Projet de Fin d'Études (PFE)
End-of-study project

Speciality : **Simulation and modeling**

Academic year : **2022-2023**

Quantifying the numerical diffusion of non-linear advection schemes

---

Privacy statement : Non-confidential report

Author : **Gabriel Mouttapa**

Promotion : **2022.5**

Tutor ENSTA : **Sonia Fliss**
(UMA laboratory)

Tutor LJK : **Laurent Debreu**
(AIRSEA team)

Internship from 11/04/2023 to 10/10/2023

Laboratoire Jean Kuntzmann

Bâtiment IMAG, Université Grenoble Alpes, 700 avenue Centrale, Domaine Universitaire, 38401 Saint-Martin-d'Hères

Non-confidentiality note

The document is non-confidential. It has been put online by the ENSTA Paris library and can therefore be consulted by the entire school community.

Version

The version of this report is not the version I was tested on for my internship defense for my school. This is a posterior version, where I add some details in the appendices and in the core of the document.

Acknowledgment

I would like to sincerely thank my supervisor, Laurent Debreu, the manager of the AIRSEA team in the Jean Kuntzmann laboratory. He helped me and guided me with a lot of encouragement during my internship and he took a lot of time to explain me the aspects of ocean modeling that I didn't know. I would also thank Florian Lemarié, a permanent member of the AIRSEA team, who helped me with another point of view at the end of my internship and Gabriel Derrida, a PhD student in AIRSEA team from ENSTA Paris, he gave me explanations on his PhD in link with my internship. Then I want to thank Annie Simon and the other members of the administrative, maintenance and cleaning staff of the IMAG building who allowed us to work in good conditions during this warm summer. Finally, I want to thank Milan Gonzalez–Thauvin, Emma Ninucci and the other interns for welcoming me at the LJK and who helped me with this report.

Abstract

In ocean modeling, the diffusion is an important physical phenomenon to analyze. Ocean numerical models, such as NEMO or CROCO, usually use the finite volume method with advection schemes to discretize transport terms on the computational grid. The schemes used for practical applications can be complex because they include certain positivity or nonlinear stability properties which are helpful for guaranteeing the robustness of numerical solutions. However, the advection schemes satisfying these properties introduce a numerical diffusion. To have a correct physical analysis of diffusive processes, it is necessary to quantify this numerical diffusion to make sure it doesn't exceed the diffusion applied for physical reasons. In this report I propose a mathematical framework for systematically evaluating the local numerical diffusion of any advection scheme. In particular, I checked that the method I propose gives results consistent with the previous analysis that had been made by hand. I illustrate these results with diagnostics on Total variation Diminishing (TVD) and Weighted Essentially Non-Oscillatory (WENO) schemes which are widely used in ocean modeling.

Contents

Introduction

Ocean modeling has a relatively recent history. Since the emergence of modern computers, many weather models were developed. Some research in mathematics and computer science aims at making these models more efficient and accurate. It changed the world's relationship with climate by warning of extreme events for example. After this period of understanding the weather science, the environmental science researchers started to do the same work with oceanography. During this time of climate change, a big challenge is to understand how the oceans are evolving, especially now that we are discovering that the interactions between climate and oceans are very important.

The purpose of the modern ocean models is to make forecasts of the oceans and sea's states with a wide variety of temporal and spatial scales : some are used to estimate the climate change for the next century, some others are used to evaluate the trajectory of a drifting boat during a day. The users and developers of these models can be both private or public. To name just a few models, in my previous internships I worked with the models ROMS (Regional Ocean Modeling System) and NEMO (Nucleus for European Modeling of the Ocean), the latter is developed by a European consortium. These numerical models are used for the environmental research purpose as well as for operational applications. The model CROCO (Coastal and Regional Ocean COmmunity model) is developed in part by the AIRSEA team of the Jean Kutzmann Laboratory where I do this internship and is specialized in the interactions between large and fine spatial scales.

The forecasts of the ocean models involve calculating the values of the physical fields on all points of a grid and at different times. The number and the variety of these fields can be very different from a model to another. The spatial grids are in two or three dimensions, and the vertical mesh is not necessarily of constant resolution. The calculation of the physical fields at different time from an initial condition implies in part the resolution of partial differential equations. The equations come from fluid mechanics, but they can involve additional terms to represent the different phenomenons. For example, the oceanographers can adjust the physical diffusion or the sub-mesh parametrization which represents the global effects of some small phenomenons. In addition of these PDE resolutions, a challenge is the adaptation of these resolutions with the complex phenomenons of reality such as the various boundary conditions, the non-linearity of the equations, the stratification of the ocean, the Coriolis force, the interaction with the atmosphere and the data assimilation of the observations.

The vast majority of PDE solvers in ocean models are based on the finite difference or finite volume method. Both of them need to choose a numeric way to calculate the derivatives. To be more specific, in this report, we focus on numerical schemes known as advection schemes. Which calculate the spatial derivative of a field on one point of the grid from its values on the other points. It exists a lot of advection schemes which approximate this derivative, important properties being their order of accuracy, which is the power on the space step of the equivalent of the error between the derivative and its numerical approximation. But these schemes have also other proprieties such as consistency, non-linear stability or total variation diminishing (TVD). These proprieties are used to maintain some physical laws and observation, for example the positivity of a concentration or the non-creation of extrema. If at the beginning of ocean modeling, the schemes were more basic, now they tend to be more complex to verify some of these proprieties and unlike some other physical models, in ocean models, they don't need to have a huge order of accuracy : the third order is usually enough. A part of the mathematical research in ocean modeling is the study of the proprieties of these schemes.

Gabriel Derrida, another ENSTA Paris student, did his final internship last year in the AIRSEA team where I am now. He worked on a propriety of transport schemes used for advection phenomenon : the numerical diffusion implied by these schemes. It is a kind of artificial diffusion created by the numerical resolution of the PDEs. Its presence blurs the study of the physical diffusion that the

oceanographers need. A part of the research of Gabriel Derrida was the estimation of this numerical diffusion for a particular type of advection schemes : the TVD schemes. During my internship I tried to generalize his analysis of numerical diffusion to any advection scheme, including the WENO schemes which are very used in ocean models.

In the part 1 of this report, I will explain the mathematical frame of my work by presenting these advection schemes and the flux form of the equations. Then, in the part 2, one can find the mathematical calculations I used to have a precise quantification of the numerical diffusion. And finally, in the part 3, I will compare some advection schemes taking into account some new tools that comes from this quantification of diffusion.

1 Advection schemes for ocean models

Here, I'm going to present the state of the art about the numerical methods for advection in ocean models that will be used in the following parts. First, I introduce the continuous problem we are working on in section 1.1. Then in section 1.2, I present the numerical methods used in the framework of this report. And finally in section 1.3 I give examples of advection schemes that we are going to analyze in 3.

1.1 The continuous problem

The basic set of equations used in ocean models is called the oceanic primitive equation. If we take $\boldsymbol{U} = (u, v, w)$ the velocity vector, ρ the water density, Θ the temperature field and S_A the salinity, these equations can be written as

$$\nabla \cdot \boldsymbol{U} = 0 \tag{1.1a}$$

$$\frac{\partial \boldsymbol{U}_h}{\partial t} + \nabla \cdot (\boldsymbol{U} \otimes \boldsymbol{U}_h) = -2\boldsymbol{\Omega} \times \boldsymbol{U} - \frac{1}{\rho_0}\nabla_h p + \nu \Delta \boldsymbol{U}_h \tag{1.1b}$$

$$\frac{\partial p}{\partial z} + \rho g = 0 \tag{1.1c}$$

$$\frac{\partial q}{\partial t} + \nabla \cdot (\boldsymbol{U} q) = F_q, \qquad \text{for } q = \Theta, S_A \tag{1.1d}$$

$$\rho = \rho_{\text{eos}}(\Theta, S_A, p_0(z)) \tag{1.1e}$$

with $\boldsymbol{U}_h = (u, v)$ the horizontal velocity vector. The first equation (1.1a) is the approximation of the conservation of mass with the Boussinesq assumption. The equation (1.1b) comes from the conservation of momentum taking into account its viscosity, the Coriolis effect with $\boldsymbol{\Omega} \times \boldsymbol{U}$, the gravity and the viscosity ν on the horizontal dimension. The third one, (1.1c), so-called the hydrostatic equation, is the vertical projection of the momentum conservation with the Boussinesq assumption. The equation (1.1d) describes the evolution of a given tracer q, which can correspond to temperature, salinity or phytoplankton concentration. Then, the equation of state (1.1e) describes the effect of these tracers on the density. In [GA08] one can find more details about this equation set.

One can note that the equations (1.1b) and (1.1d) contain advection terms. Typically, the forcing term F_q includes physical diffusion, that must be compared to the numerical diffusion inherent to discrete advection schemes. The ocean is very stratified, it means that it is made up of several vertical layers with very different densities, and that they almost never mix : the physical vertical diffusion is very small. That's why the issue of the numerical diffusion is very problematic, it can artificially mix some layers of stratification that are not mixed at all in the physical reality.

To separate the physical diffusion from the numerical one in this study, we will take $F_q = 0$, and to simplify we are going to take a constant and positive velocity c on the vertical axis x for this report. However the results presented in the following are straightforward to adapt for a non-constant velocity field. Finally, we rename the tracer of interest u and the equation used in this work is the simple transport equation

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \tag{1.2}$$

We will study this equation on the time-space $]0, T] \times [0, 1]$.

1.2 Numerical methods for advection

In ocean models, when we are working with structured grid, two methods are mainly used : the finite differences and the finite volumes. But for both of them we need to use specific schemes to represent correctly the complex oceanic flows. For example, the equation (1.1b) includes a non-linear term whose stability we want to guarantee, and (1.1d) needs to deal with the physical positivity of the tracers.

We assume a structured spatial and temporal grid. Let us choose a time step Δt and a space step Δx which allows to define the points of the time-space mesh : $t_n = n\Delta t$ and $x_i = (i-1)\Delta x$. We will note N the number of spatial points and we index them from 1 to N.

1.2.1 Finite differences

In the finite differences method, we are searching for discrete values that approximate the solution u directly by its values on the time-space mesh : $u_i^n \simeq u(x_i, t_n)$. Then the approximation of the derivatives on each point of this grid is calculated from these values (u_i^n) using Taylor expansions. If we use an explicit formula for the time derivative, we can compute the values of (u_i^{n+1}) for all i from the values of (u_i^n). The most basic example is the Euler method

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -c \left.\frac{\partial u}{\partial x}\right|_i^n$$

Then we need to choose a way to approximate this spatial derivative at the point x_i. For a sufficiently smooth function u, Taylor expansions for $\Delta x \to 0$ reads give us

$$\frac{\partial u}{\partial x}(x_i, t_n) = \frac{u(x_{i+1}, t_n) - u(x_{i-1}, t_n)}{2\Delta x} - \frac{\Delta x^2}{6}\frac{\partial^3 u}{\partial x^3}(x_i, t_n) + o\big(\Delta x^2\big)$$

We can deduce from this the simplest second-order centered scheme

$$\left.\frac{\partial u}{\partial x}\right|_i^n = \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x}$$

It is called *centered* because its stencil, the set of values (u_i^n) used for the approximation, is centered around the point x_i. Finally the finite difference method can be used as follows with the choose of a temporal initial condition and a spatial boundary condition

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

This method can be extended to higher orders with other Taylor expansions. We can introduce the *Courant number*, which is a dimensionless value useful for the calculations in part 2 and for the stability criterion.

$$\mu_C = c\frac{\Delta t}{\Delta x} \tag{1.3}$$

1.2.2 Finite volume

The finite volume method is very well adapted for hyperbolic conservation laws, which the case here. Unlike the finite differences method, where we considered derivatives approximation, the finite volume method uses the approximation of integrals, like in the finite element method, but unlike it, the finite volume method uses directly the strong formulation of the equation and not the weak one

of the variational formulations. For this method, we need to mesh our interval in *cells* centered on the points x_i and separated by the *interfaces* $x_{i+\frac{1}{2}} = \left(i - \frac{1}{2}\right)\Delta x$ for $i \in [\![1, N]\!]$. Then to represent the solution, we will use the mean values of it on each cells, hence the use of integrals :

$$u_i^n \simeq \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, t_n)\, \mathrm{d}x$$

The integration of (1.2) on the cell $\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ gives

$$\frac{\partial}{\partial t} \int_{x_{i\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t)\, \mathrm{d}x + cu\Big(x_{i+\frac{1}{2}}\Big) - cu\Big(x_{i-\frac{1}{2}}\Big) = 0$$

Like in the finite differences method, we can approximate the temporal derivative of the integral by an explicit scheme like the Euler's one with the values (u_i^n). Here, the quantity cu represent the energetic flux. That's why we introduce $F_{i+\frac{1}{2}}^n$, the *numerical flux* passing through the interface $x_{i+\frac{1}{2}}$ between t_n and t_{n+1}. Then we can write the general form of the finite volume method

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -\frac{1}{\Delta x}\Big(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\Big) \tag{1.4}$$

If we add some boundary conditions, this method allows to compute the values (u_i^n) by recurrence. Then, an approximation of the the solution u can be obtained by some reconstruction methods.

The challenge of the finite volume method lies in the calculation of the flux from the data we have : the mean values (u_i^n). To make this method conservative, we need to have the same definition for the upstream and the downstream flux. That's why we write the flux as the evaluation of a multi-variable function :

$$F_{i+\frac{1}{2}}^n = \mathcal{F}_{i+\frac{1}{2}}(u^n) = \mathcal{F}\big(u_i^n, u_{i+1}^n, u_{i-1}^n, \ldots\big) \tag{1.5}$$

The function $\mathcal{F}$ we choose, represent the *advection scheme* employed in the finite volume method. The *stencil* is the number of variables of the function $\mathcal{F}$, which means the number of points around the interface $x_{i+\frac{1}{2}}$ used to calculate $F_{i+\frac{1}{2}}$, we will note it s. If $\mathcal{F}$ is a linear function, we will have a *linear scheme*. And if $\mathcal{F}$ take the same number of points at the left and the right of the interface $x_{i+\frac{1}{2}}$, we have a *centered scheme*. We are studying an advection equation with a positive velocity c, it means that the initial condition will "move" from left to right, from the view of one fixed point, the data comes from the left. That's why we will use *decentered upwind* schemes, which take more points on the left : the side from which the information comes.

1.2.3 About time integration

In (1.4) we used the Euler's method for the time integration. But, if we take the approach to it is possible to use other methods. Actually, Euler's method induces numerical diffusion that should be corrected in the expression of the flux. Moreover, this scheme is not stable for all the advection schemes we are going to study. That's why in all this report, we are going to use the following fourth-order Runge-Kutta method (RK4), which is stable in all the situations that we are going to

face.

$$
\begin{aligned}
k_i^{(1)} &= -\Delta x\Big(\mathcal{F}_{i+\frac{1}{2}}(u^n) - \mathcal{F}_{i-\frac{1}{2}}(u^n)\Big) \\
k_i^{(2)} &= -\Delta x\Big(\mathcal{F}_{i+\frac{1}{2}}\Big(u^n + \tfrac{\Delta t}{2}k^{(1)}\Big) - \mathcal{F}_{i-\frac{1}{2}}\Big(u^n + \tfrac{\Delta t}{2}k^{(1)}\Big)\Big) \\
k_i^{(3)} &= -\Delta x\Big(\mathcal{F}_{i+\frac{1}{2}}\Big(u^n + \tfrac{\Delta t}{2}k^{(2)}\Big) - \mathcal{F}_{i-\frac{1}{2}}\Big(u^n + \tfrac{\Delta t}{2}k^{(2)}\Big)\Big) \\
k_i^{(4)} &= -\Delta x\Big(\mathcal{F}_{i+\frac{1}{2}}\Big(u^n + \Delta t k^{(3)}\Big) - \mathcal{F}_{i-\frac{1}{2}}\Big(u^n + \Delta t k^{(3)}\Big)\Big) \\
u_i^{n+1} &= u_i^n + \Delta t\Big(k_i^{(1)} + 2k_i^{(2)} + 2k_i^{(3)} + k_i^{(4)}\Big)
\end{aligned}
$$

One can notice that this method divides the temporal step Δt in 4 sub-steps where we applied at each time the finite volume method. Finally, advection schemes and flux representation can be used with finite volumes and finite differences. But we will keep the finite volume formalization for the calculations in part 2.

1.3 Examples of advection schemes

Here I present some of the advection schemes and the flux which are associated. For the sake of simplicity, since we focus here on the spatial discretization, we drop the temporal index n on (u_i^n) values.

1.3.1 Classical linear schemes

First, let us introduce a few classical and simple schemes that are often used as a first step. In figure 1, we can see the simulations done with the following advection schemes..

Upwind 1 (UP1) This scheme is the simplest one. Its associated flux is given by

$$F_{i+\frac{1}{2}}^{UP1} = cu_i \tag{1.6}$$

This scheme has good proprieties in some situations, and it is stable for $\mu_C < 1$. But the major drawback is a large numerical diffusion. It is a first order method. In [Der22], It is shown how to quantify the numerical diffusion of the UP1 scheme when it is used in the specific case of the the finite volumes with Euler's method for temporal integration. In part 2, we will recover these results with a more general approach. In figure 1 wit can be seen that the significant amount of numerical diffusion acts to flatten the initial rectangular function.

Centered scheme (CEN2) This scheme of stencil 2 is of order 2. Its flux is given by

$$F_{i+\frac{1}{2}}^{CEN2} = c\frac{u_i + u_{i+1}}{2} \tag{1.7}$$

It can be shown that this scheme is not diffusive, but it is dispersive. The dispersive effect adds errors in link with the phase, while the diffusive effect affects the amplitude, it is visible on the figure 1.

