The structure of well-balanced schemes for Friedrichs systems with linear relaxation

Bruno Després, Christophe Buet

The structure of well-balanced schemes for Friedrichs systems with linear relaxation

Bruno Després[a], Christophe Buet[b]

[a]*Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France*
tel: 33144279153, fax: 33144277200, email: despres@ann.jussieu.fr
[b]*CEA, DAM, DIF, 91297 Arpajon, France*

Abstract

We study the conservative structure of linear Friedrichs systems with linear relaxation in view of the definition of well-balanced schemes. We introduce a particular global change of basis and show that the change-of-basis matrix can be used to develop a systematic treatment of well-balanced schemes in one dimension. This algebra sheds new light on a family of schemes proposed recently by L. Gosse [15]. The application to the S_n model (a paradigm for the approximation of kinetic equations) for radiation is detailed. The discussion of the singular case is performed, and the 2D extension is shown to be equal to a specific multidimensional scheme proposed in [5]. This work is dedicated to the 2014 celebration of C. D. Munz' scientific accomplishments in the development of numerical methods for various problems in fluid mechanics.

Keywords: well balanced schemes, Friedrichs systems, conservative formulation, finite volume schemes.
2000 MSC: 65J10, 65N06, 65N99.

1. Introduction

Our interest in this work is the mathematical structure of Friedrichs systems with linear relaxation, having in mind the definition of well-balanced schemes which is a hot topic nowadays [4, 17, 21, 5, 16]. The particular case of Friedrichs systems of large size is challenging for numerical methods, and quite interesting since such large size systems are commonly encountered in the approximation of kinetic equations by moment methods [14, 6]. In this work we will consider the S_n model for radiation or neutrons propagation, which is a paradigm for approximations of kinetic equations [10]. Non linear extensions such as well-balanced schemes for shallow water equations or Euler equations are discussed in [18, 2, 23, 24, 11].

Our generic model problem is a linear system with relaxation in two dimensions

$$\partial_t U + \partial_x(A(\mathbf{x})U) + \partial_y(B(\mathbf{x})U) = -R(\mathbf{x})U, \quad U(t,\mathbf{x}) \in \mathbb{R}^n, \quad \mathbf{x} = (x,y) \in \mathbb{R}^2, \tag{1}$$

where the unknown is the function $U(t,\mathbf{x})$. The matrices $A(\mathbf{x}), B(\mathbf{x}), R(\mathbf{x}) \in \mathbb{R}^{n\times n}$ may be functions of the space variable, even if they will take as constant in most of this work. In all applications we have in mind the symmetric part of the relaxation matrix on the right hand side is non negative in the sense that $(V, RV) \geq 0$ for all $V \in \mathbb{R}^n$, that is

$$R + R^t \geq 0. \tag{2}$$

These matrices are symmetric for a Friedrichs system, that is $A = A^t$, $B = B^t$. The size n can be arbitrarily large. A more specific example which serves as a guideline in this text is the hyperbolic heat equation with $n = 2$

$$\left\{ \begin{array}{l} \partial_t p + \partial_x u = 0, \\ \partial_t u + \partial_x p = -\sigma u. \end{array} \right. \tag{3}$$

Here $U = (p,u)^t$, $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $R = \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix}$. Such linear systems are representative of the **linearization of non linear systems with relaxation**. Indeed consider the p-system with friction $\sigma \geq 0$

$$\partial_t \tau - \partial_x u = 0, \quad \partial_t u + \partial_x p(\tau) = -\sigma u.$$

Linearization $\tau = \tau_0 + \epsilon\tau_1 + \dots$ and $u = u_0 + \epsilon u_1 + \dots$ around an equilibrium $\partial_x\tau_0 = \partial_t\tau_0 = u_0 = 0$ yields

$$\partial_t\tau_1 - \partial_x u_1 = 0, \quad \partial_t u_1 - c_0^2\partial_x\tau_1 = -\sigma u_1,$$

where $c_0 > 0$ is the speed of sound. This linear system can be rewritten under the form of the Friedrichs system with linear relaxation using the symetrized variables $U = (c_0\tau_1, u_1)^t$ and the matrices $A = -c_0\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $R = \sigma\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$.

Remark 1. *Adding gravity as in $\partial_t\tau - \partial_x u = 0$ with $\partial_t u + \partial_x p(\tau) = g - \sigma u$ changes the structure of the linearized equations*

$$\partial_t\tau_1 - \partial_x u_1 = 0, \quad \partial_t u_1 - \partial_x\left(c_0(x)^2\tau_1\right) = -\sigma u_1$$

where the speed sound is no more constant $c_0 = c_0(x) = \sqrt{-p'(p^{-1}(xg))}$. For simplicity we will not consider this case hereafter, but we mention in our conclusion how the ideas developed in this work can be adapted without major difficulties.

A standard method for the discretization of problems like (1) relies on the splitting method: first, use a Finite Volume technique for the approximation of the homogeneous equation $\partial_t U + \partial_x(A(\mathbf{x})U) + \partial_y(B(\mathbf{x})U) = 0$, without right hand side; second, add the right hand side solving $\partial_t U = -R(\mathbf{x})U$. Even if efficient for most cases, a splitting method does not respect by construction the stationary states defined by

$$\mathcal{U} = \left\{\mathbf{x} \mapsto U(\mathbf{x});\ \partial_x(A(\mathbf{x})U) + \partial_y(B(\mathbf{x})U) = -R(\mathbf{x})U\right\}.$$

It has been observed in many cases that splitting methods may generate large numerical errors and unphysical oscillations, which therefore must be controlled in a way or another: a recent theoretical contribution on this topic is in [1]. In this direction the so called **well-balanced techniques** aim at combining Finite Volume techniques with the knowledge of the right-hand side so as to obtain new schemes which are exact for initial data in $\mathcal{U}$. We refer to [4, 17, 21, 5] for examples. A recent and comprehensive state-of-the-art is to be found in [16]. Moreover a efficient well-balanced scheme is often the starting point of a rigorous asymptotic preserving scheme, a topic that will not be developed in this work but for which we refer the reader to [16, 21, 5]. All these new methods are quite complex to construct and to analyze, and therefore are difficult to understand. This is why more mathematical analysis is needed to explore the fundamental structures of these techniques.

The basis of the method used hereafter tries to explore such a structure: it is reminiscent in some sense of the seminal work [7], since we begin to modify the equation in a conservative way so that usual Finite Volume schemes can be directly used to obtain methods which are exact for initial data in $\mathcal{U}$. For this purpose we will use the dual equation

$$\partial_t V + A^t(\mathbf{x})\partial_x V + B^t(\mathbf{x})\partial_y V = R^t(\mathbf{x})V, \quad V(t,\mathbf{x}) \in \mathbb{R}^n. \tag{4}$$

A fundamental property is

$$\partial_t(U,V) + \partial_x(A(\mathbf{x})U,V) + \partial_y(B(\mathbf{x})U,V) = 0 \tag{5}$$

for all U solutions of the primal equation (1) and V solutions of the dual equation (4). The remarkable fact is that identity (5) is in conservation form. It means that if one has enough knowledge of the solutions of the dual equation (4)

$$\mathcal{V} = \left\{\mathbf{x} \mapsto V(\mathbf{x});\ A(\mathbf{x})^t\partial_x V + B(\mathbf{x})^t\partial_y V = R^t(\mathbf{x})V\right\},$$

then it is possible to replace the non conservative primal equation (1) by the conservative identity (5). That is instead of analyzing the primal set $\mathcal{U}$, we put the emphasis on the dual set $\mathcal{V}$ of stationnary states of (4) which are now test functions. This is the basis of this work.

Concentrating of Finite Volume techniques and in view of formula (5), it is possible to conjecture that any well balanced Finite Volume solver for the primal equation (1) can be recast as a standard Finite Volume solver for the conservative formulation (5) (more precisely written as equation (10) in the core of the paper). In what follows we more modestly discretize directly (5) with usual conservative finite volume solvers, and deduce well-balanced

solvers for the non conservative primal formulation (1). We study two families of solvers, which are natural in our context. For the second family called the **two-states** solver, we show that it corresponds to a well balanced finite volume scheme based on space localization of the source term at the interfaces, see [16]. But in our approach there is absolutely no need of the localization method. Another asset of this method is the possibility to treat general meshes in higher dimensions. Notice that this approach is not restricted to Finite Volume techniques since the starting point is a modification of the equation.

This work is organized as follows. In section 2 we detail in one dimension the structure (5) and propose a new conservative formulation of the initial non conservative problem (1). Section 3 is dedicated to the discretization (still in dimension one) by means on standard Finite Volume techniques applied to the conservative system. We detail two different solvers. The main theorem about the well-posedness of the second family of solvers is given in this section. The example of the hyperbolic heat equation shows the second family is the same as the Gosse-Toscani scheme. The application of these ideas is performed in detail for the S_n model in section 4. The singular case $\det(A) = 0$ is briefly detailed in section 5 and a possible multiD extension in section 6. Some conclusions are drawn in section 7.

For the simplicity of the presentation (and only for that reason), we will consider that A and B are now **constant in space matrices**. Other matrices will be constructed which are non constant in space.

The interested reader can find many numerical illustrations which ilustrate the theory of this paper in [5, 11, 14, 15, 18, 21] and references therein. We particularily quote [16] for the variety of the models and of the schemes, and [6] where numerical tests are performed for Friedrichs system (including the Gosse-Toscani scheme which is a special case of our two-states solver below).

2. Conservative formulation in one dimension

In this section the matrices may be non symmetric. The set of stationary solutions of the dual equation is

$$\mathcal{V} = \left\{ x \mapsto V(x);\ A^t \partial_x V = R^t V \right\}. \tag{6}$$

This is a vectorial space of dimension p with $0 \leq p \leq n$. Let us denote a basis as $V_1(x), \ldots, V_p(x)$, so that $\mathcal{V} = \text{Span}\left(V_1(x), \ldots, V_p(x)\right)$.

Proposition 1 (Evident)**.** *Solutions of the primal equation (1) satisfy p linearly independent conservation laws*

$$\partial_t (U, V_i) + \partial_x (AU, V_i) = 0, \quad i \leq p. \tag{7}$$

One can define $\alpha_i = (U, V_i)$, the vector $\alpha = \begin{pmatrix} \alpha_1 \\ \ldots \\ \alpha_p \end{pmatrix} \in \mathbb{R}^p$ and the matrix $P(x) = \begin{pmatrix} V_1(x)^t \\ \ldots \\ V_p(x)^t \end{pmatrix} \in \mathbb{R}^{p \times n}$ so that (7) can be rewritten in a more compact form

$$\partial_t \alpha + \partial_x \left(P(x) A U \right) = 0. \tag{8}$$

An even more compact formulation is possible if A is non singular.

Proposition 2. *Assume A is non singular, that is $\det(A) \neq 0$. Then $p = n$ and the matrix P can be represented with the matrix exponential*

$$P(x) = e^{RA^{-1}x}. \tag{9}$$

Moreover the system (8) can rewritten as

$$\partial_t \alpha + \partial_x (Q(x)\alpha) = 0, \tag{10}$$

where the change of unknown is $\alpha = P(x)U \iff U = P^{-1}(x)\alpha$ and the matrix $Q(x) = P(x)AP^{-1}(x)$ is similar to A.

Proof. Any $V \in \mathcal{V}$ is solution of $\partial_x V = A^{-t} R^t V$, so can be represented as $V(x) = e^{A^{-t}R^t x} W$ where $W \in \mathbb{R}^n$ is arbitrary. So $p = n$. Moreover (5) yields $\partial_t \left(U, e^{A^{-t}R^t x} W \right) + \partial_x \left(AU, e^{A^{-t}R^t x} W \right) = 0$. Since W is arbitrary, it shows that

$$\partial_t \left(e^{RA^{-1}x} U \right) + \partial_x \left(e^{RA^{-1}x} AU \right) = 0.$$

The last part of the claim is immediate. The proof is ended. □

Remark 2. *Another proof of the identity (10) uses an idea commonly used in the numerical theory of Fokker-Planck equations. It is based on the identity*

$$A\partial_x U + RU = Ae^{-A^{-1}Rx}\partial_x\left(e^{A^{-1}Rx}U\right) = e^{-RA^{-1}x}A\partial_x\left(e^{A^{-1}Rx}U\right).$$

It yields the conservative equation $\partial_t\left(e^{RA^{-1}x}U\right) + A\partial_x\left(e^{A^{-1}Rx}U\right) = 0$ *which is exactly equal to (10). This is restricted of course to non singular matrices A. We will not develop this angle of attack since it is less clear how to generalize it to the singular case where* $det(A) = 0$ *and to the multidimensional case, as it will be done at the end of this work.*

Moreover this approach completely misses a central idea in this work which is that the matrix $P(x)$ *of the change of basis of the linear transformation is a fundamental object that needs specific examination.*

Remark 3. *The hypothesis that A is non singular appears at many places in this work. If the linear model derives from an underlying non linear model, it is equivalent to state that this underlying non-linear model is non resonant. The more difficult resonant case is briefly studied at the end of this work in section 5.*

Considering our main example which is the hyperbolic heat equation (3), one has that $RA^{-1} = \begin{pmatrix} 0 & 0 \\ \sigma & 0 \end{pmatrix}$. This matrix is nilpotent so

$$P(x) = I + xRA^{-1} = \begin{pmatrix} 1 & 0 \\ \sigma x & 1 \end{pmatrix}. \tag{11}$$

Therefore the unknown of the new formulation is $\alpha = (p, x\sigma p + u)^t$ and the matrix of equation (10) is

$$Q(x) = P(x)AP^{-1}(x) = P(x)AP(-x) = \begin{pmatrix} -x\sigma & 1 \\ 1 - x^2\sigma^2 & x\sigma \end{pmatrix}.$$

Proposition 3. *Stationary solutions of the fully conservative formulation (10) correspond to stationary solutions (i.e. well balanced solutions) of the primal non conservative equation (1).*

Proof. This is already a consequence of the previous proposition. A direct verification is as follows. Stationary solutions of the conservative formulation (10) satisfy $Q(x)\alpha(x) = W$ where $W \in \mathbb{R}^n$ is arbitrary. In terms of the primal variable it writes $P(x)AU(x) = W$, that is $e^{RA^{-1}x}AU(x) = W$. The differentiation with respect to x yields the identity $e^{RA^{-1}x}\left(RU(x) + A\partial_x U(x)\right) = 0$ which shows that $U \in \mathcal{U}$. The proof is ended. □

3. Finite volumes and Riemann solvers in one dimension

In what follows we start from (10) and shall detail the structure of some Finite Volume techniques. Additionally to the fact that A is constant, we make the assumption that A is **symmetric and non singular**

$$A = A^t \in \mathbb{R}^{n\times n} \text{ and } \det(A) \neq 0. \tag{12}$$

For the simplicity of notations the right hand side matrix will also be considered constant, $R = R^t$. However it must be noticed that the fundamental assumption is more the dissipativity of the symmetric part of R, as in (2). It is fundamental to prove the main result in theorem 1.

