} \sum_{i,j} \left[\left(\widehat{b\alpha_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}}} \right)_{x,j+\frac{1}{2}} - \left(\widehat{b\alpha_{\mathbf{l},\mathbf{i}}\phi_{\mathbf{l},\mathbf{i}}} \right)_{x,j-\frac{1}{2}} \right] dx = 0.$$

Again, summing up over all i and j the terms telescope and we obtain only one integral

$$\sum_{ij} \int_{K_{ij}} \left(u^h\right)_t\ dxdy = 0,$$

and, as before, integration in time yields the desired result.

■

3.4 Numerical experiments

3.4.1 Approximations on the sparse grid spaces

We compare the approximation in the sparse grid spaces spanned by the two set of hierarchical basis functions: the ones based on the *One* basis and the ones based on the *Haar* basis.

We observe the relative errors for different functions defined on the domain $[-1,1]^2$ on different uniform meshes.

We remind that N represents the number of intervals in which we devide each coordinate and by Z_n^0 and Z_n^S the whole space and the space of the sparse grid are denoted.

The obtained results are collected in the following table. Observe that the results that follows hold for both of the basis.

Functions		$N=8$	$N=16$	$N=32$	$N=64$	$N=128$
x^2	L^2-*error*	0.00519e-15	0.00397e-15	0.06145e-15	0.46789e-15	0.94153e-15
$x-y$	L^2-*error*	3.67682e-16	3.48507e-16	2.65130e-15	8.14088e-15	3.56748e-14
xy	L^2-*error*	0.35952	0.20814	0.11705	0.06444	0.03494
	rates	0.7885	0.8305	0.8610	0.8830	
$e^{-4y^2}\lvert x\rvert$	L^2-*error*	0.30011	0.19173	0.11350	0.06464	0.03588
	rates	0.6462	0.7566	0.8120	0.8494	
x^2y	L^2-*error*	0.42592	0.26276	0.15345	0.08658	0.04774
	rates	0.6969	0.7759	0.8257	0.8587	
x^2+y	L^2-*error*	0.0006e-12	0.00025e-12	0.00700e-12	0.05191e-12	0.10972
$\sin(\pi x)\sin(\pi x)$	L^2-*error*	0.52100	0.30744	0.22478	0.14203	0.08364
	rates	0.7610	0.45176	0.6624	0.7640	
$e^{-4y^2}\sin(x)$	L^2-*error*	0.36347	0.21183	0.12054	0.06699	0.03657
	rates	0.7789	0.8134	0.8476	0.8730	
1 if $x,y>1-2^{-4}$, 0 else	L^2-*error*	NaN	NaN	0.94373	0.90139	0.82916
	rates	NaN	NaN	0.0662	0.1205	
dof Z_n^0		64	256	1024	4096	16384
dof Z_n^S		20	48	112	256	576

For the initial projection, one notes that the functions that only depends on one variable are well approximated. Moreover, even linear combinations of functions depending only on one variables are not affected by the truncation. Finally one can note the increasing difference between the degrees of freedom involved for the different spaces and for any choice of grid.

3.4.2 2D Transport equation

In the following experiments we compare the numerical solutions of different cases of the two dimensional transport equation. The results refering to the whole space are obtained by using the *standard* basis (3.16) while the results on the sparse grid spaces refer to the set of hierarchical Haar basis functions.

Constant coefficients

We consider the two-dimensional transport equation with constants coefficients

$$\begin{aligned} u_t(x,y,t) + u_x(x,y,t) + u_y(x,y,t) &= 0, && \text{for } (x,y) \in [0,1]^2, \\ u(x,y,0) &= \sin(2\pi x)\sin(2\pi y) && \text{for } (x,y) \in [0,1]^2, \\ u(0,y,t) = u(1,y,t) \text{ and } u(x,0,t) &= u(x,1,t) && \text{for all } t. \end{aligned} \tag{3.24}$$

This problem can be solved analytically and the exact solution at any time t is given by

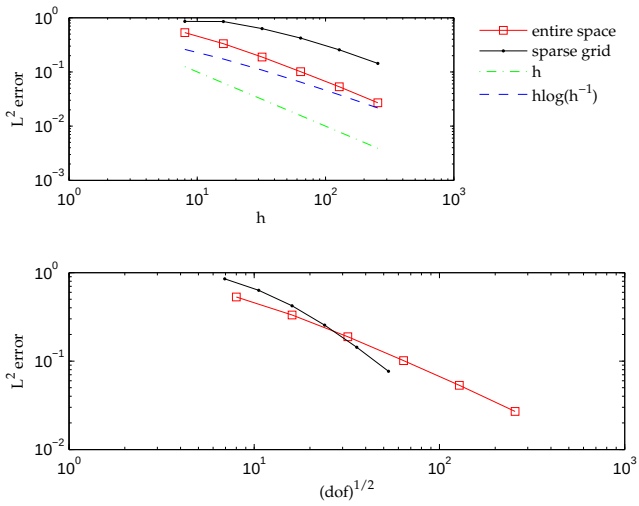
$$u_{ex}(x,y,t) = \sin(2\pi(x-t))\sin(2\pi(y-t)).$$

We use this model problem for both the whole space and the sparse grid space.
In the following table we collect the relative L^2-errors and the rates of convergence obtained for the different mesh sizes.
N indicates the number of interval in which both coordinates are devided, we performed the computations up to time $T_{end} = 0.25$.

N		8	16	32	64	128	256
Whole space	L^2*-error*	0.53033	0.33178	0.18870	0.10120	0.05320	0.026949
	rates	0.6767	0.8141	0.8989	0.9278	0.9812	
Sparse grid	L^2*-error*	0.85731	0.85202	0.63136	0.42160	0.25521	0.14331
	rates	0.0089	0.4324	0.5826	0.7242	0.8325	
Degrees of freedom	*Whole space*	64	256	1024	4096	16384	65536
	Sparse grid	20	48	112	256	576	1280

In figure 3.7 the convergence diagrams for the relative errors (in L^2-norm) of the approximate solution are given. In the diagram above we show the convergence with respect to the size of the elements in the uniform grid $h = 1/N$. In the diagram below the convergence with respect to the number of degrees of freedom are given. In both diagrams the black lines (with dots) refer to the sparse grid while the red lines (with squares) refer to the whole space. In the first diagram the dashed green line refers to the order of accuracy h and the dashed blue one to the order $h\log(h^{-1})$.

Figure 3.7: Convergence: whole space vs. sparse grid.



In the figure below we observe that taking into account the amount of degrees of freedom, at some point (already in two dimensions), we have that the error in the sparse grid is lower than the one obtained in the whole space when we consider the same number of degrees of freedom. This means that, for the same amount of degrees of freedom, the sparse grid technique becomes more accurate than the whole space after a certain number of degrees of freedom and for a fixed level of error it costs much less.

Mass Conservation

We now study how the discrete mass of a given initial condition is preserved in time.
We perform the experiment in both the whole and the sparse spaces. We compute the difference of mass between the initial data and all the future time, i.e., we look at

$$\Delta_{\text{Mass}}(t) := \frac{\int_\Omega u^h(x,y,0)dxdy - \int_\Omega u^h(x,y,t)dxdy}{\int_\Omega u^h(x,y,0)dxdy}$$

for all discrete time-steps $t \in [0, T_{end}]$.

For this experiment we slightly change the problem and we consider

$$\begin{aligned} u_t(x,y,t) + u_x(x,y,t) + u_y(x,y,t) &= 0, && \text{for } (x,y) \in [0,1]^2, \\ u(x,y,0) &= \sin(\pi x)\sin(\pi y) && \text{for } (x,y) \in [0,1]^2, \\ u(0,y,t) = u(1,y,t) \text{ and } u(x,0,t) &= u(x,1,t) && \text{for all } t. \end{aligned} \tag{3.25}$$

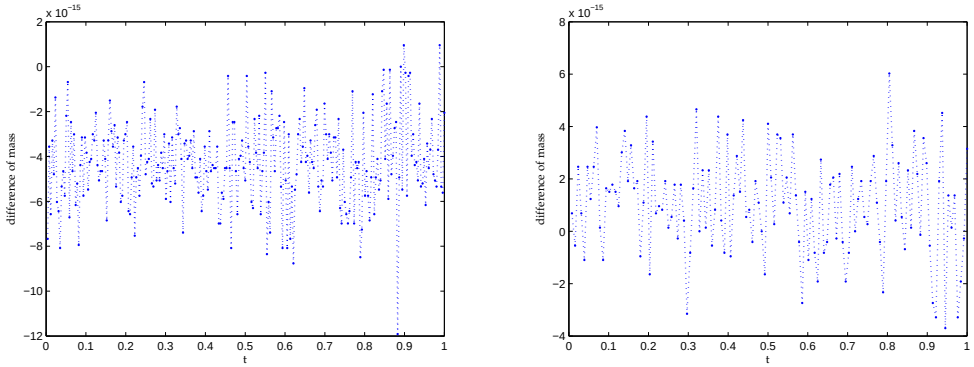
We point out that we choose a non-negative initial data and that the exact solution at any time t is given by the expression

$$u_{ex}(x,y,t) = \sin(\pi(x-t))\sin(\pi(y-t)).$$

We perform computations up to time $T_{end} = 1$ on a uniform grid 64×64.

On the left part of figure (3.8) the conservation of mass for the whole space is displayed. On the right the one of the sparse grid is given.

Figure 3.8: Conservation of mass: whole space vs. sparse grid.



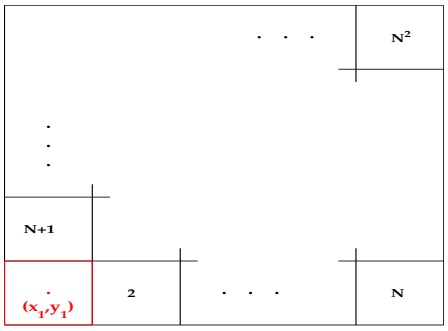
Observe that the order of magnitude in both figures is 10^{-15}, which indicates that in both cases the difference of mass between the numerical solution at time $t = 0$ and the one computed at any future discrete time $t \in (0, 1]$ is close to machine precision (hence zero), so we can conclude that the total mass is conserved. These results confirms the theoretical results we stated in previous sections.

Positivity

In the following we study the positivity of the method. When we start with a non-negative initial condition, we would like the numerical solution to be non-negative for all future times.
To this end we observe the behaviour of the numerical solution in a specific element in both the whole and sparse space for the same problem used for testing the conservation of mass (3.25).

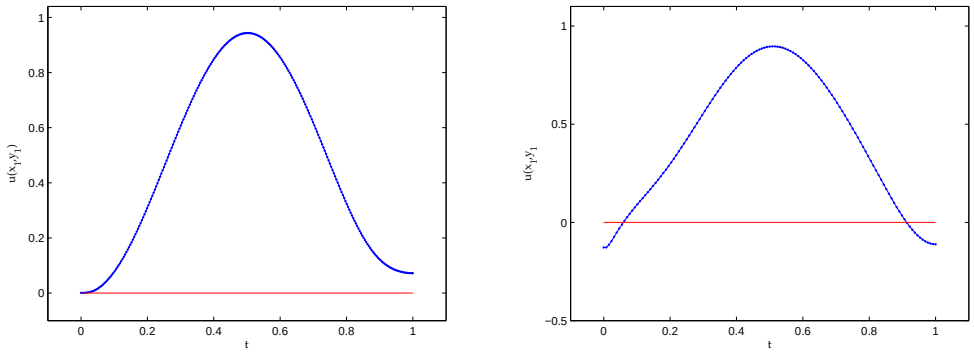
The reference element is the one depicted in figure 3.9, in which, given our initial condition, the value of the solution is close to zero.

Figure 3.9: The observed element.



In figure 3.10 two different graphs are given: we have the approximated solutions in the reference element using the whole space (left) and the sparse grid (right).

Figure 3.10: Positivity: approximate solution for the whole space (left) and the sparse grid (right)



We observe that there are time-steps in which the numerical solution on the sparse grid is negative.
In particular, at the beginning, performing the projection on the sparse space, the numerical solution is already negative. Therefore, by starting with such initial data it is hopeless to expect positivity of the approximate solution at later time.

Next, we try to modify the initial projection or the choice of spaces in such a way that the positivity (at least at the beginning) could be preserved.

We do it using two different approaches:

1. Among all subspaces that we have removed to construct the sparse grid finite element space, we select (and re-use) the ones that give a greater contribution to the projection on the entire space spanned by the complete set of hierarchical basis functions.

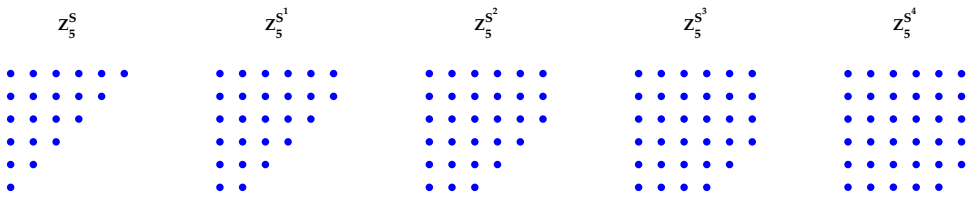
   This selection turns out to be the sequential addiction of the diagonals which were not considered after the selection of the subspaces (see figure 3.11).
   We consider the initial projection on $[0,1]^2$ of the function

   $$u(x,y,0) = \sin(\pi x)\sin(\pi y)\,. \tag{3.26}$$

   In this experiment we consider a uniform grid 32×32 ($n = 5$) and we observed the minimal value taken by the projection on the following sparse spaces:

   - $\diamond$ $Z_5^S = \bigoplus_{|\mathbf{l}|_1 \leq 5, |\mathbf{l}|_\infty \leq 5} W_{\mathbf{l}}$.
   - $\diamond$ $Z_5^{S^1} = \bigoplus_{|\mathbf{l}|_1 \leq 6, |\mathbf{l}|_\infty \leq 5} W_{\mathbf{l}}$.
   - $\diamond$ $Z_5^{S^2} = \bigoplus_{|\mathbf{l}|_1 \leq 7, |\mathbf{l}|_\infty \leq 5} W_{\mathbf{l}}$.

Figure 3.11: Sparse space considered.



- $Z_5^{S^3} = \bigoplus_{|\mathbf{l}|_1 \leq 8, |\mathbf{l}|_\infty \leq 5} W_{\mathbf{l}}$.
- $Z_5^{S^4} = \bigoplus_{|\mathbf{l}|_1 \leq 9, |\mathbf{l}|_\infty \leq 5} W_{\mathbf{l}}$.

This selection of spaces is showed in figure 3.11.
In the next table we write the minimal values taken by the projection on the sparse grids and the degrees of freedom of the corresponding spaces.