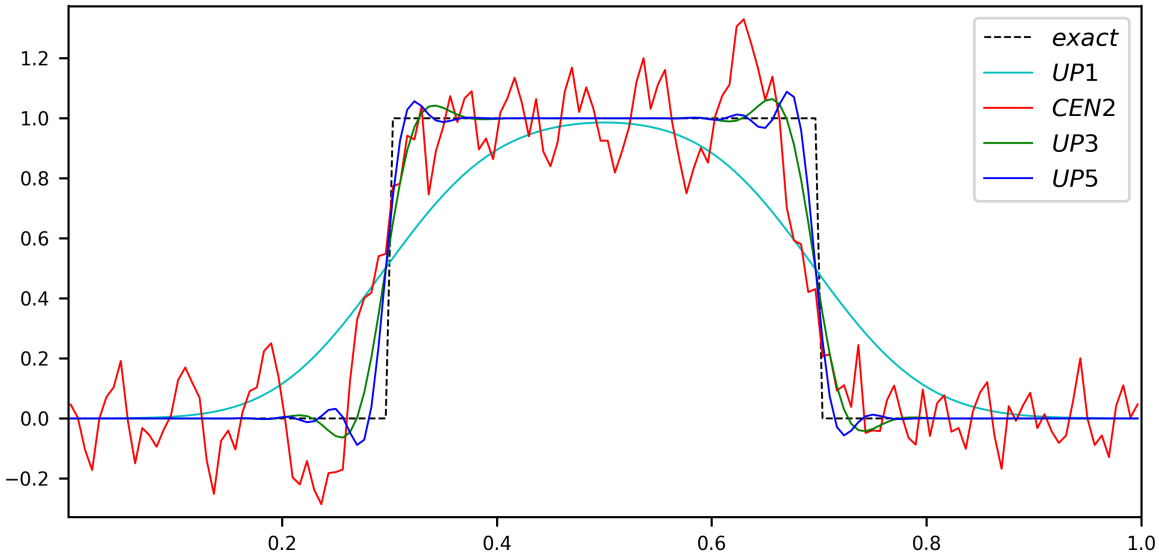

Figure 1: Simulations at time $t = 1$ for different linear advection schemes with a rectangle initial condition and $N = 150$ and $\mu_C = 0.4$ and RK4 for the time solver

Upwind 3 (UP3) This scheme is the simplest scheme of stencil 3 and order 3. Its flux is written

$$F^{UP3}_{i+\frac{1}{2}} = \frac{c}{6}(-u_{i-1} + 5u_i + 2u_{i+1}) \tag{1.8}$$

The coefficients are chosen to make it third-order accurate using Taylor expansions with the finite volumes method (above the third-order, the coefficients are different for the finite differences method). The fact that this scheme is of order 3 implies that where the solution is very regular, it will work very correctly. But where the function is not regular, like in the sides of the rectangle in figure 1, it deforms the rectangle. As UP1, this scheme is diffusive, but we'll see later that its diffusion is of another kind : it is a Bilaplacian diffusion, and it is the reason why the diffusive effects are less visible in the example of the figure 1. Indeed, Bilaplacian diffusion is more scale-selective than Laplacian diffusion inherent to UP1.

Upwind 5 (UP5) Finally, the UP5 scheme is a scheme of stencil and order 5. Its flux is

$$F^{UP5}_{i+\frac{1}{2}} = \frac{c}{60}(2u_{i-2} - 13u_{i-1} + 47u_i + 27u_{i+1} - 3u_{i+2}) \tag{1.9}$$

On figure 1, one can see that UP3 and UP5 have almost the same behavior, and UP5 is not necessarily better cause of its higher order for this particular case.

1.3.2 Total Variation Diminishing schemes (TVD)

One of the main issue of the previous schemes is that they create oscillations. In ocean models it is something that we are searching to avoid. First we need a measure of these oscillations. In the reference book [Dur10] and the set of lessons [DT05], one can find the definition of the *total variation*, which allows to measure oscillations.

TVD property The total variation of a discretized function (u_i) is defined by

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

To define the TVD property, we need to temporarily use the time dimension. A scheme has the *TVD property (Total Variation Diminishing)* if for any initial conditions, the total variation of the solution (u_i^n) it calculates decreases :

$$TV\left(u_i^{n+1}\right) \leqslant TV(u_i^n)$$

Flux limiter methods The TVD property is often obtained through flux limiting method. This method has a strong link with the famous Lax-Wendroff scheme which is internally linked with a time-space approach. But since we chose to work solely with the spatial aspect, we need to work with a version of the flux limiting method which is not the classical one. Its flux is given by

$$F_{i+\frac{1}{2}}^{TVD} = cu_i + c\frac{\Phi_i}{2}(u_{i+1} - u_i) \tag{1.10}$$

We can interpret this flux like an UP1 scheme (1.6) to which we added an anti-diffusive term to compensate the strong diffusion of the UP1. But if we rewrite the flux (1.10), we have

$$F_{i+\frac{1}{2}}^{TVD} = (1 - \Phi_i)cu_i + \Phi_i c\frac{u_i + u_{i+1}}{2}$$

Then we can interpret the TVD flux as a weighted average between an UP1 and a CEN2 scheme (1.7).

This weight is called the *flux limiter* and it is defined as a function of the ratio of the slopes : $\Phi_i = \Phi(\theta_i)$ where

$$\theta_i = \frac{u_i - u_{i-1}}{u_{i+1} - u_i}$$

This value represents the regularity of the function at the point x_i : if it is close to 1, the slopes are close and the function is regular, if it is close to 0 or very large, it indicates the presence of a maximum or a minimum and so an irregularity. If we choose correctly the function Φ, we can obtain a TVD scheme which will limit the creations of extrema. We will work with three flux-limiters, we represented them on figure 2 for a very irregular function and a smooth one. One can note that these schemes are not linear in all cases, it depends on the limiter and the regularity of the function.

Minmod limiter Introduced in [Roe86], this limiter is the simplest one which allows the scheme to have the TVD property.

$$\Phi^{mm}(\theta) = \text{minmod}(1, \theta) = \max(0, \min(\theta, 1))$$

In figure 2, one can see that it is not the most accurate cause of its diffusive effect, actually it is the most diffusive limiter. We will study it in part 2.

SuperBee limiter The SuperBee limiter is defined by

$$\Phi^{sb}(\theta) = \max(0, \min(2\theta, 1), \min(\theta, 2))$$

As we can see on figure 2, it has anti-diffusive effects : it creates angles where the function is initially smooth, it is the result of its TVD property which influences it to not create extrema. This limiter is the less diffusive one (in the meaning that it is anti-diffusive), and all other limiters are located between the Minmod and the SuperBee for this property.

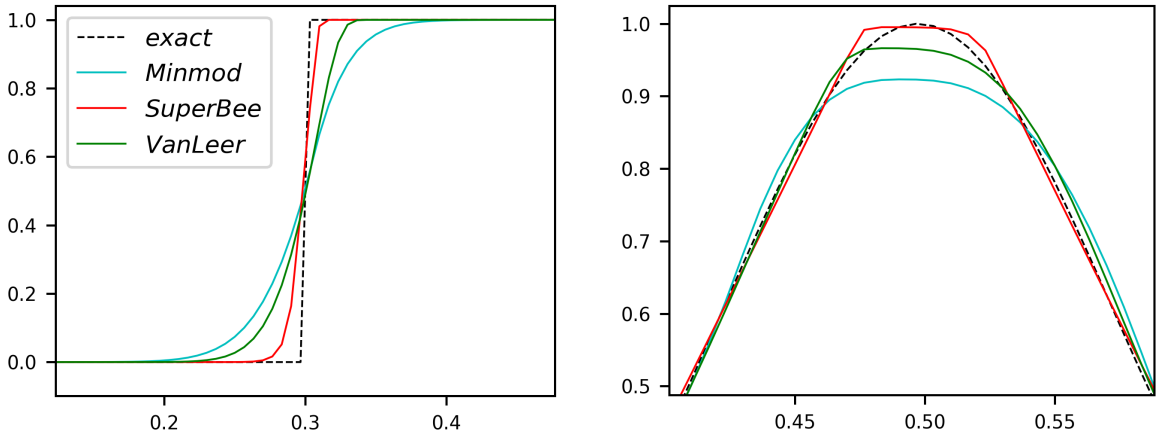

Figure 2: Simulation at $t = 1$ for some TVD schemes with rectangle initial condition (left) and Gaussian initial condition (right) ($N = 150$ and $\mu_C = 0.4$)

Van-Leer limiter Introduced in [Lee74], this limiter is a alternative one between the two previous one as we can see on 2.

$$\Phi^{vl}(\theta) = \frac{\theta + |\theta|}{1 + |\theta|}$$

1.3.3 Weighted Essentially Non-Oscillatory schemes (WENO)

A weakness of TVD schemes is their low order (1 or 2), that's why the ENO and WENO schemes, whose orders are 3 or 5, are often used in ocean models. If the WENO schemes don't have the the TVD property, they have the TVB one (Total Variation Bounded) which is usually strong enough.

ENO schemes The ENO schemes work with Lagrangian interpolations on a good stencil around the discontinuities. They were introduced for the first time in the paper [Har+87]. To sum up the theory presented in [DT05], for an order $2r-1$ scheme, the ENO fluxes are defined for $k \in [\![0, r-1]\!]$ by

$$p^{(k)}_{i+\frac{1}{2}} = c \sum_{l=0}^{r-1} c^{(k,l)} u_{i-k+l}$$

where the coefficients $c^{(k,l)}$ are calculated from interpolations and given in a table. We will work with orders 3 and 5 ($r = 2$ and $r = 3$). The ENO fluxes we are going to work with are

$$p^{(0)}_{i+\frac{1}{2}} = \frac{1}{2}u_i + \frac{1}{2}u_{i+1}$$
$$p^{(1)}_{i+\frac{1}{2}} = -\frac{1}{2}u_{i-1} + \frac{3}{2}u_i$$

for the order 3 and

$$p^{(0)}_{i+\frac{1}{2}} = \frac{1}{3}u_i + \frac{5}{6}u_{i+1} - \frac{1}{6}u_{i+2}$$
$$p^{(1)}_{i+\frac{1}{2}} = -\frac{1}{6}u_{i-1} + \frac{5}{6}u_i + \frac{1}{3}u_{i+1}$$
$$p^{(2)}_{i+\frac{1}{2}} = \frac{1}{3}u_{i-2} - \frac{7}{6}u_{i-1} + \frac{11}{6}u_i$$

for the order 5.

WENO approximation The WENO schemes are simply averages of some ENO schemes done to increase the stencil used. While the ENO method needs to call one of the different fluxes we saw, depending on the regularity of the function, the WENO method avoids these regularity test by processing an average of ENO fluxes with the good weights calculated by taking in account the regularity of the function. The general formulation of the WENO flux is

$$F^{WENO}_{i+\frac{1}{2}} = \sum_{k=0}^{r-1} \omega^{(k)}_{i+\frac{1}{2}} p^{(k)}_{i+\frac{1}{2}} \tag{1.11}$$

The weights $\omega^{(k)}_{i+\frac{1}{2}}$ being calculated by

$$\omega^{(k)}_{i+\frac{1}{2}} = \frac{\alpha^{(k)}_{i+\frac{1}{2}}}{\sum_{l=0}^{r-1} \alpha^{(l)}_{i+\frac{1}{2}}}$$

where

$$\alpha^{(k)}_{i+\frac{1}{2}} = \frac{d^{(k)}}{\left(\beta^{(k)}_{i+\frac{1}{2}} + \varepsilon\right)^2}$$

Here, $d^{(k)}$ are the optimal weights given by a table, and $\beta^{(k)}_{i+\frac{1}{2}}$ is a measure of the regularity of the function on the stencil $[\![i-k, i-k+r-1]\!]$, the ε being here just to avoid division by zero, it is chosen very small, 10^{-12} for us. Several choices exist for $\beta^{(k)}_{i+\frac{1}{2}}$, the one we have chosen comes from the minimization of the total variation of the polynomial reconstruction on the ENO stencils, one can find more information about it in [DT05].

WENO3 For the WENO of third order, we take

$$\beta^{(0)}_{i+\frac{1}{2}} = (u_i - u_{i+1})^2 \qquad \text{with} \qquad d^{(0)} = \frac{2}{3}$$
$$\beta^{(1)}_{i+\frac{1}{2}} = (u_{i-1} - u_i)^2 \qquad \text{with} \qquad d^{(0)} = \frac{1}{3}$$

WENO5 For the WENO of fifth order, we take

$$\beta^{(0)}_{i+\frac{1}{2}} = \frac{13}{12}(u_i - 2u_{i+1} + u_{i+2})^2 + \frac{1}{4}(3u_i - 4u_{i+1} + u_{i+2})^2 \qquad \text{with} \qquad d^{(0)} = \frac{3}{10}$$
$$\beta^{(1)}_{i+\frac{1}{2}} = \frac{13}{12}(u_{i-1} - 2u_i + u_{i+1})^2 + \frac{1}{4}(u_{i-1} - u_{i+1})^2 \qquad \text{with} \qquad d^{(1)} = \frac{3}{5}$$
$$\beta^{(2)}_{i+\frac{1}{2}} = \frac{13}{12}(u_{i-2} - 2u_{i-1} + u_i)^2 + \frac{1}{4}(u_{i-2} - 4u_{i-1} + 3u_i)^2 \qquad \text{with} \qquad d^{(2)} = \frac{1}{10}$$

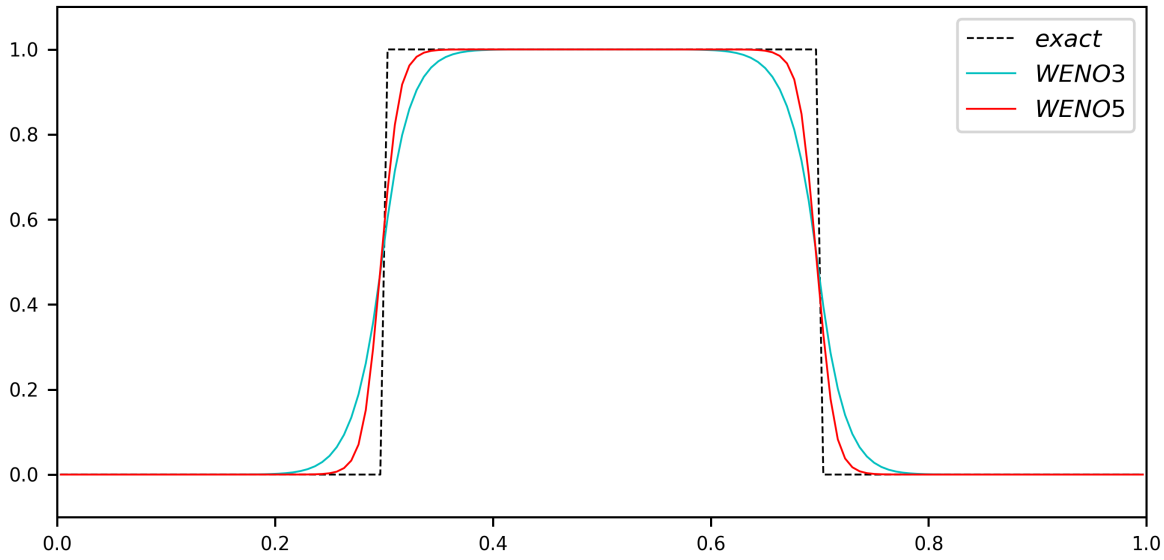

Figure 3: Simulation at $t = 1$ for WENO schemes with rectangle initial condition ($N = 150$ and $\mu_C = 0.4$)

Link with upstream schemes For a sufficiently smooth function on the full stencil of the WENO scheme, we can calculate that we have $\beta^{(k)} = 1$. The weights are then the optimal ones : $\omega^{(k)}_{i+\frac{1}{2}} = d^{(k)}$. After the calculation of the total flux $F_{i+\frac{1}{2}}$ with the formula (1.11), we notice that we have exactly the same coefficients in front of the (u_i) values that the upstream schemes (UP3 for WENO 3 and UP5 for WENO5). When there are no discontinuities, the WENO schemes behave like the upstream which are the best in this situations, but when there is a irregularity in the functions, the WENO schemes have a better behavior by limiting the creation of extrema. One can find illustration of WENO schemes on figure 3, we see that WENO5 seems to be less diffusive than WENO3, we will prove this statement in part 2.

2 A mathematical frame for the quantification of the numerical diffusion

In this section, I will present the mathematical tools I used to have a systematic quantification of the numerical diffusion for all kind of advection schemes. In the first section 2.1, I will introduce a matrix representation of the advective fluxes such as those introduced in section 1.3. Then in section 2.2, I will work on the discrete definition of the numerical diffusion. Next, in section 2.3, the heart of this work, I will show how to systematically identify the diffusive component of advection schemes. Finally, in section 2.4, I am going to address a few questions about the uniqueness of the approach I propose, and in section 2.5 I will discuss the theoretical generalization of my work to advection schemes of arbitrarily high order.

2.1 A matrix representation of the advective flux

Here, we will represent the advective flux using centered expressions with an even stencil s, thus leading to symmetric formulas that make it easier to present our work in section 2.3. To generalize our work to upwind schemes with an odd stencil s, we just have to add a zero coefficient on the right to formulate it as a scheme with an even stencil $s-1$.

2.1.1 The flux difference matrix

Let us start with a stencil $s=2$. As the flux is defined around the interface $x_{i+\frac{1}{2}}$, this stencil leads us to work with the values u_i and u_{i+1}. Then we assume that we can write the flux like this

$$F_{i+\frac{1}{2}} = c\big(a_{i+\frac{1}{2}}u_i + b_{i+\frac{1}{2}}u_{i+1}\big) \tag{2.1}$$

where $a_{i+\frac{1}{2}}$ and $b_{i+\frac{1}{2}}$ are functions of u_i and u_{i+1} which verifies

$$a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}} = 1 \tag{2.2}$$

for consistency of the corresponding numerical scheme. With this hypothesis, we can rewrite the flux difference of the finite volumes at every cell x_i as

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = c\Big[-a_{i-\frac{1}{2}}u_{i-1} + \big(-b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\big)u_i + b_{i+\frac{1}{2}}u_{i+1}\Big]$$

Let us introduce the matrix $\mathbb{F} \in \mathbb{R}^{N\times N}$, that we will call *flux difference matrix*. We define this matrix by its lines :

$$\boldsymbol{e}_i^\top \mathbb{F} = \Big(0 \cdots\cdots 0 \quad \underset{\substack{\uparrow\\ i-1}}{-a_{i-\frac{1}{2}}} \quad \underset{\substack{\uparrow\\ i}}{-b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}} \quad \underset{\substack{\uparrow\\ i+1}}{b_{i+\frac{1}{2}}} \quad 0 \cdots\cdots 0\Big)$$

Where $\boldsymbol{e}_i$ is the i-th vector of the canonical basis of $\mathbb{R}^N$. This matrix allows to write the flux difference as a matrix-vector product :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = c\boldsymbol{e}_i^\top \mathbb{F}\boldsymbol{U} \tag{2.3}$$

where $\boldsymbol{U}$ is the vector of the solution values (u_i) :

$$\boldsymbol{U} = \big(u_1 \cdots\cdots u_N\big)^\top$$

Note that it is not the same $\boldsymbol{U}$ than in the primitive equations (1.1).

2.1.2 A comment about interval edges

One can notice that we use N cells and N interfaces, which seems to be contradictory. Actually we have chosen to use periodic boundary conditions. It means the values (u_i) are also processed periodically, for example u_{N+2} is defined equal to u_2 and u_0 equal to u_{N-1}. With this assumption, we can define all the flux on the interfaces $x_{i+\frac{1}{2}}$ for $i \in [\![1, N-1]\!]$ with the formula (1.5). And in addition we can define the flux $F_{N+\frac{1}{2}}$ like the flux between the cells x_N and x_1. We end up with N interfaces and N cells numbered from 1 to N periodically, and it will facilitate the following calculations.