Let us define a new variable

$$\beta = Q(x)\alpha = P(x)AU \tag{13}$$

so that (10) recasts as $\partial_t\alpha + \partial_x\beta = 0$. The standard explicit Finite Volume discretization on a grid with varying mesh size $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ is

$$\frac{\alpha_j^{n+1} - \alpha_j^n}{\Delta t} + \frac{\beta_{j+\frac{1}{2}}^n - \beta_{j-\frac{1}{2}}^n}{\Delta x_j} = 0 \tag{14}$$

where $\beta_{j+\frac{1}{2}}^n$ is the flux at time step $t_n = n\Delta t$. The definition of the scheme relies on the definition of the flux. This will be performed in the context of simplified Riemann solvers as sketched in figure 1.

The construction of the Riemann solver is based on the spectral decomposition of the matrix $Q(x)$. Let us denote $x^* = x_{j+\frac{1}{2}}$ the point between cell $L = j$ and cell $R = j+1$. Since A is symmetric, it admits a basis of real orthonormal eigenvectors, that is $Au_p = \lambda_p u_p$ where $\lambda_p \neq 0$ since A is not singular. The point is that Q is similar to A, but is not symmetric. The (right) eigenvectors of Q are

$$Q(x^*)r_p^* = \lambda_p r_p^*, \qquad r_p^* = P(x^*)u_p. \tag{15}$$

The (left) eigenvectors are

$$Q^t(x^*)s_p^* = \lambda_p s_p^*, \qquad s_p^* = P(x^*)^{-t}u_p. \tag{16}$$

Based on this decomposition, we construct hereafter two different solvers which are referred to as the one-state solver and the two-states solvers. This distinction originates in a previous work [13].

3.1. A one-state solver

Since Q does not depend on the time variable, we also note that β is solution of the autonomous equation

$$\partial_t\beta + Q(x)\partial_x\beta = 0. \tag{17}$$

Let us assume that the local variations of matrix $x \mapsto Q(x)$ are smooth enough, so that it is reasonable to freeze the matrix at the interface $x_{j+\frac{1}{2}}$ still maintaining a correct accuracy. This idea is to be compared with similar ones in [16][pages 64-66] or [20, 26]. In our case we modify locally equation (17) by $\partial_t\beta + Q(x^*)\partial_x\beta = 0$, $x^* = x_{j+\frac{1}{2}}$. Here locally means at the interface between cell j and cell $j+1$. Solving this equation in the interval $(x^* - \epsilon, x^* + \epsilon)$ with Riemann data $\beta(t=0) = \beta_L$ for $x < x^*$ and $\beta(t=0) = \beta_R$ for $x^* < x$, one can compute the Riemann invariants $\left(s_p^*, \beta\right)$ using the equation $\partial_t\left(s_p^*, \beta\right) + \lambda_p\partial_x\left(s_p^*, \beta\right) = 0$. The solution $\beta^* = \beta(t, x^*)$, $t > 0$ is provided by the solution of the linear system

$$\begin{cases} \left(s_p^*, \beta^* - \beta_L\right) = 0, & \lambda_p > 0, \\ \left(s_p^*, \beta^* - \beta_R\right) = 0, & \lambda_p < 0. \end{cases} \tag{18}$$

This linear system is non singular since the eigenvectors (s_p^*) are linearly independent. It defines a function $\varphi : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ such that

$$\varphi(\beta_L, \beta_R, x^*) = \beta^*. \tag{19}$$

The usual consistency property of Riemann solver writes

$$\varphi(\beta, \beta, x) = \beta \qquad \forall(x, \beta). \tag{20}$$

One may call this a one-state Riemann solver because the eigenvectors are common to the same matrix $Q(x^*)$. This will not be the case in the next section. The scheme writes

$$\frac{\alpha_j^{n+1} - \alpha_j^n}{\Delta t} + \frac{\varphi(\beta_j^n, \beta_{j+1}^n, x_{j+\frac{1}{2}}) - \varphi(\beta_{j-1}^n, \beta_j^n, x_{j-\frac{1}{2}})}{\Delta x_j} = 0. \tag{21}$$

In terms of the original variable $U_j = P(x_j)^{-1}\alpha_j$, the scheme writes

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + P(x_j)^{-1}\frac{\varphi(\beta_j^n, \beta_{j+1}^n, x_{j+\frac{1}{2}}) - \varphi(\beta_{j-1}^n, \beta_j^n, x_{j-\frac{1}{2}})}{\Delta x_j} = 0. \tag{22}$$

Proposition 4. *The scheme (22) is well-balanced.*

Proof. Being well-balanced means being exact for stationary solutions. So let us assume that $U_j^0 = P(x_j)^{-1}\alpha_j^0$ with $\alpha_j^0 = Q(x_j)^{-1}\gamma$ where $\gamma \in \mathbb{R}^n$ is a given vector. By construction $\beta_j^0 = Q(x_j)\alpha_j^0 = \gamma$ is constant. Therefore $\varphi(\beta_j^0, \beta_{j+1}^0, x_{j+\frac{1}{2}}) = \gamma$, $\forall j$. In view of (22) the scheme is stationary $U_j^1 = U_j^0$ for all j. The proof is ended. □

The well-balanced scheme (22) is written as a **multiplicative modification** of a standard finite volume solver, where multiplicative means that the numerical finite volume discretization of $\partial_x(AU)$ is premultiplied on the left by the matrix $P(x_j)^{-1}$. With this respect, the initial **additive modification**, that is $+RU$, has been changed into a multiplicative one. A complementary understanding of the scheme (22) is provided after rewriting it in a more standard additive formulation. To do so we rewrite (22) as

$$\frac{U_j^{n+1}-U_j^n}{\Delta t}+\frac{P(x_{j+\frac12})^{-1}\beta^*_{j+\frac12}-P(x_{j-\frac12})^{-1}\beta^*_{j-\frac12}}{\Delta x_j}+\frac{P(x_j)^{-1}-P(x_{j+\frac12})^{-1}}{\Delta x_j}\beta^*_{j+\frac12}+\frac{P(x_{j-\frac12})^{-1}-P(x_j)^{-1}}{\Delta x_j}\beta^*_{j-\frac12}=0 \tag{23}$$

where $\beta^*_{j+\frac12}=\varphi(\beta^n_j,\beta^n_{j+1},x_{j+\frac12})$. We define

$$U^*_{j+\frac12}=A^{-1}P(x_{j+\frac12})^{-1}\beta^*_{j+\frac12}, \tag{24}$$

and $\Delta x_j^{\pm}=x_{j\pm\frac12}-x_j$ so that (23) can be rewritten in a slightly more conventional form

$$\frac{U_j^{n+1}-U_j^n}{\Delta t}+A\frac{U^*_{j+\frac12}-U^*_{j-\frac12}}{\Delta x_j}+\frac{P(x_j)^{-1}P(x_{j+\frac12})-I}{\Delta x_j}AU^*_{j+\frac12}+\frac{I-P(x_j)^{-1}P(x_{j-\frac12})}{\Delta x_j}AU^*_{j-\frac12}=0. \tag{25}$$

Here , $A\frac{U^*_{j+\frac12}-U^*_{j-\frac12}}{\Delta x_j}$ is the discretization of the divergent term $A\partial_x U$, and $\frac{P(x_j)^{-1}P(x_{j+\frac12})-I}{\Delta x_j}AU^*_{j+\frac12}+\frac{I-P(x_j)^{-1}P(x_{j-\frac12})}{\Delta x_j}AU^*_{j-\frac12}$ is the discretization of the source RU.

Proposition 5. *The quantity $U^*_{j+\frac12}$ is solution of the linear system*

$$\begin{cases}\left(u_p,U^*_{j+\frac12}-e^{-A^{-1}R\Delta x_j^+}U_j\right)=0, & \lambda_p>0,\\ \left(u_p,U^*_{j+\frac12}-e^{-A^{-1}R\Delta x^-_{j+1}}U_{j+1}\right)=0, & \lambda_p<0.\end{cases} \tag{26}$$

Proof. We make the elimination $\beta^*_{j+\frac12}=P(x_{j+\frac12})AU^*_{j+\frac12}$ and $\beta_j=P(x_j)AU_j$ in (18). Therefore the first line of (18) for $\lambda_p>0$ rewrites as $\left(P(x_{j+\frac12})^{-t}u_p,P(x_{j+\frac12})AU^*_{j+\frac12}-P(x_j)AU_j\right)=0$, or $\left(u_p,AU^*_{j+\frac12}-P(x_{j+\frac12})^{-1}P(x_j)AU_j\right)=0$, or $\left(Au_p,U^*_{j+\frac12}-A^{-1}P(x_{j+\frac12})^{-1}P(x_j)AU_j\right)=0$. Since $Au_p=\lambda_p u_p$ and

$$A^{-1}P(x_{j+\frac12})^{-1}P(x_j)A=A^{-1}e^{-RA^{-1}\left(x_{j+\frac12}-x_j\right)}A=e^{-A^{-1}R\Delta x_j^+} \tag{27}$$

it yields the first line of the claim. The second part of claim is proved in a similar fashion. □

Remark 4 (Elimination of the matrix exponentials)**.** *We notice also that the formula (27) can recast as*

$$P(x_{j+\frac12})^{-1}P(x_j)=P(x_j-x_{j+\frac12})=e^{-RA^{-1}\Delta x_j^+} \tag{28}$$

so that the linear system (26) is also rewritten as

$$\begin{cases}\left(u_p,U^*_{j+\frac12}-P(x_j)^tP(x_{j+\frac12})^{-t}U_j\right)=0, & \lambda_p>0,\\ \left(u_p,U^*_{j+\frac12}-P(x_{j+1})^tP(x_{j+\frac12})^{-t}U_{j+1}\right)=0, & \lambda_p<0.\end{cases} \tag{29}$$

It means that the matrix exponentials can be eliminated in the fluxes. We notice also that the matrix exponentials do not show up in the scheme itself (25). This remark will be fundamental in the multiD example at the end of this work: we will see that the matrix exponentials are not well defined, while the matrix P has a natural multiD definition.

Continuing to use the matrix exponentials in one dimension since there are well adapted to our purposes, the formulas (23) can now be interpreted.

• If $R = 0$ which means no relaxation, the matrix exponential terms like $e^{-A^{-1}R\Delta x_j^+}$ are equal to the identity matrix. In this case $U^*_{j+\frac{1}{2}} = U^{\text{Riemann}}_{j+\frac{1}{2}}$ is the solution of the standard Riemann solver

$$\begin{cases} \left(u_p, U^{\text{Riemann}}_{j+\frac{1}{2}} - U_j\right) = 0, & \lambda_p > 0, \\ \left(u_p, U^{\text{Riemann}}_{j+\frac{1}{2}} - U_{j+1}\right) = 0, & \lambda_p < 0. \end{cases} \tag{30}$$

This construction corresponds to figure 1.

• In all cases the matrix exponentials are related to the solution of the stationary direct problem $A\partial_x W = -RW$. The solution of this auxiliary equation can be written as $W(x_{j+\frac{1}{2}}) = \Pi(-\Delta x_j^+)W(x_j)$ where $\Pi(-\Delta x_j^+) = e^{-A^{-1}R\Delta x_j^+}$ stands for the propagator over the distance Δx_j^+. So one can rewrite (26) as the solution of the standard Riemann solver

$$\begin{cases} \left(u_p, U^{\text{Riemann}}_{j+\frac{1}{2}} - U^{\text{modified}}_{j}\right) = 0, & \lambda_p > 0, \\ \left(u_p, U^{\text{Riemann}}_{j+\frac{1}{2}} - U^{\text{modified}}_{j+1}\right) = 0, & \lambda_p < 0, \end{cases} \tag{31}$$

where the modified states are designed by propagating the cells centered states. This is illustrated in figure 2.

• The sink term in (23) is $S = S_1 + S_2$. The first term is

$$S_1 = \frac{P(x_j)^{-1}P(x_{j+\frac{1}{2}}) - I}{\Delta x_j} A U^*_{j+\frac{1}{2}} = \frac{e^{RA^{-1}\Delta x_j^+} - I}{\Delta x_j} A U^*_{j+\frac{1}{2}}.$$

A Taylor expansion with respect to mesh size yields $S_1 = \frac{\Delta x_j^+}{\Delta x_j} R U^*_{j+\frac{1}{2}} + O(h)$ where $h = \sup_j \Delta x_j$. A similar analysis yields $S_2 = \frac{-\Delta x_j^-}{\Delta x_j} R U^*_{j-\frac{1}{2}} + O(h)$. Considering that $U^*_{j+\frac{1}{2}}$ and $U^*_{j-\frac{1}{2}}$ are homogeneous to U and that $\Delta x_j^+ - \Delta x_j^- = \Delta x_j$, one sees that $S = S_1 + S_2 \approx RU$ is consistent with (the opposite of) the relaxation term.

