
UNIVERSIDAD DE MÁLAGA

Programa de Doctorado en Matemáticas

Facultad de Ciencias

Departamento de Análisis Matemático, Estadística e Investigación Operativa y Matemática Aplicada

# Numerical analysis of some nonlinear hyperbolic systems of Partial Differential Equations arising from Fluid Mechanics.

Ernesto Pimentel García

PhD Thesis

Advisors:

Carlos María Parés Madroñal, Manuel Jesús Castro Díaz

Universidad de Málaga Junio 2021



AUTOR: Ernesto Pimentel García

https://orcid.org/0000-0002-0539-0023

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

D. Carlos María Parés Madroñal, Catedrático del Departamento de Análisis Matemático, Estadística e Investigación Operativa, y Matemática Aplicada de la Universidad de Málaga, y D. Manuel Jesús Castro Díaz, Catedrático del Departamento de Análisis Matemático, Estadística e Investigación Operativa, y Matemática Aplicada de la Universidad de Málaga.

**Certifican:**

Que D. Ernesto Pimentel García, con grado y máster en Matemáticas, ha realizado en dicho Departamento, bajo nuestra dirección, el trabajo de investigación correspondiente a su Tesis Doctoral, titulada:

## Numerical analysis of some nonlinear hyperbolic systems of Partial Differential Equations arising from Fluid Mechanics

Revisado el presente trabajo, estimamos que puede ser presentado al Tribunal que ha de juzgarlo. Y para que constate a efectos de lo establecido en el artículo octavo del Real Decreto 99/2011, autorizamos la presentación de este trabajo en la Universidad de Málaga.

Málaga, a 23 de Junio de 2021

PARES MADROÑAL CARLOS MARIA
Firmado digitalmente por PARES MADROÑAL CARLOS MARIA
Fecha: 2021.06.23 13:10:51 +02'00'

Fdo: Dr. Carlos María Parés Madroñal

CASTRO DIAZ MANUEL JESUS -
Firmado digitalmente por CASTRO DIAZ MANUEL JESUS -
Fecha: 2021.06.23 13:36:55 +02'00'

Fdo: Manuel Jesús Castro Díaz

# DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña ERNESTO PIMENTEL GARCÍA
Estudiante del programa de doctorado MATEMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: NUMERICAL ANALYSIS OF SOME NONLINEAR HYPERBOLIC SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS ARISING FROM FLUID MECHANICS

Realizada bajo la tutorización de CARLOS MARÍA PARÉS MADROÑAL y dirección de CARLOS MARÍA PARÉS MADROÑAL Y MANUEL JESÚS CASTRO DÍAZ (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 23 de JUNIO de 2021

| | |
|---|---|
| PIMENTEL GARCIA ERNESTO<br>Firmado digitalmente por PIMENTEL GARCIA ERNESTO<br>Fecha: 2021.06.23 12:45:55 +02'00'<br>Fdo.: ERNESTO PIMENTEL GARCÍA<br>Doctorando/a | PARES MADROÑAL CARLOS MARIA<br>Firmado digitalmente por PARES MADROÑAL CARLOS MARIA -<br>Fecha: 2021.06.23 13:11:16 +02'00'<br>Fdo.: CARLOS MARÍA PARÉS MADROÑAL<br>Tutor/a |
| PARES MADROÑAL CARLOS MARIA -<br>Firmado digitalmente por PARES MADROÑAL CARLOS MARIA -<br>Fecha: 2021.06.23 13:11:29 +02'00'<br>Fdo.: CARLOS MARÍA PARÉS MADROÑAL | CASTRO DIAZ MANUEL JESUS<br>Firmado digitalmente por CASTRO DIAZ MANUEL JESUS -<br>Fecha: 2021.06.23 13:37:26 +02'00'<br>MANUEL JESÚS CASTRO DÍAZ |





Edificio Pabellón de Gobierno. Campus El Ejido. 29071
**Tel.:** 952 13 10 28 / 952 13 14 61 / 952 13 71 10
**E-mail:** doctorado@uma.es

| Director/es de tesis |
| --- |





Edificio Pabellón de Gobierno. Campus El Ejido. 29071
**Tel.:** 952 13 10 28 / 952 13 14 61 / 952 13 71 10
**E-mail:** doctorado@uma.es

# Agradecimientos

Una tesis no es fruto de una sola persona. Es por eso que con estas líneas me gustaría dedicar unas palabras a toda esa gente que, de una forma u otra, han contribuido a la realización de esta memoria.

En primer lugar, me gustaría agradecer a Carlos Parés toda la dedicación, ayuda, apoyo, compromiso y paciencia que ha mostrado, no solo durante el doctorado, sino también durante toda mi carrera universitaria. A él le debo buena parte de mi interés por esta maravillosa área de las matemáticas como es la Matemática Aplicada. Quiero resaltar, no solo su aporte académico, sino también personal, siempre creando un buen ambiente de trabajo y un magnífico grupo de investigación que cada día crece más.

De la misma forma, deseo agradecer a Manuel Castro todo el tiempo y apoyo dedicados durante estos años. Siempre sacando tiempo de donde no lo hay para echar una mano de forma desinteresada, no solo a mi, sino a todos los miembro de este grupo.

Es también mi deseo agradecer a Tomás Morales por sus aportes y contribuciones al desarrollo de esta memoria. Al magnífico grupo EDANYA: Mª Luz, Jorge Macías, José María Gallardo, Carlos Sánchez, Sergio Ortega, José Manuel González, Marc de la Asunción y María López. Muchas gracias a los excelentes compañeros, muchos ya doctores, que he tenido en esta etapa: Hugo Carrillo, Irene Gómez, Kleiton Schneider, Ernesto Guerrero, Juan Carlos González y Cipriano Escalante. Gracias por hacer estos años tan amenos.

Agradecer también a Christophe Chalons y Philippe G. LeFloch por todo el apoyo y aportes en esta tesis, y por hacer tan cómodas, fáciles y fructíferas las estancias de investigación en París.

Me gustaría también hacer especial mención a Julian Köllermeier por toda la colaboración recibida por su parte y por contar siempre conmigo para la realización de nuevos proyectos.

Por último, pero no menos importante, me gustaría agradecer a mi madre y a mi

padre todo el apoyo que me han brindado y toda la educación recibida, sin los cuales hubiese sido imposible llegar a este momento. También a mis hermanos y hermana por aguantarme todos estos años y hacerlos tan entretenidos. Y por supuesto, al resto de mi familia y a todos mis amigos y amigas: Los Pizarro, Hnos.Teddy, Golden Chofff, el Colesp, compañeros de la carrera y muchos más, que han conseguido que toda esta etapa haya sido tan especial y con tanto "flow".

Gracias a todos y todas.

# Contents

# List of Figures





# Resumen

En esta tesis se abordan cuatro problemas diferentes relacionados con el análisis numérico de sistemas de ecuaciones en derivadas parciales hiperbólicos no lineales. Estos problemas están relacionados con algunas de las líneas de investigación del grupo EDANYA y, más concretamente, con la resolución numérica de modelos matemáticos de la mecánica de fluidos en aplicaciones relacionadas con los flujos de aguas someras y la dinámica de gases en el contexto de la mecánica clásica o relativista. A continuación se enumeran los problemas abordados en orden cronológico.

El primer problema que se aborda es el estudio del problema de Riemann para las ecuaciones de aguas someras sobre un fondo con forma de escalón, en el caso particular en el que solo hay agua a un lado del escalón. De esta forma completamos el estudio realizado por LeFloch and Thanh en [112]. El análisis de estas situaciones de frentes secos-mojados es importantes a la hora de diseñar esquemas numéricos que traten bien los fenómenos de inundación. En el estudio de este problema surgen dos importantes dificultades. Por un lado, el término fuente que aparece en las ecuaciones es un producto no conservativo, con lo que no existe una única forma de definir las soluciones débiles del problema. En nuestro caso seguiremos la teoría de Dal Maso, LeFloch y Murat [57] para definirlas a partir de una familia de caminos. Por otra parte, aparecen casos resonantes (es decir, un autovalor de la matriz Jacobiana se anula) que implican que, una vez elegida la definición de solución débil, no haya unicidad de solución. Este estudio teórico se complementa con ensayos numéricos en los que se estudia el comportamiento de diferentes esquemas.

La siguiente cuestión que se aborda es la implementación eficiente de métodos numéricos basados en resolvedores de Riemann aproximados y, en particular, del método de Roe que se basa en la resolución de problemas de Riemann linealizados en las interceldas. El interés práctico de este capítulo reside sobre todo en la resolución numérica de sistemas de gran tamaño, como es el caso de los modelos de aguas someras multicapas [44] o basados en momentos [98]. Esta nueva propuesta de implementación se basa en la relación estrecha que hay entre este tipo de métodos y los Polynomial Viscosity Matrix (PVM) basados en la elección de un polinomio que interpola la función valor absoluto, así como en la forma de Newton de dicho polinomio.

El siguiente objetivo es hacer un estudio sistemático del comportamiento asintótico de las soluciones de las ecuaciones de Burgers y Euler relativistas basados en la métrica de Schwarzchild, usando métodos numéricos . Dicha métrica es una solución exacta de las ecuaciones de Einstein del campo gravitatorio que describe el campo generado por una estrella o una masa esférica. En estos sistemas las soluciones estacionarias y las evoluciones de sus perturbaciones juegan un papel fundamental en la comprensión de dicho comportamiento. Por tanto, es imprescindible el uso de métodos well-balanced, es decir, métodos capaces de preservar este tipo de soluciones. Aplicaremos el marco general descrito en [47] para desarrollar métodos well-balanced de orden hasta 3 para el modelo de Burgers-Schwarzschild y 2 para el modelo de Euler-Schwarzschild. Compararemos los resultados obtenidos entre estos métodos y los estándar y pondremos de manifiesto la importancia de la propiedad well-balanced.

Es sabido que, en el caso de sistemas con productos no conservativos, la consistencia, la estabilidad y el control de la entropía no son suficientes para asegurar la convergencia de las aproximaciones numéricas a soluciones débiles admisibles: es necesario, además, controlar los fenómenos de pequeña escala tales como la viscosidad numérica que afectan a la posición y amplitud de las ondas de choque (ver [109]). En [51] Chalons presentó una técnica basada en reconstrucciones discontinuas en las celdas que le permitió diseñar métodos de primer orden que capturaban bien las soluciones con choques aislados. El último problema que se plantea en la tesis es la extensión a segundo orden de esta técnica usando el formalismo de los métodos path-conservative introducidos en [132], lo que sienta las bases para su extensión a órdenes mayores. Enunciaremos y probaremos que la extensión a segundo orden mantiene la propiedad de capturar bien los choques aislados y lo comprobaremos a través de distintos tests numéricos.

Finalmente se presentan las principales contribuciones de esta tesis y las posibles líneas futuras de trabajo.

# Introducción

La Mecánica de Fluidos Computacional constituye hoy en día una de las herramientas matemáticas más importantes en la simulación de diversos fenómenos que ocurren a nuestro alrededor. Su objetivo es simular la evolución de los fluidos mediante la resolución numérica de sistemas de ecuaciones en derivadas parciales (EDPs) que gobiernan su comportamiento. El problema de estos sistemas es que en la mayor parte de los casos no es posible resolverlos de forma exacta, de ahí la necesidad de usar métodos numéricos. A través de estos métodos numéricos obtendremos simulaciones que nos permitirán comprender, predecir y controlar la evolución de los flujos de fluidos. Este tipo de herramientas tiene aplicaciones en distintos campos de estudio tales como la oceanografía, meteorología, climatología, ingeniería hidráulica, aeronáutica, biología, etc. El desarrollo de métodos numéricos adecuados requiere un conocimiento profundo de la naturaleza física de los flujos a simular y de las propiedades matemáticas de los sistemas a resolver.

En mecánica de fluidos una de las ecuaciones en derivadas parciales más generales que rigen el movimiento de los fluidos son las conocidas ecuaciones de Navier-Stokes (ver [151]). Estas ecuaciones expresan la conservación de la masa, la cantidad de movimiento y la energía, junto con una ecuación de estado que relaciona la presión, energía y densidad. A través de ciertas hipótesis sobre los fluidos a considerar que simplifican estas ecuaciones generales es posible llegar a otros modelos como, por ejemplo, las ecuaciones de *aguas poco profundas* o *shallow water*. En su versión unidimensional, estas ecuaciones fueron deducidas por Jean Claude Barré de Saint-Venant en 1843 de ahí que a veces se las llamen ecuaciones de Saint-Venant. Estas ecuaciones describen el movimiento de una capa de fluido con poco espesor. En el caso del modelo unidimensional, estas ecuaciones se obtienen de las de Navier-Stokes a partir de una serie de hipótesis:

- Se supone que el agua es homogénea e incompresible.
- La presión es hidrostática, es decir, la presión aumenta con la profundidad, siendo igual a la del aire en la superficie del fluido.
- La única fuerza interna que actúa en el fluido es la presión (se desprecian los efectos viscosos).

- Tanto el fondo sobre el que evoluciona el agua como su superficie libre pueden ser representados mediante la gráfica de una función. Es más, dichas funciones solo dependen de una de las variables horizontales, $x$, y del tiempo $t$ (en el caso de la superficie libre).
- Se supone que la velocidad del fluido solo depende de $x$ y de $t$. Además, las desviaciones de su componente horizontal con respecto a su promedio vertical se suponen despreciables.

A partir de estas hipótesis y mediante un proceso de integración vertical se obtienen las ecuaciones de la conservación de la masa y la cantidad de movimiento que son:

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + \dfrac{gh^2}{2}\right) = -gh\partial_x a, \\ \partial_t a = 0, \end{cases} \tag{I. 1}$$

donde:

- $h = h(x,t) \geq 0$ es el grosor de la capa agua;
- $u = u(x,t)$ es la velocidad horizontal promediada en la dirección vertical;
- $g$ es la intensidad del campo gravitatorio;
- $a = a(x)$ es la profundidad del fondo dado desde un nivel de referencia.

Este sistema lo podemos escribir de la siguiente forma:

$$U_t + F(U)_x = S(U)a_x,$$

donde:

$$U = \begin{pmatrix} h \\ hu \\ a \end{pmatrix}, \quad F(U) = \begin{pmatrix} hu \\ hu^2 + \dfrac{gh^2}{2} \\ 0 \end{pmatrix}, \quad S(U) = \begin{pmatrix} 0 \\ -gh \\ 0 \end{pmatrix}.$$

En esta tesis consideraremos los siguientes tres tipos de EDPs:

- *Sistemas de leyes de conservación*:

$$U_t + F(U)_x = 0, \tag{I. 2}$$

- *Sistemas de leyes de equilibrio*:

$$U_t + F(U)_x = S(U)\sigma_x, \tag{I. 3}$$

- *Sistemas con productos no conservativos*:

$$U_t + F(U)_x + B(U)U_x = S(U)\sigma_x, \tag{I. 4}$$

donde las incógnitas $U(x,t) = (u_1(x,t), ..., u_N(x,t))^T$ toma valores en un conjunto abierto y convexo $\Omega$ de $\mathbb{R}^N$, $F$ función regular de $\Omega$ a $\mathbb{R}^N$, $B$ es una función matricial regular de $\Omega$ a $\mathcal{M}_{NxN}(\mathbb{R})$, $S$ es una función de $\Omega$ a $\mathbb{R}^N$ y $\sigma(x)$ una función conocida de $\mathbb{R}$ a $\mathbb{R}$. Como hemos visto, las ecuaciones de aguas someras con topografía son un sistema de leyes de equilibrio que se considerará en el Capítulo 2. Algunos ejemplos de sistemas con productos no conservativos que veremos a lo largo de la tesis son:

- El modelo bicapa de aguas someras sin topografía (ver [44], [74]), que puede ser escrito en la forma (I. 4) con:

$$U = \begin{pmatrix} h_1 \\ q_1 \\ h_2 \\ q_2 \end{pmatrix}, \quad F(U) = \begin{pmatrix} q_1 \\ \frac{q_1^2}{h_1} + \frac{1}{2}gh_1^2 \\ q_2 \\ \frac{q_2^2}{h_2} + \frac{1}{2}gh_2^2 \end{pmatrix},$$

$$B(U) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & gh_1 & 0 \\ 0 & 0 & 0 & 0 \\ grh_2 & 0 & 0 & 0 \end{pmatrix}, \quad S(U) \equiv 0,$$

  siendo $r = \rho_1/\rho_2$, donde:

  - $h_i = h_i(x,t) \geq 0$ es el grosor de la $i$-ésima capa.
  - $q_i = q_i(x,t) = h_i(x,t)u_i(x,t)$ es el caudal de la $i$-ésima capa, donde $u_i(x,t)$ es la velocidad horizontal promediada en la dirección vertical.
  - $g$ es la intensidad del campo gravitatorio.
  - $\rho_i$ es la densidad constante de la $i$-ésima capa.

  Este sistema será considerado en el Capítulo 3.

- El sistema relativista de Burgers en el contexto de Schwarzschild (ver [114]), que puede ser escrito en la forma (I. 4) con:

$$U = \begin{pmatrix} v \\ r \end{pmatrix}, \quad F(U) = \begin{pmatrix} \left(1 - \frac{2M}{r}\right)\frac{v^2-1}{2} \\ 0 \end{pmatrix},$$

$$B(U) \equiv 0, \quad S(U) = \begin{pmatrix} \frac{2M}{r^2}(v^2-1) \\ 0 \end{pmatrix}, \quad \sigma(r) = r,$$

donde $r > 2M, v = (t, r) \in [-1, 1]$, siendo $r$ la distancia al agujero negro, $v$ la velocidad normalizada del flujo y $M > 0$ el coeficiente que representa la masa del agujero negro. Estudiaremos este sistema en el Capítulo 4.

- El sistema relativista de Euler en el contexto de Schwarzschild, que puede ser escrito en la forma (I. 4) con:

$$U = \begin{pmatrix} V^0 \\ V^1 \\ r \end{pmatrix} = \begin{pmatrix} \dfrac{1+k^2v^2}{1-v^2}\rho \\ \dfrac{1+k^2}{1-v^2}\rho v \\ r \end{pmatrix}, \qquad F(U) = \begin{pmatrix} \left(1-\dfrac{2M}{r}\right)\dfrac{1+k^2}{1-v^2}\rho v \\ \left(1-\dfrac{2M}{r}\right)\dfrac{v^2+k^2}{1-v^2}\rho \\ 0 \end{pmatrix},$$

$$B(U) \equiv 0, \quad S(U) = \begin{pmatrix} -\dfrac{2}{r}\left(1-\dfrac{2M}{r}\right)\dfrac{1+k^2}{1-v^2}\rho v \\ \dfrac{-2r+5M}{r^2}\dfrac{v^2+k^2}{1-v^2}\rho - \dfrac{M}{r^2}\dfrac{1+k^2v^2}{1-v^2}\rho + 2\dfrac{r-2M}{r^2}k^2\rho \\ 0 \end{pmatrix},$$

$$\sigma(r) = r,$$

donde

$$v = \frac{1+k^2-\sqrt{(1+k^2)^2-4k^2\left(\frac{V^1}{V^0}\right)^2}}{2k^2\frac{V^1}{V^0}}, \qquad \rho = \frac{V^1(1-v^2)}{v(1+k^2)},$$

siendo $\rho$ la densidad del fluido, $v(t, r) \in (-1, 1)$ la velocidad normalizada, $M > 0$ el coeficiente que representa la masa del agujero negro, $k \in (-1, 1)$ denota la constante de la velocidad del sonido y con $r > 2M$. Estudiaremos este modelo en el Capítulo 4.

- El sistema de Burgers acoplado (ver [44]), que tiene la forma (I. 4) con:

$$U = \begin{pmatrix} u \\ v \end{pmatrix}, \quad F(U) = \begin{pmatrix} \dfrac{u^2}{2} \\ \dfrac{v^2}{2} \end{pmatrix},$$

$$B(U) = \begin{pmatrix} 0 & u \\ v & 0 \end{pmatrix}, \quad S(U) \equiv 0,$$

donde $U = (u, v)^T \in \{U \in \mathbb{R}^2 | u + v > 0\}$. Este sistema será estudiado en el Capítulo 5.

- El sistema de aguas someras modificado introducido en [42], que tiene la forma (I. 4) con:

$$U = \begin{pmatrix} u \\ q \end{pmatrix}, \quad F(U) = \begin{pmatrix} q \\ \\ \dfrac{q^2}{h} \end{pmatrix},$$

$$B(U) = \begin{pmatrix} 0 & 0 \\ qh & 0 \end{pmatrix}, \quad S(U) \equiv 0,$$

  donde $U = (u, v)^T \in \{U \in \mathbb{R}^2 | \quad 0 < q, \ 0 < h < (16q)^{1/3}\}$. Estudiaremos este modelo en el Capítulo 5.

Podemos englobar estos tres tipos de ecuaciones en la forma

$$W_t + \mathcal{A}(W)W_x = 0, \tag{I. 5}$$

donde

$$W = \begin{pmatrix} U \\ \sigma \end{pmatrix}, \quad \mathcal{A}(W) = \left(\begin{array}{c|c} A(U) + B(U) & -S(U) \\ \hline 0 & 0 \end{array}\right),$$

con $A(U) = JF(U)$, siendo $JF(U)$ la matriz jacobiana de $F$. La particularidad de este tipo de ecuaciones es que pueden no tener una solución clásica, es decir, una función diferenciable con derivadas parciales continuas, que cumple el sistema de ecuaciones y las condiciones iniciales impuestas, incluso partiendo de condiciones iniciales muy regulares. Por tanto, la definición de solución debe extenderse al concepto más general de solución débil que se verá en el Capítulo 1. Este concepto nos permite considerar soluciones discontinuas que son consistentes con la física del problema y que corresponden con fenómenos que aparecen en la naturaleza como son los saltos hidráulicos o los frentes en el caso del modelo de aguas poco profundas o las ondas de choque en los gases. En general, dado un dato inicial no hay unicidad de solución, de ahí que sea necesario introducir un criterio que nos permita seleccionar las soluciones consistentes con la física del problema. Esta discriminación nos la dará el concepto de entropía que se verá también en el Capítulo 1.

Como dijimos, en la mayor parte de los casos, la complejidad del problema hace que no sea posible resolver el problema de forma exacta, de ahí que tengamos que valernos de métodos numéricos que nos den soluciones razonablemente buenas. Estos incluyen, entre otros, los métodos de diferencias finitas, elementos finitos y volúmenes finitos. En esta tesis nos centraremos en estos últimos. En el caso de las leyes de conservación, desde los años 80 el avance en estos métodos numéricos ha sido muy grande gracias a los trabajos de von Neumann, Courant, Friedrichs, Lax, Wendroff, Godunov, van Leer, Harten, Roe, Osher, Colella, Yee, Oleinik, entre otros tantos (ver [154], [153], [115], [116], [85]).

Un caso particular de soluciones son aquellas que no dependen del tiempo y que se denominan soluciones estacionarias. En el caso de sistemas de leyes de equilibrio (I. 3),

las soluciones estacionarias satisfacen el sistema de Ecuaciones Diferenciales Ordinarias (EDO)

$$F(U)_x = S(U)\sigma_x. \tag{I. 6}$$

En muchos casos los flujos a simular se generan por una perturbación de una solución estacionaria (como es el caso de los tsunamis), por eso es importante que los métodos numéricos capturen bien las soluciones estacionarias para, de esta forma, hacer una buena simulación de dicha perturbación. En el contexto de las ecuaciones de aguas someras, Bermúdez y Vázquez-Cendón [13] introdujeron la propiedad que llamaron *C-property*: un método numérico la posee si resuelve con exactitud las soluciones estacionarias que representan el agua en reposo, quizás las soluciones estacionarias más intuitivas de este sistema. Esta idea de construir esquemas que preservan ciertos equilibrios o soluciones estacionarias, que se llaman en general esquemas *well-balanced* o *bien equilibrados*, ha sido un campo muy activo en los últimos años: [5, 12, 28, 34, 32, 131, 50, 145, 43, 47, 123, 122, 61, 60, 80, 79]. Se dará un pequeño repaso de esta parte en el Capítulo 1.

Por otro lado, en el caso de los sistemas no conservativos de la forma (I. 4) aparecen importantes dificultades desde el punto de vista analítico y numérico cuando $B \neq 0$ y no es el jacobiano de ninguna función o cuando $S \neq 0$ y $\sigma$ es discontinuo. En estos casos, no es posible definir los términos $B(U)U_x$, $S(U)\sigma_x$ en el sentido de las distribuciones cuando $U$ es discontinuo. Se habla entonces de productos no conservativos. Se han desarrollado diferentes teorías que permiten dar un sentido matemático preciso a estos productos no conservativos, como las que se describen en [163] y [49]. En esta tesis nos centraremos en la teoría que desarrollaron Dal Maso, LeFloch y Murat [57] que nos permite definir una solución débil como una función que satisface (I. 5) en el sentido de las medidas de Borel. Esta definición está basada en una familia de caminos Lipschitz-continua $\Phi : [0,1] \times \Omega \times \Omega \to \Omega$ que satisface una serie de propiedades de regularidad y consistencia. En particular,

$$\Phi(0, U_l, U_r) = U_l, \quad \Phi(1, U_l, U_r) = U_r, \quad \forall (U_l, U_r) \in \Omega \times \Omega,$$

y

$$\Phi(\xi, U, U) = U, \quad \forall \xi \in [0,1], \quad \forall U \in \Omega.$$

. Uno de los problemas con los que nos encontramos es que este concepto de solución débil depende de la elección de la familia de caminos. Una de las preguntas que surgen es cuál es la elección correcta. Supongamos el caso en que el sistema hiperbólico es el límite de un sistema parabólico

$$W_t^\epsilon + \mathcal{A}(W^\epsilon)\, W_x^\epsilon = \epsilon(\mathcal{R}(W^\epsilon)W_x^\epsilon)_x, \tag{I. 7}$$

cuando el coeficiente viscoso $\epsilon \to 0$, siendo $\mathcal{R}(W)$ cualquier matriz definida positiva. Entonces la familia de caminos escogida debe estar relacionada con los perfiles viscosos:

una función $V$ que conecta dos estados $W_l$ y $W_r$ se dice que es un perfil viscoso para el sistema (I. 7) si satisface

$$\lim_{\chi\to-\infty} V(\chi) = W^-, \quad \lim_{\chi\to+\infty} V(\chi) = W^+, \quad \lim_{\chi\to\pm\infty} V'(\chi) = 0, \tag{I. 8}$$

y existe $\sigma \in \mathbb{R}$ tal que la onda viajera

$$W^\epsilon(x,t) = V\left(\frac{x-\sigma t}{\epsilon}\right), \tag{I. 9}$$

es una solución de (I. 7) para todo $\epsilon$. Se puede probar que, para que $V$ sea un perfil viscoso, $V$ tiene que ser solución del sistema

$$-\sigma V' + \mathcal{A}(V)\,V' = (\mathcal{R}(V)\,V')', \tag{I. 10}$$

verificando las condiciones de contorno (I. 8). En caso de existir un perfil viscoso que una a $W_l$ con $W_r$, la buena elección del camino que conecta estos estados será, tras una reparametrización, el propio perfil viscoso $V$.

En términos numéricos nos centraremos, como se ha mencionado antes, en el diseño de métodos de volúmenes finitos, y más en concreto en los llamados métodos *path-conservative* o *camino-conservativos*. Estos métodos introducidos por Parés [132], basados en la definición de solución débil desarrollada por Dal Maso, LeFloch y Murat, permiten generalizar los métodos conservativos que se usan para resolver sistemas de leyes de conservación (I. 2). De este modo se generalizan métodos conocidos como los de Godunov [86] y Roe [140] a sistemas de la forma (I. 4). Una vez elegida la familia de caminos $\Phi$, estos métodos tienen la forma:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left(\mathcal{D}^+_{i-1/2} + \mathcal{D}^-_{i+1/2}\right), \tag{I. 11}$$

donde se ha usado la siguiente notación:

- $\Delta x$ y $\Delta t$ son los pasos de discretización del dominio espacial y temporal respectivamente. Se suponen constantes por simplicidad.
- $I_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ son las celdas en las que se divide el dominio espacial, cuyo tamaño es $\Delta x$.
- $t_n = n\Delta t$, $n = 0, 1\ldots$.
- $W_i^n$ es la aproximación de la media de la solución exacta en la $i$-ésima celda en el tiempo $t_n$, es decir,

$$W_i^n \approx \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W(x,t_n)\,dx. \tag{I. 12}$$

- Finalmente,
$$\mathcal{D}^{\pm}_{i+1/2} = \mathcal{D}^{\pm}\big(W^n_i, W^n_{i+1}\big),$$
donde $\mathcal{D}^-$ y $\mathcal{D}^+$ dos funciones Lipschitz-continuas de $\Omega \times \Omega$ a $\Omega$ que satisfacen
$$\mathcal{D}^{\pm}(W, W) = 0, \quad \forall W \in \Omega, \tag{I. 13}$$
y
$$\mathcal{D}^-(W_l, W_r) + \mathcal{D}^+(W_l, W_r) = \int_0^1 \mathcal{A}\big(\Phi(s; W_l, W_r)\big)\frac{\partial \Phi}{\partial s}(s; W_l, W_r)\, ds, \tag{I. 14}$$
para cada par $W_l, W_r \in \Omega$.

Veremos cómo obtener métodos de alto orden a través de operadores de reconstrucción de estados. Para ello, siguiendo [132], se usará una forma semi-discreta de los métodos path-conservative:
$$W_i'(t) = -\frac{1}{\Delta x}\left(\mathcal{D}^-_{i+\frac{1}{2}}(t) + \mathcal{D}^+_{i-\frac{1}{2}}(t) + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}^t_i(x))\frac{\partial}{\partial x}\mathbb{P}^t_i(x)\, dx\right), \tag{I. 15}$$
donde $\mathbb{P}^t_i(x)$ es la aproximación suave de la solución en la $i$-ésima celda dada por un operador de reconstrucción de alto orden aplicado sobre la secuencia de valores en las celdas $\{W_j(t)\}$ y
$$\mathcal{D}^{\pm}_{i+1/2}(t) = \mathcal{D}^{\pm}_{i+1/2}(W^-_{i+1/2}(t), W^+_{i+1/2}(t)),$$
donde $W^-_{i+1/2}(t) = \mathbb{P}^t_i(x_{j+\frac{1}{2}})$ y $W^-_{i+1/2}(t) = \mathbb{P}^t_{i+1}(x_{i+\frac{1}{2}})$ (ver [34] para más detalles). Finalmente este sistema de ecuaciones diferenciales ordinarias lo resolveremos a través de esquemas numéricos como los TVD Runge-Kutta (ver [88] y [148]).

Dentro de los métodos path-conservative nos centraremos en aquellos que tienen la propiedad de ser well-balanced como comentamos antes. Veremos distintas técnicas que nos permitirán obtener métodos con dicha propiedad y que estarán basadas en una buena elección de la familia de caminos. En el caso de considerar el problema semi-discreto (I. 15), dada una solución estacionaria $W^*$, usaremos la siguiente terminología:

- El método numérico (I. 15) se dice que es well-balanced para $W^*$ si el vector de las medias en las celdas de $W^*$ es un equilibrio del sistema de EDO (I. 15).
- El operador de reconstrucción se dice que es well-balanced para $W^*$ si
$$\mathbb{P}_i(x) = W^*(x), \qquad x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \tag{I. 16}$$
donde $\mathbb{P}_i$ es la aproximación de $W^*$ obtenida al aplicar el operador de reconstrucción al vector de las medias en las celdas de $W^*$.

Como veremos en el Capítulo 1 se puede probar que combinando un método path-conservative de primer orden well-balanced y un operador de alto orden que sea well-balanced se pueden obtener esquemas de alto orden con dicha propiedad. En general los operadores de reconstrucción estándar no tienen por qué tener la propiedad de ser well-balanced. Sin embargo, la técnica introducida en [40] nos permitirá obtener un operador de reconstrucción well-balanced a partir de uno estándar que denotaremos por

$$\mathbb{Q}_i(x) = \mathbb{Q}_i(x; \{W_j\}_{j\in S_i}),$$

siendo $S_i$ el conjunto de índices correspondiente a las celdas vecinas a la $i$-ésima cuyos valores se usan para calcular la reconstrucción en dicha celda. Los pasos a seguir para calcular la reconstrucción $\mathbb{P}_i$ en la celda $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ dada una familia de valores en las celdas $\{W_i\}$ son los siguientes:

1. Buscar, si es posible, la solución estacionaria $W_i^*(x)$ definida en el stencil de la celda $I_i$ $(\cup_{j\in\mathcal{S}_i} I_j)$ tal que:
$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}} W_i^*(x)dx = W_i, \tag{I. 17}$$
En otro caso consideramos $W_i^* \equiv 0$.

2. Calcular las fluctuaciones $\{V_j\}_{j\in S_i}$ en el stencil $S_i$:
$$V_j = W_j - \frac{1}{\Delta r}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} W_i^*(x)dx, \quad j \in S_i. \tag{I. 18}$$

3. Aplicar el operador de reconstrucción estándar a las fluctuaciones $\{V_j\}_{j\in S_i}$:
$$\mathbb{Q}_i(x) = \mathbb{Q}_i(x; \{V_j\}_{j\in S_i}).$$

4. Calcular la reconstrucción en la celda:
$$\mathbb{P}_i(x) = W_i^*(x) + \mathbb{Q}_i(x).$$

Se prueba fácilmente que $\mathbb{P}_i$ es well-balanced para cualquier solución estacionaria si se tiene que el operador de reconstrucción $\mathbb{Q}_i$ es exacto para la función nula. Además, $\mathbb{P}_i$ es conservativo, es decir,

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbb{P}_i(x)dx = W_i, \quad \text{para todo } i,$$

si $\mathbb{Q}_i$ es conservativo, y es de alto orden en precisión si las soluciones estacionarias son suaves (ver [47] para más detalles).

Esta estrategia, basada en el concepto de operadores de reconstrucción well-balanced de alto orden, se ha aplicado a distintos modelos: sistemas de shallow water [34, 43], flujo de sangre en los vasos sanguíneos [127], sistema de Euler con gravedad [78, 97], modelo de Ripa [146], entre otros.

## Contenidos y objetivos de la tesis

A lo largo de los capítulos de esta tesis se abordan cuatro problemas diferentes relacionados con el análisis numérico de sistemas de ecuaciones en derivadas parciales hiperbólicos no lineales: la unicidad de solución, el esfuerzo computacional de los métodos numéricos, la propiedad well-balanced y el control de la viscosidad numérica. Estos problemas están relacionados con algunas de las líneas de investigación del grupo EDANYA [71] y, más concretamente, con la resolución numérica de modelos matemáticos de la mecánica de fluidos en aplicaciones relacionadas con los flujos de aguas someras y la dinámica de gases en el contexto de la mecánica clásica o relativista.

### Capítulo 1: Preliminares

El objetivo del primer capítulo es dar las principales definiciones y resultados relacionados con el estudio analítico y numérico de los sistemas de Ecuaciones en Derivadas Parciales (EDPs) hiperbólicos conservativos y no conservativos. Primero nos centraremos en sus aspectos teóricos comenzando con leyes de conservación (I. 2). En concreto estaremos interesados en el estudio de los problemas de Cauchy asociados al sistema (I. 2):

$$\begin{cases} U_t + F(U)_x = 0, \\ U(x,0) = U_0(x), \end{cases}$$

siendo $U_0(x) : \mathbb{R} \longrightarrow \Omega$ una función conocida. En particular consideraremos problemas de Riemann, es decir, problemas de Cauchy con la siguiente condición inicial:

$$U_0(x) = \begin{cases} U_l & \text{if } x < 0, \\ U_r & \text{if } x > 0, \end{cases} \tag{I. 19}$$

donde $U_l, U_r \in \Omega$. Como dijimos antes los problemas de Cauchy pueden que no tengan una solución clásica incluso cuando la condición inicial es suave. Veremos cómo extender el concepto de solución clásica cuando aparecen discontinuidades dando lugar a la definición de *solución débil*. Como hemos mencionado previamente este tipo de soluciones tienen el problema de la no unicidad con lo que se introducirá también el criterio de *entropía*, que nos permitirá seleccionar las soluciones con sentido físico. Una vez seleccionado el criterio de entropía veremos que las soluciones de los problemas de Riemann asociados a los sistemas de leyes de conservación están compuestos por *ondas simples*: *rarefacciones*,

*choques*, *discontinuidades de contacto*.

Todos estos conceptos y resultados serán generalizados a sistemas de la forma

$$W_t + \mathcal{A}(W)W_x = 0, \tag{I. 20}$$

que incluyen a las leyes de equilibrio (I. 3) y a sistemas con productos no conservativos (I. 4). La teoría desarrollada por Dal Maso, LeFloch y Murat [57], basada en la elección de una familia de caminos Lipschitz-continua, nos permitirá extender el concepto de *solución débil* y de condición de entropía a los sistemas de la forma (I. 20).

Tras el estudio teórico nos centraremos en los aspectos numéricos y, en particular, en el diseño de métodos de volúmenes finitos. Dentro de estos estudiaremos los métodos *camino-conservativos* o *path-conservative* introducidos por Parés en [132], que son una generalización de la noción de método conservativo para los sistemas de leyes de conservación (I. 2). Estos métodos, basados en la definición de solución débil dada por Dal Maso et al., nos permitirán extender los métodos de Godunov [86], Roe [93], Polinomial Viscosity matrix [140] y Resolvedores de Riemann simples [93] a sistemas de la forma (I. 20). Estos métodos, combinados con operadores de reconstrucción, serán la base para el desarrollo de métodos de alto orden shock-capturing, es decir, métodos que son de alto orden de precisión en las regiones en las que la solución es regular y que capturan correctamente la formación y propagación de discontinuidades. Describiremos con más detalle aquellos que usaremos en otros capítulos como son el operador de reconstrucción MUSCL [158] y CWENO [117]. Tras esto el concepto de método *bien equilibrado* o *well-balanced* será introducido y veremos bajo qué condiciones los métodos anteriormente mencionados satisfacen esta propiedad. Combinando los métodos well-balanced de primer orden y los operadores de reconstrucción well-balanced (I. 16), veremos cómo conseguir esquemas de alto orden que tengan además la propiedad de ser well-balanced. Finalmente haremos una pequeña discusión sobre la convergencia de estos métodos concluyendo que hay que pedirles consistencia, estabilidad, control de la entropía y control sobre la viscosidad numérica.

**Capítulo 2: El problema de Riemann para las ecuaciones de aguas someras con topografía: el caso seco-mojado**

En este capítulo estudiaremos con más detenimiento el sistema de aguas someras (I. 1). En el caso en el que el fondo $a$ es discontinuo, el término fuente de las ecuaciones de aguas someras es un producto no conservativo, por lo que aparecen las dificultades anteriormente mencionadas. En el capítulo nos centramos en los problemas de Riemann en un fondo con forma de escalón, es decir, nuestra condición inicial será de la forma:

$$(h,u,a)(x,0) = \begin{cases} (h_l,u_l,a_l), & x<0, \\ (h_r,u_r,a_r), & x>0, \end{cases} \tag{I. 21}$$

con $a_l \neq a_r$. El caso en que $a_l = a_r$ está ya totalmente resuelto (ver [153]) y, en el caso en que $a_l \neq a_r$, se han realizado diferentes estudios analíticos y numéricos de las soluciones de los problemas de Riemann: [3, 14, 141, 113, 20, 16, 21, 91], etc. En este capítulo seguiremos el estudio realizado en [112] pero centrándonos en el caso que no se estudió que corresponde a la situación en que no hay agua en alguno de los dos lados del escalón. En [41] también se consideró este problema pero solo en los casos en los que el escalón actúa como un obstáculo para el fluido, nosotros abarcaremos todas las posibles situaciones. Este problema es fundamental para el diseño de esquemas que traten bien los frentes seco-mojados.

Siguiendo [112], comenzaremos viendo los autovalores y autovectores del sistema para, de esta forma, hallar aquellas zonas en las que el sistema es estrictamente hiperbólico o solamente hiperbólico. Continuaremos viendo cómo son las ondas simples, describiendo con detalle las rarefacciones, choques y discontinuidades de contacto. Al igual que en [112] un criterio de monotonía (MC) se impondrá para seleccionar las discontinuidades de contacto admisibles sobre el escalón. Tras esto construiremos las soluciones a los distintos problemas de Riemann combiando ondas simples. Se verá que, dependiendo de las condiciones iniciales, el problema tiene 0, 1 o 2 soluciones. En el caso de que no haya ninguna solución, siguiendo [41], se reinterpretará el problema como un problema parcial de Riemann lo que nos permitirá construir una solución consistente con la física del problema. Finalmente consideraremos varios test numéricos para comparar la aproximación numérica obtenida por los distintos métodos numéricos. Al ser un sistema con productos no conservativos, nos encontraremos con los problemas que suelen aparecer en estos casos: se verá que, en los casos donde la solución no es única, distintos métodos pueden converger a diferentes soluciones. Veremos que en los casos en los que el problema de Riemann no tiene solución, los métodos convergen a la solución que proponemos. En particular observaremos que los mejores resultados vienen de la combinación del flujo de Godunov y de la técnica de la reconstrucción hidrostática generalizada que veremos en el Capítulo 1 para tratar el término fuente. El contenido de este capítulo fue publicado por Parés y Pimentel-García en 2019 en *Journal of Computational Physics*, ver [134].

### Capítulo 3: Implementación eficiente de métodos PVM y resolvedores de Riemann simples. Aplicación al método de Roe para grandes sistemas hiperbólicos

El objetivo de este capítulo es implementar de forma eficiente el método de Roe cuando tenemos un número grande de ecuaciones en el sistema considerado. La implementación estándar de dicho método se basa en el cálculo del valor absoluto de la matriz de Roe, que requiere del conocimiento explícito su espectro. Si es necesario calcularlos numéricamente (por no ser conocida su expresión analítica o ser muy costoso su cálculo), el coste computacional puede ser elevado especialmente si el tamaño de la matriz es grande.

En estos casos puede ser preferible usar otros tipos de métodos como son los *Polinomial Viscosity matrix (PVM)* y los *Resolvedores de Riemann Aproximados (ARS)*.

En el caso de los métodos PVMs lo que se hace es aproximar el valor absoluto de la matriz de Roe por su evaluación en un determinado polinomio que aproxima a la función valor absoluto. Nosotros consideraremos aquí polinomios que interpolan la función valor absoluto en algunos puntos en el sentido de Lagrange o de Hermite.

Los Resolvedores de Riemann aproximados por su parte están basados en la aproximación de las soluciones de los problemas de Riemann asociados a las interceldas. Un caso particular, que serán los que consideraremos en este capítulo, son los Resolvedores de Riemann Simples (SRS), en los que dichas aproximaciones consisten en varios estados constantes unidos por discontinuidades que viajan a una velocidad constante.

Veremos que los SRS y los PVM, introducidos en el Capítulo 1, tienen una estrecha relación. En particular, veremos que, bajo ciertas hipótesis, los SRS se pueden interpretar como PVMs y viceversa. Esta relación ya se vio en [125] y en este capítulo se volverá a ver con demostraciones simplificadas. Veremos que el método de Roe se puede interpretar como un método PVM basado en el polinomio que interpola el valor absoluto de todos sus autovalores, como ya se vio en [35]. Esto nos permitirá implementarlo usando para ello la forma de Newton del polinomio de interpolación, que es la más eficiente.

Finalmente comprobaremos que, efectivamente, esta implementación del método de Roe nos permite reducir los tiempos de cómputo. Compararemos la implementación estándar de Roe con esta nueva forma. Para ello usaremos el modelo bicapa de aguas someras [44] y el *Quadrature Based Moment Equations (QBME)* [102]. Se verá que la reducción de coste computacional crece con el número de ecuaciones. El contenido de este capítulo fue publicado por Pimentel-García, Parés, Castro y Koellermeier en 2021 en la revista *Applied Mathematics and Computations*, ver [136].

### Capítulo 4: Esquemas well-balanced para fluidos en el contexto de Schwarzschild

Este capítulo se centra en la simulación de fluidos en un contexto relativista basado en la métrica de Schwarzchild. Dicha métrica es una solución exacta de las ecuaciones de Einstein del campo gravitatorio que describe el campo generado por una estrella o una masa esférica. En concreto, nos centramos en el estudio numérico del comportamiento a largo plazo de un flujo que evoluciona alrededor de un agujero negro de Schwarzschild. Suponemos que el flujo tiene una simetría esférica, por lo que solo se consideran los modelos con una coordenada espacial (la distancia al centro del agujero negro). En concreto consideraremos, consideraremos los modelos de Burgers-Schwarzschild y Euler-Schwarzschild relativistas. Con el fin de poder realizar simulaciones numéricas fiables y precisas del comportamiento

del flujo, se diseñarán métodos de volumen finito de alto orden well-balanced para ambos sistemas. Ambos fueron estudiados en [62, 108, 114] pero aquí extenderemos la metodología introducida en [47] y presentada en el Capítulo 1 a este tipo de problemas para obtener así esquemas de alto orden well-balanced. Aunque daremos las pautas necesarias para construir esquemas de precisión arbitraria, nos centraremos en los de primer, segundo y tercer orden.

En el caso del modelo de Burgers relativista es posible calcular sus soluciones estacionarias de forma explícita lo que nos permitirá resolver fácilmente el problema (I. 17) que surge en la primera etapa del procedimiento de reconstrucción well-balanced. Usaremos los esquemas para llevar a cabo un estudio sistemático del comportamiento asintótico de las soluciones cuando el tiempo tiende a infinito. Se verá que la propiedad well-balanced es fundamental para estudiar la evolución de las perturbaciones de una solución estacionaria.

En el caso de las ecuaciones de Euler relativista, dispondremos de una expresión implícita de las soluciones estacionarias, por lo que es necesario usar un método numérico para evaluarlas en un punto: usaremos el método de Newton. Tras obtener los esquemas well-balanced de primer y segundo orden, haremos un nuevo estudio sistemático del comportamiento asintótico de las soluciones.

El contenido de este capítulo está disponible en el repositorio *arXiv* y ha sido enviado para su publicación a la revista *Journal of Scientific Computing*, ver [110], y está actualmente en fase de modificación tras los primeros informes de los revisores.

### Capítulo 5: Métodos camino-conservativos con reconstrucción discontinua en las celdas para sistemas hiperbólicos no conservativos: extensión a segundo orden

En este capítulo abordaremos el problema de la convergencia de los métodos path-conservative. Como se ha comentado, en el caso de los sistemas no conservativos de la forma (I. 20) se tiene que la consistencia, estabilidad y control de la entropía no son suficientes para asegurar la convergencia a las soluciones débiles seleccionadas. Esto es debido a que la viscosidad numérica de los métodos alteran la condición de salto que satisfacen los límites de las soluciones aproximadas: en lugar de ser las asociadas a la familia de caminos elegida, están relacionadas con los perfiles viscosos de la ecuación equivalente (ver [42], [1]). En [51] una técnica basada en reconstrucciones discontinuas en las celdas permitió al autor diseñar métodos path-conservative de primer orden que capturan correctamente las soluciones con choques aislados. El objetivo principal de este capítulo es extender esta técnica a segundo orden de precisión y establecer las bases para su generalización a orden arbitrario. Para ello usaremos la forma semi-discreta de los

métodos path-conservative (I. 15) a la que aplicaremos la reconstrucción MUSCL-Hancock [161]: se usará un operador de reconstrucción lineal de tipo *minmod* (ver [158]) en espacio y tiempo. Enunciaremos y probaremos que esta extensión sigue manteniendo la propiedad de capturar de forma exacta los choques aislados. Para validar el método numérico lo aplicaremos a los siguientes sistemas no conservativos: Couple-Burgers [44], Gas dynamics equations in Lagrangian coordinates [1] y a las modified shallow water [42]. Veremos que el método propuesto captura correctamente los choques aislados y que mejora los resultados obtenidos por el método introducido en [51].

El contenido de este capítulo está disponible en el repositorio *arXiv* y ha sido enviado para su publicación a la revista *Journal of Computational Physics*, ver [137].

### Capítulo 6: Conclusiones y trabajos futuros

### Conclusiones

- En el Capítulo 2 obtenemos las siguientes conlusiones:
  - El problema de Riemann correspondiente a un escalón con agua a uno de sus lados puede tener 0, 1 o 2 soluciones.
  - En los casos en los que el problema de Riemann no tiene solución, se ha propuesto una reinterpretación que proporciona una solución consistente con la física del problema.
  - Para capturar correctamente las discontinuidades de contacto estacionarias necesitamos un esquema numérico que las preserve. Sin embargo, esto no es suficiente para asegurar la convergencia a la solución propuesta: la solución numérica puede converger a una solución débil que no satisfaga el criterio de monotonía que se verá en este capítulo.
  - El esquema que combina el flujo numérico de Godunov y la técnica de reconstrucción hidrostática generalizada para el tratamiento de la fuente es el que parece capturar correctamente todas las soluciones propuestas.
  - En los casos en los que tenemos dos posibles soluciones al problema de Riemann, los métodos numéricos convergen a una u otra.
- En el Capítulo 3 veremos que la implementación propuesta es más eficiente que la estándar, aumentando la diferencia con el número de ecuaciones: en el caso del modelo QBME la nueva implementación llega a ser hasta 4.1 veces más rápida que la estándar. Se observa también que esta diferencia es mayor conforme se va aumentando el número de ecuaciones.
- En el Capítulo 4 resaltaremos la importancia que tienen los métodos well-balanced cuando tratamos con los modelos relativistas de Burgers y Euler. A través de una

extensa batería de tests sacaremos conclusiones del comportamiento a largo plazo de los fluidos considerados.

- En el Capítulo 5 conseguimos extender a segundo orden la técnica desarrollada en [51] manteniendo la propiedad de capturar de forma exacta los choques aislados. Además establecemos las bases para una generalización de esta técnica a esquemas de alto orden que puedan capturar más de un choque de forma exacta.

.

### Futuras líneas de trabajo

El estudio de los problemas de Riemann correspondientes a los frentes seco-mojados realizado en el Capítulo 2 pueden extenderse a otros sistemas de shallow water como el sistema de shallow water con dos velocidades considerado en [2].

La nueva implementación eficiente de los métodos PVM basados en la forma de Newton de los polinomios puede ser de gran utilidad a la hora de estudiar sistemas con un número grande de ecuaciones, como ocurre con los sistemas multicapa. En el caso del sistema multicapa considerado en [44] no conocemos de forma exacta todos los autovalores y autovectores con lo que se espera obtener una gran mejora en los tiempos de cómputo. Lo mismo ocurre con otros sistemas multicapa como los de [6, 7].

Se prevén las siguientes extensiones del estudio numérico de los sistemas relativistas en el contexto de Schwarzschild:

- Desarrollo de esquemas de orden mayor que dos well-balanced para el modelo Euler-Schwarzschild basados en la aproximación numérica de las soluciones estacionarias mediante resolvedores de sistemas de EDOs (ver [87]).

- Desarrollo de métodos de alto orden well-balanced para problemas multidimensionales.

- Desarrollo de métodos numéricos para otros modelos relativistas de mayor complejidad.

La extensión a segundo orden de la reconstrucción discontinua del Capítulo 5 nos permite establecer las bases para las siguientes líneas de investigación:

- El diseño de esquemas de volúmenes finitos de alto orden que capturen correctamente choques aislados.

- Capturar correctamente choques no aislados.

- Aplicar los métodos a modelos más complejos.
- Proponer nuevos resolvedores Discontinuous Galerkin (DG) basados en el uso de reconstrucciones discontinuas.
- Explorar la extensión a problemas multidimensionales.

# Abstract

This thesis addresses four different problems related to the numerical analysis of hyperbolic systems of nonlinear partial differential equations. These problems are related to some of the lines of research of the EDANYA group and, more specifically, to the numerical resolution of mathematical models of fluid mechanics in applications related to shallow water flows and gas dynamics in the context of classical or relativistic mechanics. The issues addressed are listed below in chronological order.

The first problem addressed is the study of the Riemann problem for the shallow water equations on a step-shaped bottom, in the particular case in which there is only water on one side of the step. In this way we will complete the study carried out by LeFloch and Thanh in [112]. The analysis of these situations of dry-wet fronts is important when designing numerical schemes that deal well with flood phenomena. Two major difficulties arise when studying this problem. On the one hand, the source term that appears in the equations is a nonconservative product, so that there is no a unique way to define the weak solutions to the problem. In our case we will follow the theory of Dal Maso, LeFloch and Murat [57] to define them from a family of paths. On the other hand, resonant cases appear (i.e., an eigenvalue of the Jacobian matrix vanishes) that implies that, once the definition of weak solution is chosen, there is no uniqueness of solution. This theoretical study is complemented with numerical tests in which the behavior of different schemes is studied.

The next question to be addressed is the efficient implementation of numerical methods based on approximate Riemann solvers and, in particular, the Roe method, which is based on solving linearized Riemann problems in the intercells. The practical interest of this chapter resides above all in the numerical resolution of large systems such as multilayer shallow water models [44] or models based on moments [98]. This new implementation is based on the close relationship between this kind of methods and the Polynomial Viscosity Matrix ones based on the election of a polynomial that interpolates the absolute value function, as well as on the Newtonian form of the interpolation polynomial.

The next objective is to make a systematic study of the asymptotic behavior of the solutions of the relativistic Burgers and Euler models based on the Schwarzchild metric,



using numerical methods. This metric is an exact solution of the Einstein equations of the gravitational field that describes the field generated by a star or a spherical mass. In these systems the stationary solutions and the evolution of its perturbations play a fundamental role in understanding the flow behavior. Therefore, the use of well-balanced methods, i.e., methods that preserve these solutions, is essential. We will apply the general framework described in [47] to develop well-balanced methods of order up to 3 for the Burgers-Schwarzschild model and 2 for the Euler-Schwarzschild model. We will compare the results obtained between these methods and the standard ones and we will show the relevance of the well-balanced property.

It is well-known that, in the case of systems with nonconservative products, consistency, stability and entropy control are not sufficient to ensure the convergence of the numerical approximations to admissible weak solutions: it is also necessary to control the small-scale effects such as the numerical viscosity that affects the position and amplitude of shock waves (see [109]). In [51] Chalons presented a technique based on in-cell discontinuous reconstructions that allowed him to design first-order methods that capture exactly isolated shocks. The last problem that arises in the thesis is the extension to second-order of this technique using the formalism of the path-conservative methods introduced in [132], what will allow to set the basis for its generalization to arbitrary order. We will state and prove that this second-order extension keeps the property of capturing exactly isolated shocks and we will verify this through different numerical tests.

In the last part we summarize the main contributions of this thesis and the possible future lines of research.

# Introduction

Computational Fluids Mechanics is today one of the most important mathematical tools in the simulation of various phenomena that happen around us. Its objective is to simulate the evolution of fluids through the numerical resolution of systems of partial differential equations (PDEs) that govern their behavior. The main difficulty of these systems is that in most cases it is not possible to solve them exactly, hence the need to use numerical methods. Through these numerical methods we will obtain simulations that will allow us to understand, predict and control the evolution of fluid flows. This type of tools has applications in different fields of study such as oceanography, meteorology, climatology, hydraulic engineering, aeronautics, biology, etc. The development of suitable numerical methods requires a deep knowledge of both the physical nature of the flows to be simulated and the mathematical properties of the systems to be solved.

In fluid mechanics one of the most general partial differential equations governing the motion of fluids are the well-known Navier-Stokes equations (see [151]). These equations express the conservation of mass, momentum and energy, together with an equation of state that relates pressure, energy, and density. Through certain hypotheses about the fluids we are considering, it is possible to simplify these general equations in order to derive simpler models such as the so-called *shallow water equations*. In their one-dimensional version, these equations were first derived by by Jean Claude Barré de Saint-Venant in 1843 hence they are called sometimes Saint-Venant equations. These equations describe the motion of a thin layer of fluid. In the one-dimensional case, they are obtained from the Navier-Stokes equations and a series of hypotheses:

- Water is assumed to be homogeneous and incompressible.
- The pressure is hydrostatic, i.e., the pressure increases with depth, being the pressure at the surface of the fluid equal to the air pressure.
- The only force acting on the fluid is pressure (viscous effects are neglected).
- Both the bottom on which the water evolves and its free surface can be represented by the graph of a function. Moreover, these functions only depend on one of the horizontal variables, $x$, and on the time $t$ (in the case of the free surface).

- It is assumed that the velocity of the fluid only depends on $x$ and $t$. Furthermore, the deviations of its horizontal component from its vertical average are assumed to be negligible.

From these hypotheses and through a vertical integration process the following equations for the conservation of mass and momentum are obtained:

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + \dfrac{gh^2}{2}\right) = -gh\partial_x a, \\ \partial_t a = 0, \end{cases} \tag{I. 22}$$

where

- $h = h(x,t) \geq 0$ is the thickness of the layer;
- $u = u(x,t)$ is the depth-averaged horizontal velocity of the water;
- $g$ is the intensity of the gravitational field;
- $a = a(x)$ is the depth of the bottom from a reference level.

This system can be written in the following form:

$$U_t + F(U)_x = S(U)a_x,$$

where:

$$U = \begin{pmatrix} h \\ hu \\ a \end{pmatrix}, \quad F(U) = \begin{pmatrix} hu \\ hu^2 + \dfrac{gh^2}{2} \\ 0 \end{pmatrix}, \quad S(U) = \begin{pmatrix} 0 \\ -gh \\ 0 \end{pmatrix}.$$

In this thesis we will consider the following three types of PDEs:

- *Systems of conservation laws*:

$$U_t + F(U)_x = 0, \tag{I. 23}$$

- *Systems of balance laws*:

$$U_t + F(U)_x = S(U)\sigma_x, \tag{I. 24}$$

- *Systems with nonconservative products*:

$$U_t + F(U)_x + B(U)U_x = S(U)\sigma_x, \tag{I. 25}$$

where the unknown $U(x,t) = (u_1(x,t), ..., u_N(x,t))^T$ takes values in an open convex set $\Omega$ of $\mathbb{R}^N$, $F$ is a regular function from $\Omega$ to $\mathbb{R}^N$, $B$ is a regular matrix function from $\Omega$ to $\mathcal{M}_{NxN}(\mathbb{R})$, $S$ is a function from $\Omega$ to $\mathbb{R}^N$, and $\sigma(x)$ is a known function from $\mathbb{R}$ to $\mathbb{R}$. As we have seen the shallow water equations with topography are a system of balance laws that will be considered in Chapter 2. Some examples of systems with nonconservative products that we will see throughout the thesis are:

- The homogeneous two-layer 1-D shallow water system (see [44], [74]) can be written in the form (I. 25) with:

$$U = \begin{pmatrix} h_1 \\ q_1 \\ h_2 \\ q_2 \end{pmatrix}, \quad F(U) = \begin{pmatrix} q_1 \\ \frac{q_1^2}{h_1} + \frac{1}{2}gh_1^2 \\ q_2 \\ \frac{q_2^2}{h_2} + \frac{1}{2}gh_2^2 \end{pmatrix},$$

$$B(U) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & gh_1 & 0 \\ 0 & 0 & 0 & 0 \\ grh_2 & 0 & 0 & 0 \end{pmatrix}, \quad S(U) \equiv 0,$$

  being $r = \rho_1/\rho_2$, where:

  - $h_i = h_i(x,t) \geq 0$ is the thickness of the $i$-th layer.
  - $q_i = q_i(x,t) = h_i(x,t)u_i(x,t)$ is the discharge of the $i$-th layer, where $u_i(x,t)$ is the depth-averaged horizontal velocity.
  - $g$ is the intensity of the gravitational field.
  - $\rho_i$ is the constant density of the $i$-th layer.

  This system will be considered in Chapter 3.

- The *relativistic Burgers-Schwarzschild model* [114] is given in the form (I. 25) by:

$$U = \begin{pmatrix} v \\ r \end{pmatrix}, \quad F(U) = \begin{pmatrix} \left(1 - \frac{2M}{r}\right)\frac{v^2-1}{2} \\ 0 \end{pmatrix},$$

$$B(U) \equiv 0, \quad S(U) = \begin{pmatrix} \frac{2M}{r^2}(v^2-1) \\ 0 \end{pmatrix}, \quad \sigma(r) = r,$$

  where $r > 2M, v = (t,r) \in [-1,1]$, being $r$ the distance to the black hole, $v$ the normalized velocity of the flow and $M > 0$ is a coefficient representing the mass of the black hole. We will study this system in Chapter 4.

- The *relativistic Euler-Schwarzschild model*, which can be written in the form (I. 25) with:

$$U = \begin{pmatrix} V^0 \\ V^1 \\ r \end{pmatrix} = \begin{pmatrix} \dfrac{1+k^2v^2}{1-v^2}\rho \\ \dfrac{1+k^2}{1-v^2}\rho v \\ r \end{pmatrix}, \qquad F(U) = \begin{pmatrix} \left(1-\dfrac{2M}{r}\right)\dfrac{1+k^2}{1-v^2}\rho v \\ \left(1-\dfrac{2M}{r}\right)\dfrac{v^2+k^2}{1-v^2}\rho \\ 0 \end{pmatrix},$$

$$B(U) \equiv 0, \quad S(U) = \begin{pmatrix} -\dfrac{2}{r}\left(1-\dfrac{2M}{r}\right)\dfrac{1+k^2}{1-v^2}\rho v \\ \dfrac{-2r+5M}{r^2}\dfrac{v^2+k^2}{1-v^2}\rho - \dfrac{M}{r^2}\dfrac{1+k^2v^2}{1-v^2}\rho + 2\dfrac{r-2M}{r^2}k^2\rho \\ 0 \end{pmatrix},$$

$$\sigma(r) = r,$$

where

$$v = \frac{1+k^2-\sqrt{(1+k^2)^2-4k^2\left(\frac{V^1}{V^0}\right)^2}}{2k^2\frac{V^1}{V^0}}, \qquad \rho = \frac{V^1(1-v^2)}{v(1+k^2)},$$

being $\rho$ the fluid density, $v(t,r) \in (-1,1)$ the normalized velocity, $M > 0$ the coefficient representing the mass of the black hole, $k \in (-1,1)$ denotes the (constant) speed of sound and with $r > 2M$. We will study this system in Chapter 4.

- The coupled Burgers system (see [44]), which can be written in the form (I. 25) with:

$$U = \begin{pmatrix} u \\ v \end{pmatrix}, \quad F(U) = \begin{pmatrix} \dfrac{u^2}{2} \\ \dfrac{v^2}{2} \end{pmatrix},$$

$$B(U) = \begin{pmatrix} 0 & u \\ v & 0 \end{pmatrix}, \quad S(U) \equiv 0,$$

where $U = (u,v)^T \in \{U \in \mathbb{R}^2 | u+v > 0\}$. This system will be studied in Chapter 5.

- The modified shallow water system introduced in [42], which has the form (I. 25) with:

$$U = \begin{pmatrix} u \\ q \end{pmatrix}, \quad F(U) = \begin{pmatrix} q \\ \dfrac{q^2}{h} \end{pmatrix},$$

$$B(U) = \begin{pmatrix} 0 & 0 \\ qh & 0 \end{pmatrix}, \quad S(U) \equiv 0,$$

where $U = (u,v)^T \in \{U \in \mathbb{R}^2 | \quad 0 < q,\ 0 < h < (16q)^{1/3}\}$.

We can write these types of equations in the form

$$W_t + \mathcal{A}(W)W_x = 0, \tag{I. 26}$$

where

$$W = \begin{pmatrix} U \\ \sigma \end{pmatrix}, \quad \mathcal{A}(W) = \left(\begin{array}{c|c} A(U) + B(U) & -S(U) \\ \hline 0 & 0 \end{array}\right),$$

with $A(U) = JF(U)$, being $JF(U)$ the jacobian matrix of $F$. The peculiarity of this type of equations is that they may not have a classical solution, i.e., a differentiable function with continuous partial derivatives, that satisfies the system of equations and the imposed initial conditions, even starting from very regular initial conditions. Therefore, the definition of solution must be extended to the more general concept of weak solution that will be seen in Chapter 1. This concept allows us to consider discontinuous solutions that are consistent with the physics of the problem and that correspond to phenomena that appears in nature such as hydraulic jumps or fronts in the case of the shallow water model or shock waves in a gas. In general, given an initial data there is no uniqueness of solution, hence it is necessary to introduce a criterion that allows us to select the ones consistent with the physics of the problem. This discrimination will be given by the concept of entropy that will be seen in Chapter 1.

As we said, in most cases, the complexity of the problem may impede its exact solution, hence we have to use numerical methods that give us reasonably good solutions. These include, among others, the finite difference, finite elements and finite volume methods. In this dissertation we will focus on the latter. In the case of conservation laws, since the 1980s the advance in these numerical methods has been great thanks to the works of von Neumann, Courant, Friedrichs, Lax, Wendroff, Godunov, van Leer, Harten, Roe, Osher, Colella, Yee, Oleinik, among many others (see [154], [153], [115], [116], [85]).

A particular case of solutions are those that do not depend on time and that are called stationary solutions. In the case of systems of balance laws the stationary solutions satisfy the Ordinary Differential Equations (ODE) system:

$$F(U)_x = S(U)\sigma_x. \tag{I. 27}$$

In many cases the flows to be simulated are generated by a perturbation of a stationary solution (as in the case of tsunamis), for this reason it is important that the numerical methods capture well the stationary solutions in order to make a good follow-up of the perturbation. In the context of shallow water equations, Bermúdez and Vázquez-Cendón [13] introduced the property that they called *C-property*: a numerical method has this property if it solves exactly the stationary solutions corresponding to water at rest, perhaps the most intuitive stationary solutions of this system. This idea of building schemes that preserve some equilibria or stationary solutions, which are generally called *well-balanced* schemes, has been a very active field in recent years:

[5, 12, 28, 34, 32, 131, 50, 145, 43, 47, 123, 122, 61, 60, 80, 79]. A short review of this subject will be given in Chapter 1.