2.1.3 Symmetric part and energy considerations

Let us get back to the matrix $\mathbb{F}$. If we write it around the position (i, i) we have

$$\mathbb{F} = \begin{pmatrix} \ddots & \ddots & \ddots & & \\ -a_{i-\frac{3}{2}} & -b_{i-\frac{3}{2}} + a_{i-\frac{1}{2}} & b_{i-\frac{1}{2}} & & \\ & -a_{i-\frac{1}{2}} & -b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}} & b_{i+\frac{1}{2}} & \\ & & -a_{i+\frac{1}{2}} & -b_{i+\frac{1}{2}} + a_{i+\frac{3}{2}} & b_{i+\frac{3}{2}} \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{matrix} \\ \leftarrow i-1 \\ \leftarrow i \\ \leftarrow i+1 \\ \\ \end{matrix}$$
$$\begin{matrix} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ i-2 & i-1 & i & i+1 & i+2 \end{matrix}$$

This way of forming the matrix allows to split $\mathbb{F}$ between its symmetric part $\mathbb{S}$ and its anti-symmetric part $\mathbb{A}$. For instance, the line i of $\mathbb{S}$ can be written as

$$\begin{aligned} \boldsymbol{e}_i^\intercal \mathbb{S} &= \frac{1}{2}(\boldsymbol{e}_i^\intercal \mathbb{M} + \mathbb{M}^\intercal \boldsymbol{e}_i) \\ &= \begin{pmatrix} \frac{1}{2}\left(-a_{i-\frac{1}{2}} + b_{i-\frac{1}{2}}\right) & -b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}} & \frac{1}{2}\left(-a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}}\right) \end{pmatrix} \\ &\qquad \begin{matrix} \uparrow & \uparrow & \uparrow \\ i-1 & i & i+1 \end{matrix} \end{aligned} \tag{2.4}$$

$\mathbb{A}$ is defined to be an anti-symmetric matrix. It means that $\mathbb{A}^\top = -\mathbb{A}$. When we use this property in a scalar product in $\mathbb{R}^N$, we have $\langle \boldsymbol{U}, \mathbb{A}\boldsymbol{U}\rangle = \langle \mathbb{A}^\top \boldsymbol{U}, \boldsymbol{U}\rangle = -\langle \boldsymbol{U}, \mathbb{A}\boldsymbol{U}\rangle$, which means $\langle \boldsymbol{U}, \mathbb{A}\boldsymbol{U}\rangle = 0$. Adding the decomposition $\mathbb{F} = \mathbb{S} + \mathbb{A}$, we obtain $\langle \boldsymbol{U}, \mathbb{F}\boldsymbol{U}\rangle = \langle \boldsymbol{U}, \mathbb{S}\boldsymbol{U}\rangle$.

The internal energy of our system can be defined as

$$E = \int_0^1 \frac{1}{2} u^2 \, \mathrm{d}x \tag{2.5}$$

Let us derive the energy with respect to time.

$$\begin{aligned} \frac{\mathrm{d}E}{\mathrm{d}t} &= \int_0^1 \frac{1}{2} \frac{\partial(u^2)}{\partial t} \, \mathrm{d}x \\ &= \int_0^1 u \frac{\partial u}{\partial t} \, \mathrm{d}x \end{aligned}$$

using the advection equation (1.2) we can replace the time derivative by a spatial derivative to get

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -c \int_0^1 u \frac{\partial u}{\partial x} \, \mathrm{d}x$$

by discretizing the integral, we obtain

$$\frac{\mathrm{d}E}{\mathrm{d}t} \simeq -c\Delta x \sum_{i=1}^{N} u_i \left.\frac{\partial u}{\partial x}\right|_i \tag{2.6}$$

And finally, using the flux representation of the finite volume method (1.4), we come to

$$\begin{aligned}\frac{\mathrm{d}E}{\mathrm{d}t} &= \sum_{i=1}^{N} u_i\big(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}\big) \\ &\overset{(2.3)}{=} c\sum_{i=1}^{N} u_i \boldsymbol{e}_i^\top \mathbb{F}\boldsymbol{U} \\ &= c\langle \boldsymbol{U}, \mathbb{F}\boldsymbol{U}\rangle\end{aligned}$$

or equivalently,

$$\frac{\mathrm{d}E}{\mathrm{d}t} = c\langle \boldsymbol{U}, \mathbb{S}\boldsymbol{U}\rangle \tag{2.7}$$

With this equation, we have a link between the variation of energy and the matrix of the difference of the flux $\mathbb{F}$. All this calculation was done by separating time and space and on one specific time t_n. Note that this result can also be found by directly discretizing the integral without using the advection equation. We conclude with (2.7) that the skew-symmetric part of the flux difference doesn't influence the energy, while the symmetric part does. That's why we will focus on this symmetric part $\mathbb{S}$ in the following.

2.1.4 Higher order stencils

So far, our derivation only works for schemes of stencils 1 or 2. But we need to go further for the schemes such as TVD or WENO. For the stencils 4 and 6, the calculation of the matrix $\mathbb{F}$ and $\mathbb{S}$ are quite easy and we just give the results here that we will use in section 2.3.

Stencil 4 We represent the schemes of stencil 3 and 4 by the following flux

$$F_{i+\frac{1}{2}} = c\big(a_{i+\frac{1}{2}}u_{i-1} + b_{i+\frac{1}{2}}u_i + c_{i+\frac{1}{2}}u_{i+1} + d_{i+\frac{1}{2}}u_{i+2}\big) \tag{2.8}$$

where, in general, $a_{i+\frac{1}{2}}, b_{i+\frac{1}{2}}, c_{i+\frac{1}{2}}$ and $d_{i+\frac{1}{2}}$ are functions of u_{i-1}, u_i, u_{i+1} and u_{i+2}. We will add the relation of consistency $a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}} + c_{i+\frac{1}{2}} + d_{i+\frac{1}{2}} = 1$. The symmetric part of the flux difference matrix of this scheme can be describe by its lines with

$$\begin{array}{ccccc} & & i-1 & i & \\ & & \downarrow & \downarrow & \\ \boldsymbol{e}_i^\intercal\mathbb{S} = \Big(\frac{1}{2}\big(d_{i-\frac{3}{2}} - a_{i-\frac{1}{2}}\big) & \frac{1}{2}\big(-d_{i-\frac{3}{2}} - b_{i-\frac{1}{2}} + c_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\big) & -c_{i-\frac{1}{2}} + b_{i+\frac{1}{2}} & & \\ & \frac{1}{2}\big(-d_{i-\frac{1}{2}} - b_{i+\frac{1}{2}} + c_{i+\frac{1}{2}} + a_{i+\frac{3}{2}}\big) & \frac{1}{2}\big(d_{i+\frac{1}{2}} - a_{i+\frac{3}{2}}\big)\Big) & & \\ & \uparrow & \uparrow & & \\ & i+1 & i+2 & & \end{array} \tag{2.9}$$

Here the colors are used to differentiate the columns. We can see the beginning of a symmetry in the organization of these coefficients.

Stencil 6 For the schemes of stencils 5 and 6 we write the flux like this

$$F_{i+\frac{1}{2}} = c\big(a_{i+\frac{1}{2}}u_{i-2} + b_{i+\frac{1}{2}}u_{i-1} + c_{i+\frac{1}{2}}u_i + d_{i+\frac{1}{2}}u_{i+1} + e_{i+\frac{1}{2}}u_{i+2} + f_{i+\frac{1}{2}}u_{i+3}\big) \tag{2.10}$$

The matrix $\mathbb{S}$ can be describe with

$$\begin{array}{cccccc} & & \overset{i-3}{\downarrow} & & \overset{i-2}{\downarrow} & \\ \boldsymbol{e}_i^\intercal\mathbb{S} = \Big(& \frac{1}{2}\big(f_{i-\frac{5}{2}} - a_{i-\frac{1}{2}}\big) & & \frac{1}{2}\big(-f_{i-\frac{5}{2}} + e_{i-\frac{3}{2}} - b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\big) & & \\ & \frac{1}{2}\big(-e_{i-\frac{3}{2}} - c_{i-\frac{1}{2}} + d_{i-\frac{1}{2}} + b_{i+\frac{1}{2}}\big) & & -d_{i-\frac{1}{2}} + c_{i+\frac{1}{2}} & & \frac{1}{2}\big(-e_{i-\frac{1}{2}} - c_{i+\frac{1}{2}} + d_{i+\frac{1}{2}} + b_{i+\frac{3}{2}}\big) \\ & & \overset{i-1}{\uparrow} & & \overset{i}{\uparrow} & & \overset{i+1}{\uparrow} \\ & & & \frac{1}{2}\big(-f_{i-\frac{1}{2}} + e_{i+\frac{1}{2}} - b_{i+\frac{3}{2}} + a_{i+\frac{5}{2}}\big) & & \frac{1}{2}\big(f_{i+\frac{1}{2}} - a_{i+\frac{5}{2}}\big)\Big) \\ & & & \overset{i+2}{\uparrow} & & \overset{i+3}{\uparrow} \end{array} \tag{2.11}$$

One can notice that the three matrices $\mathbb{S}$ we give in (2.4), (2.9) and (2.11) are consistent with each other : we can treat the schemes of stencil 2 with the stencil 6 formulas. But we will continue to separate these three for the sake of simplicity.

2.2 From the continuous to the discrete definition of the diffusion operator

In this subsection we will introduce the mathematical sense of the continuous diffusion operator with the coefficients of diffusion. And we will see how it is possible to have a discrete version of the diffusion. The goal being to represent the gradient of the flux with a diffusion operator. I detail the calculations only for the Laplacian diffusion, for the Bilaplacian and Trilaplacian diffusion, please refer to the appendix A.

2.2.1 The continuous Laplacian diffusion operator

The diffusion operator is usually introduced by a Laplacian operator Δu, or by the iteration of simple derivatives with a coefficient L (for Laplacian diffusion) between them :

$$\frac{\partial}{\partial x} L \frac{\partial u}{\partial x} \tag{2.12}$$

Here we will focus on the average effect of the diffusion effect on the energy. That's why to define the global quantity of diffusion D^L on the interval $[0,1]$, we integrate the operator of diffusion multiplied by u :

$$D^L = -\frac{\Delta x^2}{\Delta t}\int_0^1 u\frac{\partial}{\partial x}\left(L\frac{\partial u}{\partial x}\right)\mathrm{d}x \tag{2.13}$$

This kind of diffusion is also called the *harmonic diffusion*, it is a diffusion of order 2. The coefficient L is a function defined on $[0,1]$. It is mostly positive but in case of anti-diffusion effects, it can be negative. Here, we chose to make this coefficient without any physical dimension by scaling it with $\Delta x^2/\Delta t$ but in the literature the function L usually has a dimension. Also the sign here is chose to make the quantity D^L positive if L is. Indeed, assuming that u and L are periodic on $[0,1]$, with using an integration by part, we have

$$D^L = \frac{\Delta x^2}{\Delta t}\int_0^1 L\left(\frac{\partial u}{\partial x}\right)^2\mathrm{d}x \tag{2.14}$$

This other expression of the Laplacian diffusion will be useful to understand the following discrete transformation. It is also useful to understand how the diffusion effects produce a loss of energy, indeed we see that if L is positive, the quantity D^L is also positive.

2.2.2 The discretization of the integral

The definition of the diffusion we just give is global. It doesn't allow a local identification of the diffusion, which is one of the goal of this work. Let us approximate the spatial derivative by a centered scheme on the interfaces $x_{i+\frac{1}{2}}$,

$$\left.\frac{\partial u}{\partial x}\right|_{i+\frac{1}{2}} = \frac{u_{i+1}-u_i}{\Delta x} \tag{2.15}$$

To discretize the integral in (2.14), we are led to discretize the coefficient L on the interfaces and not at cell centers :

$$\begin{aligned} D^L &\simeq \frac{\Delta x^2}{\Delta t}\sum_{i=1}^{N} \Delta x L_{i+\frac{1}{2}} \left.\frac{\partial u}{\partial x}\right|^2_{i+\frac{1}{2}} \\ \frac{\Delta t}{\Delta x} D^L &= \sum_{i=1}^{N} L_{i+\frac{1}{2}}(u_i - u_{i+1})^2 \end{aligned}$$

Here we apply a manipulation of sums that can be interpreted as a discrete integration by part. We need the periodicity of the (u_i) that we discussed in subsection 2.1.2 and we also assume the periodicity of the coefficients $\left(L_{i+\frac{1}{2}}\right)$.

$$\begin{aligned} \frac{\Delta t}{\Delta x} D^L &= \sum_{i=1}^{N} u_i L_{i+\frac{1}{2}}(u_i - u_{i+1}) - \sum_{i=1}^{N} u_{i+1} L_{i+\frac{1}{2}}(u_i - u_{i+1}) \\ &= \sum_{i=1}^{N} u_i L_{i+\frac{1}{2}}(u_i - u_{i+1}) - \sum_{i=2}^{N+1} u_i L_{i-\frac{1}{2}}(u_{i-1} - u_i) \\ &= \sum_{i=1}^{L} u_i \left[-L_{i-\frac{1}{2}} u_{i-1} + \left(L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}}\right) u_i - L_{i-\frac{1}{2}} u_{i+1} \right] \end{aligned} \tag{2.16}$$

The aim of my study is to be able to estimate the coefficients $\left(L_{i+\frac{1}{2}}\right)$ from the matrix expression of the advective fluxes introduced in section 2.1. Once these coefficients are known, they can be used in tools used to study the advections schemes in part 3.

2.2.3 The diffusion matrix

Here we follow the same methodology as in section 2.1. Let us introduce the *Laplacian diffusion matrix* $\mathbb{D}^L$ defined by its lines like this

$$\boldsymbol{e}_i^\top \mathbb{D}^L = \Big(0 \cdots\cdots 0 \quad \underset{\substack{\uparrow \\ i-1}}{-L_{i-\frac{1}{2}}} \quad \underset{\substack{\uparrow \\ i}}{L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}}} \quad \underset{\substack{\uparrow \\ i+1}}{L_{i+\frac{1}{2}}} \quad 0 \cdots\cdots 0\Big) \tag{2.17}$$

Note that $\mathbb{D}^L$ is a symmetric matrix. We can write the diffusion as a scalar product :

$$D^L = \frac{\Delta x}{\Delta t}\left\langle \boldsymbol{U}, \mathbb{D}^L \boldsymbol{U}\right\rangle \tag{2.18}$$

The elements of $\mathbb{D}^L$ are obviously linked to these coefficients and to this matrix :

$$\mathbb{C}^L = \begin{pmatrix} C^L_{-\frac{1}{2},-1} & C^L_{-\frac{1}{2},0} & C^L_{-\frac{1}{2},+1} \\ C^L_{+\frac{1}{2},-1} & C^L_{+\frac{1}{2},0} & C^L_{+\frac{1}{2},+1} \\ \uparrow & \uparrow & \uparrow \\ u_{i-1} & u_i & u_{i+1} \end{pmatrix} \begin{matrix} \\ \leftarrow L_{i-\frac{1}{2}} \\ \leftarrow L_{i+\frac{1}{2}} \end{matrix} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{2.19}$$

in the sense that we can write for $i \in [\![1, N]\!]$ and $s \in [\![-1, 1]\!]$ the coefficient in position $(i, i+s)$ of $\mathbb{D}^L$ with the following sum

$$\mathbb{D}^L_{i,i+s} = \sum_{l=-\frac{1}{2}}^{\frac{1}{2}} C^L_{l,s} L_{i+l} \tag{2.20}$$

Here the sum is done on half-integers. These objects will be use later for the generalization of our approach in section 2.5.

2.2.4 Higher order diffusion operators

Here we introduce the concepts of Bilaplacian and Trilaplacian diffusion. With these kind of diffusion, we can do the same work we did with the Laplacian diffusion, but as the calculations are tedious, we put them in the appendix A.

Bilaplacian diffusion Also called the *biharmonic diffusion operator*, it can be defined by

$$\frac{\partial^2}{\partial x^2}\left(B\frac{\partial^2 u}{\partial x^2}\right) \tag{2.21}$$

This operator allows us to define a *global quantity of Bilaplacian diffusion* D^B. In the appendix A.1 we discretize the coefficient B on the cells with the family (B_i). And then we calculate the matrix $\mathbb{D}^B$ such that

$$D^B = \frac{\Delta x}{\Delta t}\left\langle \boldsymbol{U}, \mathbb{D}^B \boldsymbol{U} \right\rangle \tag{2.22}$$

It's given by its rows with

$$\boldsymbol{e}_i^\top \mathbb{D}^B = \begin{pmatrix} 0 \cdots\cdots & B_{i-1} & -2B_{i-1} - 2B_i & B_{i-1} + 4B_i + B_{i+1} & -2B_i - 2B_{i+1} & B_{i+1} \cdots\cdots 0 \\ & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ & i-2 & i-1 & i & i+1 & i+2 \end{pmatrix} \tag{2.23}$$

We can also calculate the matrix $\mathbb{C}^B$. This diffusion is of order 4.

Trilaplacian diffusion The next step is the *triharmonic diffusion* or sixth order diffusion with the operator

$$\frac{\partial^3}{\partial x^3}\left(T\frac{\partial^3 u}{\partial x^3}\right) \tag{2.24}$$

In the appendix A.2 we show that the coefficient of diffusion T, is naturally discretized on the interfaces with the family $\left(T_{i+\frac{1}{2}}\right)$. Then, as before, we can define D^T, $\mathbb{D}^T$ and $\mathbb{C}^T$.

2.3 Quantifying numerical diffusion inherent to advective flux

We focus here on the second-order case and computations for orders 4 and 6 can be find in A. We investigate two different ways to identify the diffusion coefficients from the matrix representation of the flux.

2.3.1 The expansions point of view

An intuitive approach to make a link between the diffusion and the symmetric part of the flux difference $\langle \boldsymbol{U}, \mathbb{S}\boldsymbol{U}\rangle$ amounts to find the coefficients $L_{i+\frac{1}{2}}$ such that :

$$\frac{\Delta t}{\Delta x} D^L = \langle \boldsymbol{U}, \mathbb{S}\boldsymbol{U}\rangle$$

To do so, we use the expansion of the diffusion (2.16). This expansion is a linear combination of u_{i-1}, u_i and u_{i+1}, that's why we only need a flux of stencil $s = 2$. We deduce the expansion of the symmetric part of the flux difference with (2.4). We end up with this expansions equality

$$\begin{aligned}\frac{\Delta t}{\Delta x}\sum_{i=1}^{L} u_i\Big[-L_{i-\frac{1}{2}}u_{i-1} + \big(L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}}\big)u_i - L_{i-\frac{1}{2}}u_{i+1}\Big] \\ = c\sum_{i=1}^{N} u_i\left[\frac{1}{2}\big(-a_{i-\frac{1}{2}} + b_{i-\frac{1}{2}}\big)u_{i-1} + \Big(-b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\Big)u_i + \frac{1}{2}\big(-a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}}\big)u_{i+1}\right]\end{aligned}$$

This equality leads us to write

$$\begin{aligned}-L_{i-\frac{1}{2}}u_{i-1} + \big(L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}}\big)u_i - L_{i-\frac{1}{2}}u_{i+1} \\ = \mu_C\left[\frac{1}{2}\big(-a_{i-\frac{1}{2}} + b_{i-\frac{1}{2}}\big)u_{i-1} + \Big(-b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\Big)u_i + \frac{1}{2}\big(-a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}}\big)u_{i+1}\right]\end{aligned} \tag{2.25}$$

where μ_C is the dimensionless Courant number we introduced in (1.3).