Let us now particularize these quantities for the example of the hyperbolic heat equation (3). The eigenvectors of A are $u_1 = (1, 1)$ associated to $\lambda_1 > 0$ and $u_2 = (1, -1)$ associated to $\lambda_2 = -1$. The standard Riemann solver (30) can be written under the form

$$\begin{cases} \left(p^{\text{Riemann}}_{j+\frac{1}{2}} - p_j\right) & + \left(u^{\text{Riemann}}_{j+\frac{1}{2}} - u_j\right) & = 0, \\ \left(p^{\text{Riemann}}_{j+\frac{1}{2}} - p_{j+1}\right) & - \left(u^{\text{Riemann}}_{j+\frac{1}{2}} - u_{j+1}\right) & = 0, \end{cases}$$

that is

$$p^{\text{Riemann}}_{j+\frac{1}{2}} = \frac{p_j + p_{j+1}}{2} + \frac{u_j - u_{j+1}}{2} \quad \text{and} \quad u^{\text{Riemann}}_{j+\frac{1}{2}} = \frac{u_j + u_{j+1}}{2} + \frac{p_j - p_{j+1}}{2}.$$

The modified solver (31) is function of $U^{\text{modified}}_j = e^{-A^{-1}R\Delta x_j^+} U_j$ where

$$e^{A^{-1}Ry} = I + yA^{-1}R = \begin{pmatrix} 1 & \sigma y \\ 0 & 1 \end{pmatrix}. \tag{32}$$

So $U^{\text{modified}}_j = \left(p_j - \sigma\Delta x_j^+ u_j, u_j\right)$ and $U^{\text{modified}}_{j+1} = \left(p_{j+1} - \sigma\Delta x_{j+1}^- u_{j+1}, u_{j+1}\right)$. The new solver (31) writes

$$\begin{cases} p^*_{j+\frac{1}{2}} & = \dfrac{p_j + p_{j+1}}{2} & + \dfrac{(1 - \sigma\Delta x_j^+)u_j - (1 + \sigma\Delta x_{j+1}^-)u_{j+1}}{2}, \\ u^*_{j+\frac{1}{2}} & = \dfrac{p_j - p_{j+1}}{2} & + \dfrac{(1 - \sigma\Delta x_j^+)u_j + (1 + \sigma\Delta x_{j+1}^-)u_{j+1}}{2}. \end{cases} \tag{33}$$

This solver is identical to the "Piecewise Steady Approximation", an idea that goes back to [20], see also [22].

3.2. A two-states solver more adapted to discontinuous coefficients

The previous hypothesis that $x \mapsto Q(x)$ is continuous with small variation is not always reasonable, in particular for more general problems with discontinuous coefficients as illustrated in figure 3. To illustrate this situation we consider the simple model equation

$$\partial_t U + A\partial_x U = -R(x)U \tag{34}$$

with $R(x) = R_1$ for $x < 0$ and $R = R_2$ for $0 < x$. In this simple example one has that $A = A_1 = A_2$. Still assuming that A is non singular, one gets that $P(x) = e^{R_1 A^{-1} x}$ for $x < 0$ and $P(x) = e^{R_2 A^{-1} x}$ for $0 < x$. If R_1 is very different from R_2, the matrix P does not have a continuous derivative at $x = 0$. In this case it is possible to imagine that $Q(x)$ might have strong local variation, so one may question the accuracy of the one-state solver. We now develop a method which can be used in the situation described previously but also extend to more general cases such as $A_1 \neq A_2$.

Our approach is here to upwind the eigenvectors in (18), and to consider instead

$$\begin{cases} \left(s_p^L, \beta^{**} - \beta_L\right) = 0, & \lambda_p > 0, \\ \left(s_p^R, \beta^{**} - \beta_R\right) = 0, & \lambda_p < 0, \end{cases} \tag{35}$$

where s_p^L is a generic eigenvector of the matrix $Q(x_L)$ and s_p^R is a generic eigenvector of the matrix $Q(x_R)$. We call this a two-states Riemann solver because the eigenvectors are different. The linear system is invertible if and only if the vectors s_P^L (for $\lambda_p > 0$) and s_P^R (for $\lambda_p < 0$) are linearly independent.

Theorem 1. *Assume $R + R^t \geq 0$. Then the family $\left\{s_p^L\right\}_{\lambda_p>0} \cup \left\{s_p^R\right\}_{\lambda_p<0}$ is linearly independent, and so the two-states solver (35) is well defined.*

Proof. It is clear that if R vanishes, then these eigenvectors are equal to the ones of the matrix A, and so are linearly independent. The key of the proof is the use of the dissipativity hypothesis $R + R^t \geq 0$ in an appropriate manner.

Let $s_i(x) = P(x)^{-t} u_i$ denotes a generic eigenvector of $Q(x)^t$, with the convention that s_i^L is an eigenvector of the matrix $Q(x_L)$ for $1 \leq i \leq k$ (that is $\lambda_i^L > 0$), and s_i^R is an eigenvector of the matrix $Q(x_R)$ for $k+1 \leq i \leq n$ (that is now $\lambda_i^R < 0$). One has to take care that the convention is here the opposite of the usual one since eigenvalues with low indices are positive and eigenvalues with higher indices are negative. We will show that the only real solution of the equation $\sum_{i=1}^k \alpha_i s_i^L + \sum_{i=k+1}^n \alpha_i s_i^R = 0$ is $\alpha_i = 0$ for $1 \leq i \leq n$, which proves the claim. Let us set

$$z = \sum_{i=1}^{k} \alpha_i s_i^L = -\sum_{i=k+1}^{n} \alpha_i s_i^R. \tag{36}$$

We denote x_* a point between the left cell and the right cell: $x_L < x_* < x_R$.

• Let us study the function

$$f(x) = \left(Ae^{A^{-1}Rx_*}\left(\sum_{i=1}^{k}\alpha_i s_i(x)\right), e^{A^{-1}Rx_*}\left(\sum_{i=1}^{k}\alpha_i s_i(x)\right)\right) = \left(Ae^{A^{-1}R(x_*-x)}\left(\sum_{i=1}^{k}\alpha_i u_i\right), e^{A^{-1}R(x_*-x)}\left(\sum_{i=1}^{k}\alpha_i u_i\right)\right).$$

One has that (assuming $\|u_i\| = 1$)

$$f(x_*) = \left(A\left(\sum_{i=1}^{k}\alpha_i u_i\right), \left(\sum_{i=1}^{k}\alpha_i u_i\right)\right) = \sum_{i=1}^{k}\lambda_i|\alpha_i|^2 \geq 0. \tag{37}$$

On the other hand a direct calculation shows that

$$f'(x) = -2\left(Re^{A^{-1}R(x_*-x)}\left(\sum_{i=1}^{k}\alpha_i u_i\right), e^{A^{-1}R(x_*-x)}\left(\sum_{i=1}^{k}\alpha_i u_i\right)\right) = -\left(\left(R+R^t\right)e^{A^{-1}R(x_*-x)}\left(\sum_{i=1}^{k}\alpha_i u_i\right), e^{A^{-1}R(x_*-x)}\left(\sum_{i=1}^{k}\alpha_i u_i\right)\right)$$

Here $R + R^t \geq 0$. So $f'(x) \leq 0$ and therefore

$$f(x_L) \geq f(x_*) \geq 0. \tag{38}$$

• Similarily we define

$$g(x) = \left(Ae^{A^{-1}Rx_*} \left(\sum_{i=k+1}^{n} \alpha_i s_i(x) \right), e^{A^{-1}Rx^*} \left(\sum_{i=k+1}^{n} \alpha_i s_i(x) \right) \right) = \left(Ae^{A^{-1}R(x_*-x)} \left(\sum_{i=k+1}^{n} \alpha_i u_i \right), e^{A^{-1}R(x_*-x)} \left(\sum_{i=k+1}^{n} \alpha_i u_i \right) \right).$$

One has that

$$g(x_*) = \left(A \left(\sum_{i=k+1}^{n} \alpha_i u_i \right), \left(\sum_{i=k+1}^{n} \alpha_i u_i \right) \right) = \sum_{i=k+1}^{n} \lambda_i |\alpha_i|^2 \leq 0. \tag{39}$$

We also have that $g'(x) \leq 0$. Therefore

$$0 \geq g(x_*) \geq g(x_R). \tag{40}$$

• But due to (36), one has that $f(x_L) = g(x_R)$. By comparison with (38) and (40) it yields $f(x_L) = g(x_R) = 0$. Therefore (38) implies that $f(x_*) = 0$ which shows (for example with the help of (37)) that $\alpha_i = 0$ for $i \leq k$. Similarly one has that $g(x_*) = 0$ which shows that $\alpha_i = 0$ for $k+1 \leq i$. The proof is ended. □

The solution of the linear system (35) can be written with a function $\psi : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^n$ such that

$$\psi(\beta_L, \beta_R, x_L, x_R) = \beta^{**}. \tag{41}$$

We call it the two-states solver. The usual consistency property of Riemann solvers writes

$$\psi(\beta, \beta, x, y) = \beta \qquad \forall (x, y, \beta). \tag{42}$$

It is a direct consequence that if one plugs $\beta_L = \beta_R = \beta$ in (35), then β^{**} is a solution, independently of $x_L < x_R$. Since this solution is unique by the theorem, it shows (42). With this notation for the Riemann solver, the scheme (35) writes

$$\frac{\alpha_j^{n+1} - \alpha_j^n}{\Delta t} + \frac{\psi(\beta_j^n, \beta_{j+1}^n, x_{j-1}, x_j) - \psi(\beta_{j-1}^n, \beta_j^n, x_j, x_{j+1})}{\Delta x_j} = 0. \tag{43}$$

Proposition 6. *The scheme (43) is well-balanced (same proof as the one of proposition 4).*

As before it is possible to get a more classical understanding of (43) by rewriting it as

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{P(x_{j+\frac{1}{2}})^{-1}\beta_{j+\frac{1}{2}}^{**} - P(x_{j-\frac{1}{2}})^{-1}\beta_{j-\frac{1}{2}}^{**}}{\Delta x_j} + \frac{P(x_j)^{-1} - P(x_{j+\frac{1}{2}})^{-1}}{\Delta x_j}\beta_{j+\frac{1}{2}}^{**} + \frac{P(x_{j-\frac{1}{2}})^{-1} - P(x_j)^{-1}}{\Delta x_j}\beta_{j-\frac{1}{2}}^{**} = 0 \tag{44}$$

where $\beta_{j+\frac{1}{2}}^{**} = \psi(\beta_j^n, \beta_{j+1}^n, x_{j-1}, x_j)$. The generic flux is identified as $P(x_{j+\frac{1}{2}})^{-1}\beta_{j+\frac{1}{2}}^{**}$. To make this statement more explicit we define

$$U_{j+\frac{1}{2}}^{**} = A^{-1}P(x_{j+\frac{1}{2}})^{-1}\beta_{j+\frac{1}{2}}^{**}. \tag{45}$$

Proposition 7. *The quantity $U_{j+\frac{1}{2}}^{**}$ in (45) is solution of the linear system*

$$\begin{cases} \left(u_p, e^{A^{-1}R\Delta x_j^+} U_{j+\frac{1}{2}}^{**} - U_j \right) = 0, & \lambda_p > 0, \\ \left(u_p, e^{A^{-1}R\Delta x_{j+1}^-} U_{j+\frac{1}{2}}^{**} - U_{j+1} \right) = 0, & \lambda_p < 0, \end{cases} \tag{46}$$

where the matrix exponentials are the opposite of the ones in the one-state solver (26).

Remark 5 (Elimination of the matrix exponentials)**.** *As it was done in remark 4 for the one-state solver, it is possible to rewrite (46) using only the matrix P. It yields*

$$\begin{cases} \left(u_p, P(x_j)^{-t}P(x_{j+\frac{1}{2}})^t U_{j+\frac{1}{2}}^{**} - U_j \right) = 0, & \lambda_p > 0, \\ \left(u_p, P(x_{j+1})^{-t}P(x_{j+\frac{1}{2}})^t U_{j+\frac{1}{2}}^{**} - U_{j+1} \right) = 0, & \lambda_p < 0. \end{cases} \tag{47}$$

Proof. Indeed (35) recasts as

$$\begin{cases} \left(P(x_j)^{-t}u_p, P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - P(x_j)AU_j\right) = 0, & \lambda_p > 0, \\ \left(P(x_{j+1})^{-t}u_p, P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - P(x_{j+1})AU_{j+1}\right) = 0, & \lambda_p < 0, \end{cases}$$

that is

$$\begin{cases} \left(u_p, P(x_j)^{-1}P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - AU_j\right) = 0, & \lambda_p > 0, \\ \left(u_p, P(x_{j+1})^{-1}P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - AU_{j+1}\right) = 0, & \lambda_p < 0, \end{cases}$$

or also

$$\begin{cases} \left(Au_p, A^{-1}P(x_j)^{-1}P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - U_j\right) = 0, & \lambda_p > 0, \\ \left(Au_p, A^{-1}P(x_{j+1})^{-1}P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - U_{j+1}\right) = 0, & \lambda_p < 0. \end{cases}$$

Since $Au_p = \lambda_p u_p$ it can written as

$$\begin{cases} \left(u_p, A^{-1}P(x_j)^{-1}P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - U_j\right) = 0, & \lambda_p > 0, \\ \left(u_p, A^{-1}P(x_{j+1})^{-1}P(x_{j+\frac{1}{2}})AU^{**}_{j+\frac{1}{2}} - U_{j+1}\right) = 0, & \lambda_p < 0. \end{cases}$$

One has the identity $A^{-1}P(x_j)^{-1}P(x_{j+\frac{1}{2}})A = A^{-1}e^{RA^{-1}\Delta x_j^+}A = e^{A^{-1}R\Delta x_j^+}$ from which one gets the first line of the claim. The second line is obtained similarly. The proof is ended. □

The interpretation of two-states solver, that is (44) with (46), is quite close to the interpretation of the one state solver. We nevertheless observe that the analogue of (30) is now

$$\begin{cases} \left(u_p, U^{\text{Riemann and modified}}_{j+\frac{1}{2}} - U_j\right) = 0, & \lambda_p > 0, \\ \left(u_p, U^{\text{Riemann and modified}}_{j+\frac{1}{2}} - U_{j+1}\right) = 0, & \lambda_p < 0, \end{cases} \tag{48}$$

where the modified states are the propagation of the intermediate state. The modification is the backward propagator $\Pi(\Delta x_j^+)$ which is the inverse of the direct propagator $\Pi(-\Delta x_j^+)$ that corresponds to the one-state solver. This is illustrated in figure 4.