	Z_5^S	$Z_5^{S^1}$	$Z_5^{S^2}$	$Z_5^{S^3}$	$Z_5^{S^4}$	Z_5^0
$\min(u^h(x_1, y_1))$	-0.17900	-0.09624	-0.03607	-0.00935	0.00002	0.00241
Degrees of freedom	112	192	320	512	768	1024

We observe that the only projection that preserves the positivity of the initial conditions (3.26) is the one on the space $Z_5^{S^4}$, in which the only subspace removed with respect to the whole space Z_5^0 is $W_{5,5}$.

The problem in this case is the number of degrees of freedom that one is considering. This number increased considerably: in the case observed ($n = 5$), it's almost 7 times the one of the usual sparse grid $\left(Z_5^S\right)$. Therefore, the whole idea of sparse gird technique is lost.

2. The second approach keeps the sparse space fixed (f.e. Z_5^S) and uses a sort of "brute force" to obtain non-negative initial condition: the idea is to correct the projection in each element by adding the minimal value taken by the function in the whole domain. In this way the problem of the increase of the number of degrees of freedom is overcome.

   Also in this case we consider the domain $[0, 1]^2$ and the initial conditions given by (3.26). We consider the space Z_n^S for n =3, 4, 5, 6, 7, which means that we devide each coordinate in 8, 16, 32, 64, 128 intervals respectively.
   We observe the relative error between the exact solution and the approximated ones in the sparse space by using different initial data: from one side we use the usual projection

$u^h(0) = \mathcal{P}_h(u_0)$ which, as we have seen before, yields some negative values. From the other side we use initial data $u^h(0) = \bar{\mathcal{P}}_h(u_0)$ where the new projection $\bar{\mathcal{P}}_h$ is defined by

$$u \in L^2 \mapsto \bar{\mathcal{P}}_h(u) = \mathcal{P}_h(u) + |\min(\mathcal{P}_h(u))| \in Z_n^S.$$

The relative errors are calculated using the L^2-norm, and the results are collected in the following table.

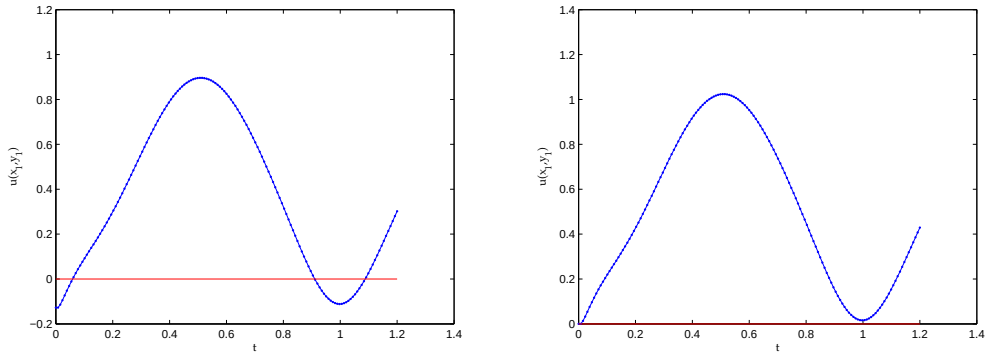
n		3	4	5	6	7
Sp.grid with $u^h(0) = \mathcal{P}_h(u_0)$	L^2-*error*	0.08947	0.06212	0.03807	0.02203	0.012339
	rates	0.5262	0.7064	0.7892	0.8364	
Sp.grid with $u^h(0) = \bar{\mathcal{P}}_h(u_0)$	L^2-*error*	0.18378	0.22075	0.18300	0.12928	0.08378
	rates	-0.2644	0.2706	0.5013	0.6258	

We notice that by using the new projection for the initial data some accuracy is lost and the order degrades, in fact the relative errors increase considerably: depending on the grid, it's from 2 to 8 times bigger.

Next we observe the behaviour in time of the numerical solution of the transport equation with constant coefficients, when we use the usual $\mathcal{P}_h$ and the modified projection $\bar{\mathcal{P}}_h$ on the initial data.
Also in this case we consider the problem (3.25). We use a uniform grid 64×64 and we observe the numerical solution computed in the same reference element as before (see figure 3.9) but we perform the experiments up to time $T_{end} = 1.2$.

In figure 3.12 the numerical solution in the case of the usual projection (left) and in the case of the modified projection (right) are given.

Figure 3.12: Behaviour in time: usual vs. modified projection.



Notice that the correction's value we add at the beginning in order to have a non-negative initial condition is enough to prevent the numerical solution to become negative in a future time. Thus, the problem of a slightly negative initial condition can be overcome with this method. However, by doing so, one has to be aware of the fact that the error in the numerical solution will increase, and the order will degrade a bit.

Variable coefficients

We consider the following forced transport equation with variable coefficients

$$\begin{aligned} u_t(x,y,t) + yu_x(x,y,t) + x^2u_y(x,y,t) &= \xi(x,y,t), && \text{for } (x,y) \in [-1,1]^2, \\ u(x,y,0) &= [1-\cos(\pi(x+1))]\, e^{-8y^2} && \text{for } (x,y) \in [-1,1]^2, \\ u(-1,y,t) = u(1,y,t) \text{ and } u(x,-1,t) &= u(x,1,t) && \text{for all } t \geq 0. \end{aligned}$$

The forced term on the right part of the first equality is defined in order for

$$u_{ex}(x,y,t) = [1-\cos(\pi(x+1)+2\pi t)]\, e^{-8y^2}$$

to be the exact solution, i.e.

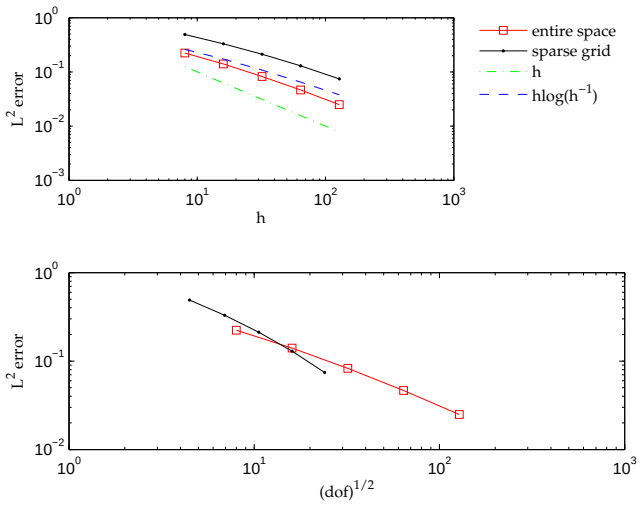
$$\xi(x,y,t) = e^{-8y^2}\left[(2+y)\pi\sin(\pi(x+1)+2\pi t) - 16yx^2\left(1-\cos(\pi(x+1)+2\pi t)\right)\right].$$

Similarly to the case in which the coefficients were constants we consider the convergence in both the whole and the sparse spaces. Because of the 1-periodicity of the exact solution we performed the experiment up to time $T_{end} = 1$, and we obtained the following (relative) L^2-errors and the corresponding rates of convergence.

N		8	16	32	64	128
Whole space	L^2*-error*	0.22328	0.14008	0.08261	0.04691	0.02486
	rates	0.6726	0.7618	0.8289	0.9038	
Sparse grid	L^2*-error*	0.48990	0.32978	0.21235	0.12962	0.07438
	rates	0.5710	0.6351	0.7121	0.8013	
Degrees of freedom	*Whole space*	64	256	1024	4096	16384
	Sparse grid	20	48	112	256	576

In figure 3.13 the convergence in the whole and in the sparse space are compared. In the figure above with respect to $h = 2/N$, while below we consider the square root of the degrees of freedom. Again the black lines refer with dots to the sparse grid while the red ones with squares to the whole space. In the first diagram the dashed green line represent the order h and the dashed blue the order of $h\log(h^{-1})$.

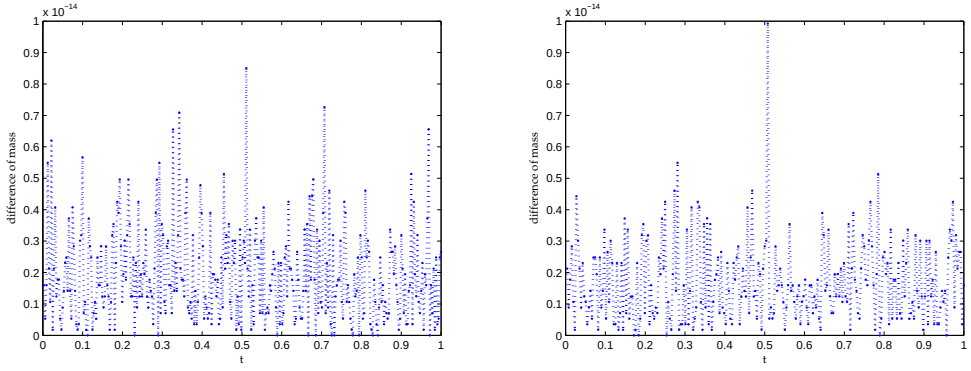
Figure 3.13: Convergence: whole space vs. sparse grid.



Mass conservation

Also for the forced transport equation with variable coefficient we check the mass conservation. We consider the same case used to show the convergence. We obtain the following diagrams, in which for both spaces the order of magnitude is 10^{-15}, which says that also in this case the method conserves the mass, up to machine precision.

Figure 3.14: Conservation of mass: whole space vs. sparse grid.



3.5 An alternative method

3.5.1 Alternative sparse grid

In this section we consider the sparse grid Z_n^S in an alternative way, which allows to reduce even more the size of the system of equations one has to solve.
In this section we will use the notation of the hierarchical One basis functions, but the whole holds also for Haar basis. We consider the following splitting of the sparse grid space:

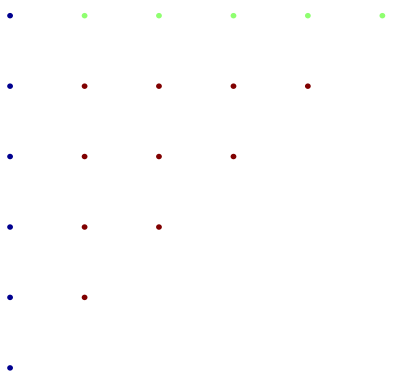
$$Z_n^S = \bigoplus_{|\mathbf{l}|_1 \leq n} W_{\mathbf{l}} = Z_n^{[1]} \oplus Z_n^{[2]} \oplus Z_n^{[R]}. \tag{3.27}$$

where

- $Z_n^{[1]} := \bigoplus_{\mathbf{l} \text{ s.t. } l_1=0} W_{\mathbf{l}} = \{w \in L^2(\Omega) \mid w \in \mathbb{P}^0(\Omega_x) \otimes \mathbb{P}^0(J_j)\ \forall j\}$.
- $Z_n^{[2]} := \bigoplus_{\mathbf{l} \text{ s.t. } l_2=0, l_1 \neq 0} W_{\mathbf{l}} = \{w \in L^2(\Omega) \mid w \in \mathbb{P}^0(I_i) \otimes \mathbb{P}^0(\Omega_y)\ \forall i\}$.
- $Z_n^{[R]} := Z_n^S \setminus \left(Z_n^{[1]} \cup Z_n^{[2]}\right)$.

A representation of these subspaces is shown in figure (3.15). Observe that the splitting (3.27)

Figure 3.15: The splitting: $Z_5^{[1]}$, $Z_5^{[2]}$ and $Z_5^{[R]}$



allows us to write a function φ^h belonging to Z_n^S as

$$\varphi^h = \varphi_{[1]}^h + \varphi_{[2]}^h + \varphi_{[R]}^h, \tag{3.28}$$

where $\varphi_{[1]}^h \in Z_n^{[1]}$, $\varphi_{[2]}^h \in Z_n^{[2]}$ and $\varphi_{[R]}^h \in Z_n^{[R]}$.

Notice that for all subspaces W_1 in $Z_n^{[1]}$ the functions only have jumps in the y-coordinate, and, likewise the functions in $Z_n^{[2]}$ only have jumps in the x-coordinate. This implies that, when we consider functions $z_{[1]}^h \in Z_n^{[1]}$ and $z_{[2]}^h \in Z_n^{[2]}$, we have,

$$\left[\!\left[z_{[1]}^h \right]\!\right]_x = 0 \qquad\qquad \left[\!\left[z_{[2]}^h \right]\!\right]_y = 0. \tag{3.29}$$

The idea is to divide the integration in time in two stages: before one computes the numerical solution considering only the spaces $Z_n^{[1]}$, $Z_n^{[2]}$, and with the numerical solution obtained one solves also the remaining system refering to the space $Z_n^{[R]}$.

More precisely, we substitute the time-derivative by the forward discretization in (3.10) and then apply (3.28) to both the terms.

We first recall the forward time-discretization:

$$\left(u^h\right)_t = \frac{\left(u^h\right)^{m+1} - \left(u^h\right)^m}{dt},$$

where the notation defined in the time integration's section has been used.