2.3.2 The energetic point of view

A second approach arises from energy considerations. The diffusion is sometimes describe as a loss of energy. In (2.5) we defined the energy as the half of the square of the L^2 norm of u. For instance, we saw in figure 1 that the effect of diffusion is to decrease this norm. Therefore it is natural to define the numerical diffusion as the temporal derivative of the energy :

$$D^L = \frac{\mathrm{d}E}{\mathrm{d}t}$$

Discretizing both sides of the equation with (2.7) and (2.18) we obtain

$$\frac{\Delta x}{\Delta t}\left\langle \boldsymbol{U}, \mathbb{D}^L\boldsymbol{U}\right\rangle = c\left\langle \boldsymbol{U}, \mathbb{S}\boldsymbol{U}\right\rangle \tag{2.26}$$

which, as we will show below, implies that

$$\mathbb{D}^L = \mu_C\mathbb{S} \tag{2.27}$$

The analysis of sections 2.1 and 2.2 are available for all values of (u_i). So the scalar product relation (2.26) is true for all $\boldsymbol{U} \in \mathbb{R}^N$, what we can rewrite

$$\forall \boldsymbol{U} \in \mathbb{R}^N, \qquad \left\langle \boldsymbol{U}, \big(\mathbb{D}^L - \mu_C\mathbb{S}\big)\boldsymbol{U}\right\rangle = 0$$

$\mathbb{S}$ is symmetric by definition, and we observed that $\mathbb{D}^L$ is also symmetric in (2.17), meaning that $\mathbb{D}^L - \mu_C\mathbb{S}$ is symmetric and thus diagonalizable. If we note Λ the diagonal matrix of the eigenvalues of $\mathbb{D}^L - \mu_C\mathbb{S}$, we can find that for all $\boldsymbol{V} \in \mathbb{R}^N$, $\langle \boldsymbol{V}, \Lambda\boldsymbol{V}\rangle = 0$, which means that $\Lambda = 0$ which proves the relation (2.27).

2.3.3 Flux-diffusion system

Whatever the viewpoint we take, both of the relations (2.25) and (2.27) lead us to have the following relationships between the diffusion coefficients and the flux coefficients

$$\forall i \in [\![1, N]\!], \qquad \begin{cases} -L_{i-\frac{1}{2}} = \dfrac{\mu_C}{2}\left(-a_{i-\frac{1}{2}} + b_{i-\frac{1}{2}}\right) \\ L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}} = \mu_C\left(-b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\right) \\ -L_{i+\frac{1}{2}} = \dfrac{\mu_C}{2}\left(-a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}}\right) \end{cases}$$

As it happens, these N systems are consistent. Indeed the third line of the system i is exactly the first line of the system i. And obviously, the sum of the first and third lines is the opposite of the second one. Finally the resolution of these systems gives us

$$\boxed{L_{i+\frac{1}{2}} = \frac{\mu_C}{2}\left(a_{i+\frac{1}{2}} - b_{i+\frac{1}{2}}\right)} \tag{2.28}$$

If we use the relation between the flux coefficients (2.2), we have $L_{i+\frac{1}{2}} = \frac{\mu_C}{2}\left(2a_{i+\frac{1}{2}} - 1\right) = \frac{\mu_C}{2}\left(1 - 2b_{i+\frac{1}{2}}\right)$, but we keep the formula (2.28) because it is more symmetric.

2.3.4 Formulas for stencils 4 and 6

For higher order diffusion, we need to expand the flux stencils.

Stencil 4 For a stencil of 3 or 4 represented by (2.8), the symmetric part of the flux (2.9) is described from the point x_{i-2} to the point x_{i+2}. We need to add a diffusion representation which includes this larger stencil : the Bilaplacian diffusion. In fact, we also keep the Laplacian diffusion : the advection schemes of stencil 3 or 4 can mix Laplacian and Bilaplacian diffusion effects. The same reasoning on energy that we did in subsection 2.3.2 allows to write

$$D^L + D^B = \frac{\mathrm{d}E}{\mathrm{d}t}$$

Using (2.22) we have

$$\mathbb{D}^L + \mathbb{D}^B = \mu_C \mathbb{S}$$

With (2.9), (2.23) and (2.17), this matrix equality leads us to the system

$$\forall i \in [\![1, N]\!], \qquad \begin{cases} B_{i-1} = \dfrac{\mu_C}{2}\left(d_{i-\frac{3}{2}} - a_{i-\frac{1}{2}}\right) \\ -L_{i-\frac{1}{2}} - 2B_{i-1} - 2B_i = \dfrac{\mu_C}{2}\left(-d_{i-\frac{3}{2}} - b_{i-\frac{1}{2}} + c_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\right) \\ L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}} + B_{i+1} + 4B_i + B_{i-1} = \mu_C\left(-c_{i-\frac{1}{2}} + b_{i+\frac{1}{2}}\right) \\ -L_{i+\frac{1}{2}} - 2B_i - 2B_{i+1} = \dfrac{\mu_C}{2}\left(-d_{i-\frac{1}{2}} - b_{i+\frac{1}{2}} + c_{i+\frac{1}{2}} + a_{i+\frac{3}{2}}\right) \\ B_{i+1} = \dfrac{\mu_C}{2}\left(d_{i+\frac{1}{2}} - a_{i+\frac{3}{2}}\right) \end{cases} \tag{2.29}$$

As we show in appendix A.1, it comes that these systems are consistent and they give us the following formulas

$$\boxed{\begin{cases} L_{i+\frac{1}{2}} = \dfrac{\mu_C}{2}\left(-d_{i-\frac{1}{2}} + 2a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}} - c_{i+\frac{1}{2}} - 2d_{i+\frac{1}{2}} + a_{i+\frac{3}{2}}\right) \\ B_i = \dfrac{\mu_C}{2}\left(d_{i-\frac{1}{2}} - a_{i+\frac{1}{2}}\right) \end{cases}} \tag{2.30}$$

The colors are supposed to show the symmetry built into these formulas

Stencil 6 For stencils of 5 or 6 represented by (2.10), we need to add sixth-order diffusion : the Trilaplacian diffusion. In the appendix A.2 we show that we obtain the following flux-diffusion formulas

$$\boxed{\begin{cases} L_{i+\frac{1}{2}} = \dfrac{\mu_C}{2}\Big(-f_{i-\frac{3}{2}} - e_{i-\frac{1}{2}} - f_{i-\frac{1}{2}} + 3a_{i+\frac{1}{2}} + 2b_{i+\frac{1}{2}} + c_{i+\frac{1}{2}} - d_{i+\frac{1}{2}} - 2e_{i+\frac{1}{2}} - 3f_{i+\frac{1}{2}} \\ \qquad\qquad + a_{i+\frac{3}{2}} + b_{i+\frac{3}{2}} + a_{i+\frac{5}{2}}\Big) \\ B_i = \dfrac{\mu_C}{2}\Big(2f_{i-\frac{3}{2}} + e_{i-\frac{1}{2}} + 3f_{i-\frac{1}{2}} - 3a_{i+\frac{1}{2}} - b_{i+\frac{1}{2}} - 2a_{i+\frac{3}{2}}\Big) \\ T_{i+\frac{1}{2}} = \dfrac{\mu_C}{2}\Big(-f_{i-\frac{1}{2}} + a_{i+\frac{3}{2}}\Big) \end{cases}} \tag{2.31}$$

With (2.30) and (2.31), we notice that for a given stencil, the lower the order of diffusion, the closer its coefficients are to the interface $x_{i+\frac{1}{2}}$ ($L_{i+\frac{1}{2}}$ depends on many coefficients close to the interface while $T_{i+\frac{1}{2}}$ depends on 2 coefficients which correspond to points far from the interface). It can be interpreted as a reason why the Laplacian diffusion is more diffusive for the very irregular functions than the Bilaplacian or the Trilaplacian one.

2.4 Other possible approaches ?

2.4.1 The uniqueness of the coefficients

The first question about the approach I propose is about the representation of the flux we did in (2.1), (2.8) and (2.10). As we said, the coefficients $a_{i+\frac{1}{2}}$, $b_{i+\frac{1}{2}}$ ... are functions of the (u_i) values of the concerned stencil. The existence of these coefficients and of these representation of the flux is implicit, indeed all the advection that are used in solvers are written in such form. But nothing assures us that there is only one way to define these coefficients.

Stencil 2 It is possible to prove the uniqueness of this flux representation for a stencil of 2. Let us write the flux

$$F_{i+\frac{1}{2}} = \mathcal{F}(u_i, u_{i+1}) = a_{i+\frac{1}{2}}u_i + b_{i+\frac{1}{2}}u_{i+1}$$

where $a_{i+\frac{1}{2}}$ is defined by a function $a(u_i, u_{i+1})$ and the same for b. The first condition we impose is the consistency $a + b = 1$: we just have to show the uniqueness of one of the two. The second condition is that we impose to the functions a and b to be independent of any offset τ on the function u : $a(u_i + \tau, u_{i+1} + \tau) = a(u_i, u_{i+1})$. Which is equivalent to say that a is a function of only one variable : the slope $u_i - u_{i+1}$. Indeed for all $x, y \in \mathbb{R}$ if we take $\tau = -y$ we have $a(x, y) = a(x - y, y - y) = a(x - y, 0)$. Then we can write the flux like this

$$\mathcal{F}(u_i, u_{i+1}) = a(u_i - u_{i+1})u_i + b(u_i - u_{i+1})u_{i+1}$$

This writing is supposed to work for any values of u_i and u_{i+1}. If we take $u_i = x \in \mathbb{R}^*$ and $u_{i+1} = 0$, we find that

$$a(x) = \frac{\mathcal{F}(x, 0)}{x}$$

For $x = 0$, we need to impose to the flux function $\mathcal{F}$ to verify $\mathcal{F}(x, 0) \underset{x\to 0}{=} O(x)$ which is the case for all advection schemes used. With this analysis, we proved the uniqueness of the writing (2.1) for any linear and non-linear advection scheme of stencil 1 or 2.

Larger stencils The question of the uniqueness of these coefficients is more difficult for stencils larger than 2. The goal is to find some properties on the coefficients $a_{i+\frac{1}{2}}, b_{i+\frac{1}{2}}, c_{i+\frac{1}{2}}$ etc. which will impose them to be unique. Actually, I didn't achieve to find them, but I present on appendix B the ideas that I tried.

2.4.2 Different definition of Bilaplacian diffusion

In some ocean models, the Bilaplacian diffusion operator, instead of the definition (2.21) is defined by

$$\frac{\partial}{\partial x}\left(\sqrt{B}\frac{\partial^2}{\partial x^2}\left(\sqrt{B}\frac{\partial u}{\partial x}\right)\right)$$

which leads to the following quantity of diffusion

$$D^L = \frac{\Delta x^4}{\Delta t}\int_0^1 u\frac{\partial}{\partial x}\sqrt{B}\frac{\partial^2}{\partial x^2}\sqrt{B}\frac{\partial u}{\partial x}\,\mathrm{d}x = \frac{\Delta x^4}{\Delta t}\int_0^1 \left(\frac{\partial}{\partial x}\left(\sqrt{B}\frac{\partial u}{\partial x}\right)\right)\mathrm{d}x$$

This formulation is useful because it allows to use a same operator for the different diffusion orders : indeed we just have to iterate the the Laplacian diffusion operator (2.12). The work of discretization is feasible with this formulation, and for instance, we have to define the coefficients $B_{i+\frac{1}{2}}$ on the interfaces and not on the cell as previously. This is also a benefit, because all the diffusion coefficients, $L_{i+\frac{1}{2}}$ and $B_{i+\frac{1}{2}}$ are defined on the same points of the grid.

With the same approach we had previously, we end up with the following system to solve

$$\forall i \in [\![1, N-1]\!], \qquad \begin{cases} B_{i-\frac{1}{2}}B_{i+\frac{1}{2}} = \dfrac{\mu_C^2}{4}\left(d_{i-\frac{1}{2}} - a_{i+\frac{1}{2}}\right)^2 \\ L_{i-\frac{1}{2}} + 2B_{i-\frac{1}{2}} = \dfrac{\mu_C}{2}\left(a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}} - c_{i+\frac{1}{2}} - d_{i+\frac{1}{2}}\right) \end{cases}$$

But after many tries, we found out that this system has an infinite number of solution. This approach of the diffusion is not appropriate to analyze the numerical diffusion of the advection schemes.

2.5 A generalization to all orders

It is possible to generalize to any size of stencil and any order of diffusion. It needs a lot of calculation and new objects, that's why one can find all the work I have done in the appendix C.

3 An analysis of the diffusion of TVD and WENO schemes

In this section, I will present the applications of the previous results. First, in section 3.1, I will analyze the diffusion coefficients for the advection schemes we presented in section 1.3 and analyze the diffusion spatially. Then in section 3.2, I will compare study the temporal evolution of the diffusion.

3.1 Spatial analysis of the numerical diffusion

3.1.1 Local quantity of diffusion

For analyzing the diffusion spatially, we need a tool : the local quantity of diffusion. First, we can define the *local quantity of Laplacian diffusion* :

$$D^L_{i+\frac{1}{2}} = \frac{\Delta x}{\Delta t} L_{i+\frac{1}{2}} (-u_i + u_{i+1})^2 \tag{3.1}$$

this is actually an approximation

$$D^L_{i+\frac{1}{2}} \simeq \frac{\Delta x^3}{\Delta t} \left(L \left(\frac{\partial u}{\partial x} \right)^2 \right) \Bigg|_{i+\frac{1}{2}}$$

We also have

$$D^L = \sum_{i=1}^{N} D^L_{i+\frac{1}{2}}$$

Like the coefficients $L_{i+\frac{1}{2}}$, we naturally have to define this discrete diffusion on the interfaces. For the Bilaplacian diffusion, we have to define the quantity of diffusion on the cells :

$$D^B_i = \frac{\Delta x}{\Delta t} B_i (u_{i-1} - 2u_i + u_{i+1})^2 \tag{3.2}$$

This is also an approximation :

$$D^B_i = \frac{\Delta x^5}{\Delta t} \left(B \left(\frac{\partial^2 u}{\partial x^2} \right)^2 \right) \Bigg|_{i+\frac{1}{2}}$$

And finally the Trilaplacian local diffusion :

$$D^T_i = \frac{\Delta x}{\Delta t} T_{i+\frac{1}{2}} (-u_{i-1} + 3u_i - 3u_{i+1} + u_{i+2})^2 \tag{3.3}$$

3.1.2 Linear schemes

First, when we apply the formulas (2.28), (2.30) and (2.31) to the schemes UP1, CEN2, UP3 and UP5 ((1.6), (1.7), (1.8) and (1.9)), we find that UP1, UP3 and UP5 have respectively diffusion of order 2, 4 and 6, and that CEN2 is not diffusive. For instance we have $L^{UP3}_{i+\frac{1}{2}} = 0$ and $B^{UP3}_i \neq 0$. Furthermore, we show that these schemes have constant diffusion coefficients : for all i

$$L^{UP1}_{i+\frac{1}{2}} = \frac{\mu_C}{2}, \qquad B^{UP3}_i = \frac{\mu_C}{4} \qquad \text{and} \qquad T_{i+\frac{1}{2}} = \frac{\mu_C}{24}$$

It is not surprising considering that these schemes are linear. Even if the coefficients of diffusion are constant, it doesn't mean that the quantity of diffusion is constant. Indeed, with these coefficients we can process the quantity of diffusion locally with (3.1).

3.1.3 TVD schemes

The stencil of TVD schemes (1.10), is 3, because Φ_i depends on u_{i-1}, u_i and u_{i+1}. But we can represent the flux only with two coefficients like in (2.1), which allows to apply the formula (2.28) :

$$L_{i+\frac{1}{2}} = \frac{\mu_C}{2}(1 - \Phi_i)$$

The TVD schemes are diffusive at the order 2, and their coefficients of diffusion are directly linked to the flux limiter Φ_i. Then we can process the quantities of diffusion $D^L_{i+\frac{1}{2}}$, what we can see on figure 4. On this figure, we chose a time close to 0, because the rectangle function is quickly deformed by the numerical diffusion, and while it is deformed differently by the different schemes, comparing the snapshot of the diffusion on one large time t doesn't have a lot of sense. Here we can note that SuperBee scheme and Van-Leer scheme have anti-diffusion effects because of their negative coefficients on the right of the rectangle. We also notice that at this specific time, the quantity of diffusion of the different flux limiters seems to be equivalent.

3.1.4 WENO schemes

The WENO schemes have higher stencils. The complete formulas of the coefficient of diffusion don't give us a lot information. But for WENO3, we can accomplish the following approximations

$$\begin{cases} B_i \simeq \dfrac{\mu_C}{4 + 2\theta_i^4} \\ L_{i+\frac{1}{2}} \simeq \dfrac{\mu_C}{2}\left(\dfrac{1}{2+\theta_i^4} - \dfrac{1}{2+\theta_{i+1}^4}\right) \end{cases}$$

We can interpret this by the fact that the diffusion might be mostly Bilaplacian where the function is regular and Laplacian where the function is irregular.

We didn't find an equivalent formulation for the WENO5 scheme. But we can say that this scheme mix Laplacian, Bilaplacian and Trilaplacian numerical diffusion, what we can see on figure 5. This figure allows to visualize the quantities of the different orders of diffusion. The Trilaplacian diffusion is the weakest one, we can also see that the three diffusion are stronger where the slope of the function u is higher.

3.2 Temporal analysis of the numerical diffusion

The last figures only shew snapshots of the diffusion quantities. But another important parameter of the diffusion is how its quantity evolves with time. That's why we computed the total quantity of diffusion on one section. The two figures below show some examples of these analyze.

In figure 3.2, we note that here the diffusion of UP3 is only Bilaplacian and UP5 is only Trilaplacian. We notice that for this periodic initial condition, the WENO3 scheme is not necessarily better than the TVD schemes. We see that its Laplacian diffusion is larger than its Bilaplacian one. And we observe what we saw in part 1 : Minmod is the more diffusive TVD scheme and Van-Leer the less one cause of its anti-diffusive effects. Finally, we can say that WENO3 and SuperBee have a very low diffusion after time $t = 0.4$, for WENO3 it is because the diffusion at the beginning was enough to flatten the solution totally

In figure 3.2 we notice that even if WENO5 has 3 types of diffusion, it is way less diffusive than the WENO3 scheme, this scheme has real benefits compared to the TVD schemes. And we also notice that even if UP5 is not very diffusive, it is not accurate because extrema are formed in the function.