Remark 6 (Interpretation of figure 4)**.** *In the language of the localization method of L. Gosse, the explanation of the figure is that a stationary state is inserted into the Riemann solver, between the left and right states: this construction can somewhat be considered as arbitrary. Other interpretations of the figure are probably possible, but we will not pursue them in this work since they are not needed. With the proposed approach, the internal stationary state does not really exists: it is just the consequence of a certain linear algebra with standard Riemann solvers and standard conservative methods.*

Going back once again to the hyperbolic heat equation and using (32), the system (46) rewrites as

$$\begin{cases} \left(p^{**}_{j+\frac{1}{2}} + \sigma\Delta x_j^+ u^{**}_{j+\frac{1}{2}} - p_j\right) + \left(u^{**}_{j+\frac{1}{2}} - u_j\right) = 0, \\ \left(p^{**}_{j+\frac{1}{2}} + \sigma\Delta x_{j+1}^- u^{**}_{j+\frac{1}{2}} - p_{j+1}\right) - \left(u^{**}_{j+\frac{1}{2}} - u_{j+1}\right) = 0. \end{cases}$$

The solution writes

$$\begin{cases} p^{**}_{j+\frac{1}{2}} = \frac{1-\sigma\Delta x_{j+1}^-}{2+\sigma\Delta x_j^+ - \sigma\Delta x_{j+1}^-}(p_j - u_j) + \frac{1+\sigma\Delta x_j^+}{2+\sigma\Delta x_j^+ - \sigma\Delta x_{j+1}^-}(p_{j+1} - u_{j+1}), \\ u^{**}_{j+\frac{1}{2}} = \frac{1}{2+\sigma\Delta x_j^+ - \sigma\Delta x_{j+1}^-}(u_j + u_{j+1} + p_j - p_{j+1}). \end{cases}$$

This solver coincides exactly with the Gosse-Toscani solver in [17].

4. Application to the S_n model

The S_n model is a natural way to approach the transfer equation

$$\partial_t I + \mu \partial_x I = \sigma (< I > - I), \quad -1 \leq \mu \leq 1, \tag{49}$$

where the mean value is $< I >= \frac{1}{2}\int_{-1}^{1} I(\mu) d\mu$ and the velocity $\mu = \cos\theta$ is representative of a direction. This is an integro-differential equation which has been widely studied in relation with the theory of propagation of light, neutrons and other types of particles. We refer to the seminal contributions of [10, 9, 12] and to the more dedicated work [27]. Most of the algebra we show come from [10] and is given to get complete interpretation of some recent schemes proposed in [15].

One usually choose a set of velocities $0 < \mu_1 < \mu_2 < \ldots < \mu_n \leq 1$ and positive weights $w_i > 0$ such that $2\sum_{i=1} w_i = 1$ and $2\sum_{i=1} w_i \mu_i^2 = \frac{1}{3}$. The velocities are actually the cosine of some angles and this is why there are less than one. The intensity is represented by

$$I(x,t,\mu) = \sum_{i=1} w_i \widehat{f_i}(x,t)\delta(\mu - \mu_i) + \sum_{i=1} w_i \widehat{g_i}(x,t)\delta(\mu + \mu_i), \tag{50}$$

where δ indicates a Dirac mass. The $\left(\widehat{f_i}\right)$ and $(\widehat{g_i})$ are solutions to

$$\begin{cases} \partial_t \widehat{f_i} + \mu_i \partial_x \widehat{f_i} = \sigma\left(\sum_{j=1} w_j (\widehat{f_j} + \widehat{g_j}) - \widehat{f_i}\right), \\ \partial_t \widehat{g_i} - \mu_i \partial_x \widehat{g_i} = \sigma\left(\sum_{j=1} w_j (\widehat{f_j} + \widehat{g_j}) - \widehat{g_i}\right). \end{cases} \tag{51}$$

Let us define the vector $\widehat{U} = \left(\widehat{f}, \widehat{g} \right)^t \in \mathbb{R}^{2n}$ where $\widehat{f} = (\widehat{f_1}, \ldots, \widehat{f_n})^t \in \mathbb{R}^n$ and $\widehat{g} = (\widehat{g_1}, \ldots, \widehat{g_n})^t \in \mathbb{R}^n$. Equations (51) are recast as a system

$$\partial_t \widehat{U} + A \partial_x \widehat{U} = -\widehat{R}\widehat{U} \tag{52}$$

where $A = \begin{pmatrix} D & 0 \\ 0 & -D \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$ (with $D = \mathrm{diag}(\mu_i) \in \mathbb{R}^{n\times n}$), and $-\widehat{R} = \mathbf{1} \otimes \widehat{\mathbf{w}} - I_d$. The notations are $\mathbf{1} = (1, \ldots 1) \in \mathbb{R}^{2n}$, $\widehat{\mathbf{w}} = (w_1, \ldots, w_n, w_1, \ldots, w_n) \in \mathbb{R}^{2n}$ and $I_d = \mathrm{diag}(1) \in \mathbb{R}^{n\times n}$ the identity matrix.

Let us define $E = \mathrm{diag}(w_i) \in \mathbb{R}^{n\times n}$. The system (52) is easily rewritten in a symmetric form with the definition of $U = (f,g)^t = (\sqrt{E}\,\widehat{f}, \sqrt{E}\,\widehat{g})^t \in \mathbb{R}^{2n}$ which is solution of

$$\partial_t U + A\partial_x U = -RU, \qquad R = \sigma\,(I_d - \mathbf{w} \otimes \mathbf{w}), \tag{53}$$

where the relaxation matrix is symmetric. Since the coefficient is non negative, that is $\sigma \geq 0$, and the new vector $\mathbf{w} = (\sqrt{w_1}, \ldots, \sqrt{w_n}, \sqrt{w_1}, \ldots, \sqrt{w_n}) \in \mathbb{R}^{2n}$ is such that $\|\mathbf{w}\| = 1$, the relaxation matrix is now non negative, that is $R \geq 0$. Under this symmetrized form, one may apply directly the previous theory since A is non singular. It remains to characterize the one-state solver and the two-states solver. This will be performed using some basic symmetry principles in combination with the transmission and reflexion operators which are fundamental in the theory of light propagation. Some parts of the following analysis can be adapted to more general systems as in [8][page 1558].

4.1. The transmission and reflexion operators

Denoting $Q = A^{-1}R$, the matrix exponential $e^{-Qx} = e^{-A^{-1}Rx}$ describes the result of the propagation of rays of light (50) through a slab $[0,x]$ of material with absorption σ. The orientation of the slab is positive, that is $x \geq 0$. One has the relation

$$\begin{pmatrix} f(x) \\ g(x) \end{pmatrix} = e^{-A^{-1}Rx} \begin{pmatrix} f(0) \\ g(0) \end{pmatrix}. \tag{54}$$

The theory of light shows it is always possible to define a transmission operator (matrix) $\mathcal{T}(x) \in \mathbb{R}^{n\times n}$ and a reflection operator (matrix) $\mathcal{S}(x) \in \mathbb{R}^{n\times n}$ so that one can compute the pair $(f(x), g(0))$ in function of $(f(0), g(x))$. These operators can be defined as a consequence of an energy identity which is obtained by integration by part for stationary solutions of (53)

$$(Df(0), f(0)) + (Dg(x), g(x)) = (Df(x), f(x)) + (Dg(0), g(0)) + \int_0^x (RU(s), U(s)) ds. \tag{55}$$

Since $R \geq 0$ and $D > 0$, it shows that all quantities vanish if $f(0) = g(x) = 0$. So the linear transformation $\begin{pmatrix} f(0) \\ g(x) \end{pmatrix} = \mathcal{L}\begin{pmatrix} f(0) \\ g(0) \end{pmatrix}$ is non singular (that is invertible), and all unknowns can be expressed in function of the vector $\begin{pmatrix} f(0) \\ g(x) \end{pmatrix}$. It means one can define operators (matrices) $\mathcal{T}_1(x), \mathcal{T}_2(x), \mathcal{S}_1(x), \mathcal{S}_2(x) \in \mathbb{R}^{n\times n}$ such

$$\begin{pmatrix} f(x) \\ g(0) \end{pmatrix} = \begin{pmatrix} \mathcal{T}_1(x) & \mathcal{S}_1(x) \\ \mathcal{S}_2(x) & \mathcal{T}_2(x) \end{pmatrix}\begin{pmatrix} f(0) \\ g(x) \end{pmatrix}.$$

Based on symmetry considerations for the physical problem, one has that $\mathcal{T}_2(x) = \mathcal{T}_1(x)$ and $\mathcal{S}_2(x) = \mathcal{S}_1(x)$. Therefore one can write

$$\begin{pmatrix} f(x) \\ g(0) \end{pmatrix} = \begin{pmatrix} \mathcal{T}(x) & \mathcal{S}(x) \\ \mathcal{S}(x) & \mathcal{T}(x) \end{pmatrix}\begin{pmatrix} f(0) \\ g(x) \end{pmatrix} \tag{56}$$

after dropping the indices. The **transmission or transfer operator** is $\mathcal{T}(x)$, the reflexion or **scattering operator** is $\mathcal{S}(x)$. The matrix in (56) is the scattering matrix. The **input** is the pair $(f(0), g(x))$ and the **output** is $(f(x), g(0))$. Of course $\mathcal{T}(0) = I_d$ and $\mathcal{S}(0) = 0$. Some basic properties can be proved, which are used to justify the fundamental composition relations (64).

Proposition 8. *The transmission operator is non singular and the reflexion operator is strictly bounded, that is*

$$det\,(\mathcal{T}(x)) \neq 0 \text{ and } \left\|D^{\frac{1}{2}}\mathcal{S}(x)D^{-\frac{1}{2}}\right\| < 1, \qquad \forall x \geq 0. \tag{57}$$

Proof. The proof is by contradiction.
• Assume $\mathcal{T}$ to be singular. There would exist a non zero vector $W \in \mathbb{R}^n$ such that

$$\begin{pmatrix} \mathcal{T}(x) & \mathcal{S}(x) \\ \mathcal{S}(x) & \mathcal{T}(x) \end{pmatrix}\begin{pmatrix} W \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \dots \end{pmatrix}, \quad W \neq 0.$$

It yields a non zero solution of (52) with vanishing Cauchy data at the end point x, that is $f(0) = W \neq 0$ and $f(x) = g(x) = 0$. But this is impossible (54). This is why the transmission operator is non singular.
• Assume now that $\left\|D^{\frac{1}{2}}\mathcal{S}(x)D^{-\frac{1}{2}}\right\| \geq 1$. There would exist a non zero vector $Z \in \mathbb{R}^n$ such that $\left\|D^{\frac{1}{2}}\mathcal{S}(x)D^{-\frac{1}{2}}Z\right\| \geq \|Z\| \neq 0$. We set $W = D^{-\frac{1}{2}}Z$ so that

$$\left\|D^{\frac{1}{2}}\mathcal{S}(x)W\right\| \geq \left\|D^{\frac{1}{2}}W\right\| \text{ with } W \neq 0. \tag{58}$$

On the other hand the identity

$$\begin{pmatrix} \mathcal{T}(x) & \mathcal{S}(x) \\ \mathcal{S}(x) & \mathcal{T}(x) \end{pmatrix}\begin{pmatrix} 0 \\ W \end{pmatrix} = \begin{pmatrix} \mathcal{S}(x)W \\ \mathcal{T}(x)W \end{pmatrix}$$

can be interpreted with the energy identity (55), where the input is $(f(0), g(x)) = (0, W)$ and the ouput is $(f(x), g(0)) = (\mathcal{S}(x)W, \mathcal{T}(x)W)$. The energy identity yields the inequality $\left\|D^{\frac{1}{2}}W\right\|^2 \geq \left\|D^{\frac{1}{2}}\mathcal{S}(x)W\right\|^2 + \left\|D^{\frac{1}{2}}\mathcal{T}(x)W\right\|^2$. Using (58), one obtains $D^{\frac{1}{2}}\mathcal{T}(x)W = 0$ and $W = 0$ since $D^{\frac{1}{2}}\mathcal{T}(x)$ is invertible. This is a contradiction. The proof is ended.

□

Proposition 9. *All coefficients of $\mathcal{T}(x)$ and $\mathcal{S}(x)$ are non negative, and one has the identity*

$$(\mathcal{T}(x) + \mathcal{S}(x))^t\,\mathbf{z} = \mathbf{z} \tag{59}$$

where $\mathbf{z} = \left(\frac{\mu_1}{\sqrt{w_1}}, \dots, \frac{\mu_n}{\sqrt{w_n}}\right) \in \mathbb{R}^n$.

Proof. A proof of (59) can be performed using the identity $R\mathbf{w} = 0$. But it is better for physical intuition to start from the stationary equations

$$\begin{cases} \mu_i\partial_x\widehat{f_i} &= \sigma\left(\sum_{j=1} w_j(\widehat{f_j} + \widehat{g_j}) - \widehat{f_i}\right), \\ -\mu_i\partial_x\widehat{g_i} &= \sigma\left(\sum_{j=1} w_j(\widehat{f_j} + \widehat{g_j}) - \widehat{g_i}\right). \end{cases} \tag{60}$$

It yields after integration $\sum_{i=1}\mu_i\left(\widehat{f_i}(0)+\widehat{g_i}(x)\right)=\sum_{i=1}\mu_i\left(\widehat{f_i}(x)+\widehat{g_i}(0)\right)$. The change of unknowns $\widehat{f_i}=w_i^{-\frac{1}{2}}f_i$ and $\widehat{g_i}=w_i^{-\frac{1}{2}}g_i$ yields $\sum_{i=1}z_i\left(f_i(0)+g_i(x)\right)=\sum_{i=1}z_i\left(f_i(x)+g_i(0)\right)$, that is in vectorial form

$$\left(\begin{pmatrix} f(x)\\ g(0)\end{pmatrix},\begin{pmatrix}\mathbf{z}\\ \mathbf{z}\end{pmatrix}\right)=\left(\begin{pmatrix} f(0)\\ g(x)\end{pmatrix},\begin{pmatrix}\mathbf{z}\\ \mathbf{z}\end{pmatrix}\right).$$

The representation (56) yields

$$\left(\begin{pmatrix} \mathcal{T}(x) & \mathcal{S}(x)\\ \mathcal{S}(x) & \mathcal{T}(x)\end{pmatrix}\begin{pmatrix} f(0)\\ g(x)\end{pmatrix},\begin{pmatrix}\mathbf{z}\\ \mathbf{z}\end{pmatrix}\right)=\left(\begin{pmatrix} f(0)\\ g(x)\end{pmatrix},\begin{pmatrix}\mathbf{z}\\ \mathbf{z}\end{pmatrix}\right)$$

which can be simplified into

$$\left(f(0)+g(x),(\mathcal{T}(x)+\mathcal{S}(x))^t\,\mathbf{z}\right)=\left(f(0)+g(x),\mathbf{z}\right).$$

Since this is true for all $f(0)+g(x)\in\mathbb{R}^n$, it shows the identity (59).