In the case of nonconservative systems of the form (I. 23) important difficulties appear from the analytical and numerical points of view when $B \neq 0$ and it is not the Jacobian of any function or when $S \neq 0$ and $\sigma$ is discontinuous. In these cases, it is not possible to define $B(U)U_x$, $S(U)\sigma_x$ in the sense of distributions when $U$ is discontinuous. We talk then about nonconservative products. Different theories have been developed that allow giving a precise mathematical sense to these nonconservative products, like those described in [163] and [49]. In this thesis we will follow the theory developed by Dal Maso, LeFloch and Murat [57] that allows one to define a weak solution as a function that satisfies (I. 26) in the sense of Borel measures. This definition is based on the choice of a family of Lipschitz continuous paths $\Phi : [0,1] \times \Omega \times \Omega \to \Omega$ satisfying certain regularity and consistency properties. In particular

$$\Phi(0, U_l, U_r) = U_l, \quad \Phi(1, U_l, U_r) = U_r, \quad \forall (U_l, U_r) \in \Omega \times \Omega,$$

and

$$\Phi(\xi, U, U) = U, \quad \forall \xi \in [0,1], \quad \forall U \in \Omega.$$

One of the problems that arise from this theory is that the concept of weak solutions depends clearly on the choice of the family of paths, which is a priori arbitrary, so that the crucial question is how to choose the correct one. In fact, when the hyperbolic system is the vanishing-viscosity limit of the parabolic problem

$$W_t^\epsilon + \mathcal{A}(W^\epsilon)\, W_x^\epsilon = \epsilon(\mathcal{R}(W^\epsilon)W_x^\epsilon)_x, \tag{I. 28}$$

where $\mathcal{R}(W)$ is any positive-definite matrix, the adequate family of paths should be related to the viscous profiles: a function $V$ is said to be a viscous profile for (I. 28) linking the states $W^-$ and $W^+$ if it satisfies

$$\lim_{\chi \to -\infty} V(\chi) = W^-, \quad \lim_{\chi \to +\infty} V(\chi) = W^+, \quad \lim_{\chi \to \pm\infty} V'(\chi) = 0 \tag{I. 29}$$

and there exists $\sigma \in \mathbb{R}$ such that the travelling wave

$$W^\epsilon(x,t) = V\left(\frac{x - \sigma t}{\epsilon}\right), \tag{I. 30}$$

is a solution of (I. 28) for every $\epsilon$. It can be easily verified that, in order to be a viscous profile, $V$ has to solve the equation

$$-\sigma V' + \mathcal{A}(V)\, V' = (\mathcal{R}(V)\, V')', \tag{I. 31}$$

with boundary conditions (I. 29). If there exists a viscous profile linking the states $W^-$ and $W^+$, the good choice for the path connecting the states would be, after a reparameterization,

the viscous profile $V$.

We will focus on the design of finite volume methods for systems of the form (I. 26), and more specifically on the so-called *path-conservative* methods. These methods introduced by Parés [132], based on the definition of weak solution developed by Dal Maso, LeFloch and Murat, allow us to generalize well-known methods for systems of conservation laws such as Godunov [86], Roe [140], Approximate Riemann solvers [93], etc. Given a family of paths $\Phi$, these methods have the form:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\big(\mathcal{D}^+_{i-1/2} + \mathcal{D}^-_{i+1/2}\big), \tag{I. 32}$$

where the following notation has been used:

- $\Delta x$ and $\Delta t$ are the space and time steps respectively. They are supposed to be constant for simplicity.
- $I_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ are the computational cells, whose length is $\Delta x$.
- $t_n = n\Delta t$, $n = 0, 1 \dots$.
- $W_i^n$ is the approximation of the average of the exact solution at the $i$th cell at time $t_n$, that is
$$W_i^n \approx \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W(x, t_n)\, dx. \tag{I. 33}$$
- Finally,
$$\mathcal{D}^\pm_{i+1/2} = \mathcal{D}^\pm\big(W_i^n, W_{i+1}^n\big),$$
where $\mathcal{D}^-$ and $\mathcal{D}^+$ two Lipschitz continuous functions from $\Omega \times \Omega$ to $\Omega$ that satisfy
$$\mathcal{D}^\pm(W, W) = 0, \quad \forall W \in \Omega, \tag{I. 34}$$
and
$$\mathcal{D}^-(W_l, W_r) + \mathcal{D}^+(W_l, W_r) = \int_0^1 \mathcal{A}\big(\Phi(s; W_l, W_r)\big)\frac{\partial \Phi}{\partial s}(s; W_l, W_r)\, ds, \tag{I. 35}$$
for every set $W_l, W_r \in \Omega$.

We will see how to obtain high-order methods through the use of reconstruction operators. In order to do this, following [132], a semi-discrete form of the path-conservative methods will be used:

$$W_i'(t) = -\frac{1}{\Delta x}\left(\mathcal{D}^-_{i+\frac{1}{2}}(t) + \mathcal{D}^+_{i-\frac{1}{2}}(t) + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^t(x))\frac{\partial}{\partial x}\mathbb{P}_i^t(x)\, dx\right), \tag{I. 36}$$

where $\mathbb{P}_i^t(x)$ is the smooth approximation of the solution at the $i$th-cell provided by a high-order reconstruction operator from the sequence of cell values $\{W_i(t)\}$ and

$$\mathcal{D}^{\pm}_{i+1/2}(t) = \mathcal{D}^{\pm}_{i+1/2}(W^-_{i+1/2}(t), W^+_{i+1/2}(t)),$$

where $W^-_{i+1/2}(t) = \mathbb{P}_i^t(x_{j+\frac{1}{2}})$ and $W^-_{i+1/2}(t) = \mathbb{P}_{i+1}^t(x_{i+\frac{1}{2}})$ (see [34] for details). Finally this system of ordinary differential equations will be solved with a high-order solver such as the TVD Runge-Kutta schemes (see [88] and [148]).

Within the path-conservative methods we will focus on those that have the property of being well-balanced as we said before. We will see different techniques that will allow us to obtain methods with this property and that will be based on a good choice of the family of paths. In the case of considering the semi-discrete problem (I. 36), given a stationary solution $W^*$, we will use the following terminology:

- The numerical method (I. 36) is said to be well-balanced for $W^*$, if the vector of cell averages of $W^*$ is an equilibrium of the ODE system (I. 36).
- The reconstruction operator is said to be well-balanced for $W^*$ if

$$\mathbb{P}_i(x) = W^*(x), \qquad x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \tag{I. 37}$$

  where $\mathbb{P}_i$ is the approximation of $W^*$ obtained by applying the reconstruction operator to the vector of cell averages of $W^*$.

As we will see in Chapter 1 it can be proven that combining a well-balanced first-order path-conservative method and a well-balanced high-order operator we can obtain high-order schemes with that property. Although standard high-order reconstruction operators are not in general well-balanced, a technique was introduced in [40] to make them have this property: let us consider a reconstruction operator that will be denoted by

$$\mathbb{Q}_i(x) = \mathbb{Q}_i(x; \{W_j\}_{j\in S_i}),$$

where $S_i$ is a set of indices corresponding to the neighbouring cells of the $i$-th one, whose values are used to compute the reconstruction in that cell. The following steps have to be performed in order to compute its well-balanced modification $\mathbb{P}_i$ at the cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ for a given family of cell values $\{W_i\}$:

1. Look for the stationary solution $W_i^*(x)$ defined in the stencil of cell $I_i$ $(\cup_{j\in\mathcal{S}_i} I_j)$ such that:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}} W_i^*(x)dx = W_i, \tag{I. 38}$$

   if possible. In other cases consider $W_i^* \equiv 0$.

2. Compute the fluctuations $\{V_j\}_{j\in S_i}$ within the stencil $S_i$:

$$V_j = W_j - \frac{1}{\Delta r}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} W_i^*(x)dx, \quad j \in S_i. \tag{I. 39}$$

3. Apply the standard reconstruction operator to the fluctuations $\{V_j\}_{j\in S_i}$:

$$\mathbb{Q}_i(x) = \mathbb{Q}_i(x; \{V_j\}_{j\in S_i}).$$

4. Compute the reconstruction in the cell:

$$\mathbb{P}_i(x) = W_i^*(x) + \mathbb{Q}_i(x).$$

$\mathbb{P}_i$ is well-balanced for every stationary solution provided that the reconstruction operator $\mathbb{Q}_i$ is exact for the null function. Moreover, it is conservative, i.e.,

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbb{P}_i(x)dx = W_i, \quad \text{for all } i,$$

if $\mathbb{Q}_i$ is conservative, and it is high-order accurate for steady solutions that are smooth (see [47] for details).

This strategy, based on the concept of well-balanced high-order operators, has been applied to different models: shallow water systems [34, 43], blood flows in vessels [127], Euler system with or without gravity [78, 97], Ripa model [146], among others.

## Outline and objectives of the thesis

Throughout the chapters of this thesis, four different problems related to the numerical analysis of hyperbolic systems of nonlinear partial differential equations are addressed: the uniqueness of solution, the computational effort of numerical methods, the well-balanced property and the control of numerical viscosity. These problems are related to some of the lines of research of the EDANYA group [71] and, more specifically, to the numerical resolution of mathematical models of fluid mechanics in applications related to shallow water flows and gas dynamics in the context of classical and relativistic mechanics.

### Chapter 1: Preliminaries

This chapter is devoted to give the main definitions and results concerning the analytical and numerical aspects of conservative and nonconservative hyperbolic systems of Partial Differential Equations (PDE). First we will focus on their theoretical aspects, starting with

conservation laws (I. 23). We are interested in the study of Cauchy problems associated to System (I. 23):

$$\begin{cases} U_t + F(U)_x = 0, \\ U(x,0) = U_0(x), \end{cases}$$

where $U_0(x) : \mathbb{R} \longrightarrow \Omega$ is a known function. In particular, we will consider Riemann problems, i.e., Cauchy problems whose initial condition is given by:

$$U_0(x) = \begin{cases} U_l & \text{if } x < 0, \\ U_r & \text{if } x > 0, \end{cases} \tag{I. 40}$$

where $U_l, U_r \in \Omega$. As we mentioned before, the Cauchy problems may not have a classical solution even when the initial condition is smooth. In this chapter we will see how to extend the concept of classical solution when discontinuities appear introducing the concept of *weak solution*. As we also said previously, these kind of solutions have the problem of non-uniqueness, so an *entropy* criterion will be also introduced, which will allow us to select the solutions with physical meaning. Once the entropy criterion has been selected, we will see that the solutions of the Riemann problems associated with the systems of conservation laws are composed of *simple waves*, i.e: *rarefactions*, *shocks* and *contact discontinuities*.

All these concepts and results will be generalized to systems of the form:

$$W_t + \mathcal{A}(W)W_x = 0, \tag{I. 41}$$

which include systems of balanced laws (I. 24) and nonconservative systems (I. 25). The theory developed by Dal Maso, LeFloch and Murat [57], based on the choice of a family of Lipschitz continuous paths, will allow us to extend the concept of weak solution and entropy to systems of the form (I. 41).

After the theoretical study, we will focus on the numerical aspects and, in particular, on the design of finite volume methods. Within these we will study the *path-conservative* methods introduced by Parés in [132], which are a generalization of the notion of conservative methods for systems of conservation laws (I. 23). These methods, based on the definition of weak solution given by Dal Maso et al., will allow us to extend the following methods: Godunov, [86], Roe [93], Polynomial Viscosity matrix [140] and Simple Riemann solvers [93] to systems of the form (I. 41). These methods, combined with reconstruction operators, will be the basis for the development of high-order shock-capturing methods, that is, methods that are high-order accurate in regions where the solution is regular and that capture correctly the generation and propagation of discontinuities. The reconstruction operators MUSCL [158] and CWENO [117], used in the remaining chapters, will be recalled. After this the concept of *well-balanced* method will be introduced and we will see

under which conditions the above mentioned methods fulfill this property. By combining first-order well-balanced methods and the well-balanced modification of the reconstruction operator, high-order schemes that also have the property of preserving the stationary solutions will be obtained. Finally, the convergence of finite volume methods to the selected weak solution will be briefly discussed, concluding that we must ask them for consistency, stability, entropy control and control of small-scale effects, including the numerical viscosity.

**Chapter 2: The Riemann problem for the shallow water equations with topography: the wet-dry case**

In this chapter we will study in more detail the shallow water system (I. 22). In the case in which the bottom $a$ is discontinuous, the source term of the shallow water equations is a nonconservative product, that is why the difficulties mentioned above appear. In the chapter we focus on the Riemann problems in a step-shaped bottom, that is, our initial condition will be of the form:

$$(h,u,a)(x,0) = \begin{cases} (h_l,u_l,a_l), & x<0, \\ (h_r,u_r,a_r), & x>0, \end{cases} \tag{I. 42}$$

with $a_l \neq a_r$. The case in which $a_l = a_r$ is already totally solved (see [153]) and, in the case in which $a_l \neq a_r$, different analytical and numerical studies of the solutions of the Riemann problems have been carried out: [3, 14, 141, 113, 20, 16, 21, 91], etc. In this chapter we will continue the study carried out in [112] but focusing on the case that was not studied there, which corresponds to the situation in which there is no water on one of the two sides of the step. In [41] this problem was also considered but only in cases where the step acts as an obstacle for the fluid, i.e. the fluid is not able to overcome the step from one side to the other, we will cover all possible situations including this one. This problem is fundamental for the design of schemes that treat dry-wet fronts well.

Following [112], we will begin by looking at the eigenvalues and eigenvectors of the system so we can find those areas in which the system is strictly hyperbolic or only hyperbolic. We will continue to see the structure of the simple waves, describing in detail the rarefactions, shocks and contact discontinuities. Like in [112] a monotonicity criterion (MC) is imposed to select the admissible stationary discontinuities over the step. After this we will build the solutions to the different Riemann problems by combining simple waves. It will be seen that, depending on the initial conditions, the problem has 0, 1 or 2 solutions. In those cases in which there is no solution, following [41], the problem will be reinterpreted as a partial Riemann problem what will allow us to construct a solution consistent with the physics of the problem. Finally several numerical tests will be considered to compare the approximation provided by different numerical methods. It will be seen that, in the cases where the solution is not unique, different methods may

converge to different solutions. In particular, we will see that the best results are given by the numerical scheme that combines the Godunov flux with the generalized hydrostatic reconstruction technique recalled in Chapter 1. The content of this chapter was published by Parés and Pimentel-García in 2019 in *Journal of Computational Physics*, see [134].

**Chapter 3: On the efficient implementation of PVM methods and simple Riemann solvers. Application to the Roe method for large hyperbolic systems**

The objective of this chapter is to implement efficiently the Roe method when we have a large number of equations in the considered system. The standard implementation of this method is based on the computation of the absolute value of the Roe matrix, which requires the explicit knowledge of its spectrum. If it is necessary to compute them numerically (because its analytical expression is not known or it is computationally expensive), the computational cost can be high, especially if the size of the matrix is big. In these cases it may be preferable to use some other methods, such as *Polynomial Viscosity matrix (PVM)* and *Approximate Riemann Solvers (ARS)*.

In the case of PVMs methods, the absolute value of the Roe matrix is approximated by its evaluation in a given polynomial that approximates the absolute value function. We will consider here polynomials that interpolate the absolute value function at some points in the Lagrange or Hermite sense.

The Approximate Riemann solvers are based on approximations of the solutions of the Riemann problems associated to the intercells. A particular case of these, which will be the ones considered in this chapter, are the Simple Riemann Solvers (SRS), whose approximated solutions consist on several constant states joined by discontinuities that travel at constant speed.

We will see that the SRS and the PVM methods, introduced in Chapter 1, are closely related. In particular, we will see that, under certain hypotheses, SRS can be interpreted as PVMs and vice versa. This relationship was already seen in [125] and in this chapter it will be reviewed with simplified proofs. We will see then that Roe method can be interpreted as a PVM based on the polynomial that interpolates the absolute value of all its eigenvalues, as already seen in [35]. This will allow us to implement it using the Newton form of the interpolation polynomial, which is the most efficient one.

Finally we will verify that, indeed, this implementation of Roe method allows us to reduce the computational times. In order to verify this we will use the bilayer shallow water model [44] and the *Quadrature Based Moment Equations (QBME)* [102]. It will be seen that the computational cost reduction grows with the number of equations. The content of this chapter was published by Pimentel-García, Parés, Castro and Koellermeier

in 2021 in the journal *Applied Mathematics and Computations*, see [136].

**Chapter 4: Well-balanced algorithms for relativistic fluids on a Schwarzschild background**

This chapter focuses on fluid simulations in a relativistic context based on the Schwarzchild metric. This metric is an exact solution of the Einstein equations of the gravitational field that describes the field generated by a star or a spherical mass. In particular, we focus on the numerical study of the long time behavior of a flow evolving around a Schwarzschild black hole. The flow is assumed to enjoy spherical symmetry so that only models with one spatial coordinate (the distance to the center of the black hole) are considered throughout. Specifically, we will introduce the relativistic Burgers-Schwarzschild and Euler-Schwarzschild models. In order to be able of running reliable and accurate numerical simulations of the flow behavior, shock-capturing high-order well-balanced finite volume methods will be designed for both systems. We build upon earlier investigations on this problem by LeFloch et al [62, 108, 114] and extend to the present problem the well-balanced methodology in Castro and Parés [47] for general systems of balance laws. Throughout this methodology we will give the necessary guidelines to build schemes of arbitrary order of accuracy, but we will focus on the first, second and third order.

In the case of the relativistic Burgers model, it is possible to compute its stationary solutions explicitly, which will allow us to easily solve the problem (I. 38) that appears in the first step of the well-balanced reconstruction procedure. We will use the schemes to perform a systematic study of the asymptotic behavior of the solutions when time tends to infinity. It will be seen that the well-balanced property is fundamental to study the evolution of the perturbations on a stationary solution.

In the case of the relativistic Euler equations, the implicit expression of the stationary solutions is available, so that a numerical method is necessary to evaluate them at a point: we will use the Newton's method. After obtaining the well-balanced first and second-order schemes, we will perform a new systematic study of the asymptotic behavior of the solutions.

The content of this chapter is available in the *arXiv* repository and has been submitted for publication to the *Journal of Scientific Computing*, see [110], and it is currently in the modification phase following the first reports from the reviewers.

**Chapter 5: In-cell Discontinuous Reconstruction path-conservative methods for non conservative hyperbolic systems: Second-order extension**

In this chapter we will address the problem of the convergence of path-conservative methods. As we mentioned before, in the case of nonconservative systems of the form (I. 41), consistency, stability and entropy control are not enough to ensure the convergence

to the selected weak solutions. This happens because the numerical viscosity of the methods alters the jump condition that the limits of the approximate solutions satisfy: instead of being those associated with the chosen family of paths, they are related to the viscous profiles of the equivalent equation (see [42], [1]). In [51] a technique based on discontinuous reconstructions inside the cell allowed the author to design first-order path-conservative methods that correctly capture solutions with isolated shocks. The main objective of this chapter is to extend this technique to second-order accuracy and to set the basis for its generalization to arbitrary order. In order to do this we will use the semi-discrete form of the path-conservative methods to which the MUSCL-Hancock reconstruction [161] will be applied: the *minmod* linear reconstruction operator (see [158]) will be used in space and time. We will state and prove that this extension still maintains the property of exactly capturing isolated shocks. In order to validate the numerical method, it will be applied to the nonconservative systems: Coupled-Burgers equations [44], Gas dynamics equations in Lagrangian coordinates [1] and the modified shallow water equations [42]. We will see that the proposed numerical method captures correctly isolated shocks and improves the results obtained with the method introduced in [51].

**Chapter 6: Conclusions and future work**

In this chapter the main contributions of this thesis are summarized and the possible future lines of research are discussed. Let us highlight the main novelties of each chapter in this thesis.

- **Chapter 2**:
  - Wet-dry Riemann problems for the shallow water system are considered.
  - 0,1, or 2 solutions are found depending on the wet state.
  - Solutions based on a reinterpretation are proposed in the case of 0 solutions.
  - This analysis is useful to better understand the behaviour of the numerical methods.
  - The correct simulation of wet-dry fronts is crucial in applications.
- **Chapter 3**:
  - An efficient implementation of PVM methods is proposed.
  - The relation between PVM methods and simple Riemann solvers is revisited.
  - A new implementation of the Roe method is proposed, named as Newton Roe method.
  - Newton Roe method is more efficient then the standard one for large hyperbolic systems.

  - Numerical test with the two-layer shallow water system and the QBME model are shown.

- **Chapter 4**:

  - A well-balanced methodology is extended to relativistic models.
  - First-, second- and third-order well-balanced methods are considered for the Burgers-Schwarzschild equation.
  - First- and second-order well-balanced methods are considered for the Euler-Schwarzschild system.
  - The well-balanced property is mandatory when dealing with these systems.
  - We perform a systematic study of the asymptotic behaviour of the solutions for reaching conclusions about the long-time convergence.

- **Chapter 5**:

  - In-cell discontinuous reconstruction overcomes convergence problem for isolated shocks.
  - The in-cell discontinuous reconstruction technique is extended to second-order.
  - The ideas are easily extended for obtaining arbitrary order of accuracy.
  - The basis for capturing exactly more than one shock is posed.

# Chapter 1

# Preliminaries

In this chapter the main definitions and results concerning the analytical and numerical aspects of conservative and nonconservative hyperbolic systems of Partial Differential Equations (PDE) are reviewed. We will focus on the difficulties that arise from the presence of nonconservative products and source terms in these systems and how we can deal with them numerically in the framework of finite volume path-conservative methods.

## 1.1 Conservative and nonconservative hyperbolic systems

The general form of a system of conservation laws in one dimension is

$$U_t + F(U)_x = 0, \tag{1.1.1}$$

where the unknown $U(x,t) = (u_1(x,t), ..., u_N(x,t))^T$ takes values in $\Omega$, being an open convex set of $\mathbb{R}^N$ called set of states and $F$ is a regular function from $\Omega$ to $\mathbb{R}^N$ called flux-function. Several flow models in physics can be written in this form. System (1.1.1) can be written in the quasi-linear form

$$U_t + A(U)U_x = 0, \tag{1.1.2}$$

where

$$A(U) = JF(U) = \begin{pmatrix} \frac{\partial F_1}{\partial u_1}(U) & \dots & \frac{\partial F_1}{\partial u_N}(U) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_N}{\partial u_1}(U) & \dots & \frac{\partial F_N}{\partial u_N}(U) \end{pmatrix} \tag{1.1.3}$$

is the Jacobian matrix of $F$. We focus on hyperbolic systems that are those satisfying that, for any $U \in \Omega$, the matrix $A(U)$ has $N$ real eigenvalues $\lambda_1(U) \leq ... \leq \lambda_N(U)$ called characteristic speeds and $N$ linearly independent corresponding eigenvectors $R_1(U), ..., R_N(U)$ that generate the so-called characteristic fields. If all the eigenvalues are

different the system is said to be strictly hyperbolic. We give now some definitions and results related to these systems.

**Definition 1.1.1.** *The characteristic field $R_i(U)$ is said to be linearly degenerated if*

$$\nabla\lambda_i(U)\cdot R_i(U)=0,\quad \forall U\in\Omega, \tag{1.1.4}$$

*where*

$$\nabla\lambda_i(U)=\begin{pmatrix}\frac{\partial\lambda_i}{\partial u_1}\\ \vdots\\ \frac{\partial\lambda_i}{\partial u_N}\end{pmatrix}.$$

**Definition 1.1.2.** *The characteristic field $R_i(U)$ is said to be genuinely nonlinear if*

$$\nabla\lambda_i(U)\cdot R_i(U)\neq 0,\quad \forall U\in\Omega. \tag{1.1.5}$$

**Definition 1.1.3.** *A regular function $w:\Omega\longrightarrow\mathbb{R}$ is said to be a Riemann invariant associated to the eigenvalue $\lambda_i(U)$ if it satisfies*

$$\nabla w(U)\cdot R_i(U)=0,\quad \forall U\in\Omega. \tag{1.1.6}$$

We are interested in the study of the Cauchy problems associated with System (1.1.1): Find a function $U:(x,t)\in\mathbb{R}\times\mathbb{R}^+\longrightarrow U(x,t)\in\Omega$ satisfying (1.1.1) and the initial condition

$$U(x,0)=U_0(x) \tag{1.1.7}$$

where $U_0(x):\mathbb{R}\longrightarrow\Omega$ is a known function. In particular we will be interested in the Riemann problems that are those Cauchy problems whose initial condition is given by

$$U_0(x)=\begin{cases}U_l & \text{if } x<0,\\ U_r & \text{if } x>0,\end{cases} \tag{1.1.8}$$

where $U_l,U_r\in\Omega$.

**Definition 1.1.4.** *A function $U:(x,t)\in\mathbb{R}\times\mathbb{R}^+\longrightarrow U(x,t)\in\Omega$ is a classical solution of the Cauchy problem (1.1.7) if $U\in C^1$ and it satisfies (1.1.1) and the initial condition.*

In general Cauchy problems may not have a classical solution even if $U_0$ is smooth, so that the concept of solution has to be extended to overcome this problem. Let us denote by $\mathcal{L}^\infty_{loc}$ the space of locally bounded measurable functions and by $C^1_0(\mathbb{R}\times\mathbb{R}^+)$ the space of $C^1$ functions $\phi$ with compact support in $\mathbb{R}\times\mathbb{R}^+$.

**Definition 1.1.5.** *Let us consider $U_0 \in \mathcal{L}^\infty_{loc}(\mathbb{R})$. A function $U \in \mathcal{L}^\infty_{loc}(\mathbb{R}\times\mathbb{R}^+)$ is said to be a weak solution of the Cauchy problem (1.1.7) if $U(x,t) \in \Omega$ almost everywhere and it satisfies*

$$\int_0^\infty \int_{\mathbb{R}} \Big( U(x,t)\cdot\frac{\partial \phi}{\partial t}(x,t) + F(U(x,t))\cdot\frac{\partial \phi}{\partial x}(x,t)\Big) dxdt = \int_{\mathbb{R}} U_0(x)\cdot\phi(x,0)dx, \quad \forall \phi \in C^1_0(\mathbb{R}\times\mathbb{R}^+). \tag{1.1.9}$$

This definition can be also written in the following equivalent way.

**Definition 1.1.6.** *A function $U : \mathbb{R}\times\mathbb{R}^+ \longrightarrow \mathbb{R}$ is said to be a weak solution of the Cauchy problem (1.1.7) if it satisfies:*

$$\int_a^b U(x,t_1)dx = \int_a^b U(x,t_0)dx + \int_{t_0}^{t_1} F(U(a,t))dt - \int_{t_0}^{t_1} F(U(b,t))dt, \tag{1.1.10}$$

*for every space-time rectangle $[a,b]\times[t_0,t_1]$.*

The weak solutions satisfy System (1.1.1) in the sense of distributions what allows us to consider solutions that are discontinuous. More precisely, we consider piecewise $C^1$ solutions according to the following definition:

**Definition 1.1.7.** *A function $U : \mathbb{R}\times\mathbb{R}^+ \longrightarrow \mathbb{R}$ is said to be piecewise $C^1$ if there exists a finite number of smooth curves $\Gamma_1, ..., \Gamma_s$ outside of which $U$ is a $C^1$ function and across which $U$ has a jump discontinuity.*

These solutions can be characterized by the so-called Rankine-Hugoniot conditions satisfied at the jumps:

**Theorem 1.1.1.** *Let us consider a piecewise $C^1$ function $U : \mathbb{R}\times\mathbb{R}^+ \longrightarrow \Omega$. We have that $U$ is a weak solution of the Cauchy problem (1.1.1)-(1.1.7) in $\mathbb{R}\times\mathbb{R}^+$ if and only if $U$ is a classical solution of (1.1.1)-(1.1.7) where $U$ is $C^1$ and it satisfies along all the discontinuities the jump conditions:*

$$\sigma[U] = [F(U)], \tag{1.1.11}$$

*where $\sigma$ is the speed of propagation of the discontinuity and*

$$[U] = U_+ - U_-, \quad [F(U)] = F(U_+) - F(U_-),$$

*being $U_+(x,t)$ and $U_-(x,t)$ the right and the left states of the discontinuity, respectively.*

See [85] for the proof. In general weak solutions are not unique and an entropy criterion is necessary to select the physically meaningful ones.

**Definition 1.1.8.** *We define an entropy pair* $(\mathcal{U}, \mathcal{F})$ *for System (1.1.1) as the pair composed by a smooth and convex function* $\mathcal{U} : \Omega \longrightarrow \mathbb{R}$*, called entropy, and a smooth function* $\mathcal{F} : \Omega \longrightarrow \mathbb{R}$*, called entropy flux, such that*

$$(\nabla \mathcal{U}(U))^T A(U) = (\nabla \mathcal{F}(U))^T, \quad \forall U \in \Omega. \tag{1.1.12}$$

Given an entropy pair $(\mathcal{U}, \mathcal{F})$, it can be easily checked that classical solutions satisfy the conservation law

$$\mathcal{U}(U)_t + \mathcal{F}(U)_x = 0. \tag{1.1.13}$$

In the case of discontinuous solutions the entropy pair is used to select the admissible solutions as follows:

**Definition 1.1.9.** *Given an entropy pair* $(\mathcal{U}, \mathcal{F})$*, a weak solution of System* (1.1.1) *is called entropy solution if it satisfies*

$$\mathcal{U}(U)_t + \mathcal{F}(U)_x \leq 0, \tag{1.1.14}$$

*in the distribution sense.*

**Theorem 1.1.2.** *Let us consider a piecewise* $C^1$ *function* $U$ *that is a weak solution of System* (1.1.1)*. We have that* $U$ *is an entropy solution if and only if*

$$\sigma(\mathcal{U}(U_+) - \mathcal{U}(U_-) \geq \mathcal{F}(U_+) - \mathcal{F}(U_-)), \tag{1.1.15}$$

*in the discontinuities.*

See [85] for the proof. An alternative to this criterion based on the entropy pairs is the so-called Lax entropy conditions:

**Definition 1.1.10.** *Let us consider a piecewise* $C^1$ *function* $U$ *that is a weak solution of System* (1.1.1)*. It is said that the discontinuity satisfies the Lax entropy conditions if there exists an index* $k \in \{1, ..., N\}$ *such that we have either*

$$\begin{cases} \lambda_k(U_+) < \sigma < \lambda_{k+1}(U_+), \\ \lambda_{k-1}(U_-) < \sigma < \lambda_k(U_-), \end{cases} \tag{1.1.16}$$

*if the* $k$*th characteristic field is genuinely nonlinear, or*

$$\lambda_k(U_-) = \sigma = \lambda_k(U_+), \tag{1.1.17}$$

*if the* $k$*th characteristic field is linearly degenerate.*

In the case of conservation laws, as we have seen, solutions are usually defined in the sense of distributions and, under the validity of an entropy inequality, existence and uniqueness results are proved for initial data close to a constant state (see for instance [119], [120], [83], [104], [105], [106]).

Some basic solutions called simple waves take an important role on the structure of the solution of Riemann problems (1.1.1)-(1.1.8). They are self-similar solutions, i.e. solutions of the form

$$U(x,t) = V(x/t),$$

where $V : \mathbb{R} \longrightarrow \mathbb{R}$ is a $C^1$ function. These solutions are the following:

- Rarefaction waves, which are solutions of System (1.1.1) associated with the genuinely nonlinear fields $\lambda_i(U)$ whose form is:

$$U(x,t) = \begin{cases} U_l & \text{if } x/t < \lambda_i(U_l), \\ V(x/t) & \text{if } \lambda_i(U_l) < x/t < \lambda_i(U_r), \\ U_r & \text{if } x > \lambda_i(U_r), \end{cases} \tag{1.1.18}$$

  being $V$ a $C^1$ function satisfying

$$A(V(\xi))V'(\xi) = \xi V'(\xi), \quad \forall \xi \in \mathbb{R}.$$

  These solutions verify:

  - The Riemann invariants are constant through them.
  - Divergence of characteristics: if such a solution can be defined, necessarily

$$\lambda_i(U_l) < \lambda_i(U_r). \tag{1.1.19}$$

- Shock waves, which are discontinuous solutions of (1.1.1) associated with the genuinely nonlinear fields $\lambda_i(U)$ whose form is:

$$U(x,t) = \begin{cases} U_l & \text{if } x < \sigma t, \\ U_r & \text{if } x > \sigma t. \end{cases} \tag{1.1.20}$$

  These solutions verify:

  - Rankine-Hugoniot conditions (1.1.11).
  - Entropy conditions (1.1.14).

- Contact discontinuities, which are discontinuous solutions of System (1.1.1) associated with the linear degenerated fields $\lambda_i(U)$ of the form:

$$U(x,t) = \begin{cases} U_l & \text{if } x < \sigma t, \\ U_r & \text{if } x > \sigma t. \end{cases} \tag{1.1.21}$$

  If such a solution can be defined, necessarily

$$\lambda_i(U_l) = \lambda_i(U_r) = \sigma. \tag{1.1.22}$$

In the case of strictly hyperbolic systems and close enough $U_l$,$U_r$, the Riemann problem (1.1.1)-(1.1.8) has an unique weak solution that is self-similar and that consists of at most $N+1$ constants states (including $U_l$ and $U_r$) connected by at most $N$ simple waves (see [85]). This unique solution of the Riemann problem (1.1.1)-(1.1.8) will be denoted by $V(x/t, U_l, U_r)$.

Let us generalize these definitions to nonconservative systems of the form

$$U_t + F(U)_x + B(U)U_x = S(U)\sigma_x, \tag{1.1.23}$$

where again the unknown $U(x,t) = (u_1(x,t), ..., u_N(x,t))^T$ takes values in an open convex set $\Omega$ of $\mathbb{R}^N$, $F$ is a regular function from $\Omega$ to $\mathbb{R}^N$, $B$ is a regular matrix function from $\Omega$ to $\mathcal{M}_{NxN}(\mathbb{R})$, $S$ is a function from $\Omega$ to $\mathbb{R}^N$, and $\sigma(x)$ is a known function from $\mathbb{R}$ to $\mathbb{R}$. The main difficulty of systems of the form (1.1.23) appears when $B \neq 0$ and it is not the Jacobian of any function or when $S \neq 0$ and $\sigma$ is discontinuous. In this cases, the corresponding terms $B(U)U_x$, $S(U)\sigma_x$ can not be defined in the sense of distributions if $U$ is discontinuous. In these cases we talk about nonconservative products.

The system (1.1.23) can be written in the form

$$W_t + \mathcal{A}(W)W_x = 0, \tag{1.1.24}$$

where

$$W = \begin{pmatrix} U \\ \sigma \end{pmatrix}, \quad \mathcal{A}(W) = \left(\begin{array}{c|c} A(U) + B(U) & -S(U) \\ \hline 0 & 0 \end{array}\right),$$

and $A(U) = JF(U)$. There are several mathematical theories that allows one to give a mathematical sense to the nonconservative product $\mathcal{A}(W)W_x$, like those described in [163] and [49]. Here the theory developed by Dal Maso, LeFloch and Murat [57] that allows a notion of weak solution which satisfies (1.1.24) in the sense of Borel measures will be followed. This definition is based on the choice of a family of Lipschitz continuous paths $\Phi : [0,1] \times \Omega \times \Omega \to \Omega$ satisfying certain regularity and consistency properties, in particular

$$\Phi(0, U_l, U_r) = U_l, \quad \Phi(1, U_l, U_r) = U_r, \quad \forall (U_l, U_r) \in \Omega \times \Omega,$$

and

$$\Phi(\xi, U, U) = U, \quad \forall \xi \in [0,1], \quad \forall U \in \Omega.$$

The family of paths can be interpreted as a tool to give sense to the integral of the nonconservative product in an interval $[a,b]$. More precisely, given a bounded variation function $W : [a,b] \longrightarrow \mathbb{R}$, we define:

$$\int_a^b \mathcal{A}(W(x))W_x(x)dx = \int_a^b \mathcal{A}(W(x)W_x(x)dx + \sum_l \int_0^1 \mathcal{A}(\Phi(\xi; W_l^-, W_l^+))\frac{\partial \Phi}{\partial \xi}(\xi; W_l^-, W_l^+)d\xi, \tag{1.1.25}$$

where $W_l^-$ and $W_l^-$ represents, respectively, the limits of $W$ to the left and the right of its $l$th discontinuity (observe that the number of discontinuities is countable because $W$ is a bounded variation function). Observe that this definition allows us to determine the weight of the Dirac measures issues of the derivative of $W$ with respect to $x$ at the discontinuities.

Once this notion of integral has been defined, it is easy to extend to nonconservative Definition 1.1.6:

**Definition 1.1.11.** *A function $W : \mathbb{R} \times \mathbb{R}^+ \longrightarrow \mathbb{R}$ is said to be a weak solution of System (1.1.24) if it satisfies*

$$\int_a^b W(x,t_1)dx = \int_a^b W(x,t_0)dx - \int_{t_0}^{t_1} \int_a^b \mathcal{A}(W(x,t))W_x(x,t)dxdt, \tag{1.1.26}$$

*for every space-time rectangle $[a,b] \times [t_0,t_1]$.*

Once a family of paths has been chosen to define weak solutions, we can characterize again the piecewise $C^1$ weak solutions by the so-called generalized Rankine-Hugoniot conditions.

**Theorem 1.1.3.** *Let us consider a piecewise $C^1$ function $W : \mathbb{R} \times \mathbb{R}^+ \to \Omega$. We have that $W$ is a weak solution of System (1.1.24) in $\mathbb{R} \times \mathbb{R}^+$ if and only if $W$ is a classical solution of (1.1.24) where $W$ is $C^1$ and it satisfies along all the discontinuities the jump conditions:*

$$\sigma(W_r - W_l) = \int_0^1 \mathcal{A}(\Phi(\xi, W_l, W_r))\frac{\partial \Phi}{\partial \xi}(\xi, W_l, W_r)d\xi \tag{1.1.27}$$

*where $\sigma$ is the speed of propagation of the discontinuity.*

If the system is conservative, i.e., if $\mathcal{A}(W)$ is the Jacobian of a function, this condition reduces to the Rankine-Hugoniot one (1.1.11) regardless of the family of paths chosen. As it happens for systems of conservation laws (1.1.1), in order to have uniqueness of solution, we need to add an entropy condition to the notion of weak solution.

**Definition 1.1.12.** *We define an entropy pair* $(\mathcal{U}, \mathcal{F})$ *for System (1.1.24) as the pair composed by a convex function* $\mathcal{U} : \Omega \longrightarrow \mathbb{R}$ *and a function* $\mathcal{F} : \Omega \longrightarrow \mathbb{R}$ *such that*

$$(\nabla \mathcal{U}(W))^T \mathcal{A}(W) = (\nabla \mathcal{F}(W))^T, \quad \forall W \in \Omega. \tag{1.1.28}$$

As it happened for the system of conservation laws, the entropy solutions of (1.1.24) should satisfy the entropy inequality

$$\mathcal{U}(W)_t + \mathcal{F}(W)_x \leq 0, \tag{1.1.29}$$

in the distribution sense.
We observe that the concept of weak solutions depends clearly on the choice of the family of paths, which is a priori arbitrary, so that the crucial question is how to choose the 'good' family of paths. In fact, when the hyperbolic system is the vanishing-viscosity limit of the parabolic problems

$$W_t^\epsilon + \mathcal{A}(W^\epsilon)\, W_x^\epsilon = \epsilon (\mathcal{R}(W^\epsilon) W_x^\epsilon)_x, \tag{1.1.30}$$

where $\mathcal{R}(W)$ is any positive-definite matrix, the adequate family of paths should be related to the viscous profiles: a function $V$ is said to be a viscous profile for (1.1.30) linking the states $W^-$ and $W^+$ if it satisfies

$$\lim_{\chi \to -\infty} V(\chi) = W^-, \quad \lim_{\chi \to +\infty} V(\chi) = W^+, \quad \lim_{\chi \to \pm\infty} V'(\chi) = 0 \tag{1.1.31}$$

and there exists $\sigma \in \mathbb{R}$ such that the travelling wave

$$W^\epsilon(x,t) = V\left(\frac{x - \sigma t}{\epsilon}\right), \tag{1.1.32}$$

is a solution of (1.1.30) for every $\epsilon$. It can be easily verified that, in order to be a viscous profile, $V$ has to solve the equation

$$-\sigma V' + \mathcal{A}(V)\, V' = (\mathcal{R}(V)\, V')', \tag{1.1.33}$$

with boundary conditions (1.1.31). If there exists a viscous profile linking the states $W^-$ and $W^+$, the good choice for the path connecting the states would be, after a reparameterization, the viscous profile $V$.

The main difference with the conservative case is that now every choice of viscous term $\mathcal{R}$ leads to different jump conditions, while for standard conservative systems the usual Rankine-Hugoniot conditions are always recovered independently of the choice of the viscous term.

Let us consider the Riemann problem:

$$\begin{cases} W_t + \mathcal{A}(W) W_x = 0, \\ W(x,0) = \begin{cases} W_l & if \quad x < 0, \\ W_r & if \quad x > 0. \end{cases} \end{cases} \tag{1.1.34}$$

Simple waves (rarefaction waves, shock waves and contact discontinuities) are defined like in the conservative case and again a solution of (1.1.34) is self-similar and it consists of at most $N+1$ constants states (including $W_l$ and $W_r$) connected by at most $N$ simple waves. We will denote it again by $V(x/t, W_l, W_r)$.

## 1.2 Numerical aspects

In this section we focus on the design of finite volume methods. Let us discretize the real line in computing cells $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ that will have constant size $\Delta x$ for simplicity. Let us also suppose that the boundaries of the cells are defined by $x_{i+\frac{1}{2}} = i\Delta x$ and we denote $x_i = (i-\frac{1}{2})\Delta x$ the center of the cell $I_i$. We will also consider $\Delta t$ as the time step and we define $t_n = n\Delta t$. Let us denote by $W_i^n$ the approximation of the cell average of the weak solution at the cell $I_i$ at time $t_n$ provided by the scheme. The initial cell values will be given by:

$$W_i^0 = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W_0(x)dx, \tag{1.2.1}$$

where $W_0(x)$ is the initial condition. Observe that according to (1.1.26), the cell averages of weak solutions satisfy:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W(x,t_{n+1})dx = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W(x,t_n)dx - \frac{\Delta t}{\Delta x}\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(W(x,t))W_x(x,t)dxdt. \tag{1.2.2}$$

The idea is then to mimic this equality at the discrete level, as it is done to define conservative methods for systems of conservation laws.

### 1.2.1 Path-conservative schemes: definition

We will focus on the so-called path-conservative methods developed by Parés [132].

**Definition 1.2.1.** *Given a family of paths $\Phi$, a numerical scheme is said to be $\Phi$-conservative if it can be written under the form:*

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\Big(\mathcal{D}^+_{i-\frac{1}{2}} + \mathcal{D}^-_{i+\frac{1}{2}}\Big), \tag{1.2.3}$$

*where*

$$\mathcal{D}^\pm_{i+\frac{1}{2}} = \mathcal{D}^\pm(W_i^n, W_{i+1}^n), \tag{1.2.4}$$

*being $\mathcal{D}^-$ and $\mathcal{D}^+$ two continuous functions from $\Omega^2$ to $\Omega$ satisfying:*

$$\mathcal{D}^\pm(W,W) = 0, \quad \forall W \in \Omega, \tag{1.2.5}$$

*and*

$$\mathcal{D}^-(W_l, W_r) + \mathcal{D}^+(W_l, W_r) = \int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_r)) \frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_r) d\xi, \tag{1.2.6}$$

*for every set* $\{W_l, W_r\} \subset \Omega$.

We observe that this definition is the discrete counterpart of (1.1.26) with $[a, b] = I_i$ and $[t_0, t_1] = [t_n, t_{n+1}]$. As we are considering a piecewise constant approximation of the solution, the dashed integral only consists of two Dirac measures placed in the intercells, whose masses are split in the two terms $\mathcal{D}^{\pm}_{i+\frac{1}{2}}$, one of them contributing to the cell $I_i$ and the other one to the cell $I_{i+1}$. This definition is a generalization of conservative methods for systems of conservation laws, i.e., if we suppose that $\mathcal{A}(W)$ is the Jacobian matrix of a flux $F(W)$, then every method that is path-conservative for some family of paths, can be rewritten as a conservative method as it was shown in [34]. Effectively, using (1.2.6) we obtain

$$\mathcal{D}^-(W_l, W_r) + \mathcal{D}^+(W_l, W_r) = F(W_r) - F(W_l).$$

Defining

$$\mathcal{F}(W_l, W_r) = \mathcal{D}^-(W_l, W_r) + F(W_l), \tag{1.2.7}$$

or, equivalently

$$\mathcal{F}(W_l, W_r) = F(W_r) - \mathcal{D}^+(W_l, W_r), \tag{1.2.8}$$

we obtain using (1.2.5)

$$\mathcal{F}(W, W) = F(W), \tag{1.2.9}$$

so $\mathcal{F}$ is numerical flux consistent with $F$ and using (1.2.7) and (1.2.8) in (1.2.3) we can write (1.2.3) equivalently in conservative form

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\Delta x}(\mathcal{F}_{i-\frac{1}{2}} - \mathcal{F}_{i+\frac{1}{2}}), \tag{1.2.10}$$

where

$$\mathcal{F}_{i+\frac{1}{2}} = \mathcal{F}(W_i^n, W_{i+1}^n).$$

Due to this, if the nonconservative system (1.1.24) contains some equations that are conservation laws, then a path-conservative method will be conservative for these laws.

Path-conservative schemes have been applied to many flow models, such as the shallow water or multilayer shallow water models, Saint Venant-Exner [36], turbidity currents [126], Ripa model [146], two-mode shallow water system [33], Baer-Nunziato model [66], Pitman-Le model [135], Savage-Hutter model [75], Bingham shallow water system [76], blood flow [127], two-phase flows [128], etc.

### 1.2.2 Path-conservative schemes: examples

Let us see that using the theory developed by Parés [132] we are able to extend to nonconservative systems some standard schemes used for solving systems of conservation laws.

#### 1.2.2.1 Godunov scheme

Let us suppose that a family of paths $\Phi$ has been chosen. Let us denote by $V(x/t, W_l, W_r)$ the self-similar solution of the Riemann problem:

$$\begin{cases} W_t + \mathcal{A}(W)W_x = 0, \\ W(x,0) = \begin{cases} W_l & if \quad x < 0, \\ W_r & if \quad x > 0, \end{cases} \end{cases} \tag{1.2.11}$$

The Godunov method consists in updating the approximation at the cells by averaging the solution of the Riemann problems associated to every intercell at the previous time step:

$$W_i^{n+1} = \frac{1}{\Delta x}\left(\int_{x_{i-\frac{1}{2}}}^{x_i} V\Big(\frac{x - x_{i-\frac{1}{2}}}{\Delta t}; W_{i-1}^n, W_i^n\Big)dx + \int_{x_i}^{x_{i+\frac{1}{2}}} V\Big(\frac{x - x_{i+\frac{1}{2}}}{\Delta t}; W_i^n, W_{i+1}^n\Big)dx\right), \tag{1.2.12}$$

under a CFL-1/2 condition, i.e.,

$$\Delta t \leq \frac{1}{2}\frac{\Delta x}{\sigma_{max}^n},$$

where $\sigma_{max}^n$ is the maximum of all speeds in absolute value of the waves present at time $t_n$, so that the solutions of the Riemann problem at two consecutive intercells do not interact in the averaging step.

In [129] it was shown that, under some assumptions over the family of paths, this method can be written in the path-conservative form (1.2.3) with

$$\mathcal{D}^-(W_l, W_r) = \int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_0^-)\frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_0^-)d\xi, \tag{1.2.13}$$

$$\mathcal{D}^+(W_l, W_r) = \int_0^1 \mathcal{A}(\Phi(\xi; W_0^+, W_r)\frac{\partial \Phi}{\partial \xi}(\xi; W_0^+, W_r)d\xi, \tag{1.2.14}$$

where

$$W_0^{\pm} = \lim_{s \to 0^{\pm}} V(s; W_l, W_r). \tag{1.2.15}$$

#### 1.2.2.2 Roe methods

Roe methods are based on the concept of a generalized Roe matrix in the sense of Toumi [157].

**Definition 1.2.2.** *Given a family of paths $\Phi$, a function $\mathcal{A}_\Phi : \Omega \times \Omega \rightarrowtail \mathcal{M}_{N\times N}(\mathbb{R})$ is called a Roe linearization if it verifies the following properties:*

- *For any $W_l, W_l \in \Omega$, $\mathcal{A}_\Phi(W_l, W_r)$ has $N$ distinct real eigenvalues.*
- *For any $W \in \Omega$, $\mathcal{A}_\Phi(W, W) = \mathcal{A}(W)$.*
- *For any $W_l, W_r \in \Omega$,*

$$\mathcal{A}_\Phi(W_l, W_r)(W_r - W_l) = \int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_r))\frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_r)d\xi. \tag{1.2.16}$$

**Remark 1.2.1.** *If the system is conservative, i.e. if $\mathcal{A}(W)$ is the Jacobian of a flux function $F$,* (1.2.16) *reduces to the usual Roe property*

$$\mathcal{A}_\Phi(W_l, W_r)(W_r - W_l) = F(W_r) - F(W_l),$$

*and thus the usual notion of Roe matrix is recovered.*

Once the linearization has been chosen, the corresponding Roe scheme can be written in the form (1.2.3) with

$$\mathcal{D}^-(W_l, W_r) = \mathcal{A}^-_\Phi(W_l, W_r)(W_r - W_l), \tag{1.2.17}$$

$$\mathcal{D}^+(W_l, W_r) = \mathcal{A}^+_\Phi(W_l, W_r)(W_r - W_l), \tag{1.2.18}$$

where

$$\mathcal{A}^\pm_\Phi(W_l, W_r) = R_\Phi(W_l, W_r)\Lambda^\pm_\Phi(W_r, W, l)R^{-1}_\Phi(W_r, W_l), \tag{1.2.19}$$

being $\Lambda^\pm_\Phi(W_r, W, l)$ the diagonal matrix whose coefficients are the positive and negative part, respectively, of the eigenvalues $\lambda_i(W_l, W_r)$, $i = 1, ..., N$, of the Roe matrix, and $R_\Phi(W_l, W_r)$ a $N \times N$ matrix whose columns are associated eigenvectors. Let us consider $a \in \mathbb{R}$, the positive and negative part of $a$ is, respectively:

$$a^+ = \max(a, 0) \quad a^- = \min(a, 0). \tag{1.2.20}$$

With this definition we have that

$$a = a^+ + a^-, \quad |a| = a^+ - a^-,$$

so

$$a^+ = \frac{1}{2}(a + |a|), \quad a^- = \frac{1}{2}(a - |a|).$$

Therefore, the following identity holds:

$$\mathcal{A}^{\pm}_{\Phi}(W_l, W_r) = \frac{1}{2}(\mathcal{A}_{\Phi}(W_l, W_r) \pm |\mathcal{A}_{\Phi}(W_l, W_r)|), \tag{1.2.21}$$

where

$$|\mathcal{A}^{\pm}_{\Phi}(W_l, W_r)| = R_{\Phi}(W_l, W_r)|\Lambda^{\pm}_{\Phi}(W_r, W, l)|R^{-1}_{\Phi}(W_r, W_l), \tag{1.2.22}$$

being $|\Lambda^{\pm}_{\Phi}(W_r, W, l)|$ the diagonal matrix whose coefficients are the absolute value of the eigenvalues of $\mathcal{A}_{\Phi}(W_l, W_r)$. Using (1.2.21), we can rewrite (1.2.17) and (1.2.18) as:

$$\mathcal{D}^{\pm}(W_l, W_r) = \frac{1}{2}\mathcal{A}_{\Phi}(W_l, W_r)(W_r - W_l) \pm \frac{1}{2}|\mathcal{A}_{\Phi}(W_l, W_r)|(W_r - W_l). \tag{1.2.23}$$

#### 1.2.2.3 Polinomial Viscosity matrix (PVM) methods

The Polinomial Viscosity matrix (PVM) methods were introduced in [35] as a generalization of the ones in [58]. These methods are based on a Roe linearization $\mathcal{A}_{\Phi}$ and on a polynomial chosen at every inter-cell:

$$p_r^{i+\frac{1}{2}}(x) = \sum_{j=0}^{r} \alpha_j^{i+\frac{1}{2}} x^j. \tag{1.2.24}$$

**Definition 1.2.3.** *The PVM method corresponding to the Roe linearization $\mathcal{A}_{\Phi}$ and the polynomials $p_r^{i+\frac{1}{2}}$ is the numerical scheme that writes in the path-conservative form (1.2.3) with:*

$$\mathcal{D}^{\pm}_{i+\frac{1}{2}} = \mathcal{D}^{\pm}(W_i^n, W_{i+1}^n) = \frac{1}{2}\mathcal{A}_{i+\frac{1}{2}}(W_{i+1}^n - W_i^n) \pm \frac{1}{2}\mathcal{Q}_{i+\frac{1}{2}}(W_{i+1}^n - W_i^n), \tag{1.2.25}$$

*with*

$$\mathcal{A}_{i+\frac{1}{2}} = \mathcal{A}_{\Phi}(W_i^n, W_{i+1}^n), \tag{1.2.26}$$

*and*

$$\mathcal{Q}_{i+\frac{1}{2}} = p_r^{i+\frac{1}{2}}(\mathcal{A}_{i+\frac{1}{2}}) = \sum_{j=0}^{r} \alpha_j^{i+\frac{1}{2}} \mathcal{A}^j_{i+\frac{1}{2}}. \tag{1.2.27}$$

The idea behind these methods is the following: if instead of a polynomial, the absolute value is chosen to compute the viscosity matrix, i.e.

$$\mathcal{Q}_{i+\frac{1}{2}} = \left|\mathcal{A}_{i+\frac{1}{2}}\right|, \tag{1.2.28}$$

the resulting numerical scheme is the standard Roe method (1.2.23). The idea is then to choose a polynomial $p_r^{i+\frac{1}{2}}$ that approximates the absolute value function so that the numerical scheme is expected to be close to the Roe method but computationally less expensive, since the evaluation of the Roe matrix at the polynomial may be cheaper than the computation of its absolute value (that requires the knowledge of the eigenstructure).

#### 1.2.2.4 Simple Riemann solvers (SRS)

According to [132], the generalized definition of simple Riemann solver (SRS) for (1.1.2) is as follows.

**Definition 1.2.4.** *Let us take a family of paths $\Phi$ in $\Omega$. We suppose that for every pair of states $W_l, W_r \in \Omega$, a finite number $s \geq 1$ of speeds*

$$\sigma_0 = -\infty < \sigma_1 < ... < \sigma_s < \sigma_{s+1} = +\infty, \tag{1.2.29}$$

*and $s-1$ intermediate states*

$$W_0 = W_l, W_1, ..., W_{s-1}, W_s = W_r, \tag{1.2.30}$$

*are chosen. The function $R : \mathbb{R} \times \Omega \times \Omega \rightarrowtail \Omega$ given by*

$$R(\sigma; W_l, W_r) = \begin{cases} W_0 = W_l, & if \quad \sigma < \sigma_1, \\ W_1, & if \quad \sigma_1 < \sigma < \sigma_2, \\ \vdots & \\ W_j, & if \quad \sigma_j < \sigma < \sigma_{j+1} \\ \vdots & \\ W_{s-1}, & if \quad \sigma_{s-1} < \sigma < \sigma_s, \\ W_s = W_r, & if \quad \sigma_s < \sigma, \end{cases} \tag{1.2.31}$$

*is said to be a SRS for (1.1.24) if it satisfies*

$$R(\sigma; W, W) = W, \quad \forall W \in \Omega, \tag{1.2.32}$$

*and*

$$\sum_{j=1}^{s} \sigma_j (W_j - W_{j-1}) = \int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_r)) \frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_r) d\xi. \tag{1.2.33}$$

Any SRS for (1.1.24) leads to a path-conservative numerical method:

$$W_i^{n+1} = W_i^n - \frac{\Delta x}{\Delta t}(\mathcal{D}^+(W_{i-1}^n, W_i^n) + \mathcal{D}^-(W_i^n, W_{i+1}^n)), \tag{1.2.34}$$

where

$$\mathcal{D}^-(W_l, W_r) = -\int_{-\infty}^{0} \big(R(\sigma; W_l, W_r) - W_l\big) d\sigma, \tag{1.2.35}$$

$$\mathcal{D}^+(W_l, W_r) = -\int_{0}^{\infty} \big(R(\sigma; W_l, W_r) - W_r\big) d\sigma, \tag{1.2.36}$$

or, equivalently,

$$\mathcal{D}^-(W_l,W_r)=\begin{cases}\displaystyle\int_0^1 \mathcal{A}(\Phi(\xi;W_l,W_r))\frac{\partial\Phi}{\partial\xi}(\xi;W_l,W_r)d\xi & if \quad \sigma_s<0,\\ \displaystyle\sum_{\sigma_{j+1}<0}\sigma_{j+1}(W_{j+1}-W_j) & if \quad \sigma_1<0<\sigma_s,\\ 0 & if \quad \sigma_1>0.\end{cases} \tag{1.2.37}$$

and

$$\mathcal{D}^+(W_l,W_r)=\begin{cases}0 & if \quad \sigma_s<0,\\ \displaystyle\sum_{\sigma_{j+1}>0}\sigma_{j+1}(W_{j+1}-W_j) & if \quad \sigma_1<0<\sigma_s,\\ \displaystyle\int_0^1 \mathcal{A}(\Phi(\xi;W_l,W_r))\frac{\partial\Phi}{\partial\xi}(\xi;W_l,W_r)d\xi & if \quad \sigma_1>0.\end{cases} \tag{1.2.38}$$

In particular, the easiest example of SRS is obtained if $s=2$ that corresponds to the extension to nonconvervative hyperbolic systems of the well known HLL solver introduced in [93]. In this case, the simple Riemann solver consists of two waves of speed $\sigma_1=S_l$, $\sigma_2=S_r$ linking three constants states $W_l,W^*,W_r$:

$$R^{HLL}(\sigma;W_l,W_r)=\begin{cases}W_l & if \quad \sigma<S_l,\\ W^* & if \quad S_l\le\sigma\le S_r,\\ W_r & if \quad \sigma>S_r.\end{cases} \tag{1.2.39}$$

In this case the consistency property (1.2.33) reduces to

$$S_l(W^*-W_l)+S_r(W_r-W^*)=\int_0^1 \mathcal{A}(\Phi(\xi;W_l,W_r))\frac{\partial\Phi}{\partial\xi}(\xi;W_l,W_r)d\xi,$$

so that the intermediate state is given by:

$$W^*=\frac{S_rW_r-S_lW_l-\displaystyle\int_0^1 \mathcal{A}(\Phi(\xi;W_l,W_r))\frac{\partial\Phi}{\partial\xi}(\xi;W_l,W_r)d\xi}{S_r-S_l}. \tag{1.2.40}$$

An extension of the HLL methods called HLLC methods were introduced in [155] for system of conservation laws and in [37] and [146] they were extended to hyperbolic nonconservative PDE systems arising in turbidity current or Ripa models.

### 1.2.3 Path-conservative schemes: high-order extension

High-order methods for (1.1.24) can be designed on the basis of a first-order path-conservative method (1.2.3) and a high-order reconstruction operator.

**Definition 1.2.5.** *A high-order reconstruction operator of order $p$ is an operator that, given a family of cell values $\{W_i\}$, provides at every cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ a smooth function that depends on the values at some neighbor cells whose indexes belong to the so-called stencil $S_i$:*

$$\mathbb{P}_i(x) = \mathbb{P}_i(x; \{W_j\}_{j\in S_i}),$$

*so that, if the cell values are the averages of a smooth function $W$:*

$$W_i = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W(x)dx, \quad \forall i \in \mathbb{Z} \tag{1.2.41}$$

*then*

$$W^-_{i+\frac{1}{2}} = W(x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^p), \tag{1.2.42}$$

$$W^+_{i-\frac{1}{2}} = W(x_{i-\frac{1}{2}}) + \mathcal{O}(\Delta x^p), \tag{1.2.43}$$

*being*

$$W^-_{i+\frac{1}{2}} = \mathbb{P}_i(x_{i+\frac{1}{2}}), \tag{1.2.44}$$

$$W^+_{i-\frac{1}{2}} = \mathbb{P}_i(x_{i-\frac{1}{2}}). \tag{1.2.45}$$

*The states $W^-_{i+\frac{1}{2}}$ and $W^+_{i-\frac{1}{2}}$ are called reconstructed states at the intercells.*

In general the functions $\mathbb{P}_i$ are computed by means of interpolation or approximation techniques. Some well-known examples are the ENO, WENO, CWENO, or hyperbolic reconstructions (see, for instance, [67], [68], [64], [92], [121], [148], [149], [117], [118]).

Let us denote by $W_i(t)$ the cell average of the solution $W$ of (1.1.24) over the cell $I_i$ at time $t$:

$$W_i(t) = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W(x,t)dx. \tag{1.2.46}$$

Weak solutions satisfy the equalities:

$$W_i'(t) = -\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(W(x,t))W_x(x,t)dx, \tag{1.2.47}$$

what, according to Parés [132], suggests the following form for a semidiscrete method:

$$W_i'(t) = -\frac{1}{\Delta x}\left(\mathcal{D}^+_{i-\frac{1}{2}}(t) + \mathcal{D}^-_{i+\frac{1}{2}}(t) + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i(x,t))\frac{\partial \mathbb{P}_i}{\partial x}(x,t)dx\right), \tag{1.2.48}$$

where

$$\mathcal{D}^{\pm}_{i+\frac{1}{2}}(t) = \mathcal{D}^{\pm}\left(W^{-}_{i+\frac{1}{2}}(t), W^{+}_{i+\frac{1}{2}}(t)\right), \tag{1.2.49}$$

being $\mathbb{P}_i(x,t) = \mathbb{P}_i(x, \{W_j(t)\}_{j\in S_i})$ the functions at the cells provided by the reconstruction operator and $\{W^{\pm}_{i+\frac{1}{2}}(t)\}$ the reconstructed states associated to $\{W_i(t)\}$.

**Remark 1.2.2.** *In* (1.2.48) *the reconstruction operators* $\{\mathbb{P}_i\}$ *are used to approximate the regular part of the weak integral in* (1.2.47) *and the terms* $\mathcal{D}^{\pm}_{i\pm\frac{1}{2}}$ *are used to split the Dirac measures corresponding to the discontinuities at the intercells.*

We observe that the semidiscrete (1.2.48) is a system of ordinary differential equations that we must solve by using a high-order numerical solver with good properties such as the TVD Runge-Kutta schemes from [88] and [148]: let us consider a problem of the form

$$W_i'(t) = -\frac{1}{\Delta x}\mathcal{H}(W(t)), \tag{1.2.50}$$

the first, second and third order in time TVD Runge-Kutta schemes to solve (1.2.50) are, respectively:

- First order in time:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\mathcal{H}(W_i^n). \tag{1.2.51}$$

- Second order in time:

$$\begin{cases} W_i^{n+\frac{1}{2}} = W_i^n - \frac{\Delta t}{\Delta x}\mathcal{H}(W_i^n). \\ W_i^{n+1} = \frac{1}{2}W_i^n + \frac{1}{2}\left(W_i^{n+\frac{1}{2}} - \frac{\Delta t}{\Delta x}\mathcal{H}\left(W_i^{n+\frac{1}{2}}\right)\right). \end{cases} \tag{1.2.52}$$

- Third order in time:

$$\begin{cases} W_i^{n+\frac{1}{3}} = W_i^n - \frac{\Delta t}{\Delta x}\mathcal{H}(W_i^n). \\ W_i^{n+\frac{2}{3}} = \frac{3}{4}W_i^n + \frac{1}{4}\left(W_i^{n+\frac{1}{3}} - \frac{\Delta t}{\Delta x}\mathcal{H}\left(W_i^{n+\frac{1}{3}}\right)\right). \\ W_i^{n+1} = \frac{1}{3}W_i^n + \frac{2}{3}\left(W_i^{n+\frac{2}{3}} - \frac{\Delta t}{\Delta x}\mathcal{H}\left(W_i^{n+\frac{2}{3}}\right)\right). \end{cases} \tag{1.2.53}$$

The order of accuracy in space of these methods depends on the election of the reconstruction operator. In the case of conservative systems the last integral of (1.2.48)

can be computed exactly in terms of the physical flux $F$ without depending on $\mathbb{P}_i$:

$$
\begin{aligned}
W_i'(t) &= -\frac{1}{\Delta x}\left(\mathcal{D}^+_{i-\frac{1}{2}}(t) + \mathcal{D}^-_{i+\frac{1}{2}}(t) + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i(x,t))\frac{\partial \mathbb{P}_i}{\partial x}(x,t)dx\right)\\
&= -\frac{1}{\Delta x}\left(\mathcal{D}^+_{i-\frac{1}{2}}(t) + \mathcal{D}^-_{i+\frac{1}{2}}(t) + F(\mathbb{P}(x_{i+\frac{1}{2}},t)) - F(\mathbb{P}(x_{i-\frac{1}{2}},t))\right)\\
&= -\frac{1}{\Delta x}\left(\mathcal{D}^+_{i-\frac{1}{2}}(t) + \mathcal{D}^-_{i+\frac{1}{2}}(t) + F(W^-_{i+\frac{1}{2}}(t)) - F(W^+_{i-\frac{1}{2}}(t))\right)\\
&= -\frac{1}{\Delta x}\left(F(W^+_{i-\frac{1}{2}}(t)) - \mathcal{F}_{i-\frac{1}{2}}(t) + \mathcal{F}_{i+\frac{1}{2}}(t) - F(W^-_{i+\frac{1}{2}}(t)) + F(W^-_{i+\frac{1}{2}}(t)) - F(W^+_{i-\frac{1}{2}}(t))\right) =\\
&= -\frac{1}{\Delta x}\left(\mathcal{F}_{i+\frac{1}{2}}(t) - \mathcal{F}_{i-\frac{1}{2}}(t)\right),
\end{aligned}
$$

with $\mathcal{F}_{i+\frac{1}{2}}(t) = \mathcal{F}\big(W^-_{i+\frac{1}{2}}(t), W^+_{i+\frac{1}{2}}(t)\big)$ and where (1.2.7)-(1.2.8) have been used. The order of accuracy in space in this case is then $p$ as a consequence of (1.2.42),(1.2.43). This is not the case for nonconservative systems in which the order of accuracy of the reconstruction and its derivative inside the cell $I_i$ has to be taken into account:

$$\mathbb{P}_i(x,t) = W(x,t) + \mathcal{O}(\Delta x^{p_1}), \quad \forall x \in I_i,$$

$$\frac{\partial}{\partial x}\mathbb{P}_i(x,t) = \frac{\partial}{\partial x}W(x,t) + \mathcal{O}(\Delta x^{p_2}), \quad \forall x \in I_i.$$

This is due to the fact that we can not write the integral appearing in (1.2.48) in terms of the physical flux. As a consequence, the order of accuracy in space will be $\min(p, p_1, p_2)$ and in general for the usual reconstruction techniques we have that $p_2 \leq p_1 \leq p$. Therefore, in the case of nonconservative systems the order of accuracy in space is $p_2$ while for conservative systems is $p$. This loss of accuracy has been detected and numerically verified for WENO-Roe methods in [31]. In [131] an interesting technique based on the use of the trapezoidal rule and Romberg extrapolation for the numerical approximation of the integrals in (1.2.48) was used to avoid the explicit computation of $\frac{\partial}{\partial x}\mathbb{P}_i(x,t)$ in order to increase the expected order of accuracy in space to $\min(p, p_1)$.

In [48] the strategy to extend high-order finite volume central schemes on staggered grids to general hyperbolic systems including nonconservative products was introduced and discussed. It was based on a path-conservative method on staggered cells and central Runge-Kutta time discretization. In [33] a second-order path-conservative central-upwind scheme for the two-mode shallow water system was proposed.

Another alternative for constructing high-order schemes called ADER-FV and ADER-DG methods were introduced in [152] and [69]. These two approaches consider the differential form of the governing PDE system to achieve high-order accuracy in time using

the Cauchy-Kowalevski procedure that substitutes time derivatives by space derivatives via successive differentiation of the system with respect to space and time. In [64] an entirely numerical approach that replaces the Cauchy-Kowalevski procedure by a local weak formulation of the governing PDE in space-time was presented. It was extended to nonconservative systems in [65]. This Cauchy-Kowalevski procedure is also used with finite differences in [139]. In [165] an alternative to this procedure was introduced and it was based on the approximation of the derivatives in the Taylor expansion in time of the solution through high order central divided difference formulas in a recursive way. A modification of this technique leading to numerical methods using stencils of minimal length was introduced in [29]. In both cases, WENO reconstructions were used to avoid spurious oscillations: see also [30]

Let us describe with more details the construction of the two reconstruction operators that will be used along some of the following Chapters: the MUSCL and the CWENO3 operators.