Therefore (3.10) becomes

$$\left(\frac{\left(u^h\right)^{m+1} - \left(u^h\right)^m}{dt}, \psi^h\right)_{K_{ij}} + \left\langle \widehat{a\left(u^h\right)^m}, \psi^h \right\rangle_{J_j} + \left\langle \widehat{b\left(u^h\right)^m}, \psi^h \right\rangle_{I_i} = 0, \qquad \forall \psi^h \in Z_n^0,$$

Applying (3.28) to the test function ψ^h, summing up over all elements and taking into account (3.29) we have the following system of equations

$$\begin{aligned}
&\left(\frac{\left(u^h\right)^{m+1} - \left(u^h\right)^m}{dt}, \psi_{[1]}^h\right)_{\mathcal{T}_h} + \left\langle \widehat{b\left(u^h\right)^m}, \psi_{[1]}^h \right\rangle_{\mathcal{I}_h} = 0 \\
&\left(\frac{\left(u^h\right)^{m+1} - \left(u^h\right)^m}{dt}, \psi_{[2]}^h\right)_{\mathcal{T}_h} + \left\langle \widehat{a\left(u^h\right)^m}, \psi_{[2]}^h \right\rangle_{\mathcal{J}_h} = 0 \\
&\left(\frac{\left(u^h\right)^{m+1} - \left(u^h\right)^m}{dt}, \psi^h\right)_{\mathcal{T}_h} + \left\langle \widehat{a\left(u^h\right)^m}, \psi_{[R]}^h \right\rangle_{\mathcal{J}_h} + \left\langle \widehat{b\left(u^h\right)^m}, \psi_{[R]}^h \right\rangle_{\mathcal{I}_h} = 0
\end{aligned} \tag{3.30}$$

where the following short hand notation has been used:

$$\begin{aligned}
\left(u^h, \psi^h\right)_{\mathcal{T}_h} &= \sum_{i,j} \int_{K_{ij}} u^h \psi^h \, dxdy, \\
\left\langle \widehat{au^h}, \psi^h \right\rangle_{\mathcal{J}_h} &= -\sum_{i,j} \int_{J_j} \left(\widehat{au^h}\right)_{i+\frac{1}{2},i} \left[\!\left[\psi^h\right]\!\right]_{i+\frac{1}{2},j} \, dy, \\
\left\langle \widehat{bu^h}, \psi^h \right\rangle_{\mathcal{I}_h} &= -\sum_{i,j} \int_{I_i} \left(\widehat{bu^h}\right)_{i,j+\frac{1}{2}} \left[\!\left[\psi^h\right]\!\right]_{i,j+\frac{1}{2}} \, dx.
\end{aligned}$$

We can also apply (3.28) to our numerical solution:

$$u^h = u_{[1]}^h + u_{[2]}^h + u_{[R]}^h.$$

This yields

$$\begin{aligned}
&\left(\frac{\left(u^h_{[1]}\right)^{m+1}+\left(u^h_{[2]}\right)^{m+1}-\left(u^h_{[1]}\right)^{m}-\left(u^h_{[2]}\right)^{m}}{dt},\psi^h_{[1]}\right)_{\mathcal{T}_h}+\left\langle b\widehat{\left(u^h_{[1]}\right)^m}+b\widehat{\left(u^h_{[2]}\right)^m},\psi^h_{[1]}\right\rangle_{\mathcal{I}_h}\\
&\qquad\qquad=-\left\langle b\widehat{\left(u^h_{[R]}\right)^m},\psi^h_{[1]}\right\rangle_{\mathcal{I}_h}\\
&\left(\frac{\left(u^h_{[1]}\right)^{m+1}+\left(u^h_{[2]}\right)^{m+1}-\left(u^h_{[1]}\right)^{m}-\left(u^h_{[2]}\right)^{m}}{dt},\psi^h_{[2]}\right)_{\mathcal{T}_h}+\left\langle a\widehat{\left(u^h_{[1]}\right)^m}+a\widehat{\left(u^h_{[2]}\right)^m},\psi^h_{[2]}\right\rangle_{\mathcal{J}_h}\\
&\qquad\qquad=-\left\langle a\widehat{\left(u^h_{[R]}\right)^m},\psi^h_{[2]}\right\rangle_{\mathcal{J}_h},
\end{aligned}$$

Observe that there are terms that simplify, more precisely

$$\left\langle b\widehat{\left(u^h_{[2]}\right)^m},\psi^h_{[1]}\right\rangle_{\mathcal{I}_h}=0\qquad\text{and}\qquad\left\langle a\widehat{\left(u^h_{[1]}\right)^m},\psi^h_{[2]}\right\rangle_{\mathcal{J}_h}=0.$$

We show now why these two equalities hold.
For the first-one we have

$$\begin{aligned}
\left\langle\widehat{bu^h_{[2]}},\psi^h_{[1]}\right\rangle_{\mathcal{I}_h}&=-\sum_{i,j}\int_{I_i}\left(\widehat{bu^h_{[2]}}\right)_{i,j+\frac12}\left[\!\left[\psi^h\right]\!\right]_{i,j+\frac12}\,dx\\
&=\sum_{i,j}\int_{I_i}\left[\left(\widehat{bu^h_{[2]}}\right)_{i,j+\frac12}\left(\psi^h_{[1]}\right)^-_{i,j+\frac12}-\left(\widehat{bu^h_{[2]}}\right)_{i,j-\frac12}\left(\psi^h_{[1]}\right)^+_{i,j-\frac12}\right]dx\\
&=\sum_{i,j}\int_{I_i}\underbrace{\left[\left(\widehat{bu^h_{[2]}}\right)_{i,j+\frac12}-\left(\widehat{bu^h_{[2]}}\right)_{i,j-\frac12}\right]}_{=0}\left(\psi^h_{[1]}\right)_{i,j}dx=0
\end{aligned}$$

The same argument applies also for the second term.

Considering this latter simplification, in order to complete the time step, we derive $\left(u^h_{[R]}\right)^{m+1}$ by using the already computed $\left(u^h_{[1]}\right)^{m+1}$ and $\left(u^h_{[2]}\right)^{m+1}$.

At the end, the system that we have to solve is

$$\begin{cases} \left(\frac{\left(u^h_{[1]}\right)^{m+1}+\left(u^h_{[2]}\right)^{m+1}-\left(u^h_{[1]}\right)^{m}-\left(u^h_{[2]}\right)^{m}}{dt}, \psi^h_{[1]}\right)_{\mathcal{T}_h} + \left\langle b\,\widehat{\left(u^h_{[1]}\right)}^{m}, \psi^h_{[1]}\right\rangle_{\mathcal{I}_h} = -\left\langle b\,\widehat{\left(u^h_{[R]}\right)}^{m}, \psi^h_{[1]}\right\rangle_{\mathcal{I}_h} \\ \left(\frac{\left(u^h_{[1]}\right)^{m+1}+\left(u^h_{[2]}\right)^{m+1}-\left(u^h_{[1]}\right)^{m}-\left(u^h_{[2]}\right)^{m}}{dt}, \psi^h_{[2]}\right)_{\mathcal{T}_h} + \left\langle a\,\widehat{\left(u^h_{[2]}\right)}^{m}, \psi^h_{[2]}\right\rangle_{\mathcal{J}_h} = -\left\langle a\,\widehat{\left(u^h_{[R]}\right)}^{m}, \psi^h_{[2]}\right\rangle_{\mathcal{J}_h} \\ \left(\frac{\left(u^h_{[R]}\right)^{m+1}-\left(u^h_{[R]}\right)^{m}}{dt}, \psi^h_{[R]}\right)_{\mathcal{T}_h} + \left\langle a\,\widehat{\left(u^h_{[R]}\right)}^{m}, \psi^h_{[R]}\right\rangle_{\mathcal{J}_h} + \left\langle b\,\widehat{\left(u^h_{[R]}\right)}^{m}, \psi^h_{[R]}\right\rangle_{\mathcal{I}_h} \\ \qquad = -\left\langle a\,\widehat{\left(u^h_{[2]}\right)}^{m+1}, \psi^h_{[R]}\right\rangle_{\mathcal{J}_h} - \left\langle b\,\widehat{\left(u^h_{[1]}\right)}^{m+1}, \psi^h_{[R]}\right\rangle_{\mathcal{I}_h} \end{cases} \tag{3.31}$$

Observe that in the last equality we have dropped the time derivative of $u^h_{[1]}$ and $u^h_{[2]}$.

3.5.2 Numerical experiments

In order to check the convergence we performed the same test done for the usual sparse space.

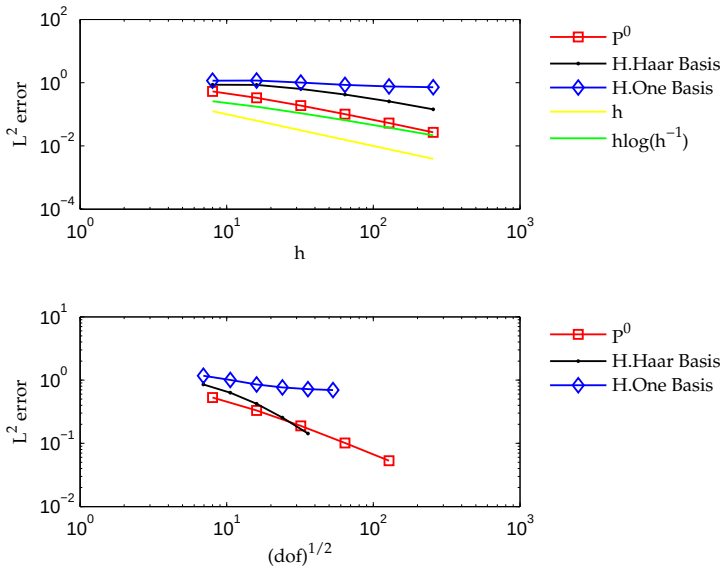
Convergence for constant coefficients

We consider the same two-dimensional transport equation with constant coefficients described in (3.24) and we obtain the following relative errors.

N		8	16	32	64	128	256
Whole space	L^2*-error*	0.53033	0.33178	0.18870	0.10120	0.05320	0.026949
	rates	0.6767	0.8141	0.8989	0.9278	0.9812	
H.Haar Basis	L^2*-error*	0.85731	0.85063	0.63137	0.42160	0.25521	0.14331
	rates	0.0113	0.4301	0.5826	0.7242	0.8325	
H.One Basis	L^2*-error*	1.16168	1.17335	1.00644	0.85008	0.76333	0.71755
	rates	-0.0144	0.2214	0.2436	0.1553	0.0892	
Degrees of freedom	*Whole space*	64	256	1024	4096	16384	65536
	Sparse grid	20	48	112	256	576	1280

We note that for the transport equation with constant coefficients the splitting presented in this section can be used only taking the hierarchical *Haar* basis. In this case the errors are almost the same as when we consider the usual sparse space. By using the hierarchical *One* basis the numerical solution does not converge (probably because of the fact that we have dropped the time derivative of $u^h_{[1]}$ and $u^h_{[2]}$ in the last equality of (3.31)).
We can see it clearly in figure 3.16.

Figure 3.16: Convergence in the alternative sparse space.



Convergence for variable coefficients

Finally, we consider the forced transport equation. We only use hierarchical *Haar* basis here because the method based on the hierarchical *One* basis already failed in the case of constant coefficient. The obtained relative errors are given in the following table.
Notice that they are the same as the one obtained using the usual sparse grid method.

N		8	16	32	64	128
Sparse Alternative	L^2-*error*	0.48990	0.32978	0.21235	0.12962	0.07438
	rates	0.5710	0.6351	0.7121	0.8013	

Chapter 4

The Vlasov-Poisson system

4.1 Motivation

The Vlasov-Poisson system is one of the basic models used to describe large ensambles of interacting particles, for this reason it has relevant applications in plasma-physics.
In general, a plasma is a gas of (partially or totally) ionized particles. In fact, by heating up a gas in a closed domain, some molecules can loose their electrons and this phenomenon produces charged particles (positive ions or negative electrons).
The Vlasov-Poisson system describes the evolution of a collisionless plasma of charged particles, that is the description of the evolution of the distribution of the particles when the only interactions considered important are the, so-called, *Coulomb interactions* that depend on the electrostatic field.

We start by describing the one-dimensional Vlasov-Poisson system of equations for which we mainly refer to [5] and [14].

4.2 The continuous problem

Let $f \in L^1_{loc}(\Omega_x \times \mathbb{R})$ to denote the *distribution function* of charged particles in a plasma. This function depends on

- ◇ the *position* $x \in \Omega_x \subset \mathbb{R}$,
- ◇ the *velocity* $v \in \mathbb{R}$,
- ◇ the *time* $t \in [0, \infty)$.

We consider the one-dimensional Vlasov equation

$$\begin{aligned} f_t + v f_x - \Phi_x(x) f_v = 0, \qquad & \forall (x,v) \in \Omega_x \times \mathbb{R} \text{ and } t \in [0,\infty), \\ f(x,v,0) = f_0(x,v), \qquad & \forall (x,v) \in \Omega_x \times \mathbb{R}. \end{aligned} \tag{4.1}$$

The electrostatic field, $E(x,t) := \Phi_x(x,t)$, is derived from a *potential* Φ that satisfies the Poisson equation

$$-\Phi_{xx}(x,t) = -E_x(x,t) = \rho(x,t) - 1, \qquad (x,v) \in \Omega_x \times [0,\infty), \tag{4.2}$$

in which $\rho(x,t)$ is the *charge density* defined by

$$\rho(x,t) = \int_{\mathbb{R}} f(x,v,t)\ dv \qquad \text{for all } (x,v) \in \Omega_x \times [0,\infty). \tag{4.3}$$

Since the plasma should be neutral, we have a further condition of *total charge neutrality*:

$$\int_{\Omega_x} \rho(x,t)dx = \int_{\Omega_x}\int_{\mathbb{R}} f(x,v,t)\ dvdx = 1 \qquad \text{for all } t \in [0,\infty). \tag{4.4}$$

Boundary conditions
From now on, we consider the physical domain in which the plasma is confined to be $\Omega_x = [0,1]$. We complete the system by imposing periodic boundary conditions on x both for the Vlasov- and for the Poisson-equation, i.e.

$$\begin{aligned} f(0,v,t) &= f(1,v,t), && \text{for all } (x,v) \in \Omega_x \times \mathbb{R}, t \in [0,\infty) \\ \Phi(0,t) &= \Phi(1,t) \qquad \text{and} \\ E(0,t) &= E(1,t), && \text{for all } t \geq 0, \end{aligned} \tag{4.5}$$

In the coordinate v we do not impose boundary conditions but we restrict the study to functions f which are compactly supported with respect to v. The reason for this restriction are given in the theorem that follows (given to Copper and Klimax) on the existence and uniqueness of classical solution of the periodic Vlasov-Poisson system (4.1)-(4.2)-(4.5).

To this end, for a fixed time interval $[0,T]$ for all $T > 0$, given a distribution function $f(x,v,t)$ we denote by

$$Q(t) = 1 + \sup\{|v| : \exists x \in \Omega_x \text{ and } \tau \in [0,t] \text{ s.t. } f(x,v,\tau) \neq 0\},$$

for all $t \in [0,\infty)$ as a measure of the support of the distribution function.
Then we have

Theorem 4.2.1 *(Well-posedness of the continuous 1DVP [19].)*
Given $f_0 \in C^1(\mathbb{R}_x \times \mathbb{R}_v)$, 1-periodic in x and compactly supported in v, $Q(0) \leq Q_0$ with $Q_0 > 1$. Then the periodic Vlasov-Poisson system (4.1)-(4.2) has a unique classical solution (f,E), that is 1-periodic in x for all time $t \in [0,T]$ for all $T > 0$.

So, in the rest of this work, we will assume that the initial data f_0 satisfies the hypotheses in the theorem, and thus, the unique classical solution to the periodic VP system (4.1)-(4.2) satisfies that there exists $L > 0$ that depends on f_0, T and Q_0 such that $\text{supp}(f(t)) \subseteq \Omega$ for all $t \in [0,T]$, where we have defined $\Omega = \Omega_x \times \Omega_v$, with $\Omega_v = [-L,L]$.

Remark 1 ⋄ Notice that, the one-dimensional Vlasov equation (4.1) is already a two-dimensional time-dependent equation. This implies that, in the three-dimensional case ($\Omega_x \subset \mathbb{R}^3$, $v \in \mathbb{R}^3$) we would face a six-dimensional time-dependent problem for which the application of Sparse grid may help in its numerical solution.