Next we define $\widehat{h_i}(x-y)=\widehat{g_i}(y)$ for $y\in[0,x]$. The system (60) recasts as $(0\le y\le x)$

$$\begin{cases}\mu_i\partial_x\widehat{f_i}(y)+\sigma\widehat{f_i}(y) &= \sigma\sum_{j=1}w_j\left(\widehat{f_j}(y)+\widehat{h_j}(x-y)\right),\\ \mu_i\partial_x\widehat{h_i}(y)+\sigma\widehat{h_i}(y) &= \sigma\sum_{j=1}w_j\left(\widehat{f_j}(x-y)+\widehat{h_j}(y)\right).\end{cases}$$

For such a system a standard property is that if $\widehat{f_i}(0)\ge 0$ and $\widehat{h_i}(0)\ge 0$ for all i, then $\widehat{f_i}(x)\ge 0$ and $\widehat{h_i}(x)\ge 0$ for all i. It shows the first part of the claim after a change of unknowns. The proof is ended. □

Proposition 10. *Assume that $x\ge 0$. One has the representation formula*

$$e^{-A^{-1}Rx}=\begin{pmatrix}\mathcal{T}-\mathcal{S}\mathcal{T}^{-1}\mathcal{S} & \mathcal{S}\mathcal{T}^{-1}\\ -\mathcal{T}^{-1}\mathcal{S} & \mathcal{T}^{-1}\end{pmatrix}(x). \tag{61}$$

The inverse formula is

$$e^{A^{-1}Rx}=\begin{pmatrix}\mathcal{T}^{-1} & -\mathcal{T}^{-1}\mathcal{S}\\ \mathcal{S}\mathcal{T}^{-1} & \mathcal{T}-\mathcal{S}\mathcal{T}^{-1}\mathcal{S}\end{pmatrix}(x). \tag{62}$$

This formula shows that the matrix on the left hand side which belongs to $\mathbb{R}^{2n\times 2n}$ can be represented with two matrices in $\mathbb{R}^{n\times n}$, thus resulting in a reduction of the size of the basic objects used in the representation formulas.

Proof. Indeed (56) shows that $g(x)=\mathcal{T}^{-1}(x)(-\mathcal{S}(x)f(0)+g(0))$ and

$$f(x)=\left(\mathcal{T}(x)-\mathcal{S}(x)\mathcal{T}^{-1}(x)\mathcal{S}(x)\right)f(0)+\mathcal{T}^{-1}(x)\mathcal{S}(x)g(0).$$

It ends the proof of (61). That (62) is inverse formula of (61) is evident. □

Proposition 11. *The transposed operators are $\mathcal{T}^t=D\mathcal{T}D^{-1}$ and $\mathcal{S}^t=D\mathcal{S}D^{-1}$.*

Proof. Since $\left(e^{-A^{-1}Rx}\right)^t=e^{-RA^{-1}x}=Ae^{-A^{-1}Rx}A^{-1}$ and A is block diagonal with D and $-D$ on the diagonal, one has from (61) $\mathcal{T}^{-t}=(-D)\mathcal{T}^{-1}(-D)^{-1}$ which yields the first relation. The second relation can be deduced from $\left(-\mathcal{T}^{-1}\mathcal{S}\right)^t=D\left(\mathcal{S}\mathcal{T}^{-1}\right)(-D)^{-1}$ which yields the second relation after elimination of the transmission operator. □

Let us consider $x\le y\le z$ together with $\mathcal{T}_1=\mathcal{T}(y-x)$, $\mathcal{S}_1=\mathcal{S}(y-x)$, $\mathcal{T}_2=\mathcal{T}(z-y)$, $\mathcal{S}_2=\mathcal{S}(z-y)$, $\mathcal{T}_3=\mathcal{T}(z-x)$ and $\mathcal{S}_3=\mathcal{S}(z-x)$. Composition formulas can be deduced for the transmission and reflexion operators. It comes from the composition-commutation relations

$$e^{-A^{-1}Rd_3}=e^{-A^{-1}Rd_2}e^{-A^{-1}Rd_1}=e^{-A^{-1}Rd_1}e^{-A^{-1}Rd_2},\quad d_3=d_1+d_2. \tag{63}$$

Proposition 12. *One has the composition formulas*

$$\left\{\begin{array}{lll} \mathcal{T}_3 & = \mathcal{T}_1 (I_d - \mathcal{S}_2\mathcal{S}_1)^{-1} \mathcal{T}_2 & = \mathcal{T}_2 (I_d - \mathcal{S}_1\mathcal{S}_2)^{-1} \mathcal{T}_1, \\ \mathcal{S}_3 & = \mathcal{S}_2 + \mathcal{T}_2\mathcal{S}_1 (I_d - \mathcal{S}_2\mathcal{S}_1)^{-1} \mathcal{T}_2 & = \mathcal{S}_2 + \mathcal{T}_2 (I_d - \mathcal{S}_1\mathcal{S}_2)^{-1} \mathcal{S}_1\mathcal{T}_2 \\ & = \mathcal{S}_1 + \mathcal{T}_1\mathcal{S}_2 (I_d - \mathcal{S}_1\mathcal{S}_2)^{-1} \mathcal{T}_1 & = \mathcal{S}_1 + \mathcal{T}_1 (I_d - \mathcal{S}_2\mathcal{S}_1)^{-1} \mathcal{S}_2\mathcal{T}_1. \end{array}\right. \tag{64}$$

where the matrices $I_d - \mathcal{S}_2\mathcal{S}_1$ and $I_d - \mathcal{S}_1\mathcal{S}_2$ can be shown non singular as consequence of the bound in (57).

Proof. The composition relation (63) yields

$$\begin{pmatrix} \mathcal{T}_3 - \mathcal{S}_3\mathcal{T}_3^{-1}\mathcal{S}_3 & \mathcal{S}_3\mathcal{T}_3^{-1} \\ -\mathcal{T}_3^{-1}\mathcal{S}_3 & \mathcal{T}_3^{-1} \end{pmatrix} = \begin{pmatrix} \mathcal{T}_2 - \mathcal{S}_2\mathcal{T}_2^{-1}\mathcal{S}_2 & \mathcal{S}_2\mathcal{T}_2^{-1} \\ -\mathcal{T}_2^{-1}\mathcal{S}_2 & \mathcal{T}_2^{-1} \end{pmatrix} \begin{pmatrix} \mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1 & \mathcal{S}_1\mathcal{T}_1^{-1} \\ -\mathcal{T}_1^{-1}\mathcal{S}_1 & \mathcal{T}_1^{-1} \end{pmatrix}. \tag{65}$$

• It yields $\mathcal{T}_3^{-1} = -\mathcal{T}_2^{-1}\mathcal{S}_2\mathcal{S}_1\mathcal{T}_1^{-1} + \mathcal{T}_2^{-1}\mathcal{T}_1^{-1} = \mathcal{T}_2^{-1} (I_d - \mathcal{S}_2\mathcal{S}_1) \mathcal{T}_1^{-1}$. It shows the first identity of the claim is deduced after inversion. Since the matrix exponentials commute, the other identity $\mathcal{T}_3 = \mathcal{T}_2 (I_d - \mathcal{S}_1\mathcal{S}_2)^{-1} \mathcal{T}_1$ is immediate after switching the indices.
• One also has from (65)

$$\mathcal{S}_3\mathcal{T}_3^{-1} = \left(\mathcal{T}_2 - \mathcal{S}_2\mathcal{T}_2^{-1}\mathcal{S}_2\right)\mathcal{S}_1\mathcal{T}_1^{-1} + \mathcal{S}_2\mathcal{T}_2^{-1}\mathcal{T}_1^{-1} = \mathcal{T}_2\mathcal{S}_1\mathcal{T}_1^{-1} + \mathcal{S}_2\mathcal{T}_2^{-1} (I_d - \mathcal{S}_2\mathcal{S}_1) \mathcal{T}_1^{-1}.$$

Therefore

$$\mathcal{S}_3 = \left[\mathcal{T}_2\mathcal{S}_1\mathcal{T}_1^{-1} + \mathcal{S}_2\mathcal{T}_2^{-1} (I_d - \mathcal{S}_2\mathcal{S}_1) \mathcal{T}_1^{-1}\right]\left[\mathcal{T}_1 (I_d - \mathcal{S}_2\mathcal{S}_1)^{-1} \mathcal{T}_2\right] = \mathcal{T}_2\mathcal{S}_1 (I_d - \mathcal{S}_2\mathcal{S}_1)^{-1} \mathcal{T}_2 + \mathcal{S}_2$$

which is the third identity of the claim. The fourth one comes from the identity $\mathcal{S}_1 (I_d - \mathcal{S}_2\mathcal{S}_1)^{-1} = (I_d - \mathcal{S}_1\mathcal{S}_2)^{-1} \mathcal{S}_1$. The two last ones are obtained by switching the indices. □

4.2. The two-states solver for S_n model

The two-states solver (22)-(25), $\frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{1}{\Delta x_j}\left(e^{RA^{-1}\Delta x_j^+} A U_{j+\frac{1}{2}}^{**} - e^{RA^{-1}\Delta x_j^-} A U_{j-\frac{1}{2}}^{**}\right) = 0$, can be rewritten under the general form

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{1}{\Delta x_j} A\left(e^{A^{-1}R\Delta x_j^+} U_{j+\frac{1}{2}}^{**} - e^{A^{-1}R\Delta x_j^-} U_{j-\frac{1}{2}}^{**}\right) = 0. \tag{66}$$

With some additional algebra, one can prove the following elegant and compact result, which shows this method is also equal to the one published in [15] and [16][formula (9.18) page 173], but is different from the one from [22].

Proposition 13. *The two-states solver for the S_n model can be written under the form*

$$\left\{\begin{array}{l} \dfrac{f_j^{n+1} - f_j^n}{\Delta t} + D\dfrac{f_j^n - \mathcal{T}_3^{j-\frac{1}{2}} f_{j-1}^n - \mathcal{S}_3^{j-\frac{1}{2}} g_j^n}{\Delta x_j} = 0, \\ \dfrac{g_j^{n+1} - g_j^n}{\Delta t} - D\dfrac{\mathcal{S}_3^{j+\frac{1}{2}} f_j^n + \mathcal{T}_3^{j+\frac{1}{2}} g_{j+1}^n - g_j^n}{\Delta x_j} = 0. \end{array}\right. \tag{67}$$

where $\mathcal{T}_3^{j-\frac{1}{2}} = \mathcal{T}_3(x_j - x_{j-1})$ and $\mathcal{S}_3^{j-\frac{1}{2}} = \mathcal{S}_3(x_j - x_{j-1})$ for all j.

Proof. In the proof we do not note the reference to the interface for the transmission and reflection operators. Let us write the flux as $U_{j+\frac{1}{2}}^{**} = \left(f_{j+\frac{1}{2}}^{**}, g_{j+\frac{1}{2}}^{**}\right)^t$, defined for all j by (46). We observe also that the eigenvectors u_p in (46) are extremely simple to determine. Indeed they are of type $u_p = (f_p, 0)$ for $\lambda_p > 0$, and of type $u_p = (0, g_p)$ for $\lambda_p < 0$. With natural notations the system (46) writes

$$\begin{pmatrix} \mathcal{T}_1^{-1} & -\mathcal{T}_1^{-1}\mathcal{S}_1 \\ -\mathcal{T}_2^{-1}\mathcal{S}_2 & \mathcal{T}_2^{-1} \end{pmatrix} \begin{pmatrix} f_{j+\frac{1}{2}}^{**} \\ g_{j+\frac{1}{2}}^{**} \end{pmatrix} = \begin{pmatrix} f_j \\ g_{j+1} \end{pmatrix}.$$

By comparison with (62) one has $e^{A^{-1}R\Delta x_j^+} = \begin{pmatrix} \mathcal{T}_1^{-1} & -\mathcal{T}_1^{-1}\mathcal{S}_1 \\ \mathcal{S}_1\mathcal{T}_1^{-1} & \mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1 \end{pmatrix}$ with the notation $\mathcal{T}_1 = \mathcal{T}(\Delta x_j^+)$ and $\mathcal{S}_1 = \mathcal{S}(\Delta x_j^+)$. Therefore we define (with natural notations)

$$\begin{pmatrix} f_{j+\frac{1}{2}}^- \\ g_{j+\frac{1}{2}}^- \end{pmatrix} = e^{A^{-1}R\Delta x_j^+} U_{j+\frac{1}{2}}^{**} = \underbrace{\begin{pmatrix} \mathcal{T}_1^{-1} & -\mathcal{T}_1^{-1}\mathcal{S}_1 \\ \mathcal{S}_1\mathcal{T}_1^{-1} & \mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1 \end{pmatrix}}_{=\mathcal{A}} \underbrace{\begin{pmatrix} \mathcal{T}_1^{-1} & -\mathcal{T}_1^{-1}\mathcal{S}_1 \\ -\mathcal{T}_2^{-1}\mathcal{S}_2 & \mathcal{T}_2^{-1} \end{pmatrix}^{-1}}_{=\mathcal{B}} \begin{pmatrix} f_j \\ g_{j+1} \end{pmatrix}.$$