#### 1.2.3.1 MUSCL reconstruction operator

The MUSCL reconstruction operator was introduced in [158] and it is second order accurate (see [159], [160] for more details). Given a family of cell values $\{W_i\}$, the reconstruction functions are linear:

$$\mathbb{P}_i(x) = W_i + \widetilde{\partial_x W}_i(x - x_i),$$

where $\widetilde{\partial_x W}_i$ is the *minmod* approximation of the first order spacial derivative at $x_i$, whose $k$th component is given by

$$\left(\widetilde{\partial_x W}_i\right)_k = \text{minmod}\left(\alpha\frac{\mathrm{W_{i+1,k}} - \mathrm{W_{i,k}}}{\Delta \mathrm{x}}, \frac{\mathrm{W_{i+1,k}} - \mathrm{W_{i-1,k}}}{2\Delta \mathrm{x}}, \alpha\frac{\mathrm{W_{i,k}} - \mathrm{W_{i-1,k}}}{\Delta \mathrm{x}}\right), \quad (1.2.54)$$

where $W^n_{i,k}$ represents the $k$th component of $W^n_i$, $\alpha$ is a parameter with $1 \leq \alpha < 2$ and

$$\text{minmod(a, b, c)} = \begin{cases} \min\{a,b,c\} & \text{if } a,b,c > 0, \\ \max\{a,b,c\} & \text{if } a,b,c < 0, \\ 0 & \text{otherwise.} \end{cases} \quad (1.2.55)$$

This reconstruction can be combined with a linear reconstruction in time that avoids the use of an ODE solver for the semidiscrete method: this is the so-called MUSCL-Hancock reconstruction that will be used in Chapter 5.

#### 1.2.3.2 Third order CWENO reconstruction operator

The CWENO reconstructions were introduced in [117]. We consider here the expression of the operator described in [55] and [56]. Given a family of cell values $\{W_i\}$, the expression

of the reconstruction on the $i$th cell is a polynomial

$$\mathbb{P}_i^{CWENO3} = CWENO(P_{Opt}, P_1, P_2) \tag{1.2.56}$$

of degree 2 that depends on:

- $P_{Opt}$: polynomial of degree 2 that interpolates $\{W_{i-1}, W_i, W_{i+1}\}$.
- $P_1$: polynomial of degree 1 that interpolates $\{W_{i-1}, W_i\}$.
- $P_2$: polynomial of degree 1 that interpolates $\{W_i, W_{i+1}\}$.

By interpolating consecutive cell averages we mean that

$$\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} P(x)dx = W_j, \quad \forall j \in S_i. \tag{1.2.57}$$

Let us see how to obtain this polynomial for a stencil $S_i$ composed by $l$ cell averages: Let us consider $\overline{W}(x)$ a primitive of $W(x)$,

$$\overline{W}(x) = \int_{-\infty}^{x} W(x)dx, \tag{1.2.58}$$

where $W$ is the function of which the cell averages come from. We see that we can obtain the values $\overline{W}(x_{i+\frac{1}{2}})$ in terms of the cell averages:

$$\overline{W}(x_{i+\frac{1}{2}}) = \sum_{j=-\infty}^{i}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} W(x)dx = \sum_{j=-\infty}^{i}\Delta x W_j. \tag{1.2.59}$$

Now that we know the values $\overline{W}(x_{i+\frac{1}{2}})$, we denote by $\overline{P}$ the unique polynomial of degree at most $l$ that interpolates the function $\overline{W}$ evaluated in the intercells of the stencil. The polynomial of degree at most $l-1$ we are looking for is the derivative of $\overline{P}$:

$$P(x) = \overline{P}'(x). \tag{1.2.60}$$

Effectively:

$$\begin{aligned}
\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} P(x)dx &= \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \overline{P}'(x)dx \\
&= \frac{1}{\Delta x}\left(\overline{P}(x_{j+\frac{1}{2}}) - \overline{P}(x_{j-\frac{1}{2}})\right) \\
&= \frac{1}{\Delta x}\left(\overline{W}(x_{j+\frac{1}{2}}) - \overline{W}(x_{j-\frac{1}{2}})\right) \\
&= \frac{1}{\Delta x}\left(\int_{\infty}^{x_{j+\frac{1}{2}}} W(x)dx - \int_{\infty}^{x_{j-\frac{1}{2}}} W(x)dx\right) \\
&= \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} W(x)dx = W_j, \quad j \in S_i.
\end{aligned} \tag{1.2.61}$$

The definition of $\mathcal{P}_i^{CWENO3}$ depends on the choice of a set of positive real coefficients $d_0, d_1, d_2 \in [0,1]$ such that $\sum_{k=0}^{2} d_k = 1$, $d_0 \neq 0$, called linear coefficients, in the following way:

1. We introduce the second degree polynomial $P_0$ defined as:

$$P_0(x) = \frac{1}{d_0}\left(P_{Opt}(x) - d_1 P_1(x) - d_2 P_2(x)\right). \tag{1.2.62}$$

2. The nonlinear coefficients $w_k$ are computed from the linear ones as:

$$\alpha_k = \frac{d_k}{(I[P_k] + \epsilon)^t} \quad w_k = \frac{\alpha_k}{\sum_{j=0}^{2} \alpha_j}, \tag{1.2.63}$$

where $I[P_k]$ denotes a suitable smoothness indicator that in our case will be the Jiang-Shu indicator (see [96]):

$$I[P_k] = \sum_{l \geq 1} diam(I_i)^{2l-1} \int_{I_i} \left(\frac{d^l}{dx^l} P_k(x)\right) dx, \tag{1.2.64}$$

being $\frac{d^l}{dx^l} P_k(x)$ the $l$-th derivative of $P_k$, $\epsilon$ is a small positive quantity and $t \geq 2$. Different smoothness indicators have been used in the literature, see for instance [9], [30].

3. Finally, $P_i^{CWENO3}$ is given by:

$$\mathbb{P}_i^{CWENO3}(x) = \sum_{k=0}^{2} w_k P_k(x). \tag{1.2.65}$$

The following choice of linear coefficients will be used in this thesis:

$$d_0 = 0.7, \quad d_1 = 0.15 \quad d_2 = 0.15.$$

This third order reconstruction operator was also considered in [56] and [138] but using different coefficients.

### 1.2.4 Path-conservative schemes: well-balancing

Systems of the form (1.1.24) may have non-trivial stationary solutions. This is the case of solutions that balance the flux, nonconservative products, and source terms in systems of the form (1.1.23). The design of numerical methods that preserve all or a significant set of stationary solutions is crucial when the waves generated by a perturbation of an equilibrium have to be simulated: numerical methods with this property are called in

general well-balanced. Since [13], where this property was firstly called the $C$-Property and it was only related to the water at rest for the shallow water equations, the study and design of well-balanced numerical methods have been a very active front of research: see for instance [5, 12, 28, 34, 32, 131, 50, 145, 43, 47, 123, 122, 61, 60, 80, 79, 23]. In this section we will review the definition of well-balanced methods and how to obtain high-order well-balanced methods through the use of high-order well-balanced reconstruction operators.

#### 1.2.4.1 Definition of the well-balanced property

The stationary solutions of the system (1.1.24) are those solutions of the system such that $W_t = 0$ and verify

$$\mathcal{A}(W(x))W_x(x) = 0, \tag{1.2.66}$$

for all $x$ where $W$ is defined. Let us take a regular stationary solution $W$, we observe from (1.2.66) that 0 is an eigenvalue of $\mathcal{A}(W(x))$ and $W_x(x)$ is an associated eigenvector for every $x$. Therefore, $x \rightarrowtail W(x)$ can be interpreted as a parametrization of an integral curve of a linearly degenerated characteristic field whose corresponding eigenvalue takes the value 0 through the curve. We denote by $\Gamma$ the set of all the integral curves $\gamma$ of a linearly degenerated field of $\mathcal{A}(W)$ such that the corresponding eigenvalue vanishes on $\Gamma$. We are now able to state the general definition of a well-balanced method.

**Definition 1.2.6.** *A numerical scheme* (1.2.3) *is said to be well-balanced if, given any pair of states $W_l$ and $W_r$ belonging to $\gamma \in \Gamma$ one has*

$$\mathcal{D}^{\pm}(W_l, W_r) = 0. \tag{1.2.67}$$

If the numerical method with this definition is applied to the initial condition

$$W_i^0 = W^*(x_i), \quad \forall i,$$

being $W^*$ a stationary solution, then

$$W_i^n = W_i^0, \quad \forall i.$$

In practice if we have two states $W_l$ and $W_r$ belonging to the same integral curve $\gamma \in \Gamma$, then we want the stationary contact discontinuity at $x^*$

$$W(x,t) = \begin{cases} W_l, & x < x^*, \\ W_r, & x > x^*, \end{cases} \tag{1.2.68}$$

to be a weak solution. In the case of conservative systems the fact of belonging to the same integral curve is equivalent to the preservation of the corresponding Riemann invariant so it will be an admissible weak solution. The problem when having a nonconservative

system is that we need to choose a notion of weak solution where (1.2.68) is an admissible solution. Therefore, the family of paths has to be chosen so that (1.2.68) satisfies the Rankine-Hugoniot condition (1.1.27). In order to do this, it is enough to ensure that, in the case we have two states $W_l$ and $W_r$ belonging to the same integral curve of a linearly degenerated field, then the corresponding path is a parametrization of the arc of this integral curve linking $W_l$ and $W_r$.

Let us check this in the particular case of systems of the form (1.1.23) that can be written in the form (1.1.24) with:

$$W = \begin{pmatrix} U \\ \sigma \end{pmatrix}, \quad \mathcal{A}(W) \equiv \mathcal{A}(U) = \left(\begin{array}{c|c} A(U)+B(U) & -S(U) \\ \hline 0 & 0 \end{array}\right), \tag{1.2.69}$$

being $A(U) = JF(U)$ the Jacobian of the flux $F$. Let us suppose that the matrix $A(U) + B(U)$ has $N-1$ real eigenvalues

$$\lambda_1(U) < ... < \lambda_{N-1}(U), \tag{1.2.70}$$

and associated eigenvectors $r_j(U), j = 1, ..., N-1$. If none of these eigenvalues is null, then System (1.1.24) is strictly hyperbolic: $\mathcal{A}(W)$ has $N$ different real eigenvalues

$$\lambda_1(U), ..., \lambda_{N-1}(U), 0, \tag{1.2.71}$$

with associated eigenvectors $R_j(U), j = 1, ..., N$ given by

$$R_j(U) = \begin{pmatrix} r_j(U) \\ 0 \end{pmatrix}, \quad j = 1, ..., N-1, \quad R_N(U) = \begin{pmatrix} (A(U)+B(U))^{-1} S(U) \\ 1 \end{pmatrix}. \tag{1.2.72}$$

In this case the set $\Gamma$ is composed by the integral curves of the linearly degenerated field $R_N(U)$, i.e, the integral curves of the ODE system

$$\frac{dW}{ds} = R_N(W), \tag{1.2.73}$$

i.e.,

$$\begin{cases} \dfrac{dU}{ds} = (A(U)+B(U))^{-1} S(U), \\ \dfrac{d\sigma}{ds} = 1, \end{cases} \tag{1.2.74}$$

that is equivalent, choosing $\sigma$ as the parameter and using the last equation of (1.2.74), to:

$$\frac{dU}{d\sigma} = (A(U)+B(U))^{-1} S(U), \tag{1.2.75}$$

what implies

$$(A(U) + B(U))\,\frac{dU}{d\sigma} = S(U), \tag{1.2.76}$$

that is clearly a reparametrization of the equation satisfied by the stationary solutions

$$(A(U) + B(U))\,\frac{dU}{dx} = S(U)\frac{d\sigma}{dx}. \tag{1.2.77}$$

Therefore a pair of states $W_l = \begin{pmatrix} U_l \\ \sigma_l \end{pmatrix}$ and $W_r = \begin{pmatrix} U_r \\ \sigma_r \end{pmatrix}$ belong to the same integral curve $\gamma \in \Gamma$ if and only if there exists a solution of the ODE system

$$(A(V) + B(V))\,\frac{dV}{d\sigma} = S(V) \tag{1.2.78}$$

such that

$$\begin{cases} V(\sigma_l) = U_l, \\ V(\sigma_r) = U_r. \end{cases} \tag{1.2.79}$$

**Remark 1.2.3.** *If $V(\sigma)$ is a solution of* (1.2.78)*, then $U(x) = V(\sigma(x))$ is a stationary solution of* (1.1.23)*: it satisfies*

$$F(U)_x + B(U)U_x = S(U)\sigma_x$$

*in the smooth regions and* (1.2.78)-(1.2.79) *in the discontinuities.*

As we said before, we want to choose a family of paths that, in the case we have two states belonging to the same integral curve, they can be connected by an admissible stationary contact discontinuity. A possible election is the following one:

$$\Phi(\xi; W_l, W_r) = \begin{pmatrix} \Phi_U(\xi; W_l, W_r) \\ \Phi_\sigma(\xi; W_l, W_r) \end{pmatrix} = \begin{pmatrix} V(\sigma_l + \xi(\sigma_r - \sigma_l)) \\ \sigma_l + \xi(\sigma_r - \sigma_l) \end{pmatrix}, \quad \xi \in [0,1]. \tag{1.2.80}$$

Effectively, if we choose this path the solution (1.2.68) verifies the Rankine-Hugoniot condition with null speed:

$$\begin{aligned} &\int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_r))\frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_r)d\xi = \\ &\int_0^1 \left[(A(\Phi_U(\xi; W_l, W_r)) + B(\Phi_U(\xi; W_l, W_r)))\frac{\partial \Phi_U}{\partial \xi}(\xi; W_l, W_r) - S(\Phi_U(\xi; W_l, W_r))\frac{\partial \Phi_\sigma}{\partial \xi}(\xi; W_l, W_r)\right] d\xi \\ &= \int_{\sigma_l}^{\sigma_r} \left[(A(V(\sigma)) + B(V(\sigma)))\frac{dV}{d\sigma}(\sigma) - S(V(\sigma))\right] d\sigma = 0, \end{aligned}$$

where $V$ verifies (1.2.78)-(1.2.79).

As we have seen if the eigenvalues of $A(U_l) + B(U_l)$ do not vanish, the jump condition (1.2.78)-(1.2.79) is equivalent to state that the solution of the Cauchy problem

$$\begin{cases} \dfrac{dV}{d\sigma} = (A(V) + B(V))^{-1} S(V) \\ V(\sigma_l) = U_l \end{cases} \tag{1.2.81}$$

is defined in $\sigma_r$ and satisfies $V(\sigma_r) = U_r$.

**Remark 1.2.4.** *Solving* (1.2.81) *we obtain all the possible right states that can be linked through an admissible stationary discontinuity at* $x^*$ *with* $W_l = \begin{pmatrix} U_l \\ \sigma_l \end{pmatrix}$.

**Remark 1.2.5.** *In the resonant case, i.e., when one of the eigenvalues of* $A(U_l) + B(U_l)$ *vanishes, the Cauchy problem* (1.2.81) *may have no solution or to have more than one. In the case of multiple solutions a criterion is needed to decide what are the admissible discontinuities to be preserved by the numerical method (see, for instance, [84]). In Chapter 2 we will see the Monotonicity criterion for the shallow Water equations with topography introduced in [112] that is similar to the one used in [111].*

At this point we have determined the path (1.2.80) linking pairs of states that can be the limits of an admissible jump at the discontinuity points of $\sigma$. Let us suppose that our family of paths

$$\Phi(\xi; W_l, W_r) = \begin{pmatrix} \Phi_U(\xi; W_l, W_r) \\ \Phi_\sigma(\xi; W_l, W_r) \end{pmatrix}, \quad \xi \in [0, 1]. \tag{1.2.82}$$

reduces to (1.2.80) when the states belong to the same integral curve of the linearly degenerated field associated with the null eigenvalue (see [34] for a possible choice). Let us see how, with this election of family of paths, the methods presented in the previous section can have the well-balanced property.

## 1.2.5 Well-balanced path-conservative methods

Let us see the requirements under which the numerical methods described above are well-balanced if the family of paths is such that the path linking the states that belong to the same integral curve $\gamma \in \Gamma$ is a parametrization of the arc of this curve.

### 1.2.5.1 Well-balanced property of Godunov and Roe

Let us take a pair of states $W_l$ and $W_r$ belonging to $\gamma \in \Gamma$. In this case the solution of the Riemann problem (1.2.11) is the stationary contact discontinuity so that $W_0^-$ and $W_0^+$ in (1.2.15) coincide with $W_l$ and $W_r$, respectively, then in the Godunov method (1.2.13)-(1.2.14) we obtain:

$$\mathcal{D}_{GOD}^-(W_l, W_r) = \int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_l) \frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_l) d\xi = 0,$$

$$\mathcal{D}^{+}_{GOD}(W_l, W_r) = \int_0^1 \mathcal{A}(\Phi(\xi; W_r, W_r)\frac{\partial \Phi}{\partial \xi}(\xi; W_r, W_r)d\xi = 0,$$

so that the Godunov method is well-balanced with this family of paths. In the case of the Roe methods the property of the Roe matrix (1.2.16) reduces to:

$$\mathcal{A}_\Phi(W_l, W_r)(W_r - W_l) = \int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_r))\frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_r)d\xi = 0, \tag{1.2.83}$$

so 0 is an eigenvalue of $\mathcal{A}_\Phi$ with associated eigenvector $W_r - W_l$. Then by definition, $W_r - W_l$ is also an eigenvector associated to the eigenvalue 0 for $\mathcal{A}^-_\Phi$ and $\mathcal{A}^+_\Phi$ and therefore we obtain from (1.2.17)-(1.2.18):

$$\mathcal{D}^{\pm}_{ROE}(W_l, W_r) = \mathcal{A}^{\pm}_\Phi(W_l, W_r)(W_r - W_l) = 0,$$

so that the Roe method is well-balanced with this family of paths.

#### 1.2.5.2 Well-balanced Polinomial Viscosity matrix methods

Let us consider now PVMs methods whose fluctuations are given by:

$$\mathcal{D}^{\pm}_{PVM}(W_l, W_r) = \frac{1}{2}\mathcal{A}_\Phi(W_r - W_l) \pm \frac{1}{2}\mathcal{Q}_\Phi(W_r - W_l).$$

Using again the property of the Roe matrix (1.2.83) and the expression of the viscosity matrix $\mathcal{Q}_\Phi = p_r(\mathcal{A}_\Phi)$ (1.2.27) we obtain:

$$\mathcal{D}^{\pm}_{PVM}(W_l, W_r) = \pm\frac{1}{2}\sum_{j=0}^{r}\alpha_j\mathcal{A}^j_\Phi(W_r - W_l) = \pm\frac{1}{2}\alpha_0 I(W_r - W_l),$$

and this is 0 if and only if we have $\alpha_0 = 0$, i.e., if and only if the polynomial $p_r$ verifies $p_r(0) = 0$. A modification of the PVM methods that allows one to obtain well-balanced methods when $\alpha_0 = 0$ was introduced in [45]. This modification is based on the use of a modified identity matrix: in the term $\alpha_0 I$, the matrix is replaced by

$$\tilde{I}(W_l, W_r) = R_\Phi(W_l, W_r)\tilde{Id}(W_r, W, l)R^{-1}_\Phi(W_r, W_l), \tag{1.2.84}$$

where again $R_\Phi(W_l, W_r)$ is a matrix whose columns are eigenvectors of $\mathcal{A}_\Phi(W_l, W_r)$, and $\tilde{Id}$ is the diagonal matrix whose $i$th coefficient is 1 if $\lambda_i(W_l, W_r) \neq 0$, or 0 if $\lambda_i(W_l, W_r) = 0$. With this modification, if the states $W_l$ and $W_r$ belong to $\gamma \in \Gamma$, $W_r - W_l$ is an eigenvector associated to 0 for $\mathcal{A}_\Phi$ and also for $\tilde{I}(W_l, W_r)$, then

$$\mathcal{D}^{\pm}_{PVM}(W_l, W_r) = \pm\frac{1}{2}\alpha_0\tilde{I}(W_l, W_r)(W_r - W_l) = 0,$$

so that with this modification the PVM method is well-balanced with this family of paths.

**Remark 1.2.6.** *In the case of problems of the form* (1.1.23)*, some algebraic calculations allow us to show that the matrix* $\tilde{I}(W_l, W_r)$ *has the following block structure:*

$$\tilde{I}(W_l, W_r) = \left( \begin{array}{c|c} I & -(A_\Phi + B_\Phi)^{-1} S_\Phi \\ \hline 0 & 0 \end{array} \right). \tag{1.2.85}$$

*This modification is equivalent to write the first term of the viscosity part as:*

$$\alpha_0 (W_r - W_l - (A_\Phi + B_\Phi)^{-1} S_\Phi (\sigma_r - \sigma_l)). \tag{1.2.86}$$

### 1.2.5.3 Well-balanced simple Riemann solvers

As it was pointed out in [34] and [37], from the expression (1.2.35)-(1.2.36) of the fluctuations, it can be easily deduced that a sufficient condition to have the well-balanced property is the following: given two states $W_l$ and $W_r$ belonging to the same integral curve of a linearly degenerate field, the corresponding simple Riemann solver is given by:

$$R(\sigma; W_l, W_r) = \begin{cases} W_l & if \quad \sigma < 0, \\ W_r & if \quad \sigma > 0. \end{cases}$$

Let us consider a SRS $R(\sigma; W_l, W_r)$ for (1.1.24) consisting of $s$ speeds (including 0)

$$\sigma_0 = -\infty < \sigma_1 < ... < \sigma_{k-1} < \sigma_k = 0 < \sigma_{k+1} < ... < \sigma_s < \sigma_{s+1} = +\infty, \tag{1.2.87}$$

and $s-1$ intermediate states: $W_j^l = \begin{pmatrix} U_j \\ \sigma_l \end{pmatrix}$ for $j = 1, ..., k-1$ and $W_j^r = \begin{pmatrix} U_j \\ \sigma_r \end{pmatrix}$ for $j = k, ..., s-1$ such that

$$R(\sigma; W_l, W_r) = \begin{cases} W_0 = W_l, & if \quad \sigma < \sigma_1, \\ W_1^l, & if \quad \sigma_1 < \sigma < \sigma_2, \\ \vdots & \\ W_{k-1}^l, & if \quad \sigma_{k-1} < \sigma < 0 \\ W_k^r, & if \quad 0 < \sigma < \sigma_k \\ \vdots & \\ W_{s-1}^r, & if \quad \sigma_{s-1} < \sigma < \sigma_s, \\ W_s = W_r, & if \quad \sigma_s < \sigma. \end{cases} \tag{1.2.88}$$

The corresponding method will be well-balanced provided that $W_j^l = W_l$ for all $j = 1, ..., k-1$ and $W_j^r = W_r$ for all $j = k, ..., s-1$ whenever the states $W_l$ and $W_r$ belongs to the same integral curve of a linearly degenerate field. The HLL scheme (1.2.39) does not verify this property in general. Nevertheless, it can be generalized in the following form:

$$R^{HLL}(\sigma; W_l, W_r) = \begin{cases} W_l & if \quad \sigma < S_l, \\ W_l^* & if \quad S_l \leq \sigma \leq 0, \\ W_r^* & if \quad 0 \leq \sigma \leq S_r, \\ W_r & if \quad \sigma > S_r, \end{cases} \tag{1.2.89}$$

where the states $W_l^* = (U_l^*, \sigma_l)^T$ and $W_r^* = (U_r^*, \sigma_r)^T$ should satisfy the consistency property (1.2.33), that in this case of system of the form (1.1.23) it reduces to

$$\begin{aligned} S_l(U_l^* - U_l) + S_r(U_r - U_r^*) = F(U_r) - F(U_l) \\ + \int_0^1 B(\Phi_U)\frac{\partial \Phi_U}{\partial \xi}d\xi - \int_0^1 S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi}d\xi, \end{aligned} \tag{1.2.90}$$

where we have skipped the variables of the paths for shortness. As we have seen, to obtain a well-balanced method we should have $W_l^* = W_l$ and $W_r^* = W_r$ whenever $W_l$ and $W_r$ belong to the same integral curve $\gamma \in \Gamma$. If this happen, integrating (1.2.75), we obtain

$$U_r - U_l = \int_0^1 (A+B)^{-1}(\Phi_U)S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi}d\xi. \tag{1.2.91}$$

Therefore, if $W_l^* = W_l$ and $W_r^* = W_r$ one has

$$U_l^* - U_l + U_r - U_r^* = U_r - U_l - \int_0^1 (A+B)^{-1}(\Phi_U)S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi}d\xi. \tag{1.2.92}$$

Imposing (1.2.90) and (1.2.92) for any pair of states, some easy computations lead to the definitions:

$$U_l^* = U^* - \frac{S_r}{S_r - S_l}\int_0^1 (A+B)^{-1}(\Phi_U)S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi}d\xi, \tag{1.2.93}$$

$$U_r^* = U^* - \frac{S_r}{S_r - S_l}\int_0^1 (A+B)^{-1}(\Phi_U)S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi}d\xi, \tag{1.2.94}$$

where

$$U^* = \frac{S_r W_r - S_l W_l - \left(F(U_r) - F(U_l) + \int_0^1 B(\Phi_U)\frac{\partial \Phi_U}{\partial \xi}d\xi - \int_0^1 S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi}d\xi\right)}{S_r - S_l}, \tag{1.2.95}$$

is the intermediate state of the HLL method (1.2.40). This gives a well-balanced HLL method.

This idea can be generalized to define well-balanced SRSs: Let us suppose that, for any pair of states $W_l$ and $W_r$, $s$ speeds (including 0)

$$\sigma_0 = -\infty < \sigma_1 < ... < \sigma_{k-1} < \sigma_k = 0 < \sigma_{k+1} < ... < \sigma_s < \sigma_{s+1} = +\infty, \tag{1.2.96}$$

and $s-1$ states $V_1, V_2, ..., V_k, ...V_{s-1} \in \mathbb{R}^N$ can be chosen so that

$$\sum_{j=1}^{k-1} V_j + \sum_{j=k+1}^{s-1} V_j = U_r - U_l - \int_0^1 (A+B)^{-1}(\Phi_U)S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi}d\xi, \tag{1.2.97}$$

$$\sum_{j=1}^{s-1} \sigma_j V_j = F(U_r) - F(U_l) + \int_0^1 B(\Phi_U)\frac{\partial \Phi_U}{\partial \xi} d\xi - \int_0^1 S(\Phi_U)\frac{\partial \Phi_\sigma}{\partial \xi} d\xi, \tag{1.2.98}$$

and $V_j = 0$ for $j \neq k$ whenever $W_r$ and $W_l$ belong to the same integral curve $\gamma \in \Gamma$. Then, define the states $W_j \in \mathbb{R}^N$ for $j = 0, 1, ..., s$ as follows:

$$W_j = \begin{pmatrix} U_j \\ \sigma_l \end{pmatrix} \quad for \quad j = 1, ...k-1, \quad and \quad W_j = \begin{pmatrix} U_j \\ \sigma_r \end{pmatrix} \quad for \quad j = k, ...s, \tag{1.2.99}$$

where

$$U_0 = U_l, \quad U_j = V_j + U_{j-1} \quad for \quad j = 1, ...k-1, \tag{1.2.100}$$

$$U_s = U_r, \quad U_j = U_{j+1} - V_j \quad for \quad j = s-1, ...k. \tag{1.2.101}$$

The speeds $\sigma_j$ and the intermediate states $W_j$ define a well-balanced SRS.

In [37] and [146] a well-balanced SRS has been applied to turbidity currents and Ripa models, respectively.

#### 1.2.5.4 Generalized Hydrostatic Reconstruction

An alternative strategy for defining well-balanced methods is the so-called generalized hydrostatic reconstruction (GHR) introduced in [46] as a generalization of the hydrostatic reconstruction technique introduced in [5] (which was further enhanced in [19]) to obtain schemes that preserve the solutions corresponding to water at rest for the shallow water equations. It is based not only in a specific choice of the family of paths but also in an election of the fluctuation terms. Let us consider two arbitrary states $W_l = \begin{pmatrix} U_l \\ \sigma_l \end{pmatrix}$ and $W_r = \begin{pmatrix} U_r \\ \sigma_r \end{pmatrix}$, the GHR follows the next steps:

1. Let us consider a family of paths $\varphi(\xi; U_l, U_r)$ and a path-conservative method $\mathcal{D}_H^{\pm}$ for the homogeneous problem:

$$U_t + F(U)_x + B(U)U_x = 0. \tag{1.2.102}$$

2. Now we define the family of paths $\Phi$ for the non-homogeneous problem following this procedure:

   - First we choose an intermediate value $\sigma_0$ such as $\sigma_l = \sigma_r = \sigma_0$ when $\sigma_l = \sigma_r$. In Chapter 2 we have used

$$\sigma_0 = min(\sigma_l, \sigma_r). \tag{1.2.103}$$

- Next we link (if possible) the states $W_l$ and $W_r$ with two states $W_0^- = (U_0^-, \sigma_0)^T$, $W_0^+ = (U_0^+, \sigma_0)^T$, respectively, through the integral curve associated with the linearly degenerated field whose corresponding eigenvalue is null, i.e, we solve the system (1.2.78) with initial condition $V_l(\sigma_l) = U_l$ and $V_r(\sigma_r) = U_r$, respectively, and we denote these solutions by $V_l(\sigma)$ and $V_r(\sigma)$, then:

$$U_0^- = V_l(\sigma_0), \quad U_0^- = V_r(\sigma_0). \tag{1.2.104}$$

- Finally we use the path $\varphi$ to connect the states $U_0^-$ and $U_0^+$.

The family of paths $\Phi$ linking $W_l$ and $W_r$ will be the union of the two arcs of the integral curve and the curve $s \to (\varphi(\xi; U_0^-, U_0^+), \sigma_0)^T$.

3. We consider the following fluctuation:

$$\mathcal{D}^{\pm}_{GHR}(W_l, W_r) = \mathcal{D}^{\pm}_{H}(U_0^-, U_0^+). \tag{1.2.105}$$

**Remark 1.2.7.** *We observe that the path $\Phi$ reduces to* (1.2.80) *if $W_l$ and $W_r$ belong to the same integral curve of a linearly degenerated field: observe that, in this case, $U_0^- = U_0^+$.*

This method is well-balanced whatever the choice of path and of fluctuations is made in the homogeneous case. Effectively, if $W_l$ and $W_r$ belong to the same integral curve $\gamma \in \Gamma$, then $W_0^- = W_0^+$, so

$$\mathcal{D}^{\pm}_{GHR}(W_l, W_r) = \mathcal{D}^{\pm}_{H}(U_0^-, U_0^+) = 0,$$

where we have used (1.2.5). If the homogeneous problem is conservative, it does not matter the family of paths $\varphi$ we chose: we can simply chose the family of segments:

$$\varphi(\xi; W_l, W_r) = W_l + \xi(W_r - W_l). \tag{1.2.106}$$

**Remark 1.2.8.** *This alternative strategy to generate well-balanced methods avoids, for example, the computation of the inverse of $A + B$ in the PVM case that can lead to difficulties in the sonic points.*

### 1.2.6 Path-conservative schemes: high-order well-balanced reconstruction operators

The well-balanced property of a first path-conservative method can be lost if it is extended to high-order using a reconstruction operator: see (1.2.48). Nevertheless, as pointed out in [40] and more recently in [47], the high-order semidiscrete method is still well-balanced if the reconstruction operator is well-balanced in the following sense:

**Definition 1.2.7.** *The reconstruction operator is said to be well-balanced for a stationary solution $W(x)$ if*

$$\mathbb{P}_i(x;\{W_j\}_{j\in S_i}) = W(x), \quad \forall x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \forall i, \tag{1.2.107}$$

*where*

$$W_i \cong \frac{1}{\Delta x}\int_{I_i} W(x)dx.$$

Therefore, the main ingredients to obtain high-order well-balanced methods are:

- A first-order well-balanced path-conservative scheme like the ones discussed in the previous subsection for computing the $\mathcal{D}^{\pm}$ part of the semidiscrete method (1.2.48).
- A well-balanced high-order reconstruction operator to compute the integral appearing in the semidiscrete method (1.2.48) and the reconstructed states $\{W^{\pm}_{i\pm\frac{1}{2}}\}$.

**Theorem 1.2.1.** *Let $W$ be a stationary solution such that $x \to W(x)$ is a parametrization of a curve $\gamma \in \Gamma$. Let us consider a first order path-conservative method $\mathcal{D}^{\pm}$ which is well-balanced for $\gamma$ and a reconstruction operator that is well-balanced for $W$. Then, the numerical method* (1.2.48) *is well-balanced for $W$ in the sense that the vector of its cell-averages $\{W_i\}$ is an equilibrium of the ODE system* (1.2.48).

*Proof.* Let us consider the right-hand side of (1.2.48) corresponding to the cell averages of $W$. Since the reconstruction operator is well-balanced, one has

$$\mathbb{P}^0_i(x) = W(x), \quad \forall x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \forall i. \tag{1.2.108}$$

Moreover,

$$W^{0,-}_{i+\frac{1}{2}} = \mathbb{P}^0_i(x_{i+\frac{1}{2}}), \quad W^{0,+}_{i+\frac{1}{2}} = \mathbb{P}^0_{i+1}(x_{i+\frac{1}{2}}), \tag{1.2.109}$$

then $W^{0,-}_{i+\frac{1}{2}}$ and $W^{0,+}_{i+\frac{1}{2}}$ belong to $\gamma \in \Gamma$. Therefore:

$$\begin{aligned}
&\mathcal{D}^+_{i-\frac{1}{2}} + \mathcal{D}^-_{i+\frac{1}{2}} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}^0_i(x))\frac{\partial \mathbb{P}^0_i}{\partial x}(x)dx \\
&= \mathcal{D}^+(W^{0,-}_{i-\frac{1}{2}}, W^{0,+}_{i-\frac{1}{2}}) + \mathcal{D}^-(W^{0,-}_{i+\frac{1}{2}}, W^{0,+}_{i+\frac{1}{2}}) + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(W(x))\frac{\partial W}{\partial x}(x)dx = 0,
\end{aligned}$$

as we wanted to prove. □

A standard reconstruction operator is not expected in general to be well-balanced: the functions $\mathbb{P}_i$ are usually computed by interpolation techniques within a particular class of functions, such as polynomials, hyperbolas, etc, and , in general, the stationary solutions are not expected to belong to that class. Nevertheless, the technique introduced in [40]

can be used in order to make well-balanced a standard reconstruction operator that will be denoted by

$$\mathbb{Q}_i(x) = \mathbb{Q}_i(x; \{W_j\}_{j\in S_i}).$$

The following steps have to be performed in order to compute a well-balanced reconstruction operator $\mathbb{P}_i$ at the cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ for a given family of cell values $\{W_i\}$:

1. Look for the stationary solution $W_i^*(x)$ defined in the stencil of cell $I_i$ $(\cup_{j\in\mathcal{S}_i} I_j)$ such that:
$$\frac{1}{\Delta x}\int_{x_{i+\frac{1}{2}}}^{x_{i-\frac{1}{2}}} W_i^*(x)dx = W_i, \tag{1.2.110}$$
if possible. In other cases consider $W_i^* \equiv 0$.

2. Compute the fluctuations $\{V_j\}_{j\in S_i}$ within the stencil $S_i$:
$$V_j = W_j - \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} W_i^*(x)dx, \quad j \in S_i. \tag{1.2.111}$$

3. Apply the standard reconstruction operator to the fluctuations $\{V_j\}_{j\in S_i}$:
$$\mathbb{Q}_i(x) = \mathbb{Q}_i(x; \{V_j\}_{j\in S_i}).$$

4. Define the well-balanced operator:
$$\mathbb{P}_i(x) = W_i^*(x) + \mathbb{Q}_i(x).$$

$\mathbb{P}_i$ is well-balanced for every stationary solution provided that the reconstruction operator $\mathbb{Q}_i$ is exact for the null function. Moreover, it is conservative, i.e.,

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbb{P}_i(x)dx = W_i, \quad \text{for all } i,$$

provided that $\mathbb{Q}_i$ is conservative, and it is high-order accurate provided that the steady solutions are smooth (see [47] for details).

The key point of this procedure is the first step where in general, if we are able to obtain the stationary solution at least in its implicit form, we need to solve the $N \times N$ nonlinear system (1.2.110). Observe that, if it is impossible to find a stationary solution defined in the stencil that satisfies (1.2.110) then the standard reconstruction is used. Please note that this choice does not spoil the well-balanced character of the numerical method: in this case, the cell values in the stencil cannot be the averages of a stationary solution (otherwise there would be at least one solution $W_i^*$) and thus there is no local

equilibrium to preserve. On the other hand, if there is more than one stationary solution defined on the stencil that satisfies (1.2.110), a criterion is needed to select one of them depending on the problem, as we will see in Chapter 4.

In general the averages of the initial condition are computed using a quadrature formula:

$$W_{0,i} = \sum_{k=0}^{M} \alpha_k^i W_0(x_k^i), \quad \forall i, \tag{1.2.112}$$

where $\alpha_0^i, ..., \alpha_M^i$ and $x_0^i, ..., x_M^i$ represent, respectively, the weights and the nodes of the chosen quadrature formula, whose order of accuracy must be greater or equal to the one of the reconstruction operator. In this case the two first steps of the well-balanced reconstruction procedure have to be modified to obtain well-balanced methods:

1. Look for the stationary solution $W_i^*(x)$ defined in the stencil of cell $I_i$ $(\cup_{j\in\mathcal{S}_i} I_j)$ such that:

$$\sum_{k=0}^{M} \alpha_k^i W_i^*(x_k^i) = W_i, \tag{1.2.113}$$

   if possible. In other cases consider $W_i^* \equiv 0$.

2. Compute the fluctuations $\{V_j\}_{j\in S_i}$ within the stencil $S_i$:

$$V_j = W_j - \sum_{k=0}^{M} \alpha_k^j W_i^*(x_k^j), \quad j \in S_i. \tag{1.2.114}$$

The well-balanced property can also be lost if the quadrature formula is used to compute the integral appearing in (1.2.48). In order to circumvent this difficulty, the semi-discrete scheme is first rewritten as proposed in [47] taking into account the non-conservative part

$$\frac{dW_i}{dt} = -\frac{1}{\Delta x}\Big(\mathcal{D}^-_{i+\frac{1}{2}} + \mathcal{D}^+_{i-\frac{1}{2}} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\mathcal{A}(\mathbb{P}_i(x))\frac{\partial}{\partial x}\mathbb{P}_i(x) - \mathcal{A}(W_i^*(x))\frac{\partial}{\partial x}W_i^*(x)\right) dx, \tag{1.2.115}$$

where $W_i^*(x)$ is the stationary solution found at the first stage of the reconstruction operator when applied to $\{W_i(t)\}$. Let us consider the particular case of systems of the form (1.1.23) that can be written in the form (1.1.24) with:

$$W = \begin{pmatrix} U \\ \sigma \end{pmatrix}, \quad \mathcal{A}(W) \equiv \mathcal{A}(U) = \left(\begin{array}{c|c} A(U) + B(U) & -S(U) \\ \hline 0 & 0 \end{array}\right), \tag{1.2.116}$$

being $A(U) = JF(U)$ the Jacobian of the flux $F$. In this case we obtain in (1.2.115):

$$\begin{aligned}\frac{dW_i}{dt} = &-\frac{1}{\Delta x}\Big(\mathcal{D}^-_{i+\frac{1}{2}} + \mathcal{D}^+_{i-\frac{1}{2}} + F(\mathbb{P}_i(x_{i+\frac{1}{2}})) - F(U_i^*(x_{i+\frac{1}{2}})) + F(U_i^*(x_{i-\frac{1}{2}})) - F(\mathbb{P}_i(x_{i-\frac{1}{2}}))\\ &+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(B(\mathbb{P}_i(x))\frac{\partial}{\partial x}\mathbb{P}_i(x) - B(U_i^*(x))\frac{\partial}{\partial x}U_i^*(x)\right)dx\\ &+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left((S(\mathbb{P}_i(x)) - S(U_i^*(x)))\frac{\partial}{\partial x}\sigma(x)\right)dx.\end{aligned}$$

Once this equivalent form is obtained, the quadrature formula can be applied to the integrals without losing the well-balanced property, and this leads to a numerical method of the form:

$$\begin{aligned}W_i'(t) = &-\frac{1}{\Delta x}\Bigg(\mathcal{D}^-_{i+\frac{1}{2}} + \mathcal{D}^+_{i-\frac{1}{2}} + F(\mathbb{P}_i(x_{i+\frac{1}{2}})) - F(U_i^*(x_{i+\frac{1}{2}})) + F(U_i^*(x_{i-\frac{1}{2}})) - F(\mathbb{P}_i(x_{i-\frac{1}{2}}))\\ &+ \sum_{k=0}^{M}\alpha_k^i\left(B(\mathbb{P}_i(x_k^i))\frac{\partial}{\partial x}\mathbb{P}_i(x_k^i) - B(U_i^*(x_k^i))\frac{\partial}{\partial x}U_i^*(x_k^i)\right)\\ &+ \sum_{k=0}^{M}\alpha_k^i\left(S(\mathbb{P}_i(x_k^i)) - S(U_i^*(x_k^i)))\frac{\partial}{\partial x}\sigma(x_k^i)\right)\Bigg),\end{aligned} \tag{1.2.117}$$

where $\alpha_0^i, ..., \alpha_M^i$ and $x_0^i, ..., x_M^i$ represent, respectively, the weights and the nodes of the chosen quadrature formula in the $i$th cell, it can be easily checked that the numerical method is still well-balanced for every stationary solution.

**Remark 1.2.9.** *In the case $B$ and $S$ are null, the scheme reduces to*

$$W_i'(t) = -\frac{1}{\Delta x}(\mathcal{F}_{i+\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}), \tag{1.2.118}$$

*and, therefore, it is conservative.*

Finally let us mention that this methodology can be extended to systems for which the solutions of the ODE system (1.2.66) are not available neither in explicit or implicit form: in this case, the nonlinear problems arising at Step 1 have to be solved numerically. In [87] a control-based strategy combined with a standard ODE solver is used to find solutions of (1.2.66) that satisfy conditions over their averages like (1.2.110) or (1.2.113).

This strategy, based on the concept of well-balanced high-order operators, has been applied to different models: shallow water systems [34, 43], blood flows in vessels [127], Euler system with gravity [78, 97], Ripa model [146], among others.

### 1.2.7 Path-conservative schemes: convergence issues

In [42] it was shown that, if the numerical solutions provided by a path-conservative method converge uniformly in the sense of graphs as $\Delta x \to 0$, the limit is a weak solution according to the chosen family of paths. Nevertheless, this notion of convergence (see [57], [107]) is too strong and the numerical solutions provided by finite-difference or finite-volume methods do not converge usually in this sense. This is not to say that path-conservative methods do not converge: in practice, it can be observed that numerical methods like the extensions of Godunov or Roe schemes described in the previous section converge in $L^1$-norm under the usual CFL condition. What happens is that the limit may be a weak solution according to a different family of paths, i.e. it is a classical solution in the smoothness regions but its discontinuities satisfy a jump condition (1.1.27) different of the expected one: see [42], [1]. In fact, the family of paths that controls the jump conditions satisfied by the limits of the numerical solutions is related to the viscous profiles of the equivalent equation of the method: see [42]. If, for instance, the family of paths is based on the viscous profiles related to a regularization (1.1.30), the leading terms in the equivalent equation that represent the numerical viscosity of the scheme may not match the viscous term in (1.1.30).

The definition of path-conservative method is a formal notion of consistency and, although Lax's equivalence Theorem ensures that consistency and stability implies convergence for linear systems, this is not the case in general for nonlinear problems. For instance, in the case of systems of conservation laws, stable conservative methods may converge to solutions that are not admissible weak solutions: this is the case for Roe methods that may converge to weak solutions that are not entropy solutions. In order to ensure the convergence to the right weak solutions, besides consistency and stability, entropy has to be well controlled: for instance, entropy-fix techniques have to be added to Roe methods (see [39]). In the case of nonconservative systems, consistency, stability, and control of the entropy are not enough: the numerical viscosity and, in general, the numerical dissipation effects, have to be well-controlled (see [109] for a review on this topic).

The design of finite-difference or finite-volume methods satisfying these four properties is difficult in general. Nevertheless, different techniques have been introduced to overcome, at least partially, this convergence issue: [15], [10], [4], [17], [39], [52], [53], [77], [132]. In particular the path-conservative entropy stable methods introduced in [39] and extended to DG high-order methods in [94] significantly reduce the convergence error: to do this, entropy-conservative numerical methods are first introduced that are stabilized by means of a discretization of the viscous term of the regularized equation (1.1.30). More recently, in [51], an in-cell discontinuous reconstruction technique has been added to first-order path-conservative methods that allows one to capture correctly weak solutions with isolated shock waves. An extension of this technique to second order will be developed in Chapter 5.

# Chapter 2

# The Riemann problem for the shallow water equations with topography: the wet-dry case

In this chapter we study the solutions of Riemann problems for the shallow water equations with topography. More precisely we consider problems whose initial conditions correspond to a wet-dry front, i.e, problems where there is vacuum on the right or on the left of a step. In the homogeneous case, i.e., when the bottom is flat, the equations have the structure of a system of conservation laws and it has been already discussed, including wet-dry situations, in [153]. When the bottom is not flat, they have the structure of a system of balance laws and new difficulties arise. On the one hand, nontrivial stationary solutions appear and have to be correctly handled in order to be able to design well-balanced methods. On the other hand, due to the initial condition or to the interaction of the fluid with the varying topography, wet-dry fronts can develop. These fronts appear very frequently in practical applications: floods, dam-breaks, breaking of waves on beaches, etc; and the design of numerical methods handling correctly with them is challenging: see [8, 38, 41, 54, 59, 124, 164, 24].

As we have seen in Chapter 1, Godunov-type methods are based on the exact or approximate solution of Riemann problems at the intercells. A good knowledge of the solutions of the Riemann problem is therefore necessary to approximate them correctly. In the case of a varying topography, different analytical and numerical studies of the solution of the Riemann problem have been performed: [3, 14, 141, 112, 113, 20, 16, 21, 91], etc. In [41] the solution of the Riemann problem corresponding to a wet-dry situation over a step has been partially studied in order to improve the Roe method. In some cases, the Riemann problem is interpreted as a Partial Riemann problem (in the sense proposed in [63]) associated to the homogeneous system.



In this chapter we study the complete solution of the wet-dry Reimann problem. Besides the theoretical interest of this analysis, the results may be useful to design numerical methods and/or to produce reference solutions to compare different schemes. The chapter is structured as follows: in Section 2.1 we introduce the model in consideration. Then, in Section 2.2, the simple waves of the system are described. In Section 2.3 the solutions of the Riemann problems are built by composing these simple waves. Depending on the initial conditions, we find zero, one, or two solutions: the data sets leading to one or another situations are specified. Moreover, following [41], problems with zero solutions will be reinterpreted as Partial Riemann problems associated to the homogeneous system what will allow us to build a solution. Finally in Section 2.4 some numerical results will be shown, where different numerical methods are compared. In particular, the behavior of the numerical methods in the non-uniqueness cases will be studied: as it will be seen, they can converge to one or to the other, what is the reason of the huge differences discussed in [124]. The content of this chapter was published by Parés and Pimentel-García in 2019 by the *Journal of Computational Physics*, see [134].

## 2.1 Model

We consider the shallow water system of Partial Differential Equations

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + \dfrac{gh^2}{2}\right) = gh\partial_x a, \\ \partial_t a = 0, \end{cases} \tag{2.1.1}$$

that governs the flow of a shallow layer of fluid, with the following notation:

- $h = h(x,t) \geq 0$ is the height of the water from the bottom to the surface;
- $u = u(x,t)$ is the depth-averaged horizontal velocity of the water;
- $g$ is the intensity of the gravitational field;
- $a = a(x)$ is the depth of the bottom from a reference level;
- $\eta = \eta(x,t)$ is the elevation of the surface of the water ($h(x,t) = \eta(x,t) + a(x)$).

See Figure 2.1 to clarify the notation.

A Riemann problem associated with (2.1.1) is a Cauchy Problem with initial condition:

$$(h,u,a)(x,0) = \begin{cases} (h_l,u_l,a_l), & x < 0, \\ (h_r,u_r,a_r), & x > 0. \end{cases} \tag{2.1.2}$$

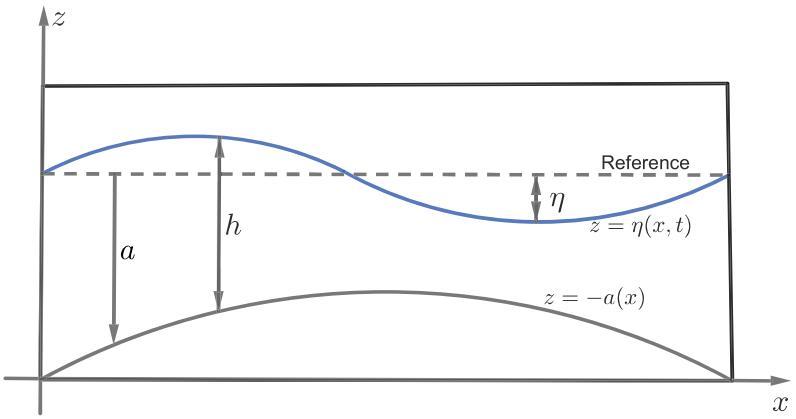



Figure 2.1: Shallow water system: notations

The goal of this chapter is to study the solution of Riemann problems corresponding to wet-dry fronts, i.e. Riemann problems whose initial condition is such that $h_l = 0$ or $h_r = 0$. When $a$ is discontinuous (and this is the case for Riemann problems) the source term $gh\partial_x a$ is a nonconservative product that, as we have seen in Chapter 1, cannot be defined in the distributions sense but it can be interpreted in the sense of measures using the theory developed in [57], but this interpretation is not unique. The definition of weak solutions of the system depends thus on the interpretation of the nonconservative products as measures. The notion of weak solution considered in the literature when addressing this type of Riemann problems is not always the same. We follow here the one adopted in [112] and [113] which is the one we used in Section 1.2.4: according to this definition, weak solutions develop stationary waves over the step across which Riemann invariants are preserved. Since these stationary waves are associated to a linearly degenerate characteristic field, as it will be seen in next section, they can be understood as contact discontinuities, so that the preservation of Riemann invariants is natural. According to this definition, in [112] and [113] it is shown that the wet-wet Riemann problem over a step may have 1, 2, or 3 solutions depending on the initial data. The conclusions are different for other definitions of weak solution: for instance, in [14] 0 or 1 solutions are found.

## 2.2 Simple waves

System (2.1.1) can be written in nonconservative form (1.1.24) as follows:

$$\partial_t W + \mathcal{A}(W)W_x = 0, \tag{2.2.1}$$

where:

$$W = \begin{pmatrix} h \\ u \\ a \end{pmatrix}, \quad \mathcal{A}(W) = \begin{pmatrix} u & h & 0 \\ g & u & g \\ 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of $\mathcal{A}$ are:

$$\lambda_1(W) = u - \sqrt{gh}, \quad \lambda_2(W) = u + \sqrt{gh}, \quad \lambda_3(W) = 0, \tag{2.2.2}$$

and:

$$R_1(W) = \begin{pmatrix} h \\ -\sqrt{gh} \\ 0 \end{pmatrix}, \quad R_2(W) = \begin{pmatrix} h \\ \sqrt{gh} \\ 0 \end{pmatrix}, \quad R_3(W) = \begin{pmatrix} gh \\ -gu \\ u^2 - gh \end{pmatrix}, \tag{2.2.3}$$

are associated eigenvectors.

We note that $\lambda_1(W) = \lambda_3(W)$ in the surface:

$$C^+ = \{(h, u, a) : u = \sqrt{gh}\}, \tag{2.2.4}$$

$\lambda_2(W) = \lambda_3(W)$ in the surface:

$$C^- = \{(h, u, a) : u = -\sqrt{gh}\}, \tag{2.2.5}$$

and $\lambda_1(W) = \lambda_2(W)$ when $h = 0$. Therefore, system (2.1.1) is non-strictly hyperbolic.

The characteristic fields associated to the eigenvalues $\lambda_1$ and $\lambda_2$ are genuinely nonlinear when $h > 0$, while the one associated to $\lambda_3$ is linearly degenerate.

The surface $C := C^+ \cup C^-$ contains the critical states and it divides the half-space $h \geq 0$ of the $(h, u, a)$-space in the following regions:

$$\begin{array}{l} A_1 := \{W = (h, u, a) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} : \lambda_2(W) > \lambda_1(W) > \lambda_3(W)\}, \\ A_2 := \{W = (h, u, a) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} : \lambda_2(W) > \lambda_3(W) > \lambda_1(W)\}, \\ A_2^+ := \{W = (h, u, a) \in A_2 : u > 0\}, \\ A_2^- := \{W = (h, u, a) \in A_2 : u < 0\}, \\ A_3 := \{W = (h, u, a) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} : \lambda_3(W) > \lambda_2(W) > \lambda_1(W)\} \end{array} \tag{2.2.6}$$

(see Figure 2.2). System (2.1.1) is strictly hyperbolic in the interior of these regions. Moreover, states belonging to $A_1 \cup A_3$ are supercritical and those belonging to $A_2$ subcritical.

Following [112] and using what we have seen in Chapter 1, the simple waves for system (2.1.1) are:

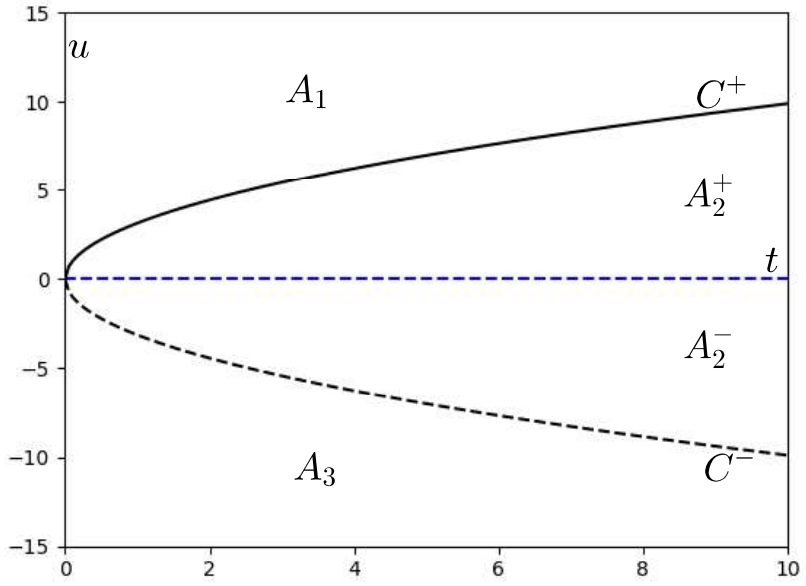



Figure 2.2: Projection of the regions $A_i$, $i = 1, 2, 3$ on the $(h, u)$-plane

- Rarefaction waves, which are smooth solutions of (2.1.1) with constant value of $a$ associated to the genuinely nonlinear fields. These waves depend only on the self-similar variable $x/t$, and satisfy:
  - Riemann invariants preservation: the Riemann invariants
    $$u + 2\sqrt{gh}, \quad u - 2\sqrt{gh} \tag{2.2.7}$$
    are constant along 1-rarefactions and 2-rarefactions, respectively.
  - Divergence of the characteristics:
    $$\lambda_i(W_l) < \lambda_i(W_r), \tag{2.2.8}$$
    where $W_l$ and $W_r$ are the left and the right states of the i-rarefaction, $i = 1, 2$.
- Shock waves, which are discontinuous solutions of (2.1.1) with constant value of $a$ associated to the genuinely nonlinear fields. These waves satisfy:
  - Rankine Hugoniot conditions:
    $$\begin{aligned} &\sigma_i[h] = [hu], \\ &\sigma_i[hu] = [u^2h + gh^2/2], \end{aligned} \tag{2.2.9}$$
    where $\sigma_i$ is the speed of the shock associated to the $\lambda_i$-field and $[w]$ represents the jump at a discontinuity of the variable $w$.

– Lax entropy conditions:

$$\lambda_i(W_l) > \sigma_i > \lambda_i(W_r). \tag{2.2.10}$$

- Stationary contact discontinuities, which are discontinuous solutions of (2.1.1) with discontinuous value of $a$, associated to the linearly degenerated field corresponding to the null eigenvalue that satisfy:

$$\begin{aligned} &[hu] = 0, \\ &[u^2/2 + g(h-a)] = 0. \end{aligned} \tag{2.2.11}$$

**Remark 2.2.1.** *The first equality in* (2.2.11) *can be understood as the preservation of the mass-flow through stationary contact discontinuities and the second one as the preservation of the mechanical energy:* $u^2/2$ *is the kinetic energy and* $g(h-a)$ *the potential one. As it has been mentioned in the previous section, the definition of weak-solution is not unique and different choices may lead to different jump conditions for these stationary discontinuities: this is the case in [14] or [141]. The jump conditions chosen by these authors imply the dissipation of mechanical energy through stationary contact discontinuities.*

Given a left-hand state $W_l$, the 1-shock $\mathcal{S}_1(W_l)$ and the 2-shock $\mathcal{S}_2(W_l)$ consisting of all right-hand states $W$ that can be connected to $W_l$ by a shock associated with $\lambda_1$ and $\lambda_2$, respectively, are:

$$\mathcal{S}_1(W_l) : u = u_l - \sqrt{\frac{g}{2}}(h-h_l)\sqrt{\frac{1}{h}+\frac{1}{h_l}}, \quad h > h_l, \tag{2.2.12}$$

$$\mathcal{S}_2(W_l) : u = u_l + \sqrt{\frac{g}{2}}(h-h_l)\sqrt{\frac{1}{h}+\frac{1}{h_l}}, \quad h < h_l. \tag{2.2.13}$$

Given a right-hand state $W_r$, the backward 1-shock curve $\mathcal{S}_1^B(W_r)$ and the backward 2-shock curve $\mathcal{S}_2^B(W_r)$, consisting of all left-hand states $W$ that can be connected to $W_r$ by a shock associated with $\lambda_1$ and $\lambda_2$, respectively, are:

$$\mathcal{S}_1^B(W_r) : u = u_r - \sqrt{\frac{g}{2}}(h-h_r)\sqrt{\frac{1}{h}+\frac{1}{h_r}}, \quad h < h_r, \tag{2.2.14}$$

$$\mathcal{S}_2^B(W_r) : u = u_r + \sqrt{\frac{g}{2}}(h-h_r)\sqrt{\frac{1}{h}+\frac{1}{h_r}}, \quad h > h_r. \tag{2.2.15}$$

Moreover, given two states $W_0$ and $W$ connected by a 1-shock wave or by a 2-shock wave, the speed of that shock will be, respectively:

$$\sigma_1(W_0, W) = u_0 - \sqrt{\frac{g}{2}\left(h + \frac{h^2}{h_0}\right)}, \tag{2.2.16}$$

$$\sigma_2(W_0, W) = u_0 + \sqrt{\frac{g}{2}\left(h + \frac{h^2}{h_0}\right)}. \tag{2.2.17}$$

The following result, whose proof can be found in [112], gives information about the sign of the speed of shocks:

**Proposition 2.2.1.**

1. *If $W_0 = (h_0, u_0, a_0) \in A_1$, then there exists $\hat{W}_0 = (\hat{h}_0, \hat{u}_0, a_0) \in \mathcal{S}_1(W_0) \cap A_2^+$ with $\hat{h}_0 > h_0$ such that:*
   - *(a) $\sigma_1(W_0, \hat{W}_0) = 0$,*
   - *(b) $\sigma_1(W_0, W) > 0$ for all $W \in \mathcal{S}_1(W_0)$ such that $h \in (h_0, \hat{h}_0)$,*
   - *(c) $\sigma_1(W_0, W) < 0$ for all $W \in \mathcal{S}_1(W_0)$ such that $h \in (\hat{h}_0, +\infty)$.*
2. *If $W_0 \in A_2 \cup A_3$, then $\sigma_1(W_0, W) < 0$ for all $W \in \mathcal{S}_1(W_0)$.*
3. *If $W_0 = (h_0, u_0, a_0) \in A_3$, then there exists $\hat{W}_0 = (\hat{h}_0, \hat{u}_0, a_0) \in \mathcal{S}_2^B(W_0) \cap A_2^-$ with $\hat{h}_0 > h_0$ such that:*
   - *(a) $\sigma_2(W_0, \hat{W}_0) = 0$,*
   - *(b) $\sigma_2(W_0, W) < 0$ for all $W \in \mathcal{S}_2^B(W_0)$ such that $h \in (h_0, \hat{h}_0)$,*
   - *(c) $\sigma_2(W_0, W) > 0$ for all $W \in \mathcal{S}_2^B(W_0)$ such that $h \in (\hat{h}_0, +\infty)$.*
4. *If $W_0 \in A_1 \cup A_2$, then $\sigma_2(W_0, W) > 0$ for all $W \in \mathcal{S}_2^B(W_0)$.*

Given a left-hand state $W_l$, the 1-rarefaction $\mathcal{R}_1(W_l)$ and the 2-rarefaction $\mathcal{R}_2(W_l)$ consisting of all right-hand states $W$ that can be connected to $W_l$ by a rarefaction associated with $\lambda_1$ and $\lambda_2$, respectively, are:

$$\mathcal{R}_1(W_l) : u = u_l - 2(\sqrt{gh} - \sqrt{gh_l}), \quad h \leq h_l, \tag{2.2.18}$$

$$\mathcal{R}_2(W_l) : u = u_l + 2(\sqrt{gh} - \sqrt{gh_l}), \quad h \geq h_l. \tag{2.2.19}$$

Given a right-hand state $W_r$, the backward 1-rarefaction curve $\mathcal{R}_1^B(W_r)$ and the backward 2-rarefaction curve $\mathcal{R}_2^B(W_r)$, consisting of all left-hand states $W$ that can be connected to $W_r$ by a rarefaction associated with $\lambda_1$ and $\lambda_2$, respectively, are:

$$\mathcal{R}_1^B(W_r) : u = u_r - 2(\sqrt{gh} - \sqrt{gh_r}), \quad h \geq h_r, \tag{2.2.20}$$

$$\mathcal{R}_2^B(W_r) : u = u_r + 2(\sqrt{gh} - \sqrt{gh_r}), \quad h \leq h_r. \tag{2.2.21}$$

Moreover, given two states $W_l$ and $W_r$ connected by a 1-rarefaction or by a 2-rarefaction, the speed of the head and the tail of these rarefactions are, respectively:

$$\begin{aligned} S_{H_1} &= \lambda_1(W_l) = u_l - \sqrt{gh_l}, \\ S_{T_1} &= \lambda_1(W_r) = u_r - \sqrt{gh_r}, \end{aligned} \tag{2.2.22}$$

$$\begin{aligned} S_{H_2} &= \lambda_2(W_r) = u_r + \sqrt{gh_r}, \\ S_{T_2} &= \lambda_2(W_l) = u_l + \sqrt{gh_l}. \end{aligned} \tag{2.2.23}$$

Using again the notation in [112], we define the curves:

$$\mathcal{W}_1(W_l) = \mathcal{S}_1(W_l) \cup \mathcal{R}_1(W_l), \tag{2.2.24}$$
$$\mathcal{W}_1^B(W_r) = \mathcal{S}_1^B(W_r) \cup \mathcal{R}_1^B(W_r), \tag{2.2.25}$$
$$\mathcal{W}_2(W_l) = \mathcal{S}_2(W_l) \cup \mathcal{R}_2(W_l), \tag{2.2.26}$$
$$\mathcal{W}_2^B(W_r) = \mathcal{S}_2^B(W_r) \cup \mathcal{R}_2^B(W_r). \tag{2.2.27}$$

The four curves can be parameterized in the form $u = u(h), h > 0$, where the function is:

- strictly convex and strictly decreasing for $\mathcal{W}_1(W_l)$ and $\mathcal{W}_1^B(W_l)$;
- strictly concave and strictly increasing for $\mathcal{W}_2(W_l)$ and $\mathcal{W}_2^B(W_l)$.

An example of $\mathcal{W}_1(W_l)$ and $\mathcal{W}_2(W_l)$ is shown in Figure 2.3 and another of $\mathcal{W}_1^B(W_r)$ and $\mathcal{W}_2^B(W_r)$ in Figure 2.4.