⋄ Observe that the third condition in (4.5) is equivalent to

$$\int_0^1 (-E_x(x,t))\, dx = 0, \qquad \forall\ t \geq 0.$$

This implies

$$\int_0^1 \left(\int_{-\infty}^{\infty} f(x,v,t)\ dv - 1 \right)\ dx = 0$$

which is exactly (4.4), i.e.,

$$\int_0^1 \int_{-\infty}^{\infty} f(x,v,t)\ dvdx = 1. \tag{4.6}$$

4.2.1 The weak formulation of the Vlasov equation

Notice that the one-dimensional Vlasov equation (4.1) can be seen as a two-dimensional transport equation (3.1) with non-linear coefficients defined as

$$a(v) = v \quad \text{and} \quad b(x) = -E(x,t).$$

Therefore, the derivation of the weak formulation is similar to the one carried out in the previous section.
We multiply (4.1) by a test function

$$\psi \in C_0^\infty(\Omega) = \{w \in C^\infty(\Omega) \text{ and } w \text{ periodic in } x \text{ and compact w.r.t } v\}$$

and then integrate over Ω.
Because of periodic boundary conditions in x and compact support with respect to v, the boundary terms disappear and the weak formulation reads: *Find f such that*

$$(f_t, \psi) - (vf, \psi_x) + (Ef, \psi_x) = 0, \qquad \forall\ \psi \in C_0^\infty(\Omega).$$

4.2.2 Properties of the system

L^2-conservation

The function f continuous solution of problem (4.1)-(4.2)-(4.5) satisfies the L^2-conservation property, i.e.

$$\frac{d}{dt}||f(\cdot,\cdot,t)||_0 = 0 \qquad \forall t > 0.$$

Proof: See the proof of the L^2-conservation for the 2D transport equation.

Mass conservation

The function f, continuous solution of the problem (4.1)-(4.2)-(4.5), conserves the total mass, i.e.

$$\frac{d}{dt}\int_\Omega f(x,y,t)\ dxdv = 0, \qquad \forall t > 0.$$

Proof: See the proof of the Mass conservation for the 2D transport equation.

Energy conservation

The **total energy** ε_{tot} of the Vlasov-Poisson system is the sum of the **kinetic** (ε_k) and **potential** (ε_p) **energy**:

$$\varepsilon_{tot}(t) = \varepsilon_k(t) + \varepsilon_p(t),$$

where

$$\varepsilon_k(t) = \frac{1}{2}\int_\Omega |v|^2 f(x,v,t)\ dxdv,$$
$$\varepsilon_p(t) = \frac{1}{2}\int_{\Omega_x} |E(x,t)|^2\ dx.$$

The solution of the Vlasov-Poisson system of equation conserves the total energy, i.e.

$$\begin{aligned}\frac{d}{dt}\varepsilon_{tot}(t) &= \frac{d}{dt}\left[\varepsilon_k(t) + \varepsilon_p(t)\right] \\ &= \frac{1}{2}\frac{d}{dt}\left[\int_\Omega |v|^2 f(x,v,t)\ dxdv + \int_{\Omega_x} |E(x,t)|^2\ dx\right] = 0.\end{aligned}$$

Proof: see [18].

4.3 The discrete Vlasov equation

4.3.1 Notation

We fix $n \in \mathbb{N}$ and divide the domain

$$\Omega = \Omega_x \times \Omega_v = [0,1] \times [-L, L],$$

in a **uniform partition**

$$\mathcal{T}_h = \mathcal{I}_h \times \mathcal{J}_h = \{K_{ij} = I_i \times J_j\}_{1 \le i,j \le 2^n}.$$

The one-dimensional partitions $\mathcal{I}_h$ and $\mathcal{J}_h$ are the sets of all intervals

$$I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \qquad J_i = [v_{j-\frac{1}{2}}, v_{j+\frac{1}{2}}] \quad \text{respectively,}$$

and since they are uniform we have

$$h_x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} \qquad h_v = v_{j+\frac{1}{2}} - v_{j-\frac{1}{2}} \qquad \text{and } h = \min(h_x, h_v) \qquad \text{for all } 1 \le i,j \le 2^n.$$

We use the same notation defined in (3.5) for the value taken at cell interfaces and the same definition of the trace operators (3.4). Also the **finite element space** we consider is the one used before, i.e.

$$Z_n^0 = \{z \in L^2(\Omega)\ :\ z \in \mathbb{Q}^0(K_{ij}) = \mathbb{P}^0(I_i) \otimes \mathbb{P}^0(J_j),\ \ 1 \le i,j \le 2^n\},$$

4.3.2 The Discontinuous Galerkin formulation

We let $f^h(0) = \mathcal{P}_h(f_0)$ and since the derivation of the method is close to the one done for the two-dimensional transport equation we only write the resulting DG-method which turns out to be:
find $f^h : [0, T_{end}] \to Z_n^0$ *such that*

$$\sum_{ij} \left[\left(\left(f^h\right)_t, \psi^h \right)_{K_{ij}} + \left\langle \widehat{vf^h}, \psi^h \right\rangle_{J_j} - \left\langle \widehat{E^h f^h}, \psi^h \right\rangle_{I_i} \right] = 0, \qquad \forall \psi^h \in Z_n^0, \tag{4.7}$$

where, in order to ensure stability, the numerical fluxes are defined as

$$\left(\widehat{vf^h}\right)_{i+\frac{1}{2}} = \begin{cases} vf^h(x^-_{i+\frac{1}{2}}, v) & \text{if } v \geq 0 \\ vf^h(x^+_{i+\frac{1}{2}}, v) & \text{if } v < 0 \end{cases} \quad \left(\widehat{E^h f^h}\right)_{j+\frac{1}{2}} = \begin{cases} E^h(x) f^h(x, v^-_{j+\frac{1}{2}}) & \text{if } E^h(x) \leq 0 \\ E^h(x) f^h(x, v^+_{j+\frac{1}{2}}) & \text{if } E^h(x) > 0. \end{cases} \tag{4.8}$$

Notice that the numerical flux defined on the horizontal edges depends on the sign of $E^h(x)$ which is the approximation of the electrostatic field, obtained by solving the Poisson equation at each time-step. We will see in the next section that we use linear polynomials for the approximation of the electrostatic field, so it can happen that E^h changes of sign in an element. We solve this issue by projecting the approximation into the space of constant functions: $P^0(E^h)$. Notice that this is nothing but the average of the function on the observed element.

4.3.3 L^2-stability

Let $f^h : [0, T_{end}] \to Z_n^0$ be the approximation (4.7) of problem (4.1), with the numerical fluxes as in (4.8). Then

$$||f^h(t)||_{0,\mathcal{T}_h} \leq ||f^h(0)||_{0,\mathcal{T}_h} \qquad \forall\, t \in [0, T_{end}].$$

Proof: The proof follows exactly the same steps as for the two-dimensional transport equation, except for the fact that here we have compact support in v instead periodic boundary conditions.

4.3.4 Mass conservation

Let $f^h : [0, T_{end}] \to Z_n^0$ be the numerical solution of (4.7) then the following equalities hold

$$\sum_{i,j} \int_{K_{ij}} f^h(t)\, dxdv = \sum_{i,j} \int_{K_{ij}} f^h(0)\, dxdv = \sum_{i,j} \int_{K_{ij}} f_0\, dxdv = 1, \qquad \forall t \in [0, T_{end}]. \tag{4.9}$$

Proof: Here also we remind to the section in the two-dimensional transport equation for the proof of the first two equalities. The last equalitity follows directly from (4.6).

4.4 The Poisson equation

In this section we consider the one-dimensional Poisson equation (4.2) with the periodic boundary conditions defined by (4.5)

$$\begin{aligned} -\Phi_{xx}(x,t) &= \rho(x,t) - 1, && \text{for } x \in \Omega_x \text{ and a fixed } t, \\ \Phi(0,t) &= \Phi(1,t) && \text{and} \\ \Phi_x(0,t) &= \Phi_x(1,t), && \text{for all } t \geq 0. \end{aligned} \tag{4.10}$$

We recall that from the Vlasov-Poisson equation we have

$$\rho(x,t) = \int_{\mathbb{R}} f(x,v,t)\ dv. \tag{4.11}$$

Remark 2

- ◇ In order to ensure uniqueness of the solution, one has to fix the value of Φ at a point

$$\Phi(0,t) = 0 \qquad \forall\ t \geq 0.$$

- ◇ Observe that what is really needed in the Vlasov-Poisson system is not the approximation of the potential Φ but the approximation of its derivative Φ_x $(= E(x,t))$.
  For this reason, the use of a mixed method (such as the Local Discontinuous Galerkin method) is particularly suitable [8], because one approximates not only Φ^h but also E^h.
- ◇ The variable t is a fixed value here. For this reason, in this section, we will drop it, in order to keep the expressions clearer.

Considering these remarks and the fact that $\Omega_x = [0,1]$, we rewrite (4.10) as the following system of ordinary differential equations

$$E(x) = \Phi_x(x), \qquad \text{for } x \in \Omega_x, \tag{4.12}$$

$$-E_x(x) = \rho(x) - 1, \qquad \text{for } x \in \Omega_x, \tag{4.13}$$

$$\Phi(0) = \Phi(1) \text{ and } E(0) = E(1). \tag{4.14}$$

4.4.1 The weak formulation of the Poisson equation

In order to obtain the weak formulation we multiply (4.12) and (4.13) by test functions p and $w \in C^\infty_{\text{per}}(\Omega_x)$ and integrate both equations over Ω_x

$$\int_{\Omega_x} Ep\ dx = \int_{\Omega_x} \Phi_x p\ dx,$$

$$-\int_{\Omega_x} E_x w\ dx = \int_{\Omega_x} (\rho - 1) w\ dx.$$

Integration by parts yields

$$\int_{\Omega_x} Ep\ dx = -\int_{\Omega_x} \Phi p_x\ dx + [\Phi(1)p(1) - \Phi(0)p(0)],$$

$$\int_{\Omega_x} Ew_x\ dx - [E(1)w(1) - E(0)w(0)] = \int_{\Omega_x} (\rho - 1) w\ dx.$$

Because of (4.14) and since $p, w \in C^\infty_{\text{per}}(\Omega_x)$, the boundary terms disappear.
Therefore the weak formulation of the system reads: *Find* (E, Φ) *such that*

$$\begin{aligned} \int_{\Omega_x} Ep \; dx &= -\int_{\Omega_x} \Phi p_x \; dx, \\ \int_{\Omega_x} Ew_x \; dx &= \int_{\Omega_x} (\rho - 1) w \; dx, \qquad \textit{for all } p, w \in C^\infty_{\text{per}}(\Omega_x). \end{aligned}$$

4.4.2 The LDG-formulation

First we need to define the partition and the finite element space.
We use the same **partition** $\mathcal{I}_h$ defined for the discretization of the one-dimensional transport equation.
Thus we fix $N > 0$ and consider uniform intervals $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ for $i = 1, \dots, N$ with $h = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$.

The **finite element space** is the following

$$V_h^1(\mathcal{I}_h) = \{w \in L^2(\Omega_x) \mid w \in \mathbb{P}^1(I_i), \text{ for all } i = 1, \dots, N\},$$

in which we consider the L^2-functions which are linear polynomals on any interval of the partition.

In order to obtain the LDG-formulation we substitute E, Φ, p, and w by E^h, Φ^h, p^h, and w^h in the weak formulation defined on an interval I_i.
Reorganizing the terms one obtains

$$\begin{aligned} \int_{I_i} E^h p^h \; dx + \int_{I_i} \Phi^h \left(p^h\right)_x \; dx - \left[\left(\widehat{\Phi^h} p^h\right)_{i+\frac{1}{2}} - \left(\widehat{\Phi^h} p^h\right)_{i-\frac{1}{2}}\right] = 0, \\ + \int_{I_i} E^h \left(w^h\right)_x \; dx - \left[\left(\widehat{E^h} w^h\right)_{i+\frac{1}{2}} - \left(\widehat{E^h} w^h\right)_{i-\frac{1}{2}}\right] = \int_{I_i} (\rho^h - 1) w \; dx, \qquad \forall\, p^h, w^h \in V_h^1, \end{aligned} \tag{4.15}$$

where the discrete density ρ^h is given by

$$\rho^h(x,t) = \sum_j \int_{J_j} f^h(x, v, t) dv \qquad \forall\, x \in \mathcal{I}_h, \;\; \forall t \in [0, T]. \tag{4.16}$$

Here, again, the numerical fluxes appear.
Recalling (3.4) and following [8] we define them as follows

$$\begin{aligned} \widehat{E^h} &= \left\{E^h\right\} - C_{12} \left[\!\left[E^h\right]\!\right] + C_{11} \left[\!\left[\Phi^h\right]\!\right], \\ \widehat{\Phi^h} &= \left\{\Phi^h\right\} + C_{12} \left[\!\left[\Phi^h\right]\!\right]. \end{aligned}$$

For the boundary nodes, due to periodicity we enforce

$$\left(\widehat{E^h}\right)_{\frac{1}{2}} = \left(\widehat{E^h}\right)_{N+\frac{1}{2}} \qquad \text{and} \qquad \left(\widehat{\Phi^h}\right)_{\frac{1}{2}} = \left(\widehat{\Phi^h}\right)_{N+\frac{1}{2}}.$$

Finally, we sum up over all elements and rewrite (4.15) in a short hand format

$$\begin{aligned} \mathcal{A}\left(E^h, p^h\right)_{\mathcal{I}_h} + \mathcal{B}\left(\Phi^h, p^h\right)_{\mathcal{I}_h} = 0, \\ \mathcal{C}\left(E^h, w^h\right)_{\mathcal{I}_h} = \mathcal{A}\left(\rho^h - 1, w^h\right)_{\mathcal{I}_h}, \qquad \forall\, p^h, w^h \in V_h^1, \end{aligned} \tag{4.17}$$

where the bilinear form $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are defined as follows:

$$\mathcal{A}(E,p)_{\mathcal{I}_h} = \sum_i \int_{I_i} Ep\ dx,$$
$$\mathcal{B}(\Phi,p)_{\mathcal{I}_h} = \sum_i \int_{I_i} \Phi p_x\ dx + \sum_i \widehat{\Phi}_{i+\frac{1}{2}} [\![p]\!]_{i+\frac{1}{2}}$$
$$\mathcal{C}(E,w)_{\mathcal{I}_h} = \sum_i \int_{I_i} Ew_x\ dx + \sum_i \widehat{E}_{i+\frac{1}{2}} [\![w]\!]_{i+\frac{1}{2}}$$

4.5 Error analysis for VP-system

Theorem 4.5.1 A-priori Error Estimates
Let $f \in C^1((0,T];H^1(\Omega)\cap C_0^0(\Omega))$ with $f_v \in L^\infty(\Omega)$ be the solution at time $t\in[0,T]$ of the Vlasov-Poisson system (4.1)-(4.2) and let $E \in C^0([0,T];H^1(\mathcal{I}_h))$ be the associated electrostatic potential. Let $f^h \in C^1([0,T];Z_h^0)$ be the DG approximation to f, solution of (4.7)-(4.8).
Let $(E^h,\Phi^h) \in C^0([0,T];V_h^1(\mathcal{I}_h)\times V_h^1(\mathcal{I}_h))$ be the LDG approximation to (4.10) solution of (4.15) with $c_{12}=1/2$ and $c_{11}=ch^{-1}$. Then, the following error estimate holds for all $t\in[0,T]$

$$\|f(t)-f^h(t)\|_{0,\emptyset} \le C_{err}h^{1/2}\ , \tag{4.18}$$

where C_{err} depends on the final time t, the shape regularity of the partition and depends also on f (and therefore on f_0) through the norms

$$C_{err} = C_{err}(\|f(t)\|_{1,\Omega}^2,\ \|f_t(t)\|_{1,\Omega},\ \|f(t)\|_{1,\Omega}^2\|f_v(t)\|_{0,\infty,\Omega},\ (\|E\|_{1,\Omega_x}^2+\|\Phi\|_{2,\Omega_x}^2),\ e^{\|f_v(t)\|_{0,\infty,\Omega}})\ .$$

Proof: The proof can be found in [20].