One has the algebra

$$\mathcal{A}\mathcal{B} = \begin{pmatrix} \mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1 & \mathcal{S}_1\mathcal{T}_1^{-1} \\ -\mathcal{T}_1^{-1}\mathcal{S}_1 & \mathcal{T}_1^{-1} \end{pmatrix}^{-1} \mathcal{B} = \left[\begin{pmatrix} \mathcal{T}_1^{-1} & -\mathcal{T}_1^{-1}\mathcal{S}_1 \\ -\mathcal{T}_2^{-1}\mathcal{S}_2 & \mathcal{T}_2^{-1} \end{pmatrix} \begin{pmatrix} \mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1 & \mathcal{S}_1\mathcal{T}_1^{-1} \\ -\mathcal{T}_1^{-1}\mathcal{S}_1 & \mathcal{T}_1^{-1} \end{pmatrix} \right]^{-1}$$

$$= \begin{pmatrix} I & 0 \\ -\mathcal{S} & \mathcal{T}_2^{-1}(I - \mathcal{S}_2\mathcal{S}_1)\mathcal{T}_1^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -\mathcal{S} & \mathcal{T}_3^{-1} \end{pmatrix}^{-1}$$

with $\mathcal{S} = \mathcal{T}_2^{-1}\mathcal{S}_2\mathcal{T}_1 + \mathcal{T}_2^{-1}(I - \mathcal{S}_2\mathcal{S}_1))\mathcal{T}_1^{-1}\mathcal{S}_1$. So $\mathcal{A}\mathcal{B} = \begin{pmatrix} I & 0 \\ \mathcal{T}_3\mathcal{S} & \mathcal{T}_3 \end{pmatrix}$. One can check that

$$\mathcal{T}_3\mathcal{S} = \left[\mathcal{T}_1(I_d - \mathcal{S}_2\mathcal{S}_1)^{-1}\mathcal{T}_2\right]\left[\mathcal{T}_2^{-1}\mathcal{S}_2\mathcal{T}_1 + \mathcal{T}_2^{-1}(I - \mathcal{S}_2\mathcal{S}_1))\mathcal{T}_1^{-1}\mathcal{S}_1\right] = \mathcal{T}_1(I_d - \mathcal{S}_2\mathcal{S}_1)^{-1}\mathcal{S}_2\mathcal{T}_1 + \mathcal{S}_1 = \mathcal{S}_3$$

using the last identity of (64). Therefore one can write $\begin{pmatrix} f_{j+\frac{1}{2}}^- \\ g_{j+\frac{1}{2}}^- \end{pmatrix} = \begin{pmatrix} f_j \\ \mathcal{S}_3 f_j + \mathcal{T}_3 g_{j+1} \end{pmatrix}$. By symmetry one gets $\begin{pmatrix} f_{j-\frac{1}{2}}^+ \\ g_{j-\frac{1}{2}}^+ \end{pmatrix} = \begin{pmatrix} \mathcal{S}_3 f_{j-1} + \mathcal{S}_3 g_j \\ g_j \end{pmatrix}$. The proof is ended. □

A similar algebra can be performed for the one-state solver, but the results are less interesting for the moment. One starts from

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{1}{\Delta x_j} A\left(e^{A^{-1}R\Delta x_j^+} U_{j+\frac{1}{2}}^* - e^{A^{-1}R\Delta x_j^-} U_{j-\frac{1}{2}}^*\right) = 0. \tag{68}$$

It remains to identify the flux $U_{j+\frac{1}{2}}^* = \left(f_{j+\frac{1}{2}}^*, g_{j+\frac{1}{2}}^*\right)^t$ which is defined for all j by (29). The operator that shows up in the first line of (29) is simply $e^{-A^{-1}R\Delta x_j^+}$ for which we can use the direct formula (61). Same kind of remarks hold for the second line of (29) but with the second line of (62) used with $-x = -\Delta x_{j+1}^- > 0$. So one finds out that

$$\begin{cases} f_{j+\frac{1}{2}}^* = \left(\mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1\right) f_j + \mathcal{S}_1\mathcal{T}_1^{-1} g_j, \\ g_{j+\frac{1}{2}}^* = \mathcal{S}_2\mathcal{T}_2^{-1} f_{j+1} + \left(\mathcal{T}_2 - \mathcal{S}_2\mathcal{T}_2^{-1}\mathcal{S}_2\right) g_{j+1}, \end{cases}$$

where $\mathcal{T}_2 = \mathcal{T}(-\Delta x_{j+1}^-)$ and $\mathcal{S}_1 = \mathcal{S}(-\Delta x_{j+1}^-)$. Therefore we define with natural notations

$$\begin{pmatrix} f_{j+\frac{1}{2}}^- \\ g_{j+\frac{1}{2}}^- \end{pmatrix} = e^{A^{-1}R\Delta x_j^+} U_{j-\frac{1}{2}}^* = \begin{pmatrix} \mathcal{T}_1^{-1} & -\mathcal{T}_1^{-1}\mathcal{S}_1 \\ \mathcal{S}_1\mathcal{T}_1^{-1} & \mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1 \end{pmatrix} \begin{pmatrix} \left(\mathcal{T}_1 - \mathcal{S}_1\mathcal{T}_1^{-1}\mathcal{S}_1\right) f_j + \mathcal{S}_1\mathcal{T}_1^{-1} g_j \\ \mathcal{S}_2\mathcal{T}_2^{-1} f_{j+1} + \left(\mathcal{T}_2 - \mathcal{S}_2\mathcal{T}_2^{-1}\mathcal{S}_2\right) g_{j+1} \end{pmatrix}.$$

It yields the expression of the fluxes in terms of the reflexion and transmission operators. However this formulation is still quite involved and will not be pursued in this work.

Theorem 2. *The scheme (67) is positive, L^∞ and L^1 stable under CFL*

Proof. The positivity of the first line of (67) is a consequence of the explicit Euler reformulation

$$f_j^{n+1} = \left(I - \frac{\Delta t}{\Delta x_j} D\right) f_j^n + \frac{\Delta t}{\Delta x_j} D\mathcal{T}_3^{j-\frac{1}{2}} f_{j-1}^n + \frac{\Delta t}{\Delta x_j} D\mathcal{S}_3^{j-\frac{1}{2}} g_j^n. \tag{69}$$

Assuming the CFL condition $\frac{\Delta t}{\Delta x_j} \times$ spectral radius$(D) \leq 1$, all matrices are non negative. Assume the vectors f_j^n, f_{j-1}^n and g_j^n have non negative coefficients. Therefore the vector f_j^{n+1} has also non negative coefficients. A similar algebra yields for the second line of (67)

$$g_j^{n+1} = \left(I - \frac{\Delta t}{\Delta x_j} D\right) g_j^n + \frac{\Delta t}{\Delta x_j} D\mathcal{S}_3^{j+\frac{1}{2}} f_j^n + \frac{\Delta t}{\Delta x_j} D\mathcal{T}_3^{j+\frac{1}{2}} g_j^n. \tag{70}$$

Let $\mathbf{z} \in \mathbb{R}^n$ be the vector with positive coefficients defined in proposition 9. Using the property of the proposition, one can rewrite (69-70) as

$$\begin{cases} \left(c\mathbf{z} - f_j^{n+1}\right) = \left(I - \frac{\Delta t}{\Delta x_j} D\right)\left(c\mathbf{z} - f_j^n\right) + \frac{\Delta t}{\Delta x_j} D\mathcal{T}_3^{j-\frac{1}{2}} \left(c\mathbf{z} - f_{j-1}^n\right) + \frac{\Delta t}{\Delta x_j} D\mathcal{S}_3^{j-\frac{1}{2}} \left(c\mathbf{z} - g_j^n\right), \\ \left(c\mathbf{z} - g_j^{n+1}\right) = \left(I - \frac{\Delta t}{\Delta x_j} D\right)\left(c\mathbf{z} - g_j^n\right) + \frac{\Delta t}{\Delta x_j} D\mathcal{S}_3^{j+\frac{1}{2}} \left(c\mathbf{z} - f_j^n\right) + \frac{\Delta t}{\Delta x_j} D\mathcal{T}_3^{j+\frac{1}{2}} \left(c\mathbf{z} - g_j^n\right). \end{cases} \tag{71}$$

Assuming that data are bounded in L^∞ at step t_n, it is possible to find $c > 0$ large enough so that all quantities are non negative at t_n, and remain non negative at time t_{n+1}. It shows an upper bound which implies the L^∞ stability (under CFL).

To show the stability in L^1 it is sufficient to consider that non negative solutions of (69-70) satisfy one conservation law. In view of proposition 9, one readily proves that (69-70) imply

$$\sum_j \Delta x_j \left(D^{-1}\mathbf{z}, f_j^{n+1} + g_j^{n+1}\right) = \sum_j \Delta x_j D^{-1} \left(D^{-1}\mathbf{z}, f_j^n + g_j^n\right). \tag{72}$$

Therefore non negative initial data which are bounded in L^1 remain bounded in L^1. The case of a general initial data bounded in L^1 is as usual treated with a splitting of between positive and negative parts.

□

Remark 7. *To prove the stability in L^2, it is possible to redo the analysis using identities like $\mathcal{T}^t D\mathcal{T} + \mathcal{S}^t D\mathcal{S} \leq D$ which are consequences of the energy identity (55). This strategy is more involved and so is not developed in this work. Another possibility is to use the Riesz-Thorin theorem in interpolation theory [28] which automatically yields that the stability in L^1 and L^∞ yields the stability in any L^p, $1 \leq p \leq \infty$.*

The conservation law (72) in the proof corresponds to the preservation of the total energy $\frac{d}{dt}\int_{\mathbb{R}}\int_{-1}^{1} I dx d\mu = 0$ that comes from (49). Additional ones stemming from the general structure and rewriting in terms of the unknown α are given in proposition 14.

As stressed previously, all these schemes have a fully conservative interpretation using the variable α. A verification for the two-states S_n solver is as followed.

Proposition 14. *Consider the two-states solver (67). Assume for simplicity that the initialization has compact support, that is $\left(f_j^0, g_j^0\right) = 0$ for $|j| \geq M \in \mathbb{N}$. Then the solution satisfies $2n$ conservation which can be expressed as*

$$\sum_{j\in\mathbb{Z}} e^{RA^{-1}j\Delta x} \begin{pmatrix} f_j^n \\ g_j^n \end{pmatrix} = \sum_{j\in\mathbb{Z}} e^{RA^{-1}j\Delta x} \begin{pmatrix} f_j^0 \\ g_j^0 \end{pmatrix}, \quad \forall n \in \mathbb{N}.$$

Proof. The equality rewrites as $\sum_j \alpha_j^n = \sum_j \alpha_j^0$ which the conservation laws for the variable α. It is possible to express the powers $e^{RA^{-1}j\Delta x} = \left(e^{RA^{-1}\Delta x}\right) j$ in function of the transmission and reflexion operators. To prove the equality we consider the difference of two iterates. One has

$$\Delta x \sum_j e^{RA^{-1}j\Delta x} \begin{pmatrix} f_j^{n+1} - f_j^n \\ g_j^{n+1} - g_j^n \end{pmatrix} = -\sum_j e^{RA^{-1}j\Delta x} A \begin{pmatrix} f_j^n - \mathcal{T}_3 f_{j-1}^n - \mathcal{S}_3 g_j^n \\ \mathcal{S}_3 f_j^n + \mathcal{T}_3 g_{j+1}^n - g_j^n \end{pmatrix}$$

$$= -A \sum_j e^{A^{-1}Rj\Delta x} \left[\begin{pmatrix} f_j^n - \mathcal{S}_3 g_j^n \\ \mathcal{S}_3 f_j^n - g_j^n \end{pmatrix} + \begin{pmatrix} -\mathcal{T}_3 f_{j-1}^n \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \mathcal{T}_3 g_{j+1}^n \end{pmatrix} \right]$$

$$= -A \sum_j e^{A^{-1}Rj\Delta x} \left[\underbrace{\begin{pmatrix} f_j^n - \mathcal{S}_3 g_j^n \\ \mathcal{S}_3 f_j^n - g_j^n \end{pmatrix} + e^{A^{-1}R\Delta x} \begin{pmatrix} -\mathcal{T}_3 f_j^n \\ 0 \end{pmatrix} + e^{-A^{-1}R\Delta x} \begin{pmatrix} 0 \\ \mathcal{T}_3 g_j^n \end{pmatrix}}_{=W_j^n} \right].$$

The representation formulas (61-62) show that

$$W_j^n = \begin{pmatrix} f_j^n - \mathcal{S}_3 g_j^n \\ \mathcal{S}_3 f_j^n - g_j^n \end{pmatrix} + e^{A^{-1}R\Delta x} \begin{pmatrix} -\mathcal{T}_3 f_j^n \\ 0 \end{pmatrix} + e^{-A^{-1}R\Delta x} \begin{pmatrix} 0 \\ \mathcal{T}_3 g_j^n \end{pmatrix}$$

$$= \begin{pmatrix} f_j^n - \mathcal{S}_3 g_j^n \\ \mathcal{S}_3 f_j^n - g_j^n \end{pmatrix} + \begin{pmatrix} \mathcal{T}_3^{-1} & -\mathcal{T}_3^{-1}\mathcal{S}_3 \\ \mathcal{S}_3\mathcal{T}_3^{-1} & \mathcal{T}_3 - \mathcal{S}_3\mathcal{T}_3^{-1}\mathcal{S}_3 \end{pmatrix} \begin{pmatrix} -\mathcal{T}_3 f_j^n \\ 0 \end{pmatrix} + \begin{pmatrix} \mathcal{T}_3 - \mathcal{S}_3\mathcal{T}_3^{-1}\mathcal{S}_3 & \mathcal{S}_3\mathcal{T}_3^{-1} \\ -\mathcal{T}_3^{-1}\mathcal{S}_3 & \mathcal{T}_3^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ \mathcal{T}_3 g_j^n \end{pmatrix}$$

$$= \begin{pmatrix} f_j^n - \mathcal{S}_3 g_j^n \\ \mathcal{S}_3 f_j^n - g_j^n \end{pmatrix} + \begin{pmatrix} -f_j^n \\ -\mathcal{S}_3 f_j^n \end{pmatrix} + \begin{pmatrix} \mathcal{S}_3 g_j^n \\ g_j^n \end{pmatrix} = 0.$$

Therefore $\Delta x \sum_j e^{RA^{-1}j\Delta x} \begin{pmatrix} f_j^{n+1} - f_j^n \\ g_j^{n+1} - g_j^n \end{pmatrix} = 0$. It shows the result by recurrence. The proof is ended. □

One of these $2n$ conservation laws is actually the physical one $\sum_{j\in\mathbb{Z}} \left(f_j^n + g_j^n, D^{-1}\mathbf{z}\right) = \sum_{j\in\mathbb{Z}} \left(f_j^0 + g_j^0, D^{-1}\mathbf{z}\right)$. It is easily seen using the definition of the scheme (67) and the identity (59).