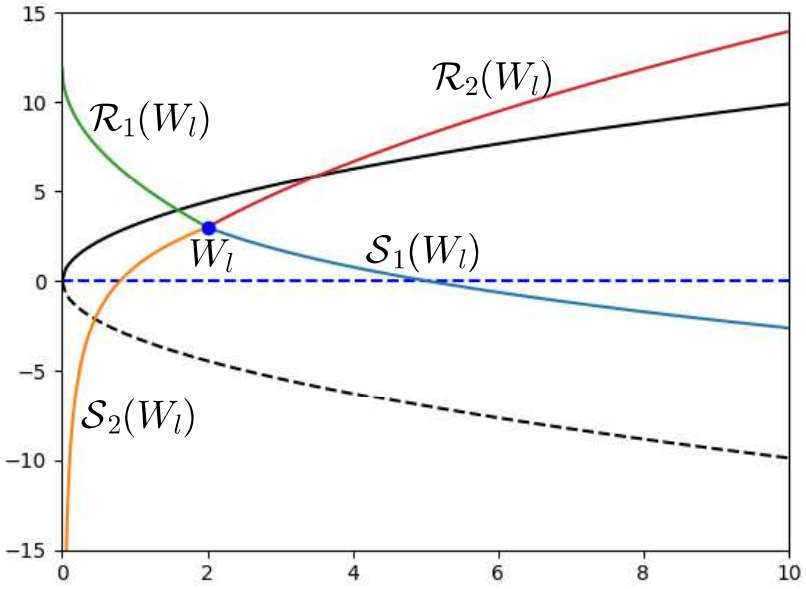



Figure 2.3: $\mathcal{W}_1$ and $\mathcal{W}_2$ curves for a state $W_l$

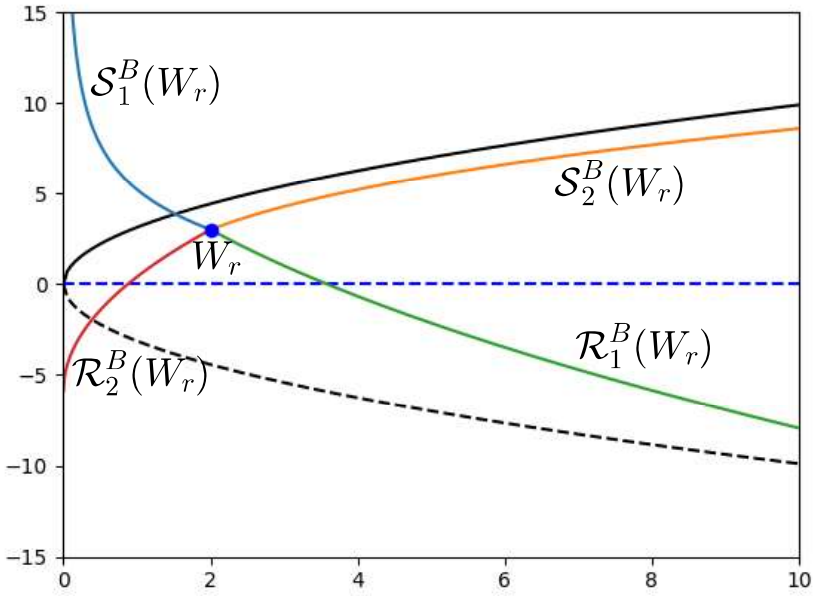



Figure 2.4: $\mathcal{W}_1^B$ and $\mathcal{W}_2^B$ curves for a state $W_r$

Let us consider now the set $\mathcal{W}_3(W_0)$ of all the states $W$ that can be connected to $W_0$ by an admissible 3-stationary wave. Using the jump conditions (2.2.11), we obtain:

$$(h,u,a) \in \mathcal{W}_3(W_0) \quad \Longrightarrow \quad u = \frac{u_0 h_0}{h}, \quad \varphi(h) = 0, \tag{2.2.28}$$

where $\varphi : (0,\infty) \to \mathbb{R}$ is given by:

$$\varphi(h) = a_0 - a + \frac{u_0^2}{2g}\left(\frac{h_0^2}{h^2} - 1\right) + h - h_0. \tag{2.2.29}$$

Therefore, given $a$ and $a_0$, in order to find the states that can be linked through a 3-stationary wave to a state $W_0$, one has to look for the roots of the function $\varphi$. The following results hold (see [112]):

**Lemma 2.2.1.** *Suppose that $W_0 = (h_0, u_0, a_0)$ and $a$ are given with $u_0 \neq 0$. Let us define:*

$$h_{min}(W_0) = \left(\frac{u_0^2 h_0^2}{g}\right)^{\frac{1}{3}}, \tag{2.2.30}$$

$$a_{min}(W_0) = a_0 + \frac{u_0^2}{2g}\left(\frac{h_0^2}{h_{min}^2} - 1\right) + h_{min} - h_0. \tag{2.2.31}$$

*Then*

- *if* $a > a_{min}$ *the function* $\varphi$ *has two roots* $h_*(W_0), h^*(W_0)$ *with* $h_* \leq h_{min} \leq h^*$;
- *if* $a = a_{min}$ *it has only one root* $h_{min}$;
- *if* $a < a_{min}$ *it has no roots.*

**Proposition 2.2.2.** *Given a left-hand state* $W_0 = (h_0, u_0, a_0)$ *and a right-hand bottom level* $a$*:*

1. *If* $u_0 \neq 0$ *and* $a > a_{min}(W_0)$*, there are two distinct right-hand states* $W_* = (h_*, u_*, a)$ *and* $W^* = (h^*, u^*, a)$ *that can be connected to* $W_0$ *by a 3-stationary wave. Here,* $h_*, h^*$ *are the roots of* $\varphi$ *and*
$$u_*^* = \frac{h_0 u_0}{h_*^*}.$$
*Moreover,* $W_*$ *is subcritical and* $W^*$ *is supercritical.*

2. *If* $u_0 \neq 0$ *and* $a = a_{min}(W_0)$*, there is only a state that can be connected to* $W_0$ *by a 3-stationary wave:* $(h_{min}, u_{min}, a_{min})$ *with*
$$u_{min} = \frac{u_0 h_0}{h_{min}}.$$
*This state is critical.*

3. *If* $u_0 \neq 0$ *and* $a < a_{min}(W_0)$*, there is no stationary wave from* $W_0$ *to a state with level* $a$*.*

4. *If* $u_0 = 0$ *and* $a \geq a_0 - h_0$*, there is only a state that can be connected to* $W_0$ *by a 3-stationary wave:*
$$u = u_0 = 0, \quad h = h_0 + a - a_0.$$
*This state is subcritical.*

**Remark 2.2.2.** *According to Remark 2.2.1, items 3 and 4 of the proposition can be understood as follows: if* $u_0 \neq 0$ *and* $a < a_{min}(W_0)$ *or if* $u_0 = 0$ *and* $a < a_0 - h_0$*, the mechanical energy of the water is not enough to go up the step.*

Following [112] not all the possible stationary waves are considered to be admissible. The following criterion is imposed:

**Monotonicity criterion** (MC): *Along* $\mathcal{W}_3(W_0)$*, the bottom level* $a$ *is a monotone function of* $h$*.*

Accordingly, the 3-stationary waves connecting a supercritical and a subcritical state are not admissible.

Given a state $W_0 = (h_0, u_0, a_0) \in A_i, i = 1, 2, 3$, and a level $a \geq a_{min}(W_0)$, according to (MC) there is only one state that can be connected to $W_0$ with an admissible 3-stationary

wave: as in [112], it will be represented by $SW(W, a)$. And given a critical state $W_0 \in C$ and a level $a > a_0$ there are two states that can be connected to $W_0$ with an admissible 3-stationary wave: we will represent by $SW_{sup}(W_0, a)$ (resp. $SW_{sub}(W_0, a)$) the supercritical (resp. subcritical) one.

The following notation will be used to represent the structure of the solution of the Riemann problems: the symbol

$$\mathcal{W}_{i_1}(W_0, W_1) \oplus \cdots \oplus \mathcal{W}_{i_k}(W_{k-1}, W_k)$$

will indicate that the solution of the Riemann problems is composed by $k$ simple waves: $W_0$ is the left state; $W_k$, the right state; and $W_j$, $j = 1, \ldots, k-1$ the intermediate states. Finally, the indexes $i_j \in \{1, 2, 3\}$ indicate the characteristic field to which the $j$th wave is associated, i.e.

$$W_{j+1} \in \mathcal{W}_{i_j}(W_j), \quad j = 0, \ldots, k-1.$$

Moreover, in the case of 1 and 2-waves, the type of simple wave (rarefaction or shock) will be specified by replacing $\mathcal{W}_i$ by $\mathcal{R}_i$ or $\mathcal{S}_i$.

## 2.3 The wet-dry Riemann Problem

We consider the shallow water system (2.1.1) with initial conditions:

$$W(x,0) = \begin{cases} W_l = (h_l, u_l, a_l) & \text{if } x < 0, \\ W_r \in \{(0, u, a_r) : \ u \in \mathbb{R}\} & \text{if } x > 0, \end{cases} \tag{2.3.1}$$

or

$$W(x,0) = \begin{cases} W_l \in \{(0, u, a_l) : \ u \in \mathbb{R}\} & \text{if } x < 0, \\ W_r = (h_r, u_r, a_r) & \text{if } x > 0, \end{cases} \tag{2.3.2}$$

with $h_l, h_r > 0$. These initial conditions correspond to a situation in which there is a step at $x = 0$, the bottom is flat to the left and to the right of the step, and there is no water to the left or to the right of the step: see Figure 2.5.

Although from the physical point of view the value of $u$ at a dry state is meaningless, from the mathematical point of view, a dry state can be represented by any point of the plane $h = 0$ in the $(h, u, a)$-space. From this point of view, the considered problem is a Partial Riemann Problem (see [63]), as only the belonging to a given set is imposed at the right or at the left of $x = 0$. Moreover, as it will be seen, the value of $u$ at the side that initially is dry will be the limit of the velocity at the wet-dry front.

In the case of a flat bottom (i.e. if $a_l = a_r$) the following result holds (see [153]):

**Theorem 2.3.1.** *Let us suppose that $a_l = a_r$. Then, the partial Riemann problem corresponding to the homogeneous shallow water system with initial conditions* (2.3.1) *(resp.* (2.3.2)*) has a unique solution consisting of a 1-rarefaction (resp. a 2-rarefaction) connecting $W_l$ (resp. $W_r$) to vacuum.*

Without loss of generality we will consider that $a_l < a_r$, i.e. the right side of the step is deeper than the left one.

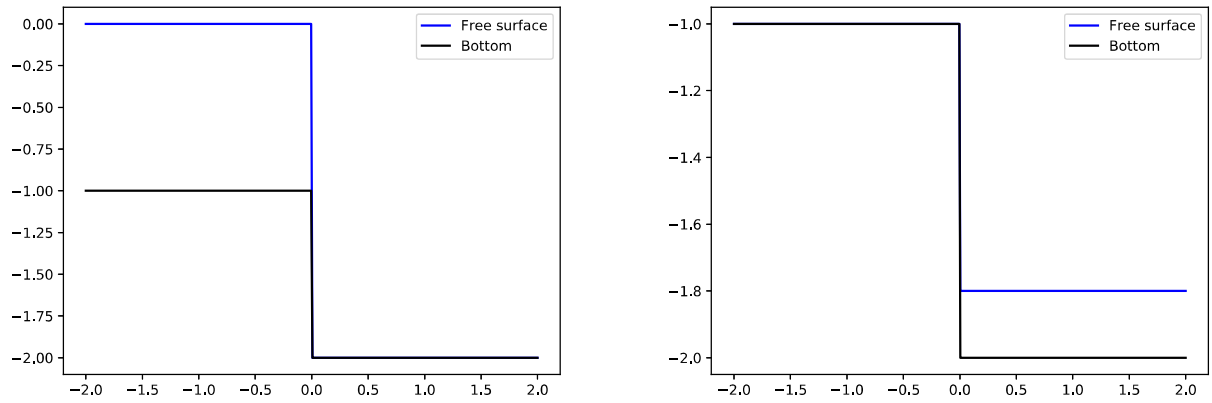



Figure 2.5: Left: initial condition of the form (2.3.1). Right: initial condition of the form (2.3.2).

### 2.3.1 Case 1: initial condition (2.3.1)

To find a solution of the Riemann problem, we have to connect the state $W_l$ to a state $W_r$ belonging to the line $\{h = 0\}$ of the plane $a = a_r$ through admissible simple waves. The following result holds:

**Theorem 2.3.2.** *Suppose that the Riemann problem with initial condition* (2.3.1) *has a self-similar solution $W$ composed by admissible simple waves. Let us denote by $W_0^-$ and $W_0^+$ the limits of $W$ to the left and to the right of $x = 0$. Then, necessarily $W_0^+ \in A_1$ and $W_0^- \in A_1 \cup C^+$.*

*Proof.* Since the source term vanishes at $(0, \infty)$, the function

$$V(x,t) = \begin{cases} W_0^+ & \text{if } x < 0, \\ W(x,t) & \text{if } x > 0, \end{cases}$$

has to be a solution of the homogeneous shallow water system. Moreover, $V$ has to be the self-similar solution of the Partial Riemann problem linking $W_0^+$ to vacuum. Then, due to Theorem 2.3.2, $V$ necessarily consists of a 1-rarefaction linking $W_0^+$ to the line $h = 0$. If $W_0^+ \in A_2 \cup A_3 \cup C^-$, the head of the 1-rarefaction $S_{H_1}$ would be negative, but then the 1-rarefaction could not follow the stationary wave linking $W_0^-$ and $W_0^+$ in the solution of the Riemann problem. Therefore $W_0^+$ belongs to $C^+ \cup A_1$.

The stationary wave linking $W_0^-$ and $W_0^+$ can be a 3-stationary wave or the composition of 3-stationary wave and stationary shocks. Due to Theorem 2.2.1, since $W_0^+ \in C^+ \cup A_1$, it cannot be the right state of a stationary shock in $a = a_r$ (it should belong to $A_2^+$ to be the right state of a 1-shock or to $A_3$ to be the right state of a 2-shock). Therefore, $W_0^+$ has to be linked to a state $W_0^*$ in $a = a_l$ through a 3-stationary wave. Due to (MC) the

only possibility for having an admissible 3-stationary wave is $W_0^* \in A_1 \cup C^+$and $W_0^+ \in A_1$. Again, $W_0^*$ cannot be the right state of a stationary shock in $a = a_l$. Therefore $W_0^* = W_0^-$ and thus $W_0^- \in A_1 \cup C^+$ and $W_0^+ \in A_1$, as we wanted to prove.

□

We are going to show that, if $a_l < a_r$, the Partial Riemann problem with initial condition (2.3.1) has always one solution. To construct it, let us consider the following three regions of the plane $a = a_l$ (see Fig 2.6):

- **Region I**: $A_1$.
- **Region II**: $A_2 \cup C^- \cup A_3 \cap \{(h,u) : \ u > -2\sqrt{gh}\}$.
- **Region III**: $\{(h,u) : \ u < -2\sqrt{gh}\}$.

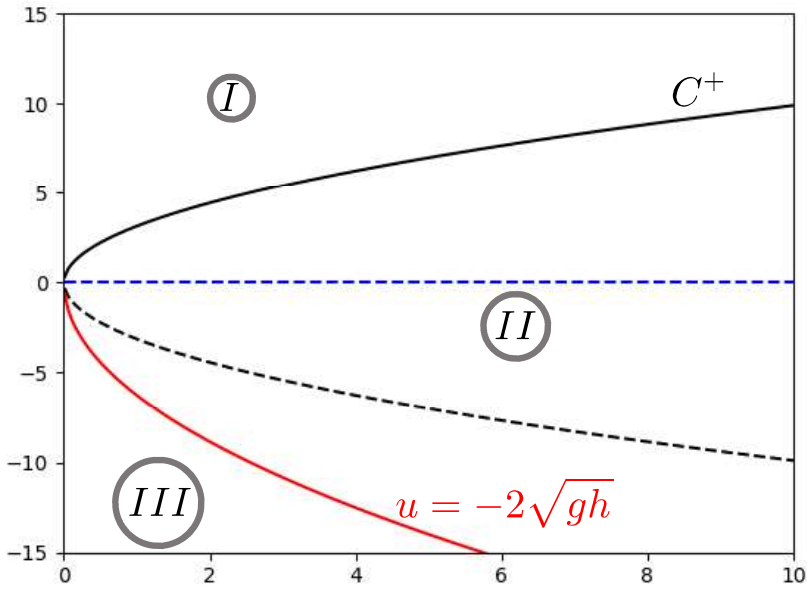



Figure 2.6: Regions of the plane $a = a_l$ for the partial Riemann problem with initial conditions (2.3.1).

Let us study the Riemann problem with initial condition (2.3.1) for a left state $W_l$ belonging to any of these 3 regions and their boundaries.

- **Region I**: Given a state $W_l \in A_1$, we first consider the 3-stationary wave that connects $W_l$ to $W_0^+ = SW(W_l, a_r) \in A_1$ and then the 1-rarefaction connecting $W_0^+$ to a state $W_r$ belonging to the line $h = 0$. Since $W_l, W_0^+ \in A_1$, both the head and

the tail of the 1-rarefaction are positive, so that it can follow the stationary wave. In conclusion, we have found a solution of the form:

$$\mathcal{W}_3(W_l, W_0^+) \oplus \mathcal{R}_1(W_0^+, W_r). \tag{2.3.3}$$

Figure 2.7 shows the projection of the intermediate states and the simple waves on the $(h, u)$-plane and a sketch of the free surface at a time $t > 0$.

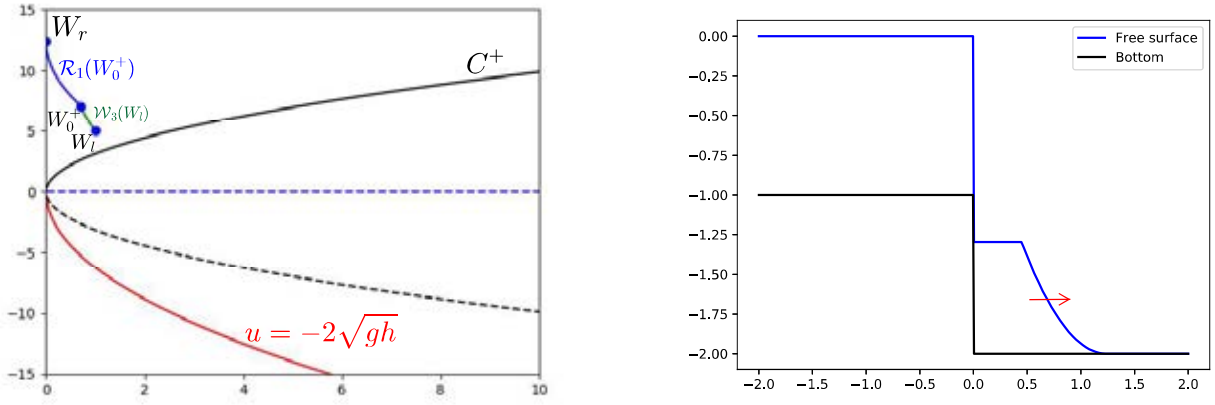



Figure 2.7: Solution in Region I. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

- **Region II**: Given a state $W_l$ in region II, we first consider the state $W_0^- \in \mathcal{R}_1(W_l) \cap C^+$. This state has to satisfy:

$$\begin{cases} u_0^- = \sqrt{gh_0^-}, \\ u_0^- = u_l - 2\left(\sqrt{gh_0^-} - \sqrt{gh_l}\right). \end{cases} \tag{2.3.4}$$

Some easy computations allow us to solve this system:

$$h_0^- = \frac{1}{g}\left(\frac{u_l + 2\sqrt{gh_l}}{3}\right)^2, \quad u_0^- = \frac{u_l + 2\sqrt{gh_l}}{3}. \tag{2.3.5}$$

We consider then the 1-rarefaction linking $W_l$ to $W_0^-$, followed of the 3-stationary wave linking this latter state to $W_0^+ = SW_{sup}(W_0^-, a_r) \in A_1$. Finally $W_0^+$ is connected to a state $W_r$ in the line $h = 0$ through a 1-rarefaction: see Figure 2.8. Since $W_l \in A_2 \cup A_3$ and $W_0^- \in C^+$, the head of the first 1-rarefaction is negative and the tail is 0, so that it can be followed by the 3-stationary wave. In turns, since

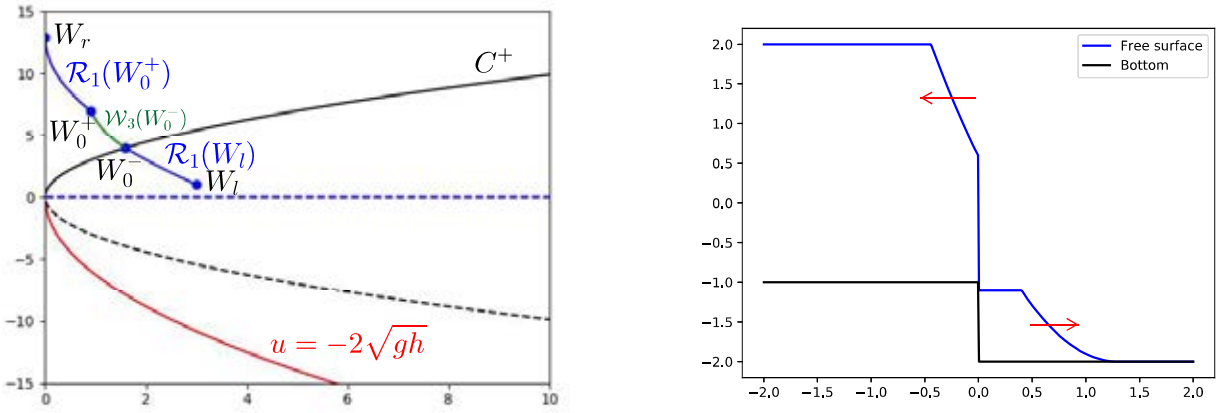



Figure 2.8: Solution in Region II. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

$W_0^+ \in A_1$ and $W_r \in A_1$, the second 1-rarefaction has positive head and tail, so that it can follows the 3-stationary wave. We obtain thus a solution whose structure is

$$\mathcal{R}_1(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_0^+) \oplus \mathcal{R}_1(W_0^+, W_r). \tag{2.3.6}$$

- **Region III**: A state $W_l$ in Region III can be linked through a 1-rarefaction to the state $W_r = (0, u_r, a_l)$, where $u_r = u_l + 2\sqrt{gh_l}$: see Figure 2.9. The wet-dry front travels at the speed $u_r < 0$, so that at time $t > 0$ there is vacuum in $x > u_r t$. The structure of the solution is thus

$$\mathcal{R}_1(W_l, W_r). \tag{2.3.7}$$

- **Boundary between Region I and Region II**: For states $W_l \in C^+$ a solution can be constructed like in Region I, i.e. (2.3.3) with $W_0^+ = SW_{sup}(W_l, a_r)$.
- **Boundary between Region II and Region III**: For states such that $\{u_l = -2\sqrt{gh_l}\}$ a solution can be constructed like in Region III, i.e. (2.3.7). In this case, there is vacuum at $x > 0$ for every $t > 0$.

### 2.3.2 Case 2: initial condition (2.3.2)

To find a solution of the Riemann problem, we have to connect the state $W_r$ to a state $W_l$ belonging to the line $\{h = 0\}$ of the plane $a = a_l$ through admissible simple waves. The following result holds:

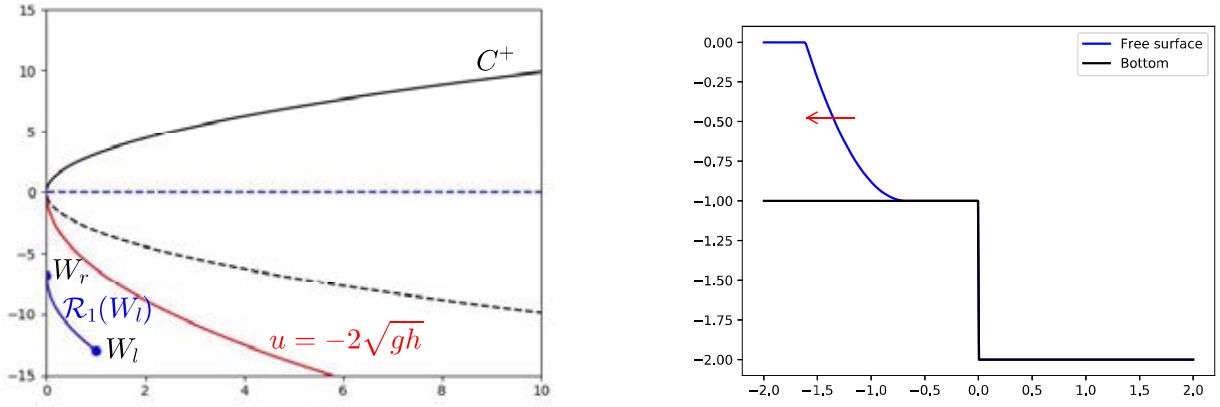



Figure 2.9: Solution in Region III. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

**Theorem 2.3.3.** *Suppose that the Riemann problem with initial condition* (2.3.2) *has a self-similar solution $W$ composed by admissible simple waves. Let us denote by $W_0^-$ and $W_0^+$ the limits of $W$ to the left and to the right of $x = 0$. Then, necessarily $W_0^- \in A_3 \cup C^-$ and $W_0^+ \in A_3 \cup SW_{sub}(C^-, a_r)$.*

*Proof.* Similar to the proof of Theorem 2.3.2. □

We are going to show that, if $a_l < a_r$, the Partial Riemann problem with initial condition (2.3.2) may have zero, one, or two solutions. To see it, let us consider the following six regions of the plane $a = a_r$ (see Fig 2.10):

- **Region I**: $A_1 \cap \{(h, u) : \ u > 2\sqrt{gh}\}$.
- **Region II**: States with $u > 0$ that are between the curves $\{(h, u) : \ u = 2\sqrt{gh}\}$ and $\mathcal{W}_2(W_{rest})$, where
$$W_{rest} = (a_r - a_l, 0, a_r)$$
is the state corresponding to a situation of water at rest with $\eta_r = -a_l$.
- **Region III**: States between the curves $SW_{sub}(C^-, a_r)$ and $\mathcal{W}_2(W_{rest})$.
- **Region IV**: States with $u < 0$ that are between the curves $\mathcal{W}_2(W_{rest})$ and $SW_{sup}(C^-, a_r)$.
- **Region V**: States between the curves $SW_{sub}(C^-, a_r)$, $SW_{sup}(C^-, a_r)$, and $\mathcal{W}_2(W_{rest})$.
- **Region VI**: States below the curve $SW_{sup}(C^-, a_r)$.

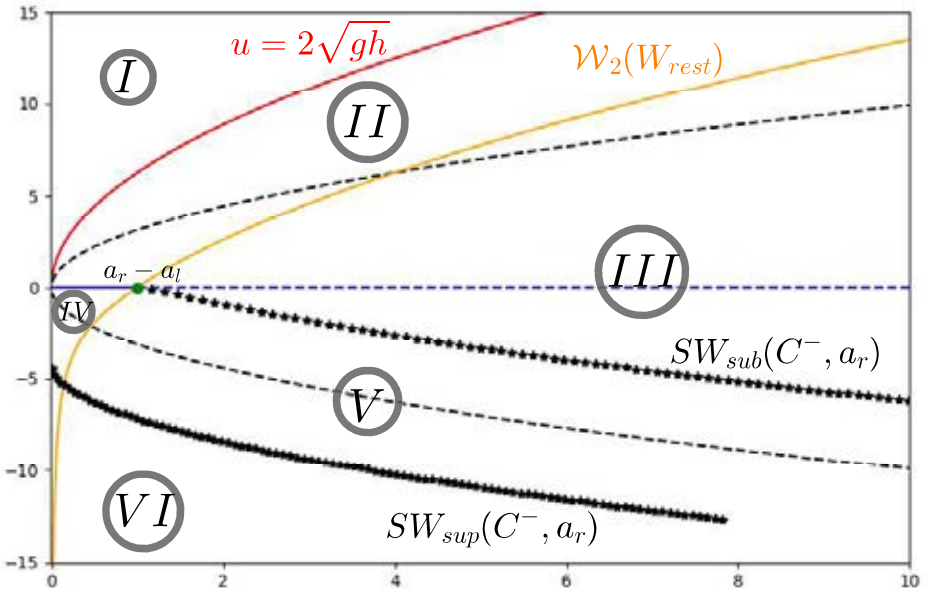



Figure 2.10: Regions of the plane $a = a_r$ for the partial Riemann problem with initial conditions (2.3.2).

**Remark 2.3.1.** *The mechanical energy of the states belonging to Regions IV and V is not enough to go up the step (see Remark 2.2.2).*

Let us study the Riemann problem with initial condition (2.3.2) for a right state belonging to any of these 6 regions and their boundaries.

- **Region I**: A state $W_r$ in Region I can be linked through a 2-rarefaction to the state $W_l = (0, u_l, a_r)$, where $u_l = u_r - 2\sqrt{gh_r}$: see Figure 2.11. The wet-dry front travels at the speed $u_l > 0$, so that at time $t > 0$ there is vacuum in $x < u_l t$. The structure of the solution is thus
$$\mathcal{R}_2(W_l, W_r). \tag{2.3.8}$$

- **Region III**: Let us assume that the curves $\mathcal{R}_2^B(W_r)$ and $SW_{sub}(C^-, a_r)$ intersect at one point $W_0^+$: see Figure 2.12. To construct a solution of the Riemann problem, we consider the 2-rarefaction linking $W_r$ to $W_0^+$ followed by the 3-stationary wave linking this latter state to $W_0^- = SW(W_0^+, a_l)$. Finally $W_0^-$ is connected to a state $W_l$ in $h = 0$ through a 2-rarefaction. Since $W_r \in A_1 \cup A_2$ and $W_0^+ \in A_2^-$, the head and tail of the first 2-rarefaction are positive, so that it can follow the 3-stationary wave. In turns, since $W_0^- \in C^-$ and $W_l \in A_3$, the second 2-rarefaction has negative

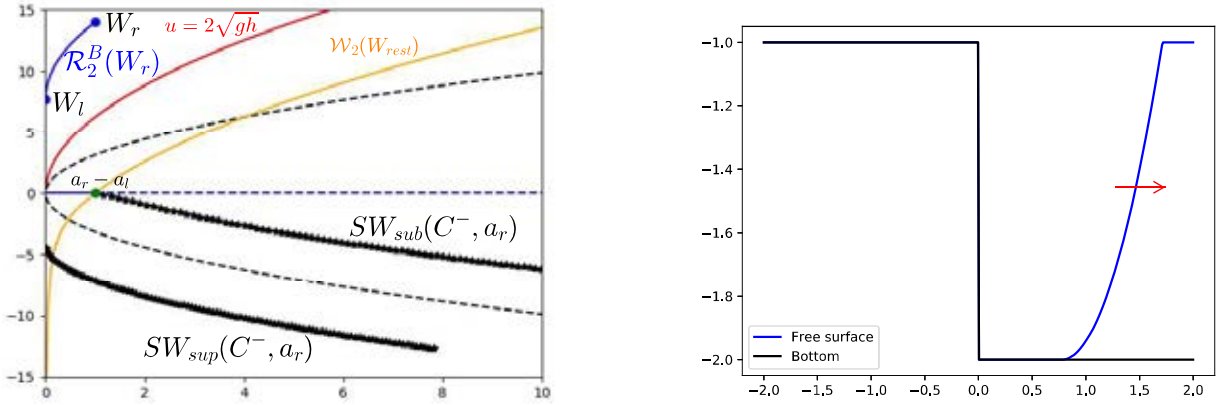



Figure 2.11: Solution for states in region I. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

head and the tail is equal to 0, so that it can be followed by the 3-stationary wave. We obtain thus a solution whose structure is

$$\mathcal{R}_2(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_0^+) \oplus \mathcal{R}_2(W_0^+, W_r). \tag{2.3.9}$$

See Figure 2.12.

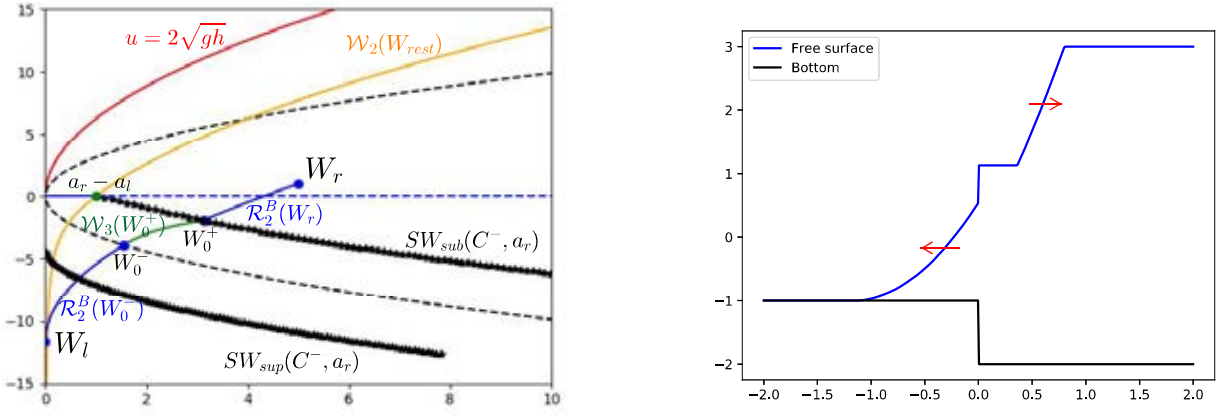



Figure 2.12: Solution of a state in Region III. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

To finish, let us check that $\mathcal{R}_2^B(W_r)$ and $SW_{sub}(C^-, a_r)$ intersects at one point:

**Proposition 2.3.1.** *Suppose that $a_l < a_r$ and $W_r$ belong to Region III. Then, the intersection of the curves $\mathcal{R}_2^B(W_r)$ and $SW_{sub}(C^-, a_r)$ is non-empty and consists of only one state $W_0^+$.*

*Proof.* A state $W_0^+ \in \mathcal{R}_2^B(W_r) \cap SW_{sub}(C^-, a_r)$ has to satisfy:

$$\begin{cases} u_0^+ = u_r + 2\left(\sqrt{gh_0^+} - \sqrt{gh_r}\right), \\ a_l = a_{min}(W_0^+) = a_r + \dfrac{(u_0^+)^2}{2g}\left(\dfrac{(h_0^+)^2}{h_{min}(W_0^+)^2} - 1\right) + h_{min}(W_0^+) - h_0^+, \end{cases} \tag{2.3.10}$$

where:

$$h_{min}(W_0^+) = \left(\frac{(u_0^+)^2(h_0^+)^2}{g}\right)^{\frac{1}{3}}.$$

Therefore $h_0^+$ has to be a root of the function

$$\psi(h) = a_r - a_l + \frac{3}{2}\left(\frac{u(h)^2h^2}{g}\right)^{\frac{1}{3}} - \frac{u(h)^2}{2g} - h, \tag{2.3.11}$$

where:

$$u(h) = u_r + 2(\sqrt{gh} - \sqrt{gh_r}).$$

Let us study the roots of $\psi$. To do this, we consider first the state $W_c^r \in \mathcal{R}_2^B(W_r) \cap C^-$. This state has to satisfy:

$$\begin{cases} u_c^r = -\sqrt{gh_c^r}, \\ u_c^r = u_r + 2(\sqrt{gh_c^r} - \sqrt{gh_r}). \end{cases} \tag{2.3.12}$$

Some easy computations show that:

$$h_c^r = \frac{1}{9g}(u_r - 2\sqrt{gh_r})^2, \quad u_c^r = -\frac{1}{3}(2\sqrt{gh_r} - u_r). \tag{2.3.13}$$

Next, we consider the state $W_0^r = (h_0^r, 0, a_r) \in (\mathcal{R}_2(W_r) \cup \mathcal{R}_2^B(W_r)) \cap \{u = 0\}$. This state has to satisfy:

$$\begin{cases} u_0^r = 0, \\ u_0^r = u_r + 2(\sqrt{gh_0^r} - \sqrt{gh_r}). \end{cases} \tag{2.3.14}$$

Some easy computations show that:

$$h_0^r = \frac{\left(\sqrt{gh_r} - \dfrac{u_r}{2}\right)^2}{g}. \tag{2.3.15}$$

Observe that, in the interval $[h_c^r, h_0^r]$, the curve $\mathcal{R}_2^B(W_r)$ travels from $C^-$ to $u = 0$. Therefore, if there is an intersection with $SW_{sub}(C^-, a_r)$, the corresponding value of $h$ has to be in this interval. Now, since $W_c^r \in C^-$, it verifies:

$$a_{min}(W_c^r) = a_r. \tag{2.3.16}$$

and, as a consequence:

$$\psi(h_c^r) = a_r - a_l > 0. \tag{2.3.17}$$

On the other hand, since $W_r$ is in Region III, $\mathcal{R}_2^B(W_r)$ intersects the line $u = 0$ at the right of the state $W_{rest}$ and thus:

$$\psi(h_0^r) = a_r - a_l - h_0^r < h_0^r - h_0^r = 0. \tag{2.3.18}$$

Therefore, there is at least one root in $[h_c^r, h_0^r]$.

Let us see that there is only one. First, we rewrite $\psi$ as $\psi(h) = f(h) - p(h)$ with

$$f(h) = a_r - a_l + \frac{3}{2}\left(\frac{u(h)^2 h^2}{g}\right)^{\frac{1}{3}}, \tag{2.3.19}$$

and:

$$p(h) = \frac{u(h)^2}{2g} + h. \tag{2.3.20}$$

The derivatives of these two functions are:

$$f'(h) = \frac{1}{g^{\frac{1}{3}}} \frac{u'(h)h + u(h)}{u(h)^{\frac{1}{3}} h^{\frac{1}{3}}}, \tag{2.3.21}$$

$$p'(h) = \frac{u(h)u'(h)}{g} + 1, \tag{2.3.22}$$

where $u'(h) = \sqrt{\dfrac{g}{h}}$ is the derivative of $u(h)$. The denominator of $f'$ is zero when $u(h) = 0$ and this only happens when $h = h_0^r$. Therefore, in $(h_c^r, h_0^r)$ the function is differentiable and its derivative vanishes if $-\sqrt{gh} = u(h)$, what only happens for $h = h_c^r$. So, the sign of $f'$ has to be constant in $(h_c^r, h_0^r)$. It is easy to see that:

$$\lim_{h \to (h_0^r)^-} f'(h) = -\infty, \tag{2.3.23}$$

so that $f'$ is negative in $(h_c^r, h_0^r)$, i.e., $f$ is strictly decreasing in this interval.
On the other hand, $p'(h)$ only vanishes at $h_c$, so that the sign of $p'$ remains constant in $(h_c^r, h_0^r)$. It is easy to see that:

$$p'(h_0^r) = 1 > 0, \tag{2.3.24}$$

so that $p'$ is positive in $(h_c^r, h_0^r)$, i.e., $p$ is strictly increasing in the interval.
Since $f$ is strictly decreasing and $p$ is strictly increasing in $(h_c^r, h_0^r)$, there is only a root of $\psi$ in this interval, as we wanted to prove. □

- **Region V**: Let us assume that the curves $\mathcal{S}_2^B(W_r)$ and $SW_{sub}(C^-, a_r)$ intersect at one point $W_0^+$: see Figure 2.13. To construct the solution of the Riemann problem, we consider the 2-shock linking $W_r$ to $W_0^+$ followed by the 3-stationary wave linking this latter state to $W_0^- = SW(W_0^+, a_l)$. Finally $W_0^-$ is connected to a state $W_l$ in $h = 0$ through a 2-rarefaction: see Figure 2.13. Since $W_0^- \in C^-$ and $W_l \in A_3$, the 2-rarefaction has negative head and the tail is 0, so that it can be followed by the 3-stationary wave. We obtain thus a solution whose structure is

$$\mathcal{R}_2(W_0, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_0^+) \oplus \mathcal{S}_2(W_0^+, W_r), \tag{2.3.25}$$

*provided* that the speed of the 2-shock linking $W_r$ to $W_0^+$ is positive, so that the shock can follow the 3-stationary wave. This is true for the states of Region V belonging to $A_2 \cup C^-$ because of Proposition 2.2.1 but it has to be proved for those belonging to $A_3$. Let us prove first that $W_0^+$ exists and is unique:

**Proposition 2.3.2.** *Suppose that $a_l < a_r$ and $W_r$ in Region V or Region VI. Then, the intersection of the curves $\mathcal{S}_2^B(W_r)$ and $SW_{sub}(C^-, a_r)$ is non-empty and consists of only one state $W_0^+$.*

*Proof.* A state $W_0^+ \in \mathcal{S}_2^B(W_r) \cap SW_{sub}(C^-, a_r)$ has to satisfy:

$$\begin{cases} u_0^+ = u_r + \sqrt{\dfrac{g}{2}}(h_0^+ - h_r)\sqrt{\dfrac{1}{h_0^+} + \dfrac{1}{h_r}}, \\ \\ a_l = a_{min}(W_0^+) = a_r + \dfrac{(u_0^+)^2}{2g}\left(\dfrac{(h_0^+)^2}{h_{min}(W_0^+)^2} - 1\right) + h_{min}(W_0^+) - h_0^+. \end{cases} \tag{2.3.26}$$

Then, $h_0^+$ has to be a root of the function

$$\phi(h) = a_r - a_l + \frac{3}{2}\left(\frac{u(h)^2 h^2}{g}\right)^{\frac{1}{3}} - \frac{u(h)^2}{2g} - h, \tag{2.3.27}$$

where:

$$u(h) = u_r + \sqrt{\frac{g}{2}}(h - h_r)\sqrt{\frac{1}{h} + \frac{1}{h_r}}.$$

From this point, the proof is similar to that of Theorem 2.3.1, replacing $W_0^r$ and $W_c^r$ by $W_0^s \in \mathcal{S}_2^B(W_r) \cap \{u = 0\}$ and $W_c^s \in (\mathcal{S}_2(W_r) \cup \mathcal{S}_2^B(W_r)) \cap C^-$. □

The positiveness of the speed of the 2-shock linking $W_r$ to $W_0^+$ for states of Region V belonging to $A_3$ is a consequence of the following result:

**Proposition 2.3.3.** *Suppose $a_l < a_r$ and that the states $W_r = (h_r, u_r, a_r) \in A_3$ and $\hat{W}_0 = (\hat{h}_0, \hat{u}_0, a_r) \in \mathcal{S}_2^B(W_r)$ can be linked by a stationary shock. Then $a_{min}(W_r) < a_{min}(\hat{W}_0)$.*

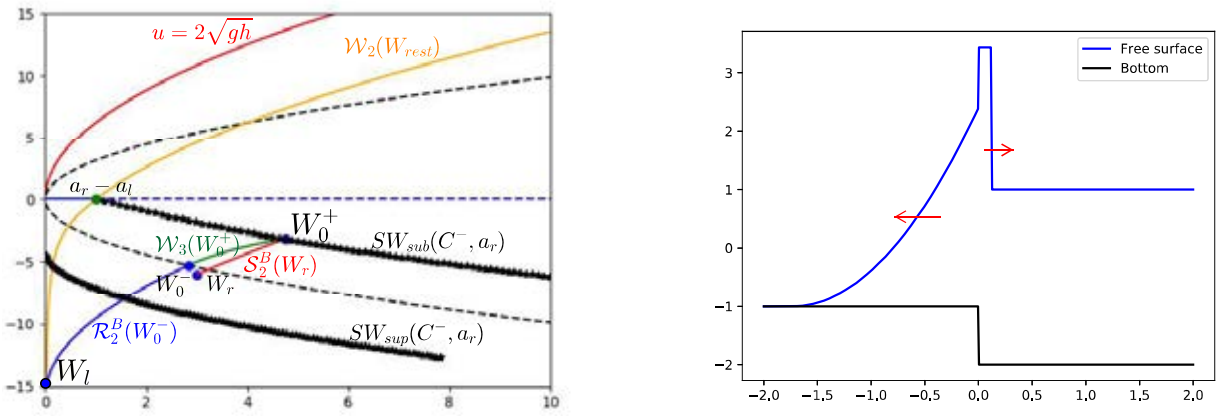



Figure 2.13: Solution of a state in Region V. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

*Proof.* Since the shock linking $\hat{W}_0$ and $W_r$ is stationary, we can express $\hat{W}_0$ in terms of $W_r$:

$$\hat{h}_0 = \frac{h_r}{2}\left(-1 + \sqrt{1 + \frac{8u_r^2}{gh_r}}\right), \tag{2.3.28}$$

$$\hat{u}_0 = \frac{u_r h_r}{\hat{h}_0}. \tag{2.3.29}$$

Let's see that $a_{min}(W_r) < a_{min}(\hat{W}_0)$ using (2.3.28), (2.3.29) and $h_r < \hat{h}_0$:

$$\begin{aligned}
& a_{min}(W_r) < a_{min}(\hat{W}_0) \Leftrightarrow \\
\Leftrightarrow\ & a_r + \frac{3}{2}\left(\frac{u_r^2 h_r^2}{g}\right)^{\frac{1}{3}} - \frac{u_r^2}{2g} - h_r < a_r + \frac{3}{2}\left(\frac{(\hat{u}_0)^2(\hat{h}_0)^2}{g}\right)^{\frac{1}{3}} - \frac{(\hat{u}_0)^2}{2g} - \hat{h}_0 \\
\Leftrightarrow\ & \frac{3}{2}\left(\frac{u_r^2 h_r^2}{g}\right)^{\frac{1}{3}} - \frac{u_r^2}{2g} - h_r < \frac{3}{2}\left(\frac{u_r^2 h_r^2}{g}\right)^{\frac{1}{3}} - \frac{(\hat{u}_0)^2}{2g} - \hat{h}_0 \\
\Leftrightarrow\ & -\frac{u_r^2}{2g} - h_r < -\frac{(\hat{u}_0)^2}{2g} - \hat{h}_0 \\
\Leftrightarrow\ & \frac{u_r^2}{2}(\hat{h}_0 + h_r) > g(\hat{h}_0)^2 \\
\Leftrightarrow\ & \frac{u_r^2 h_r}{4}\left(1 + \sqrt{1 + \frac{8u_r^2}{gh_r}}\right) > g\frac{h_r^2}{4}\left(2 + \frac{8u_r^2}{gh_r} - 2\sqrt{1 + \frac{8u_r^2}{gh_r}}\right) \\
\Leftrightarrow\ & \frac{u_r^2}{gh_r}\left(1 + \sqrt{1 + \frac{8u_r^2}{gh_r}}\right) > 2\left(1 + \frac{4u_r^2}{gh_r} - \sqrt{1 + \frac{8u_r^2}{gh_r}}\right).
\end{aligned}$$

If we define $z$ by

$$z^2 = 1 + \frac{8u_r^2}{gh_r},$$

we must prove:

$$\frac{1}{8}(z^2-1)(z+1) > 2\left(1 + \frac{1}{2}(z^2-1) - z\right)$$

what is equivalent to

$$(z-1)(z-3)^2 > 0$$

and this is true, since $z > 1$: $W_r \in A_3$, so $u_r \neq 0$. □

**Corollary 2.3.1.** *Suppose that $a_l < a_r$ and that $W_r \in A_3$ belongs to Region V. Then, the 2-shock linking $W_r$ to the state $W_0^+$ given by Proposition 2.3.2 has positive speed.*

*Proof.* Observe that, in the half-plane $u \leq 0$ of the plane $a = a_r$, the states such that

$$a_{min}(W) = a_l$$

are those belonging to the curves $SW_{sub}(C^-, a_r)$ and $SW_{sup}(C^-, a_r)$. On the other hand, since

$$a_{min}(W) = a_r > a_l, \quad \forall W \in C^- \cap \{a = a_r\},$$

in the region of $a = a_r$ between $SW_{sub}(C^-, a_r)$ and $SW_{sup}(C^-, a_r)$ the following inequality holds

$$a_{min}(W) > a_l,$$

and this is the case, in particular, for $W_r$. Using Theorem 2.3.3, we have

$$a_l < a_{min}(W_r) < a_{min}(\hat{W}_0),$$

where $\hat{W}_0$ is the state corresponding to the 2-stationary shock. As a consequence, $\hat{W}_0$ has to be in the region between $SW_{sub}(C^-, a_r)$ and $SW_{sup}(C^-, a_r)$. More specifically, since $W_0^+ = S_2^B(W_r) \cap S_{sub}(C^-, a_r)$, $\hat{W}_0$ has to belong to the arc of $S_2^B(W_r)$ that links $W_r$ to $W_0^+$ what implies $\hat{h}_0 < h_0^+$. Using Proposition 2.2.1 we have that the speed of $\mathcal{S}_2(W_0^+, W_r)$ is positive. □

- **Region VI**: Given a state $W_r$ in Region VI, we first consider the 3-stationary wave that connects $W_r$ to $W_0^- = SW(W_r, a_r)$ and then the 2-rarefaction connecting $W_0^-$ to a state $W_l$ belonging to the line $h = 0$: see Figure 2.14. Since $W_r \in A_3$, by (MC) $W_0^- \in A_3$ and thus the head and the tail of the 2-rarefaction are negative, so that it can be followed by the stationary wave. In conclusion, we have found a solution of the form:

$$\mathcal{R}_2(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_r). \tag{2.3.30}$$

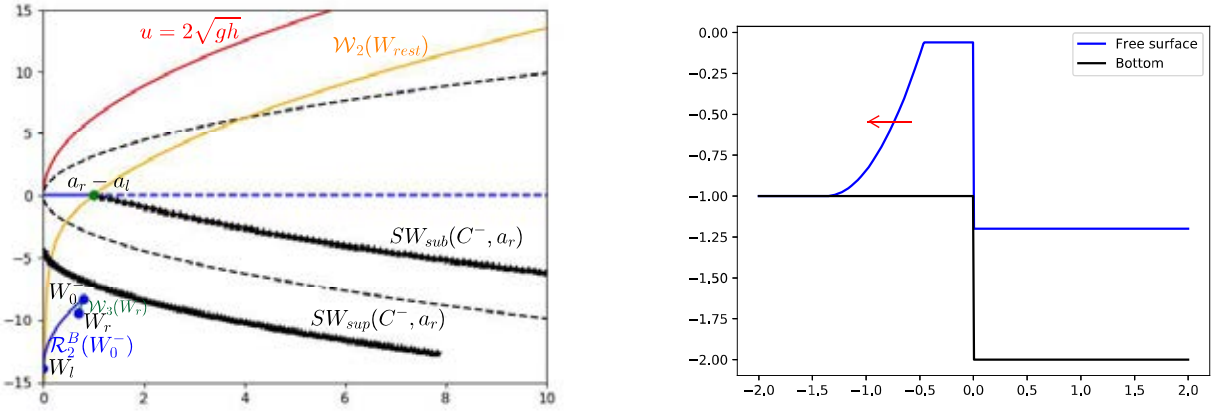



Figure 2.14: Solution of a state in Region VI. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

For some states of this region we can also construct a solution of the form (2.3.25). This solution is possible whenever the 2-shock has positive speed. Region VI can be split into two subregions, Regions VIa and VIb, whose boundary is the curve

$$\Gamma = \{W \in A_3 : \ \sigma_2(W, W_0^+) = 0\},$$

composed by the states of $A_3$ such that the 2-shock linking them with the state $W_0^+$ given by Proposition 2.3.2 is stationary, i.e. $W_0^+ = \hat{W}_0$ (see Figure 2.15). In Region VIa (including the boundary) a solution of the form (2.3.25) can be constructed.

- **Regions II and IV**: A self-similar solution of the Riemann problem with initial conditions (2.3.2) composed by admissible simple waves cannot be constructed in these cases: the states belonging to these regions cannot be linked through a 2-wave with positive or null velocity to a state whose mechanical energy is enough to go up the step (see Remarks 2.2.1 and 2.2.2). In other words, the step acts in these cases as an obstacle for the fluid. Therefore, in order to construct a physically meaningful solution, we consider $x = 0$ as a boundary of the problem and impose the boundary condition $u = 0$: the flow is arrested by the obstacle. To impose properly this boundary condition, following [41], we consider the partial Riemann problem consisting of the homogeneous shallow water system with initial condition:

$$\begin{cases} W(x, 0) = W_r, & \text{if } x > 0, \\ u = 0, & \text{if } x < 0. \end{cases} \tag{2.3.31}$$

To solve this problem, from $W_r$ we have to reach a state in the line $\{u = 0\}$. In Region II we consider the 2-rarefaction that connects $W_r$ to $W_0^r \in \mathcal{R}_2^B(W_r) \cap \{u = 0\}$

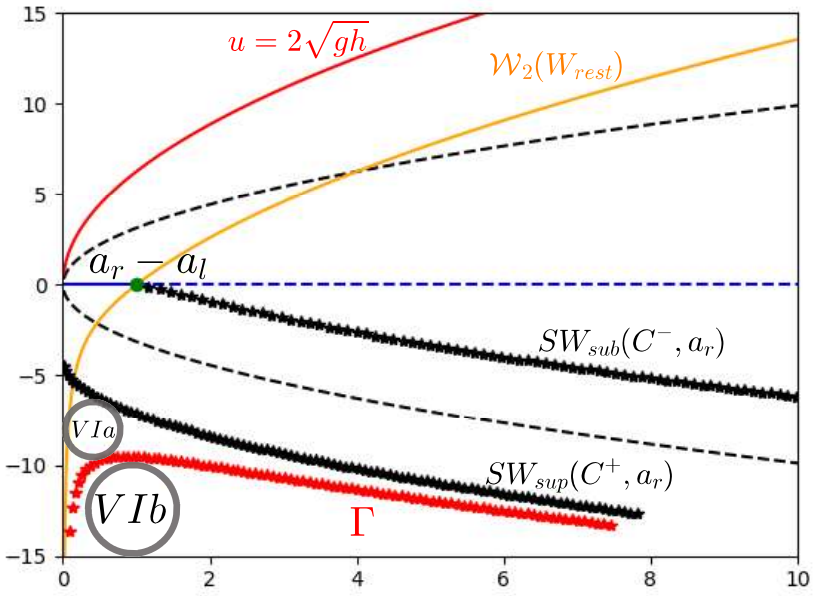



Figure 2.15: The two subregions of Region VI.

given by (2.3.15): see Fig 2.16. Since $W_r \in A_1 \cup A_2$ and $W_0^r \in A_2$, the tail and the head of the 2-rarefaction are positive. The structure of the solution is thus:

$$\mathcal{R}_2(W_0^r, W_r). \tag{2.3.32}$$

In Region IV we consider the 2-shock linking $W_r$ to $W_0^s \in \mathcal{S}_2^B(W_r) \cap \{u = 0\}$: see Fig 2.17. This 2-shock has positive speed. The structure of the solution is thus:

$$\mathcal{S}_2(W_0^s, W_r). \tag{2.3.33}$$

- **Boundary between Region I and Region II**: for states such that $\{u_r = 2\sqrt{gh_r}\}$ a solution can be constructed like in Region I, i.e. (2.3.8). In this case, there is vacuum at $x < 0$ for every $t > 0$.
- **Boundary between Region II and Region III**: for states $W_r \in \mathcal{R}_2(W_{rest})$ a solution can be constructed consisting of the 2-rarefaction connecting $W_r$ to $W_{rest}$ followed by the 3-stationary wave connecting $W_{rest}$ to the vacuum state $(0, 0, a_l)$. The structure of this solution is

$$W_3((0, 0, a_l), W_{rest}) \oplus R_2(W_{rest}, W_r).$$

- **Boundary between Region II and Region IV**: the partial Riemann problem for the homogeneous shallow water system with initial condition (2.3.31) in considered. The solution is stationary in this case.

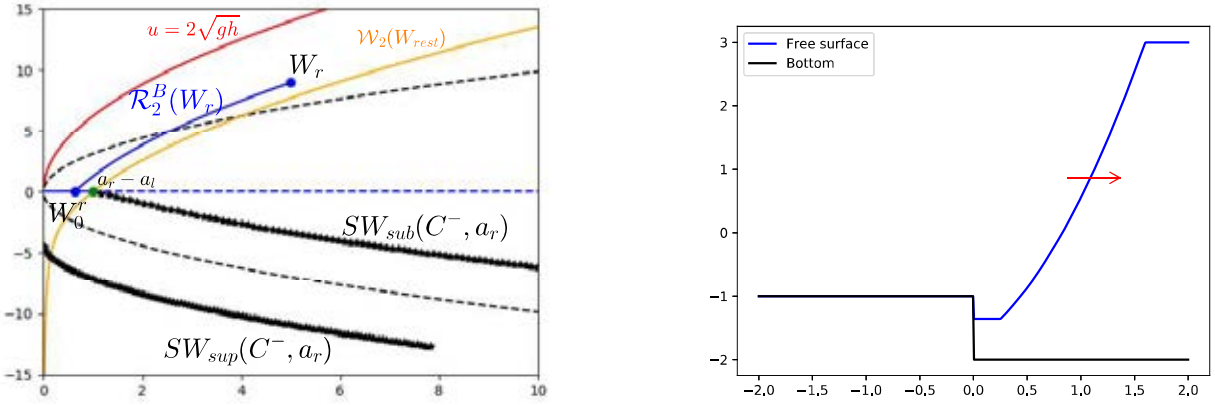



Figure 2.16: Solution of a state in Region II. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

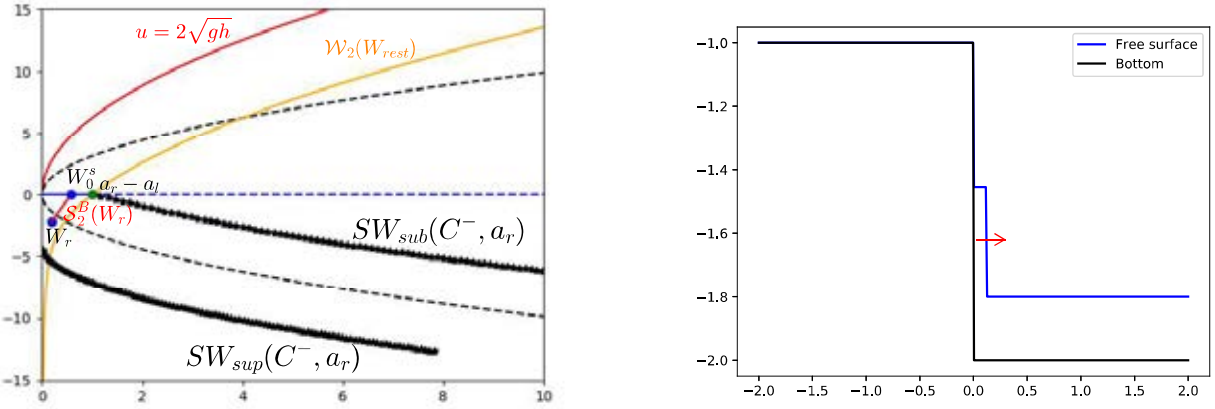



Figure 2.17: Solution of a state in Region IV. Left: projection of the intermediate states and the simple waves on the $(h, u)$-plane. Right: sketch of the free surface at a time $t > 0$. The direction of the movement of the simple waves is indicated with a red arrow.

- **Boundary between Region III and Region V**: for states $W_r \in SW_{sub}(C^-, a_r)$ we consider the 3-stationary wave that connects $W_r$ to $W_0^- = SW(W_r, a_l)$ and then the 2-rarefaction connecting $W_0^-$ to a state $W_l$ belonging to the line $\{h = 0\}$. Since $W_r \in SW_{sub}(C^-, a_r)$, $W_0^- \in C^-$ and thus the head of the 2-rarefaction has zero speed and its tail is positive. Thereupon we can construct a solution of the form:

$$\mathcal{R}_2(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_r). \tag{2.3.34}$$

- **Boundary between Region IV and Region V**: for states $W_r \in \mathcal{S}_2(W_{rest})$ a

solution can be constructed consisting of the 2-shock connecting $W_r$ to $W_{rest}$ followed by the 3-stationary wave connecting $W_{rest}$ to the vacuum state $(0, 0, a_l)$. The structure of this solution is

$$W_3((0, 0, a_l), W_{rest}) \oplus S_2(W_{rest}, W_r).$$

- **Boundary between Region IV and Region VI**: for states $W_r \in SW_{sup}(C^-, a_r)$ that are above the curve $\mathcal{W}_2(W_{rest})$ a solution can be constructed like in Region VI, i.e (2.3.30).

- **Boundary between Region V and Region VI**: For states $W_r \in SW_{sup}(C^-, a_r)$ that are below the curve $\mathcal{W}_2(W_{rest})$ two solutions like in Region VIa can be constructed, one of the form (2.3.30) and another one of the form (2.3.25).

### 2.3.3 Summary

Taking into account the order of the eigenvalues at every region, the sign of the velocities of the waves, and Theorems 2.3.2 and 2.3.3, it can be checked that the solutions constructed above are the only possible self-similar solutions consisting of simple waves linking some intermediate states. We summarize the results in this table:

| Case 1: initial condition (2.3.1) | | |
|---|---|---|
| Regions | No. solutions | Form of the solutions |
| Region I | 1 | $\mathcal{W}_3(W_l, W_0^+) \oplus \mathcal{R}_1(W_0^+, W_r)$ |
| Region II | 1 | $\mathcal{R}_1(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_0^+) \oplus \mathcal{R}_1(W_0^+, W_r)$ |
| Region III | 1 | $\mathcal{R}_1(W_l, W_r)$ |
| Case 2: initial condition (2.3.2) | | |
| Regions | No. solutions | Form of the solutions |
| Region I | 1 | $\mathcal{R}_2(W_l, W_r)$ |
| Region II | 0 | $\mathcal{R}_2(W_0^r, W_r)$ |
| Region III | 1 | $\mathcal{R}_2(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_0^+) \oplus \mathcal{R}_2(W_0^+, W_r)$ |
| Region IV | 0 | $\mathcal{S}_2(W_0^s, W_r)$ |
| Region V | 1 | $\mathcal{R}_2(W_0, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_0^+) \oplus \mathcal{S}_2(W_0^+, W_r)$ |
| Region VIa | 2 | $\begin{cases} \mathcal{R}_2(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_r) \quad \text{and} \\ \mathcal{R}_2(W_0, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_0^+) \oplus \mathcal{S}_2(W_0^+, W_r) \end{cases}$ |
| Region VIb | 1 | $\mathcal{R}_2(W_l, W_0^-) \oplus \mathcal{W}_3(W_0^-, W_r)$ |

In Regions where the number of solutions is 0, the structure of the solution corresponds to that of the considered partial Riemann problem for the homogeneous shallow water system.

## 2.4 Numerical tests

In the previous section we have seen that in some cases the wet-dry Riemann problem may have zero, one, or two solutions. The goal of this section is to study how some Godunov-type methods behave in these different situations. We consider the following numerical fluxes for the homogeneous shallow water system:

- Roe flux (ROE): we take in (1.2.17)-(1.2.18) the following Roe matrix:

$$A_\Phi(W_l, W_r) = \begin{pmatrix} \bar{u} & \bar{h} \\ g & \bar{u} \end{pmatrix}, \tag{2.4.1}$$

  where

$$\bar{h} = \frac{h_l + h_r}{2}, \quad \bar{u} = \frac{\sqrt{h_l}u_l + \sqrt{h_r}u_r}{\sqrt{h_l} + \sqrt{h_r}}.$$

  (See [82]).

- Godunov flux (GODUNOV): see [153] for the solution of the Riemann problems in the homogeneous case;

and the following numerical treatment of the source term:

- Upwind discretization (UPW): we take the straight segments (1.2.106) as our family of paths (see [13]);
- Hydrostatic reconstruction (HR): this is a particular case of the generalized hydrostatic reconstruction based on the integral curves of the linearly degenerate field in the particular case we have $u = 0$, i.e., the case of water at rest (see [5]);
- Generalized hydrostatic reconstruction (GHR): we take the family of paths and the fluctuations described in Chapter 1 Section 1.2.5.4 (see [46]).

For more details in these numerical techniques the interested readers are addressed to the cited references. Let us only mention that, while the three numerical treatments of the source terms lead to methods that preserve water-at-rest stationary solutions (i.e. they satisfy the $C$-property, according to [13]), only GHR leads to schemes that preserve any stationary solution and, in particular, stationary contact discontinuities. The following combination of these techniques will be used:

- ROE_UPW.
- ROE_HR.
- ROE_GHR.

- GOD_GHR.

The numerical treatment of wet-dry fronts proposed in [38] is used in ROE_UPW.

In all the tests below, uniform meshes of the interval [-2,2], CFL = 0.9, and free boundary conditions are considered.

### 2.4.1 Tests 1 and 2

The goal of these tests is to study the behavior of the numerical methods in cases where the Riemann problem has no solution, i.e. when the step acts like an obstacle for the fluid.

We consider first the initial condition:

$$W_1(x,0) = \begin{cases} W_l = (0,0,1), \ if \ x < 0, \\ W_r = (0.2,-2,2), \ if \ x \geq 0. \end{cases} \tag{2.4.2}$$

It can be easily checked that the problem is of the form (2.3.2) and that the state $W_r$ is in Region IV: a solution of the form (2.3.33) was proposed in this case. In Figure 2.18 the numerical solutions obtained with a mesh of 400 cells are shown: all of them seem to converge to the proposed solution. Due to the numerical error there is a small amount of water that goes up the step and travel leftward for the ROE_UPW (see the bump on the left in Figure 2.18).

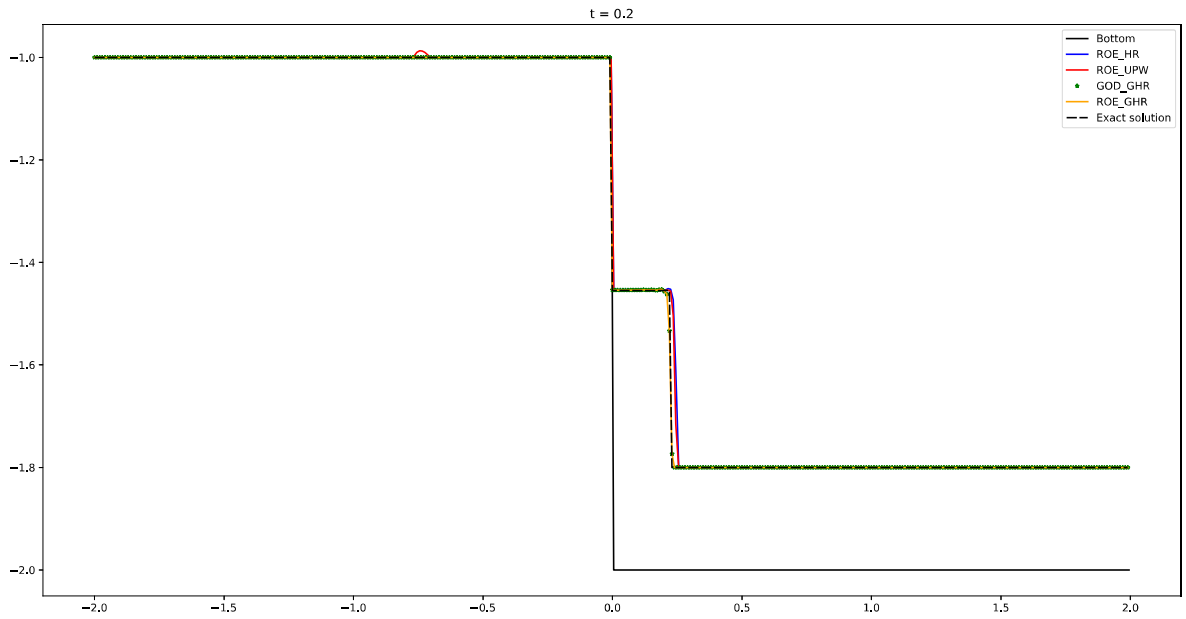



Figure 2.18: Numerical results of the $h$ component for the initial condition (2.4.2).

Secondly, we consider the initial condition:

$$W_2(x,0) = \begin{cases} W_l = (0,0,1), \ if \ x < 0, \\ W_r = (2,5,2), \ if \ x \geq 0. \end{cases} \tag{2.4.3}$$

It can be easily checked that in this case the state $W_r$ is in Region II: a solution of the form (2.3.32) was proposed in this case. In Figure 2.19 the numerical solutions obtained with a mesh of 400 cells are shown while in Figure 2.20 we use a mesh of 800 cells: all of them seem to converge again to the proposed solution.