Corollary 1 In the same hypothesis of Theorem 4.5.1, the following error estimates hold

$$\|E-E^h\|_{0,\mathcal{I}}^2 + c_{11}\|[\![\Phi_h]\!]\|_{0,\gamma_x}^2 \le C_{err}^2h + Ch^2(\|E\|_{1,\Omega_x}^2+\|\Phi\|_{2,\Omega_x}^2),$$

where C_{err} is the constant of Theorem 4.5.1.

4.6 Numerical results

We consider the following forced Vlasov-Poisson system in $\Omega = [-4,4] \times [-5,5]$:

$$f_t(x,v,t) + v f_x(x,v,t) - E(x,t) f_v(x,v,t) = \xi(x,v,t) \qquad (x,v) \in \Omega, \tag{4.19}$$

$$-E_x(x,t) = \rho(x,t) - \sqrt{\pi} \qquad x \in [-4,4], \tag{4.20}$$

with periodic boundary condition in x.
The function ξ, defined as

$$\xi(x,v,t) = \sin\left(\frac{\pi}{2}x + 2\pi t\right) e^{-4v^2} \left[2\pi + \frac{v\pi}{2} + \frac{8v}{\sqrt{\pi}}\left(2 - \cos\left(\frac{\pi}{2}x + 2\pi t\right)\right)\right],$$

is chosen so that the exact solution (f, E) of the problem (4.19)-(4.20) is given by

$$f(x,v,t) = \left(2 - \cos\left(\frac{\pi}{2}x + 2\pi t\right)\right) e^{-4v^2} \qquad (x,v) \in \Omega,$$

$$E(x,t) = \frac{1}{\sqrt{\pi}} \sin\left(\frac{\pi}{2}x + 2\pi t\right) \qquad x \in [-4,4].$$

Observe that this definition of f implies $f(x,v,t)|_{\Omega_v} \simeq 0$ (compact support v).

4.6.1 Convergence

We now study the convergence of the method for different mesh-sizes. We divide each coordinate in N subintervals for $N = 16$, 32, 64, 128 and 256.
Since the exact solution (f, E) is 1-periodic in time we performed the computations up to time $T_{end} = 1$. As in the case of transport equations the errors are computed using the discrete L^2-norms.
In the following table the relative errors for f and E are given. For f the errors are computed using the L^2-norm, for E we use the L^2-norm and the Q-seminorm.
This latter is defined (see [5]) as

$$|(E,\Phi)|^2_{Q,\mathcal{T}_h} := ||E||_{0,\mathcal{I}_h} + \sum_i c_{11} |[\![\Phi]\!]_{i+\frac{1}{2}}|^2,$$

where Φ is the solution of the Poisson problem (4.20) and $c_{11} = 2h^{-1}$ is the positive constant which comes from the LDG method used to solve it.

N		16	32	64	128	256
Whole space	*L^2-error for f*	0.2546935	0.2204524	0.1375240	0.0773834	0.0412853
	rates	0.2083	0.6808	0.8296	0.9064	
	L^2-error for E	0.1898788	0.1065016	0.0558869	0.0286738	0.0145184
	rates	0.8342	0.9303	0.9628	0.9818	
	Q-error for E	0.2039074	0.1098290	0.05671024	0.0288789	0.0145697
	rates	0.8371	0.9397	0.9701	0.9870	
Sparse grid	*L^2-error for f*	0.4100648	0.3853896	0.2960314	0.1850251	0.1077873
	rates	0.0895	0.3806	0.6780	0.7795	
	L^2-error for E	0.1165185	0.2508481	0.1913409	0.0841518	0.0420354
	rates	-1.1063	0.3907	1.1851	1.001	
	Q-error for E	0.1422369	0.2530979	0.1916933	0.0842347	0.0420542
	rates	-0.8314	0.4009	1.1863	1.002	

In the next figures the obtained results are displayed. First, the relative errors for the distribution function f are given (4.1), then (4.2) and (4.3) are the ones concerning the electrostatic field E, in the L^2-norm and Q-seminorm respectively. For all cases, in the figure above, the number of intervals N is the reference on the x-axis, while in the figure below the reference is the square-root of the degrees of freedom.
In all diagrams the red lines (with squares) refer to the relative errors when the whole space is considered, while the pink lines (with circles) refers to the ones obtained with the sparse grid. The dashed black lines refer to the order $h^{1/2}$, the dotted blue lines to the order $h\log(h^{-1})$ and the dotted green ones to the order h.

Figure 4.1: Convergence: whole vs. sparse space. Error: $\frac{||f-f^h||_{0,\mathcal{T}_h}}{||f||_{0,\mathcal{T}_h}}$

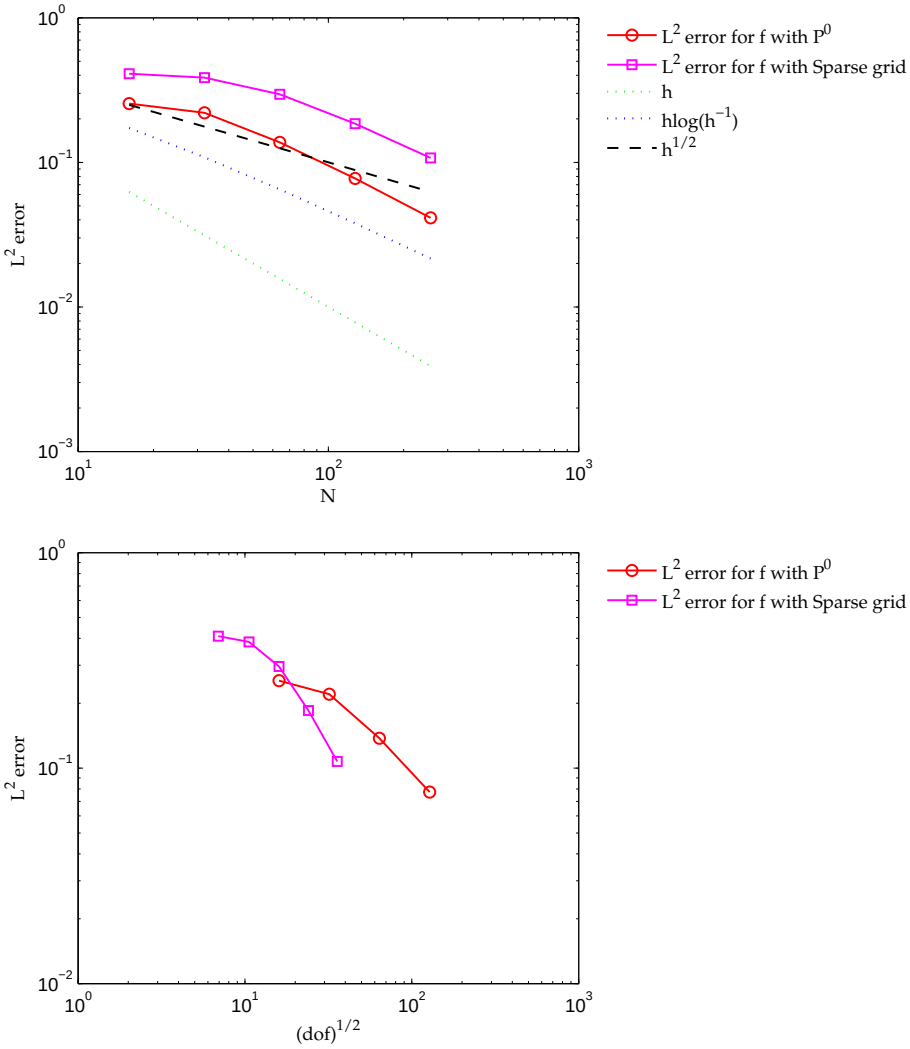


Figure 4.2: Convergence: whole vs. sparse space. Error: $\frac{||E-E^h||_{0,\mathcal{T}_h}}{||E||_{0,\mathcal{T}_h}}$

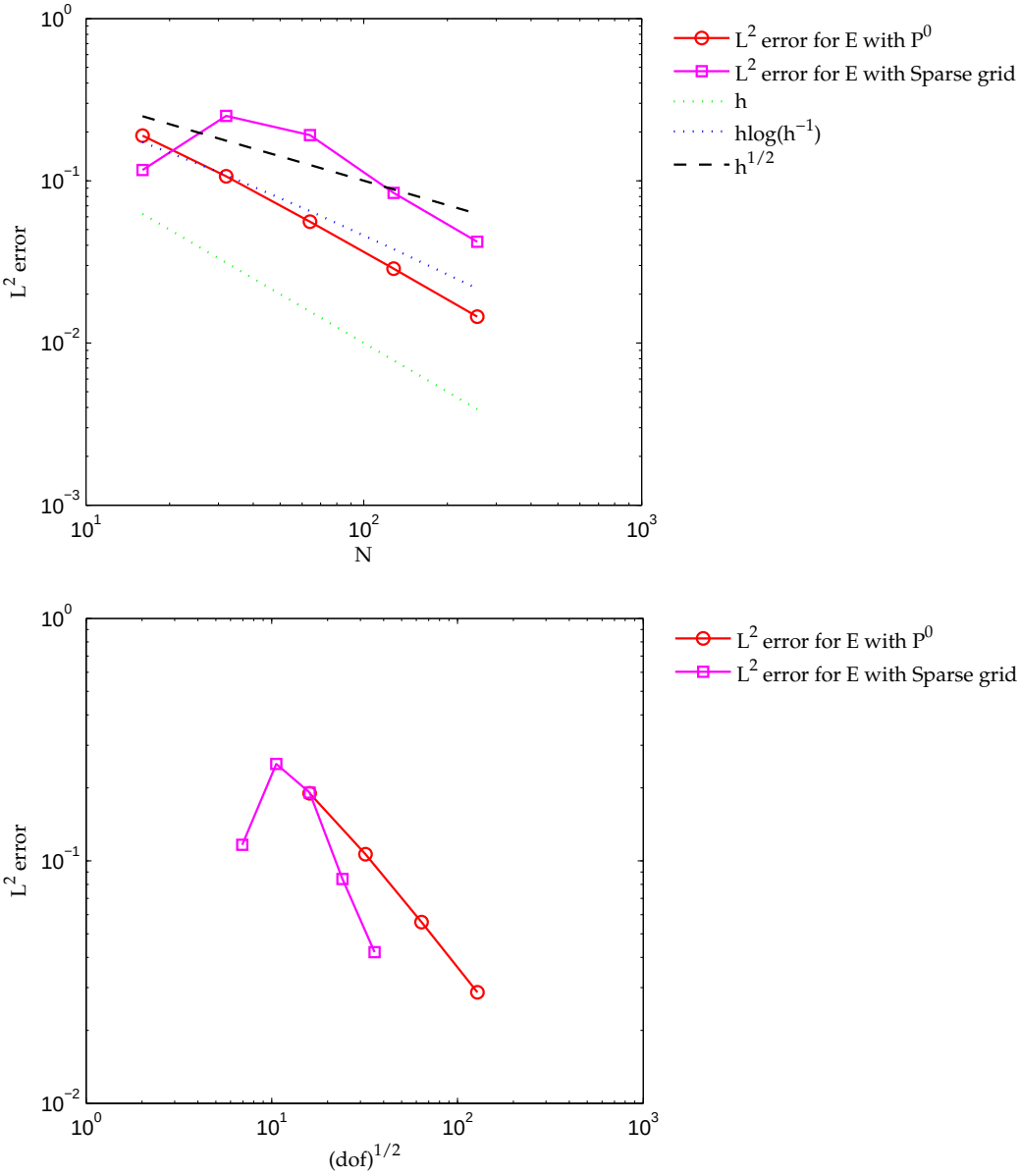
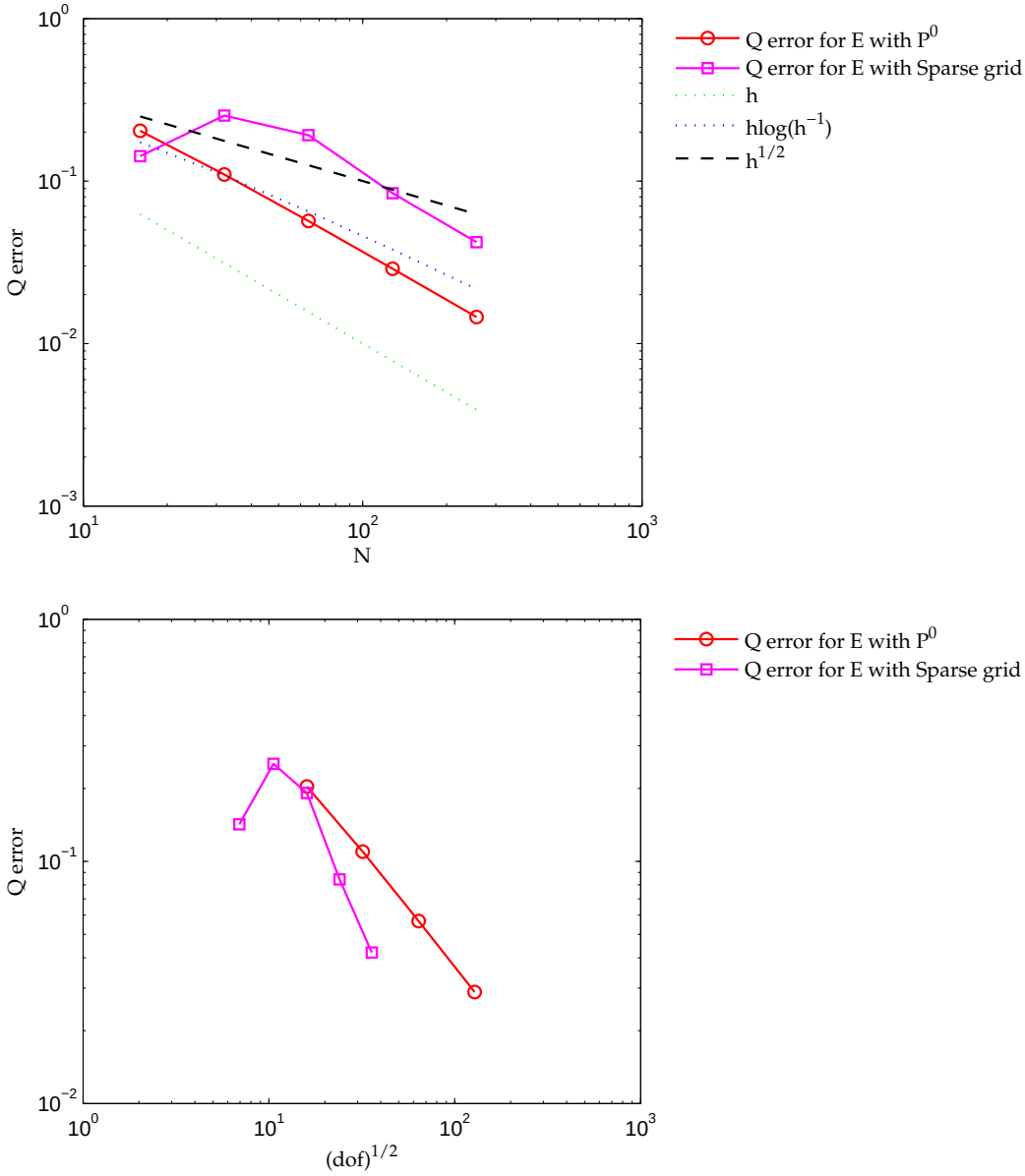