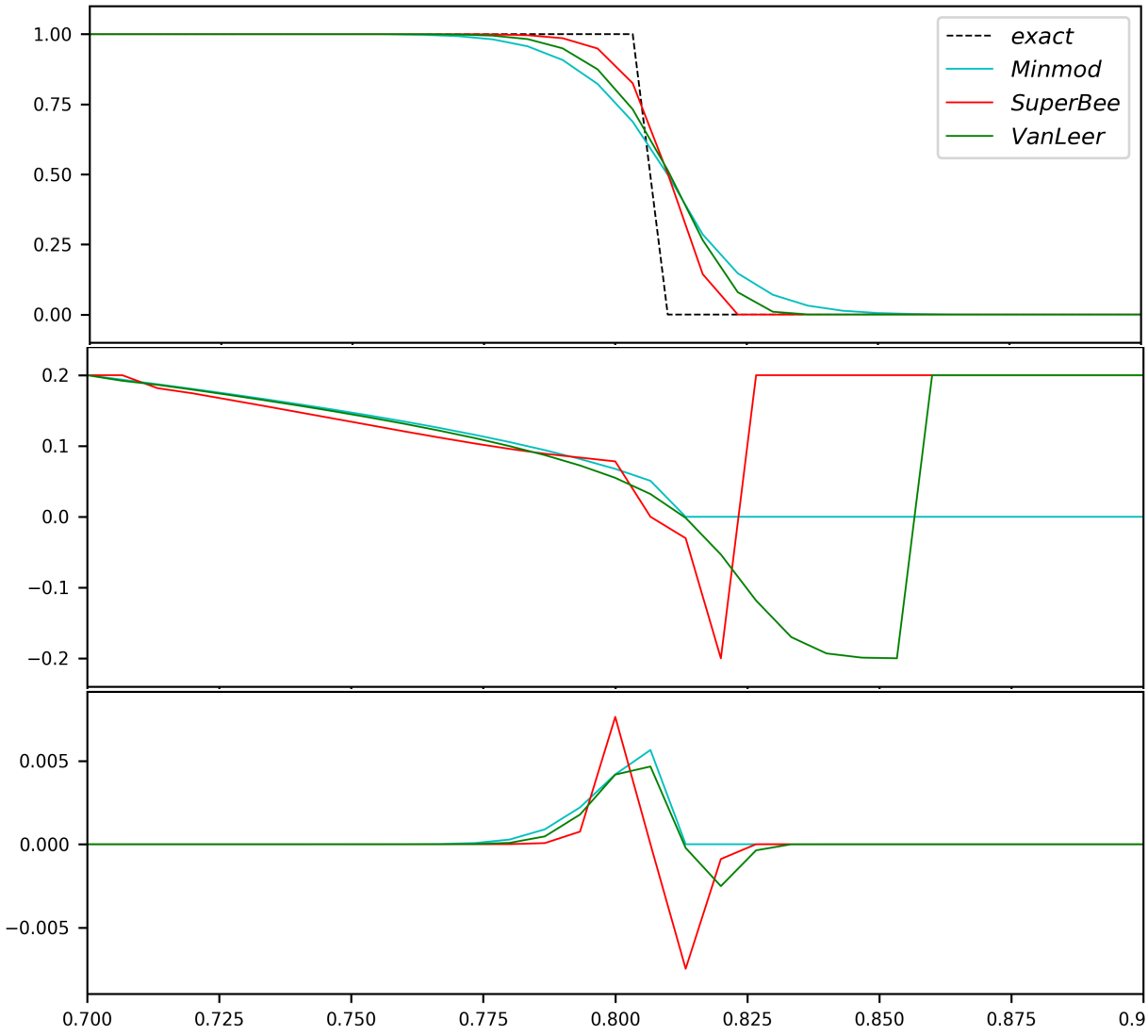

Figure 4: This figure shows the solution (up), the coefficients of diffusion $L_{i+\frac{1}{2}}$ (middle) and the local quantity diffusion $D^L_{i+\frac{1}{2}}$ (3.1) (down) for a section of the rectangle function at the time $t = 0.1$ for different TVD schemes. The other parameters are the same than in figure 1

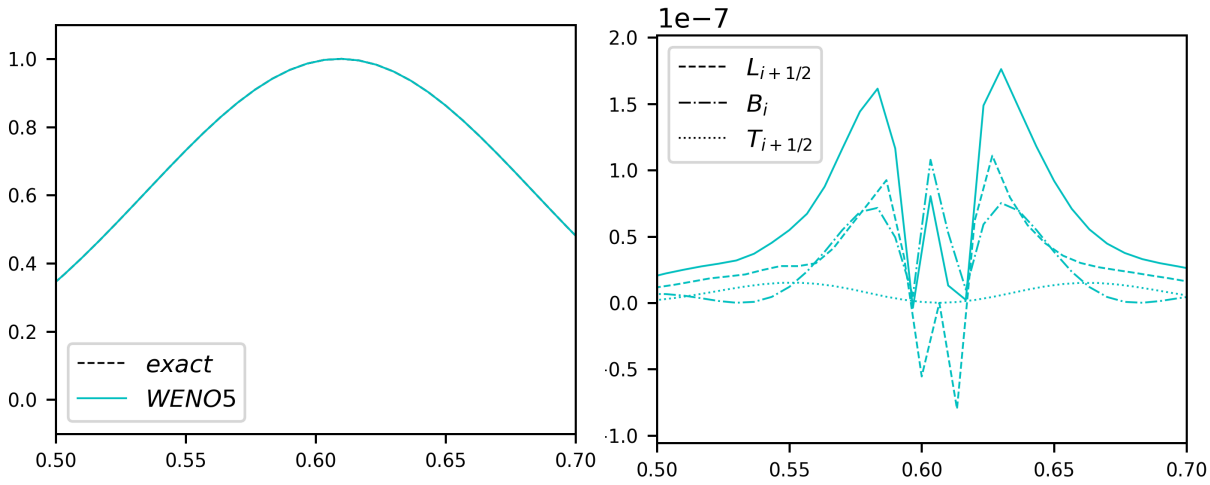

Figure 5: This figure shows the solution u_i(left) and the local quantity diffusion $D^L_{i+\frac{1}{2}}$, D^B_i and $D^T_{i+\frac{1}{2}}$ ((3.1), (3.2) and (3.3)) (right) for the middle section of a Gaussian function at the time $t = 0.1$ for the WENO5 scheme. The other parameters are the same than in figure 1

Conclusion and perspectives

During my internship I was interested in the discretization of advection terms in the context of ocean modeling with the aim of estimating their impact in terms of energy diffusion. In the part 1 I introduced standard advection schemes used in ocean models and based on the finite volume and finite differences methods. In particular, the TVD and WENO schemes which minimize the creation of local extrema are presented. But these schemes involve the appearance of numerical diffusion which is problematic because it can artificially mix the stratified layers of the ocean. It is essential to quantify this numerical diffusion to make sure it does not exceed the physical diffusion. In [Der22], the numerical diffusion of the TVD schemes was studied by manipulating the fluxes by hand. To generalize this approach, we initially thought, as suggested by [BSA21], that we could use neural networks. But it turned out that the mathematical framework developed in part 2 of the report was sufficiently satisfactory to meet the objectives of the internship. As a reminder, in this part, we made a connection between the symmetrical part of the gradient of the flux, and the generic representation of diffusion, which is naturally symmetrical. This work leads to the formulas (2.28), (2.30) and (2.31) which express the diffusion coefficients as a function of the flux coefficients. Finally, in part 3, we apply this generic analysis to the TVD and WENO schemes. We noticed that different orders of diffusion can coexist for a same scheme, and that the different schemes have various behaviors according to the initial conditions. This numerical analysis is a new tool that we can take into account when choosing an advection scheme to put alongside the cost of calculation of the accuracy for example.

The next step of this internship could be to use this tool to analyze the numerical diffusion inherent to some real ocean model solvers. For instance, we could take some outputs of the CROCO model, which is developed in part by the AIRSEA team where I work, to quantify a posteriori the numerical diffusion it has induced to the simulations. In addition, there's still a lot of work to be done to apply this work in 3 dimensions. Indeed, as shown in [Lem+12], the stability and accuracy of the diffusion operators in ocean models is important but more importantly the orientation of the diffusion operator in 3D has great impact on the quality of the solution (the diffusion should be along surfaces of constant density). This should be taken into account in any 3-dimensional analysis

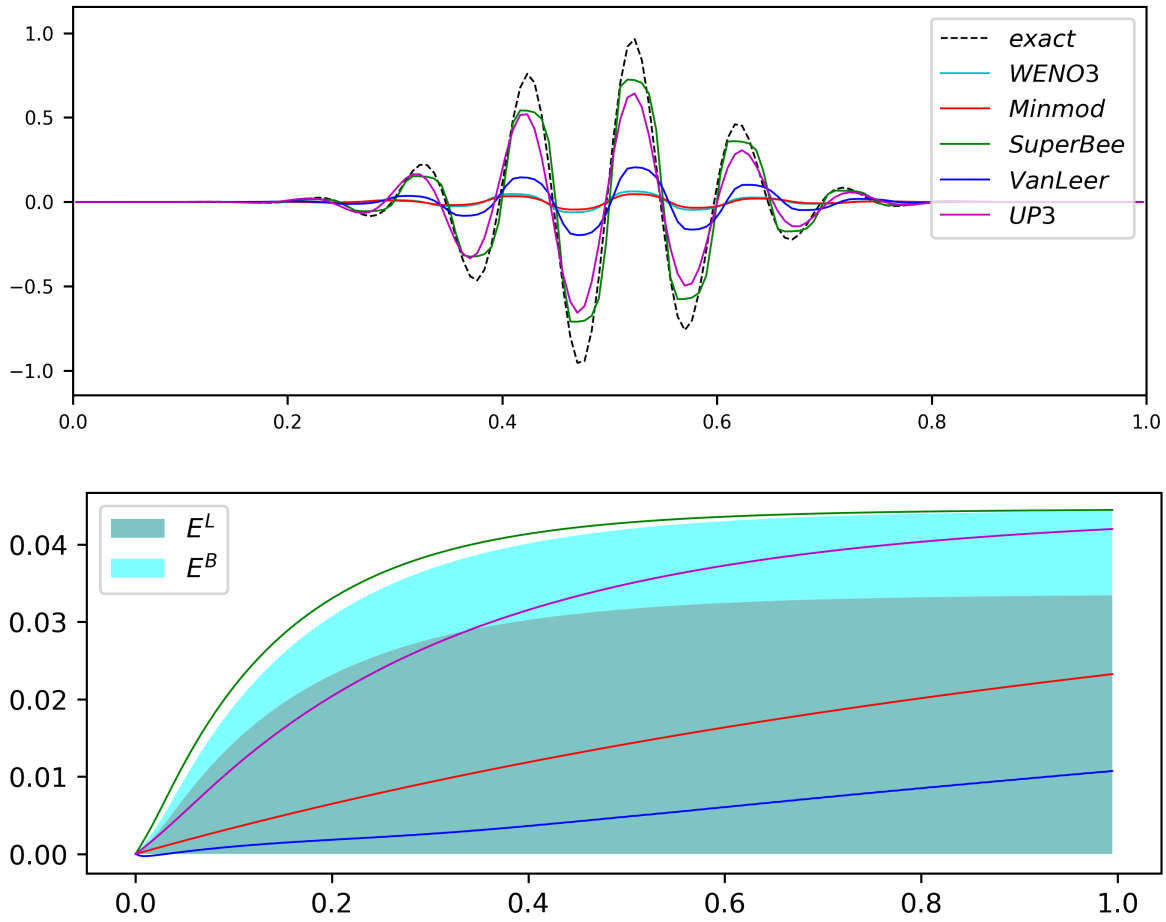

Figure 6: This figure show the evolution with time of the total quantity of diffusion integrated on the represented section for the WENO3 and the TVD schemes. The colored areas represent the different diffusion of WENO3. The colored areas represent the different diffusions of WENO5.

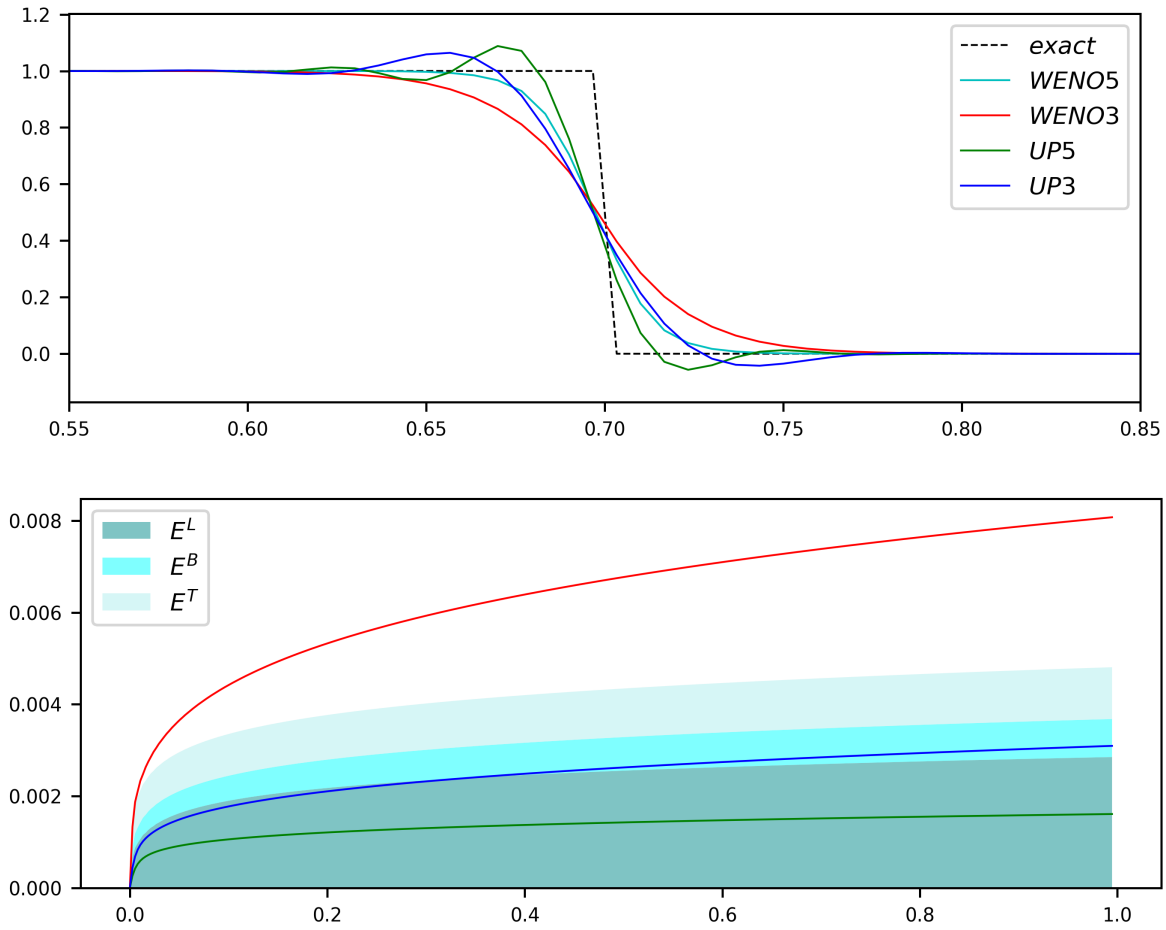

Figure 7: This figure show the evolution with time of the total quantity of diffusion integrated on the represented section for the WENO3 and the WENO5 schemes. The colored areas represent the different diffusion of WENO5.

of the numerical diffusion as the orientation of the diffusion is a key aspect, not only its intensity.

A Calculations for Bilaplacian and Trilaplacian diffusion

In this first appendix, I present the calculations used in part 2 for the Bilaplacian diffusion (appendix A.1) and the Trilaplacian diffusion ((appendix A.2).

A.1 Bilaplacian diffusion

The operator Let us get back to the the definition of the Bilaplacian diffusion operator in (2.21)

$$\frac{\partial^2}{\partial x^2}\left(B\frac{\partial^2 u}{\partial x^2}\right)$$

As in (2.13), we can define the *global quantity of Bilaplacian diffusion* D^B by an average of this operator :

$$D^B = \frac{\Delta x^4}{\Delta t}\int_0^1 u\frac{\partial^2}{\partial x^2}\left(B\frac{\partial^2 u}{\partial x^2}\right)\mathrm{d}x$$

The sign and the multiplicative coefficients are chosen to make D^B positive when B and to make it adimensional. A first integration by part gives us

$$D^B = -\frac{\Delta x^4}{\Delta t}\int_0^1 \frac{\partial u}{\partial x}\frac{\partial}{\partial x}\left(B\frac{\partial^2 u}{\partial x^2}\right)\mathrm{d}x$$

and a second one

$$D^B = \frac{\Delta x^4}{\Delta t}\int_0^1 B\left(\frac{\partial^2 u}{\partial x^2}\right)^2\mathrm{d}x$$

These integrations by part need the hypothesis that u, $\dfrac{\partial u}{\partial x}$ and B are periodic on $[0,1]$.