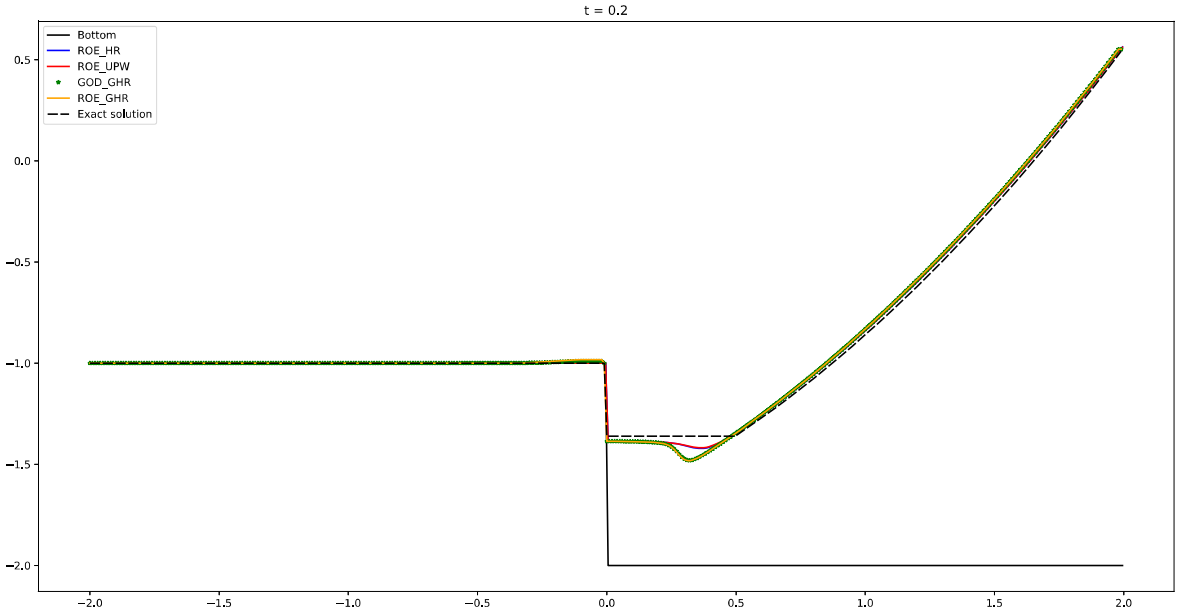



Figure 2.19: Numerical results with 400 cells of the $h$ component for the initial condition (2.4.2).

### 2.4.2 Test 3

In this test we check the ability of the numerical methods to correctly capture stationary contact discontinuities. We consider the next initial condition:

$$W_3(x,0) = \begin{cases} W_l = (0,0,1), \ if \ x < 0, \\ W_r = (5,1,2), \ if \ x \geq 0. \end{cases} \tag{2.4.4}$$

Since the right state is in Region III, the exact solution is of the form (2.3.9). The numerical results obtained in a mesh of 400 cells are shown in Figure 2.21 and in Figure 2.22 we make a zoom on the stationary contact discontinuity. The numerical methods based on UPW or HR, as expected, are not able to correctly capture the stationary contact discontinuities: among them, ROE_HR gives the numerical solution farthest to

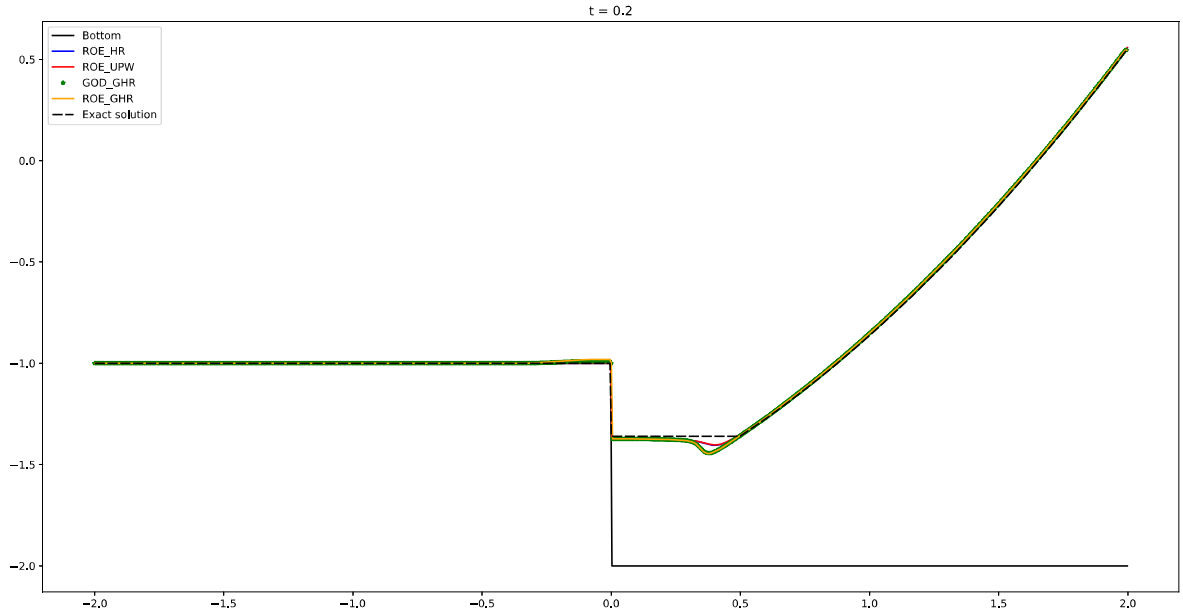



Figure 2.20: Numerical results with 800 cells of the $h$ component for the initial condition (2.4.2).

the proposed one. Although GHR leads to schemes that preserve these discontinuities, only GOD_GHR seems to converge to the proposed weak solution: ROE_GHR captures a contact discontinuity that does not satisfy the (MC) criterion (there is a transition from sub to supercritical through the stationary contact discontinuity).

### 2.4.3 Test 4

In [124] a test was shown where the numerical solutions obtained with ROE_HR were very far of those produced by other methods. In fact, this initial condition was in Region VIa where two possible solutions have been found: one of the form (2.3.25) and the other of the form (2.3.30): let us check that ROE_HR converges to the former while the other numerical methods converge to the latter. To do this, we consider the initial condition:

$$W_4(x,0) = \begin{cases} W_l = (0,0,1), \ if \ x < 0, \\ W_r = (1,-8,2), \ if \ x \geq 0. \end{cases} \tag{2.4.5}$$

It can be checked that the right state is in Region VIa. Figure 2.23 shows the result obtained with a mesh of 400 cells. We observe that the only method that converges to the solution of the form (2.3.25) is ROE_HR, while all the others schemes converge to the solution of the form (2.3.30).

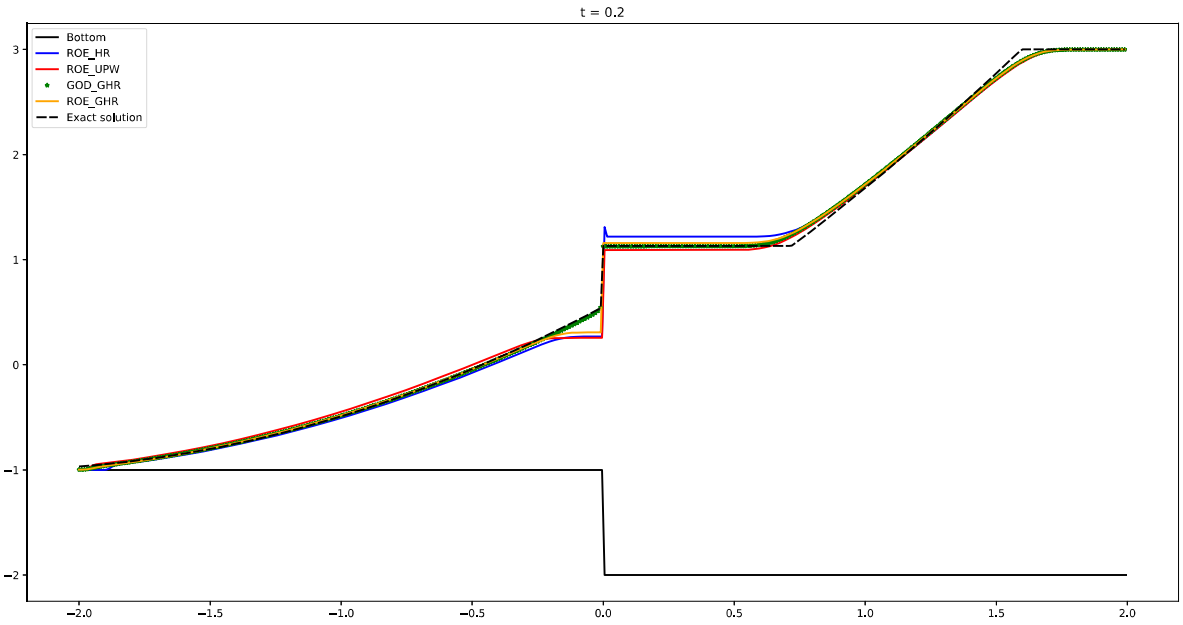



Figure 2.21: Numerical results of the $h$ component for the initial condition (2.4.4).

### 2.4.4 Summary of numerical results

We have compared the numerical solutions obtained with several standard numerical fluxes and treatment of the source terms with the exact solutions we obtained before. The main conclusions are the following:

- In cases where the Riemann problem does not have solutions, the numerical methods converge to the proposed solutions based on a reinterpretation of the problem.
- In order to capture correctly the stationary contact discontinuities it is necessary to have a numerical method that preserve them. Nevertheless, this is not sufficient to ensure the convergence to the proposed solutions: the numerical solution may converge to weak solutions containing stationary discontinuities over the step that do not satisfy the (MC) criterion.
- The combination of Godunov numerical flux and the Generalized Hydrostatic Reconstruction technique introduced in [51] seems to produce a numerical method that correctly captures all the proposed solutions.
- In cases where the Riemann problem has two solutions, the numerical methods may converge to one or to the other.

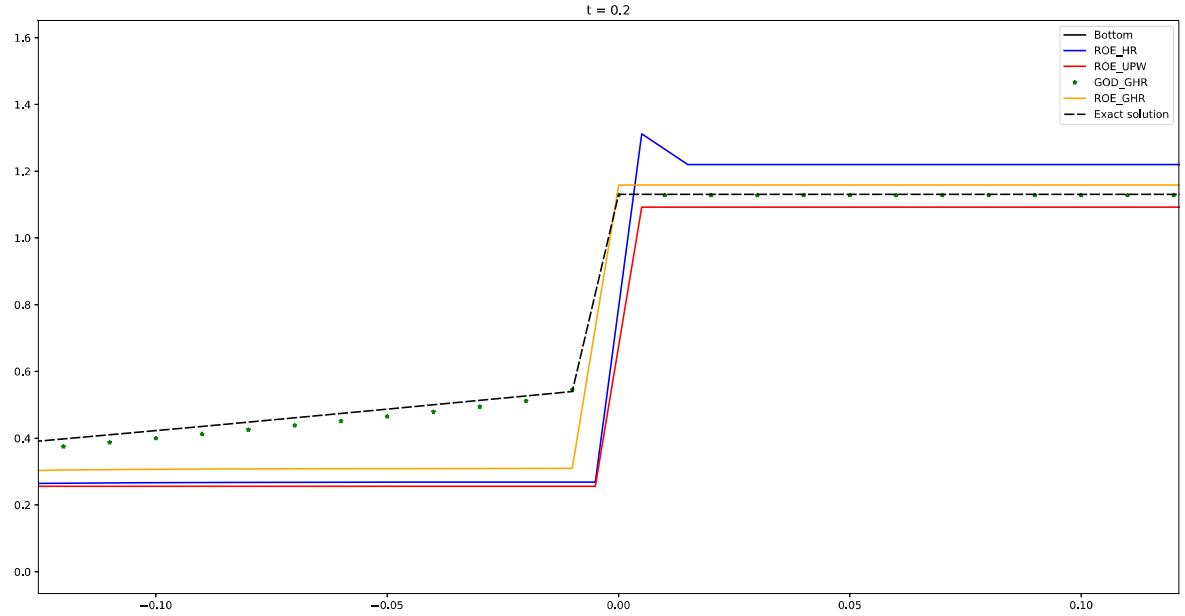



Figure 2.22: Zoom of the numerical results of the $h$ component for the initial condition (2.4.4).

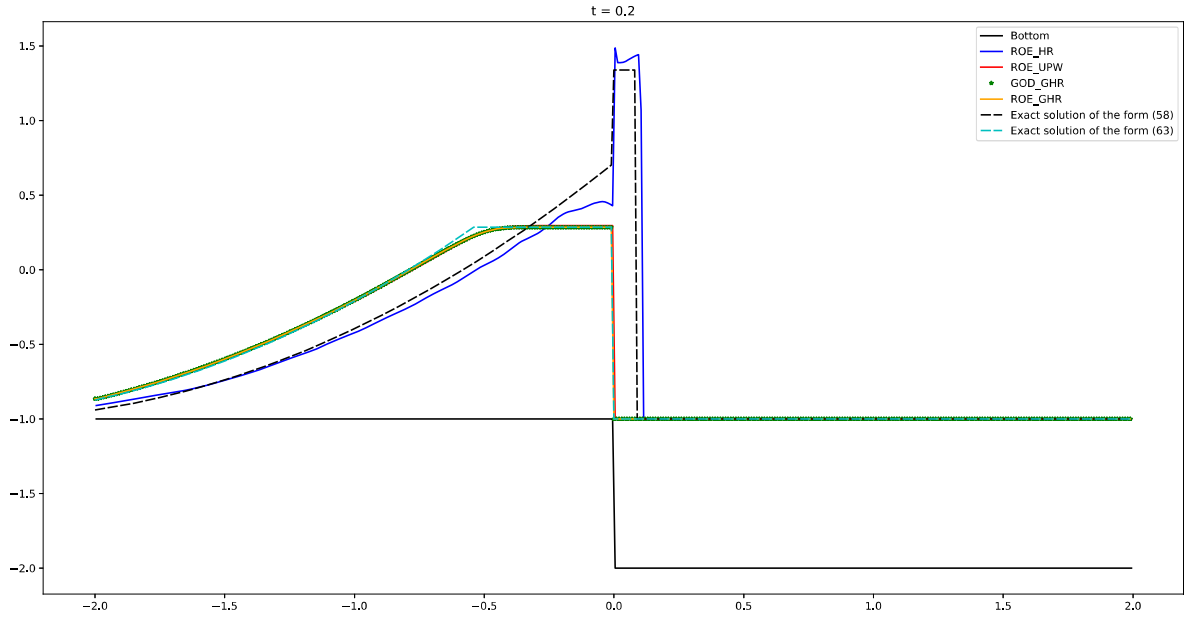



Figure 2.23: Numerical results of the $h$ component for the initial condition (2.4.5).

# Chapter 3

# On the efficient implementation of PVM methods and simple Riemann solvers. Application to the Roe method for large hyperbolic systems

In Chapter 1 it has been seen that the implementation of the Roe method requires the computation of the absolute value of the intermediate matrices whose definition depends on their eigenvalues and eigenvectors. An implementation based on this definition may be computationally expensive if the eigenvalues and eigenvectors are not explicitly known especially if the size of the matrices is large. To overcome this difficulty, different families of methods have been proposed in the literature. Two of them are the Approximate Riemann solvers (ARS), in the sense of Harten et al. [93], and the Polynomial Viscosity matrix (PVM) methods introduced in Chapter 1. As we saw in (1.2.25), in PVM the absolute value of the matrix is approximated by the evaluation of the Roe matrix at a chosen polynomial, which avoids the necessity of computing the eigenvalues and eigenvectors. In many cases, the chosen polynomial interpolates the absolute value function at some points (either in the Lagrange or the Hermite sense): we will call interpolatory PVM to these methods for simplicity. A new implementation of interpolatory PVM methods is presented here. This new implementation is based on the Newton form of the polynomials that is known to be the more efficient way to evaluate the interpolation polynomials.

Approximate Riemann solvers on their side are based on the approximation of the solutions of the Riemann problems associated at each intercell. In the case of the so-called simple Riemann solvers (SRS) these approximations consist of some constant states linked by jump discontinuities that travel at constant speed: see (1.2.31). In [125] the close connection between PVM and SRS methods was studied. In particular, under certain assumptions, a SRS method can be interpreted as an interpolatory PVM. In this case, the

efficient implementation of PVM presented here can also be applied to SRS. New shorter proofs of the results shown in [125] about the relations between PVM and SRS are given here for the sake of completeness and clarity.

Roe method can be interpreted as a complete SRS and therefore as a PVM, as it was pointed out in [35]. The implementation based on the Newton form of the corresponding interpolating polynomial can be then used: this new implementation will be called *Newton Roe method* here. The advantage of Newton Roe method compared to the standard implementation is that (a) the eigenvectors have not to be computed and (b) no matrix inversion is required. This new implementation shows a significant speedup if the number of equations of the hyperbolic systems is large enough. In particular we are interested in solving large nonconservative hyperbolic systems like those corresponding to the multilayer shallow water system (see, for example, [44]) or the Quadrature-Based Moment Equations (QBME, see [98] and the references therein) where the number of equations is a parameter of the model, either given in terms of the number of layers in the first case or the number of moments in the second one. Although the application of the methods to general nonconservative systems is considered here for the sake of generality, all of them reduce to conservative methods for systems of conservation laws so that the new implementations can be used as well in this particular case.

The chapter is structured as follows: in Section 3.1 the implementation of interpolatory PVM methods using the Newton form of the interpolating polynomial is introduced together with a study of the complexity. Then, in Section 3.2, the relation between PVM methods and SRS is analyzed. Section 3.3 is devoted to describe the Roe method as a PVM solver so it will allow us to apply the Newton form of the interpolation polynomial to improve the computational efficiency of its implementation. Finally, in Section 3.4 some numerical results will be shown to compare the standard form of the Roe scheme and the new Newton form. The two-layer shallow water system and the QBME moment models in primitive and partially conservative variables are considered. The content of this chapter was published by Pimentel-García, Parés, Castro and Koellermeier in 2021 by the journal *Applied Mathematics and Computation*, see [136].

## 3.1 Newton form of PVM methods

In this chapter, recalling the expression of PVM methods (1.2.25)-(1.2.26)-(1.2.27), we only consider polynomial approximations of the absolute value function that are based on Lagrange or Hermite interpolation, i.e. the polynomial $p_r^{i+\frac{1}{2}}$ in (1.2.24) is the polynomial of degree less or equal than $r-1$ that interpolates the absolute value function at $r$ different

points $\sigma_j^{i+1/2}$, $j = 1, \ldots, r$.:

$$p_r^{i+\frac{1}{2}}(\sigma_j^{i+1/2}) = \left|\sigma_j^{i+1/2}\right|, \quad j = 1, \ldots, r,$$

or the polynomial $p_{2r}^{i+\frac{1}{2}}$ of degree $2r-1$ that interpolates the absolute value function and its derivative at $r$ different points $\sigma_j^{i+1/2}$, $j = 1, \ldots, r$:

$$p_{2r}^{i+\frac{1}{2}}(\sigma_j^{i+1/2}) = \left|\sigma_j^{i+1/2}\right|, \quad (p_{2r}^{i+\frac{1}{2}})'(\sigma_j^{i+1/2}) = sign(\sigma_j^{i+1/2}), \quad j = 1, \ldots, r,$$

where

$$sign(x) = \begin{cases} -1 & x < 0, \\ 0 & x = 0, \\ 1 & x > 0. \end{cases}$$

Well known conservative methods like Rusanov, Lax-Friedrichs, HLL, FORCE, GFORCE, etc. can be interpreted as PVM methods based on interpolating polynomials: see [35].

**Remark 3.1.1.** *When one of the interpolation points $\sigma_j^{i+1/2}$ is equal to 0 and a smooth transonic regime is detected, the interpolated value is taken to be $\epsilon > 0$ instead of 0 as an entropy fix technique to avoid the appearance of 'dog-leg' phenomena. The adequate value of $\epsilon$ depends on the problem, but good choices of this parameter can be done using the smallest non-zero eigenvalue. For instance, if $\lambda = \sigma_k^{i+1/2}$ is the smallest non-zero eigenvalue, then a good choice is given by $\epsilon = 0.5|\lambda|$.*

### 3.1.1 Implementation: the Lagrange case

Once the interpolation points and values have been chosen, the most efficient way to evaluate the interpolation polynomials $p_r^{i+1/2}$ is using its Newton form, based on the well-known divided differences. Let us describe an algorithm to compute the product

$$\mathcal{Q}_{i+\frac{1}{2}}(W_{i+1}^n - W_i^n) \tag{3.1.1}$$

based on this form of the polynomial. For the sake of simplicity, the dependency on the intercell will not be explicitly written, so that indexes and super-indexes $i + 1/2$ will be dropped. Moreover, $W_l$ and $W_r$ will be used instead of $W_i^n$ and $W_{i+1}^n$ for simplicity. Using this notation, one has that, in the case of Lagrange interpolation, (3.1.1) writes as follows:

$$\begin{aligned} \mathcal{Q}(W_r - W_l) &= p_r(\mathcal{A})(W_r - W_l) \\ &= [\sigma_1](W_r - W_l) + \sum_{i=2}^{r} [\sigma_1, \ldots, \sigma_i] \prod_{j=1}^{i-1} \left(\mathcal{A}(W_r - W_l) - \sigma_j(W_r - W_l)\right), \end{aligned}$$

where the divided differences are recursively defined as follows:

- $[\sigma_i] = |\sigma_i|, \quad i = 1, \ldots, r.$
- Given $k+1$ indexes $\{i_0, ..., i_k\} \subset \{1, ..., r\}$,

$$[\sigma_{i_0}, \sigma_{i_1}, ..., \sigma_{i_k}] = \frac{[\sigma_{i_1}, ..., \sigma_{i_k}] - [\sigma_{i_0}, ..., \sigma_{i_{k-1}}]}{\sigma_{i_k} - \sigma_{i_0}}. \tag{3.1.2}$$

Once the divided differences have been computed, the following algorithm can be used to compute (3.1.1) in an optimal way:

- $V_0 = W_r - W_l$,
- For i = 1 to r:

$$V_i = \mathcal{A}V_{i-1} - \sigma_i V_{i-1},$$

and finally,

$$p_r(\mathcal{A})(W_r - W_l) = [\sigma_1]V_0 + [\sigma_1, \sigma_2]V_1 + ... + [\sigma_1, ..., \sigma_r]V_{r-1}. \tag{3.1.3}$$

The operations needed to compute (3.1.1) are thus the following:

- $\frac{3}{2}r(r+1)$ operations to compute the divided differences;
- $r$ matrix-vector products;
- $2r$ scalar-vector products;
- $2r$ vector sums.

Therefore, the total number of operations is

$$\frac{3}{2}r(r+1) + r(2N^2 - N) + 4rN.$$

Since in practice $r$ is at most $O(N)$ the complexity of the algorithm is

$$O\left(2rN^2\right). \tag{3.1.4}$$

The complexity of the computation of $\mathcal{D}^{\pm}$, that involves another matrix/vector product is the same.

### 3.1.2 Implementation: the Hermite case

In the case of Hermite interpolation, (3.1.1) writes as follows:

$$\begin{aligned}\mathcal{Q}(W_r - W_l) &= p_{2r}(\mathcal{A})(W_r - W_l)\\ &= [\tilde{\sigma}_1](W_r - W_l) + \sum_{i=2}^{2r}[\tilde{\sigma}_1,\ldots,\tilde{\sigma}_i]\prod_{j=1}^{i-1}\left(\mathcal{A}(W_r - W_l) - \tilde{\sigma}_j(W_r - W_l)\right),\end{aligned}$$

where

$$\tilde{\sigma}_{2j-1} = \tilde{\sigma}_{2j} = \sigma_j, \quad j = 1,\ldots,r,$$

and the divided differences are defined in the same way with the following exceptions:

- $[\tilde{\sigma}_{2j-1}, \tilde{\sigma}_{2j}] = sign(\sigma_j), \quad j = 1,\ldots r.$

The algorithm to compute (3.1.1) is then the same and its complexity is, in this case:

$$O\left(4rN^2\right). \tag{3.1.5}$$

**Remark 3.1.2.** *This algorithm can be easily adapted to PVM based on a polynomial that interpolates the absolute value and its derivative at some points and only the absolute value at some other points.*

## 3.2 Relation between PVM and SRS methods

In [125] the relation between PVM and SRS methods was studied: the main results shown there are revisited here and new shorter proofs are given. The following result in [136] gives a necessary and sufficient condition to have the equivalence between a PVM method and a SRS.

**Theorem 3.2.1.** *Given a PVM and a SRS method based on the same family of paths, the following statements are equivalent:*

1. $\mathcal{D}^{\pm}_{PVM}(W_l, W_r) = \mathcal{D}^{\pm}_{SRS}(W_l, W_r), \quad \forall W_l, W_r \in \Omega,$
2. *For every* $W_l, W_r \in \Omega$*:*

$$\sum_{i=1}^{r}|\sigma_i|(W_i - W_{i-1}) = p(A_\Phi(W_l, W_r))(W_r - W_l). \tag{3.2.1}$$

*Proof.* Taking into account (1.2.33) and (1.2.16) we have the following equality:

$$\sum_{j=1}^{r}\sigma_j(W_j - W_{j-1}) = \mathcal{A}_\Phi(W_l, W_r)(W_r - W_l), \tag{3.2.2}$$

where $\mathcal{A}_\Phi$ is the Roe linearization chosen to define the PVM. Now, given $W_l, W_r \in \Omega$, we have:

$$\begin{aligned}
&\mathcal{D}^+_{PVM}(W_l, W_r) = \mathcal{D}^+_{SRS}(W_l, W_r) \\
&\quad \Leftrightarrow \frac{1}{2}\mathcal{A}_\Phi(W_l, W_r)(W_r - W_l) + \frac{1}{2}p(\mathcal{A}_\Phi(W_l, W_r))(W_r - W_l) \\
&\qquad = \sum_{\sigma_{j+1}>0} \sigma_{j+1}(W_{j+1} - W_i) \\
&\quad \Leftrightarrow \frac{1}{2}\mathcal{A}_\Phi(W_l, W_r)(W_r - W_l) + \frac{1}{2}p(\mathcal{A}_\Phi(W_l, W_r))(W_r - W_l) \\
&\qquad = \frac{1}{2}\sum_{j=1}^{r} \sigma_j(W_j - W_{j-1}) + \frac{1}{2}\sum_{j=1}^{r} |\sigma_j|(W_j - W_{j-1}) \\
&\quad \Leftrightarrow p(\mathcal{A}_\Phi(W_l, W_r))(W_r - W_l) = \sum_{j=1}^{r} |\sigma_j|(W_j - W_{j-1}),
\end{aligned}$$

where (3.2.2) has been used. The proof of the equivalence between

$$\mathcal{D}^-_{PVM}(W_l, W_r) = \mathcal{D}^-_{SRS}(W_l, W_r)$$

and (3.2.1) is similar. □

### 3.2.1 PVM based on Lagrange interpolation

We first show that any PVM method based on Lagrange polynomial interpolation can be seen as a SRS. More precisely, the following result holds:

**Theorem 3.2.2.** *Any PVM based on a polynomials $p_r$ of degree less or equal than r-1 that interpolates the graph of the absolute value at r different points $\sigma_i$, $i = 1, ..., r$ can be interpreted as a SRS with speeds $\sigma_i$, $i = 1, ..., r$.*

*Proof.* Let us consider the Lagrange polynomial basis $l_i(\lambda), i = 1, ..., r$, where $l_i$ is the polynomial of degree $r - 1$ such that:

$$l_i(\sigma_j) = \delta_{i,j} = \begin{cases} 1 & if \quad i = j, \\ 0 & if \quad i \neq j. \end{cases}$$

Then $p_r$ can be written in its Lagrange form:

$$p(\lambda) = \sum_{i=1}^{r} |\sigma_i| l_i(\lambda).$$

Let us define:

$$V_i = l_i(\mathcal{A}_\Phi(W_l, W_r))(W_r - W_l), \quad i = 1, ..., r.$$

The intermediate states $W_i$, $i = 0, ..., r$ given by

- $W_0 = W_l$;
- $W_i = V_i + W_{i-1}, \quad j = 1, \dots, r$

together with the speeds $\sigma_i$, $i = 1, \dots, r$ define a SRS. In effect, from the equality

$$\sum_{i=1}^{r} l_i(\lambda) = 1, \quad \forall \lambda \in \mathbb{R}$$

we deduce

$$W_r = (W_r - W_0) + W_0 = \sum_{i=1}^{r} V_i + W_0 = (W_r - W_l) + W_0 = W_r.$$

Now, (1.2.32) is trivially checked and (1.2.33) is easily deduced from the Roe property (1.2.16) and

$$\sum_{i=1}^{r} \sigma_i l_i(\lambda) = \lambda, \quad \forall \lambda \in \mathbb{R}.$$

Finally, we have:

$$\sum_{i=1}^{r} |\sigma_i| V_i = \sum_{i=1}^{r} |\sigma_i| l_i(\mathcal{A}_\Phi(W_l, W_r))(W_r - W_l) = p(\mathcal{A}_\Phi(W_l, W_r))(W_r - W_l).$$

Therefore, using Theorem 3.2.1, the SRS solver coincides with the PVM method. □

Let us prove now a kind of reciprocal:

**Theorem 3.2.3.** *A SRS with $r$ speeds $\sigma_i$, $i = 1, ..., r$ such that $V_i = W_i - W_{i-1}$, $i = 1, ..., r$, are linearly independent, can be interpreted as the PVM based on the polynomial $p$ of degree less or equal than $r - 1$ that interpolates the points*

$$\{(\sigma_i, |\sigma_i|)\}, \quad i = 1, \cdots, r.$$

*Proof.* First, if $r < N$, the set of linearly independent vectors $\{V_1, ..., V_r\}$ is completed to obtain a basis $\{V_1, ..., V_N\}$ of $\mathbb{R}_N$ and the set of real numbers $\sigma_1, ..., \sigma_r$ is completed to obtain a family of pairwise different numbers $\sigma_1, ..., \sigma_N$. Then we consider the matrix $\mathcal{A}_\Phi(W_L, W_R)$ whose eigenvalues are $\sigma_1, ..., \sigma_N$ and the corresponding eigenvectors are $\{V_1, ..., V_N\}$. We also consider the polynomial $p$ such that $p(\sigma_i) = |\sigma_i|$, $i = 1, ..., r$. Let us see that the PVM associated with $\mathcal{A}_\Phi$ and $p$ is equivalent to the SRS. Using the property (1.2.33) we have that:

$$\begin{aligned} \mathcal{A}_\Phi(W_l, W_r)(W_r - W_l) &= \mathcal{A}_\Phi(W_l, W_r) \left( \sum_{i=1}^{r} V_i \right) && (3.2.3) \\ &= \sum_{i=1}^{r} \sigma_i V_i && (3.2.4) \\ &= \int_0^1 \mathcal{A}(\Phi(\xi; W_l, W_r)) \frac{\partial \Phi}{\partial \xi}(\xi; W_l, W_r) d\xi. && (3.2.5) \end{aligned}$$

Therefore, $\mathcal{A}_\Phi$ defines a Roe matrix. Moreover:

$$p(\mathcal{A}_\Phi(W_l,W_r))(W_r-W_l)=p(\mathcal{A}_\Phi(W_l,W_r))\sum_{i=1}^{r}V_i=\sum_{i=1}^{r}p(\sigma_i)V_i=\sum_{i=1}^{r}|\sigma_i|V_i,$$

and using again Theorem 3.2.1, we have that the PVM is equivalent to the SRS. □

Therefore, when a SRS satisfies the hypothesis of Theorem 3.2.3 it can be interpreted as an interpolatory PVM and the implementation described in the previous section can be thus applied.

### 3.2.2 PVM based on Hermite interpolation

We show now that any PVM method based on Hermite interpolation can be seen as a SRS.

**Theorem 3.2.4.** *Any PVM based on a polynomial $p_{2r}$ of degree less or equal than $2r-1$ such that*

$$p_{2r}(\sigma_i)=|\sigma_i|,\ i=1,...,r,\quad p'_{2r}(\sigma_i)=sign(\sigma_i),\ i=1,...,r,$$

*where $\sigma_1<\sigma_2<...<\sigma_I<0<\sigma_{I+1}<...<\sigma_r$, can be interpreted as a SRS with $r+1$ speeds $\sigma_i$, $i=1,...,r$, and 0.*

*Proof.* Let us consider the Hermite polynomial basis $h_i(\lambda)$, $i=1,...,r$, $k_i(\lambda)$, $i=1,...,r$, i.e. the polynomials of degree less or equal than $2r-1$ such that

$$\begin{aligned}h_i(\sigma_j)=\delta_{i,j},\ h'_i(\sigma_j)=0,\quad \forall i,j,\\ k_i(\sigma_j)=0,\ k'_i(\sigma_j)=\delta_{i,j},\quad \forall i,j.\end{aligned}$$

Using this basis, the interpolating polynomial can be written as follows:

$$p_{2r}(\lambda)=\sum_{i=1}^{r}|\sigma_i|h_i(\lambda)+\sum_{i=1}^{r}sign(\sigma_i)k_i(\lambda),\quad \forall\lambda. \tag{3.2.6}$$

Let us define:

$$\begin{aligned}
&V_i^0=h_i(\mathcal{A}_\Phi)(W_r-W_l),\quad i=1,\dots,r,\\
&V_i^1=k_i(\mathcal{A}_\Phi)(W_r-W_l),\quad i=1,\dots,r,\\
&W_i^0=W_{i-1}^0+V_i^0,\quad i=1,..,r,\\
&W_i^1=-\frac{1}{\sigma_{i+1}-\sigma_i}V_i^1,\quad i=1,...,I-1,\\
&W_I^{1,-}=\frac{1}{\sigma_I}V_I^1,\\
&W_I^{1,+}=-\frac{1}{\sigma_{I+1}}V_{I+1}^1,\\
&W_i^1=-\frac{1}{\sigma_{i+1}-\sigma_i}V_{i+1}^1,\quad i=I+1,...,r-1.
\end{aligned}$$

Let us consider the function:

$$R(\sigma)=\begin{cases} W_l & if \quad \sigma<\sigma_1,\\ W_1=W_1^0+W_1^1 & if \quad \sigma_1<\sigma<\sigma_2,\\ \vdots & \\ W_{I-1}=W_{I-1}^0+W_{I-1}^1 & if \quad \sigma_{I-1}<\sigma<\sigma_I,\\ W_I^-=W_I^0+W_I^{1,-} & if \quad \sigma_I<\sigma<0,\\ W_I^+=W_I^0+W_I^{1,+} & if \quad 0<\sigma<\sigma_{I+1},\\ W_{I+1}=W_{I+1}^0+W_{I+1}^1 & if \quad \sigma_{I+1}<\sigma<\sigma_{I+2}\\ \vdots & \\ W_{r-1}=W_{r-1}^0+W_{r-1}^1 & if \quad \sigma_{r-1}<\sigma<\sigma_r\\ W_r^0 & if \quad \sigma_r<\sigma. \end{cases} \tag{3.2.7}$$

Let us verify that $R$ is a SRS. First, reasoning like in the proof of Theorem 3.2.2 and taking into account the equality

$$\sum_{i=1}^{r} h_i(\lambda)=1, \quad \forall\lambda,$$

we obtain

$$W_r^0=W_r.$$

Next, (1.2.32) can be trivially checked. Let us check property (1.2.33):

$$\begin{aligned}
&\sigma_1\left(W_1^0+W_1^1-W_L\right)+\sum_{i=2}^{I-1}\sigma_i\left(W_i^0+W_i^1-W_{i-1}^0-W_{i-1}^1\right)\\
&\quad+\sigma_I\left(W_I^0+W_I^{1,-}-W_{I-1}^0-W_{I-1}^0\right)+\sigma_{I+1}\left(W_{I+1}^0+W_{I+1}^1-W_I^0-W_I^{1,+}\right)\\
&\quad+\sum_{i=I+2}^{r-1}\sigma_i\left(W_i^0+W_i^1-W_{i-1}^0-W_{i-1}^1\right)+\sigma_r\left(W_r^0-W_{r-1}^0-W_{r-1}^1\right)\\
&\quad=\sum_{i=1}^{r}\sigma_i(W_i^0-W_{i-1}^0)+\sum_{i=1}^{I-1}(\sigma_i-\sigma_{i-1})W_i^1+\sigma_I W_I^{1,-}-\sigma_{I+1}W_I^{1,+}\\
&\qquad+\sum_{i=I+2}^{r-1}(\sigma_i-\sigma_{i-1})W_i^1\\
&\quad=\sum_{i=1}^{r}\sigma_i V_i^0+\sum_{i=1}^{r}V_i^1=\left(\sum_{i=1}^{r}(\sigma_i h_i+k_i)\right)(\mathcal{A}_\Phi)(W_r-W_l)=\mathcal{A}_\Phi(W_r-W_l)\\
&\quad=\int_0^1\mathcal{A}(\Phi(\xi;W_l,W_r))\frac{\partial\Phi}{\partial\xi}(\xi;W_l,W_r)d\xi.
\end{aligned}$$

where the Roe property and the equality

$$\left(\sum_{i=1}^{r}(\sigma_i h_i + k_i)\right)(\lambda) = \lambda, \quad \forall\lambda$$

have been used. Therefore, $R$ is a SRS. Let us finally check (3.2.1):

$$\begin{aligned}
&|\sigma_1|\left(W_1^0 + W_1^1 - W_l\right) + \sum_{i=2}^{I-1}|\sigma_i|\left(W_i^0 + W_i^1 - W_{i-1}^0 - W_{i-1}^1\right) \\
&\quad + |\sigma_I|\left(W_I^0 + W_I^{1,-} - W_{I-1}^0 - W_{I-1}^0\right) + |\sigma_{I+1}|\left(W_{I+1}^0 + W_{I+1}^1 - W_I^0 - W_I^{1,+}\right) \\
&\quad + \sum_{i=I+2}^{r-1}|\sigma_i|\left(W_i^0 + W_i^1 - W_{i-1}^0 - W_{i-1}^1\right) + |\sigma_r|\left(W_r^0 - W_{r-1}^0 - W_{r-1}^1\right) \\
&= \sum_{i=1}^{r}|\sigma_i|(W_i^0 - W_{i-1}^0) + \sum_{i=1}^{I-1}(|\sigma_i| - |\sigma_{i-1}|)W_i^1 + |\sigma_I|W_I^{1,-} - |\sigma_{I+1}|W_I^{1,+} \\
&\qquad + \sum_{i=I+2}^{r-1}(|\sigma_i| - |\sigma_{i-1}|)W_i^1 \\
&= \sum_{i=1}^{r}|\sigma_i|V_i^0 - \sum_{i=1}^{I}V_i^1 + \sum_{i=I+1}^{r}V_i^1 \\
&= \left(\sum_{i=1}^{r}(|\sigma_i|h_i + sign(\sigma_i)k_i)\right)(\mathcal{A}_\Phi)(W_r - W_l) \\
&= p_{2r}(\mathcal{A}_\Phi)(W_r - W_l),
\end{aligned}$$

and thus the PVM is equivalent to the SRS.

□

**Remark 3.2.1.** *The main advantage of having an expression of a PVM method as a SRS is that this form makes easier to prove properties such as the positivity of the method: see [125].*

## 3.3 Application to the Roe method

### 3.3.1 Standard form

As we saw in (1.2.23), given a Roe linearization $\mathcal{A}_\Phi$, the standard form of the Roe method is given by (1.2.3) with:

$$\mathcal{D}^{\pm}_{i+\frac{1}{2}} = \mathcal{D}^{\pm}(W_i^n, W_{i+1}^n) = \frac{1}{2}\mathcal{A}_{i+\frac{1}{2}}(W_{i+1}^n - W_i^n) \pm \frac{1}{2}\left|\mathcal{A}_\Phi(W_i^n, W_{i+1}^n)\right|(W_{i+1}^n - W_i^n),$$

where

$$|\mathcal{A}_\Phi(W_i^n, W_{i+1}^n)| = R_\Phi(W_i^n, W_{i+1}^n)|\Lambda_\Phi(W_i^n, W_{i+1}^n)|R_\Phi^{-1}(W_i^n, W_{i+1}^n), \tag{3.3.1}$$

being $|\Lambda_\Phi(W_i^n, W_{i+1}^n)|$ the diagonal matrix whose coefficients are the absolute value of the eigenvalues of $\mathcal{A}_\Phi(W_i^n, W_{i+1}^n)$, and $R_\Phi(W_i^n, W_{i+1}^n)$ is the matrix whose $i$-th column is a right-eigenvector associated to the $i$-th eigenvalue.

### 3.3.2 SRS form

The Roe method can be interpreted as the scheme corresponding to the complete SRS

$$R(\lambda; W_l, W_r) = \begin{cases} W_0 = W_l, & if \quad \lambda < \lambda_1, \\ W_j, & if \quad \lambda_j < \lambda < \lambda_{j+1}, \\ W_N = W_r, & if \quad \lambda_N < \lambda, \end{cases} \tag{3.3.2}$$

where $\lambda_i$, $i = 1, ..., N$, are the eigenvalues of $\mathcal{A}_\Phi$ and the intermediates states $W_i$ verify:

$$W_i - W_{i-1} = \alpha_i R_i, \quad i = 1, ..., N,$$

where $R_i$ is the right eigenvector associated to $\lambda_i$ and $\alpha_i$ is the $i$-coordinate of the vector $W_r - W_l$ in the $\mathbb{R}^{N+1}$ basis defined by the eigenvectors. The reciprocal is also true: any complete SRS, i.e. any SRS with $N$ speeds, is equivalent to a Roe method.

### 3.3.3 PVM form

According to Theorem 3.2.3, a Roe method can be written as the PVM method based on the polynomial $p_N^{i+\frac{1}{2}}$ of degree less or equal than $N-1$ that verifies:

$$p_N(\lambda_i) = |\lambda_i|, \quad i = 1, ..., N. \tag{3.3.3}$$

Therefore, it can be implemented following the algorithm proposed in Section 3.1.1 what leads to the Newton Roe method. Since $r = N$, in this case the complexity of the algorithm is (see (3.1.4)):

$$O\left(2N^3\right).$$

Observe that, with this implementation, it is not necessary to compute the eigenvectors what, in many cases, may constitute an important saving of computational time. If the eigenvectors are explicitly known or easy to compute, still the computation of the absolute value of the matrix requires the inversion of the eigenvectors matrix. If, for instance, the inverse is computed by solving $N$ linear systems using the LU factorization, the complexity would be

$$O\left(\frac{3}{2}N^3 + 2N^3\right) = O\left(\frac{7}{2}N^3\right),$$

so that still the Newton Roe implementation is cheaper.

**Remark 3.3.1.** *Let us suppose that a Roe matrix $\mathcal{A}_\Phi$ and some approximations $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$ of the wave speeds are available that do not coincide with the eigenvalues of the Roe matrix. One could then implement the PVM based on the Roe matrix and the polynomial that interpolates the absolute value function at the approximated speeds. The resulting PVM method would be then equivalent to a new Roe method whose matrix $\widetilde{\mathcal{A}}_\Phi$ has the approximated speeds as eigenvalues and the states $V_i$ given in the proof of Theorem 3.2.2 as the corresponding eigenvectors.*

### 3.3.4 Close or double eigenvalues

In some cases, the Roe matrix has eigenvalues that are very close or even identical (if the system is not strictly hyperbolic). In these cases, the divided differences are not well-defined or involves close to zero denominators. In order to fix this difficulty, if $\lambda_k \simeq \lambda_{k+1}$ or $\lambda_k = \lambda_{k+1}$, then instead of considering the Lagrange interpolation

$$p(\lambda_i) = |\lambda_i|, \quad i = 1, ..., N$$

the following interpolation is used:

$$p(\lambda_i) = |\lambda_i|, \ i \neq k+1, \quad p'(\lambda_k) = sign(\lambda_k).$$

This is again an advantage compared to the implementation of the standard form of the Roe method, since in the case of a double eigenvalue the computation of the Jordan form is necessary to compute the absolute value of the matrix.

## 3.4 Models and numerical tests

In this section Roe method and Newton Roe method will be applied to the following models:

- the two-layer shallow water equations,
- the Quadrature-Based Moment equations for rarefied gas.

Different implementations of the Roe method for the first model have been considered so far: see [133], [103]. The combination of the Newton Roe formulation with the use of Ferrari's formula to compute the eigenvalues as proposed in the latter reference leads to an optimal implementation of the Roe method for this system in which concerns the number of operations. Although it will be shown that, as expected, this implementation reduces the computational time, systems with more than 4 variables are needed to observe a drastic reduction of the CPU time. In QBME models, the number of equations is a parameter of the model so that it constitutes an excellent test problem to measure the CPU reduction due to the Newton Roe implementation as a function of the size of the

system.

The methods have been implemented in C++ and run on the linux-subsystem of a 64-bit Windows 10 YOGA 720 with Intel Core i7-7200 2.5 GHz machine. The Eigen library [90] has been used to compute all matrix-vectors operations.

### 3.4.1 Two-layer shallow water equations

We consider the homogeneous two-layer 1-D shallow water system (see [44]):

$$\begin{cases} \dfrac{\partial h_1}{\partial t} + \dfrac{\partial q_1}{\partial x} = 0, \\ \dfrac{\partial q_1}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{q_1^2}{h_1} + \dfrac{1}{2}gh_1^2\right) = -gh_1\dfrac{\partial h_2}{\partial x}, \\ \dfrac{\partial h_2}{\partial t} + \dfrac{\partial q_2}{\partial x} = 0, \\ \dfrac{\partial q_2}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{q_2^2}{h_2} + \dfrac{1}{2}gh_2^2\right) = -\dfrac{\rho_1}{\rho_2}gh_2\dfrac{\partial h_1}{\partial x}. \end{cases} \tag{3.4.1}$$

Index 1 refers to the upper layer while index 2 refers to the lower layer. This system uses the following notation:

- $h_i = h_i(x,t) \geq 0$ is the thickness of the $i$-th layer at the section of coordinate $x$ at time $t$.
- $q_i = q_i(x,t)$ is the discharge of the $i$-th layer at the section of coordinate $x$ at time $t$.
- $g$ is the intensity of the gravitational field.
- $\rho_i$ refers to the constant density of the $i$-th layer.

The bottom is assumed to be flat. Following [74], this system can be written in the form

$$\partial_t W + F(W)_x + B(W)W_x = 0, \tag{3.4.2}$$

where

$$W = \begin{pmatrix} h_1 \\ q_1 \\ h_2 \\ q_2 \end{pmatrix}, \quad F(W) = \begin{pmatrix} q_1 \\ \dfrac{q_1^2}{h_1} + \dfrac{1}{2}gh_1^2 \\ q_2 \\ \dfrac{q_2^2}{h_2} + \dfrac{1}{2}gh_2^2 \end{pmatrix},$$

$$B(W) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & gh_1 & 0 \\ 0 & 0 & 0 & 0 \\ grh_2 & 0 & 0 & 0 \end{pmatrix},$$

with $r = \rho_1/\rho_2$. System (3.4.1) can be rewritten in the form (1.1.2):

$$\partial_t W + \mathcal{A}(W)\partial_x W = 0,$$

with

$$\mathcal{A}(W) = A(W) + B(W) = JF(W) + B(W).$$

The eigenvalues of $\mathcal{A}(W)$ are the roots of the characteristic polynomial:

$$p(\lambda) = (\lambda^2 - 2u_1\lambda + u_1^2 - gh_1)(\lambda^2 - 2u_2\lambda + u_2^2 - gh_2) - rgh_1gh_2, \tag{3.4.3}$$

where $u_i = q_i/h_i$, $i = 1, 2$. When $r \cong 1$, first order approximations of the eigenvalues were given in [147]:

$$\lambda_{ext}^{\pm} = \frac{u_1h_1 + u_2h_2}{h_1 + h_2} \pm \sqrt{g(h_1 + h_2)}, \tag{3.4.4}$$

$$\lambda_{int}^{\pm} = \frac{u_1h_2 + u_2h_1}{h_1 + h_2} \pm \sqrt{g'\frac{h_1h_2}{h_1 + h_2}\left(1 - \frac{(u_1 - u_2)^2}{g'(h_1 + h_2)}\right)}, \tag{3.4.5}$$

where $g' = (1 - r)g$.

The exact expression of the eigenvalues can be obtained by using Ferrari's method to find an analytical solution for quartic equations. Following [103], this expression is as follows:

$$\lambda_{ext}^{\pm} = \frac{\frac{a}{2} \pm \sqrt{Z} \pm \sqrt{-A - Z \mp \frac{B}{\sqrt{Z}}}}{2}. \tag{3.4.6}$$

$$\lambda_{int}^{\pm} = \frac{\frac{a}{2} \pm \sqrt{Z} \mp \sqrt{-A - Z \mp \frac{B}{\sqrt{Z}}}}{2}, \tag{3.4.7}$$

with

$$Z = \frac{1}{3}\left(2\sqrt{\Delta_0}cos\left(\frac{\theta}{3}\right) - A\right),$$

$$\theta = arccos\left(\frac{\Delta_1}{2\sqrt{\Delta_0^3}}\right),$$

$$A = 2b - \frac{3a^2}{4},$$

$$B = 2c - ab + a^3 4,$$

$$\Delta_0 = b^2 + 12d - 3ac,$$
$$\Delta_1 = 27a^2d - 9abc + 2b^3 - 72bd + 27c^2,$$

where

$$a = -2(u_1 + u_2),$$
$$b = u_1 - gh_1 + 4u_1u_2 + u_2^2 - gh_2,$$
$$c = -2u_2(u_1^2 - gh_1) - 2u_1(u_2^2 - gh_2),$$
$$d = (u_1^2 - gh_1)(u_2^2 - gh_2) - rgh_1gh_2,$$

being $a, b, c, d$ the coefficients of the polynomial $p$ written in the form:

$$p(\lambda) = \lambda^4 + a\lambda^3 + b\lambda^2 + c\lambda + d.$$

Given an eigenvalue $\lambda$, an associated eigenvector is given by:

$$R_i = \begin{pmatrix} 1 \\ \lambda \\ \mu \\ \lambda\mu \end{pmatrix}, \tag{3.4.8}$$

where:

$$\mu = \frac{(\lambda - u_1)^2}{gh_1} - 1.$$

A more detailed description of the calculation of the eigenvalues and eigenvectors of this system can be found in [103].

We consider the Roe matrix of the system based on the family of straight segments

$$\Phi(\xi; W_l, W_r) = W_l + \xi(W_r - W_l)$$

described in [44].

#### 3.4.1.1 Test 1: Dam-break problem

We consider the internal dam-break test introduced in [74]: the equations are solved in the space interval $[0, 10]$ with initial conditions:

$$h_1(x, 0) = \begin{cases} 0.2, & if \quad x < 5, \\ 0.8, & if \quad x \geq 5, \end{cases}$$

$$h_2(x, 0) = \begin{cases} 0.8, & if \quad x < 5, \\ 0.2, & if \quad x \geq 5, \end{cases}$$

$$q_1(x, 0) = q_2(x, 0) = 0.$$

| #Cells | Standard Roe | Newton Roe |
|---|---|---|
| 1250 | 10.09 | 9.80 |
| 2500 | 28.96 | 28.73 |
| 5000 | 102.85 | 100.09 |
| 10000 | 362.65 | 349.47 |

Table 3.1: Test 1: CPU times in (s) for meshes with different number of cells for the standard Roe scheme and the Newton Roe scheme.

Roe and Newton Roe methods are run in the time interval $[0, 10]$ with CFL $= 0.9$. Since both methods are equivalent, the numerical results are identical to machine precision and therefore we don't compare them. In Table 3.1 the CPU times in (s) corresponding to both implementations using meshes with different number of cells are shown.

It can be seen that here is a small speedup for the Newton Roe method for this system, for which $N = 4$.

### 3.4.2 Hyperbolic Moment Models for rarefied gases

An application that involves larger systems of hyperbolic PDEs is given by continuum models for rarefied gases. In addition to the standard conservation laws, the accurate description of non-equilibrium flows in rarefied gases requires further equations [150]. These equations model the evolution of higher order moments of the underlying particle distribution function, see e.g. [156]. The resulting so-called *moment models* based on the work in [89] are not always hyperbolic, but have been modified to yield hyperbolicity using different frameworks, see [27, 72, 99]. There are many applications of these models, e.g. for shock tube computations and 2D flows, see [101, 73]. The modifications lead to nonconservative, hyperbolic PDE systems, such as the QBME system, which was numerically investigated in [102]. The QBME model is very suitable for the application of the Newton Roe solver, because the eigenvalues are analytically known, whereas the full eigenvector decomposition is difficult to compute. In addition, the model contains more equations leading to a large potential speedup.

#### 3.4.2.1 QBME moment models in primitive variables

In the following, we will briefly describe the QBME model equations, for more details see [98]. The general form of the model equations is already given in the form

$$\partial_t W_M + \mathcal{A}_M(W_M)\,\partial_x W_M = S(W_M), \tag{3.4.9}$$

where $W_M \in \mathbb{R}^{M+1}$ is the vector of unknown variables given by

$$W_M = (\rho, v, \theta, f_3, \ldots, f_M)^T \tag{3.4.10}$$

with density $\rho$, velocity $v$, temperature $\theta$, and the $f_i, i = 0, \dots, M$ with $f_0 = 1, f_1 = 0, f_2 = 0$ are the so-called *moments*, coefficients from the expansion of an underlying distribution function, and the transport matrix $\mathcal{A}_M \in \mathbb{R}^{(M+1)\times(M+1)}$ depends on $W_M$.

The right-hand side $S \in \mathbb{R}^{M+1}$ denotes the collision term, which will be the same for all models. It will be approximated by the BGK model [22] and reads

$$S(W_M) = -\frac{1}{\tau} B W_M, \tag{3.4.11}$$

for non-linear relaxation time $\tau = \frac{\mathrm{Kn}}{\rho}$ with Knudsen number Kn and diagonal matrix

$$B = \mathrm{diag}\,(0, 0, 0, 1, \dots, 1) \in \mathbb{R}^{(M+1)\times(M+1)}, \tag{3.4.12}$$

corresponding to conservation of mass, momentum and energy and a relaxation of the higher order coefficients towards equilibrium.

For the QBME [98, 99] the system matrix can be written as $\mathcal{A}_M =$

$$\begin{pmatrix}
v & \rho & & & & & \\
\frac{\theta}{\rho} & v & 1 & & & & \\
 & 2\theta & v & \frac{6}{\rho} & & & \\
 & 4f_3 & \frac{\rho\theta}{2} & v & 4 & & \\
-\frac{\theta f_3}{\rho} & 5f_4 & \frac{3f_3}{2} & \theta & v & 5 & \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\
-\frac{\theta f_{M-2}}{\rho} & M f_{M-1} & \frac{(M-2)f_{M-2}+\theta f_{M-4}}{2} & -\frac{M(M+1)f_M}{2\theta} & -\frac{3f_{M-3}}{\rho} & \theta & v & M \\
-\frac{\theta f_{M-1}}{\rho} & (M+1)f_M & -f_{M-1} + \frac{\theta f_{M-3}}{2} & \frac{3(M+1)f_M}{\rho\theta} & -\frac{3f_{M-2}}{\rho} & & \theta & v
\end{pmatrix}. \tag{3.4.13}$$

The characteristic polynomial of the system matrix $A_\mathrm{M}$ reads

$$\chi_M(\lambda) = \theta^{\frac{M+1}{2}} \mathrm{He}_{M+1}\left(\frac{\lambda - v}{\sqrt{\theta}}\right). \tag{3.4.14}$$

The real eigenvalues $\lambda_j$ are thus the shifted and scaled roots of the probabilists' Hermite polynomial $\mathrm{He}_{M+1}$

$$\lambda_j = v + \sqrt{\theta} c_j, \;\; j = 1, \dots, M+1, \tag{3.4.15}$$

where the $c_j$ are now the Hermite roots such that $\mathrm{He}_{M+1}(c_j) = 0$. The system is thus globally hyperbolic.

The entries of the eigenvectors $v_j$ can be derived as

$$\begin{aligned}
v_{j,1} &= \rho \mathrm{He}_0(c_j), \\
v_{j,2} &= \theta^{1/2} \mathrm{He}_1(c_j), \\
v_{j,3} &= \theta \mathrm{He}_2(c_j), \\
v_{j,4} &= \frac{1}{3!} \rho \theta^{3/2} \mathrm{He}_3(c_j), \\
v_{j,5} &= \frac{1}{4!} \rho \theta^{4/2} \mathrm{He}_4(c_j) - f_3 \sqrt{\theta} c_j, \\
v_{j,6} &= \frac{1}{5!} \rho \theta^{5/2} \mathrm{He}_5(c_j) + f_3 \theta \frac{1 - c_j^2}{2} - f_4 \sqrt{\theta} c_j, \\
&\vdots \\
v_{j,M} &= \frac{1}{(M-1)!} \rho \theta^{(M-1)/2} \mathrm{He}_{M-1}(c_j) + f_{M-3} \theta \frac{1 - c_j^2}{2} - f_{M-2} \sqrt{\theta} c_j, \\
v_{j,M+1} &= \frac{1}{M!} \rho \theta^{M/2} \mathrm{He}_M(c_j) + f_{M-2} \theta \frac{1 - c_j^2}{2} - f_{M-1} \sqrt{\theta} c_j + (M+1) f_M \frac{c_j^2 - 1}{2}.
\end{aligned}$$

**Remark 3.4.1.** *Other hyperbolic moment models such as the ones mentioned in [26, 100] exist, but the models are all given in a similar form with analytical eigenvalues and eigenvectors in the primitive variables. We thus chose the QBME model exemplarily to focus on the numerical method.*

#### 3.4.2.2 QBME model in partially-conservative variables

To ensure exact conservation of mass, momentum and energy, the moment model above needs to be written in terms of the partially-convective/partially-conservative variables $U_M \in \mathbb{R}^{M+1}$ with

$$U_M = \left(\rho, \rho v, (v^2 + \theta)\rho, v(v^2 + 3\theta)\rho + 6 f_3, f_4, \ldots, f_M\right)^T. \tag{3.4.16}$$

The transformation matrix from primitive to partially-conservative variables is

$$\frac{\partial U}{\partial W} = \begin{pmatrix}
1 & 0 & & & & \\
v & \rho & & & 0 & \\
v^2 + \theta & 2v\rho & \rho & & & \\
v(v^2 + 3\theta) & 3(v^2 + \theta)\rho & 3v\rho & 6 & & \\
 & & & 1 & & \\
 & 0 & & & \ddots & \\
 & & & & & 1
\end{pmatrix}, \tag{3.4.17}$$

and its inverse

$$\left(\frac{\partial U}{\partial W}\right)^{-1} = \begin{pmatrix} 1 & 0 & & & & \\ -\frac{v}{\rho} & \frac{1}{\rho} & & & 0 & \\ \frac{v^2-\theta}{\rho} & -\frac{2v}{\rho} & \frac{1}{\rho} & & & \\ -\frac{v(v^2-3\theta)}{6} & \frac{v^2-\theta}{2} & -\frac{v}{2} & \frac{1}{6} & & \\ & & & 1 & & \\ & 0 & & & \ddots & \\ & & & & & 1 \end{pmatrix}. \tag{3.4.18}$$

Choosing $U_M$ as the vector of unknowns results in a new system of equations that describes the same physical behavior but with different variables, with the benefit of exact conservation of mass, momentum and energy as the first three variables in $U_M$. Assuming the new variables $U_M$ can be computed from $W_M$ with Jacobian $\frac{\partial U_M}{\partial W_M} \in \mathbb{R}^{(M+1)\times(M+1)}$, which is a lower triangular matrix depending on the entries of $W_M$, the change of variables can be performed. We use $\left(\frac{\partial U_M}{\partial W_M}\right) BW_M = \widetilde{B}W_M$, with

$$\widetilde{B} = diag\,(0,0,0,6,1,\ldots,1) \in \mathbb{R}^{(M+1)\times(M+1)}. \tag{3.4.19}$$

The resulting system

$$\partial_t U_M + \widetilde{\mathcal{A}_M}\ \partial_x U_M = -\frac{1}{\tau}\widetilde{B}W_M. \tag{3.4.20}$$

uses a new transformed system matrix

$$\widetilde{\mathcal{A}_M} = \left(\frac{\partial U_M}{\partial W_M}\right) \mathcal{A}_M \left(\frac{\partial U_M}{\partial W_M}\right)^{-1} \tag{3.4.21}$$

with the same eigenvalues as $\mathcal{A}_M$.

We know from the previous section, that we can write the primitive system matrix using its eigenvalues (3.4.15) and corresponding eigenvectors as

$$\mathcal{A}_M = V^{-1}\Lambda V, \tag{3.4.22}$$

where $V$ is the matrix containing the eigenvectors and $\Lambda$ is a diagonal matrix containing the eigenvalues. The corresponding eigenvectors and eigenvalues are given in section 3.4.2.1.

The eigenvalues of the partially-conservative models will be the same as those of the primitive variables. However, the eigenvectors change according to the following computation

$$\begin{aligned} \widetilde{\mathcal{A}_M} &= \frac{\partial U}{\partial W}\mathcal{A}\left(\frac{\partial U}{\partial W}\right)^{-1} \\ &= \frac{\partial U}{\partial W}V^{-1}\Lambda V\left(\frac{\partial U}{\partial W}\right)^{-1} \\ &= \widetilde{V}^{-1}\Lambda\widetilde{V} \end{aligned}$$

for $\widetilde{V}$ the corresponding eigenvector matrix of the partially-conservative system matrix. This matrix is thus given by

$$\widetilde{V} = V \left( \frac{\partial U}{\partial W} \right)^{-1}. \tag{3.4.23}$$

As $V$ and $\left(\frac{\partial U}{\partial W}\right)^{-1}$ are analytically given, the exact eigenvectors can be computed analytically during the simulation, even though already their evaluation might be computationally expensive during a standard Roe scheme.

The QBME model in partially-conservative variables can be computed analytically by following the transformation of the variables. The final system reads

$$\widetilde{\mathcal{A}}_{\text{QBME}} = \widetilde{\mathcal{A}}_{M,1} + \widetilde{\mathcal{A}}_{M,2}. \tag{3.4.24}$$

with $\widetilde{\mathcal{A}}_{M,1} =$

$$\begin{pmatrix}
0 & 1 & & & & & & & \\
0 & 0 & 1 & & & & & & \\
0 & 0 & 0 & 1 & & & & & \\
-v^4 + 6v^2\theta - 3\theta^2 - \frac{24vf_3}{\rho} & 4\left(v^3 - 3v\theta + \frac{6f_3}{\rho}\right) & -6v^2 + 6\theta & 4v & 24 & & & & \\
\frac{3v^2 f_3 - 5\theta f_3 - 10vf_4}{2\rho} - \frac{v(v^2-3\theta)\theta}{6} & \frac{\theta}{2}(v^2-\theta)\frac{-3vf_3+5f_4}{\rho} & -\frac{v\theta}{2} + \frac{3f_3}{2\rho} & \frac{\theta}{6} & v & 5 & & & \\
b_{1,5} & b_{2,5} & b_{3,5} & b_{4,5} & \theta & v & 6 & & \\
\vdots & \vdots & \vdots & \vdots & & \ddots & \ddots & \ddots & \\
b_{1,M-1} & b_{2,M-1} & b_{3,M-1} & b_{4,M-1} & & & \theta & v & M \\
b_{1,M} & b_{2,M} & b_{3,M} & b_{4,M} & & & & \theta & v
\end{pmatrix}, \tag{3.4.25}$$

and the following entries in the first four columns

$$\begin{aligned}
b_{1,i} &= \frac{f_{i-2}\left(u^3 - 3\theta u\right) + \theta f_{i-3}\left(u^2 - \theta\right) + f_{i-1}\left((i-1)u^2 - \theta(i+1)\right) - 2(i+1)uf_i}{2\rho}, \\
b_{2,i} &= \frac{3f_{i-2}\left(\theta - u^2\right) - 2\theta u f_{i-3} - 2(i-1)uf_{i-1} + 2(i+1)f_i}{2\rho}, \\
b_{3,i} &= \frac{\theta f_{i-3} + 3uf_{i-2} + (i-1)f_{i-1}}{2\rho}, \\
b_{4,i} &= -\frac{f_{i-2}}{2\rho}.
\end{aligned}$$

Note that for entries $b_{1,i}$, we need to use the coefficient $f_2 = 0$.

The second matrix in (3.4.24) contains seven additional terms in the last two rows

such that the modification is given by

$$\widetilde{\mathcal{A}}_{M,2} = \frac{M\,(M+1)}{2\rho\theta} \begin{pmatrix} & & & & \emptyset \\ \widehat{m}_{M-1,1} & \widehat{m}_{M-1,2} & \widehat{m}_{M-1,3} & & \\ \widehat{m}_{M,1} & \widehat{m}_{M,2} & \widehat{m}_{M,3} & \widehat{m}_{M,4} & \end{pmatrix}. \tag{3.4.26}$$

The terms in the second but last row are given by

$$\begin{aligned} \widehat{m}_{M-1,1} &= f_M\left(\theta - u^2\right), \\ \widehat{m}_{M-1,2} &= 2uf_M, \\ \widehat{m}_{M-1,3} &= -f_M. \end{aligned}$$

Note, that for $M = 4$, the three entries above need to be multiplied by 6, due to the variable transformation that basically scales the fourth equation. This does not affect the other cases with $M \neq 4$.

The additional entries in the last row are defined as

$$\begin{aligned} \widehat{m}_{M,1} &= \frac{\theta^2 f_{M-1} - u^3 f_M - \theta u(uf_{M-1} - 3f_M)}{M}, \\ \widehat{m}_{M,2} &= \frac{(-3\theta f_M + 3u^2 f_M + 2\theta u f_{M-1})}{M}, \\ \widehat{m}_{M,3} &= -\frac{(\theta f_{M-1} + 3uf_M)}{M}, \\ \widehat{m}_{M,4} &= \frac{f_M}{M}. \end{aligned}$$

Similarly as in the shallow water model, we use a linear path with one Gauss quadrature point to compute the Roe linearization of the system matrix and for the numerical treatment of the source term $S_{i+\frac{1}{2}}$ we simply evaluate $S$ in $\frac{W_{M_i}+W_{M_{i+1}}}{2}$. For more details on the proper numerical discretization of the nonconservative QBME system, we refer to [102].

#### 3.4.2.3 Test 2: Shock tube case

The one-dimensional shock tube problem is considered by choosing the initial conditions

$$W_M(0,x) = \begin{cases} W_M^l & \text{if } x < 0, \\ W_M^r & \text{if } x > 0, \end{cases} \tag{3.4.27}$$

and non-linear relaxation time $\tau = \frac{\mathrm{Kn}}{\rho}$.

We consider the system in primitive and partially-conservative variables. According to the tests in [26], the left and right states are chosen as

$$W_M^l = (7,0,1,0,\ldots,0)^T, \qquad W_M^r = (1,0,1,0,\ldots,0)^T, \tag{3.4.28}$$

or equivalently

$$U_M^l = (7, 0, 7, 0, \ldots, 0)^T, \qquad U_M^r = (1, 0, 1, 0, \ldots, 0)^T, \tag{3.4.29}$$

corresponding to a jump in density at the discontinuity at $x = 0$.

We consider $\mathrm{Kn} = 0.05$ representing a relatively small Knudsen number close to the continuum flow regime and we will take $x \in [-2, 2]$, $CFL = 0.3$, and $t_{end} = 0.3s$.

### Primitive variables

We are going to compare the runtime for $M = 5$ (i.e. the system has 6 equations) and $M = 11$ (i.e. the system has 12 equations) in primitive variables. We show in Figure 3.1 the numerical solution of the problem (3.4.27) using both ways of writing the Roe scheme in order to see that both give the same result. A detailed comparison of numerical solutions for this test case can be found in [102]. In Tables 3.2 and 3.3 we show the CPU runtimes in (s) for different discretizations of the domain and for $M = 5$ and $M = 11$, respectively, using the standard Roe scheme and the new Newton form. Here we observe a big difference in CPU times between both ways of writing the Roe scheme, due to the larger complexity of the model and the additional number of equations. We see that taking 12 variables ($M = 11$) the difference is getting bigger.