Figure 4.3: Convergence: whole vs. sparse space. Error: $\frac{|(E-E^h,\Phi-\Phi^h)|_{Q,\mathcal{T}_h}}{|(E,\Phi)|_{Q,\mathcal{T}_h}}$

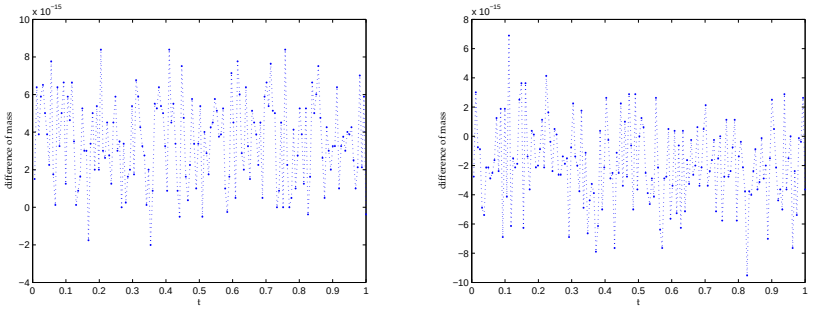


Note that the first mesh-size we consider here is the one that divides each coordinate in 16 intervals. This implies $h_x = 0.5$ which is a large mesh-size. This is probably the reason why the errors concerning the numerical solution of the electrostatic field E^h behaves in a strange way in this case.

4.6.2 Mass conservation

We study the conservation of the total mass also for the Vlasov-Poisson system. Using the same data as in the convergence analysis in a uniform grid 64×64, we observe that also in this case the magnitude of the difference of mass between the initial condition and any future time is of the order of 10^{-15} (see figure 4.4), which indicates that we have mass conservation.

Figure 4.4: Conservation of mass: whole space vs. sparse grid.



4.6.3 Energy conservation

We study the conservation of the discrete total energy of the Vlasov-Poisson system.
First let us introduce the discrete energies that arise at this stage of the analysis.
We have

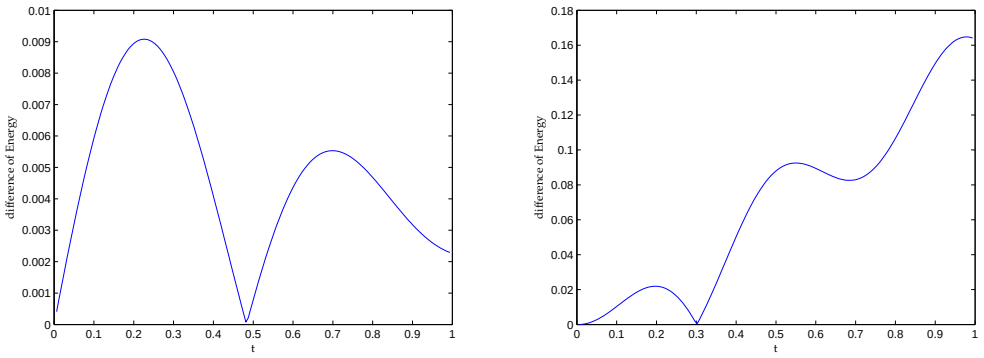
- discrete kinetic energy $\quad \varepsilon_k^d(t) = \frac{1}{2}\sum_{i,j}\int_{K_{ij}} |v|^2 f^h(x,v,t)\, dxdv,$
- discrete potential energy $\quad \varepsilon_p^d(t) = \frac{1}{2}\sum_{i}\int_{I_i} |E^h(x,t)|^2\, dx + c_{11}\sum_i \left[\!\left[\Phi^h(x,t)\right]\!\right]^2_{i+\frac{1}{2}},$
- discrete total energy $\quad \varepsilon_{tot}^d(t) = \varepsilon_k^d(t) + \varepsilon_p^d(t).$

We observe the difference between the initial total energy $\varepsilon_{tot}(0)$ and the discrete total energy computed at any following time-step t

$$\Delta_{\varepsilon_{tot}^d}(t) = \frac{|\varepsilon_{tot}^d(0) - \varepsilon_{tot}^d(t)|}{\varepsilon_{tot}^d(0)}.$$

These differences are shown in figure (4.5), left for the whole space, right for the sparse grid.

Figure 4.5: Energy conservation in a 64×64 grid.



We observe that the orders of magnitude are 10^{-2} for the whole space and 10^{-1} for the sparse grid. Thus, in both cases, the energy is not conserved.

Chapter 5

Appendix - The implementation

This chapter is dedicated to the implementation of the methods presented in this work. We refer to previous chapters for the missing definitions.
Before entering into details, we want to highlight the main difficulties that we faced in the implementation. In the two-dimensional case, it has not been easy to implement an algorithm to create the needed matrices when hierarchical basis functions were used. The main issue is the fact that the supports of these functions (usually) encompass many elements K_{ij}. This causes some complications in the assembly of the needed matrices, especially in the treatment of the interelement boundaries.
We also noticed that, because of our implementation of the flux-matrices, in the case when a hierarchical basis is used, we need to choose the boundaries of the observed domain Ω to be rational. This requirement comes from the fact that for fine meshes ($h \leq 2^{-5}$) our implementation fails in selecting the element's boundaries that give a contribution to the flux-matrices when irrational boundaries are chosen.

5.1 One dimension

5.1.1 Transport equation with standard basis

We denote by $\bar{\mathbf{u}} = (\bar{u}_1, \ldots, \bar{u}_N)^T$ the vector of the coefficients with respect to the standard basis (2.10),

$$u^h(x) = \sum_i \bar{u}_i \chi_i(x),$$

then (2.9) becomes the following linear system

$$\mathbf{M^S} \left(\bar{\mathbf{u}}\right)_t + \mathbf{Flux^S}\ \bar{\mathbf{u}} = \mathbf{0}, \tag{5.1}$$

where the *mass-matrix* is $\mathbf{M^S} = h\mathbf{I}$, with $\mathbf{I}$ the $N \times N$ *identity* matrix.
The *flux-matrix* $\mathbf{Flux^S}$ depends on the value of a:

◇ if $a > 0$,

$$\mathbf{Flux^S} = \begin{bmatrix} 1 & & & -1 \\ -1 & \ddots & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}$$

◇ if $a < 0$,

$$\mathbf{Flux^S} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 1 & & & -1 \end{bmatrix}.$$

5.1.2 Transport equation with hierarchical basis

In the following we use hierarchical One basis functions $\phi_{l,i}$ for the space-discretization of our problem.
The same calculation could be carried out with $\theta_{l,i}$ instead.

Using this basis we can expand u^h as follows

$$u^h(x,t) = \sum_{l=0}^{n} \sum_{i=1}^{2^l-1} \alpha_{l,i}(t)\phi_{l,i}(x), \qquad i \text{ odd}.$$

We substitute u^h in (2.9) and so for all $k = 1 \ldots, 2^n$,

$$\begin{aligned} \int_{I_k} \left(u^h\right)_t \psi^h dx + \left[\min(a,0)(u^h_{k+1} - u^h_k) + \max(a,0)(u^h_k - u^h_{k-1})\right] \psi^h = \\ = \sum_l \sum_i \left(\alpha_{l,i}(t)\right)_t \int_{I_k} \phi_{l,k}(x_k)\psi^h(x_k)\, dx \\ + \sum_l \sum_i \min(a,0)\alpha_{l,i}(t) \left[\phi_{l,i}(x_{k+1}) - \phi_{l,i}(x_k)\right] \psi^h(x_k) \\ + \sum_l \sum_i \max(a,0)\alpha_{l,i}(t) \left[\phi_{l,i}(x_k) - \phi_{l,i}(x_{k-1})\right] \psi^h(x_k) = 0, \end{aligned}$$

for all $\psi^h \in \bigoplus_{l \le n} W_l$.

We obtain, again, a system of equations:

$$\mathbf{M^H} \left(\bar{\mathbf{a}}\right)_t + \mathbf{Flux^H}\ \bar{\mathbf{a}} = \mathbf{0},$$

where

◇ $\mathbf{M^H}$ is the *mass-matrix*.

◇ $\mathbf{Flux^H}$ is the *flux-matrix*.

◇ $\bar{\mathbf{a}} = (\alpha_{0,1}, \ldots, \alpha_{n,2^n-1})^T$ is the *vector of the coefficients*.

5.1.3 Time integration

By taking into account the results of the previous section and using the explicit Euler method we have, first

$$\bar{\mathbf{w}}^1 = \bar{\mathbf{w}}^0 - dt \ \left[\mathbf{M^{-1}Flux}\right] \ \bar{\mathbf{w}}^0,$$

and, for any $m > 0$,

$$\bar{\mathbf{w}}^{m+1} = \bar{\mathbf{w}}^m - dt \ \left[\mathbf{M^{-1}Flux}\right] \ \bar{\mathbf{w}}^m,$$

which can be rewritten as

$$\mathbf{M} \ \bar{\mathbf{w}}^{m+1} = \left[\mathbf{M} - dt \ \mathbf{Flux}\right] \bar{\mathbf{w}}^m,$$

where $(\mathbf{M}, \mathbf{Flux}, \bar{\mathbf{w}})$ is equal to, either $(\mathbf{M^S}, \mathbf{Flux^S}, \bar{\mathbf{u}})$ or $(\mathbf{M^H}, \mathbf{Flux^H}, \bar{\mathbf{a}})$ depending on which basis is used.

Remark 3 ◇ Notice that $\bar{\mathbf{w}}^0$ are the coefficients of the given initial data.

◇ Obsere that the method requires the matrix $\mathbf{M}$ to be inverted. When standard or hierarchical Haar basis functions are used, this won't be an issue since $\mathbf{M}$ is a diagonal matrix. On the other hand when we use hierarchical One basis functions $\mathbf{M}$ is not diagonal anymore but still symmetric and positive definite. For these cases a *Cholesky*-decomposition is used:

$$\mathbf{M} = \mathbf{R}\mathbf{R^T},$$

this computation is done once for all at the beginning.
Then, at each time step, we have to solve two systems

$$\begin{aligned} \mathbf{R} \ \mathbf{z} &= \left[\mathbf{M} - dt \ \mathbf{Flux}\right] \bar{\mathbf{w}}^n, \\ \mathbf{R^T} \ \bar{\mathbf{w}}^{n+1} &= \mathbf{z}. \end{aligned}$$

5.1.4 The LDG-implementation

The basis functions

Since E^h and $\Phi^h \in \mathbb{P}^1(I_i) \ \forall i$, we use first order Lagrange basis functions on each element.
On the reference element $\bar{I} = [0, 1]$ these are defined as

$$\begin{aligned} &\bar{\varphi}_1(\bar{x}) = 1 - \bar{x}, \\ &\bar{\varphi}_2(\bar{x}) = \bar{x}, \qquad\qquad \text{for all } \bar{x} \in [0, 1]. \end{aligned}$$

We can define the basis functions for an arbitrary element I_i using the affine map

$$x(\bar{x}) := x_{i-\frac{1}{2}} + h_x \bar{x},$$

where $x(\bar{x}) \in I_i$, $\bar{x} \in \bar{I}$ and $h = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$.

Then, for $k = 1, 2$, the basis functions on the element I_i will look like

$$\varphi_k^i(x(\bar{x})) = \begin{cases} \bar{\varphi}_k(\bar{x}), & \text{if } \bar{x} \in \bar{I}, \\ 0, & \text{else.} \end{cases} \tag{5.2}$$

We show the basis function in the element I_i in figure 5.1, while in figure 5.2 the basis functions for $N = 8$ are shown.