5. The singular case

For many applications, the matrix A may be singular. We will detail some consequences on a example, which is the P^1 model coupled a linear temperature equation. It writes

$$\begin{cases} \partial_t p & +\partial_x u & = \tau(T - p), \\ \partial_t u & +\partial_x p & = -\sigma u, \\ \partial_t T & & = \tau(p - T), \end{cases} \tag{73}$$

for which

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } R = \begin{pmatrix} \tau & 0 & -\tau \\ 0 & \sigma & 0 \\ -\tau & 0 & \tau \end{pmatrix}.$$

We notice that $A = A^t$ is singular and $R = R^t \geq 0$. Assuming that $\sigma, \tau > 0$, the solutions of the adjoint stationary equation satisfy

$$\partial_x \widehat{u} = 0, \quad \partial_x \widehat{p} = \sigma \widehat{u}, \quad \widehat{T} = \widehat{p}.$$

Notice the additional inequality $\sigma \geq \tau$ for physically motivated problems. One can take $V_1 = (1, 0, 1)$ and $V_2 = (\sigma x, 1, \sigma x)$ which yields two and only two linearly independent functions. So $\dim(\mathcal{V}) = 2 < 3$ which indicates a degeneracy.

Therefore the general method developed previously cannot be used directly and must be adapted. It can be done at least in two directions.

5.1. Shifting of the spectrum

The idea is that stationary solutions (of the adjoint equation) are just particular special solutions. We can use more general particular solutions (A and R being symmetric)

$$\begin{cases} \partial_t V + A\partial_x V = RV, \\ \partial_t V + \xi\partial_x V = 0 \end{cases} \tag{74}$$

where $\xi \in \mathbb{R}$ is an arbitrary real number. It yields $A_\xi \partial_x V = RV$ where $A_\xi = A - \xi I$ is a shift of the matrix A with $\det(A_\xi) \neq 0$. The spectrum is shifted as well. One has to take care that the test functions $V_1^\xi(x,t), \dots, V_n^\xi(x,t)$ particular solutions of (74) depend also on the time variable.

However one can expect some kind of degeneracy as $\xi \to 0$. Since it is difficult to determine in advance the behavior of the method in this regime, we do not pursue in this direction.

5.2. Ad-hoc integration in time

We present the idea on the example (73). Since $\dim(\mathcal{V}) = 2 < 3$, one can write two conservative equations for $\alpha_1 = (U, V_1) = p + T$ and $\alpha_2 = (U, V_2) = \sigma x(p + T) + u$

$$\begin{cases} \partial_t \alpha_1 + \partial_x u = 0, \\ \partial_t \alpha_2 + \partial_x (ux\sigma + p) = 0. \end{cases}$$

Defining arbitrarily $\alpha_3 = T$, one obtains the system

$$\begin{cases} \partial_t \alpha_1 + \partial_x(-\sigma x \alpha_1 + \alpha_2) = 0, \\ \partial_t \alpha_2 + \partial_x \left((1 - \sigma^2 x^2)\alpha_1 + \sigma x \alpha_2 - \alpha_3 \right) = 0, \\ \partial_t \alpha_3 = \tau(\alpha_1 - 2\alpha_3). \end{cases} \tag{75}$$

The last equation is more an ODE and can be integrated in time since $(e^{2\tau t}\alpha_3)' = \tau e^{2\tau t}\alpha_1$. The solution is $\alpha_3(x,t) = e^{-2\tau t}\int_0^t \tau e^{2\tau s}\alpha_1(x,s)ds + e^{-2\tau t}\alpha_3(x,0)$. One can now perform a partial discretization in time during a time step of length Δt. With an explicit Euler formulation for example, it yields

$$\alpha_3(x,\Delta t) \approx \left(e^{-2\tau\Delta t} \int_0^{\Delta t} \tau e^{2\tau s} ds \right) \alpha_1(x,0) + e^{-2\tau\Delta t}\alpha_3(x,0) = \frac{1}{2}\left(1 - e^{-2\tau\Delta t}\right)\alpha_1(x,0) + e^{-2\tau\Delta t}\alpha_3(x,0).$$

Plugging this approximation in (75) yields the system of two conservation laws

$$\begin{cases} \partial_t \alpha_1 + \partial_x(-\sigma x \alpha_1 + \alpha_2) = 0, \\ \partial_t \alpha_2 + \partial_x \left((1 - \sigma^2 x^2)\alpha_1 + \sigma x \alpha_2 - \frac{1}{2}\left(1 - e^{-2\tau\Delta t}\right)\alpha_1 - e^{-2\tau\Delta t}\alpha_3(0) \right) = 0. \end{cases} \tag{76}$$

This approximation is valid during the time step Δt. Within this time step it is possible to use a standard Riemann solver for systems of conservation laws to get a well-balanced scheme with explicit contribution of the time step, as in [16][page 185].

6. MultiD

This section is devoted to show that the definition of a **multidimensional** well-balanced scheme displays a much richer structure than in the one dimensional case. Indeed the set of adjoint stationary states may have an infinite dimension, while the dimension is only finite in one dimension. A solution will be detailed on the example of the 2D hyperbolic heat equation with the change-of-basis matrix $P(\mathbf{x})$ which seems to be the correct object to manipulate in higher dimensions.

6.1. 2D hyperbolic heat equation

We consider the two dimensional hyperbolic heat equation. The primal formulation writes

$$\left\{\begin{array}{lll} \partial_t p & +\partial_x u & +\partial_y v & = 0, \\ \partial_t u & +\partial_x p & & = -\sigma u, \\ \partial_t v & & +\partial_y p & = -\sigma v, \end{array}\right. \tag{77}$$

where we take that $\sigma > 0$ is constant. The adjoint stationary states $(\widehat{p}, \widehat{u}, \widehat{v})$ that correspond to (4-5) are solutions of

$$\left\{\begin{array}{lll} \partial_x \widehat{u} & +\partial_y \widehat{v} & = 0, \\ \partial_x \widehat{p} & & = \sigma \widehat{u}, \\ & \partial_y \widehat{p} & = \sigma \widehat{v}. \end{array}\right.$$

So $(\widehat{u}, \widehat{v}) = \frac{1}{\sigma}\nabla\widehat{p}$ and $\Delta\widehat{p} = 0$. Therefore

$$\mathcal{V} = \left\{(\widehat{p}, \widehat{u}, \widehat{v});\ \widehat{p} \text{ is harmonic and } (\widehat{u}, \widehat{v}) = \frac{1}{\sigma}\nabla\widehat{p}\right\}.$$

In one dimension, the set of harmonic functions reduces to affine functions and so coincides with (11), as observed in [3] in the context of stationary transport equations. In two dimension, we define

$$\mathcal{V}_n = \mathcal{V} \cap \{\widehat{p} \text{ is an harmonic polynomial of degree } \leq n\}.$$

Clearly $p \in \mathcal{V}_1$ is equivalent to $\widehat{p} = a + bx + cy$. It yields three test functions

$$V_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad V_2 = \begin{pmatrix} \sigma x \\ 1 \\ 0 \end{pmatrix}, \quad V_3 = \begin{pmatrix} \sigma y \\ 0 \\ 1 \end{pmatrix}$$

and three conservation laws for $\alpha_1 = p$, $\alpha_2 = \sigma x p + u$ and $\alpha_3 = \sigma y p + v$. The system writes

$$\partial_t \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \partial_x \begin{pmatrix} m_1 \\ \alpha_1 + \sigma x m_1 \\ \sigma y m_1 \end{pmatrix} + \partial_y \begin{pmatrix} m_2 \\ \sigma x m_2 \\ 1 + \sigma y m_2 \end{pmatrix} = 0, \tag{78}$$

where $m_1 = -\sigma x \alpha_1 + \alpha_2$ and $m_2 = -\sigma y \alpha_1 + \alpha_3$. This system can be used directly to discretize the two dimensional hyperbolic heat equation with a conservative method. Let us note the 2D variable as $\mathbf{x} = (x, y)$. The original variable can be recovered at any time with the inverse formula $U = P(\mathbf{x})^{-1}\alpha$ where the matrix is triangular

$$P(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 \\ \sigma x & 1 & 0 \\ \sigma y & 0 & 1 \end{pmatrix} \tag{79}$$

and non singular matrix. We notice the inverse formula

$$P(\mathbf{x})^{-1} = P(-\mathbf{x}) \tag{80}$$

and the commutation and composition property

$$P(\mathbf{x})P(\mathbf{y}) = P(\mathbf{y})P(\mathbf{x}) = P(\mathbf{x} + \mathbf{y}). \tag{81}$$

The identity (78) is also

$$\partial_t \alpha + \partial_x \left(P(\mathbf{x}) A P(\mathbf{x})^{-1}\alpha\right) + \partial_y \left(P(\mathbf{x}) B P(\mathbf{x})^{-1}\alpha\right) = 0. \tag{82}$$

Starting from $p \in \mathcal{V}_2$ is equivalent to the decomposition

$$\widehat{p} = a + bx + cy + d(x^2 - y^2) + exy, \quad a, b, c, d, e \in \mathbb{R}.$$

This set yields 5 conservation laws. It means that two additional algebraic conservational laws are satisfied by the solutions of the system (78). We do not know for the moment how to use this information for the development of numerical methods, even if it seems clear that discontinuous Galerkin methods [19, 25] may take advantage of it.

6.2. Discretization of the 2D hyperbolic heat equation

Using the structure (82), we show how to recover a well-balanced (and asymptotic diffusion preserving) scheme with corner fluxes recently designed in [5]. We start from a nodal based finite volume scheme for the homogeneous equation. The corresponding semi-discrete scheme writes with original notations on a 2D grid

$$\left\{ \begin{array}{l} s_j p_j'(t) + \sum_r l_{jr} \mathbf{n}_{jr} \cdot \mathbf{u}_r = 0, \\ s_j \mathbf{u}_j'(t) + \sum_r l_{jr} \mathbf{n}_{jr} p_{jr} = 0. \end{array} \right. \tag{83}$$

Denoting $\mathbf{n}_{jr} = \Big(\cos\theta_{jr}, \sin\theta_{jr}\Big)$, it admits the reformulation more adapted to our purposes

$$s_j U_j'(t) + \sum_r l_{jr} \widetilde{A}_{jr} U_{jr} = 0, \tag{84}$$

where $U_j = (p_j, \mathbf{u}_j)$, $U_{jr} = (p_{jr}, \mathbf{u}_r)$ and $A_{jr} = \cos\theta_{jr} A + \sin\theta_{jr} B = A_{jr}^t$ is symmetrix matrix. We now desire to modify this scheme in order to discretized the 2D conservative formulation (78), by means of an extension of the so-called two states solver.

We first write the analogue of the multiplicative scheme (21-22) as

$$s_j U_j'(t) + P\Big(\mathbf{x}_j\Big)^{-1} \sum_r l_{jr} \beta_{jr} = 0. \tag{85}$$

The associated additive formulation (25) writes

$$s_j U_j'(t) + \sum_r l_{jr} \widetilde{A}_{jr} \widehat{U}_{jr} + \sum_r \Big(P(\mathbf{x}_j)^{-1} P(\mathbf{x}_r) - I\Big) \widetilde{A}_{jr} U_{jr} = 0. \tag{86}$$

The product of matrices can be rewritten using the composition formula (81). It writes under a more local formula $P(\mathbf{x}_j)^{-1} P(\mathbf{x}_r) = P(\mathbf{x}_r - \mathbf{x}_j)$. These two formulations are the same since we assume the generalization of (24) under the form

$$P(\mathbf{x}_r) A_{jr} U_{jr} = \beta_{jr}. \tag{87}$$

It remains to define the fluxes to close the system.

To this end we decide first to preserve the structure of the fluxes defined in [5]. That is we keep $U_{jr} = (p_{jr}, \mathbf{u}_r)$ which means that the first component $p_{jr} \in \mathbb{R}$ is delocalized around the node $\mathbf{x}_r$ while the two other components $\mathbf{u}_r \in \mathbb{R}^2$ are "attached" to the node. For convenience we generalize the two states solver starting from (47).