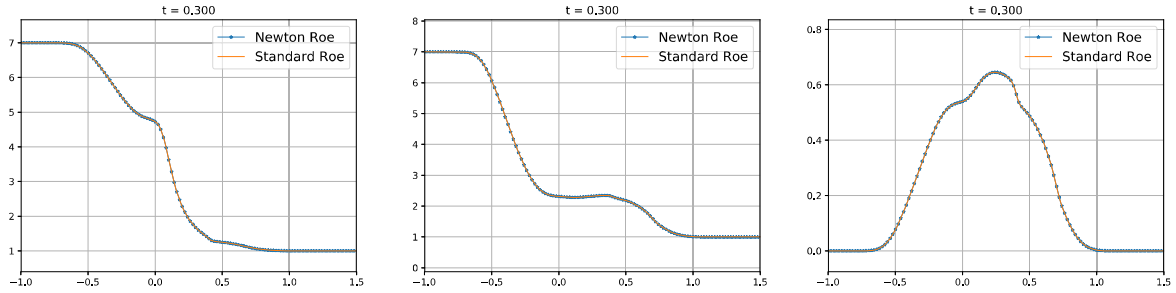



Figure 3.1: Numerical solution of the problem (3.4.27) computed in the primitive variables with $M = 5$, 1000 cells, $CFL = 0.3$ at time $t = 0.3$. Starting from the left the density $\rho$, pressure $p = \rho\theta$ and velocity $u$ are plotted.

In Table 3.4 and in Figure 3.2 we observe that for 1000 cells of the domain, as we increase the number of moments $M$, the speedup of the new Newton Roe scheme is increasing.

### Partially-conservative variables

We are going to compare the runtime for $M = 5$ (i.e. the system has 6 equations) and $M = 13$ (i.e. the system has 14 equations) in partially-conservative variables. This time

| #Cells | Standard Roe | Newton Roe | Speedup |
|---|---|---|---|
| 125 | 0.31 | 0.22 | 1.40 |
| 250 | 0.62 | 0.36 | 1.74 |
| 500 | 1.61 | 0.63 | 2.56 |
| 1000 | 4.14 | 1.49 | 2.79 |

Table 3.2: Test 2: CPU times in (s) for different number of cells of the domain obtained for the usual Roe scheme and the Newton Roe scheme with $M = 5$ (average of 5 runs) and using primitives variables.

| #Cells | Standard Roe | Newton Roe | Speedup |
|---|---|---|---|
| 125 | 0.77 | 0.38 | 2.04 |
| 250 | 2.10 | 0.73 | 2.88 |
| 500 | 7.40 | 2.55 | 2.91 |
| 1000 | 21.40 | 5.92 | 3.61 |

Table 3.3: Test 2: CPU times in (s) for different number of cells of the domain obtained for the usual Roe scheme and the Newton Roe scheme with $M = 11$ (average of 5 runs) and using primitives variables.

| M | Standard Roe | Newton Roe | Speedup |
|---|---|---|---|
| 5 | 4.14 | 1.44 | 2.03 |
| 7 | 8.14 | 2.85 | 2.86 |
| 9 | 14.78 | 4.60 | 3.22 |
| 11 | 21.40 | 5.92 | 3.61 |

Table 3.4: Test 2: CPU times in (s) for 1000 cells of the domain obtained for the usual Roe scheme and the Newton Roe scheme with different number of moments $M$ (average of 5 runs) and using primitives variables.

we are not going to show the results of the numerical schemes because the standard Roe scheme and the Newton Roe scheme again give the same solution. A detailed comparison of numerical solutions for this test case can be found in [102]. In Tables 3.5 and 3.6 we show the CPU runtimes in (s) for different discretizations of the domain and for $M = 5$ and $M = 13$, respectively, using the standard Roe scheme and the new Newton form. Again we observe a big difference in CPU times between both ways of writing the Roe scheme and this difference is even bigger than when we were using primitive variables because with partially-conservative variables we have to add the computation of the eigenvectors as was seen in (3.4.23). We see that taking 14 variables ($M = 13$) the difference is increasing.

In Table 3.7 and in Figure 3.3 we observe that for 1000 cells of the domain, as we increase the number of moments $M$, the speedup of the new Newton Roe scheme is

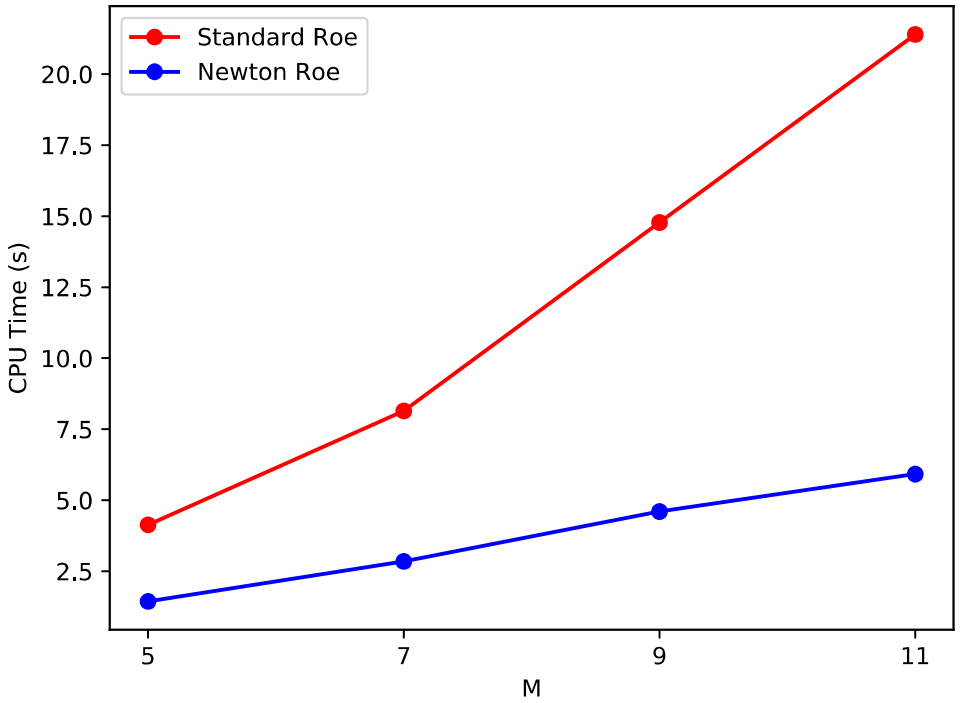



Figure 3.2: Number of moments $M$ vs CPU time (s) for the standard Roe scheme and its Newton's form using primitive variables.

| #Cells | Standard Roe | Newton Roe | Speedup |
|---|---|---|---|
| 250 | 0.72 | 0.30 | 2.37 |
| 500 | 2.15 | 0.65 | 3.29 |
| 1000 | 5.30 | 1.65 | 3.21 |
| 2000 | 18.32 | 5.53 | 3.32 |

Table 3.5: Test 2: CPU times in (s) for different number of cells of the domain obtained for the usual Roe scheme and the Newton Roe scheme with $M = 5$ (average of 5 runs) and using partially-conservative variables.

increasing. As observed before, the differences between both ways of writing the Roe scheme are bigger using the partially-conservative variables. The new Newton Roe solver yields a significant speedup in all test cases for the QBME model.

**Remark 3.4.2.** *We only present results for odd $M$ here as the computation using even $M$ leads to very different runtimes in our simulation software, due to specifications of the underlying linear algebra libraries. However, the resulting flow solutions are still the same*

| #Cells | Standard Roe | Newton Roe | Speedup |
|---|---|---|---|
| 250 | 3.70 | 1.06 | 3.49 |
| 500 | 12.23 | 2.87 | 4.27 |
| 1000 | 40.62 | 9.09 | 4.47 |
| 2000 | 166.48 | 39.10 | 4.26 |

Table 3.6: Test 2: CPU times in (s) for different number of cells of the domain obtained for the usual Roe scheme and the Newton Roe scheme with $M = 13$ (average of 5 runs) and using partially-conservative variables.

| M | Standard Roe | Newton Roe | Speedup |
|---|---|---|---|
| 5 | 18.32 | 5.53 | 3.32 |
| 7 | 39.10 | 11.10 | 3.52 |
| 9 | 71.97 | 19.23 | 3.74 |
| 11 | 106.73 | 25.84 | 4.13 |
| 13 | 166.48 | 39.10 | 4.26 |

Table 3.7: Test 2: CPU times in (s) for 2000 number of cells of the domain obtained for the usual Roe scheme and the Newton Roe scheme with different number of moments $M$ (average of 5 runs) and using partially-conservative variables.

*as for the standard Roe scheme.*



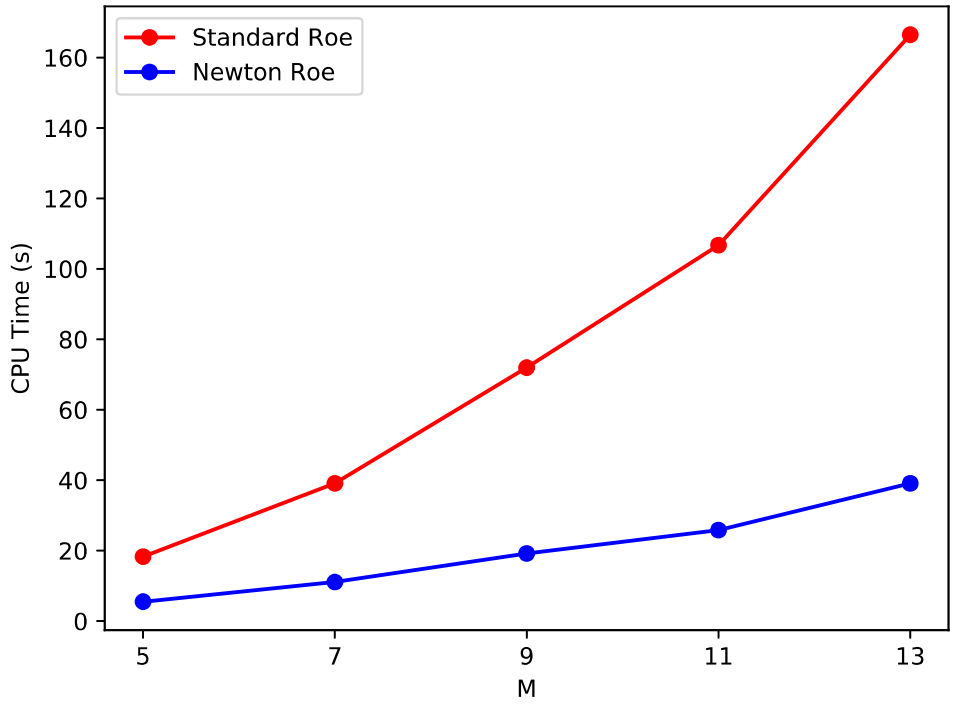



Figure 3.3: Number of moments $M$ vs CPU time (s) for the standard Roe scheme and its Newton's form using partially-conservative variables.

# Chapter 4

# Well-balanced methods for relativistic fluids on a Schwarzschild background

In this chapter we are interested in the long time behavior of relativistic compressible fluid flows on a Schwarzschild black hole background. The flow is assumed to enjoy spherical symmetry and therefore we deal with nonlinear hyperbolic systems of partial differential equations (PDEs) in one space variable. The goal is, first, designing and testing numerically finite volume methods that are well-balanced; and, in second place, to perform a through investigation of the behavior of the solutions and numerically infer definite conclusions about the long-time behavior of such flows. Our study should provide first and useful insights for, on the one hand, further development concerning the mathematical analysis of the models and, on the other hand, further investigations to the same problem in higher dimensions without symmetry restriction.

We will treat first a simplified model, namely the *relativistic Burgers-Schwarzschild model* [114], and next the *relativistic Euler-Schwarzschild model*. The flow is assumed to enjoy spherical symmetry so that only models with one spatial coordinate (the distance to the center of the black hole) are considered throughout. Our purpose in this chapter is to design shock-capturing schemes that are high-order accurate and well-balanced. These results are based on earlier investigations on this problem by LeFloch et al [62, 108, 114] and extend to the present problem the well-balanced methodology in Castro and Parés [47], recalled in Section 1.2.6, in order to properly take the Schwarzschild curved geometry into account. For earlier work on well-balanced schemes we also refer to [40, 143, 144] and, concerning the design of geometry-preserving schemes, we refer for instance to [11, 62, 70, 81, 142, 162, 25, 34] and the references therein.

The chapter is structured as follows: in Section 4.1 we introduce the two models in

consideration, the relativistic Burgers equation and the relativistic Euler equations posed on a Schwarzschild background. Then, in Section 4.2 we recall the strategy explained in Section 1.2.6 and we extend it for developing high-order well-balanced schemes for both models in consideration. In Section 4.3 this strategy is applied to the relativistic Burgers-Schwarzschild equation to obtain first, second and third order well-balanced methods, and in Section 4.4 several numerical tests allow us to investigate the efficiency, accuracy, and robustness of the proposed algorithms and to obtain some conclusions. Finally, in sections 4.5 and 4.6 we do the same with the relativistic Euler-Schwarzschild equations. In particular, by relying on these extensive tests we demonstrate that the proposed schemes are numerically well-balanced and we discuss the importance of this property in order to have reliable results. We study the long time behavior of the Burgers or Euler equations and discuss the roles of initial condition imposed at the boundary. We also describe how steady shocks behave under small or large perturbations. The content of this chapter was introduced by LeFloch, Parés and Pimentel-García, it is available in the *arXiv* repository and was submitted in December 2020 to *Journal of Scientific Computing*, see [110]. It is currently in the modification phase after the first reports from the reviewers.

## 4.1 Models of interest

The *relativistic Burgers-Schwarzschild model* [114] is given by:

$$\partial_t v + \partial_r \left( \left( 1 - \frac{2M}{r} \right) \frac{v^2 - 1}{2} \right) = \frac{2M}{r^2}(v^2 - 1), \qquad r > 2M,\ v = v(t,r) \in [-1,1]. \tag{4.1.1}$$

Here the unknown is a scalar function which represents the normalized velocity of the flow. This equation clearly takes the form

$$v_t + F(v,r)_r = S(v,r), \tag{4.1.2}$$

where

$$F(v,r) = \left( 1 - \frac{2M}{r} \right) \frac{v^2 - 1}{2}, \qquad S(v,r) = \frac{2M}{r^2}(v^2 - 1), \tag{4.1.3}$$

$M > 0$ is a coefficient representing the mass of the black hole.

We will next consider the *relativistic Euler-Schwarzschild model*

$$\begin{aligned}
&\partial_t \left( \frac{1+k^2v^2}{1-v^2}\rho \right) + \partial_x \left( \left( 1 - \frac{2M}{r} \right) \frac{1+k^2}{1-v^2}\rho v \right) = -\frac{2}{r} \left( 1 - \frac{2M}{r} \right) \frac{1+k^2}{1-v^2}\rho v, \\
&\partial_t \left( \frac{1+k^2}{1-v^2}\rho v \right) + \partial_x \left( \left( 1 - \frac{2M}{r} \right) \frac{v^2+k^2}{1-v^2}\rho \right) = \frac{-2r+5M}{r^2} \frac{v^2+k^2}{1-v^2}\rho - \frac{M}{r^2} \frac{1+k^2v^2}{1-v^2}\rho + 2\frac{r-2M}{r^2}k^2\rho,
\end{aligned} \tag{4.1.4}$$

in which the unknowns are the fluid density $\rho$ and the normalized velocity $v(t,r) \in (-1,1)$, defined for all $r > 2M$, while $k \in (-1,1)$ denotes the (constant) speed of sound. (The

limiting values $v = \pm 1$ will be reached at the boundary $r = 2M$ only.) We can write this system in the form

$$V_t + F(V, r)_r = S(V, r), \tag{4.1.5}$$

where

$$V = \begin{pmatrix} V^0 \\ V^1 \end{pmatrix} = \begin{pmatrix} \dfrac{1+k^2v^2}{1-v^2}\rho \\ \dfrac{1+k^2}{1-v^2}\rho v \end{pmatrix}, \qquad F(V, r) = \begin{pmatrix} \left(1 - \dfrac{2M}{r}\right)\dfrac{1+k^2}{1-v^2}\rho v \\ \left(1 - \dfrac{2M}{r}\right)\dfrac{v^2+k^2}{1-v^2}\rho \end{pmatrix}, \tag{4.1.6}$$

$$S(V, r) = \begin{pmatrix} -\dfrac{2}{r}\left(1 - \dfrac{2M}{r}\right)\dfrac{1+k^2}{1-v^2}\rho v \\ \dfrac{-2r+5M}{r^2}\dfrac{v^2+k^2}{1-v^2}\rho - \dfrac{M}{r^2}\dfrac{1+k^2v^2}{1-v^2}\rho + 2\dfrac{r-2M}{r^2}k^2\rho \end{pmatrix}, \tag{4.1.7}$$

where

$$v = \frac{1 + k^2 - \sqrt{(1+k^2)^2 - 4k^2\left(\frac{V^1}{V^0}\right)^2}}{2k^2\frac{V^1}{V^0}}, \qquad \rho = \frac{V^1(1-v^2)}{v(1+k^2)}. \tag{4.1.8}$$

This is a strictly hyperbolic system and the eigenvalues of the Jacobian of the flux function are (cf. for instance [114])

$$\mu_\pm = \left(1 - \frac{2M}{r}\right)\frac{v \pm k}{1 \pm k^2 v}. \tag{4.1.9}$$

As usual, a state $(\rho, v)$, by definition, is

- *sonic* if one of the eigenvalues vanishes, i.e. if $|v| = k$;
- *supersonic* if both eigenvalues have the same sign, i.e. if $|v| > k$;
- *subsonic* if the eigenvalues have different signs, i.e. if $|v| < k$.

We will need to distinguish between these regimes in order to design a robust scheme.

## 4.2 A well-balanced methodology

Both problems of interest are of the general form

$$V_t + F(V, r)_r = S(V, r), \qquad r > 2M, \tag{4.2.1}$$

with unknown $V = V(t, r) \in \mathbb{R}^N$ and $N = 1$ or 2. Systems of this form have non-trivial stationary solutions, which satisfy the ODE

$$F(V, r)_r = S(V, r). \tag{4.2.2}$$

In this section we recall the family of well-balanced numerical methods that we introduced in Section 1.2.6 following the strategy in [47] and we extend it to these models.

We consider semi-discrete finite volume numerical methods of the form

$$\frac{dV_i}{dt} = -\frac{1}{\Delta r}\Big(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} - \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} S(\mathbb{P}_i^t(r), r)\, dr\Big), \tag{4.2.3}$$

and the following notation is used:

- $I_i = \Big[r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}\Big]$ are the computational cells, whose length $\Delta r$ is supposed to be constant for simplicity.
- $V_i(t)$ is the approximation of the average of the exact solution at the $i$th cell at time $t$, that is,
$$V_i(t) \cong \frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} V(r,t)\, dr. \tag{4.2.4}$$
- $\mathbb{P}_i^t(r)$ is the approximation of the solution at the $i$th cell given by a high-order reconstruction operator from the sequence of cell averages $\{V_i(t)\}$:
$$\mathbb{P}_i^t(r) = \mathbb{P}_i^t(r; \{V_j(t)\}_{j\in\mathcal{S}_i}), \tag{4.2.5}$$
where $\mathcal{S}_i$ denotes the set of indices of the cells belonging to the stencil of the $i$th cell.
- $F_{i+\frac{1}{2}} = \mathbb{F}(V_{i+\frac{1}{2}}^{t,-}, V_{i+\frac{1}{2}}^{t,+}, r_{i+\frac{1}{2}})$, where $V_{i+\frac{1}{2}}^{t,\pm}$ are the reconstructed states at the interfaces, i.e.
$$V_{i+\frac{1}{2}}^{t,-} = \mathbb{P}_i^t(r_{i+\frac{1}{2}}), \quad V_{i+\frac{1}{2}}^{t,+} = \mathbb{P}_{i+1}^t(r_{i+\frac{1}{2}}), \tag{4.2.6}$$
and $\mathbb{F}$ is a consistent first-order numerical flux.

As we proved in Section 1.2.6, if the reconstruction operator is well-balanced for a continuous stationary solution $V^*$ of (4.2.2) then the numerical method is also well-balanced for $V^*$. Let us recall some definitions: given a stationary solution $V^*$ of (4.2.2):

- the numerical method (4.2.3) is said to be well-balanced for $V^*$ if the vector of cell averages of $V^*$ is an equilibrium of the ODE system (4.2.3);
- the reconstruction operator is said to be well-balanced for $V^*$ if
$$\mathbb{P}_i(r) = V^*(r), \qquad r \in [r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}], \tag{4.2.7}$$
where $\mathbb{P}_i$ is the approximation of $V^*$ obtained by applying the reconstruction operator to the vector of cell averages of $V^*$.

We follow the procedure explained in Section 1.2.6 in order to design a well-balanced reconstruction operator $\mathbb{P}_i$ on the basis of a standard operator $\mathbb{Q}_i$.

1. Seek the stationary solution $V_i^*(x)$ such that

$$\frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} V_i^*(r)\,dr = V_i. \tag{4.2.8}$$

2. Apply the reconstruction operator to the cell values $\{W_j\}_{j\in S_i}$ given by

$$W_j = V_j - \frac{1}{\Delta r}\int_{r_{j-\frac{1}{2}}}^{r_{j+\frac{1}{2}}} V_i^*(r)\,dr, \quad j\in\mathcal{S}_i, \tag{4.2.9}$$

   in order to obtain $\mathbb{Q}_i(r) = \mathbb{Q}_i(r;\{W_j\}_{j\in\mathcal{S}_i})$.

3. Define finally

$$\mathbb{P}_i(r) = V_i^*(r) + \mathbb{Q}_i(r). \tag{4.2.10}$$

As we said, the reconstruction operator $\mathbb{P}_i$ in (4.2.10) is well-balanced for every stationary solution provided that the reconstruction operator $\mathbb{Q}_i$ is exact for the zero function. Moreover, if $\mathbb{Q}_i$ is conservative, then $\mathbb{P}_i$ is conservative, in the sense that

$$\frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} \mathbb{P}_i(r)\,dr = V_i, \tag{4.2.11}$$

and $\mathbb{P}_i$ has the same accuracy as $\mathbb{Q}_i$ if the stationary solutions are sufficiently regular.

As we pointed out in Section 1.2.6, the well-balanced property of the method can be lost if a quadrature formula is used to compute the integral appearing at the right-hand side of (4.2.3). In order to circumvent this difficulty, in [47] it is proposed to rewrite the methods as in (1.2.115), that in this case reduces to:

$$\begin{aligned}\frac{dV_i}{dt} = &-\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F\left(V_i^{t,*}(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right) - F_{i-\frac{1}{2}} + F\left(V_i^{t,*}(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}}\right)\right)\\ &+\frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}}\left(S(\mathbb{P}_i^t(r), r) - S(V_i^{t,*}(r), r)\right)\,dr,\end{aligned} \tag{4.2.12}$$

where $V_i^{t,*}$ is the stationary solution found in (4.2.8) at the $i$th cell at time $t$. In this equivalent form, a quadrature formula can be applied to the integral without losing the well-balanced property , and this leads to a numerical method of the form (1.2.117), that in this case it reduces to:

$$\begin{aligned}\frac{dV_i}{dt} = &-\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F\left(V_i^{t,*}(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right) - F_{i-\frac{1}{2}} + F\left(V_i^{t,*}(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}}\right)\right)\\ &+\sum_{l=0}^{q}\alpha_l\left(S(\mathbb{P}_i^t(r_{i,l}), r_{i,l}) - S(V_i^{t,*}(r_{i,l}), r_{i,l})\right).\end{aligned} \tag{4.2.13}$$

Here, $\alpha_0, \ldots, \alpha_q$, $r_{i,0}, \ldots, r_{i,q}$ represent the weights and the nodes of the chosen quadrature formula, whose order of accuracy must be greater or equal to the one of the reconstruction operator.

As we saw, if the quadrature formula is used to compute the averages of the initial condition as well: $V_i^0 = \frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} V_0(r)\,dr$, the reconstruction procedure has to be modified to preserve the well-balanced property: Steps 1 and 2 have to be replaced by the following:

1. Seek the stationary solution $V_i^*(x)$ such that

$$\sum_{l=0}^{q} \alpha_l V_i^*(r_{i,l}) = V_i. \tag{4.2.14}$$

2. Apply the reconstruction operator to the cell values $\{W_j\}_{j\in S_i}$ given by

$$W_j = V_j - \sum_{l=0}^{q} \alpha_l V_i^*(r_{j,l}), \quad j \in \mathcal{S}_i.$$

Finally, for a first-order method we consider the trivial reconstruction operator, given the cell averages $\{V_i\}$ of a function $V$, and we consider the piecewise constant approximation of $V$

$$\mathbb{Q}_i(r, V_i) = V_i, \qquad r \in [r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}]. \tag{4.2.15}$$

It can be easily checked that the numerical method then reduces to

$$\frac{dV_i}{dt} = -\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F\left(V_i^{t,*}(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right) - F_{i-\frac{1}{2}} + F\left(V_i^{t,*}(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}}\right)\right), \tag{4.2.16}$$

where $F_{i+\frac{1}{2}} = \mathbb{F}\left(V_i^*(r_{i+\frac{1}{2}}), V_{i+1}^*(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right)$. Moreover, if the initial averages are computed with the midpoint rule, namely

$$V_{i,0} = \frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} V_0(r)\,dr \cong V_0(r_i), \tag{4.2.17}$$

then at the first step of the reconstruction procedure, the problem (4.2.14) reduces to finding the stationary solution satisfying

$$V_i^*(r_i) = V_i. \tag{4.2.18}$$

For a second-order method the mid-point rule can also be selected, so that the problem (4.2.14) reduces again to (4.2.18).

## 4.3 Burgers-Schwarzschild model: designing the numerical algorithm

### 4.3.1 Preliminaries

For the Burgers-Schwarzschild equation (4.1.1), the steady state solutions are of the form

$$v^*(r) = \pm\sqrt{1 - K^2\left(1 - \frac{2M}{r}\right)}, \quad K > 0. \tag{4.3.1}$$

In Figure 4.1 we plot the steady solutions for several values of $K^2$. The domain of definition of these stationary solutions is

$$D_K = \begin{cases} [2M, \infty), & K^2 \leq 1, \\ \left[2M, \dfrac{2MK^2}{K^2 - 1}\right], & K^2 > 1. \end{cases} \tag{4.3.2}$$

It can be easily checked that, given a pair $(K, r^*)$ such that $r^* \in D_K$, the discontinuous function defined in $D_K$ by

$$w^*(r) = \begin{cases} \sqrt{1 - K^2\left(1 - \frac{2M}{r}\right)}, & r \leq r^*, \\ -\sqrt{1 - K^2\left(1 - \frac{2M}{r}\right)}, & \text{otherwise}, \end{cases} \tag{4.3.3}$$

is an entropy weak stationary solution of the Burgers-Schwarzschild model.

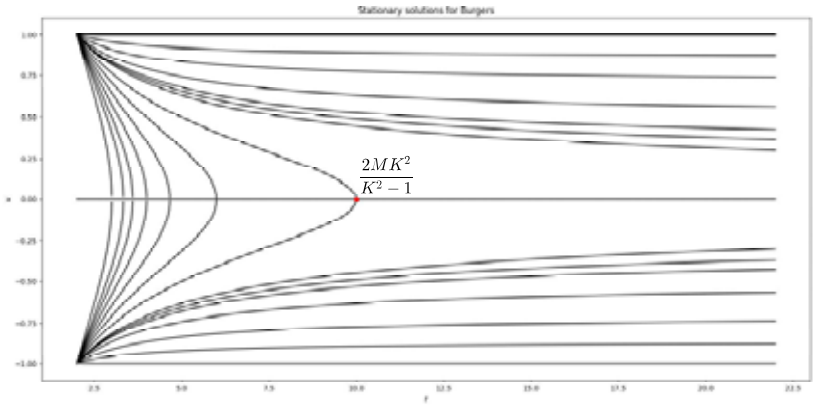



Figure 4.1: Steady solutions to the Burgers model.

### 4.3.2 First-order method

If the midpoint rule is used to compute the initial averages, at the first step of the reconstruction procedure one has to search for $K_i^2$ such that

$$\sqrt{1 - K_i^2 \left(1 - \frac{2M}{r_i}\right)} = |v_i|. \tag{4.3.4}$$

There is always a unique solution given by

$$\widetilde{K}_i^2 = \frac{1 - v_i^2}{1 - \dfrac{2M}{r_i}}, \tag{4.3.5}$$

so that the stationary solution is

$$v_i^*(r) = \operatorname{sgn}(v_i)\sqrt{1 - \widetilde{K}_i^2 \left(1 - \frac{2M}{r}\right)}. \tag{4.3.6}$$

In order to apply the numerical method (4.2.12), this stationary solution has to be computed at $r_{i\pm\frac{1}{2}}$ and this is only possible if $r_{i+\frac{1}{2}} \in D_{\widetilde{K}_i}$, that is, provided

$$\widetilde{K}_i^2 \leq 1 \text{ or } \left(\widetilde{K}_i^2 > 1 \text{ and } r_{i+\frac{1}{2}} \leq \frac{2M\widetilde{K}_i^2}{\widetilde{K}_i^2 - 1}\right). \tag{4.3.7}$$

If this condition is satisfied, then the numerical method (4.2.16) can be used.

If this condition is not satisfied, then the standard trivial reconstruction is considered, i.e. $\mathbb{Q}_i(r) = v_i$. The numerical method writes then as follows:

$$\frac{dv_i}{dt} = -\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} - S(v_i, r_i)\right), \tag{4.3.8}$$

where $F_{i+\frac{1}{2}} = \mathbb{F}(v_i, v_{i+1}, r_{i+\frac{1}{2}})$.

Summing up, the expression of the semi-discrete first-order method reads

$$\frac{dv_i}{dt} = -\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} - S_i\right), \tag{4.3.9}$$

where

$$S_i = \begin{cases} F(v_i^*(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}) - F(v_i^*(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}})), & \text{if (4.3.7) is satisfied;} \\ S(v_i, r_i), & \text{otherwise.} \end{cases} \tag{4.3.10}$$

The forward Euler method (1.2.51) is used for the time discretization.

We emphasize that, if (4.3.7) is not satisfied, then $v_i$ cannot be the point value of a stationary solution defined in the computational domain, so that the use of the standard reconstruction does not destroy the well-balanced property of the method, since in this case there is no stationary solution to preserve.

### 4.3.3 Second-order method

Let us suppose again that the midpoint rule is used to compute the cell averages and that the minmod reconstruction operator (1.2.55) is considered. The stationary solution $v_i^*$ selected at the first stage of the reconstruction procedure is again (4.3.6) with $\tilde{K}_i$ given by (4.3.5). In order to compute the reconstructions, this stationary solution has to be computed at the points $r_{i-1}$, $r_{i-\frac{1}{2}}$, $r_{i+\frac{1}{2}}$, $r_{i+1}$ so that the following condition has to be satisfied $r_{i+1} \in D_{\widetilde{K}_i}$, i.e.

$$\widetilde{K}_i^2 \leq 1 \text{ or } \left( \widetilde{K}_i^2 > 1 \text{ and } r_{i+1} \leq \frac{2M\widetilde{K}_i^2}{\widetilde{K}_i^2 - 1} \right). \tag{4.3.11}$$

If this condition is satisfied, the following step of the reconstruction procedure consists in computing the fluctuations:

$$\begin{aligned} w_{i-1} &= v_{i-1} - \operatorname{sgn}(v_{i-1}) \sqrt{1 - \widetilde{K}_i^2 \left(1 - \frac{2M}{r_{i-1}}\right)}, \\ w_i &= v_i - \operatorname{sgn}(v_i) \sqrt{1 - \widetilde{K}_i^2 \left(1 - \frac{2M}{r_i}\right)} = 0, \\ w_{i+1} &= v_{i+1} - \operatorname{sgn}(v_{i+1}) \sqrt{1 - \widetilde{K}_i^2 \left(1 - \frac{2M}{r_{i+1}}\right)}. \end{aligned} \tag{4.3.12}$$

Then the reconstruction is defined as

$$\mathbb{P}_i(r) = v_i^*(r) + \operatorname{minmod}\left( \frac{w_{i+1} - w_i}{\Delta r}, \frac{w_{i+1} - w_{i-1}}{2\Delta r}, \frac{w_i - w_{i-1}}{\Delta r} \right) (r - r_i), \tag{4.3.13}$$

where

$$\operatorname{minmod}(a,b,c) = \begin{cases} \min\{a,b,c\} & if \quad a,b,c > 0, \\ \max\{a,b,c\} & if \quad a,b,c < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3.14}$$

Once the well-balanced reconstruction operator has been computed, the form (4.2.12) of the numerical method is considered and the midpoint rule is used again to approximate the integral. Observe however that, in this case:

$$\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} \left( S(\mathbb{P}_i^t(r), r) - S(V_i^{t,*}(r), r) \right) dr \cong \Delta r \left( S(\mathbb{P}_i^t(r_i), r_i) - S(V_i^{t,*}(r_i), r_i) \right) = 0.$$

Therefore, the expression (4.2.13) reduces again to (4.2.16) with $F_{i+\frac{1}{2}} = \mathbb{F}(v_{i+\frac{1}{2}}^{t,-}, v_{i+\frac{1}{2}}^{t,+}, r_{i+\frac{1}{2}})$.

If (4.3.11) is not satisfied, then the standard MUSCL reconstruction is applied, namely

$$\mathbb{Q}_i(r) = v_i + \text{minmod}\left(\frac{v_{i+1}-v_i}{\Delta r}, \frac{v_{i+1}-v_{i-1}}{2\Delta r}, \frac{v_i - v_{i-1}}{\Delta r}\right)(r - r_i). \tag{4.3.15}$$

The expression of the numerical method is given again by (4.3.9)-(4.3.10) with the difference that the second-order reconstructions are used now to compute the numerical fluxes. The TVDRK2 method (1.2.52) is used in order to discretize the equations in time.

Observe that, according to the well-balanced reconstruction procedure described in the previous section, the fluctuations to be reconstructed should be in this case

$$w_j = v_j - v_i^*(r_j) = v_j - \text{sgn}(v_i)\sqrt{1 - \widetilde{K}_i^2\left(1 - \frac{2M}{r_j}\right)}, \quad j = i-1, i, i+1,$$

but in (4.3.12) the sign of $v_i$ has been replaced by that of $v_j$. This modification allows one to preserve not only the continuous stationary solutions solution but also the discontinuous stationary solutions of the family (4.3.3).

### 4.3.4 Third-order method

In order to design a third-order method the *CWENO* reconstruction of order 3 (1.2.65) will be considered and the two-point Gauss quadrature will be used to compute the initial averages and the integrals coming from the source term:

$$v_i^0 = \frac{1}{2}(v_0(r_{i,0}) + v_0(r_{i,1})),$$

where

$$r_{i,0} = r_{i-\frac{1}{2}} + \frac{\Delta r}{2}\left(-\sqrt{\frac{1}{3}} + 1\right), \quad r_{i,1} = r_{i-\frac{1}{2}} + \frac{\Delta r}{2}\left(\sqrt{\frac{1}{3}} + 1\right).$$

Therefore, at the first step of the reconstruction procedure one has to look for $K_i^2$ such that:

$$\frac{1}{2}\left(\sqrt{1 - K_i^2\left(1 - \frac{2M}{r_{i,0}}\right)} + \sqrt{1 - K_i^2\left(1 - \frac{2M}{r_{i,1}}\right)}\right) = |v_i|. \tag{4.3.16}$$

If we define the function

$$g(x) = \frac{1}{2}\left(\sqrt{1 - x\left(1 - \frac{2M}{r_{i,0}}\right)} + \sqrt{1 - x\left(1 - \frac{2M}{r_{i,1}}\right)}\right), \quad x \geq 0,$$

it can be easily verified that $g$ is a positive decreasing function defined in the interval $[0, K_{i,c}^2]$ where $K_{i,c}^2 = \left(1 - \frac{2M}{r_{i,1}}\right)^{-1}$. Therefore there are two possibilities:

- if $|v_i| \in [g(K_{i,c}^2), 1]$, there is a unique $\widetilde{K}_i^2$ satisfying (4.3.16);
- in other case, (4.3.16) has no solution.

If (4.3.16) is satisfied, the corresponding stationary solution

$$v_i^*(r) = \mathrm{sgn}(v_i)\sqrt{1 - \widetilde{K}_i^2\left(1 - \frac{2M}{r}\right)}$$

has to be computed in the points $\{r_{i-1,0}, r_{i-1,1}, r_{i+1,0}, r_{i+1,1}\}$ in the reconstruction procedure. Therefore, these points have to be in the interval of definition of $v_i^*$, and this happens if $r_{i+1,1} \in D_{\widetilde{K}_i}$. Therefore, the well-balanced reconstruction can be computed only if the following condition is satisfied:

$$|v_i| \in [g(K_{i,c}^2), 1] \text{ and } \left(\widetilde{K}_i^2 \leq 1 \text{ or } \left(\widetilde{K}_i^2 > 1 \text{ and } r_{i+1,1} \leq \frac{2M\widetilde{K}_i^2}{\widetilde{K}_i^2 - 1}\right)\right). \tag{4.3.17}$$

If this condition is satisfied, the fluctuations can be then computed:

$$w_j = v_j - \mathrm{sgn}(v_j)\frac{1}{2}\left(\sqrt{1 - \widetilde{K}_i^2\left(1 - \frac{2M}{r_{j,0}}\right)} + \sqrt{1 - \widetilde{K}_i^2\left(1 - \frac{2M}{r_{j,1}}\right)}\right), \quad j = i-1, i, i+1,$$

and the well-balanced reconstruction is given by

$$\mathbb{P}_i(r) = v_i^*(r) + \mathbb{Q}_i(r; w_{i-1}, w_i, w_{i+1}),$$

where $\mathbb{Q}$ represents the CWENO approximation (1.2.65).

If (4.3.17) is not satisfied, the standard CWENO reconstruction is applied:

$$\mathbb{Q}_i(r) = \mathbb{Q}_i(r; v_{i-1}, v_i, v_{i+1}).$$

The expression of the semi-discrete method is finally (4.3.9) where the numerical fluxes are computed at the reconstructed states and

$$S_i = \begin{cases} F(v_i^*(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}) - F(v_i^*(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}}) \\ + \quad \dfrac{\Delta r}{2}\displaystyle\sum_{j=0,1}\left(S(\mathbb{P}_i(r_{i,j}), r_{i,j}) - S(v_i^*(r_{i,j}), r_{i,j})\right) & \text{if (4.3.17) is satisfied,} \\ \dfrac{\Delta r}{2}\displaystyle\sum_{j=0,1} S(\mathbb{Q}_i(r_{i,j}), r_{i,j}), & \text{otherwise.} \end{cases} \tag{4.3.18}$$

The TVDRK3 method of order 3 (1.2.53) will be used for the time discretization.

### 4.3.5 Preserving the exact averages of the stationary solutions

The three methods presented above can be modified to preserve the exact averages of the stationary solutions instead of its approximation computed with a quadrature formula. To do this, the problem to be solved at the first stage of the well-balanced reconstruction procedure is the following one: find $K_i^2$ such that:

$$\frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i-\frac{1}{2}}}\sqrt{1-K_i^2\left(1-\frac{2M}{r}\right)}dr=|v_i|. \tag{4.3.19}$$

If we define the function

$$g(x)=\frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i-\frac{1}{2}}}\sqrt{1-x\left(1-\frac{2M}{r}\right)}dr=|v_i|,$$

it can be easily checked that it is a decreasing function defined in $[0,K_{e,i}^2]$ where $K_{e,i}^2=\left(1-2M/r_{i+\frac{1}{2}}\right)^{-1}$ and $g(0)=1$. Therefore, (4.3.19) has a unique solution $\widetilde{K}_i^2$ if

$$|v_i|\leq g(K_{e,i}). \tag{4.3.20}$$

The explicit expression of $g$ can be obtained: $g(x)=\frac{1}{\Delta r}\left(f(x,r_{i+\frac{1}{2}})-f(x,r_{i-\frac{1}{2}})\right)$, where the function $f(x,r)$ is equal to

$$\begin{cases} r\sqrt{1-x\left(1-\frac{2M}{r}\right)}+\dfrac{xM}{\sqrt{1-x}}\log\left(x(M-r)+r+\sqrt{1-x}r\sqrt{1-x\left(1-\dfrac{2M}{r}\right)}\right), & 0\leq x<1,\\ 2r\sqrt{\dfrac{2M}{r}}, & x=1,\\ r\sqrt{1-x}\left(1-\frac{2M}{r}\right)-\dfrac{2xM}{\sqrt{x-1}}\tan^{-1}\left(\dfrac{\sqrt{1-x\left(1-\frac{2M}{r}\right)}}{\sqrt{x-1}}\right), & x>1. \end{cases}$$

is a primitive of $\sqrt{1-x\left(1-\frac{2M}{r}\right)}$. Therefore $g(K_{e,i})$ can be explicitly computed.
The well-balanced reconstruction can thus be computed if (4.3.20) is satisfied and the cells of the stencil $\mathcal{S}_i$ are contained in the domain of definition $D_{\widetilde{K}_i}$ of the corresponding stationary solution. Otherwise, the standard reconstruction is applied. The expression of the numerical methods is the same that the ones above.

## 4.4 Burgers-Schwarzschild model: a numerical study

### 4.4.1 Preliminaries

In this section several numerical tests are applied to check the performance of the well-balanced numerical methods introduced in the previous section. We consider the spatial

interval $[2M, L]$ with $M = 1$ and $L = 4$, a 256-point uniform mesh and the CFL number is set to 0.5. At $x = 2M$ we impose $F_{-\frac{1}{2}} = 0$ as boundary condition because $\left(1 - \frac{2M}{r}\right) = 0$. At $x = L$ we will use a transmissive boundary condition using ghost-cells if the initial condition is not a stationary solution; otherwise, the stationary solution is imposed in the ghost-cells. The following numerical flux will be used:

$$F_{i+\frac{1}{2}} = \mathbb{F}(r_{i+\frac{1}{2}}, v_i, v_{i+1}) = \left(1 - \frac{2M}{r_{i+\frac{1}{2}}}\right) \frac{q^2(0; v_i, v_{i+1}) - 1}{2},$$

where $q(\cdot; v_L, v_R)$ is the self-similar solution of the Riemann problem for the standard Burgers equation with the initial condition

$$v_0(r) = \begin{cases} v_L, & r < 0, \\ v_R, & r > 0. \end{cases}$$

In order to check the relevance of the well-balanced property, these methods will be compared with those based on the same numerical fluxes and the standard first-, second- or third-order reconstructions.

### 4.4.2 Stationary solutions

#### Positive stationary solution

We consider the initial condition

$$v_0(r) = \sqrt{\frac{3}{4} + \frac{1}{2r}} \tag{4.4.1}$$

corresponding to the positive stationary solution with $K = \frac{1}{2}$. Table 4.1 shows the error in $L^1$ norm between the initial condition and the numerical solution at time $t = 50$.

| Scheme (256 cells) | Error (1st) | Error (2nd) | Error (3rd) |
|---|---|---|---|
| Well-balanced | 1.13E-14 | 8.72Ee-17 | 7.22E-14 |
| Non well-balanced | 1.89 | 1.61 | 8.78E-02 |

Table 4.1: Well-balanced versus non-well-balanced schemes: $L^1$ errors at $t = 50$ for the Burgers model with the initial condition (4.4.1).

Figure 4.2 compares the numerical solutions obtained with the well-balanced and the non-well-balanced methods: while the three well-balanced methods capture the stationary solution at machine accuracy (what make their graphs indistinguishable), the numerical solutions provided by the non-well-balanced ones depart from the steady state (after a time that increases with the order) and converge to a different equilibrium that takes the value -1 at $r = 2M$. This unexpected behaviour of the non-well-balanced methods is mainly due to the lack of boundary condition at $r = 2M$ that makes particularly difficult to preserve the stationary solutions.

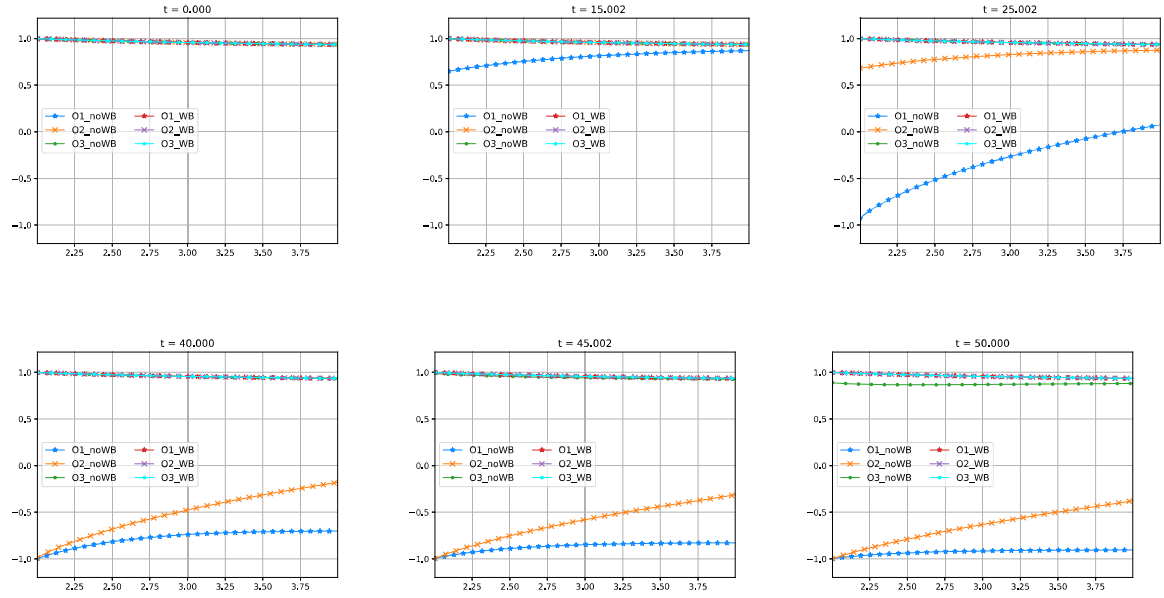



Figure 4.2: Burgers-Schwarzschild model with the initial condition (4.4.1): first-, second- and third-order well-balanced and not-well-balanced methods at various times for variable $v$.

### Negative stationary solution

Let us consider now as initial condition the negative stationary condition corresponding to $K = \frac{1}{2}$:

$$v_0(r) = -\sqrt{\frac{3}{4} + \frac{1}{2r}}. \tag{4.4.2}$$

Figure 4.3 shows the numerical solutions obtained with the different numerical methods. Notice that the scale of the vertical axis is not the same as the one in Figure 4.2: it has been changed so that the difference between the numerical solutions can be better seen. Table 4.2 shows the error in $L^1$ norm between the initial condition and the numerical solution at time $t = 50$.

| Scheme (256 cells) | Error (1st) | Error (2nd) | Error (3rd) |
|---|---|---|---|
| Well-balanced | 6.98E-16 | 1.24E-16 | 4.03E-16 |
| Non well-balanced | 3.92E-02 | 3.20E-07 | 1.63E-10 |

Table 4.2: Well-balanced versus non-well-balanced schemes: $L^1$ errors at $t = 50$ for the Burgers model with the initial condition (4.4.2)

According to Figure 4.3 and Table 4.2 we need more time to see the differences between

the well-balanced and non-well-balanced schemes of order 1,2 and 3 but the errors are again much smaller with the well-balanced schemes for this test. In this case we need more time to see these differences because this negative stationary solution is close to the constant state $v(r) = -1$ where it seems that the non-well-balanced schemes converge.

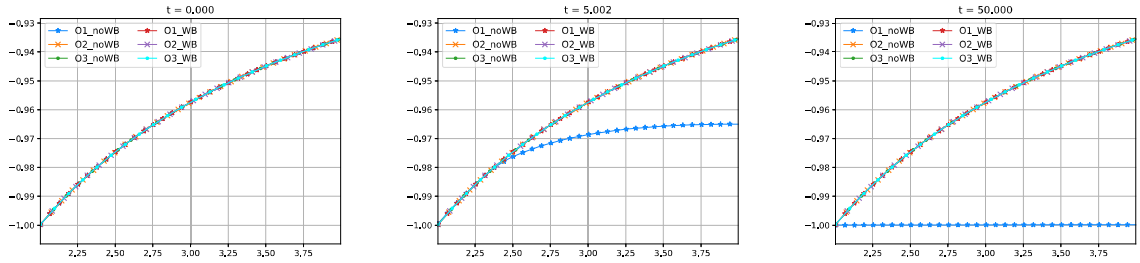



Figure 4.3: Burgers-Schwarzschild model with the initial condition (4.4.2): first-, second- and third-order well-balanced and non-well-balanced methods at selected times for variable $v$.

### Discontinuous stationary entropy weak solution

Let us consider finally the discontinuous initial condition

$$v_0(r) = \begin{cases} \sqrt{\dfrac{3}{4} + \dfrac{1}{2r}} & \text{if } 2 < r < 3, \\ -\sqrt{\dfrac{3}{4} + \dfrac{1}{2r}} & \text{otherwise,} \end{cases} \tag{4.4.3}$$

that is a stationary entropy weak solution of the family (4.3.3). Table 4.3 shows the error in $L^1$ norm between the initial condition and the numerical solution at time $t = 50$.

| Scheme (256 cells) | Error (1st) | Error (2nd) | Error (3rd) |
|---|---|---|---|
| Well-balanced | 8.68E-15 | 8.54E-17 | 7.90E-14 |
| Non well-balanced | 1.02 | 1.09 | 1.09 |

Table 4.3: Well-balanced versus non-well-balanced schemes: $L^1$ errors at $t = 50$ for the Burgers model with the initial condition (4.4.3)

Figure 4.4 shows the differences between the numerical solutions obtained with well-balanced and non-well-balanced methods: again the latter depart from the stationary solution at time that decrease with the order. Tests 1-3 clearly show the need of using well-balanced methods for this equation.

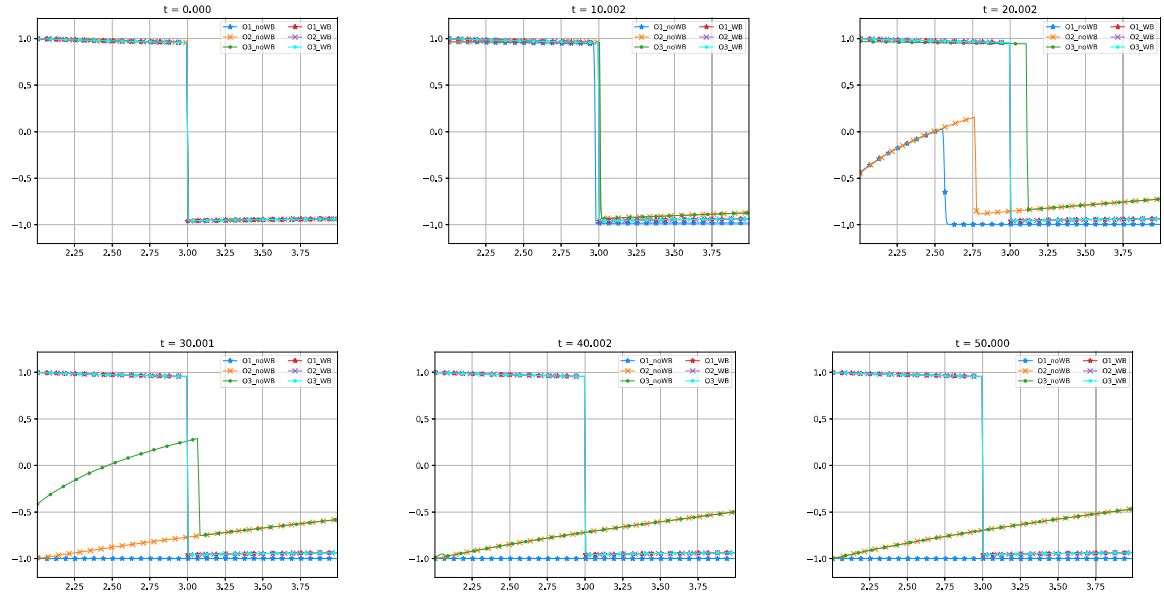



Figure 4.4: Burgers-Schwarzschild model with the initial condition (4.4.3): first-, second- and third-order well-balanced and non-well-balanced methods at selected times for variable $v$.

### 4.4.3 Moving shocks connecting two steady profiles

#### Right-moving shock

We consider now the initial condition

$$v_0(r) = \begin{cases} \sqrt{\dfrac{1}{2} + \dfrac{1}{r}} & \text{if } 2 < r < 2.5, \\ \sqrt{\dfrac{2}{r}} & \text{otherwise} \end{cases} \tag{4.4.4}$$

The corresponding solution consists of a right-moving shock connecting two branches of stationary solutions. Figure 4.5 shows the numerical solutions obtained with the first-, second- and third-order well-balanced methods and a reference solution computed with the first-order standard method using a mesh of 10000 cells. As it can be seen, the well-balanced methods capture correctly the shock with a resolution that increases with the order as expected.

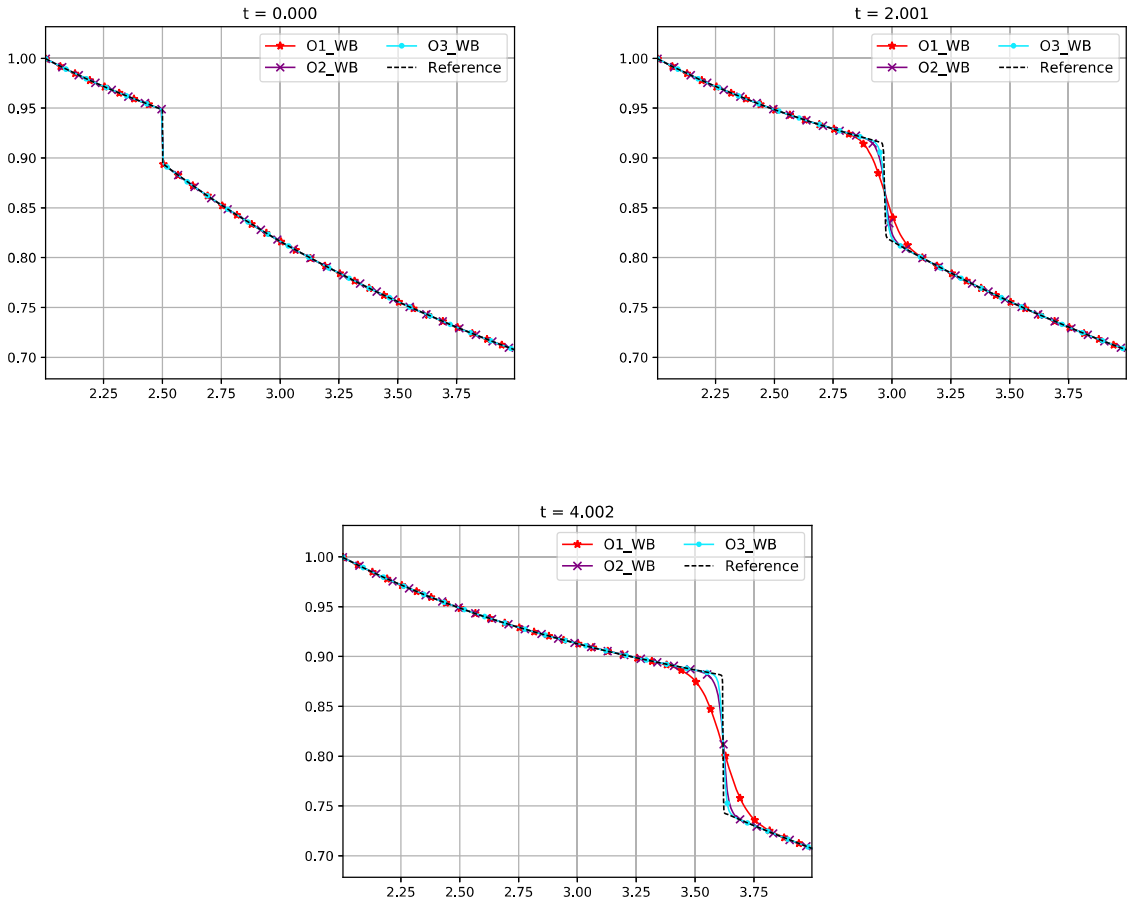



Figure 4.5: Burgers-Schwarzschild model with the initial condition (4.4.4): first-, second- and third-order well-balanced methods at selected times for variable $v$.

### Left-moving shock

Similar conclusions can be drawn for the left-moving shock linking two branches of stationary solutions that generates from the initial condition:

$$v_0(r) = \begin{cases} -\sqrt{\dfrac{2}{r}} & \text{if } 2 < r < 2.5, \\ -\sqrt{\dfrac{3}{4} + \dfrac{1}{2r}} & \text{otherwise,} \end{cases} \tag{4.4.5}$$

see Figure 4.6. A reference solution computed with the first-order standard method has been computed again using a mesh of 10000 cells.

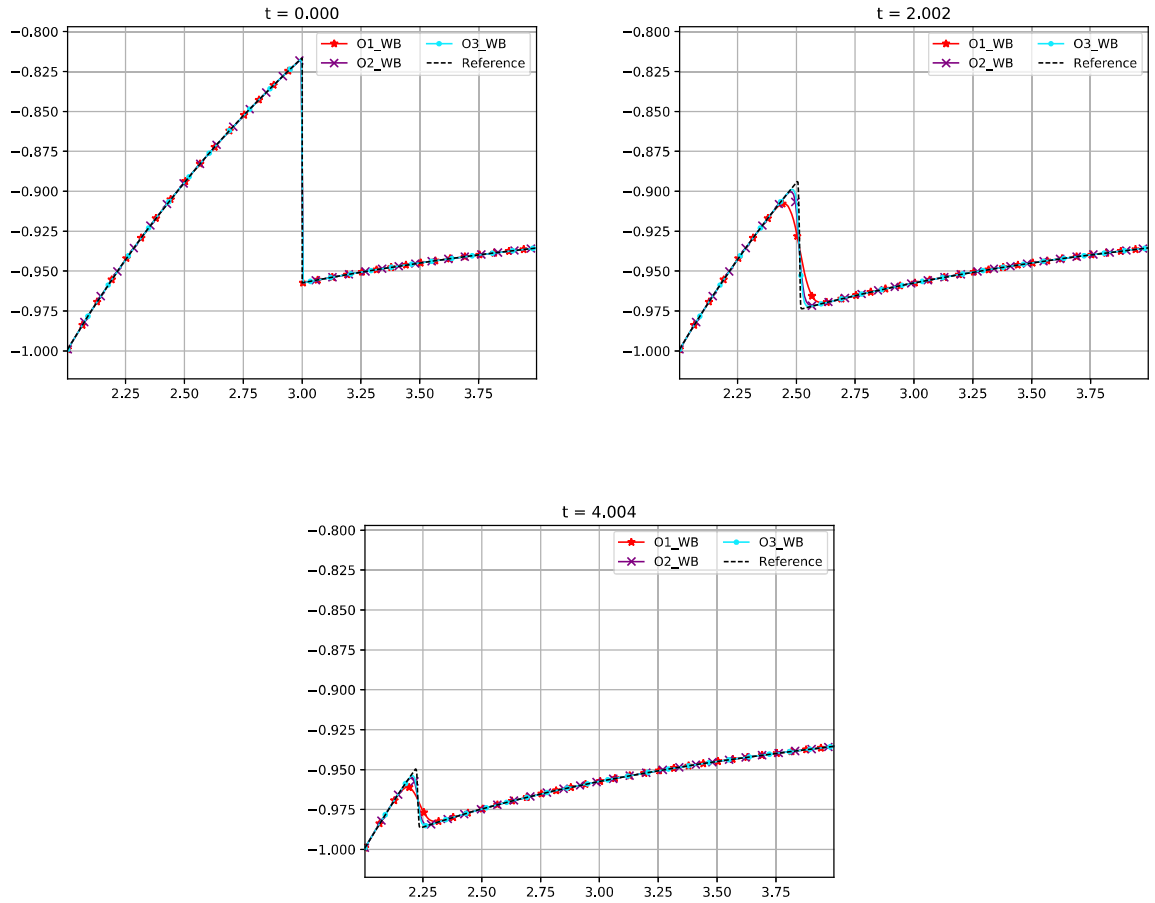



Figure 4.6: Burgers-Schwarzschild model with the initial condition (4.4.5): first-, second- and third-order well-balanced methods at selected times for variable $v$.

### 4.4.4 Perturbation of a steady shock solution

#### Left side perturbation

In this test case we consider the initial condition:

$$\tilde{v}_0(r) = v_0(r) + p_L(r), \tag{4.4.6}$$

where $v_0$ is the steady shock solution given by (4.4.3) and

$$p_L(r) = \begin{cases} -\dfrac{1}{5} e^{-200(r-2.5)^2} & \text{if } 2.2 < r < 2.8, \\ 0 & \text{otherwise.} \end{cases} \tag{4.4.7}$$

The first-, second- and third-order well-balanced method have been applied to this problem. In Figure 4.7 it can be observed that, after the wave generated by the initial perturbation

leaves the computational domain, the stationary solution (4.4.3) is not recovered: a different stationary solution of the family (4.3.3) is obtained whose shock is placed at a different location. Observe that all the three methods capture the same stationary solution.

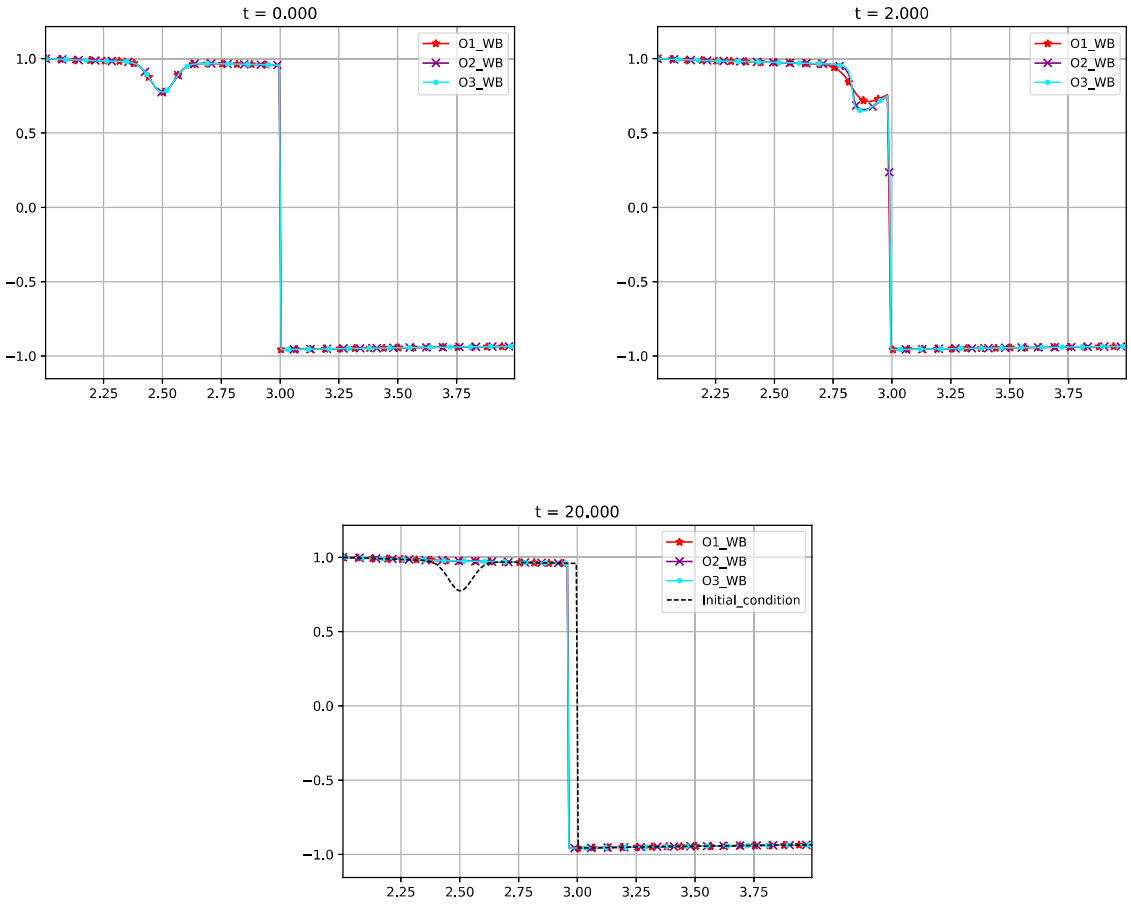



Figure 4.7: Burgers-Schwarzschild model with the initial condition (4.4.6)-(4.4.3)-(4.4.7): first-, second- and third-order well-balanced methods at selected times for variable $v$.

### Right side perturbation

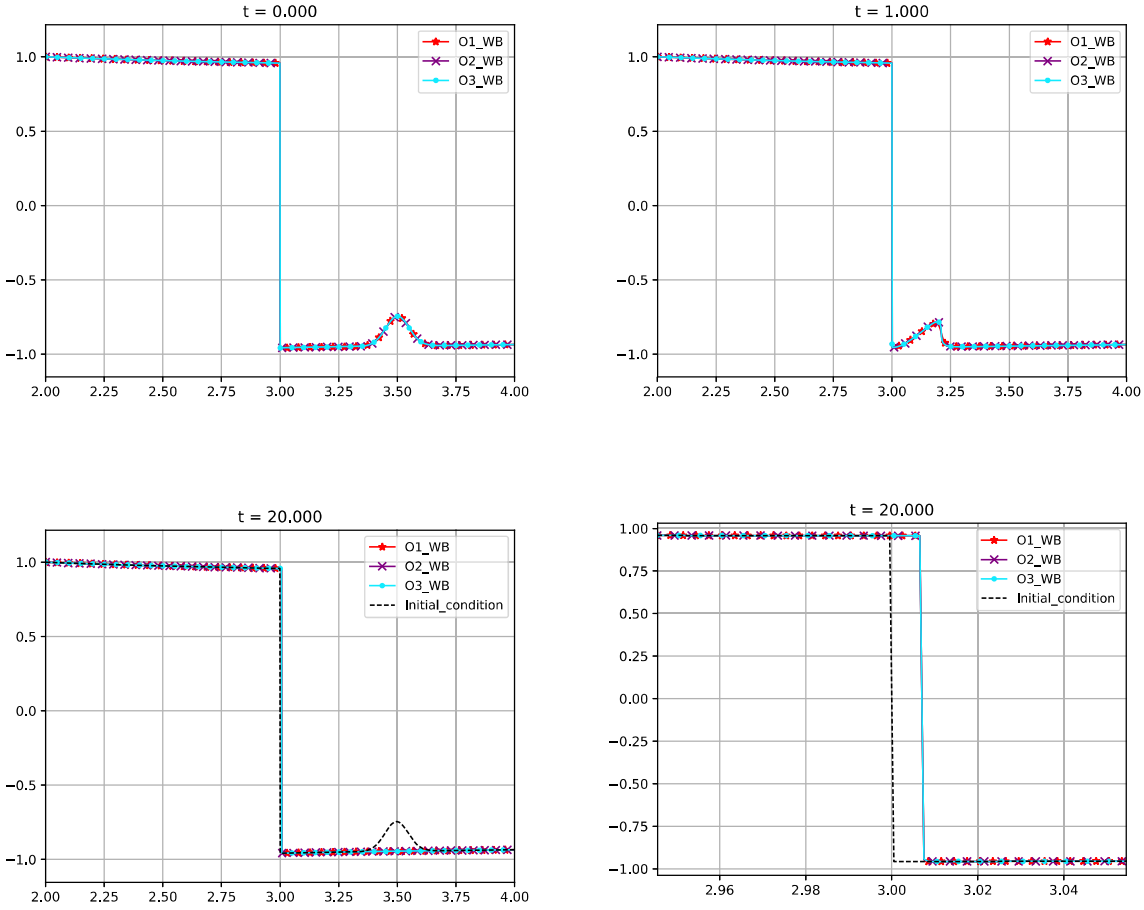
Similar conclusions can be drawn if a perturbation at the right side of the shock is superposed to the stationary solution $v_0$ given by (4.4.3):

$$\tilde{v}_0(r) = v_0(r) + p_R(r), \tag{4.4.8}$$

with

$$p_R(r) = \begin{cases} \frac{1}{5} e^{-200(r-3.5)^2} & \text{if } 3.2 < r < 3.8, \\ 0 & \text{otherwise,} \end{cases} \tag{4.4.9}$$

see Figure 4.8. In this case we have used a 2000-point uniform mesh because the displacement of the shock is smaller in this case and more points in the mesh are needed in order to see that the steady shock is not recovered.