Figure 5.1: Basis functions on the element I_i

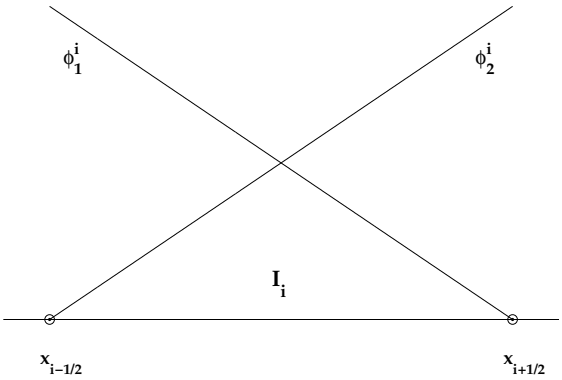
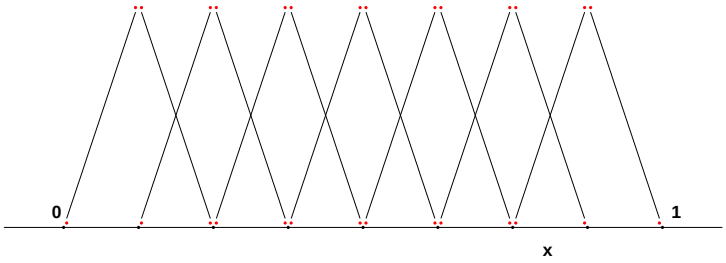


Figure 5.2: Basis functions on the partition with $N = 8$



So the approximated solutions E^h, Φ^h can be expressed in term of the basis functions

$$E^h(x) = \sum_{i=1}^{N}\sum_{k=1}^{2} \beta_k^i \varphi_k^i(x) \quad \text{and} \quad \Phi^h(x) = \sum_{i=1}^{N}\sum_{k=1}^{2} \omega_k^i \varphi_k^i(x).$$

The different parts

We describe the implementation of the different bilinear forms in (4.17):

◇ $\mathcal{A}$ is nothing but the usual *mass-matrix*, which turns out to be block-diagonal since for any interval I_i

$$E^h(x) = \sum_{k=1}^{2} \beta_k^i \varphi_k^i(x), \qquad \forall x \in I_i.$$

The two dimensional block that refers to the interval I_i is given by

$$\int_{I_i} E^h \varphi_j^i \, dx = \int_{I_i} \left(\sum_{k=1}^{2} \beta_k^i \varphi_k^i \right) \varphi_j^i \, dx = \sum_{k=1}^{2} \beta_k^i \underbrace{\int_{I_i} \varphi_k^i \varphi_j^i \, dx}_{(\mathbf{M}_{loc})_{jk}}.$$

The *local mass-matrix* is a 2×2 matrix whose entries are obtained solving

$$(\mathbf{M}_{loc})_{jk} = \int_{I_i} \varphi_k^i \varphi_j^i \, dx = \int_{\bar{I}} \bar{\varphi}_k \bar{\varphi}_j h \, d\bar{x} = h \int_{\bar{I}} \bar{\varphi}_k \bar{\varphi}_j \, d\bar{x}.$$

◇ In $\mathcal{B}$ and $\mathcal{C}$ there is a common term: the *gradient-matrix*.
This has also a block-structure and we describe the implementation of the one in $\mathcal{B}$: for the interval I_i we have

$$\int_{I_i} E^h \left(\varphi_j^i\right)_x \, dx = \int_{I_i} \left(\sum_{k=1}^{2} \beta_k^i \varphi_k^i \right) \left(\varphi_j^i\right)_x \, dx = \sum_{k=1}^{2} \beta_k^i \underbrace{\int_{I_i} \varphi_k^i \left(\varphi_j^i\right)_x \, dx}_{(\mathbf{G}_{loc})_{jk}}.$$

Therefore, similarly to the local mass-matrix, we can compute the *local gradient-matrix* as follows

$$(\mathbf{G}_{loc})_{jk} = \int_{I_i} \varphi_k^i (\varphi_j^i)_x \, dx = \int_{\bar{I}} \bar{\varphi}_k h^{-1} \left(\bar{\varphi}_j\right)_x h \, d\bar{x} = \int_{\bar{I}} \bar{\varphi}_k \left(\bar{\varphi}_j\right)_x \, d\bar{x}.$$

◇ Notice that for $\mathcal{B}$ and $\mathcal{C}$ we also have to take into account the contribution of the numerical fluxes.
For the constants c_{11} and c_{12} we follow [5], thus we take $c_{12} = 1/2$ and $c_{11} = 2h^{-1}$. This choice implies

$$\widehat{E^h} = \left\{E^h\right\} - \frac{1}{2} \left[\!\left[E^h\right]\!\right] + \frac{2}{h} \left[\!\left[\Phi^h\right]\!\right] = \left(E^h\right)^- + \frac{2}{h} \left[\!\left[\Phi^h\right]\!\right],$$
$$\widehat{\Phi^h} = \left\{\Phi^h\right\} + \frac{1}{2} \left[\!\left[\Phi^h\right]\!\right] = \left(\Phi^h\right)^+.$$

Considering $\mathcal{B}$, this implies that, for an element I_i, we also have to consider the contribution given by

$$\left(\Phi^h p^h\right)_{i+\frac{1}{2}} - \left(\Phi^h p^h\right)_{i-\frac{1}{2}} = \left(\Phi^h\right)^+_{i+\frac{1}{2}} \left(p^h\right)^-_{i+\frac{1}{2}} - \left(\Phi^h\right)^+_{i-\frac{1}{2}} \left(p^h\right)^+_{i-\frac{1}{2}} \qquad \text{(see (4.15))}.$$

Notice that by taking $p^h = \varphi_1^i$ we have

$$\left(p^h\right)^-_{i+\frac{1}{2}} = 0 \text{ and } \left(p^h\right)^+_{i-\frac{1}{2}} = 1,$$

and, on the other hand, $p^h = \varphi_2^i$ implies

$$\left(p^h\right)^-_{i+\frac{1}{2}} = 1 \text{ and } \left(p^h\right)^+_{i-\frac{1}{2}} = 1.$$

Furthermore, observe that this also implies

$$\left(\Phi^h\right)^+_{i+\frac{1}{2}} = \omega_1^{i+1}\varphi_1^{i+1}(x_{i+\frac{1}{2}}) \quad \text{and} \quad \left(\Phi^h\right)^+_{i-\frac{1}{2}} = \omega_1^{i}\varphi_1^{i}(x_{i-\frac{1}{2}}).$$

Therefore

$$\left(\Phi^h p^h\right)_{i+\frac{1}{2}} - \left(\Phi^h p^h\right)_{i-\frac{1}{2}} = \omega_1^{i+1}\varphi_1^{i+1}\varphi_2^i|_{x=x_{i+\frac{1}{2}}} - \omega_1^{i}\varphi_1^{i}\varphi_1^i|_{x=x_{i-\frac{1}{2}}}.$$

Next, we observe the contribution of the flux's term of the bilinear form $\mathcal{C}$ in the element I_i.
Proceeding as above we have

$$\begin{aligned}
&\left(\widehat{E^h}w^h\right)_{i+\frac{1}{2}} - \left(\widehat{E^h}w^h\right)_{i-\frac{1}{2}} + \frac{2}{h}\left(\left[\!\left[\Phi^h\right]\!\right] w^h\right)_{i+\frac{1}{2}} - \frac{2}{h}\left(\left[\!\left[\Phi^h\right]\!\right] w^h\right)_{i-\frac{1}{2}} \\
&\quad = \left(\widehat{E^h}\right)^-_{i+\frac{1}{2}} \left(w^h\right)^-_{i+\frac{1}{2}} - \left(\widehat{E^h}\right)^-_{i-\frac{1}{2}} \left(w^h\right)^+_{i-\frac{1}{2}} + \frac{2}{h}\left[\!\left[\Phi^h\right]\!\right]_{i+\frac{1}{2}} \left(w^h\right)^-_{i+\frac{1}{2}} - \frac{2}{h}\left[\!\left[\Phi^h\right]\!\right]_{i-\frac{1}{2}} \left(w^h\right)^+_{i-\frac{1}{2}} \\
&\quad = \beta_2^i\varphi_2^i\varphi_2^i|_{x=x_{i+\frac{1}{2}}} - \beta_2^{i-1}\varphi_2^{i-1}\varphi_1^i|_{x=x_{i-\frac{1}{2}}} + \frac{2}{h}\left[\!\left[\Phi^h\right]\!\right]_{i+\frac{1}{2}} \varphi_2^i(x_{i+\frac{1}{2}}) - \frac{2}{h}\left[\!\left[\Phi^h\right]\!\right]_{i-\frac{1}{2}} \varphi_1^i(x_{i-\frac{1}{2}})
\end{aligned} \tag{5.3}$$

After having created and assembled all these parts, the following linear system remains to be solved

$$\mathbf{A}\ \mathbf{x} = \mathbf{b},$$

in which

- $\mathbf{x} = (\beta, \omega)^T$, with $\beta = (\beta_1, \ldots, \beta_{2N})$ and $\omega = (\omega_1, \ldots, \omega_{2N})$.
- $\mathbf{A}$ belongs to $\mathbb{R}^{4N\times 4N}$ and has the following structure:

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{M} & \mathbf{G_{up}} \\ \hline \mathbf{G_{down}} & \mathbf{D} \end{array}\right]$$

  where $\mathbf{M}$ respresents the *global mass matrix*, $\mathbf{G_{up}}$ and $\mathbf{G_{down}}$ are the sum of the *global gradient matrices* and the contribution given by the flux's terms $\left(\Phi^h\right)^+$ and $\left(E^h\right)^-$ respectively, while $\mathbf{D}$ is the square matrix filled with the contribution of the part related to the jumps $\left(\left[\!\left[\Phi^h\right]\!\right]\right)$ that appears in (5.3).

- The right hand-side is $\mathbf{b} = (0, \ldots, 0, \rho_1, \ldots, \rho_{2N})^T \in \mathbb{R}^{4N}$, where for any $i = 1, \ldots, N$, ρ_{2i-1} and ρ_{2i} are the following integrals

$$\rho_{2i-1} = \int_{I_i} (\rho^h - 1)\varphi_1^i(x)\ dx \quad \text{and} \quad \rho_{2i} = \int_{I_i} (\rho^h - 1)\varphi_2^i(x)\ dx.$$

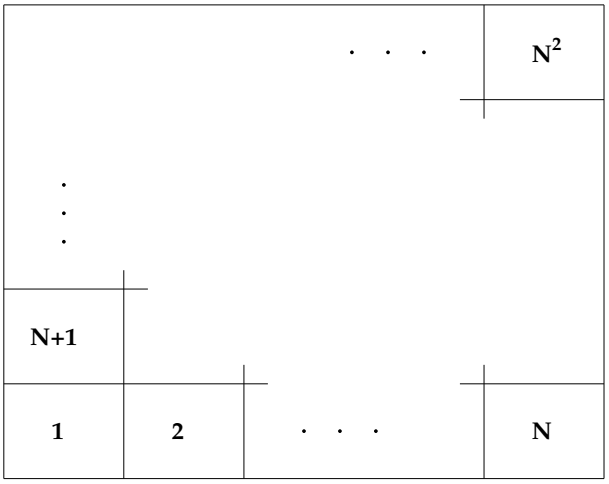
  The function ρ^h is the discrete charge density given by (4.16).

5.2 Two dimensions

5.2.1 Transport equation with standard basis

In two dimensions we need to order the elements of our partition.
We start by the element at the buttom-left corner and we number all elements along the x-coordinate. Once the $N^{\cdot th}$ element in the buttom-right corner is reached, we go up one line and we start again numbering from the left. In this way, the $(N+1)^{\cdot th}$ element results to be the one above the first one and we end with the $\left(N^2\right)^{\cdot th}$ element in the top-right corner (see also figure 5.3).

Figure 5.3: The numbering of the elements



Then, similarly to the one dimensional case we can write our system of equations in the following matrix-formulation

$$\mathbf{M}\ \partial_t \bar{\mathbf{u}} + (\mathbf{Flux_a} + \mathbf{Flux_b})\ \bar{\mathbf{u}} = \mathbf{0},$$

where we have:

- ◇ The *mass-matrix* $\mathbf{M} = h_x h_y\ \mathbf{I}$, with $\mathbf{I}$ the identity.
- ◇ The *flux-matrix* that depends on a: $\mathbf{Flux_a}$.
- ◇ The *flux-matrix* that depends on b: $\mathbf{Flux_b}$.
- ◇ The *vector of the coefficients*: $\bar{\mathbf{u}} = (\bar{u}_1, \ldots, \bar{u}_{N^2})^T$.

  Remark 4 In order to build both flux-matrices one needs to know the connections among the elements. To this end we construct two *connectivity*-matrices:

```
connEW = [ N, 2 ;          connSN = [ (N-1)*N+1, N+1 ;
           1, 3 ;                     (N-1)*N+2, N+2 ;
           2, 4 ;                     (N-1)*N+3, N+3 ;
           ...  ]                         ...        ]
```

In these matrices each row corresponds to the respective element and the values in the columns represent its neighbours:

- In `connEW` (East-West) the first column indicates the left-neighbours and the second one the right-neighbours.
- In `connSN` (South-North) the first column indicates the below-neighbours and the second one the upper-neighbours.

Then for the $k.th$ element K_{ij}, depending on the sign of $a(y_j)$ and $b(x_i)$, the contribution of the respective integral has to appear in the entry of the fluxes matrices which corresponds to the correct neighbour:
Assuming $a(y_j) = 1$ and $b(x_i) = -1$ we would have

```
FluxA(connEW(i,1),k)   =  hy          FluxB(connEW(j+1,2),k) =  hx
FluxA(connEW(i-1,1),k) = -hy          FluxB(connEW(j,2),k)   = -hx
```

5.2.2 Transport equation with hierarchical basis

In general, when a hierarchical basis is used, the implementation of a method becomes more complicated. The main difference with respect to the usual standard basis is the fact that in the support of a hierarchical basis function there could be more than one element.

To solve this issue we build structures (matrices) in which we define connections among elements, edges and each hierarchical basis function. Once all these structures exist, they are used to implement the matrices (*mass*, *fluxes*,...) that arise from the discretized problem. Let us see how this process is carried out.