The discretization To approximate this integral, we need an approximation of the second derivative of u by a centered scheme, which is doable on the cells and not the interfaces :

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_i = \frac{u_{i-1} - 2u_i + u_{i+1}}{\Delta x^2} \tag{A.1}$$

That's why we are obliged to discretize the coefficients on the cells : $B_i \simeq B(x_i)$. Let us discretize the integral and do the equivalent of the integration by part in the discretized world :

$$
\begin{aligned}
D^B &\simeq \frac{\Delta x^4}{\Delta t}\sum_{i=1}^{N}\Delta x B_i \left.\frac{\partial^2 u}{\partial x^2}\right|_i^2 \\
\frac{\Delta t}{\Delta x}D^B &= \sum_{i=1}^{N} B_i(u_{i-1}-2u_i+u_{i+1})^2 \\
&= \sum_{i=1}^{N} u_{i-1}B_i(u_{i-1}-2u_i+u_{i+1}) - 2\sum_{i=1}^{N} u_i B_i(u_{i-1}-2u_i+u_{i+1}) \\
&\qquad\qquad + \sum_{i=1}^{N} u_{i+1}B_i(u_{i-1}-2u_i+u_{i+1}) \\
&= \sum_{i=0}^{N-1} u_i B_{i+1}(u_i-2u_{i+1}+u_{i+2}) - 2\sum_{i=1}^{N} u_i B_i(u_{i-1}-2u_i+u_{i+1}) \\
&\qquad\qquad + \sum_{i=2}^{N+1} u_i B_{i-1}(u_{i-2}-2u_{i-1}+u_i)
\end{aligned}
$$

with the hypothesis that (B_i) and (u_i) are periodic families for i, we have

$$
\begin{aligned}
\frac{\Delta t}{\Delta x}D^B &= \sum_{i=1}^{N} u_i B_{i+1}(u_i-2u_{i+1}+u_{i+2}) - 2\sum_{i=1}^{N} u_i B_i(u_{i-1}-2u_i+u_{i+1}) \\
&\qquad\qquad + \sum_{i=1}^{N} u_i B_{i-1}(u_{i-2}-2u_{i-1}+u_i) \\
&= \sum_{i=1}^{N} u_i \Big[B_{i-1}u_{i-2} - 2(B_{i-1}+B_i)u_{i-1} + (B_{i-1}+4B_i+B_{i+1})u_i \\
&\qquad\qquad - 2(B_i+B_{i+1})u_{i+1} + B_{i+1}u_{i+2} \Big] \qquad \text{(A.2)}
\end{aligned}
$$

The matrix This expansions leads us to introduce the *Bilaplacian diffusion matrix* $\mathbb{D}^B$ defined by its lines in (2.23)

$$
\underset{\qquad\qquad\qquad\qquad}{\boldsymbol{e}_i^\top \mathbb{D}^B = \big(0\cdots\cdots} \underset{\substack{\uparrow\\ i-2}}{B_{i-1}} \quad \underset{\substack{\uparrow\\ i-1}}{-2B_{i-1}-2B_i} \quad \underset{\substack{\uparrow\\ i}}{B_{i-1}+4B_i+B_{i+1}} \quad \underset{\substack{\uparrow\\ i+1}}{-2B_i-2B_{i+1}} \quad \underset{\substack{\uparrow\\ i+2}}{B_{i+1}} \underset{\qquad\qquad\qquad\qquad}{\cdots\cdots 0\big)}
$$

So that

$$
D^B = \frac{\Delta x}{\Delta t}\langle U, \mathbb{D}^B U\rangle
$$

We observe that this matrix is symmetric, we will give a general proof of this on the appendices C. As we did with (2.19) we can introduce the matrix $\mathbb{C}^B$ by

$$
\mathbb{C}^B = \begin{pmatrix} C^B_{-1,-2} & C^B_{-1,-1} & C^B_{-1,0} & C^B_{-1,+1} & C^B_{-1,+2} \\ C^B_{0,-2} & C^B_{0,-1} & C^B_{0,0} & C^B_{0,+1} & C^B_{0,+2} \\ C^B_{+1,-2} & C^B_{+1,-1} & C^B_{+1,0} & C^B_{+1,0} & C^B_{+1,+2} \end{pmatrix} \begin{matrix} \leftarrow B_{i-1} \\ \leftarrow B_i \\ \leftarrow B_{i+1} \end{matrix} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & -2 & 4 & -2 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix} \qquad \text{(A.3)}
$$

$$
\begin{matrix} \uparrow \\ u_{i-2} \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i-1} \end{matrix} \quad \begin{matrix} \uparrow \\ u_i \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i+1} \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i+2} \end{matrix}
$$

so that for $i \in [\![1, N]\!]$ and $s \in [\![-2, 2]\!]$,

$$\mathbb{D}^B_{i,i+s} = \sum_{l=-1}^{1} C^B_{l,s} B_{i+l}$$

which is an equivalent formulation as (2.20), but with the whole numbers and not half-integers.

The systems This expression of $\mathbb{D}^B$ leads us to write the system (2.29) that we recall

$$\forall i \in [\![1, N]\!], \quad \begin{cases} B_{i-1} = \frac{\mu_C}{2}\left(d_{i-\frac{3}{2}} - a_{i-\frac{1}{2}}\right) \\ -L_{i-\frac{1}{2}} - 2B_{i-1} - 2B_i = \frac{\mu_C}{2}\left(-d_{i-\frac{3}{2}} - b_{i-\frac{1}{2}} + c_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\right) \\ L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}} + B_{i+1} + 4B_i + B_{i-1} = \mu_C\left(-c_{i-\frac{1}{2}} + b_{i+\frac{1}{2}}\right) \\ -L_{i+\frac{1}{2}} - 2B_i - 2B_{i+1} = \frac{\mu_C}{2}\left(-d_{i-\frac{1}{2}} - b_{i+\frac{1}{2}} + c_{i+\frac{1}{2}} + a_{i+\frac{3}{2}}\right) \\ B_{i+1} = \frac{\mu_C}{2}\left(d_{i+\frac{1}{2}} - a_{i+\frac{3}{2}}\right) \end{cases}$$

First we notice that the first line and the last lines are the same for two different values of i : we can keep only one of the two (if we consider the (B_i) and the coefficients $\left(a_{i+\frac{1}{2}}\right)$, $\left(a_{i+\frac{1}{2}}\right)$ etc. periodic for $i \in [\![1, N]\!]$ we don't considerate the values). We have the same result for the second and the fourth line. We also notice that when we add the five lines together, we obtain $0 = 0$ so we can ignore the central line. We end with the two last lines to study. The last one at the range $i - 1$ gives

$$\forall i \in [\![1, N]\!], \quad B_i = \frac{\mu_C}{2}\left(d_{i-\frac{1}{2}} - a_{i+\frac{1}{2}}\right)$$

When we replace the values of B_i and B_{i+1} in the fourth line with this formula we get

$$\forall i \in [\![1, N]\!], \quad L_{i+\frac{1}{2}} = \frac{\mu_C}{2}\left(a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}} - c_{i+\frac{1}{2}} - d_{i+\frac{1}{2}}\right)$$

We w=show that this set of systems is consistent and we get the formulas (2.30).

A.2 Trilaplacian diffusion

The operator We gave the definition of the Trilaplacian diffusion operator in (2.24)

$$\frac{\partial^3}{\partial x^3}\left(T\frac{\partial^3 u}{\partial x^3}\right)$$

we can define from this the *global quantity of Trilaplacian diffusion* D^T by

$$D^T = -\frac{\Delta x^6}{\Delta t}\int_0^1 u\frac{\partial^3}{\partial x^3}\left(T\frac{\partial^3 u}{\partial x^3}\right)\mathrm{d}x$$

The sign and the multiplicative coefficients are chosen to make D^T positive when T and to make it adimensional. A first integration by part gives us

$$D^T = -\frac{\Delta x^6}{\Delta t}\int_0^1 \frac{\partial u}{\partial x}\frac{\partial^2}{\partial x^2}\left(T\frac{\partial^3 u}{\partial x^3}\right)\mathrm{d}x$$

a second one

$$D^T = \frac{\Delta x^6}{\Delta t}\int_0^1 \frac{\partial^2 u}{\partial x^2}\frac{\partial}{\partial x}\left(T\frac{\partial^3 u}{\partial x^3}\right)\mathrm{d}x$$

aand a last one

$$D^T = \frac{\Delta x^6}{\Delta t}\int_0^1 T\left(\frac{\partial^3 u}{\partial x^3}\right)^2\mathrm{d}x$$

These integrations by part need the hypothesis that $u, \dfrac{\partial u}{\partial x}, \dfrac{\partial^2 u}{\partial x^2}$ and T are periodic on $[0,1]$.

The discretization To approximate this integral, we need an approximation of the third derivative of u by a centered scheme, which is doable on the interfaces :

$$\left.\frac{\partial^3 u}{\partial x^3}\right|_{i+\frac{1}{2}} = \frac{-u_{i-1}+3u_i-3u_{i+1}+u_{i+2}}{\Delta x^3} \tag{A.4}$$

That's why we are obliged to discretize the coefficients on the interfaces : $T_{i+\frac{1}{2}} \simeq B\big(x_{i+\frac{1}{2}}\big)$. Let us discretize the integral and do the equivalent of the integration by part in the discretized world :

$$D^T \simeq \frac{\Delta x^6}{\Delta t}\sum_{i=1}^{N}\Delta x T_{i+\frac{1}{2}}\left.\frac{\partial^3 u}{\partial x^3}\right|^2_{i+\frac{1}{2}}$$

with a discrete integration by part

$$\begin{aligned}
\frac{\Delta t}{\Delta x}D^T &= -\sum_{i=1}^{N} u_{i-1}T_{i+\frac{1}{2}}(-u_{i-1}+3u_i-3u_{i+1}+u_{i+2}) + 3\sum_{i=1}^{N} u_i T_{i+\frac{1}{2}}(-u_{i-1}+3u_i-3u_{i+1}+u_{i+2})\\
&\quad - 3\sum_{i=1}^{N} u_{i+1}T_{i+\frac{1}{2}}(-u_{i-1}+3u_i-3u_{i+1}+u_{i+2}) + \sum_{i=1}^{N} u_{i+2}T_{i+\frac{1}{2}}(-u_{i-1}+3u_i-3u_{i+1}+u_{i+2})\\
\frac{\Delta t}{\Delta x}D^T &= -\sum_{i=0}^{N-1} u_i T_{i+\frac{3}{2}}(-u_i+3u_{i+1}-3u_{i+2}+u_{i+3}) + 3\sum_{i=1}^{N} u_i T_{i+\frac{1}{2}}(-u_{i-1}+3u_i-3u_{i+1}+u_{i+2})\\
&\quad - 3\sum_{i=2}^{N+1} u_i T_{i-\frac{1}{2}}(-u_{i-2}+3u_{i-1}-3u_i+u_{i+1}) + \sum_{i=3}^{N+2} u_i T_{i-\frac{3}{2}}(-u_{i-3}+3u_{i-2}-3u_{i-1}+u_i)
\end{aligned}$$

because $\big(T_{i+\frac{1}{2}}\big)$ and u_i are periodic for i :

$$\begin{aligned}
\frac{\Delta t}{\Delta x}D^T &= -\sum_{i=1}^{N} u_i T_{i+\frac{3}{2}}(-u_i+3u_{i+1}-3u_{i+2}+u_{i+3}) + 3\sum_{i=1}^{N} u_i T_{i+\frac{1}{2}}(-u_{i-1}+3u_i-3u_{i+1}+u_{i+2})\\
&\quad - 3\sum_{i=1}^{N} u_i T_{i-\frac{1}{2}}(-u_{i-2}+3u_{i-1}-3u_i+u_{i+1}) + \sum_{i=1}^{N} u_i T_{i-\frac{3}{2}}(-u_{i-3}+3u_{i-2}-3u_{i-1}+u_i)\\
&= \sum_{i=1}^{N} u_i\Big[-T_{i-\frac{3}{2}}u_{i-3} + 3\big(T_{i-\frac{3}{2}}+T_{i-\frac{1}{2}}\big)u_{i-2} - 3\big(T_{i-\frac{3}{2}}+3T_{i-\frac{1}{2}}+T_{i+\frac{1}{2}}\big)u_{i-1}\\
&\qquad + \big(T_{i-\frac{3}{2}}+9T_{i-\frac{1}{2}}+9T_{i+\frac{1}{2}}+T_{i+\frac{3}{2}}\big)u_i - 3\big(T_{i-\frac{1}{2}}+3T_{i+\frac{1}{2}}+T_{i+\frac{3}{2}}\big)u_{i+1}\\
&\qquad + 3\big(T_{i+\frac{1}{2}}+T_{i+\frac{3}{2}}\big)u_{i+2} - T_{i+\frac{3}{2}}u_{i+3}\Big]
\end{aligned}$$

The matrix This expansion leads us to introduce the *Trilaplacian diffusion matrix* $\mathbb{D}^T$ defined by its lines :

$$\boldsymbol{e}_i^\top \mathbb{D}^T = \Big(\underset{\downarrow}{i-3} \quad \underset{\downarrow}{i-2} \quad \underset{\downarrow}{i-1} \quad \underset{\downarrow}{i} \quad \underset{\uparrow}{i+1} \quad \underset{\uparrow}{i+2} \quad \underset{\uparrow}{i+3} \Big)$$
$$\begin{pmatrix} -T_{i-\frac{3}{2}} & 3T_{i-\frac{3}{2}} + 3T_{i-\frac{1}{2}} & -3T_{i-\frac{3}{2}} - 9T_{i-\frac{1}{2}} - 3T_{i+\frac{1}{2}} & T_{i-\frac{3}{2}} + 9T_{i-\frac{1}{2}} + 9T_{i+\frac{1}{2}} + T_{i+\frac{3}{2}} \\ & & -3T_{i-\frac{1}{2}} - 9T_{i+\frac{1}{2}} - 3T_{i+\frac{3}{2}} & 3T_{i+\frac{1}{2}} + 3T_{i+\frac{3}{2}} \end{pmatrix} \quad -T_{i+\frac{3}{2}} \Big)$$
(A.5)

So that

$$D^T = \frac{\Delta x}{\Delta t} \langle U, \mathbb{D}^T U \rangle$$

We observe that this matrix is also symmetric. We can introduce the matrix $\mathbb{C}^B$ by

$$\mathbb{C}^B = \begin{pmatrix} C^T_{-\frac{3}{2},-3} & C^T_{-\frac{3}{2},-2} & C^T_{-\frac{3}{2},-1} & C^T_{-\frac{3}{2},0} & C^T_{-\frac{3}{2},+1} & C^T_{-\frac{3}{2},+2} & C^T_{-\frac{3}{2},+3} \\ C^T_{-\frac{1}{2},-3} & C^T_{-\frac{1}{2},-2} & C^T_{-\frac{1}{2},-1} & C^T_{-\frac{1}{2},0} & C^T_{-\frac{1}{2},+1} & C^T_{-\frac{1}{2},+2} & C^T_{-\frac{1}{2},+3} \\ C^T_{+\frac{1}{2},-3} & C^T_{+\frac{1}{2},-2} & C^T_{+\frac{1}{2},-1} & C^T_{+\frac{1}{2},0} & C^T_{+\frac{1}{2},+1} & C^T_{+\frac{1}{2},+2} & C^T_{+\frac{1}{2},+3} \\ C^T_{+\frac{3}{2},-3} & C^T_{+\frac{3}{2},-2} & C^T_{+\frac{3}{2},-1} & C^T_{+\frac{3}{2},0} & C^T_{+\frac{3}{2},+1} & C^T_{+\frac{3}{2},+2} & C^T_{+\frac{3}{2},+3} \end{pmatrix} \begin{matrix} \leftarrow T_{i-\frac{3}{2}} \\ \leftarrow T_{i-\frac{1}{2}} \\ \leftarrow T_{i+\frac{1}{2}} \\ \leftarrow T_{i+\frac{3}{2}} \end{matrix}$$
$$\begin{matrix} \uparrow \\ u_{i-3} \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i-2} \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i-1} \end{matrix} \quad \begin{matrix} \uparrow \\ u_i \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i+1} \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i+2} \end{matrix} \quad \begin{matrix} \uparrow \\ u_{i-3} \end{matrix}$$
$$= \begin{pmatrix} -1 & 3 & -3 & 3 & 0 & 0 & 0 \\ 0 & 3 & -9 & 9 & -3 & 0 & 0 \\ 0 & 0 & -3 & 9 & -9 & 3 & 0 \\ 0 & 0 & 0 & 3 & -3 & 3 & -1 \end{pmatrix}$$
(A.6)

so that for $i \in [\![1, N]\!]$ and $s \in [\![-3, 3]\!]$,

$$\mathbb{D}^B_{i,i+s} = \sum_{l=-\frac{3}{2}}^{\frac{3}{2}} C^T_{l,s} T_{i+l}$$

The system As we did for a stencil of 4, we represent the variation of energy with three types of diffusion :

$$D^L + D^B + D^L = \frac{\mathrm{d}E}{\mathrm{d}t}$$

In terms of matrix, we have

$$\mathbb{D}^L + \mathbb{D}^B + \mathbb{D}^T = \mu_C \mathbb{S}$$

where $\mathbb{S}$ is the matrix of the symmetric part of the differences of the fluxes given by (2.11). The equations (2.17), (2.23) and (A.5) give us the set of systems for all $i \in [\![1, N]\!]$

$$\begin{cases} -T_{i-\frac{3}{2}} = \frac{\mu_C}{2}\left(f_{i-\frac{5}{2}} - a_{i-\frac{1}{2}}\right) \\ B_{i-1} + 3T_{i-\frac{3}{2}} + 3T_{i-\frac{1}{2}} = \frac{\mu_C}{2}\left(-f_{i-\frac{5}{2}} + e_{i-\frac{3}{2}} - b_{i-\frac{1}{2}} + a_{i+\frac{1}{2}}\right) \\ -L_{i-\frac{1}{2}} - 2B_{i-1} - 2B_i - 3T_{i-\frac{3}{2}} - 9T_{i-\frac{1}{2}} - 3T_{i+\frac{1}{2}} = \frac{\mu_C}{2}\left(-e_{i-\frac{3}{2}} - c_{i-\frac{1}{2}} + d_{i-\frac{1}{2}} + b_{i+\frac{1}{2}}\right) \\ L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}} + B_{i+1} + 4B_i + B_{i-1} + T_{i-\frac{3}{2}} + 9T_{i-\frac{1}{2}} + 9T_{i+\frac{1}{2}} + T_{i+\frac{3}{2}} = \mu_C\left(-d_{i-\frac{1}{2}} + c_{i+\frac{1}{2}}\right) \\ -L_{i+\frac{1}{2}} - 2B_i - 2B_{i+1} - 3T_{i-\frac{1}{2}} - 9T_{i+\frac{1}{2}} - 3T_{i+\frac{3}{2}} = \frac{\mu_C}{2}\left(-e_{i-\frac{1}{2}} - c_{i+\frac{1}{2}} + d_{i+\frac{1}{2}} + b_{i+\frac{3}{2}}\right) \\ B_{i+1} + 3T_{i+\frac{1}{2}} + 3T_{i+\frac{3}{2}} = \frac{\mu_C}{2}\left(-f_{i-\frac{1}{2}} + e_{i+\frac{1}{2}} - b_{i+\frac{3}{2}} + a_{i+\frac{5}{2}}\right) \\ -T_{i+\frac{3}{2}} = \frac{\mu_C}{2}\left(f_{i+\frac{1}{2}} - a_{i+\frac{5}{2}}\right) \end{cases}$$

As in the last section, we notice that it's a consistent set of systems : the lines 1 and 7, 2 and 6 and 3 and 5 are the same ones at different ranks of i, and the sum of all the lines of each systems is equivalent to $0 = 0$. Although we can see that we can calculate $T_{i+\frac{1}{2}}$ for all i with the last line, then it allows us to calculate B_i for all i with the line 6 and finally we can have $L_{i+\frac{1}{2}}$ for all i with the line 5. We find like this the formulas (2.31).

B The uniqueness of the flux coefficients

In section 2.4, we shew that it is possible to show the uniqueness of the linear flux decomposition for a stencil $s = 2$. On the first subappendix B.1, I try to apply the work I did on a stencil of $s = 2$ for higher stencils with special conditions. And then in the subappendix B.2 I explore the possibility of the uniqueness of the diffusion coefficients without the uniqueness of the flux ones. Unfortunately, none of these works achieved to a convincing conclusion.