Zoom of the solution

Figure 4.8: Burgers-Schwarzschild model with the initial condition (4.4.8)-(4.4.3)-(4.4.9): first-, second- and third-order well-balanced methods at selected times for variable $v$.

### Left side perturbation with zero average

Now we consider an initial condition of the form (4.4.6) with a perturbation $p_L$ such that:

$$p_L(r) = \begin{cases} 0.1\cos(-25.5\pi + 10\pi x)e^{-200(x-2.8)^2} & \text{if } 2.7 < r < 2.9, \\ 0 & \text{otherwise.} \end{cases} \tag{4.4.10}$$

In Figure 4.9 it can be observed that now, after the wave generated by the initial perturbation leaves the computational domain, the stationary solution (4.4.3) is recovered. Here we have used again a 2000-point uniform mesh to verify that the steady state does not move.

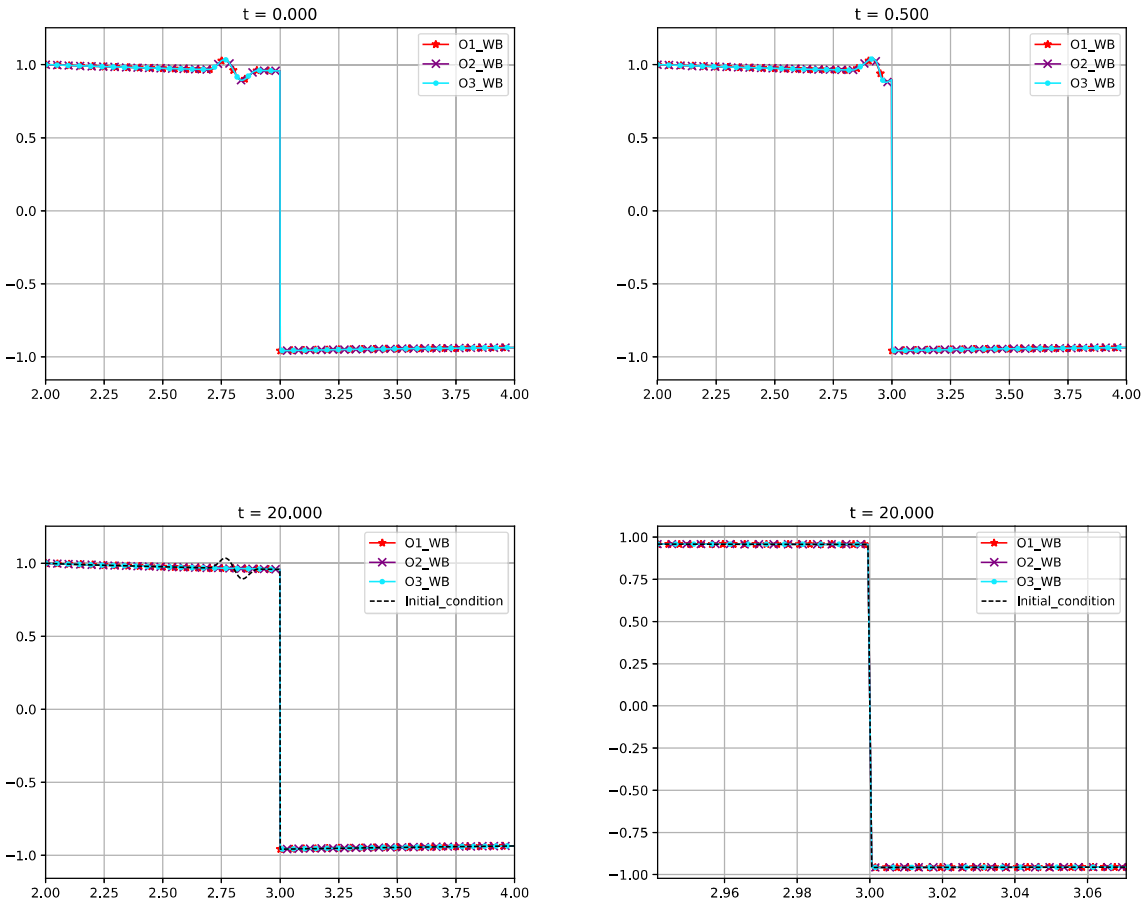


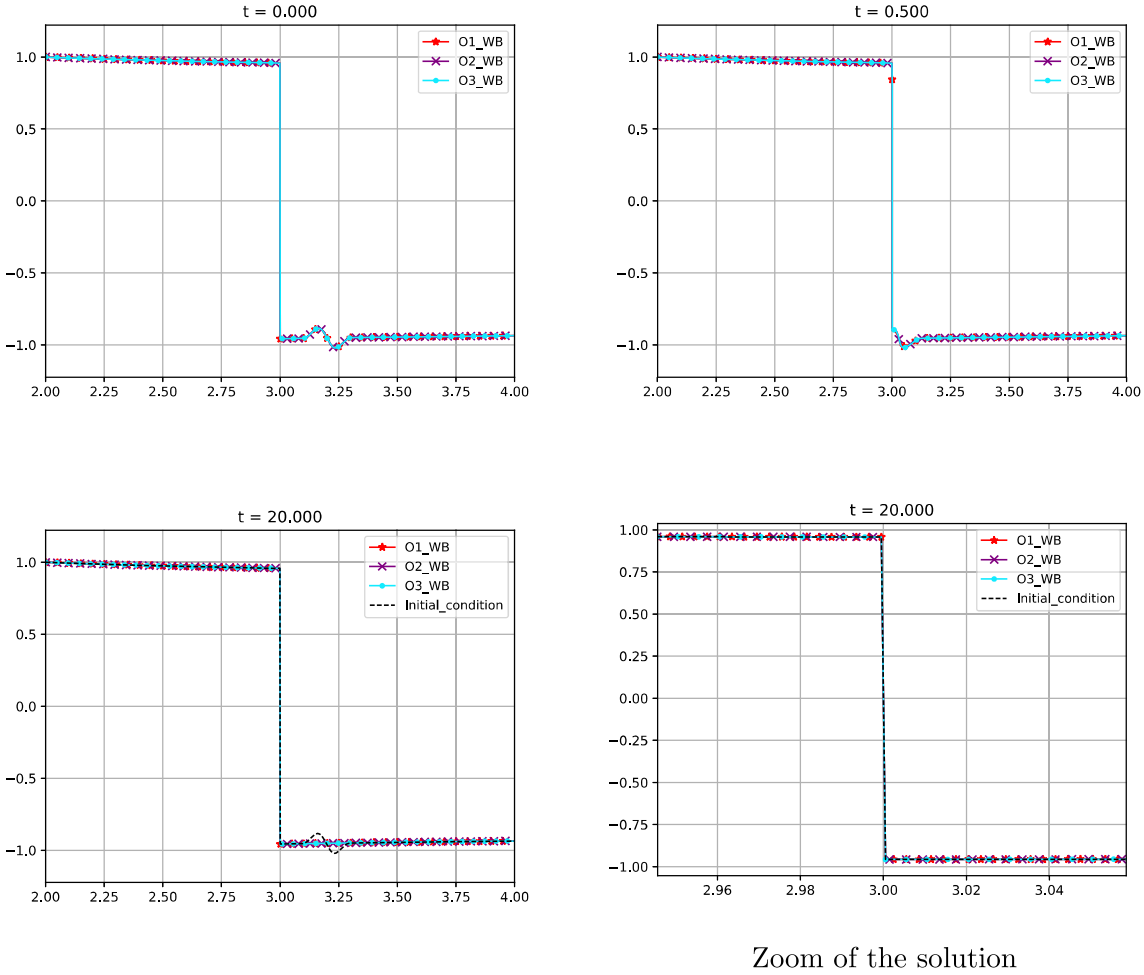

Figure 4.9: Burgers-Schwarzschild model with the initial condition (4.4.6)-(4.4.3)-(4.4.10): first-, second- and third-order well-balanced methods at selected times and zoom of the initial and final stationary shocks (right-down) for variable $v$.

### Right side perturbation with zero average

Similar conclusions can be drawn if we consider an initial condition of the form (4.4.8) with $\int p_R(r)dr = 0$. In particular we take

$$p_R(r) = \begin{cases} 0.1\cos(-29.5\pi + 10\pi x)e^{-200(x-3.2)^2} & \text{if } 3.1 < r < 3.3, \\ 0 & \text{otherwise,} \end{cases} \tag{4.4.11}$$

see Figure 4.10. Here we have used again a 2000-point uniform mesh to verify that the steady state does not move.




Zoom of the solution

Figure 4.10: Burgers-Schwarzschild model with the initial condition (4.4.8)-(4.4.3)-(4.4.11): first-, second- and third-order well-balanced methods at selected times, and zoom of the initial and final stationary shocks (right-down) for variable $v$.

**Left and right side perturbations with zero average**

In order to study the relation between the amplitude of the perturbation and the distance between the initial and the final stationary shocks, we consider the initial condition:

$$\tilde{v}_0(r) = v_0(r) + p_L(r) + p_R(r), \tag{4.4.12}$$

where $v_0$ is the steady shock solution given by (4.4.3) and $\int (p_L(r) + p_R(r))dr = 0$. In particular we take $p_L(r)$ as in (4.4.7) and $p_R(r)$ as in (4.4.9). In Figure 4.11 it can be observed that, after the wave generated by the initial perturbation leaves the computational domain, the stationary solution (4.4.3) is not recovered: a different stationary solution of the family (4.3.3) with the shock is placed at a different location. This is a natural result because as we saw before the right perturbation creates a lower displacement than the left perturbation. Here we have used again a 2000-point uniform mesh.

**Relation between the perturbation and the displacement of the shock**

In order to study the relationship between the amplitude of the perturbation and the distance between the initial and the final shock locations, we consider the family of initial conditions:

$$\tilde{v}_0(r) = v_0(r) + \delta_v(\alpha, r), \tag{4.4.13}$$

where $v_0$ is given again by (4.4.3) and

$$\delta_v(\alpha, r) = \begin{cases} \alpha \cos(5\pi r - 12\pi) e^{-200(r-2.8)^2} & \text{if } 2.7 < r < 2.9, \\ 0 & \text{otherwise,} \end{cases} \tag{4.4.14}$$

with $\alpha > 0$.

The amplitude of the perturbation is measured by $\int \delta_v(\alpha, r)\, dr$ and the distance between the shocks are measured by

$$\lim_{t\to\infty} \int |v(r,t) - v_0(r)|\, dr.$$

See Figure 4.12. Table 4.4 and Figure 4.13 show the relationship between those magnitudes: it is clearly linear.

### 4.4.5 Long-time behavior of the solutions

In this section we consider different initial conditions and investigate the long-time behavior of the corresponding solutions using the first-order well-balanced scheme. A large number of tests have been performed with the first-order methods (that is the less costly one) considering different initial conditions, different meshes, and different lengths of the computational domain: the observed behavior of the numerical solutions have been always one of the four ones shown here depending on the value at $2M$ (1 or lower) and at the right boundary (positive or negative).

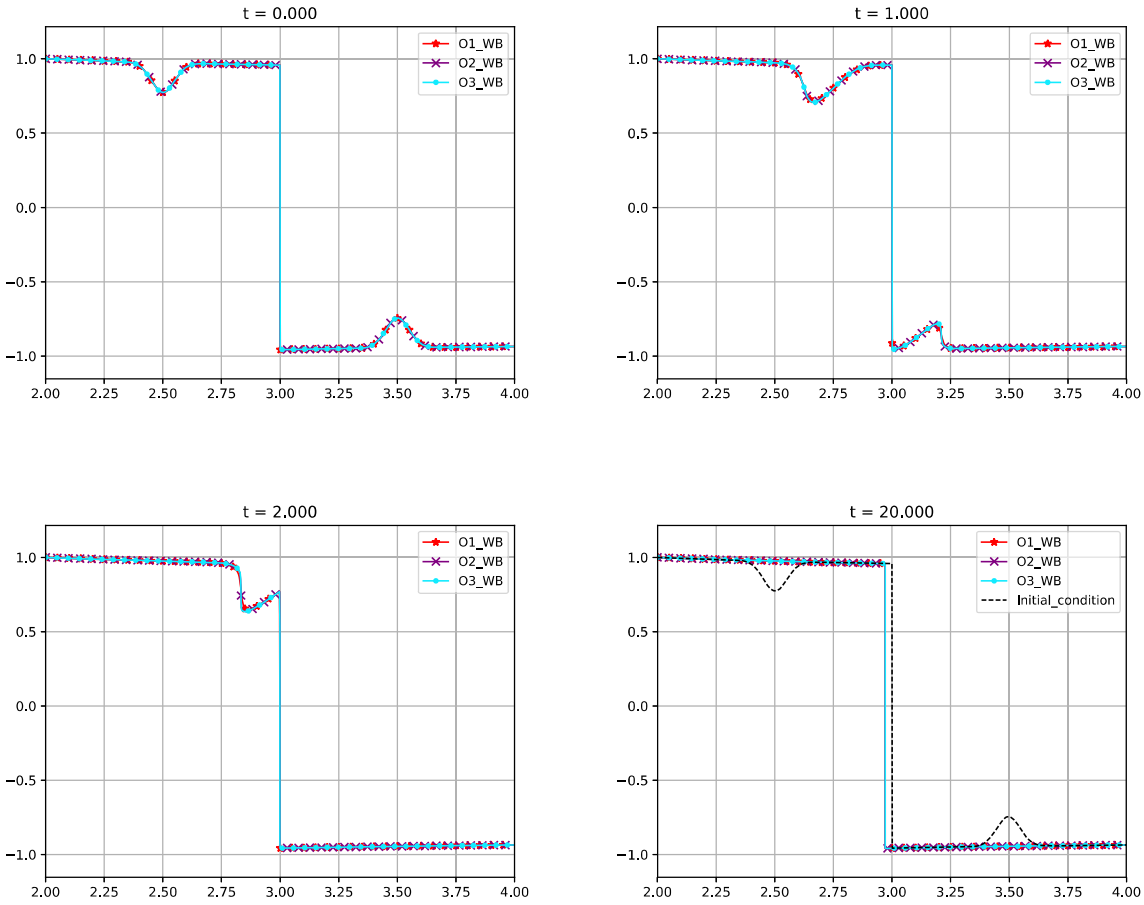
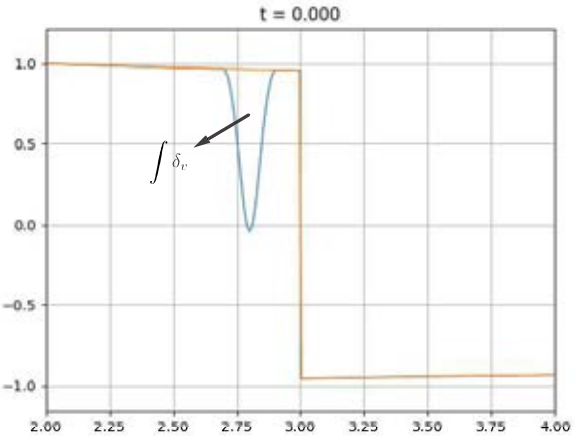



Figure 4.11: Burgers-Schwarzschild model with the initial condition (4.4.8)-(4.4.3)-(4.4.11): comparison between the first-, second- and third-order well-balanced methods at selected times for variable $v$.

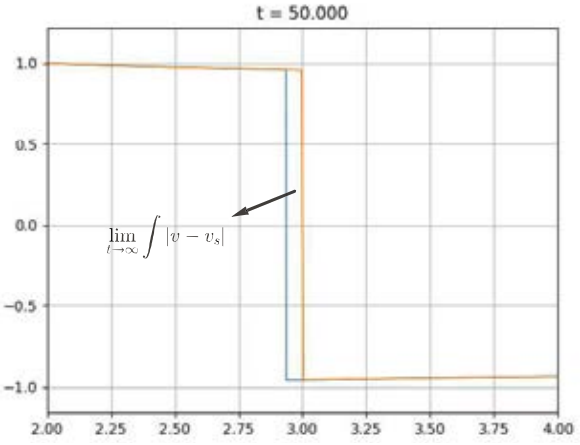
1. Initial condition satisfying $v_0(2M) = 1$ and $v_0(L) \geq 0$: let us consider the initial condition

$$v_0(r) = \begin{cases} 1 & \text{if } 2 < r < 2.1, \\ \cos(30r)e^{\frac{-1}{(x-2.5)^2}} & \text{otherwise,} \end{cases} \tag{4.4.15}$$

that takes value 1 in a neighborhood of $2M = 2$ and a positive value at the right boundary of the computational domain $x = 4$. As it can be observed in Figure 4.14 after a transient regime, the numerical solution takes the form of a right-moving shock linking the stationary solution $v \equiv 1$ with the negative stationary solution that takes value $-1$ at $x = 2M$ and value 0 at $x = 4$. Once this shock leaves the domain, the stationary solution $v \equiv 1$ is reached in the whole computational domain.




(a) Area of the perturbation for $\alpha = 1$




(b) Area between the initial and the final steady shock and the final steady shocks for $\alpha = 1$

Figure 4.12: Burgers-Schwarzschild model with the initial condition (4.4.13)-(4.4.3)-(4.4.14): measures of the perturbation and the shock displacement for $\alpha = 1$.

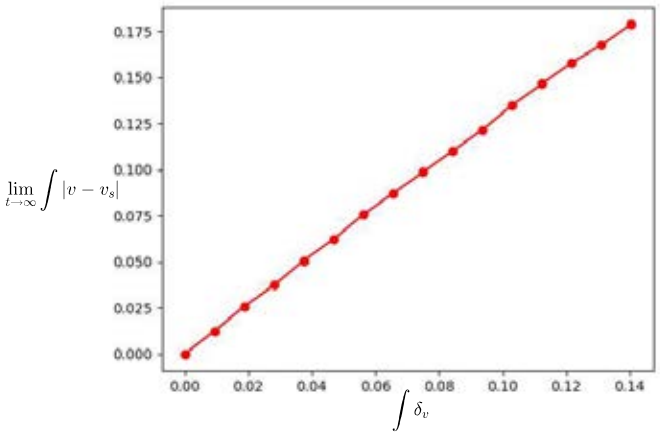



Figure 4.13: Burgers-Schwarzschild model with the initial condition (4.4.13)-(4.4.3)-(4.4.14): values of $\lim_{t\to\infty} \int |v - v_s|$ as a function of $\int \delta_v$.

2. Initial condition satisfying $v_0(2M) = 1$ and $v_0(L) < 0$: we consider now the

$$v_0(r) = \begin{cases} 1 & \text{if } 2 < r < 2.1, \\ \cos(20r)e^{\frac{-1}{(x-2.5)^2}} & \text{otherwise,} \end{cases} \tag{4.4.16}$$

that takes value 1 in a neighborhood of $2M = 2$ and negative value at the right boundary of the computational domain $x = 4$. As it can be observed in Figure 4.15

| $\alpha$ | $\int \delta_v$ | $\lim_{t\to\infty} \int \lvert v - v_s \rvert$ |
|---|---|---|
| 0.0 | 0.00000 | 0.00000 |
| 0.1 | 0.00936 | 0.01245 |
| 0.2 | 0.01873 | 0.02586 |
| 0.3 | 0.02809 | 0.03735 |
| 0.4 | 0.03745 | 0.05076 |
| 0.5 | 0.04682 | 0.06225 |
| 0.6 | 0.05618 | 0.07566 |
| 0.7 | 0.06554 | 0.08715 |
| 0.8 | 0.07491 | 0.09864 |
| 0.9 | 0.08427 | 0.11013 |
| 1.0 | 0.09364 | 0.12163 |
| 1.1 | 0.10300 | 0.13503 |
| 1.2 | 0.11236 | 0.14653 |
| 1.3 | 0.12172 | 0.15802 |
| 1.4 | 0.13109 | 0.16760 |
| 1.5 | 0.14045 | 0.17909 |

Table 4.4: Burgers-Schwarzschild model with the initial condition (4.4.13)-(4.4.3)-(4.4.14): measures of the perturbation and the shock displacement for different values of $\alpha$.

after a transient period, the numerical solution takes the form of a right-moving shock linking the stationary solution $v \equiv 1$ with the negative stationary solution that takes value $-1$ at $x = 2M$ and value $v_0(4)$ at $x = 4$. Once this shock leaves the domain, the stationary solution $v \equiv 1$ is reached in the whole computational domain.

3. Initial condition satisfying $v_0(2M) < 1$ and $v_0(L) \geq 0$: we consider now the initial condition

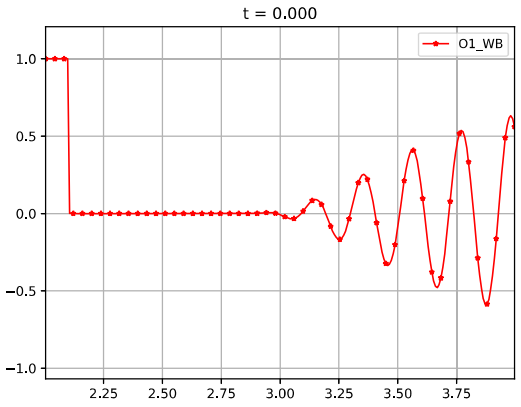
$$v_0(r) = \begin{cases} 0.8 & \text{if } 2 < r < 2.1, \\ \cos(30r)e^{\frac{-1}{(x-2.5)^2}} & \text{otherwise.} \end{cases} \tag{4.4.17}$$

In this case the numerical solution reaches in finite time the negative stationary solution $v^*$ such that $v^*(2) = -1$ and $v^*(4) = 0$: see Figure 4.16.

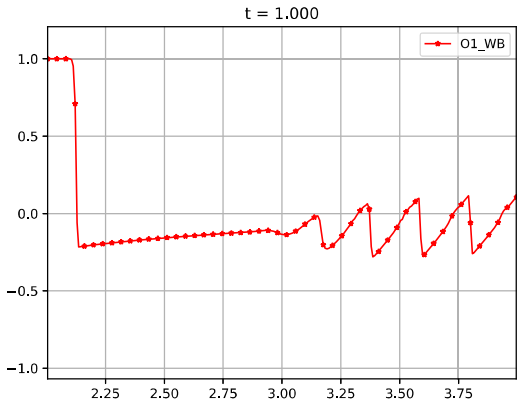
4. Initial condition satisfying $v_0(2M) < 1$ and $v_0(L) < 0$: we finally consider the initial condition

$$v_0(r) = \begin{cases} 0.8 & \text{if } 2 < r < 2.1, \\ \cos(20r)e^{\frac{-1}{(x-2.5)^2}} & \text{otherwise.} \end{cases} \tag{4.4.18}$$

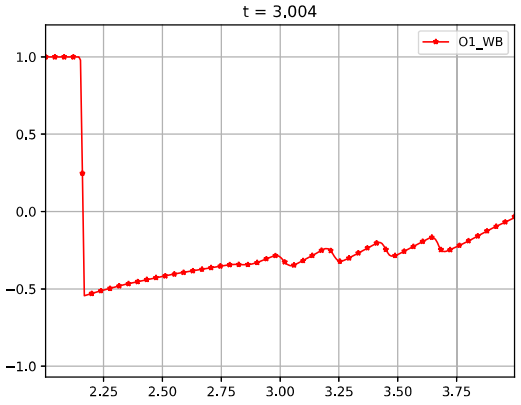
The numerical solution reaches in finite time the negative stationary solution $v^*$ such that $v^*(2) = -1$ and $v^*(4) = v_0(4)$: see Figure 4.17.



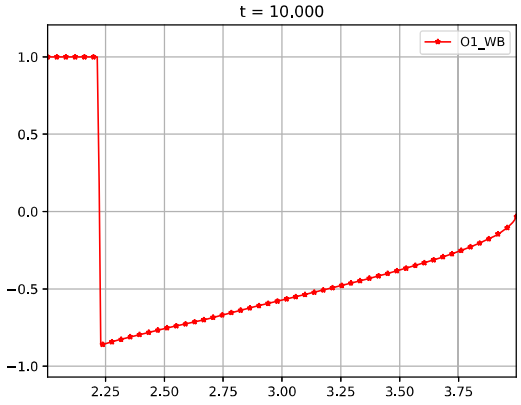

(a) $t = 0$



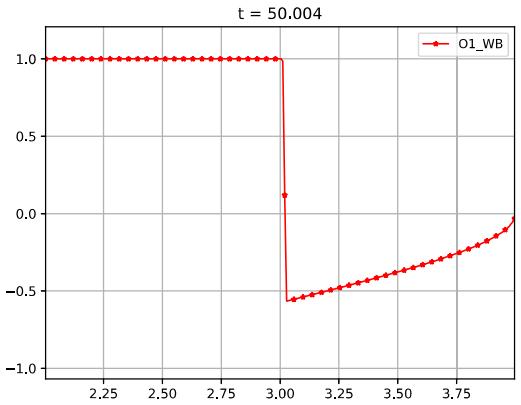

(b) $t = 1$



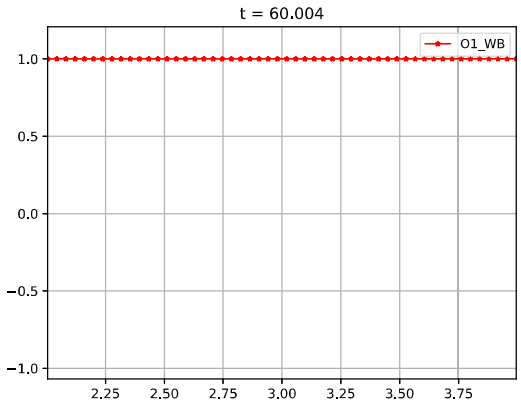

(c) $t = 3$




(d) $t = 10$

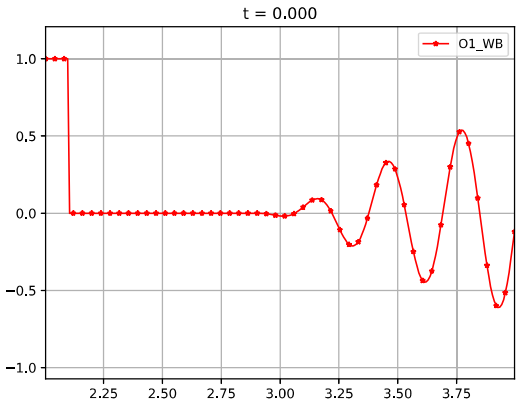



(e) $t = 50$

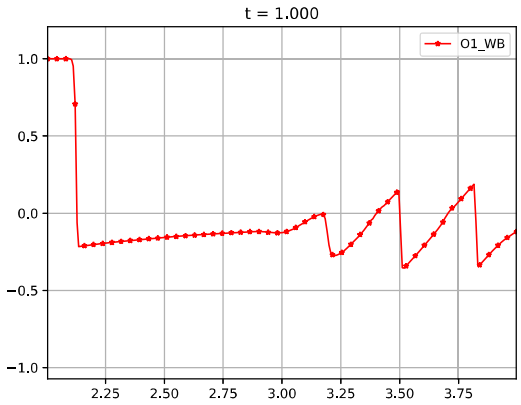



(f) $t = 60$

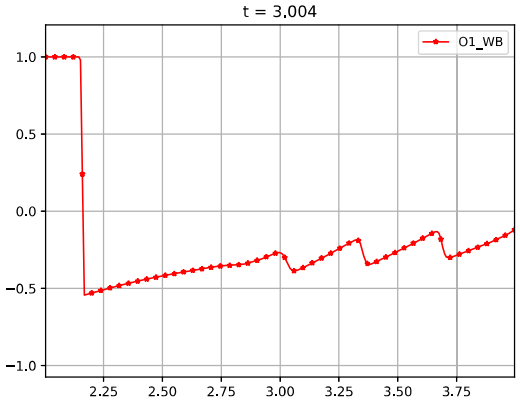
Figure 4.14: Burgers-Schwarzschild model with the initial condition (4.4.15): first-order well-balanced scheme at selected times for variable $v$.



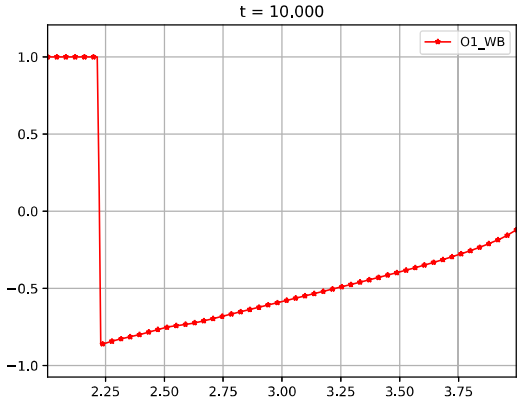

(a) $t = 0$



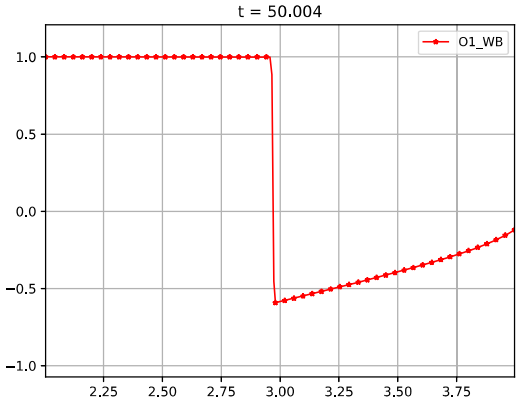

(b) $t = 1$



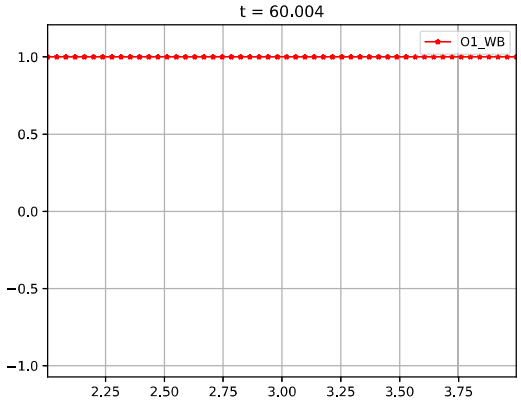

(c) $t = 3$




(d) $t = 10$

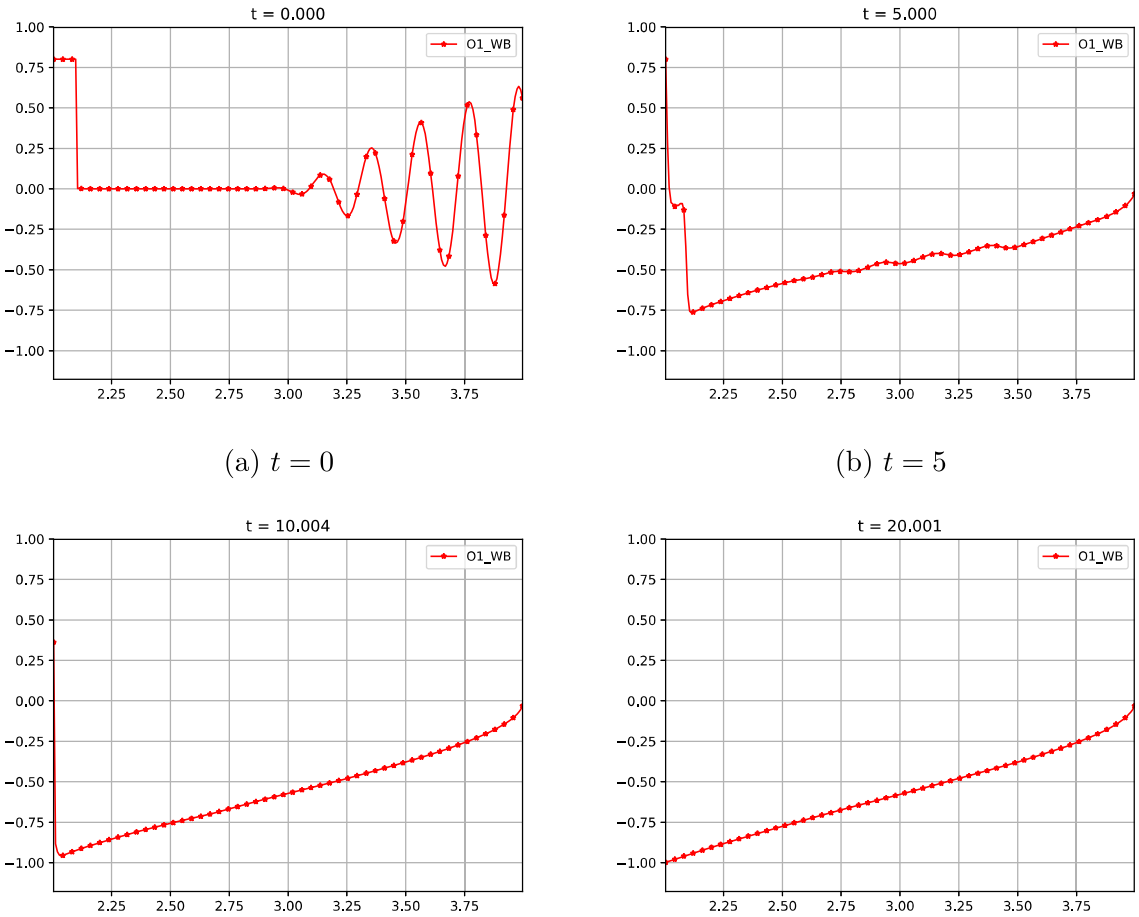



(e) $t = 50$




(f) $t = 60$

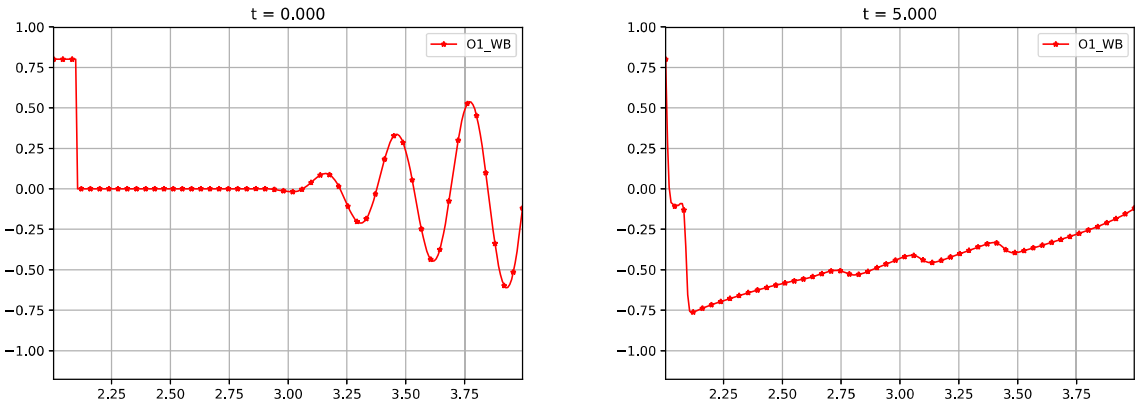
Figure 4.15: Burgers-Schwarzschild model with the initial condition (4.4.16): first-order well-balanced scheme at selected times for variable $v$.



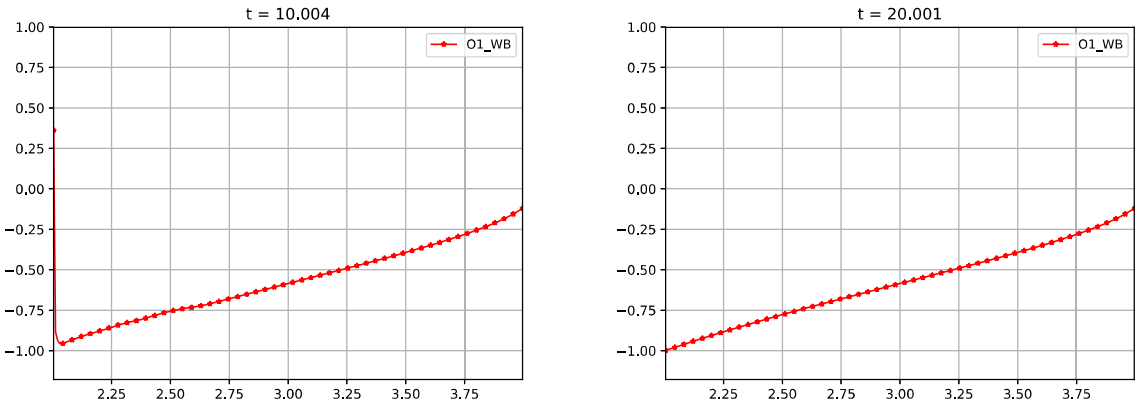

(a) $t = 0$ (b) $t = 5$

(c) $t = 10$ (d) $t = 20$

Figure 4.16: Burgers-Schwarzschild model with the initial condition (4.4.17): first-order well-balanced scheme at selected times for variable $v$.




(a) $t = 0$ (b) $t = 5$




(c) $t = 10$ (d) $t = 20$

Figure 4.17: Burgers-Schwarzschild model with the initial condition (4.4.18): numerical solution obtained with the first-order well-balanced scheme at selected times for variable $v$.

### 4.4.6 Main conclusions for the Burgers-Schwarzschild model

From Figures 4.7 to 4.13 and Table 4.4 we can conclude the following:

**Conclusion 1.** *If a perturbation $\delta_v$ whose support consists on only one interval is added to a steady shock solution of the form*

$$v_0(r) = \begin{cases} \sqrt{1 - K_0^2(1 - \frac{2M}{r})} & \text{if } 2M < r < r_0, \\ -\sqrt{1 - K_0^2(1 - \frac{2M}{r})} & \text{otherwise,} \end{cases}$$

*then the solution reaches at finite time another steady shock solution of the form:*

$$v(r) = \begin{cases} \sqrt{1 - K_0^2(1 - \frac{2M}{r})} & \text{if } 2M < r < r_1, \\ -\sqrt{1 - K_0^2(1 - \frac{2M}{r})} & \text{othewise,} \end{cases}$$

*where*

1. *If $\int_{2M}^{r_0} \delta_v = 0$ and $\int_{r_0}^{\infty} \delta_v = 0$, then $r_1 = r_0$, i.e. the initial stationary solution is recovered.*
2. *If $\int_{2M}^{\infty} \delta_v = 0$ and $\int_{2M}^{r_0} \delta_v = -\int_{r_0}^{\infty} \delta_v$, then $r_1 \neq r_0$ and a different stationary solution is obtained.*
3. *If $\int \delta_v \neq 0$, then $r_1 \neq r_0$ and a different stationary solution is obtained. In this case the distance between $r_0$ and $r_1$ depends linearly on the amplitude of the perturbation: see Table 4.4 and Figure 4.13.*

In view of Figures 4.14 to 4.17 we have reached the following.

**Conclusion 2.** *1. For a bounded domain $[2M, L]$:*

*(a) If $v_0(r) = 1$ for $r \in [2M, 2M + \epsilon)$, with $\epsilon > 0$, $v_0(L) \geq 0$ and $v_0 \neq 1$, in finite time the solution has the form of a right-moving shock that links the stationary solution $v \equiv 1$ and the negative steady solution $v^*$ such that $v^*(2M) = -1$ and $v^*(L) = 0$, that is,*

$$v_0^*(r) = -\sqrt{1 - \frac{1}{1 - \frac{2M}{L}}\left(1 - \frac{2M}{r}\right)}.$$

*(b) If $v_0(r) = 1$ for $r \in [2M, 2M + \epsilon)$, with $\epsilon > 0$ and $v_0(L) = a$, with $a < 0$, then in finite time the solution has the form of a right-moving shock that links the stationary solution $v \equiv 1$ and the negative steady solution $v*$ such that $v^*(2M) = -1$ and $v_0^*(L) = a$, that is:*

$$v_0^*(r) = -\sqrt{1 - \frac{1 - a^2}{1 - \frac{2M}{L}}\left(1 - \frac{2M}{r}\right)}.$$

(c) *If* $v_0(2M) < 1$ *and* $v_0(L) \geq 0$*, then in finite time the solution coincides with the negative steady solution such that* $v^*(2M) = -1$ *and* $v_0^*(L) = 0$*, that is:*

$$v_0^*(r) = -\sqrt{1 - \frac{1}{1 - \frac{2M}{L}}\left(1 - \frac{2M}{r}\right)}.$$

(d) *If* $v_0(2M) < 1$ *and* $v_0(L) = a$*, with* $a < 0$*, then in finite time the solution coincides with the negative stationary solution* $v^*$ *such that* $v^*(L) = a$*, that is:*

$$v_0^*(r) = -\sqrt{1 - \frac{1 - a^2}{1 - \frac{2M}{L}}\left(1 - \frac{2M}{r}\right)}.$$

2. *For the unbounded domain* $[2M, \infty)$ *the following conclusions can be drawn by passing to the limit when* $L \to \infty$*:*

(a) *If* $v_0(r) = 1$ *for* $r \in [2M, 2M + \epsilon)$*, with* $\epsilon > 0$*,* $\lim_{r\to\infty} v_0(r) \geq 0$ *and* $v_0 \neq 1$*, in finite time the solution has the form of a right-moving shock that links the stationary solution* $v \equiv 1$ *and the negative stationary solution*

$$v_0^*(r) = -\sqrt{\frac{2M}{r}},$$

*corresponding to* $K^2 = 1$*.*

(b) *If* $v_0(r) = 1$ *for* $r \in [2M, 2M + \epsilon)$*, with* $\epsilon > 0$ *and* $\lim_{r\to\infty} v_0(r) = a$*, with* $a < 0$*, then in finite time* $t_0$ *the solution has the form of a right- moving shock that links the stationary solution* $v \equiv 1$ *and the negative stationary solution* $v^*$ *such that* $v^*(2M) = -1$ *and* $\lim_{r\to\infty} v_0^*(r) = a$*, that is:*

$$v_0^*(r) = -\sqrt{1 - (1 - a^2)\left(1 - \frac{2M}{r}\right)}.$$

(c) *If* $v_0(2M) < 1$ *and* $\lim_{r\to\infty} v_0(r) \geq 0$*, then the solution converges as* $t \to \infty$ *to the negative stationary solution* $v^*$ *such that* $v^*(2M) = -1$ *and* $\lim_{r\to\infty} v_0^*(r) = 0$*, that is:*

$$v_0^*(r) = -\sqrt{\frac{2M}{r}}.$$

(d) *If* $v_0(2M) < 1$ *and* $\lim_{r\to\infty} v_0(r) = a$*, with* $a < 0$*, then the solution converges as* $t \to \infty$ *to the negative stationary solution* $v^*$ *such that* $v^*(2M) = -1$ *and* $\lim_{r\to\infty} v_0^*(r) = a$*, that is:*

$$v_0^*(r) = -\sqrt{1 - (1 - a^2)\left(1 - \frac{2M}{r}\right)}.$$

# 4.5 Euler-Schwarzschild model: designing the numerical algorithm

## 4.5.1 Preliminaries

In the case of the Euler-Schwarzschild equations (4.1.4), the stationary solutions are implicitly given by the equations:

$$\begin{cases} \dfrac{\mathrm{sgn}(v)(1-v^2)|v|^{\frac{2k^2}{1-k^2}} r^{\frac{4k^2}{1-k^2}}}{\left(1-\frac{2M}{r}\right)} = C_1 \equiv constant, \\ \\ r(r-2M)\rho \dfrac{v}{1-v^2} = C_2 \equiv constant. \end{cases} \tag{4.5.1}$$

The pair $(v,\rho)$ of a stationary solution satisfies the ODE system:

$$\frac{dv}{dr} = v\frac{(1-v^2)(1-k^2)}{r(r-2M)}\left(\frac{2k^2}{1-k^2}(r-2M)-M\right)\Big/(v^2-k^2), \tag{4.5.2}$$

$$\frac{d\rho}{dr} = -\frac{2(r-M)}{r(r-2M)}\rho - \rho\frac{(1+v^2)(1-k^2)}{r(r-2M)}\left(\frac{2k^2}{1-k^2}(r-2M)-M\right)\Big/(v^2-k^2), \tag{4.5.3}$$

see [114]. Figure 4.18 shows the graph of $v$ for some of them. When these functions are defined in $(2M,\infty)$, they have a maximum or a minimum in

$$r_c = \frac{M(1-k^2)}{2k^2} + 2M, \tag{4.5.4}$$

that comes from solving $\frac{dv}{dr}=0$. In Figure 4.18 the red stationary solutions are those such that at $r=r_c$ the take the value $v=\pm k$.

Given two constants $C_1$ and $C_2$, in order to compute the stationary solution given by (4.5.1) in a point $r=a$, the following non linear system has to be solved:

$$\mathrm{sgn}(v)(1-v^2)|v|^{\frac{2k^2}{1-k^2}} = \frac{\left(1-\frac{2M}{a}\right)}{a^{\frac{4k^2}{1-k^2}}}C_1, \qquad \rho = \frac{1-v^2}{va(a-2M)}C_2. \tag{4.5.5}$$

It is enough thus to solve, if it is possible, the nonlinear equation

$$g(v) = K_a, \tag{4.5.6}$$

with

$$g(v) = \mathrm{sgn}(v)(1-v^2)|v|^{\frac{2k^2}{1-k^2}}, \quad v\in[-1,1], \tag{4.5.7}$$

$$K_a = \frac{\left(1-\frac{2M}{a}\right)}{a^{\frac{4k^2}{1-k^2}}}C_1, \tag{4.5.8}$$

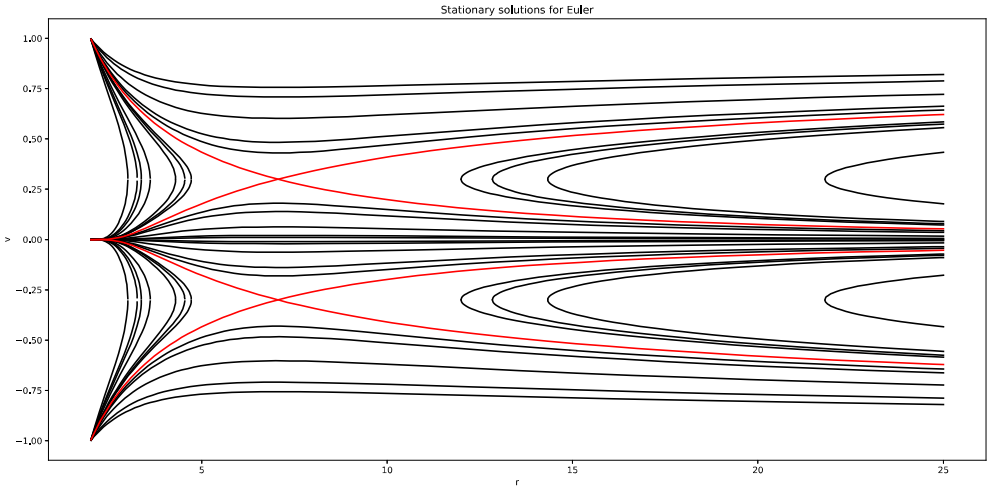



Figure 4.18: Euler-Schwarzschild model with $k = 0.3$: $v$ variable for some stationary solutions

to compute $v$. Once this equation is solved, $\rho$ is computed using the second equation of (4.5.5).

It can be easily checked that $g$ satisfies:

$$-(1-k^2)k^{\frac{2k^2}{1-k^2}} = g(-k) \leq g(v) \leq g(k) = (1-k^2)k^{\frac{2k^2}{1-k^2}}, \qquad v \in [-1,1].$$

Moreover, $g$ is strictly monotone in $[-1,-k)$, $(-k,k)$, and $(k,1]$. As a consequence we can conclude:

- if $|K_a| > g(k)$ equation (4.5.6) has no solution, i.e. a stationary solution given by $C_1$ and $C_2$ cannot be defined at $r = a$;
- if $|K_a| = g(k)$ then equation (4.5.6) has only one solution ($k$ if $K_a > 0$, $-k$ if $K_a < 0$). Therefore, (4.5.5) has only one solution that is a sonic state;
- otherwise, (4.5.6) has two possible solutions. Therefore there are two states $(\rho, v)$ that solve (4.5.5), one supersonic and one subsonic.

For the sake of clarity, together with the representation

$$V = [\rho(1+k^2v^2)/(1-v^2), \rho v(1+k^2)/(1-v^2)]^T,$$

for the states, we will use

$$\widetilde{V} = [\rho, v]^T.$$

$V$ can be easily computed from $\widetilde{V}$ and, given $V$, $\widetilde{V}$ is also easily computed by (4.1.8) that comes from solving a second-degree equation.

### 4.5.2 First-order method

If the midpoint rule is used to compute the initial averages, given a family of cell values $\widetilde{V}_i$, in the first step of the well-balanced reconstruction procedure one has to find, if it is possible, a stationary solution $\widetilde{V}_i^*$ defined in $[r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}]$ such that

$$\widetilde{V}_i^*(r_i) = \widetilde{V}_i = [\rho_i, v_i]^T.$$

Obviously such a stationary solution would correspond to the choice of constants:

$$C_{i,1} = \frac{\operatorname{sgn}(v_i)(1-v_i^2)|v_i|^{\frac{2k^2}{1-k^2}} r_i^{\frac{4k^2}{1-k^2}}}{\left(1-\frac{2M}{r_i}\right)}, \tag{4.5.9}$$

$$C_{i,2} = r_i(r_i-2M)\rho_i \frac{v_i}{1-v_i^2}. \tag{4.5.10}$$

According to the discussion above, the corresponding stationary solution is defined in $r_{i\pm\frac{1}{2}}$ provided that:

$$|K_{i\pm\frac{1}{2}}| \leq g(k), \tag{4.5.11}$$

where

$$K_{i\pm\frac{1}{2}} = \left(1-\frac{2M}{r_{i\pm\frac{1}{2}}}\right) r_{i\pm\frac{1}{2}}^{-\frac{4k^2}{1-k^2}} C_{i,1}. \tag{4.5.12}$$

When $|K_{i\pm\frac{1}{2}}| < g(k)$ there are two possible values for $\widetilde{V}_i^*(r_{i\pm\frac{1}{2}})$, one subsonic and one supersonic. Therefore, a criterion is needed to select one or the other. The following criterion will be used here:

- if $\widetilde{V}_i$ is not sonic, then the state whose regime (sub or supersonic) is the same as $\widetilde{V}_i$ is selected for $\widetilde{V}_i^*(r_{i\pm\frac{1}{2}})$.
- if $\widetilde{V}_i$ is sonic, then the state whose regime is the same as $\widetilde{V}_{i+1}$ is selected for $\widetilde{V}_i^*(r_{i+\frac{1}{2}})$ and the state whose regime is the same as $\widetilde{V}_{i-1}$ is selected for $\widetilde{V}_i^*(r_{i-\frac{1}{2}})$.

Observe that this criterion aims to preserve the regime of the given cell values.

If condition (4.5.11) is satisfied, then the numerical method (4.2.16) is used. Otherwise the standard trivial reconstruction is considered.

The expression of the semi-discrete first-order method is then as follows:

$$\frac{dV_i}{dt} = -\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} - S_i\right), \tag{4.5.13}$$

where

$$S_i = \begin{cases} F(V_i^*(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}) - F(V_i^*(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}})), & \text{if (4.5.11) is satisfied;} \\ S(V_i, r_i), & \text{otherwise.} \end{cases} \tag{4.5.14}$$

The forward Euler method (1.2.51) is used again for the time discretization.

### 4.5.3 Second-order method

Let us use again the midpoint rule to compute cell averages and the minmod reconstruction operator. The stationary solution sought at the first stage of the well-balanced reconstruction procedure is again characterized by the constants (4.5.9)-(4.5.10). This time, this stationary solution has to be computed at the points $r_{i-1}$, $r_{i-\frac{1}{2}}$, $r_{i+\frac{1}{2}}$, $r_{i+1}$ so that the following condition has to be satisfied:

$$|K_{i+j}| \leq g(k), \quad j = -1, -\frac{1}{2}, \frac{1}{2}, 1, \tag{4.5.15}$$

where $K_{i\pm\frac{1}{2}}$ are given by (4.5.12) and

$$K_{i\pm1} = \left(1 - \frac{2M}{r_{i\pm1}}\right) r_{i\pm1}^{-\frac{4k^2}{1-k^2}} C_{i,1}. \tag{4.5.16}$$

If this condition is satisfied, the reconstruction is defined as follows:

$$\mathbb{P}_i(r) = V_i^*(r) + \text{minmod}\left(\frac{W_{i+1} - W_i}{\Delta r}, \frac{W_{i+1} - W_{i-1}}{2\Delta r}, \frac{W_i - W_{i-1}}{\Delta r}\right)(r - r_i),$$

where the minmod function is applied component by component and

$$W_j = V_j - V_i^*(r_j), \quad j = i-1, i, i+1.$$

Observe that the *conserved* variables $V$ are used in the reconstruction procedure.

If (4.5.15) is not satisfied, then the standard MUSCL reconstruction is applied:

$$\mathbb{Q}_i(r) = V_i + \text{minmod}\left(\frac{V_{i+1} - V_i}{\Delta r}, \frac{V_{i+1} - V_{i-1}}{2\Delta r}, \frac{V_i - V_{i-1}}{\Delta r}\right)(r - r_i).$$

The expression of the numerical method is given again by (4.5.13)-(4.5.14) with the difference that the second-order reconstructions are used now to compute the numerical fluxes. The TVDRK2 (1.2.52) method is used now to discretize the equations in time.

### 4.5.4 Third-order method

Although it will not be implemented in this chapter, let us briefly describe the first step of a third-order well-balanced reconstruction procedure based on the two-point Gauss quadrature in order to compute averages: it consists on finding $C_1$ and $C_2$ such that

$$\frac{1}{2}V^*(r_{i,0}; C_1, C_2) + \frac{1}{2}V^*(r_{i,1}; C_1, C_2) = V_i, \tag{4.5.17}$$

where $V^*(r; C_1, C_2)$ represents the value at $r$ of a stationary solution characterized by the constants $C_1$ and $C_2$.

# 4.6 Euler-Schwarzschild model: a numerical study

## 4.6.1 Preliminaries

In this section several tests are considered to check the performance of the first- and second-order well-balanced numerical methods for Euler-Schwarzschild model introduced in the previous section. We consider the spatial interval $[2M, L]$ with $M = 1$ and $L = 10$, a 500-point uniform mesh, $k = 0.3$ and the CFL number is set to 0.5 again. At $x = 2M$ we impose $F_{-\frac{1}{2}} = 0$ as boundary condition because $\left(1 - \frac{2M}{r}\right) = 0$. At $x = L$ we will use a transmissive boundary condition in the case we are not in a stationary solution or we will expand the stationary solution if we are in one.

In order to test the dependency of the results on the numerical method, two different first-order numerical fluxes are considered: the Lax-Friedrichs numerical flux

$$F_{i+\frac{1}{2}} = \frac{1}{2}(F(V_i) + F(V_{i+1})) - \frac{1}{2}\frac{\Delta t}{\Delta x}(V_{i+1} - V_i), \tag{4.6.1}$$

and a HLL-like numerical flux in PVM form (1.2.25):

$$F_{i+\frac{1}{2}} = \frac{1}{2}(F(V_i) + F(V_{i+1})) - \frac{1}{2}\left(\alpha_0(V_{i+1} - V_i) + \alpha_1(F(V_{i+1}) - F(V_i))\right), \tag{4.6.2}$$

with

$$\alpha_0 = \frac{\overline{\lambda_2}|\overline{\lambda_1}| - \overline{\lambda_1}|\overline{\lambda_2}|}{\overline{\lambda_2} - \overline{\lambda_1}}, \quad \alpha_1 = \frac{|\overline{\lambda_2}| - |\overline{\lambda_1}|}{\overline{\lambda_2} - \overline{\lambda_1}}, \tag{4.6.3}$$

where $\overline{\lambda_1}$ and $\overline{\lambda_2}$ are the eigenvalues of some intermediate matrix $J_{i+\frac{1}{2}}$ of the form

$$J_{i+\frac{1}{2}} = \left(1 - \frac{2M}{r_{i+\frac{1}{2}}}\right)\begin{bmatrix} 0 & 1 \\ \dfrac{k^2 - v_m^2}{1 - k^2 v_m^2} & \dfrac{2(1-k^2)v_m}{1 - k^2 v_m^2} \end{bmatrix}, \tag{4.6.4}$$

where $v_m$ is some intermediate value between $v_i^n$ and $v_{i+1}^n$.

Given two states $V_L$ and $V_R$, in order to choose an adequate intermediate value $v_m$, we look for $v$ such that the following Roe-type property (1.2.16) is satisfied:

$$\begin{bmatrix} 0 & 1 \\ \dfrac{k^2 - v^2}{1 - k^2 v^2} & \dfrac{2(1-k^2)v}{1 - k^2 v^2} \end{bmatrix} \cdot (V_R - V_L) = \widehat{F}_R - \widehat{F}_L, \tag{4.6.5}$$

where

$$\widehat{F}_\alpha = \begin{pmatrix} \dfrac{1 + k^2}{1 - v_\alpha^2}\rho_\alpha v_\alpha \\ \dfrac{v_\alpha^2 + k^2}{1 - v_\alpha^2}\rho_\alpha \end{pmatrix}, \quad \alpha = L, R,$$

i.e. the factor $(1-2M/r)$ is neglected for simplicity. Due to the form of the matrix, it is enough to find $v$ such that

$$\frac{k^2-v^2}{1-k^2v^2}(V_{1,R}-V_{1,L})+\frac{2(1-k^2)v}{1-k^2v^2}(V_{2,R}-V_{2,L})=F_{2,R}-F_{2,L}.$$

This equality is equivalent to the second-order degree equation for $v$:

$$\alpha v^2+\beta v+\gamma=0,$$

where

$$\begin{aligned}\alpha&=\rho_R(1-v_L^2)-\rho_L(1-v_R^2),\\ \beta&=-2\left(\rho_R v_R(1-v_L^2)-\rho_L v_L(1-v_R^2)\right),\\ \gamma&=\rho_R v_R^2(1-v_L^2)-\rho_L v_L^2(1-v_R^2).\end{aligned}$$

Since the discriminant

$$D=\rho_L\rho_R(1-v_L^2)(1-v_R^2)(v_R-v_L)^2$$

is always positive, there are always two real solutions:

$$v_\pm=\frac{\rho_R v_R(1-v_L^2)-\rho_L v_L(1-v_R^2)\pm|v_R-v_L|\sqrt{\rho_L\rho_R(1-v_L^2)(1-v_R^2)}}{\rho_R(1-v_L^2)-\rho_L(1-v_R^2)}$$

and it can be proven that:

- if $v_L<v_R$, then $v_-\in(v_L,v_R)$ and $v_+\notin(v_L,v_R)$, so that we will take $v_m=v_-$;
- if $v_L>v_R$ then $v_-\notin(v_R,v_L)$ and $v_+\notin(v_R,v_L)$, so that we will take $v_m=v_+$.

Finally, in the case $\alpha=0$ and $V_R\neq V_L$, we take $v_m=-\frac{\gamma}{\beta}$ and in the case $||V_R-V_L||_\infty<\epsilon$ we take $v_m=\frac{v_L+v_R}{2}$.

Once $v_m$ has been chosen, the expression of $\overline{\lambda_j}$, $j=1,2$ in (4.6.3) is as follows:

$$\begin{aligned}\overline{\lambda_1}=\lambda_1(v_m)&=\left(1-\tfrac{2M}{r_{i+\frac12}}\right)\tfrac{v_m-k}{1-k^2v_m},\\ \overline{\lambda_2}=\lambda_2(v_m)&=\left(1-\tfrac{2M}{r_{i+\frac12}}\right)\tfrac{v_m+k}{1+k^2v_m}.\end{aligned}$$

Since for a 2-systems HLL and Roe methods are equivalent and the intermediate value chosen to compute the wave speeds satisfies a Roe-type property, this numerical flux will be called Roe-type numerical flux in what follows.

The numerical methods will be compared with those based on the same numerical flux and the standard first- and second-order reconstructions.

### 4.6.2 Stationary solutions

#### Positive stationary solution

We take as initial condition the positive supersonic stationary solution satisfying

$$\rho^*(10) = 1, \quad v^*(10) = 0.6. \tag{4.6.6}$$

Table 4.5 shows the error in $L^1$ norm between the numerical solution at time $t = 50$ for the well-balanced and non-well-balanced methods using the Roe-type numerical flux. Figures 4.19 and 4.20 compare the numerical solutions obtained with the well-balanced and the non-well-balanced methods: as it happened for the Burgers-Schwarzschild model, the numerical solutions obtained with non-well-balanced methods depart from the initial steady state.

| Scheme (500 cells) | Error $v$ (1st) | Error $\rho$ (1st) | Error $v$ (2nd) | Error $\rho$ (2nd) |
|---|---|---|---|---|
| Well-balanced | 3.34E-13 | 5.61e-12 | 3.43e-13 | 7.12e-12 |
| Non well-balanced | 0.94 | 5.79 | 0.93 | 5.75 |

Table 4.5: Well-balanced versus non-well-balanced schemes: $L^1$ errors at time $t = 50$ for the Euler-Schwarzschild model with the initial condition (4.6.6).

#### Negative stationary solution

Let us consider now as initial condition the negative supersonic stationary solution $V^*$ that satisfies

$$\rho^*(10) = 1, \quad v^*(10) = -0.8. \tag{4.6.7}$$

Table 4.6 shows the error in $L^1$ norm between the numerical solution at time $t = 50$. Figures 4.21, 4.22 show the difference between the numerical results given by well-balanced and non-well-balanced methods using the Roe-type numerical flux. Again the numerical solutions obtained with non-well-balanced methods depart from the initial steady state.

| Scheme (500 cells) | Error $v$ (1st) | Error $\rho$ (1st) | Error $v$ (2nd) | Error $\rho$ (2nd) |
|---|---|---|---|---|
| Well-balanced | 1.54e-15 | 7.02e-13 | 1.35e-15 | 5.01e-13 |
| Non well-balanced | 0.01 | 2240.72 | 0.01 | 2250.77 |

Table 4.6: Well-balanced versus non-well-balanced schemes: $L^1$ errors at time $t = 50$ for the Burgers-Schwarzschild model with the initial condition (4.6.7)

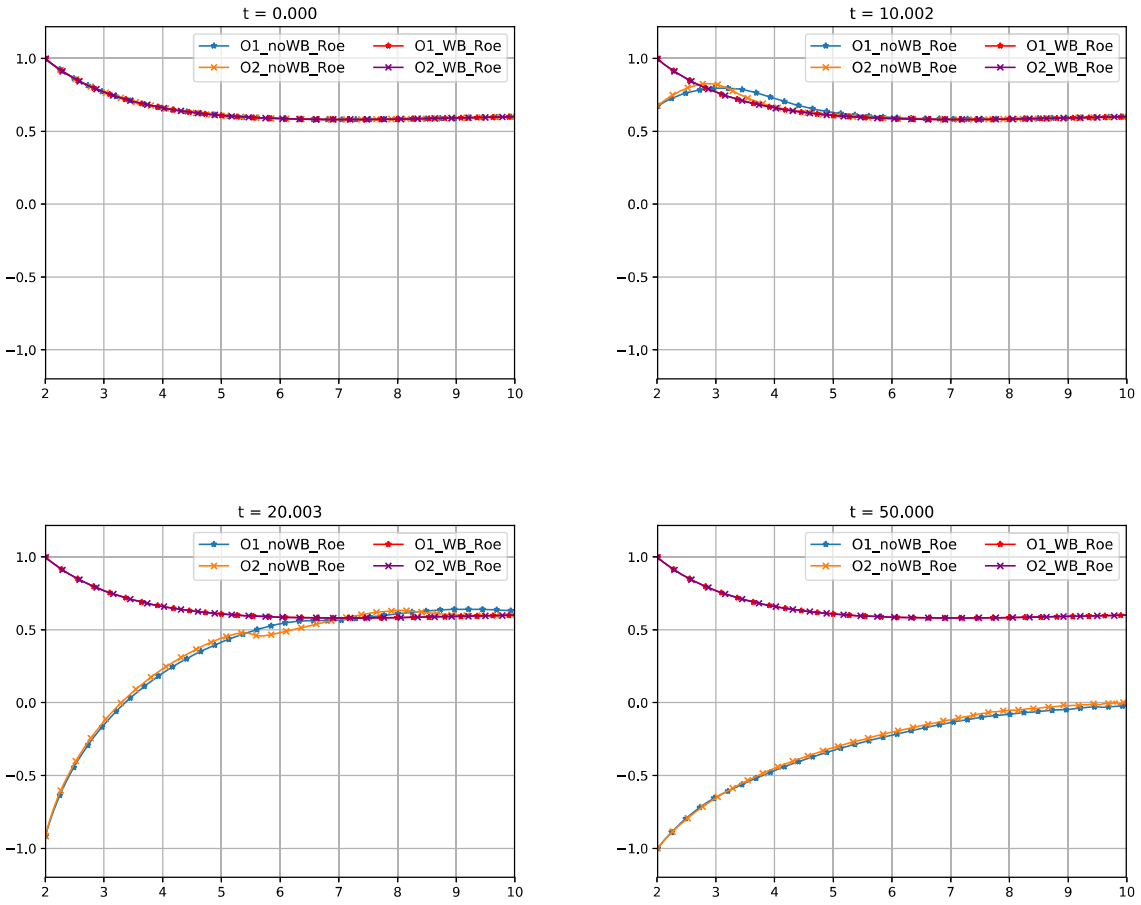



Figure 4.19: Euler-Schwarzschild model with the initial condition (4.6.6): first- and second-order well-balanced and non-well-balanced methods at selected times for the variable $v$.