Using the hierarchical One basis we can write the numerical solution as

$$u^h(x,y,t) = \sum_{\mathbf{l}} \sum_{\mathbf{i}} \alpha_{\mathbf{l},\mathbf{i}}(t)\phi_{\mathbf{l},\mathbf{i}}(x,y), \qquad \text{for all } \mathbf{l} \text{ s.t. } |\mathbf{l}|_\infty \leq n \text{ and } i_1, i_2 \text{ odd},$$

Then the substitution of this expression in (3.10) yields the following

$$
\begin{aligned}
\int_{K_{ij}} \left(u^h\right)_t \psi^h_{ij}\, dx &+ \int_{J_j} \left[\min(a,0)(u^h_{i+1,j} - u^h_{ij}) + \max(a,0)(u^h_{ij} - u^h_{i-1,j})\right] \psi^h_{ij}\, dy \\
&+ \int_{I_i} \left[\min(b,0)(u^h_{i,j+1} - u^h_{ij}) + \max(b,0)(u^h_{ij} - u^h_{i,j-1})\right] \psi^h_{ij}\, dx \\
&= \sum_{\mathbf{l}} \sum_{\mathbf{i}} \left(\alpha_{\mathbf{l},\mathbf{i}}(t)\right)_t \int_{K_{ij}} \phi_{\mathbf{l},\mathbf{i}}(x_i, y_j) \psi^h_{ij}\, dxdy \\
&+ \sum_{\mathbf{l}} \sum_{\mathbf{i}} \min(a,0)\alpha_{\mathbf{l},\mathbf{i}}(t) \int_{J_j} \left[\phi_{\mathbf{l},\mathbf{i}}(x_{i+1}, y_j) - \phi_{\mathbf{l},\mathbf{i}}(x_i, y_j)\right] \psi^h_{ij}\, dy \\
&+ \sum_{\mathbf{l}} \sum_{\mathbf{i}} \max(a,0)\alpha_{\mathbf{l},\mathbf{i}}(t) \int_{J_j} \left[\phi_{\mathbf{l},\mathbf{i}}(x_{i,y_j}) - \phi_{\mathbf{l},\mathbf{i}}(x_{i-1}, y_j)\right] \psi^h_{ij}\, dy \\
&+ \sum_{\mathbf{l}} \sum_{\mathbf{i}} \min(b,0)\alpha_{\mathbf{l},\mathbf{i}}(t) \int_{I_i} \left[\phi_{\mathbf{l},\mathbf{i}}(x_i, y_{j+1}) - \phi_{\mathbf{l},\mathbf{i}}(x_i, y_j)\right] \psi^h_{ij}\, dx \\
&+ \sum_{\mathbf{l}} \sum_{\mathbf{i}} \max(b,0)\alpha_{\mathbf{l},\mathbf{i}}(t) \int_{I_i} \left[\phi_{\mathbf{l},\mathbf{i}}(x_i, y_j) - \phi_{\mathbf{l},\mathbf{i}}(x_i y_{j-1})\right] \psi^h_{ij}\, dx = 0,
\end{aligned}
$$

for all $\psi^h \in \bigoplus_{l \le n} W_{\mathbf{l}}$.

This yields again a system of equations of the form:

$$\mathbf{M}\, \partial_t \bar{\mathbf{a}} + \left(\mathbf{Flux_a} + \mathbf{Flux_b}\right) \bar{\mathbf{a}} = \mathbf{0},$$

where, always the same components appear:

- ◇ The *mass- and fluxes-matrices* $\mathbf{M}$, $\mathbf{Flux_a}$ and $\mathbf{Flux_b}$.
- ◇ The *vector of the coefficients* $\bar{\mathbf{a}} = (\alpha_{(0,0),(1,1)}, \ldots, \alpha_{(n,n),(2^n-1,2^n-1)})^T$.

Observe that also here, when the hierarchical *One* basis is used, $\mathbf{M}$ is not diagonal.

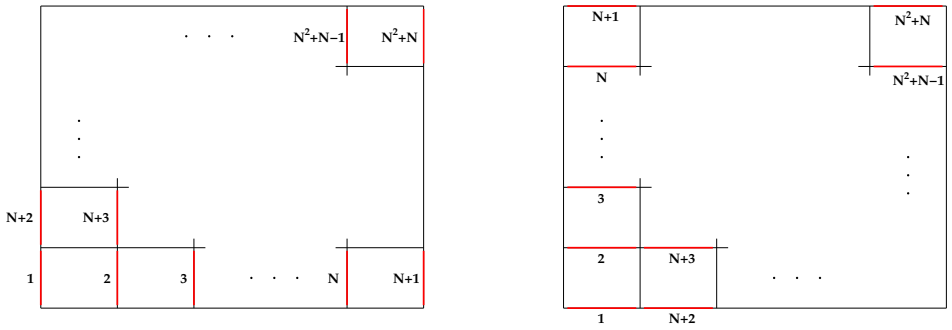
Next, we present the programs which have been implemented in order to build these matrices:

- ◇ `GridPoints` :
  This program creates a matrix in which the *grid points* associated to each hierarchical basis function are defined.
  The output (`xOrd`) is a matrix with N^2 rows (the number of hierarchical basis functions) and four columns. For any row, the first column indicates the x-coordinates, the second one the y-coordinate, the third one the level in x (l_1) and the forth the level in y (l_2) of the corresponding function.

```
xOrd = [ x_01, y_01, 0, 0 ;
         x_11, y_01, 1, 0 ;
         x_21, y_01, 2, 0 ;
         x_22, y_01, 2, 0 ;
               ...        ]
```

◇ `EdgesX/Y` :
These two programs create matrices which indicate the elements shared by any edge of the partition: `EdgesX` refers to the vertical edges, while `EdgesY` to the horizontal ones. The numbering of the edges is showed in figure 5.4.

Figure 5.4: The numbering of the edges



◇ `ElSupport` :
Notice that, the number of hierarchical basis functions in a fixed subspace $W_{\mathbf{l}}$ increases as the level increases. Since the supports of the functions in a subspace are non-overlapping, in each subspace, any element has at most one basis function in which it is different from zero. Therefore it is worthed to know in advance in which basis function's support does any element is.

This program creates two matrices of N^2 rows (number of elements) and $(n+1)^2$ columns (number of subspaces $W_{\mathbf{l}}$, recall $|\mathbf{l}|_\infty \leq n$): `SupportEl` and `Value`.
In each row of the first matrix, there are the indices[1] hierarchical basis functions in which the corresponding element is non-zero. In each row of the second matrix, there are the values taken by the corresponding element when it's evaluated by a hierarchical basis function.

◇ `JumpSupportX/Y` :
The same as for the elements happens with the edges.
With these two programs one creates the matrices that contain the indices of the functions and the values of the jumps for which the edge corresponding to the row is non-zero. Therefore the matrices have N^2+N rows (number of edges) and $(n+1)^2$ columns (number of subspaces).

With these structures the assembly of the *mass-* and *fluxes-matrices* simplifies:

◇ To build the *mass*-matrix we use the data given by `ElSupport`: for each element we know which functions are involved (`SupportEl`) and also the value taken by the functions `Value`. In this way, for any element, the update of the entries of $\mathbf{M}$ is done by a loop over all elements.

◇ The construction of $\mathbf{Flux_a}$ and $\mathbf{Flux_b}$ is similar.
Observe that, by summing up over all elements the weak formulation (3.10) can be rewritten

[1] the two-dimensional indices has to be reorganized in a linear ordering (f.e. lexicographical order).

as follows:

$$\left(\left(u^h\right)_t, \psi^h\right)_{\mathcal{T}_h} + \left\langle \widehat{au^h}, \psi \right\rangle_{\mathcal{J}_h} + \left\langle \widehat{bu^h}, \psi \right\rangle_{\mathcal{I}_h} = 0,$$

in which the notations used mean the following:

$$\begin{aligned}
\left(\left(u^h\right)_t, \psi^h\right)_{\mathcal{T}_h} &= \sum_{i,j} \left(\left(u^h\right)_t, \psi^h\right)_{K_{ij}}, \\
\left\langle \widehat{au^h}, \psi \right\rangle_{\mathcal{J}_h} &= \sum_{i,j} \int_{J_j} \widehat{au^h}_{i+\frac{1}{2},y} [\![\psi]\!]_{i+\frac{1}{2},y} \, dy, \\
\left\langle \widehat{bu^h}, \psi \right\rangle_{\mathcal{I}_h} &= \sum_{i,j} \int_{I_i} \widehat{bu^h}_{x,j+\frac{1}{2}} [\![\psi]\!]_{x,j+\frac{1}{2}} \, dx.
\end{aligned}$$

Thus, we can implement the *fluxes* matrices by considering loops over vertical (for $\mathbf{Flux_a}$) and horizontal edges (for $\mathbf{Flux_b}$).
Thanks to `EdgesX/Y` we know exactly which elements are shared by any edge, thus, by checking the sign of $a(y)$ (or $b(x)$) we know exactly which element we have to consider for the numerical fluxes. Furthermore, by looking at `JumpSupportX/Y` we also know the basis functions that have a non-zero jump on any edge and their corresponding values. Thus, with a loop over all edges, at each step the right entries of the matrices are updated.

Remark 5 (The implementation of the Vlasov-Poisson system)

The spatial discretization we use is almost the same. Here, we just want to point out the changes one has to do to adapt the code to the new problem. In fact, to pass from the two-dimensional transport equation (3.1) to the one-dimensional Vlasov equation (4.1) with the respective initial data and boundary conditions only two small changes are needed:

1. We do not have periodic boundary condition in v but compact support instead, so we change the matrix `EdgesV` by setting to zero the element's neighbours that refer to boundary edges.

2. In the time integration, we have to compute at any time step the electrostatic field E^h and then compute the corresponding flux matrix, checking whether E^h is negative or positive in any interval I_i.
   This check is done by taking the average of the approximations of E^h on the interval I_i, i.e.

$$E_i^h = \frac{E^h(x_{i-\frac{1}{2}}) + E^h(x_{i+\frac{1}{2}})}{2},$$

   where by E_i^h we denote the value of E^h that refers to the interval I_i. Observe that this is nothing but a projection on the space of constant functions.

5.2.3 Alternative method

For this case, in order to write the matrix formulation of the system we want to solve, we use a different notation for the hierarchical basis function:
We denote by $\{\phi_i^{[1]}\}$, $\{\phi_j^{[2]}\}$ and $\{\phi_k^{[R]}\}$ the set of basis[2] of $Z_n^{[1]}$, $Z_n^{[2]}$ and $Z_n^{[R]}$ respectively. Furthermore, we let

$$\bar{\mathbf{a}}_{[s]} = \left(\alpha_1^{[s]}, \ldots, \alpha_{|Z_n^{[s]}|}^{[s]}\right) \qquad \text{for } s = \{1, 2, R\}$$

to be the *vectors of the coefficients*, such that

$$u_{[1]}^h = \sum_i \alpha_i^{[1]} \phi_i^{[1]}, \qquad u_{[2]}^h = \sum_j \alpha_j^{[2]} \phi_j^{[2]} \qquad \text{and} \qquad u_{[R]}^h = \sum_k \alpha_k^{[R]} \phi_k^{[R]}.$$

Then the system can be written as

$$\begin{aligned} \mathbf{M_1} \left(\bar{\mathbf{a}}_{[1]}^{n+1}, \bar{\mathbf{a}}_{[2]}^{n+1}\right)^T &= [\mathbf{M_1} - dt\ \mathbf{Flux_{1L}}] \left(\bar{\mathbf{a}}_{[1]}^{n}, \bar{\mathbf{a}}_{[2]}^{n}\right)^T + dt\ \mathbf{Flux_{1R}} \left(\bar{\mathbf{a}}_{[R]}^{n}\right)^T, \\ \mathbf{M_2} \left(\bar{\mathbf{a}}_{[R]}^{n+1}\right)^T &= [\mathbf{M_2} - dt\ \mathbf{Flux_{2L}}] \left(\bar{\mathbf{a}}_{[R]}^{n}\right)^T + dt\ \mathbf{Flux_{2R}} \left(\bar{\mathbf{a}}_{[1]}^{n+1}, \bar{\mathbf{a}}_{[2]}^{n+1}\right)^T, \end{aligned}$$

where again *mass-* and *fluxes-matrices* appear.

[2] Here, again, we use the framework of the hierarchical One basis, but the same holds also for the hierarchical Haar basis.

References

[1] H.J. Bungartz, M. Griebel, *"Sparse grids"*, Acta Numerica, 13:147-269, 2004.

[2] D. Pflüger, *"Spatially Adaptive Sparse Grids for High-Dimensional Problems"*, Ph.D thesis, Technischen Universität München, 2010.

[3] H.J. Bungartz, T. Dornseifer, *"Sparse Grids: Recent Developments For Elliptic Partial Differential Equations"*, TUM-I9702, SFB-Bericht Nr. 342/02/97 A, 1997.

[4] R.A. Adams, *"Sobolev Spaces"*, Pure and Applied Mathematics, Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-San Francisco-London, 1975.

[5] B. Ayuso, J.A. Carrillo, C.W. Shu, *"Discontinuous Galerkin methods for the one-dimensional Vlasov-Poisson system"*, Kinetic and Related Models, Volume **4**, Number **4**, 2011.

[6] S. Hajian, *"An energy preserving discontinuous Galerkin method for Vlasov-Poisson system"*, M.Sc thesis, Universitat Autónoma de Barcelona, 2011.

[7] B. Cockburn, *"Discontinuous Galerkin Methods"*, School of Mathematics, University of Minnesota, 2003.

[8] P. Castillo, B. Cockburn, I. Perugia, D. Schötzau, *"An a priori analysis of the local discontinuous Galerkin method for elliptic problems"*, SIAM J. Numer. Anal., Vol. 38, No. 5, pp. 1676–1706, 2000

[9] L.C. Evans, *"Partial Differential Equations"*, Graduate Studies in Mathematics, Volume 19, American Mathematical Society, 1998.

[10] A. Quarteroni, R. Sacco, F. Saleri, *"Numerical Mathematics"*, Texts in Applied Mathematics, Springer, 2000.

[11] C.W. Shu, *"Discontinuous Galerkin Methods: General Approach and Stability"*, Division of Applied Mathematics, Brown University Providence, USA.

[12] D.N. Arnold, F. Brezzi, B. Cockburn, L.D. Marini, *"Unified analysis of discontinuous Galerkin Methods for elliptic problems"*, SIAM J. Numer. Anal., Vol. 39, No. 5, pp. 1749–1779, 2002.

[13] S.R. Groot, W.A. van Leeuwen, Ch.G. van Weert, *"Relativistic Kinetic Theory (Principle and applications)"*, North-Holland publishing company, Amsterdam-New York-Oxford, 1980.

[14] S. Wollmann, E. Ozizmir, *"Numerical approximations of the one-dimensional Vlasov-Poisson system with periodic boundary conditions"*, SIAM J. Numer. Anal., Vol. 33, No. 4, pp. 1377–1409, 1996.

[15] C. Villani, *"A review of mathematical topics in collisional kinetic theory"*, Handbook of Mathematical Fluid Dynamics (Vol. 1), Elsevier Science, 2002.

[16] J.B. Goodman, R.J. Leveque, *"On the Accuracy of Stable Schemes for 2D Scalar Conservation Laws"*, Mathematics of Computation, Vol. 45, No. 171, pp. 15-21, American Mathematical Society, 1985.

[17] C.W. Shu, *"Numerical Methods for Hyperbolic Conservation Laws"*, Lecture notes (semester I), Division of Applied Mathematics, Brown University Providence, USA, 2006.

[18] J. Dolbeault, *"An introduction to kinetic equations: the Vlasov-Poisson system and the Boltzmann equation"*, Current developments in partial differential equations (Temuco, 1999), Discrete Contin. Dyn. Syst., 8(2) (2002), pp. 361–380.

[19] J. Cooper, Klimas, *"Vlasov-Maxwell and Vlasov-Poisson equations as models of a one-dimensional electron plasma"*, Phys. Fluids - The Physics of Fluids, Vol. 26, 1983.

[20] B. Ayuso de Dios, S. Castelanelli, *"Application of sparse gird techniques to DG-approximations"*, (Work in preparation).