B.1 The flux coefficients

Here we show the calculation we did for a stencil of $s = 4$, but we obtain the same kind results for $s = 3$. We follow the idea of the subsection 2.4.1 but with a higher stencil.

Flux writing We suppose that we can write the flux at the interface $x_{i+\frac{1}{2}}$ like

$$F_{i+\frac{1}{2}} = a_{i+\frac{1}{2}}u_{i-1} + b_{i+\frac{1}{2}}u_i + c_{i+\frac{1}{2}}ui + 1 + d_{i+\frac{1}{2}}u_{i+2}$$

for all $i \in [\![1, N]\!]$, where $a_{i+\frac{1}{2}}, b_{i+\frac{1}{2}}, c_{i+\frac{1}{2}}$ and $d_{i+\frac{1}{2}}$ are functions describe by the functions $\mathscr{A}, \mathscr{B}, \mathscr{C}$ and $\mathscr{D}$ evaluated on u_{i-1}, u_i, u_{i+1} and u_{i+2}, and where they verify the relation of consistency

$$a_{i+\frac{1}{2}} + b_{i+\frac{1}{2}} + c_{i+\frac{1}{2}} + d_{i+\frac{1}{2}} = 1$$

To work on a possible uniqueness of this writing, we can suppose that there exists other coefficients which verify it. If we study the difference between this two sets of coefficients, and if we redefine $\mathscr{A}, \mathscr{B}, \mathscr{C}$ and $\mathscr{D}$ by this difference, we have to study the following relations : for all $u_{i-1}, u_i, u_{i+1}, u_{i+2} \in \mathbb{R}$,

$$\begin{aligned}\mathcal{A}(u_{i-1}, u_i, u_{i+1}, u_{i+2})u_{i-1} + \mathcal{B}(u_{i-1}, u_i, u_{i+1}, u_{i+2})u_i \\ \mathcal{C}(u_{i-1}, u_i, u_{i+1}, u_{i+2})u_{i+1} + \mathcal{D}(u_{i-1}, u_i, u_{i+1}, u_{i+2})u_{i+2} = 0\end{aligned}$$

with $\mathscr{A} + \mathscr{B} + \mathscr{C} + \mathscr{D} = 0$.

Independence of an offset The first condition is to make these coefficients independents of any offset τ on the function (u_i), which has sense because the energy which cross the interface $x_{i+\frac{1}{2}}$ that the flux represent, shouldn't depend on an offset of the function. As earlier, if we take $\tau = u_{i-1}$ we show that

$$a_{i+\frac{1}{2}} = a(u_{i-1}, u_i, u_{i+1}, u_{i+2}) = \mathscr{A}(0, u_i - u_{i-1}, u_{i+1} - u_{i-1}, u_{i+2} - u_{i-1})$$

At this point, let us rename $(u_{i-1}, u_i, u_{i+1}, u_{i+2})$ by t, u, v, w. If rearrange the variables, we show that a is a variable of the three slopes $u - t, v - u$ and $w - v$:

$$a_{i+\frac{1}{2}} = \mathscr{A}(u - t, v - u, w - v)$$

We have the same result for b, c and d. Then we are working on the following relation : $\forall t, u, v, w \in \mathbb{R}$,

$$\mathscr{A}(u-t, v-u, w-v)t+\mathscr{B}(u-t, v-u, w-v)u+\mathscr{C}(u-t, v-u, w-v)v+\mathscr{D}(u-t, v-u, w-v)w = 0 \quad \text{(B.1)}$$

with $\mathscr{A} + \mathscr{B} + \mathscr{C} + \mathscr{D} = 0$, the goal being to prove that $\mathscr{A} = \mathscr{B} = \mathscr{C} = \mathscr{D} = 0$.

Analyze on subspaces Let us apply (B.1) for three zeros in the set (t, u, v, w). For example, with $(-x, 0, 0, 0)$ where $x \neq 0$, we obtain $-\mathscr{A}(x, 0, 0)x = 0$. Which means $\mathscr{A}(x, 0, 0) = 0$ for all $x \in \mathbb{R}^*$. With the same reasoning, we have some relations like $\mathscr{B}(x, 0, 0) = 0$ or $\mathscr{B}(0, y, 0) = 0$ for example. We can show that the functions $\mathscr{A}, \mathscr{B}, \mathscr{C}$ and $\mathscr{D}$ are null on some straight lines of $\mathbb{R}^3$. Now, if we put two zeros, we have some relations like

$$\mathscr{A}(x, y, 0)(x + y) + \mathscr{B}(x, y, 0)y = 0$$

for $(t, u, v, w) = (-x - y, -y, 0, 0)$ for example. We get relations between the functions evaluated on some hyperplanes of $\mathbb{R}3*$. Even if we achieve to show from these relations that the functions $\mathscr{A}, \mathscr{B}, \mathscr{C}$ and $\mathscr{D}$ are null on some hyperplanes, we would be far to show that they are null everywhere. Actually it doesn't help us at all : we can have a function of $\mathbb{R}^3$ which is null on as many as possible lines and hyperplanes, it can be not null outside them, and there is no obvious condition we could impose to it to make it null. For instance, it exists some functions that are of class $\mathscr{C}^\infty$ and null everywhere except in a region, we just have to chose one function like this for which the not null region is outside the hyperplanes where $\mathscr{A}$ is supposed to be null.

The skew-symmetric system Applying (B.1) with three or two zeros is not sufficient, we have to keep only one. Let us start with $(t, u, v, w) = (0, x, x + y, x + y + z)$ where $x, y, z \in \mathbb{R}^*$. We get

$$\mathscr{B}(x, y, z)x + \mathscr{C}(x, y, z)(x + y) + \mathscr{D}(x, y, z)(x + y + z) = 0$$

If we add the three other possible relations, we get the skew-symmetric system

$$\forall x, y, z \in \mathbb{R}, \quad \begin{pmatrix} 0 & x & x+y & x+y+z \\ -x & 0 & y & y+z \\ -x-y & -y & 0 & y \\ -x-y-z & -y-z & -z & 0 \end{pmatrix} \begin{pmatrix} \mathscr{A}(x,y,z) \\ \mathscr{B}(x,y,z) \\ \mathscr{C}(x,y,z) \\ \mathscr{D}(x,y,z) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{B.2}$$

But unfortunately when we calculate it's determinant, we get

$$y^2(x + y + z)^2 + x^2z^2 + 2xyz(x + y + z) - (x + y)^2(y + z)^2$$

which is equal to the zero polynomial. By the way that's why we present the work for a stencil of 4 : or a stencil of 3, we would have end up with a third dimension skew-symmetric matrix whose determinant would have been null for sure because 3 is odd. In (B.2), if we replace one line of the matrix by $(1, 1, 1, 1)$ to represent the consistency relation, we also get a zero determinant. Actually the rank of this matrix is of 3 : the functions $\mathscr{A}, \mathscr{B}, \mathscr{C}$ and $\mathscr{D}$ are in a line. But with only these conditions, we can't show that they are null.

Reduction of each stencil An idea is to reduce the stencil of each function $\mathscr{A}, \mathscr{B}, \mathscr{C}$ and $\mathscr{D}$. For example let us say that $a_{i+\frac{1}{2}}$ depends only on u_{i-1}, u_i and u_{i+1} because it is in front of u_{i-1} in (2.8). With this idea, $b_{i+\frac{1}{2}}$ and $c_{i+\frac{1}{2}}$ would depend on all the stencil and $d_{i+\frac{1}{2}}$ on u_i, u_{i+1} and u_{i+2}. This idea comes from the fact that it is a valid condition for some advection schemes, the TVD one for sure, and the WENO it is almost true. With the framework of this section, these conditions would replace the relation (B.1) by

$$\mathscr{A}(u - t, v - u)t + \mathscr{B}(u - t, v - u, w - v)u + \mathscr{C}(u - t, v - u, w - v)v + \mathscr{D}(v - u, w - v)w = 0$$

We reduces the number of variables of some functions. This relation leads us to write some relations like

$$\mathscr{A}(x, y) = -\frac{y}{x + y}\mathscr{B}(x, y, 0)$$

We link one function on its whole domain by another one on a hyperplane of its domain. With some manipulations, it is possible to show that for all$x, z \in \mathbb{R}^*$,

$$\mathscr{A}(x,0) = \mathscr{B}(x,0,z) = \mathscr{C}(x,0,z) = \mathscr{D}(0,z)$$

But we just with these relations we are back at the previous issue : we can make these functions not null outside this hyperplane.

B.2 The diffusion coefficients

An idea could be to show that event if the flux coefficients are not unique, maybe the diffusion coefficients are. The idea is to write the symmetric part of flux difference for a stencil of 4 as an expansion with (2.16) and (A.2) :

$$\begin{aligned}\operatorname{sym}\big(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}\big) = B_{i-1}u_{i-2} &- \big(2B_{i-1} + 2B_i + L_{i-\frac{1}{2}}\big)u_{i-1} \\ &+ \big(B_{i-1} + 4B_i + B_{i+1} + L_{i-\frac{1}{2}} + L_{i+\frac{1}{2}}\big)u_i - \big(2B_i + 2B_{i+1} + L_{i+\frac{1}{2}}\big)u_{i+1} + B_{i+1}u_{i+2}\end{aligned}$$

Now, let write the diffusion coefficients like functions of their given stencils that we can deduce from (2.30) :

$$\begin{aligned}B_i &= \mathscr{B}(u_{i-2}, u_{i-1}, u_i, u_{i+1}, u_{i+2}) \\ L_{i+\frac{1}{2}} &= \mathscr{L}(u_{i-2}, u_{i-1}, u_i, u_{i+1}, u_{i+2}, u_{i+3})\end{aligned}$$

As in subappendix B.1, we can reduce the number of variables of $\mathcal{B}$ and $\mathcal{L}$ by one if we consider that the diffusion coefficients ignore the offset and we can study only the difference between two pair of candidates functions. The system to study is too long to be written her. It implies too much variables, and unfortunately, it seems that as before it doesn't have a unique solution.

C A generalization to all orders

The formulas (2.28), (2.30) and (2.31) make us think that a generalization to any stencils and order of diffusion is possible. The goal of this appendix is to express these formula to any order. We will work on a stencil of $2d$ with diffusions of orders 2 to $2d$. In the first subappendix C.1 we will generalize the work on the flux, then in subappendix C.2 we will introduce a general definition of the diffusion and finally, in subappendix C.3 we will study the flux diffusion system.

C.1 Flux at any stencil

First we need to generalize the expressions of the flux (2.1), (2.8) and (2.10). To represent the flux at the interface $x_{i+\frac{1}{2}}$ for a stencil of $2d$ by a linear combination we write

$$F_{i+\frac{1}{2}} = \sum_{k=-d+\frac{1}{2}}^{d-\frac{1}{2}} a^{(k)}_{i+\frac{1}{2}} u_{i+\frac{1}{2}+k}$$

where the $s^{(l)}_{i+\frac{1}{2}}$ are the *flux coefficients* and where the sum is on half-integers. For instance in the flux expression for a stencil 4 (2.8) we would have $a^{(-\frac{3}{2})}_{i+\frac{1}{2}} = ca_{i+\frac{1}{2}}$, $a^{(-\frac{1}{2})}_{i+\frac{1}{2}} = cb_{i+\frac{1}{2}}$, $a^{(\frac{1}{2})}_{i+\frac{1}{2}} = cc_{i+\frac{1}{2}}$ and

$a_{i+\frac{1}{2}}^{(\frac{3}{2})} = cd_{i+\frac{1}{2}}$. Here the velocity is include in the flux coefficients, which is by the way useful if we want it to be variable. We will suppose in addition the consistency relation

$$\sum_{k=-d+\frac{1}{2}}^{d-\frac{1}{2}} a_{i+\frac{1}{2}}^{(k)} = c \tag{C.1}$$

where c is the constant velocity. Let us compute the flux difference at the point x_i with this framework :

$$\begin{aligned} F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} &= \sum_{k=-d+\frac{1}{2}}^{d-\frac{1}{2}} a_{i+\frac{1}{2}}^{(k)} u_{i+\frac{1}{2}+k} - \sum_{k=-d+\frac{1}{2}}^{d-\frac{1}{2}} a_{i-\frac{1}{2}}^{(k)} u_{i-\frac{1}{2}+k} \\ &= \sum_{k=-d+\frac{1}{2}}^{d-\frac{1}{2}} a_{i+\frac{1}{2}}^{(k)} u_{i+\frac{1}{2}+k} - \sum_{k=-d-\frac{1}{2}}^{d-\frac{3}{2}} a_{i-\frac{1}{2}}^{(k+1)} u_{i+\frac{1}{2}+k} \end{aligned}$$

if we add that $a_{i+\frac{1}{2}}^{(-d-\frac{1}{2})} = a_{i+\frac{1}{2}}^{(d+\frac{1}{2})} = 0$,

$$\begin{aligned} F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} &= \sum_{k=-d-\frac{1}{2}}^{d-\frac{1}{2}} \left(a_{i+\frac{1}{2}}^{(k)} - a_{i-\frac{1}{2}}^{(k+1)} \right) u_{i+\frac{1}{2}+k} \\ &= \sum_{k=-d}^{d} \left(a_{i+\frac{1}{2}}^{(k-\frac{1}{2})} - a_{i-\frac{1}{2}}^{(k+\frac{1}{2})} \right) u_{i+k} \end{aligned}$$

Then we can define the flux matrix for all $i \in [\![1, N]\!], k \in [\![-d, d]\!]$ with

$$\mathbb{F}_{i,i+k} = a_{i+\frac{1}{2}}^{(k-\frac{1}{2})} - a_{i-\frac{1}{2}}^{(k+\frac{1}{2})}$$

so that

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = \sum_{l=-d}^{d} \mathbb{F}_{i,i+k} u_{i+k}$$

the other coefficients being zeros. Then, as we did in subsection 2.6, we have

$$\begin{aligned} \frac{\mathrm{d}E}{\mathrm{d}t} &= \sum_{i=1}^{N} u_i (F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}) \\ &= \sum_{i=1}^{N} \sum_{l=-d}^{d} u_i \mathbb{F}_{i,i+k} u_{i+k} \\ &= \langle \boldsymbol{U}, \mathbb{F}\boldsymbol{U} \rangle \\ &= \langle \boldsymbol{U}, \mathbb{S}\boldsymbol{U} \rangle \end{aligned}$$

where $\mathbb{S}$ is the symmetric part of $\mathbb{F}$ that we can compute :

$$\begin{aligned} \mathbb{S}_{i,i+k} &= \frac{1}{2}(\mathbb{F}_{i,i+k} - \mathbb{F}_{i+k,i}) \\ &= \frac{1}{2}(\mathbb{F}_{i,i+k} - \mathbb{F}_{i+k,i+k-k}) \\ \mathbb{S}_{i,i+k} &= \frac{1}{2}\left(a_{i+\frac{1}{2}}^{(k-\frac{1}{2})} - a_{i-\frac{1}{2}}^{(k+\frac{1}{2})} + a_{i+k+\frac{1}{2}}^{(-k-\frac{1}{2})} - a_{i+k-\frac{1}{2}}^{(-k+\frac{1}{2})} \right) \end{aligned} \tag{C.2}$$

C.2 Diffusion at any order

The operator Let generalize the diffusion operators we defined in (2.12), (2.21) and (2.24). For $r \in [\![1, d$, the *operator of diffusion of order* $2r$ is

$$\Delta^{(r)}(u) = \frac{\partial^r}{\partial x^r}\left(K^{(r)}\frac{\partial^r u}{\partial x^r}\right)$$

where $K^{(r)}$ is the *coefficient of diffusion of order* $2r$, it's what we want to compute from the flux coefficients. From this we can define the *global quantity of diffusion of order* $2r$ by

$$D^{(r)} = (-1)^r \Delta x^{2r-1} \int_0^1 u\Delta^{(r)}(u)\,\mathrm{d}x$$

the multiplicative constant is useful to make $K^{(r)}$ with the same dimension that the velocity c. With integrations by part with induction, we can show that

$$D^{(r)} = \Delta x^{2r-1} \int_0^1 K^{(r)}\left(\frac{\partial^r u}{\partial x^r}\right)^2 \mathrm{d}x \tag{C.3}$$

The hypothesis for these integrations by part are that for all $l \in [\![0, 2r-1]\!]$, $\dfrac{\partial^k u}{\partial x^k}$ is periodic and for all $l \in [\![0, r]\!]$, $\dfrac{\partial^k K^{(r)}}{\partial x^k}$ is periodic.

The derivative discretization To discretize the integral (C.3), we need to find a centered scheme to approximate the r-th derivative of u. We need to use the following formulas to calculate a one order derivative of a function f on the interfaces and the cells

$$\left.\frac{\partial f}{\partial x}\right|_i = \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x} \qquad \text{and} \qquad \left.\frac{\partial f}{\partial x}\right|_{i+\frac{1}{2}} = \frac{u_{i+1} - u_i}{\Delta x} \tag{C.4}$$

As we start from the values of u on the cells : (u_i), we quickly see that when r is an even number, we will work on the cells and when it is an odd number, we will work on the interfaces. The first three steps with this method to compute a centered approximation of the derivative give us (2.15), (A.1) and (A.4). If we continue the calculation by induction, we end up with

$$\left.\frac{\partial^{2p} u}{\partial x^{2p}}\right|_i = \frac{(-1)^p}{\Delta x^{2p}} \sum_{k=-p}^{p} (-1)^l \binom{2p}{p+l} u_{i+l}$$

for an even number $r = 2p$ and

$$\left.\frac{\partial^{2p+1} u}{\partial x^{2p+1}}\right|_{i+\frac{1}{2}} = \frac{(-1)^{p+1}}{\Delta x^{2p+1}} \sum_{k=-p}^{p+1} (-1)^l \binom{2p+1}{p+l} u_{i+l}$$

for an odd number $r = 2p+1$.