Proposition 15. *The corner based 2D scheme based on the two-states solver is equal to the one published in [5]. The corner based linear solver writes*

$$\left\{ \begin{array}{l} p_{j_r} + \sigma\Big(\mathbf{x}_r - \mathbf{x}_j, \mathbf{u}_r\Big) - p_j + \Big(\mathbf{n}_{jr}, \mathbf{u}_r - \mathbf{u}_j\Big) = 0, \\ \sum_j l_{jr} \mathbf{n}_{jr} p_{jr} = 0. \end{array} \right. \tag{88}$$

Remark 8. *What we call the linear vertex-based solver (around $\mathbf{x}_r$) is made of the linear equations (89)-(90). The unknowns are $\mathbf{u}_r$ and the p_{jr}s for all cells j around the node $\mathbf{x}_r$. Unfortunately one cannot rely on theorem 1 (in 1D) to show this linear system is well posed. However it is been shown in [5] that this system is in practice non singular, for a wide range of meshes. Moreover the final scheme has been proved to be diffusion asymptotic preserving.*

Proof. Indeed the matrix $P(x_j)^{-t} P(x_{j+\frac{1}{2}})^t$ in one dimension becomes

$$P(\mathbf{x}_j)^{-t} P(\mathbf{x}_r)^t = P(\mathbf{x}_r - \mathbf{x}_j)^t = \left(\begin{array}{ccc} 1 & \sigma(x_r - x_j) & \sigma(y_r - y_j) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right).$$

The eigenvector u_p for $\lambda_p > 0$ is now in this context the eigenvector u_{jr} of the matrix $\widetilde{A}_{jr}$ associated a positive eigenvalue. There is no ambiguity since there is only one positive eigenvalue (equal to 1). One obtains the eigenvector $u_{jr} = (1, \mathbf{n}_{jr})^t$ since

$$\widetilde{A}_{jr} u_{jr} = \begin{pmatrix} 0 & \cos\theta_{jr} & \sin\theta_{jr} \\ \cos\theta_{jr} & 0 & 0 \\ \sin\theta_{jr} & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ \cos\theta_{jr} \\ \sin\theta_{jr} \end{pmatrix} = \begin{pmatrix} 1 \\ \cos\theta_{jr} \\ \sin\theta_{jr} \end{pmatrix} = u_{jr}.$$

With these notations the linear equations (47) become

$$\left(u_{jr}, P(\mathbf{x}_r - \mathbf{x}_j)^t U_{jr} - U_j\right) = 0 \text{ for all cells around the node } \mathbf{x}_r.$$

We obtain more specifically for all cells (that is all j) around the node $\mathbf{x}_r$

$$p_{j_r} + \sigma\left(\mathbf{x}_r - \mathbf{x}_j, \mathbf{u}_r\right) - p_j + \left(\mathbf{n}_{jr}, \mathbf{u}_r - \mathbf{u}_j\right) = 0. \tag{89}$$

To close we add a formula that enforces the conservativity of the divergent part. It writes $\sum_j l_{jr} \widetilde{A}_{jr} \widehat{U}_{jr} = 0$ for all cells around the node $\mathbf{x}_r$. A consequence is

$$\sum_r \left(\sum_j l_{jr} \widetilde{A}_{jr} \widehat{U}_{jr} \right) = \sum_j \left(\sum_r l_{jr} \widetilde{A}_{jr} \widehat{U}_{jr} \right) = 0$$

which indeed guarantees the conservativity of the divergent part of the scheme. The formula $\sum_j l_{jr} \widetilde{A}_{jr} \widehat{U}_{jr} = 0$ can be decomposed in two different equations. The first one is $\sum_j l_{jr} \left(\mathbf{n}_{jr}, \mathbf{u}_r\right) = \left(\sum_j l_{jr} \mathbf{n}_{jr}, \mathbf{u}_r\right) = 0$ and is trivially true since $\sum_j l_{jr} \mathbf{n}_{jr} = 0$ (see [5]). The second equation is

$$\sum_j l_{jr} \mathbf{n}_{jr} p_{jr} = 0 \in \mathbb{R}^2. \tag{90}$$

The corner based linear system (88) is made of (89) and (90). So the proof is ended. □

7. Conclusion and perspectives

The structure of well-balanced schemes has been detailed for Friedrichs systems with linear relaxation, starting from an original fully conservative formulation of the equations which stressed the idea of a change of basis with a duality method where the change-of-basis matrix plays the main role.

A more general family of Friedrichs type is

$$A_0(\mathbf{x})\partial_t U + \partial_x(A(\mathbf{x})U) + \partial_y(B(\mathbf{x})U) = S(\mathbf{x}) - R(\mathbf{x})U$$

for which the same ideas of duality and change of basis should apply identically. It can be used to treat the example evoked in remark 1. But the method can be developed in many other directions which are briefly detailed below.

First directions already evoked in this work are: asymptotic preserving methods which usually start from well-balanced techniques plus small parameters; and 2D formulations which display an interesting structure where the matrix P plays a major role. An open problem is to determine wether the commutation-composition relations (81) are necessary or if there are true only for our example. The effective calculation of the matrix exponentials for the S_n model can be performed starting from the references [10, 9, 16, 27]. This technical difficulty does not show up for the hyperbolic heat equation since the matrix exponential of a nilpotent matrix has only a finite number of terms.

Other directions concern the numerical development of high order discretization Finite Volume Method or of completely different discretization technics such as Finite Element Method for example. The numerical analysis of the singular case is needed to understand the influence of low wave velocities for the S_n model and similar models.

Since any hyperbolic system can be locally linearized, it is reasonable to extend the techniques developed in this work to more general non linear problems. In the same vein, problems with variable coefficients can be addressed systematically with this formulation.

A fully open problem is the use of the conservative structure for the convergence analysis although stability is evoked briefly in remark 7, § 4.

Acknowledgments: The authors warmly thank Laurent Gosse for valuable comments and discussions and for pointing out some references in the literature about well-balanced schemes for various equations.

[1] D. Amadori and L. Gosse. Error estimates for well-balanced and time-split schemes on a damped semilinear wave equation. Hal preprint hal-00959775v2, 2014.
[2] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein anD B. Perthame, A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows, SIAM . Sci. Comput., (2004), 2050-2065.
[3] G. Birkhoff and I. Abu-Shumays, Harmonic solutions of transport equations, Journal of Mathematical Analysis and Applications Volume 28, Issue 1, October 1969, Pages 211-221.
[4] F. Bouchut. Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources, Frontiers in Mathematics series, Birkhuser, 2004.
[5] C. Buet, B. Després and E. Franck Design of asymptotic preserving schemes for the hyperbolic heat equation on unstructured meshes Numerish Mathematik, Volume 122, Issue 2, pp 227-278, 2012.
[6] C. Buet, B. Després and E. Franck, Asymptotic Preserving Schemes on Distorted Meshes for Friedrichs Systems with Stiff Relaxation: Application to Angular Models in Linear Transport, Journal of Scientific Computing, online, 2014.
[7] P. Cargo and A.-Y. LeRoux, Un schéma équilibre adapté au modèle d'atmosphère avec termes de gravité. (in French) [A well-balanced scheme for a model of an atmosphere with gravity] C. R. Acad. Sci. Paris Sér. I Math. 318 (1994), no. 1, 73-76.
[8] A. Arnold, J.-A. Carrillo, M.D. Tidriri, Large-time behavior of discrete equations with nonsymmetric interactions. Math. Mod. Meth. in Appl. Sci. 12 (2002) 1555-1564.
[9] K. M. Case and P. F. Zweifel, Linear transport theory, Addison-Wesley Pub. Co., 1967.
[10] S. Chandraseckhar, Radiative transfer, Dover publication, 1960.
[11] F. Coquel and E. Godlewski. Asymptotic preserving scheme for Euler system with large friction. J. Sci. Comput. 48 (2011), no. 1-3, 164-172.
[12] R. Dautray, J.L. Lions, J.-L. and I. N. Sneddon. 1999. Mathematical Analysis and Numerical Methods for Science and Technology - Volumes 1 to 6, Berlin, Springer.
[13] B. Després, Lois de conservation eulériennes, lagrangiennes et méthodes numériques (in French), Springer, 2009.
[14] L. Gosse, Well-balanced schemes using elementary solutions for linear models of the Boltzmann equation in one space dimension, Kinetic Relat. Mod., 5 283-323, 2012.
[15] L. Gosse, Transient radiative transfer in the grey case: Well-balanced and asymptotic-preserving schemes built on Case's elementary solutions Journal of Quantitative Spectroscopy and Radiative Transfer 112, pp 1995-2012, 2011.
[16] L. Gosse, Computing Qualitatively Correct Approximations of Balance Laws, Exponential-Fit, Well-Balanced and Asymptotic-Preserving, SEMA SIMAI Springer Series, Vol. 2, 2013.
[17] L. Gosse and G. Toscani, An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations. C. R. Acad. Sci. Paris Ser. I 334, 337-342 (2002)
[18] J.-M. Greenberg and A.-Y. Leroux, A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. SIAM J. Numer. Anal. 33 (1996), no. 1, 1-16.
[19] F. Hindenlang, G.-J. Gassner, C. Altmann, A. Beck, Andrea, M. Staudenmaier, Marc and C.-D. Munz, Explicit discontinuous Galerkin methods for unsteady problems. Comput. & Fluids 61 (2012), 86-93.
[20] L. Huang and T. P. Liu, A conservative, piecewise-steady difference scheme for transonic nozzle flow, Computers & Mathematics with Applications Volume 12, Issues 4-5, Part A, 1986, Pages 377-388.
[21] S. Jin, Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review, Lecture Notes for Summer School on Methods and Models of Kinetic Theory (M and MKT), Porto Ercole (Grosseto, Italy), 2010.
[22] S. Jin and D. Levermore, Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. JCP 126, 449-467 (1996).
[23] T. Muller and A. Pfeiffer, Well-balanced simulation of geophysical flows via the shallow water equations with bottom topography: consistency and numerical computations, n Hyperbolic problems: Theory, Numerics, Applications, AIMS, 2014, 801-808.
[24] S. Ortleb and A. Meister, A well-balanced DG scheme with unconditionally positive implicit time integration, in Hyperbolic problems: Theory, Numerics, Applications, AIMS, 2014, 823-830.
[25] M. Tavelli and M. Dumbser, A high order semi-implicit discontinuous Galerkin method for the two dimensional shallow water equations on staggered unstructured meshes. Appl. Math. Comput. 234 (2014), 623-644.
[26] B. VanLeer, On the relation between the upwind differencing schemes of Engquist-Osher, Godunov and Roe. SIAM J Sci. Comp. 5 (1984) 1-20.
[27] P.C. Waterman, Matrix exponential description of radiative transfer, J. Opt. Soc. Am., Vol 71, 4, 1981, 410-422.
[28] A. Zygmund, Trigonometric series, Cambridge University Press, New York, 1959.

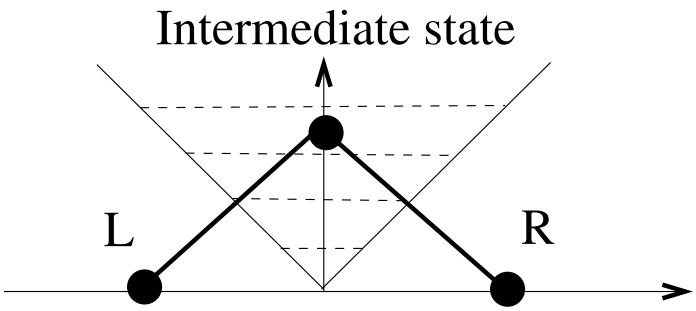

Figure 1: Schematic of a Riemann solver. The intermediate state is computed in function of a left state L and a right state R

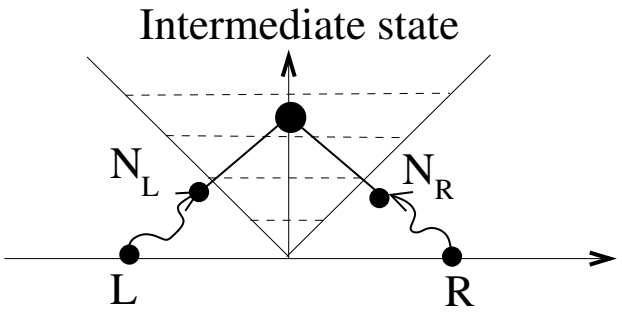

Figure 2: Schematic of the modified Riemann solver (31). The New-left and New-right states are modification of the initial Left and Right states.

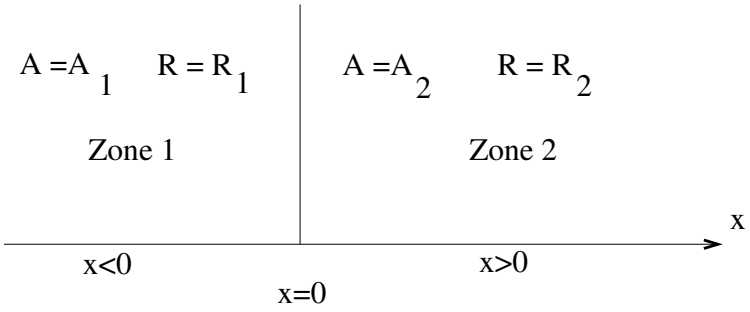

Figure 3: Discontinuity of the coefficients. The analysis is performed for $A = A_1 = A_2$. On physical grounds, such problems are very similar to scattering problems that can be analyzed with transmissions and reflections operators. This correspondance will be evidenced in section 4.

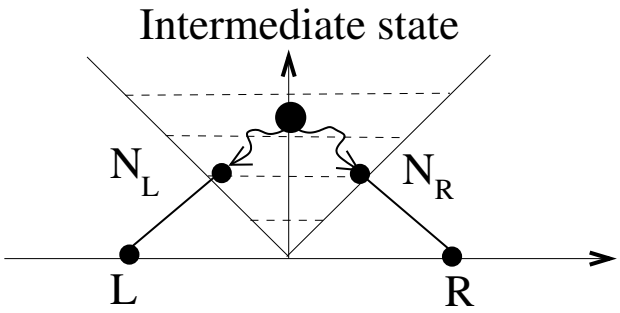

Figure 4: Schematic of the modified Riemann solver (46)-(48). The New-left and New-right internal states are modification of the intermediate state by the backward propagator. This illustration is to be compared with figure 2.

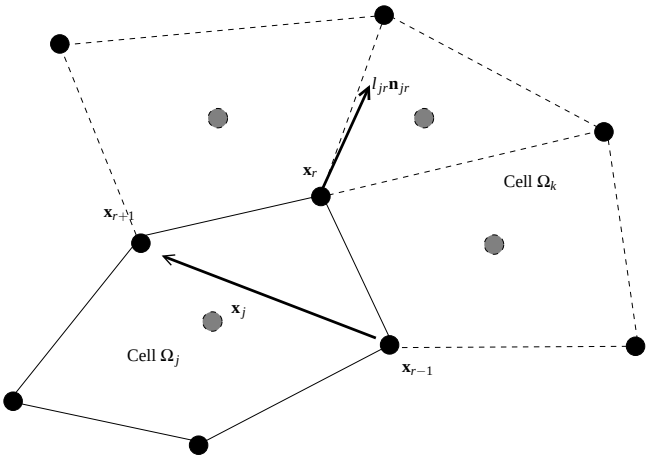

Figure 5: Notation for nodal formulation: the corner length l_{jr} and the corner normal $\mathbf{n}_{jr}$. Notice that $l_{jr}\mathbf{n}_{jr}$ is equal to the orthogonal vector to the half of the vector that starts at $\mathbf{x}_{r-1}$ and finish at $\mathbf{x}_{r+1}$. The center of the cell is an arbitrary point inside the cell.