Proof. The initializations are easily proved by the calculation of the coefficients for the rank $2p = 0$ and by the second assumption on the discretizations (C.4) for the rank $2p+1 = 1$. Let assume that

the formula at the rank $2p$ is true. By hypothesis (C.4) and with the Pascal formula, we have

$$\begin{aligned}
\left.\frac{\partial^{2p+1}u}{\partial x^{2p+1}}\right|_{i+\frac{1}{2}} &= \frac{1}{\Delta x}\left[\left.\frac{\partial^{2p}u}{\partial x^{2p}}\right|_{i+1} - \left.\frac{\partial^{2p}u}{\partial x^{2p}}\right|_{i}\right] \\
&= \frac{(-1)^p}{\Delta x^{2p+1}}\left[\sum_{l=-p}^{p}(-1)^l\binom{2p}{p+l}u_{i+1+l} - \sum_{l=-p}^{p}(-1)^l\binom{2p}{p+l}u_{i+l}\right] \\
&= \frac{(-1)^p}{\Delta x^{2p+1}}\left[(-1)^p\binom{2p}{2p}u_{i+p+1} - (-1)^{-p}\binom{2p}{0}u_{i-p}\right. \\
&\qquad \left. + \sum_{l=-p+1}^{p}\left((-1)^{l-1}\binom{2p}{p+l-1} - (-1)^l\binom{2p}{p+l}\right)u_{i+l}\right] \\
&= \frac{(-1)^{p+1}}{\Delta x^{2p+1}}\left[(-1)^{-p}u_{i-p} + (-1)^{p+1}u_{i+p+1}\right. \\
&\qquad \left. + \sum_{l=-p+1}^{p}(-1)^l\left(\binom{2p}{p+l-1} + \binom{2p}{p+l}\right)u_{i+l}\right] \\
&= \frac{(-1)^{p+1}}{\Delta x^{2p+1}}\left[(-1)^{-p}\binom{2p+1}{p-p}u_{i-p} + (-1)^{p+1}\binom{2p+1}{p+p+1}u_{i+p+1}\right. \\
&\qquad \left. + \sum_{l=-p+1}^{p}(-1)^l\binom{2p+1}{p+l}u_{i+l}\right] \\
&= \frac{(-1)^{p+1}}{\Delta x^{2p+1}}\sum_{l=-p}^{p+1}(-1)^l\binom{2p+1}{p+l}u_{i+l}
\end{aligned}$$

To finish the induction we have to do the same work on the rank $2p+2$. With (C.4) we can write again

$$\left.\frac{\partial^{2p+2}u}{\partial x^{2p+2}}\right|_{i} = \frac{1}{\Delta x}\left[\left.\frac{\partial^{2p+1}u}{\partial x^{2p+1}}\right|_{i+\frac{1}{2}} - \left.\frac{\partial^{2p+1}u}{\partial x^{2p+1}}\right|_{i-\frac{1}{2}}\right]$$

we can show with the same manipulations that

$$\left.\frac{\partial^{2(p+1)}u}{\partial x^{2(p+1)}}\right|_{i} = \frac{(-1)^{p+1}}{\Delta x^{2(p+1)}}\sum_{k=-(p+1)}^{p+1}(-1)^l\binom{2(p+1)}{p+1+l}u_{i+l}$$

□

Discretized integration by part of the diffusion Now that we know how to calculate the value of the derivative from the values of u on the cells, we can discretize the integral in (C.3). We

start by the even stencil number case : $r = 2p$.

$$\frac{1}{\Delta x^{4p-1}} D^{(2p)} \simeq \sum_{i=1}^{N} \Delta x \left(K^{(2p)} \left(\frac{\partial^{2p} u}{\partial x^{2p}} \right)^2 \right) \Bigg|_i$$

$$\frac{1}{\Delta x^{4p}} D^{(2p)} \simeq \sum_{i=1}^{N} K_i^{(2p)} \left(\frac{\partial^{2p} u}{\partial x^{2p}} \Bigg|_i \right)^2$$

$$\frac{1}{\Delta x^{4p}} D^{(2p)} = \sum_{i=1}^{N} K_i^{(2p)} \left(\frac{(-1)^p}{\Delta x^{2p}} \sum_{l=-p}^{p} (-1)^l \binom{2p}{p+l} u_{i+l} \right) \left(\frac{(-1)^p}{\Delta x^{2p}} \sum_{q=-p}^{p} (-1)^q \binom{2p}{p+q} u_{i+q} \right)$$

$$D^{(2p)} = \sum_{l=-p}^{p} \sum_{q=-p}^{p} (-1)^{l+q+2p} \binom{2p}{p+l} \binom{2p}{p+q} \sum_{i=1}^{N} K_i^{(2p)} u_{i+l} u_{i+q}$$

$$= \sum_{l=-p}^{p} \sum_{q=-p}^{p} (-1)^{l+q} \binom{2p}{p+l} \binom{2p}{p+q} \sum_{i=1+l}^{N+l} K_{i-l}^{(2p)} u_i u_{i+q-l}$$

by supposing that (u_i) and $\left(K_i^{(r)}\right)$ are periodic families, we get

$$= \sum_{l=-p}^{p} \sum_{q=-p}^{p} (-1)^{l+q} \binom{2p}{p+l} \binom{2p}{p+q} \sum_{i=1}^{N} K_{i-l}^{(2p)} u_i u_{i+q-l}$$

$$= \sum_{i=1}^{N} u_i \sum_{l=-p}^{p} \sum_{q=-p}^{p} (-1)^{l+q} \binom{2p}{p+l} \binom{2p}{p+q} K_{i-l}^{(2p)} u_{i+q-l}$$

Let introduce the quantity $CL_i^{(2p)}$ by

$$CL_i^{(2p)} = \sum_{l=-p}^{p} \sum_{q=-p}^{p} (-1)^{l+q} \binom{2p}{p+l} \binom{2p}{p+q} K_{i-l}^{(2p)} u_{i+q-l}$$

So that

$$D^{(2p)} = \sum_{i=1}^{N} u_i CL_i^{(2p)} \tag{C.5}$$

We have

$$CL_i^{(2p)} = \sum_{l=-p}^{p} \sum_{q=-p}^{p} (-1)^{q-l} \binom{2p}{p-l} \binom{2p}{p+q} K_{i+l}^{(2p)} u_{i+q+l}$$

$$= \sum_{s=-2p}^{2p} \sum_{q+l=s,\, -p \leqslant q,l, \leqslant p} (-1)^{q-l} \binom{2p}{p-l} \binom{2p}{p+q} K_{i+l}^{(2p)} u_{i+q+l}$$

In one hand, $(-1)^{l-q} = (-1)^{s-2k} = (-1)^s$, and in the other hand,

$$\left\{ \begin{array}{l} q+l=s \\ -p \leqslant l \leqslant p \\ -p \leqslant q \leqslant p \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} q=s-l \\ -p \leqslant l \leqslant p \\ -p+s \leqslant l \leqslant p+s \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} q=s-l \\ -p+s^+ \leqslant l \leqslant p+s^- \end{array} \right.$$

where $s^+ = \max(0, s)$ and $s^- = \min(0, s)$. We have then

$$CL_i^{(2p)} = \sum_{s=-2p}^{2p} u_{i+s} (-1)^s \sum_{l=-p+s^+}^{p+s^-} \binom{2p}{p-l} \binom{2p}{p+s-l} K_{i+l}^{(2p)}$$

Matrix of diffusion Here we can define the *matrix of diffusion of order* $2p$ for $i \in [\![1, N]\!]$ and $s \in [\![-2p, 2p]\!]$ by

$$\mathbb{D}^{(2p)}_{i,i+s} = (-1)^s \sum_{l=-p+s^+}^{p+s^-} \binom{2p}{p-l}\binom{2p}{p+s-l} K^{(2p)}_{i+l} \tag{C.6}$$

the other coefficients being zero. We have

$$CL^{(2p)}_i = \sum_{s=-2p}^{2p} u_{i+s}\mathbb{D}^{(2p)}_{i,i+s} = \sum_{i=n}^{N} u_n \mathbb{D}^{(2p)}_{i,n}$$

if we recall (C.5), we have

$$D^{(2p)} = \sum_{i=1}^{N}\sum_{n=1}^{N} u_i \mathbb{D}^{(2p)}_{i,n} u_n$$

which means that

$$D^{(2p)} = \langle \boldsymbol{U}, \mathbb{D}\boldsymbol{U}\rangle \tag{C.7}$$

Let introduce the coefficients that we already see in (2.19), (A.3) and (A.6) for $l \in [\![-p, p]\!]$ and $s \in [\![-2p, 2p]\!]$:

$$C^{(2p)}_{l,s} = (-1)^s \binom{2p}{p+l}\binom{2p}{p+s-l} \tag{C.8}$$

First, let notice that $\binom{2p}{p+l} = \binom{2p}{p-l}$, so that the matrix of diffusion (C.6) can be written

$$\mathbb{D}^{(2p)}_{i,i+s} = \sum_{l=-p+s^+}^{p+s^-} C^{(2p)}_{l,s} K^{(2p)}_{i+l}$$

When s is positive, $\binom{2p}{p+s-l} = 0$ for $p + s - l > 2p$ which means $l < -p + s = -p + s^+$, and $s^- = 0$, so we have

$$\mathbb{D}^{(2p)}_{i,i+s} = \sum_{l=-p}^{p} C^{(2p)}_{l,s} K^{(2p)}_{i+l} \tag{C.9}$$

The same result appears for $s < 0$.

Some properties We have the following properties

1. $$C^{(2p)}_{p,2p} = 1 \tag{C.10}$$

2. $$C^{(2p)}_{l,s} = C^{(2p)}_{-l,-s} = C^{(2p)}_{l-s,-s} = C^{(2p)}_{s-l,s}$$

3. $$\sum_{s=-2p}^{2p} C^{(2p)}_{l,s} = 0 \tag{C.11}$$

4. $\mathbb{D}^{(2p)}$ is symmetric

Proof. 1. A simple apply of (C.8)

2. This property comes from the fact $\binom{2p}{p+l} = \binom{2p}{p-l}$ and $\binom{2p}{p+s-l} = \binom{2p}{p+l-s}$ and the intervention of the binomials

3.
$$\begin{aligned}\sum_{s=-2p}^{2p} C_{l,s}^{(2p)} &= \sum_{s=-2p}^{2p} (-1)^s \binom{2p}{p+l}\binom{2p}{p+s-l} \\ &= \binom{2p}{p+l} \sum_{s=-p-l<0}^{3p-l>2p} (-1)^{s-p+l}\binom{2p}{s} = (-1)^{l-p}\binom{2p}{p+l}\sum_{s=0}^{2p}(-1)^s\binom{2p}{s} \\ &= 0\end{aligned}$$

4. If $s > 0$, we have
$$\mathbb{D}_{i,i+s}^{(2p)} = \sum_{l=-p+s}^{p} C_{l,s}^{(2p)} K_{i+l}^{(2p)} = \sum_{l=-p}^{p-s} C_{l+s,s}^{(2p)} K_{i+s+l}^{(2p)} = \sum_{l=-p}^{p} C_{l,-s}^{(2p)} K_{i+s+l}^{(2p)} = \mathbb{D}_{i+s,i+s-s}^{(2p)} = \mathbb{D}_{i+s,i}^{(2p)}$$

□

The odd number case We can do the same work for an odd number $r = 2p+1$. But in this case we have to define the coefficients $K_{i+\frac{1}{2}}^{(2p+1)}$ on the interfaces. The difference we have to keep in mind, is that we define the coefficients $C_{k,s}^{(2p+1)}$ for half inter values of k : if we note $p' = p + \frac{1}{2}$, we have

$$C_{k,s}^{(2p+1)} = (-1)^s \binom{2p'}{p'+k}\binom{2p'}{p'+s-k}$$

then we can define the diffusion matrix by

$$\mathbb{D}_{i,i+s}^{(2p+1)} = \sum_{k=-p'}^{p'} C_{k,s}^{(2p+1)} K_{i+k}^{(2p+1)} \tag{C.12}$$

Finally, we have the same properties that we shew for the even case.

C.3 Study of the system

Origin of the system We defined the diffusion at any order, now we can define the total diffusion matrix

$$\mathbb{D} = \sum_{r=1}^{d} \mathbb{D}^{(r)}$$

As a sum of symmetrical matrix, $\mathbb{D}$ is symmetric. And verifies

$$D = \sum_{r=1}^{d} D^{(r)} = \langle \boldsymbol{U}, \mathbb{D}\boldsymbol{U} \rangle$$

as we defined the diffusion matrix for (C.7). So if we want to represent the loss of energy by diffusion :

$$D = \frac{\mathrm{d}E}{\mathrm{d}t}$$

we have with (C.2)

$$\forall \boldsymbol{U} \in \mathbb{R}^N, \qquad \langle \boldsymbol{U}, \mathbb{D}\boldsymbol{U} \rangle = \langle \boldsymbol{U}, \mathbb{S}\boldsymbol{U} \rangle$$

as both of these matrices are symmetrical

$$\mathbb{D} = \mathbb{S}$$

because we are searching a system like (2.29), we study

$$\forall i \in [\![1, N]\!],\ \forall s \in [\![-d, d]\!], \qquad \mathbb{D}_{i,i+s} = \mathbb{S}_{i,i+s} \tag{C.13}$$

Consistency of the system Let us analyze more $\mathbb{D}$.

$$\mathbb{D}_{i,i+s} = \sum_{r=1}^{d} \sum_{k=-\frac{r}{2}}^{\frac{r}{2}} C_{k,s}^{(r)} K_{i+k}^{(r)}$$

Here the second sum is on whole numbers when r is even and on half numbers when it is odd. This writing allows us to mix the two expressions (C.9) and (C.12). In one hand we have

$$\sum_{s=-d}^{d} \mathbb{D}_{i,i+s} = \sum_{r=1}^{d} \sum_{k=-\frac{r}{2}}^{\frac{r}{2}} \sum_{s=-d}^{d} C_{k,s}^{(r)} K_{i+k}^{(r)} \overset{(C.11)}{=} 0$$

and in the other hand,

$$\begin{aligned} \sum_{s=-d}^{d} \mathbb{D}_{i,i+s} &= \frac{1}{2} \left(\sum_{s=-d}^{d} a_{i+\frac{1}{2}}^{(s-\frac{1}{2})} - \sum_{s=-d}^{d} a_{i-\frac{1}{2}}^{(s+\frac{1}{2})} + \sum_{s=-d}^{d} a_{i+s+\frac{1}{2}}^{(-s-\frac{1}{2})} - \sum_{s=-d}^{d} a_{i+s-\frac{1}{2}}^{(-s+\frac{1}{2})} \right) \\ &\overset{(C.1)}{=} \frac{1}{2}(c - c + c - c) = 0 \end{aligned}$$

It means that we can ignore one equality on the middle of (C.13). And as the two matrix are symmetrical, we can study only the half of the line equality, which means that we take $s \in [\![1, d]\!]$.

Resolution of the system From the definition (C.8), it's easy to find that $C_{k,s}^{(r)} = 0$ if $k < s - \frac{r}{2}$. Furthermore, $\frac{r}{2} \geqslant k \geqslant s - \frac{r}{2} \Rightarrow r \geqslant s$. So we have

$$\begin{aligned} \mathbb{S}_{i,i+s} &= \sum_{r=s}^{d} \sum_{k=s-\frac{r}{2}}^{\frac{r}{2}} C_{k,s}^{(r)} K_{i+k}^{(r)} \\ &= \sum_{r=s+1}^{d} \sum_{k=s-\frac{r}{2}}^{\frac{r}{2}} C_{k,s}^{(r)} K_{i+k}^{(r)} + \sum_{k=s-\frac{s}{2}}^{\frac{s}{2}} C_{k,s}^{(s)} K_{i+k}^{(s)} \\ &= \sum_{r=s+1}^{d} \sum_{k=s-\frac{r}{2}}^{\frac{r}{2}} C_{k,s}^{(r)} K_{i+k}^{(r)} + C_{\frac{s}{2},s}^{(s)} K_{i+\frac{s}{2}}^{(s)} \\ &\overset{(C.10)}{=} \sum_{r=s+1}^{d} \sum_{k=s-\frac{r}{2}}^{\frac{r}{2}} C_{k,s}^{(r)} K_{i+k}^{(r)} + (-1)^s K_{i+\frac{s}{2}}^{(s)} \end{aligned}$$

Finally is we shift i by $\frac{s}{2}$ we have the following relation for all $i \in [\![1, N]\!]$ and $s \in [\![1, d]\!]$:

$$K_i^{(s)} = (-1)^{s+1} \sum_{r=s+1}^{d} \sum_{k=s-\frac{r}{2}}^{\frac{r}{2}} C_{k,s}^{(r)} K_{i-\frac{s}{2}+k}^{(r)} + \frac{(-1)^{s+1}}{2} \left(a_{i-\frac{s-1}{2}}^{(s-\frac{1}{2})} - a_{i-\frac{s+1}{2}}^{(s+\frac{1}{2})} + a_{i+\frac{s+1}{2}}^{(-s-\frac{1}{2})} - a_{i+\frac{s-1}{2}}^{(-s+\frac{1}{2})} \right)$$

This relation allows us to compute the different coefficients of diffusion from the coefficients of flux by recurrence, starting by the higher orders of diffusion.

References

[Lee74] B. Van Leer. "Towards the ultimate conservative difference scheme II. Monotonicity and conservation combined in a second-order scheme". In: *Journal of Computational Physics* 14 (4 1974). Ed. by Elsevier, pp. 361–370. URL: https://www.sciencedirect.com/science/article/abs/pii/0021999174900199.

[Roe86] P. L. Roe. "Characteristic-Based Schemes for the Euler Equations". In: *Annual Review of Fluid Mechanics* 18 (1986). Ed. by Annual Reviews, pp. 337–365. URL: https://www.annualreviews.org/doi/10.1146/annurev.fl.18.010186.002005.

[Har+87] A. Harten et al. "Uniformly high order accuracy essentially non-oscillatory schemes III". In: *Journal of Computational Physics* 131 (1 1987). Ed. by Elsevier, pp. 3–47. URL: https://www.sciencedirect.com/science/article/abs/pii/S0021999196956326.

[DT05] V. Daru and C. Tenaud. "Approximations d'ordre élevé pour les écoulements compressibles avec discontinuités". In: *École de printemps de Mécanique des Fluides Numérique*. Roscoff, 2005. URL: https://perso.limsi.fr/tenaud/Files/MFNschool_Daru_Tenaud.pdf.

[GA08] S. M. Griffies and A. J. Adcroft. "Formulating the Equations of Ocean Models". In: *Ocean Modeling in an Eddying Regime*. Ed. by American Geophysical Union. Vol. 177. Geophysical Monograph Series. 2008, pp. 281–317. URL: https://doi.org/10.1029/177GM18.

[Dur10] D. R. Durran. *Numerical Methods for Fluid Dynamics. With Applications to Geophysics*. Vol. 32. Texts in Applied Mathematics. Springer New York, NY, 2010. URL: https://link.springer.com/book/10.1007/978-1-4419-6412-0.

[Lem+12] F. Lemarié et al. "On the stability and accuracy of the harmonic and biharmonic isoneutral mixing operators in ocean models". In: *Ocean Modeling* 52-53 (2012). Ed. by Elsevier, pp. 9–35. URL: https://www.sciencedirect.com/science/article/abs/pii/S1463500312000674.

[BSA21] D. A. Bezgin, S. J. Schmidt, and N. A. Adams. "WENO3-NN: A maximum-order three-point data-driven weighted essentially non-oscillatory scheme". In: *Journal of Computational Physics* 452 (2021). Ed. by Elsevier. URL: https://www.sciencedirect.com/science/article/abs/pii/S0021999121008159.

[Der22] G. Derrida. "Towards a better control of numerical mixing in ocean models". Internship report. ENSTA Paris, Laboratoire Jean Kuntzmann - AIRSEA team, 2022.