### Discontinuous stationary entropy weak solution

Let us consider finally the initial condition

$$V_0(r) = \begin{cases} V_-^*(r) & \text{if } r \leq 6, \\ V_+^*(r) & \text{otherwise,} \end{cases} \tag{4.6.8}$$

where $V_-^*(r)$ is the supersonic stationary solution such that

$$\rho_-^*(6) = 4, \quad v_-^*(6) = 0.6 \tag{4.6.9}$$

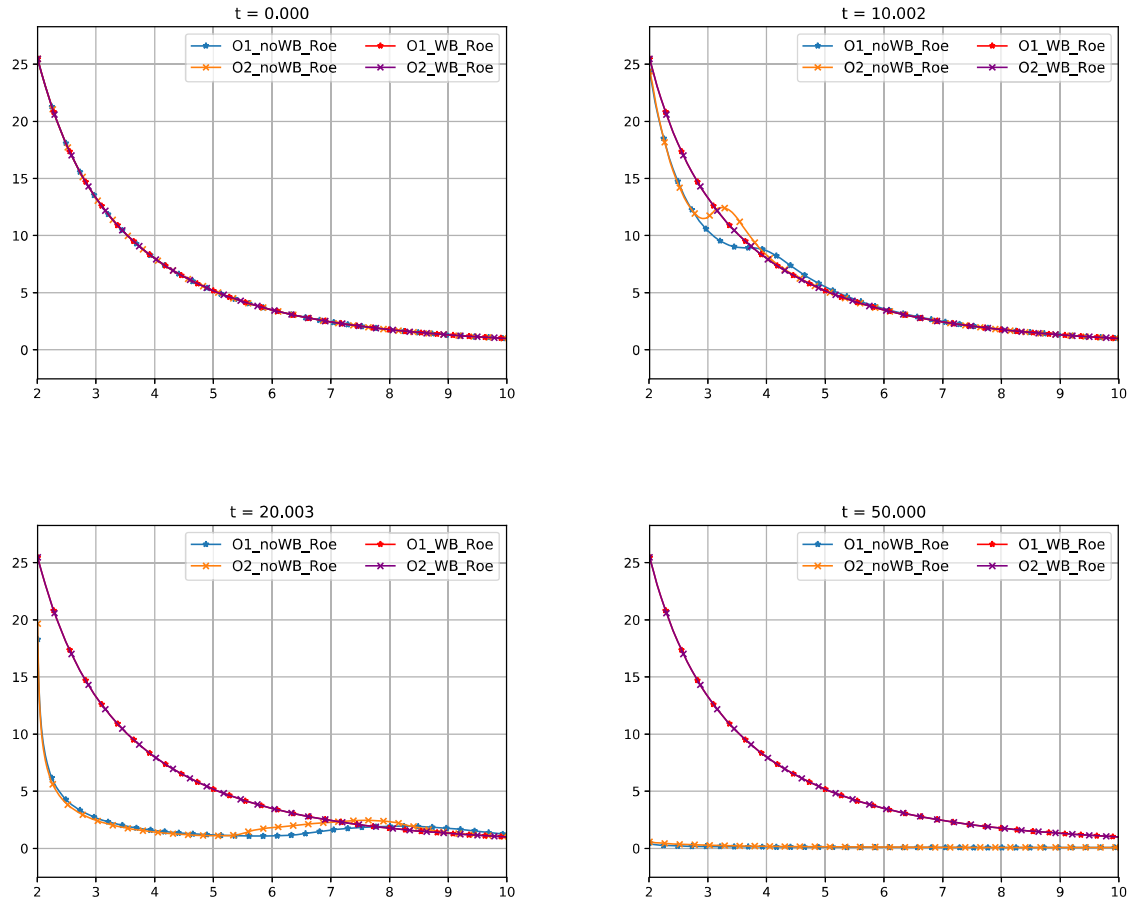



Figure 4.20: Euler-Schwarzschild model with the initial condition (4.6.6): first- and second-order well-balanced and non-well-balanced methods at selected times for the variable $\rho$.

and $V_+^*(r)$ is the subsonic one such that

$$\rho_+^*(6) = \frac{\rho_-^*(6)(v_-^*(6)^2 - k^4)}{k^2(1 - v_-^*(6)^2)}, \quad v_+^*(6) = \frac{k^2}{v_-^*(6)}. \tag{4.6.10}$$

$V_0$ is an entropy weak stationary solution of the system: see [114]. Table 4.7 shows the error in $L^1$ norm between the numerical solution at time $t = 50$ and Figures 4.23, 4.24 show the comparison of the numerical results obtained with well-balanced and non-well-balanced methods at selected times.

The numerical results of this section put on evidence, as for the Burgers-Schwarzschild system, the relevance of using well-balanced methods for the Euler-Schwarzschild model.

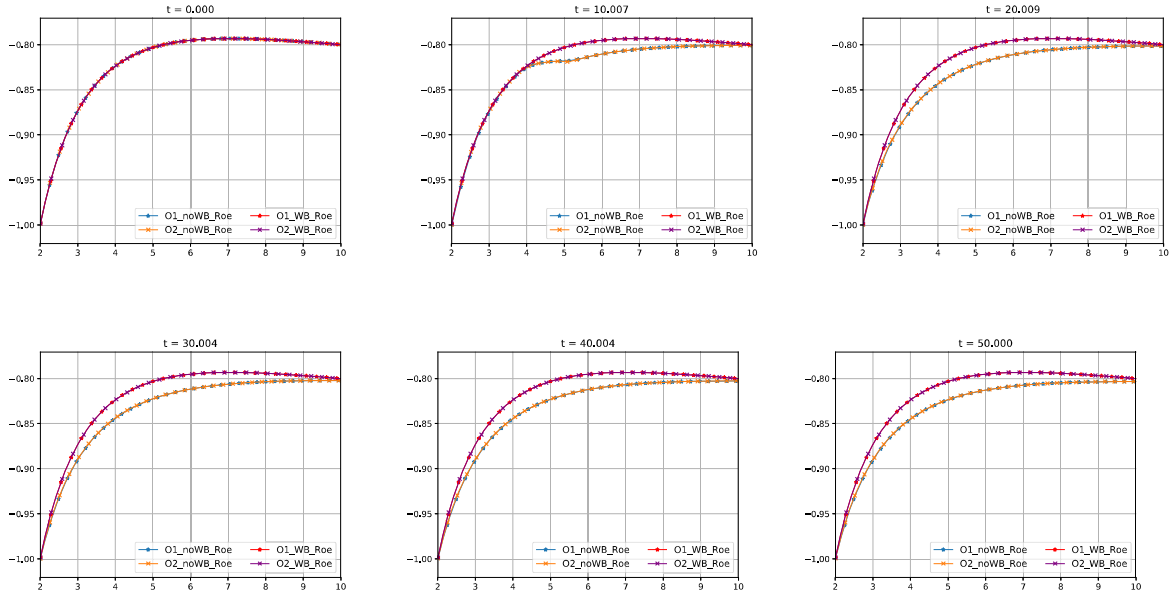



Figure 4.21: Euler-Schwarzschild model with the initial condition the stationary solution satisfying (4.6.7): first- and second-order well-balanced and non-well-balanced methods at selected times for the variable $v$.

| Scheme (500 cells) | Error $v$ (1st) | Error $\rho$ (1st) | Error $v$ (2nd) | Error $\rho$ (2nd) |
|---|---|---|---|---|
| Well-balanced | 2.20e-13 | 1.25e-11 | 1.92e-13 | 1.03e-11 |
| Non well-balanced | 0.89 | 3.94 | 0.89 | 3.92 |

Table 4.7: Well-balanced versus non-well-balanced schemes: $L^1$ errors at time $t = 50$ for the Burgers-Schwarzschild model with the initial condition (4.6.8)

### 4.6.3 Perturbation of a regular stationary solution

In this test we consider the initial condition

$$\widetilde{V}_0(r) = \widetilde{V}^*(r) + \delta(r), \tag{4.6.11}$$

where $\widetilde{V}^*$ is the supersonic stationary solution such that

$$\rho^*(10) = 1, \qquad v^*(10) = 0.9 \tag{4.6.12}$$

and

$$\delta(r) = [\delta_v(r), \delta_\rho(r)]^T = \begin{cases} [-0.01e^{-200(r-6)^2}, 0]^T & \text{if } 5 < r < 7, \\ [0,0]^T & \text{otherwise.} \end{cases} \tag{4.6.13}$$

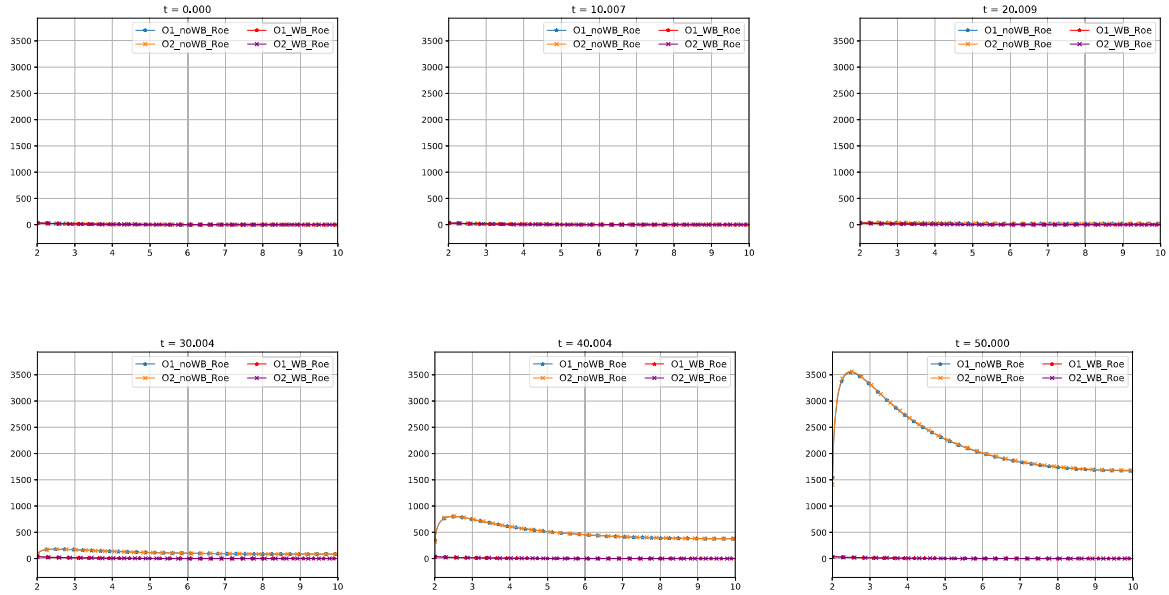



Figure 4.22: Euler-Schwarzschild model with the initial condition (4.6.7): first- and second-order well-balanced and non-well-balanced methods at selected times for the variable $\rho$.

It can be observed in Figure 4.25 that the stationary solution $V^*$ is recovered once the perturbation has left the domain. In this Figure, the numerical results obtained with the first- and second-order well-balanced methods are compared with a reference solution computed using the first-order well-balanced method with a 5000-point mesh (the Roe-type numerical flux is used again). As expected, the second-order method is less diffusive.

### 4.6.4 Perturbation of a steady shock solution

#### Left side perturbation

We consider the initial condition:

$$\widetilde{V}_0(r) = \widetilde{V}^*(r) + \delta_L(r), \tag{4.6.14}$$

where

$$\delta_L(r) = [\delta_{v,L}(r), \delta_{\rho,L}(r)]^T = \begin{cases} [0.2e^{-200(r-4)^2}, 0]^T & \text{if } 3 < r < 5, \\ [0,0]^T & \text{otherwise,} \end{cases} \tag{4.6.15}$$

and $V^*(r)$ is the stationary solution given by (4.6.8)-(4.6.10).

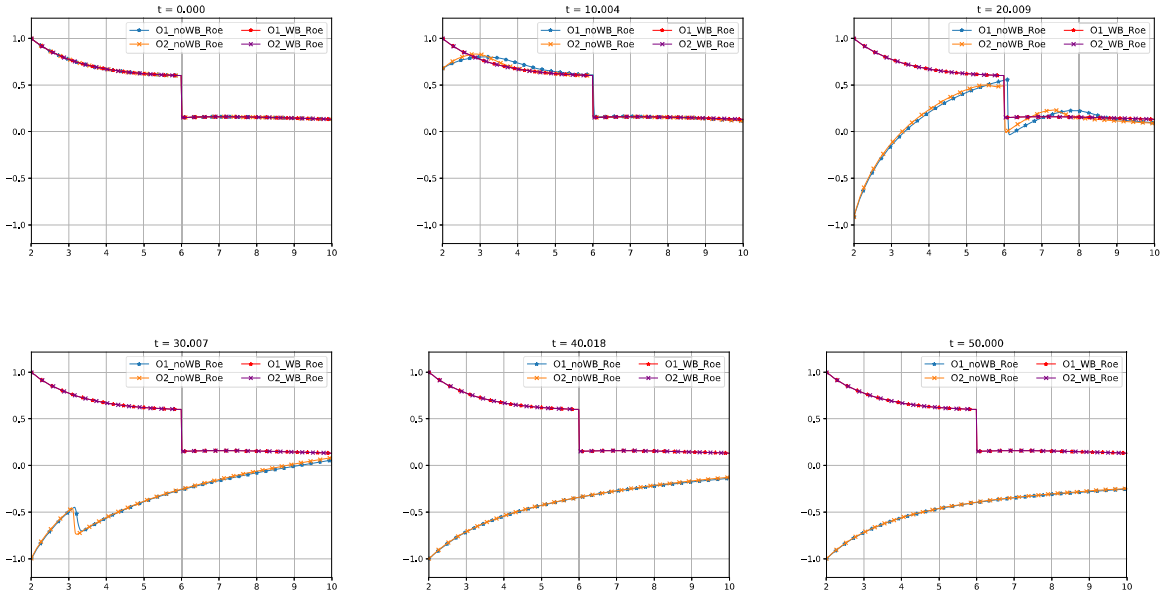



Figure 4.23: Euler-Schwarzschild model with the initial condition (4.6.8): first- and second-order well-balanced and non-well-balanced methods at selected times for the variable $v$.

In Figures 4.26 and 4.27 the numerical results obtained with the first- and second-order well-balanced methods using the Lax-Friedrichs and the Roe-type numerical methods with different meshes are compared. As it happened for the Burgers-Schwarzschild model, the location of the stationary shock changes after the passage of the wave generated by the perturbation. Nevertheless in this case the movement of the shock is slower. Different numerical methods have been applied to check the dependency of the motion on the scheme: although the evolution of the shock slightly depends on the number of points of the mesh, all the numerical solutions capture the same final location of the shock. In Figure 4.28 the evolution of the shock given by the first-order WB method with different number of cells are compared. The location of the shock at every time step has been detected by using the condition $\dfrac{v_i - v_{i-1}}{v_{i+1} - v_i} \geq 0.8$.

### Right side perturbation

Let us consider now two different initial conditions: on the one hand

$$\widetilde{V}_{0,1}(r) = \widetilde{V}^*(r) + \delta_R(r), \tag{4.6.16}$$

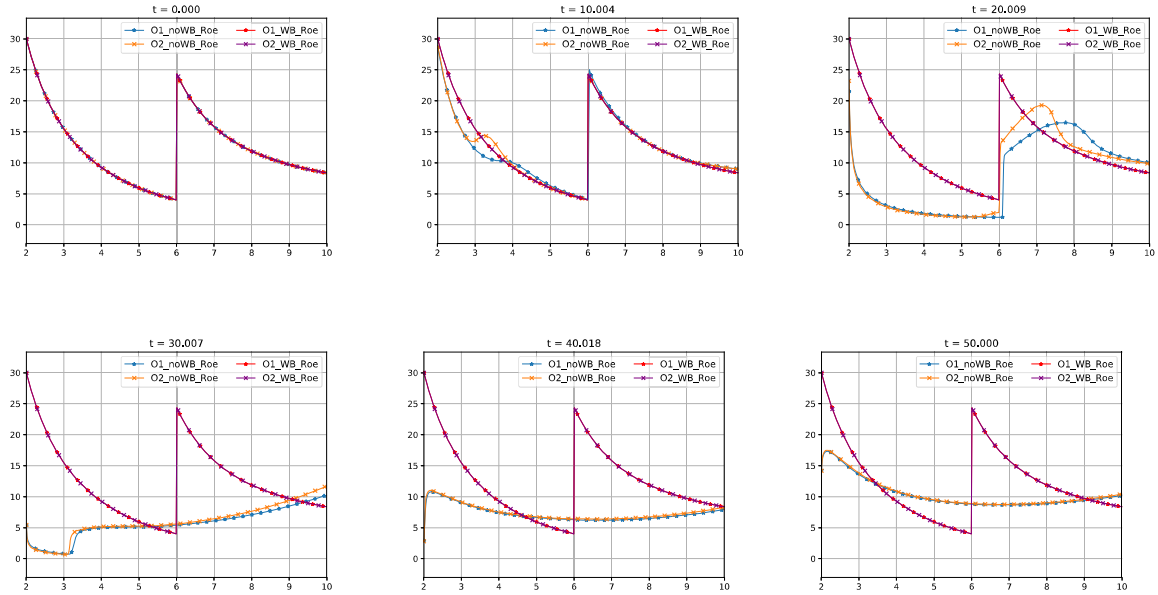



Figure 4.24: Euler-Schwarzschild model with the initial condition (4.6.8): first- and second-order well-balanced and non-well-balanced methods at selected times for the variable $\rho$.

where $\widetilde{V}^*(r)$ is again the discontinuous stationay solution given by (4.6.8)-(4.6.10) and

$$\delta_R(r) = [\delta_{v,R}(r), \delta_{\rho,R}(r)]^T = \begin{cases} [-0.05e^{-200(r-8)^2}, 0]^T & \text{if } 7 < r < 9, \\ [0,0]^T & \text{otherwise.} \end{cases} \tag{4.6.17}$$

On the other hand,

$$\widetilde{V}_{0,2}(r) = \widetilde{V}_2^*(r) + \delta_R(r), \tag{4.6.18}$$

where $\delta_R$ is given again by (4.6.16) and $\widetilde{V}_2^*(r)$ is the steady shock of the form (4.6.8) satisfying

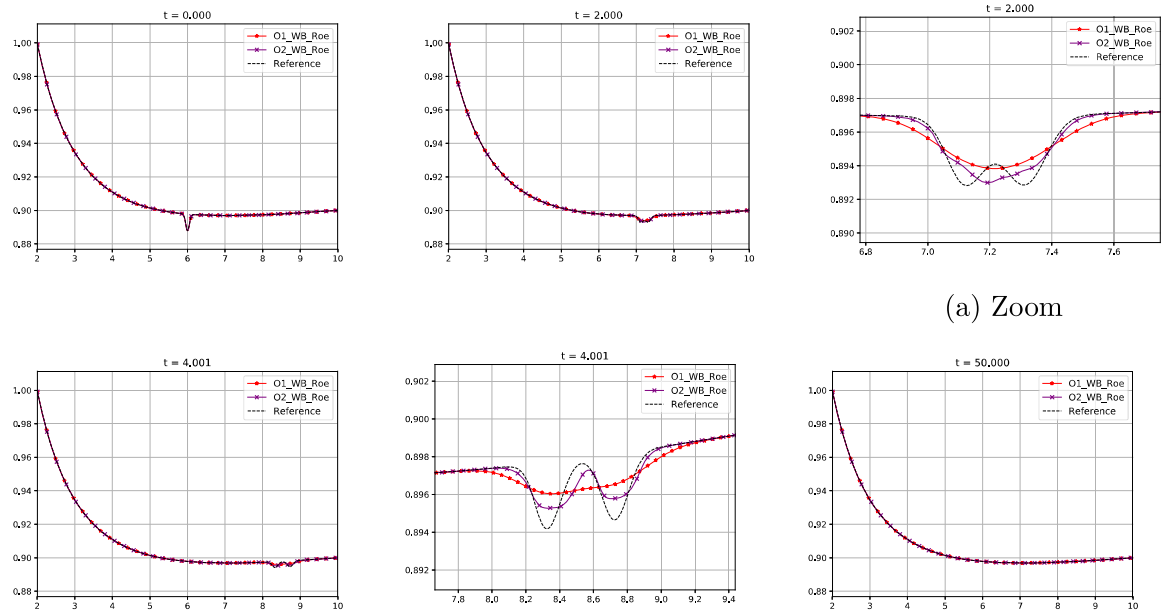
$$\rho_-^*(6) = 5, \quad v_-^*(6) = 0.6. \tag{4.6.19}$$

Observe that the definition of $v$ is identical for both stationary solutions but $\rho$ is different.

After the passage of the perturbation, the shock starts moving leftward and, in both cases, the numerical solution converges to a smooth transonic stationary solution of the form:

$$V^*(r) = \begin{cases} V_-^*(r) & \text{if } r \le r_c, \\ V_+^*(r) & \text{otherwise,} \end{cases} \tag{4.6.20}$$

where $r_c$ is given by (4.5.4); $V_-^*(r)$ and $V_+^*(r)$ are respectively a subsonic and a supersonic stationary solution satisfying $v_\pm^*(r_c) = -k$: see Figure 4.29. Nevertheless, the limits




(a) Zoom




(b) Zoom

Figure 4.25: Euler-Schwarzschild model with the initial condition (4.6.11): first- and second-order well-balanced at selected times for the variable $v$.

in time of the approximations of $\rho$ are different: see Figure 4.30. Observe that, in the Euler-Schwarzschild model (4.5.1), there are infinitely many stationary solutions with the same function $v$ and different $\rho$.

### Relation between the perturbation and the displacement of the shock

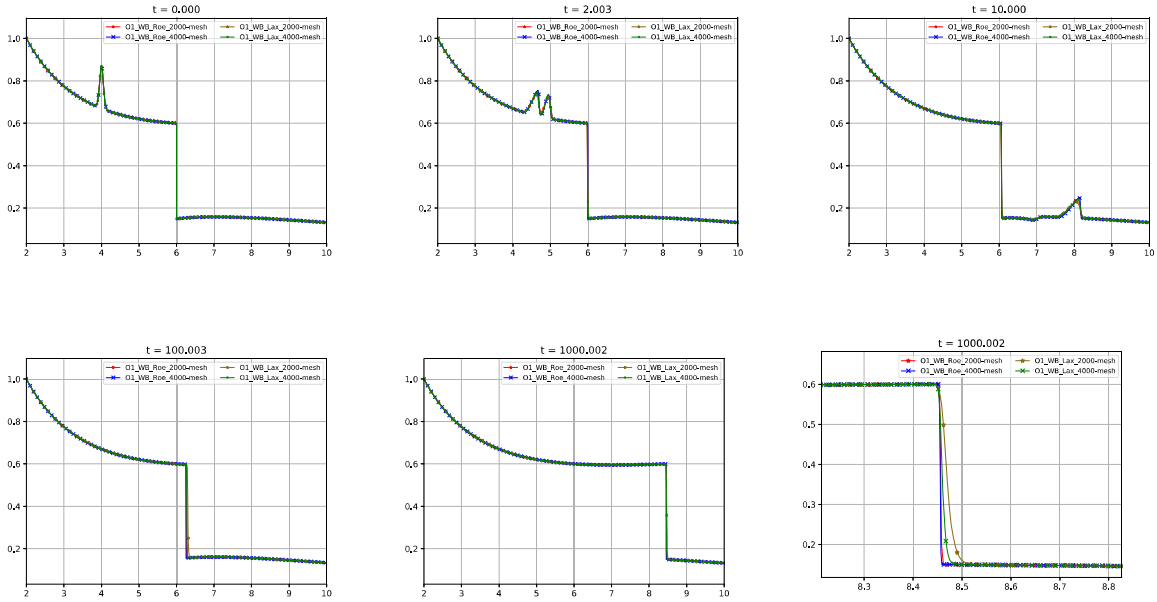
In order to study the relationship between the amplitude of the perturbation and the distance between the initial and the final shock locations, we consider the family of initial conditions:

$$\widetilde{V}_0(r) = \widetilde{V}^*(r) + \delta(\alpha, r), \tag{4.6.21}$$

where $\widetilde{V}^*$ is the steady shock solution given by (4.6.8)-(4.6.10) and

$$\delta(\alpha, r) = [\delta_v(\alpha, r), \delta_\rho(\alpha, r)]^T = \begin{cases} [\alpha e^{-200(r-4)^2}, 0]^T & \text{if } 3 < r < 5, \\ [0, 0]^T & \text{otherwise,} \end{cases} \tag{4.6.22}$$

with $\alpha > 0$. In this case we will also use the Roe-type numerical flux and a 2000-point uniform mesh. Figures 4.31 and 4.32 show the numerical solution for different values of




(a) Zoom

Figure 4.26: Euler-Schwarzschild model with the initial condition (4.6.14): comparison between the first-order well-balanced method with different meshes using the Roe-type and the Lax numerical fluxes at selected times for the variable $v$.

$\alpha$ and we observe that depending on the amplitude of the perturbation the numerical solutions converge in time to different steady shock solutions.

The amplitude of the perturbation is measured with $\int \delta_v(\alpha, r)\, dr$ and the distance between the shocks are measured by

$$\lim_{t\to\infty} \int |v(r,t) - v^*(r)|\, dr,$$

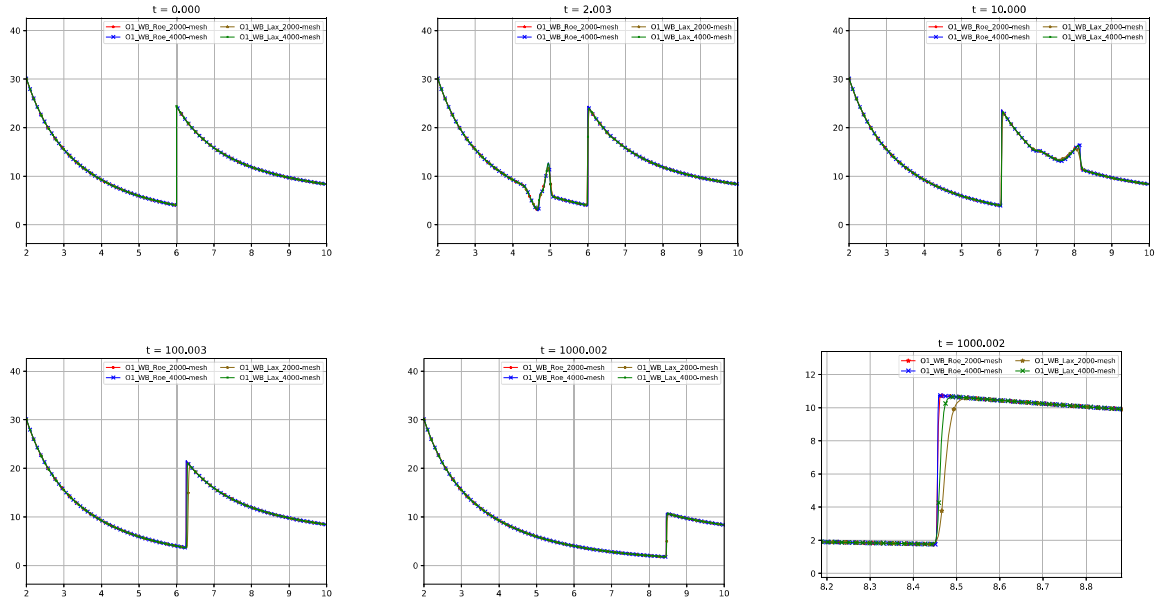
as we did for the Burgers-Schwarzschild model. Table 4.8 and Figure 4.33 show the relationship between those magnitudes: the displacement of the shock seems to grow exponentially with the amplitude.

Let us finally consider a family of initial conditions that generate leftward movement of the initial steady shock :

$$\widetilde{V}_0(r) = \widetilde{V}^*(r) + \delta(\beta, r), \tag{4.6.23}$$

where $\widetilde{V}^*$ is again the steady shock solution given by (4.6.8)- (4.6.10) and

$$\delta(\beta, r) = [\beta v(\beta, r), \delta_\rho(\beta, r)]^T = \begin{cases} [\beta e^{-200(r-8)^2}, 0]^T & \text{if } 7 < r < 8, \\ [0, 0]^T & \text{otherwise,} \end{cases} \tag{4.6.24}$$




(a) Zoom

Figure 4.27: Euler-Schwarzschild model with the initial condition (4.6.14): first-order well-balanced method with different meshes using the Roe-type and the Lax numerical fluxes at selected times for the variable $\rho$.

| $\alpha$ | $\int \delta_v$ | $\lim_{t\to\infty} \int \lvert v - v^* \rvert$ |
|---|---|---|
| 0.05 | 0.0063 | 1.0952 |
| 0.1 | 0.0125 | 1.0969 |
| 0.15 | 0.0188 | 1.0987 |
| 0.2 | 0.0251 | 1.1023 |
| 0.25 | 0.0313 | 1.1077 |
| 0.3 | 0.0376 | 1.1327 |

Table 4.8: Euler-Schwarzschild model with the initial condition (4.6.21): measures of the perturbation and the shock displacement for different values of $\alpha$

with $\beta < 0$. In this case we will use the Roe-type numerical flux and a 2000-point uniform mesh. Figures 4.34 and 4.35 show the numerical solution for different values of $\beta$ and we observe that it converges to same stationary solution without depending on the perturbation.

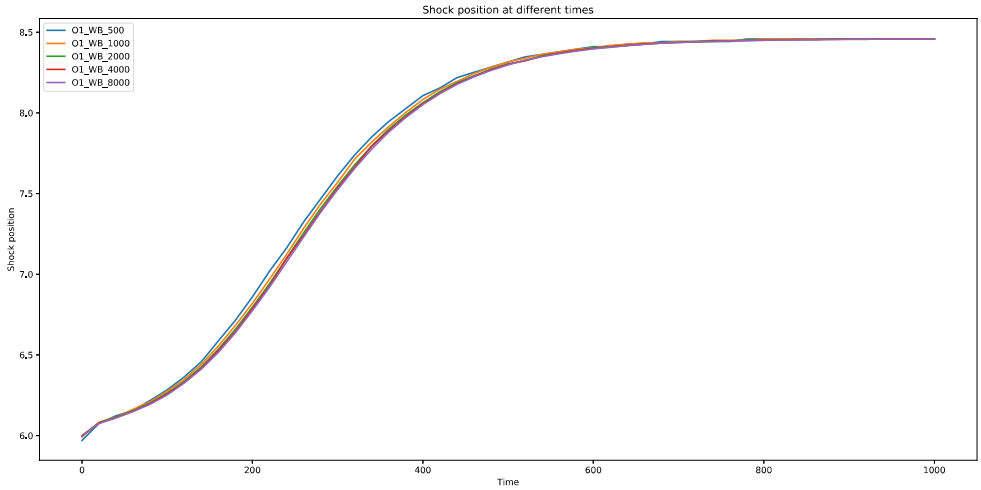



Figure 4.28: Euler-Schwarzschild model taking as initial condition (4.6.14): evolution of the shock position with time obtained with the first-order well-balanced method using the Roe-type numerical flux with different meshes.

### 4.6.5 Main conclusions for the Euler-Schwarzschild model

From Figure 4.25 we can conclude the following:

**Conclusion 3.** *If a smooth stationary solution of the Euler system (4.1.5) is perturbed, the solution is restored once the wave generated by the perturbation goes away.*

From Figures 4.26 to 4.35 and Table 4.8 we can conclude the following:

**Conclusion 4.** *If a perturbation $\delta = (\delta_v, \delta_\rho)$ is added to a steady shock solution of the form*

$$V_0(r) = \begin{cases} V_-^*(r) & \text{if } r \leq r_0, \\ V_+^*(r) & \text{otherwise,} \end{cases}$$

*then:*

1. *If the perturbation moves the steady shock to the right, then a different stationary solution of the form*

$$V(r) = \begin{cases} V_-^*(r) & \text{if } r \leq r_1, \\ V_+^*(r) & \text{otherwise,} \end{cases}$$

*with $r_0 \neq r_1$, is obtained, and the distance between $r_0$ and $r_1$ seems to depend exponentially on the amplitude of the perturbation: see Table 4.8 and Figure 4.33.*

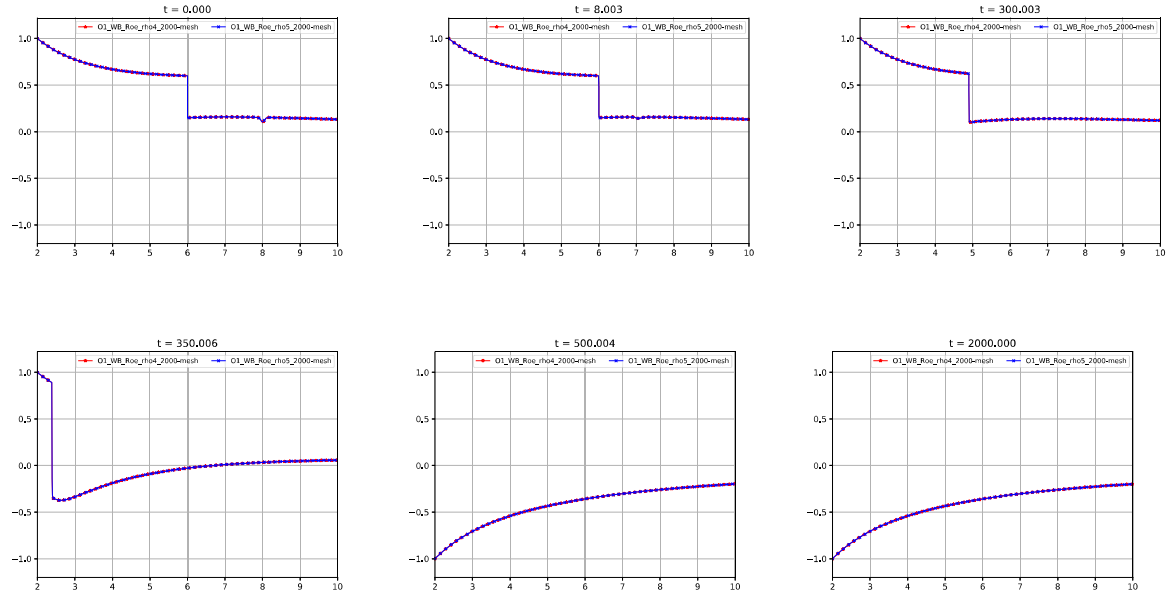



Figure 4.29: Euler-Schwarzschild model with the initial conditions (4.6.16) and (4.6.18): first-order well-balanced method with a 2000-point mesh using the Roe-type numerical flux at selected times for the variable $v$: the numerical solutions coincide.

2. *If the perturbation moves the steady shock to the left, then a steady shock solution of the form* (4.6.20) *is obtained.*

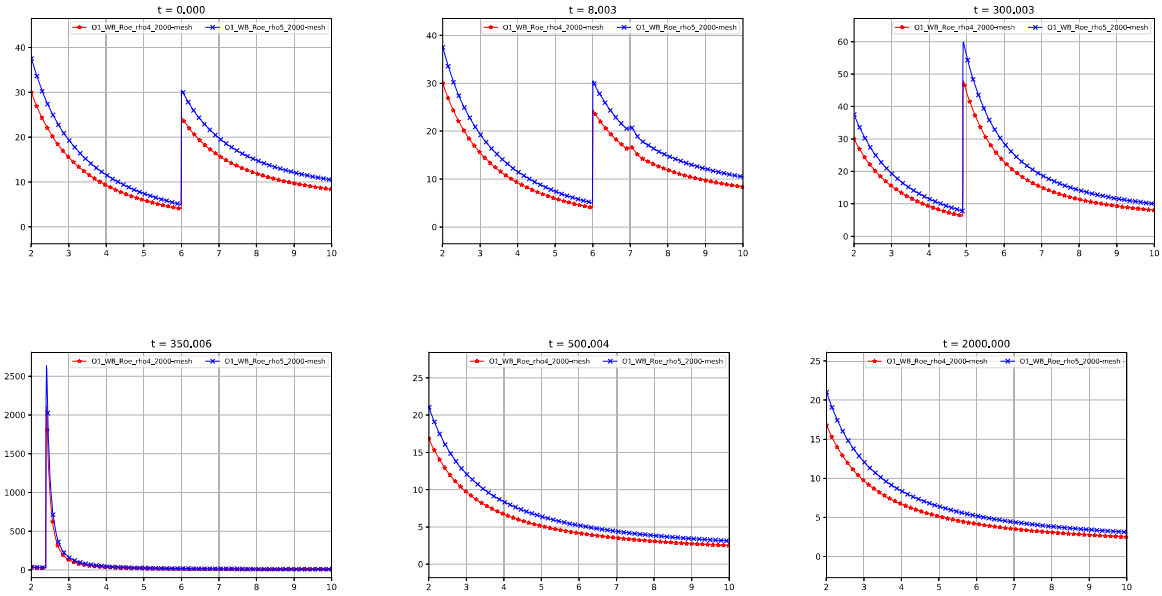
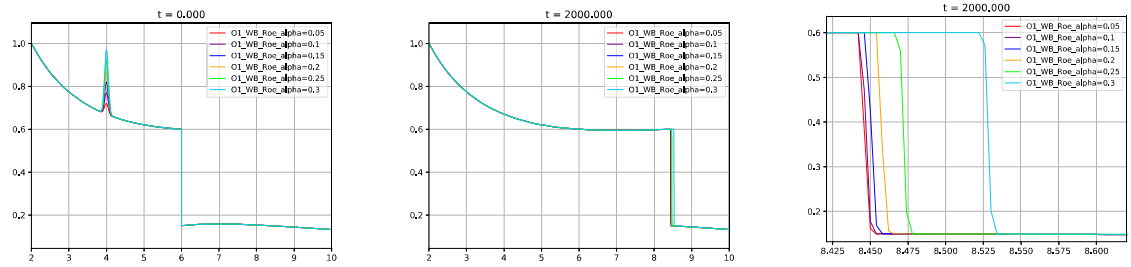
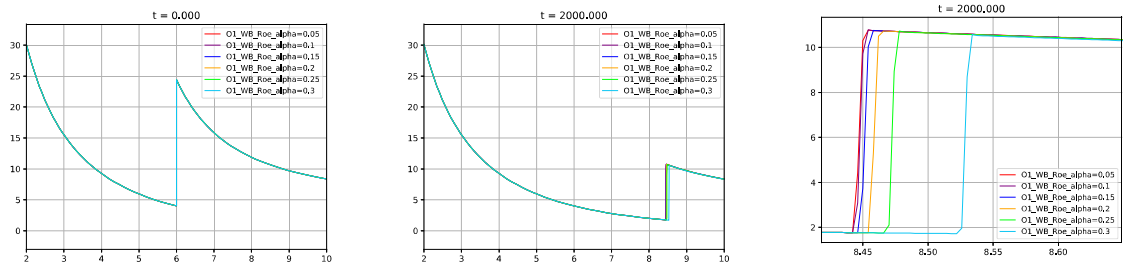



Figure 4.30: Euler-Schwarzschild model with the initial conditions (4.6.16) and (4.6.18): first-order well-balanced method with a 2000-point mesh using the Roe-type numerical flux at selected times for the variable $\rho$.




(a) Zoom

Figure 4.31: Euler-Schwarzschild model with the initial condition (4.6.21): first-order well-balanced method taking different values of $\alpha$ for variable $v$.




(a) Zoom

Figure 4.32: Euler-Schwarzschild model with the initial condition (4.6.21): first-order well-balanced method taking different values of $\alpha$ for variable $\rho$.

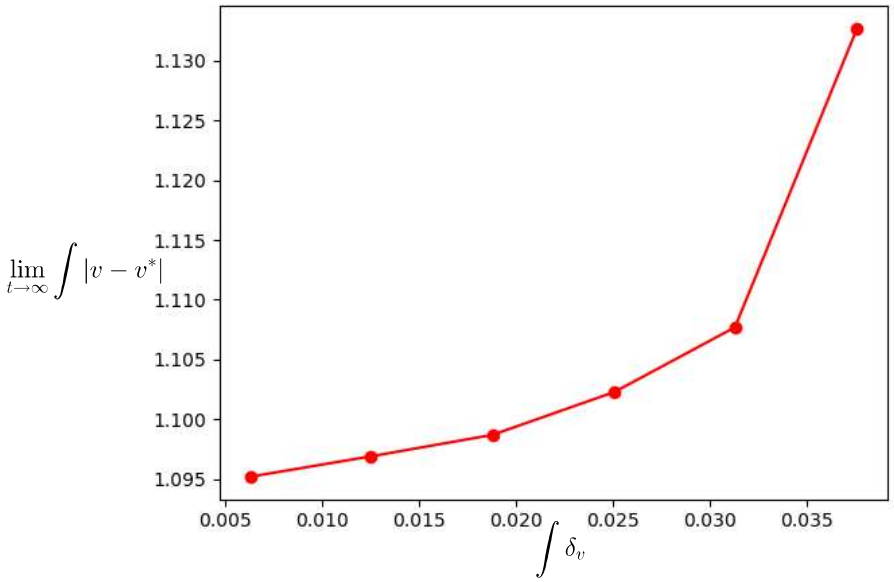



Figure 4.33: Euler-Schwarzschild model with the initial condition (4.6.21): values of $\lim_{t\to\infty} \int |v - v^*|$ as a function of $\int \delta_v$.

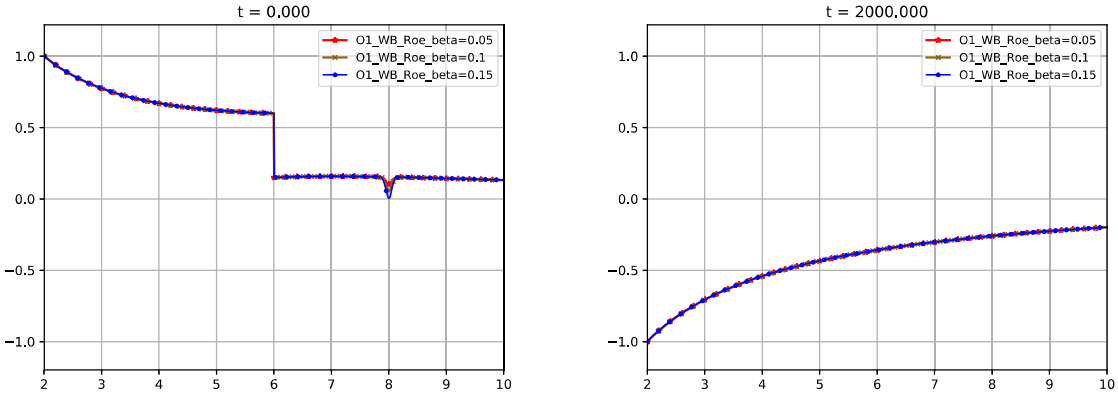



Figure 4.34: Euler-Schwarzschild model with the initial condition (4.6.23): first-order well-balanced method taking different values of $\beta$ for variable $v$.

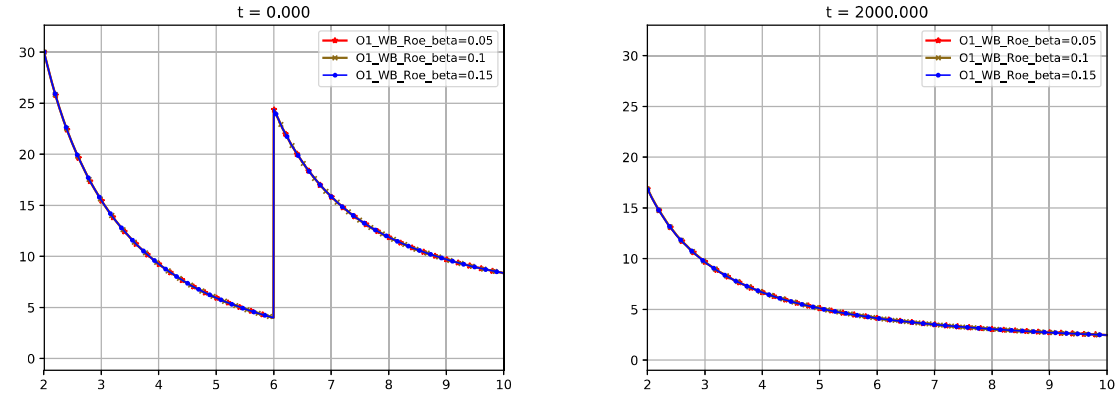



Figure 4.35: Euler-Schwarzschild model with the initial condition (4.6.23): first-order well-balanced method taking different values of $\beta$ for variable $\rho$.

# Chapter 5

# In-cell Discontinuous Reconstruction path-conservative methods for non conservative hyperbolic systems: Second-order extension

It is a well-known fact that, in the case of systems of conservation laws, in order to ensure the convergence to the right weak solutions of the approximations provided by a method, besides consistency and stability, entropy has to be well controlled: for instance, entropy-fix techniques have to be added to Roe methods (see [39]). However, in the case of nonconservative systems, consistency, stability, and control of the entropy are not enough: the numerical viscosity and, in general, the numerical dissipation effects, have to be well controlled (see [109] for a review on this topic). Even the Godunov method, consisting of averaging the exact solutions of the Riemann problems, fails in general to converge to the selected weak solution: see for instance [95], [42], [18], [109] and the references therein.

The design of finite-difference or finite-volume methods satisfying these four properties is difficult in general. The theoretical framework of path-conservative methods introduced in [132] facilitates the design of schemes that are formally consistent with the definition of weak solution based on the theory of Dal Maso, LeFloch and Murat [57] that depends on the election of a family of paths. As we have seen in Chapter 1, these schemes facilitate the extension of well-known families of conservative methods such as Godunov and Roe to nonconservative systems, preserving its stability properties. This framework helps also to obtain high-order methods and to design well-balanced schemes. Nevertheless, as the concept of path-conservative method is just a formal definition of consistency, it is not enough to ensure the convergence to the right solutions if the entropy and the dissipation are not well controlled: see [42], [1]. Different techniques have been introduced to overcome, at least partially, this convergence issue: [15], [10], [4], [17], [39], [52], [53],

[77], [132]. In particular the path-conservative entropy stable methods introduced in [39] and extended to DG high-order methods in [94] significantly reduce the convergence error. More recently, in [51], an in-cell discontinuous reconstruction technique has been added to first-order path-conservative methods that allows one to capture correctly weak solutions with isolated shock waves.

The main objective of this chapter is to extend the in-cell discontinuous reconstruction methods introduced in [51] to second-order accuracy and to set the basis of an extension to high-order methods following this in-cell methodology and a way for capturing exactly more than one shock. To do this, these numerical methods will be first written as a high-order path-conservative scheme (see for example [31, 34]) and then, depending of the smoothness of the numerical solution, a standard MUSCL-Hancock reconstruction (see [158] and [161]) or a discontinuous one is used in the cell to update the numerical solution.

The chapter is organized as follows: In Section 5.1 a brief recall of some contents from Chapter 1 is given, then in Section 5.2 the new family of second-order in-cell discontinuous reconstruction methods is presented. First the semi-discrete method is introduced including the description of the reconstructions in the cells; then a temporal discretization based on a second order Taylor development is introduced. The shock-capturing property of the method is then enunciated and proved. Section 5.3 is devoted to show numerically the efficiency of the proposed numerical scheme. More precisely, first the Coupled-Burgers nonconservative system introduced as a toy problem in [44] is considered: the application of the method to this system is described and several numerical tests are shown to validate the methods. Next, we focus on the gas dynamics equations in Lagrangian coordinates and the modified shallow water system introduced in [42]. These systems were used in [1] and [42] respectively to illustrate the convergence issue of path-conservative methods when small-scale effects are not controlled: the method proposed in this chapter is applied to these system to put on evidence that the convergence issue is corrected.

The content of this chapter was introduced by Pimentel-García, Castro, Chalons, Morales de Luna, Parés, it is available in the *arXiv* repository and was submitted in April 2021 to *Journal of Computational Physics*, see [137].

## 5.1 Preliminaries

Let us remember some concepts from Chapter 1 that we will use. Let us consider first order quasi-linear PDE systems of the form (1.1.23)

$$\partial_t W + \mathcal{A}(W)\partial_x W = 0, \quad x \in \mathbb{R}, \quad t \in \mathbb{R}^+, \tag{5.1.1}$$

in which the unknown $W(x,t)$ takes values in an open convex set $\Omega$ of $\mathbb{R}^N$, and $\mathcal{A}(W)$ is a smooth locally bounded map from $\Omega$ to $\mathcal{M}_{N\times N}(\mathbb{R})$. The system is supposed to be strictly

hyperbolic and the characteristic fields $R_i(W)$, $i = 1, \dots, N$, are supposed to be either genuinely nonlinear:

$$\nabla\lambda_i(W) \cdot R_i(W) \neq 0, \quad \forall W \in \Omega,$$

or linearly degenerate:

$$\nabla\lambda_i(W) \cdot R_i(W) = 0, \quad \forall W \in \Omega.$$

Here, $\lambda_1(W), \dots, \lambda_N(W)$ represent the eigenvalues of $\mathcal{A}(W)$ (in increasing order) and $R_1(W), \dots, R_N(W)$ a set of associated eigenvectors. In order to give sense to the nonconservative products we will follow again the theory of Dal Maso, LeFloch and Murat [57]: it allows one to define the nonconservative product $\mathcal{A}(W)\,W_x$ as a bounded measure for functions $W$ with bounded variation. Let us consider a family of Lipschitz continuous paths $\Phi$. We remember that the family of paths can be understood as a tool to give a sense to integrals of the form

$$\int_a^b \mathcal{A}(W(x))\,W_x(x)\,dx,$$

for functions $W$ with jump discontinuities. More precisely, given a bounded variation function $W : [a, b] \mapsto \Omega$, we remember the following definition (1.1.25):

$$\int_a^b \mathcal{A}(W(x))\,W_x(x)\,dx = \int_a^b \mathcal{A}(W(x))\,W_x(x)\,dx + \sum_m \int_0^1 \mathcal{A}(\Phi(s; W_m^-, W_m^+))\frac{\partial\Phi}{\partial s}(s; W_m^-, W_m^+)\,ds. \tag{5.1.2}$$

In this definition, $W_m^-$ and $W_m^+$ represent, respectively, the limits of $W$ to the left and right of its $m$th discontinuity. Observe that, in (5.1.2), the family of paths has been used to determine the Dirac measures placed at the discontinuities of $W$.

Once this definition of the nonconservative products is assumed to define the concept of weak solution, we recall the generalized Rankine-Hugoniot condition (1.1.27):

$$\int_0^1 \mathcal{A}(\Phi(s; W^-W^+))\frac{\partial\Phi}{\partial s}(s; W^-, W^+)\,ds = \sigma(W^+ - W^-), \tag{5.1.3}$$

which has to be satisfied across an admissible discontinuity. Here, $\sigma$ is the speed of propagation of the discontinuity, and $W^-$ and $W^+$ are the left and right limits of the solution at the discontinuity.

Once the family of paths has been prescribed, a concept of entropy is required, as it happens for systems of conservation laws, that may be given by an entropy pair (see Definition 1.1.12) or by Lax's entropy criterion (see Definition 1.1.10).

## 5.2 Second-order in-cell discontinuous reconstruction path-conservative methods

In this section, a numerical method of the form (1.2.48) is described in which a first-order path-conservative numerical method with fluctuation functions $\mathcal{D}^{\pm}$ is combined with a

standard second-order reconstruction operator in smoothness regions and a discontinuous reconstruction operator close to discontinuities so that no numerical viscosity is added in the non-smooth regions.

### 5.2.1 Semi-discrete method

Once the numerical approximations $W_i^n$ of the averages of the solutions have been computed at time $t_n = n\Delta t$, the first step is to mark the cells $I_i$ such that the solution of the Riemann problem consisting of (5.1.1) with initial conditions

$$W(x,0) = \begin{cases} W_{i-1}^n & \text{if } x < 0, \\ W_{i+1}^n & \text{if } x > 0, \end{cases} \tag{5.2.1}$$

involves a shock wave. Let us denote by $\mathcal{M}_n$ the set of indices of the marked cells, i.e.

$$\mathcal{M}_n = \{i \text{ s.t. the solution of the Riemann problem (5.1.1), (5.2.1) involves a shock wave}\}. \tag{5.2.2}$$

To advance in time the semi-discrete numerical method (1.2.48) is considered:

$$W_i'(t) = -\frac{1}{\Delta x}\left(\mathcal{D}^-_{i+\frac{1}{2}}(t) + \mathcal{D}^+_{i-\frac{1}{2}}(t) + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^n(x,t))\frac{\partial}{\partial x}\mathbb{P}_i^n(x,t)dx\right), \quad t \geq t_n, \tag{5.2.3}$$

where

$$W_i(t) \approx \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} W(x,t)dx,$$

$$\mathcal{D}^{\pm}_{i+1/2}(t) = \mathcal{D}^{\pm}_{i+1/2}(W^-_{i+1/2}(t), W^+_{i+1/2}(t)),$$

with

$$W^-_{i+1/2}(t) = \mathbb{P}_i^n(x_{i+\frac{1}{2}}, t), \quad W^+_{i+1/2}(t) = \mathbb{P}_{i+1}^n(x_{i+\frac{1}{2}}, t),$$

and $\mathbb{P}_i^n(x,t)$ is defined as follows:

- If $i-1, i, i+1 \notin \mathcal{M}_n$ then $\mathbb{P}_i^n$ is the approximation of the first degree Taylor polynomial of the solution given by:

$$\mathbb{P}_i^n(x,t) = W_i^n + \widetilde{\partial_x W}_i^n(x - x_i) - \mathcal{A}(W_i^n)\widetilde{\partial_x W}_i^n(t - t^n).$$

  Here, $\widetilde{\partial_x W}_i^n$ is the *minmod* approximation of the first order spacial derivative of $W$ at $x_i$ at time $t_n$, whose $k$th component is given by

$$\left(\widetilde{\partial_x W}_i^n\right)_k = \text{minmod}\left(\alpha\frac{\mathrm{w^n_{i+1,k}} - \mathrm{w^n_{i,k}}}{\Delta \mathrm{x}}, \frac{\mathrm{w^n_{i+1,k}} - \mathrm{w^n_{i-1,k}}}{2\Delta \mathrm{x}}, \alpha\frac{\mathrm{w^n_{i,k}} - \mathrm{w^n_{i-1,k}}}{\Delta \mathrm{x}}\right),$$

where $w_{i,k}^n$ represents the $k$th component of $W_i^n$, $\alpha$ is a parameter with $1 \leq \alpha < 2$ and

$$\text{minmod(a, b, c)} = \begin{cases} \min\{a,b,c\} & \text{if } a,b,c > 0, \\ \max\{a,b,c\} & \text{if } a,b,c < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that, using the equation:

$$\partial_t W(x_i,t_n) = -\mathcal{A}(W(x_i,t_n))\partial_x W(x_i,t_n) \approx -\mathcal{A}(W_i^n)\widetilde{\partial_x W}_i^n.$$

- If $i \in \mathcal{M}_n$ then

$$\mathbb{P}_i^n(x,t) = \begin{cases} W_{i,l}^n & \text{if } x \leq x_{i-1/2} + d_i^n \Delta x + \sigma_i^n (t - t_n), \\ W_{i,r}^n & \text{otherwise.} \end{cases}.$$

where $d_i^n$ is chosen so that

$$d_i^n w_{i,l,k}^n + (1 - d_i^n) w_{i,r,k}^n = w_{i,k}^n, \tag{5.2.4}$$

for some index $k \in \{1, \ldots, N\}$; and $\sigma_i^n$, $W_{i,l}^n$, and $W_{i,r}^n$ are chosen so that if $W_{i-1}^n$ and $W_{i+1}^n$ may be linked by an admissible discontinuity with speed $\sigma$, then

$$W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_{i+1}^n, \quad \sigma_i^n = \sigma. \tag{5.2.5}$$

Observe that this in-cell discontinuous reconstruction can only be done if $0 \leq d_i^n \leq 1$, i.e. if

$$0 \leq \frac{w_{i,r,k}^n - w_{i,k}^n}{w_{i,r,k}^n - w_{i,l,k}^n} \leq 1,$$

otherwise the index $i$ is removed from the set $\mathcal{M}_n$ and the MUSCL-Hancock reconstruction is applied in the cell. Moreover, if $d_i^n = 1$ and $\sigma_i^n > 0$ (resp. $d_i^n = 0$ and $\sigma_i^n < 0$) the cell is unmarked and the cell $I_{i+1}$ (resp. $I_{i-1}$) is marked if necessary: note that in these cases, the discontinuity leaves the cell $I_i$ for any $t > t_n$.

- Otherwise (i.e. if $i \notin \mathcal{M}_n$ but $i - 1 \in \mathcal{M}_n$ or $i + 1 \in \mathcal{M}_n$)

$$\mathbb{P}_i^n(x,t) = W_i^n.$$

**Remark 5.2.1.** *In the case $i \in \mathcal{M}_n$, if one of the equations of system* (5.1.1)*, say the $k$th one, is a conservation law, the index $k$ is selected in* (5.2.4)*, so that the corresponding variable is conserved. Moreover, if there is a linear combination of the unknowns $\sum_{k=1}^N \alpha_k w_k$ that is conserved,* (5.2.4) *may be replaced by:*

$$d_i^n \sum_{k=1}^N \alpha_k w_{i,l,k}^n + (1 - d_i^n) \sum_{k=1}^N \alpha_k w_{i,r,k}^n = \sum_{k=1}^N \alpha_k w_{i,k}^n. \tag{5.2.6}$$

*If there are more than one conservation laws, the index $k$ corresponding to one of them is selected in* (5.2.4)*.*

### 5.2.2 Choice of $\sigma_i^n$, $W_{i,l}^n$, $W_{i,r}^n$

Two different strategies are considered here, the first one is based on the exact solutions of the Riemann problems and the second one on a Roe linearization:

- **First strategy:** If the solutions of the Riemann problems are explicitly known, in a marked cell $\sigma_i^n$, $W_{i,l}^n$, $W_{i,r}^n$ can be chosen as the speed, the left, and the right states of (one of the) discontinuous waves appearing in the solution of the Riemann problem with initial data $W_{i-1}^n$, $W_{i+1}^n$.

- **Second strategy:** If a Roe matrix is available, in a marked cell $\sigma_i^n$, $W_{i,l}^n$, $W_{i,r}^n$ can be chosen as the speed, the left, and the right states of one of the discontinuities appearing in the solution of the linearized Riemann problem with initial data $W_{i-1}^n$, $W_{i+1}^n$. More explicitly, an index $k^*$ has to be selected and then

$$\sigma_i^n = \lambda_{k^*}(W_{i-1}^n, W_{i+1}^n),$$

$$W_{i,l}^n = W_{i-1}^n + \sum_{k=1}^{k^*-1} \alpha_k R_k(W_{i-1}^n, W_{i+1}^n),$$

$$W_{i,r}^n = W_{i,l}^n + \alpha_{k^*} R_{k^*}(W_{i-1}^n, W_{i+1}^n),$$

  where $\alpha_k$, $k = 1, \dots, N$ represent the coordinates of $W_{i+1}^n - W_{i-1}^n$ on the basis of eigenvectors of $\mathcal{A}_\Phi(W_{i-1}^n, W_{i+1}^n)$, i.e.

$$W_{i+1}^n - W_{i-1}^n = \sum_{k=1}^{N} \alpha_k R_k(W_{i-1}^n, W_{i+1}^n).$$

It can be easily checked that both strategies satisfy (5.2.5) if the solution of the Riemann problem with initial data $W_{i-1}^n$, $W_{i+1}^n$ consist of only one discontinuous wave. These two strategies can be easily extended to any approximate Riemann solver.

### 5.2.3 Time step

The time step $\Delta t_n$ is chosen as follows:

$$\Delta t_n = \min(\Delta t_n^c, \Delta t_n^r), \tag{5.2.7}$$

where

$$\Delta t_n^c = CFL \min\left(\frac{\Delta x}{\max_{i,l} |\lambda_{i,l}|}\right), \tag{5.2.8}$$

where $CFL \in (0,1)$ is the stability parameter and $\lambda_{i,l}, \dots, \lambda_{i,N}$ represent the eigenvalues of $\mathcal{A}(W_i^n)$; and

$$\Delta t_n^r = \min_{i \in \mathcal{M}_n} \begin{cases} \dfrac{1-d_i^n}{|\sigma_i^n|}\Delta x, & if \quad \sigma_i^n > 0, \\ \\ \dfrac{d_i^n}{|\sigma_i^n|}\Delta x, & if \quad \sigma_i^n < 0. \end{cases} \tag{5.2.9}$$

Observe that, besides the stability requirement, this choice of time step ensures that no discontinuous reconstruction leaves a marked cell.

### 5.2.4 Fully discrete method

Once the time step is chosen, (5.2.3) is integrated in the interval $[t^n, t^{n+1}]$, with $t^{n+1} = t^n + \Delta t_n$, to obtain:

$$W_i^{n+1} = W_i^n - \frac{1}{\Delta x}\int_{t^n}^{t^{n+1}} \left( \mathcal{D}^-_{i+\frac{1}{2}}(t) + \mathcal{D}^+_{i-\frac{1}{2}}(t) + \fint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^n(x,t))\partial_x \mathbb{P}_i^n(x,t) dx \right) dt,$$

and the mid-point rule is used to approximate the integrals in time:

$$W_i^{n+1} = W_i^n - \frac{\Delta t_n}{\Delta x}\left( \mathcal{D}^-_{i+\frac{1}{2}}(t^{n+\frac{1}{2}}) + \mathcal{D}^+_{i-\frac{1}{2}}(t^{n+\frac{1}{2}}) + \fint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^n(x,t^{n+1/2}))\partial_x \mathbb{P}_i^n(x,t^{n+1/2})\, dx \right). \tag{5.2.10}$$

The computation of the dashed integral in this expression depends on the cell:

1. If $i-1, i, i+1 \notin \mathcal{M}_n$ the mid-point rule is used again to approximate the integral:

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^n(x,t^{n+1/2}))\partial_x \mathbb{P}_i^n(x,t^{n+1/2})\, dx \approx \Delta x \mathcal{A}(W_i^{n+\frac{1}{2}})\widetilde{\partial_x W}_i^n, \tag{5.2.11}$$

where

$$W_i^{n+\frac{1}{2}} = \mathbb{P}_i^n(x_i, t^{n+\frac{1}{2}}) = W_i^n - \frac{\Delta t}{2}\mathcal{A}(W_i^n)\widetilde{\partial_x W}_i^n.$$

2. If $i \in \mathcal{M}_n$, taking into account the definition of the dashed integrals (5.1.2), one has:

$$\fint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^n(x,t^{n+1/2}))\partial_x \mathbb{P}_i^n(x,t^{n+1/2})\, dx = \int_0^1 \mathcal{A}(\Phi(s; W_{i,l}^n, W_{i,r}^n))\partial_s \Phi(s; W_{i,l}^n, W_{i,r}^n)\, ds. \tag{5.2.12}$$

Observe that, if $W_{i,1}^n$ and $W_{i,r}^n$ can be linked by a shock whose speed is $\sigma_i^n$, then the generalized Rankine-Hugoniot condition (5.1.3) leads to

$$\fint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^n(x,t^{n+1/2}))\partial_x \mathbb{P}_i^n(x,t^{n+1/2})\, dx = \sigma_j^n \left(W_{i,r}^n - W_{i,l}^n\right). \tag{5.2.13}$$

3. If $i \notin \mathcal{M}_n$ but $i-1 \in \mathcal{M}_n$ or $i+1 \in \mathcal{M}_n$ then

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{A}(\mathbb{P}_i^n(x, t^{n+1/2}))\partial_x \mathbb{P}_i^n(x, t^{n+1/2})\, dx = 0. \tag{5.2.14}$$

The final expression of the fully discrete numerical method is as follows:

$$W_i^{n+1} = W_i^n - \frac{\Delta t_n}{\Delta x}\left(\mathcal{D}^-_{i+\frac{1}{2}}(t^{n+\frac{1}{2}}) + \mathcal{D}^+_{i-\frac{1}{2}}(t^{n+\frac{1}{2}}) + \mathcal{D}_i\right), \tag{5.2.15}$$

where

$$\mathcal{D}_i = \begin{cases} \Delta x \mathcal{A}(W_i^{n+\frac{1}{2}})\widetilde{\partial_x W}_i^n & \text{if } i-1, i, i+1 \notin \mathcal{M}_n; \\ \displaystyle\int_0^1 \mathcal{A}(\Phi(s; W_{i,l}^n, W_{i,r}^n))\partial_s \Phi(s; W_{i,l}^n, W_{i,r}^n)\, ds & \text{if } i \in \mathcal{M}_n; \\ 0 & \text{otherwise.} \end{cases} \tag{5.2.16}$$

Observe that the numerical method coincides with the standard MUSCL-Hancock far from discontinuities and with the numerical method introduced in [51] close to discontinuities, with a difference: in this reference, the discontinuities are allowed to leave the marked cells and the contribution to the neighbor cells are then taken into account. While this technique allows one to avoid additional restrictions to the time step, makes more difficult the implementation of the numerical method, nevertheless could be implemented as in [51].

### 5.2.5 Shock-capturing property

Let us prove that isolated shock waves are exactly captured by the scheme and contain no spurious numerical diffusion. Although the proof is essentially the same in [51], it is included for the sake of completeness.

**Theorem 5.2.1.** *Assume that $W_l$ and $W_r$ can be joined by an entropy shock of speed $\sigma$. Then, the numerical method provides an exact numerical solution of the Riemann problem with initial conditions*

$$W(x,0) = \begin{cases} W_l & \text{if } x < 0, \\ W_r & \text{otherwise,} \end{cases}$$

*in the sense that*

$$W_i^n = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} W(x, t^n)\, dx, \quad \forall i, n \tag{5.2.17}$$

*where $W(x,t)$ is the exact solution.*

*Proof.* Let us suppose that $0 \in I_{i^*}$ and $0 = x_{i^*-1/2} + d\Delta x$, with $0 \leq d \leq 1$. Then the initial cell averages are:

$$W_i^0 = \begin{cases} W_l & \text{if } i < i^*; \\ dW_l + (1-d)W_r & \text{if } i = i^*; \\ W_r & \text{otherwise.} \end{cases}$$

If $0 < d < 1$ the only marked cell at time $t^0 = 0$ is $I_{i^*}$, i.e. $\mathcal{M}_0 = \{i^*\}$. The only non-constant reconstruction is then $\mathbb{P}_0^0$ and the equalities that

$$W_i^1 = W_i^0 = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} W(x, t^1)\, dx, \quad \forall i \neq i^*$$

can be easily deduced from the definition of the numerical method.

Let us compute $W_{i^*}^1$. Observe that, in order to have (5.2.4), necessarily $d_0^0 = d$. Therefore, since $W_l$ and $W_r$ can be linked by an admissible discontinuity of speed $\sigma$, using (5.2.5) one has:

$$\mathbb{P}_0^0(x,t) = \begin{cases} W_l & \text{if } x \leq \sigma t, \\ W_r & \text{otherwise.} \end{cases}$$

Observe that $\mathbb{P}_0^0$ coincides with the exact solution. We have now:

$$\begin{aligned} W_{i^*}^1 &= W_{i^*}^0 - \frac{\Delta t_0}{\Delta x}\left(\mathcal{D}_{\frac{1}{2}}^-(t^{\frac{1}{2}}) + \mathcal{D}_{-\frac{1}{2}}^+(t^{\frac{1}{2}}) + \mathcal{D}_0\right) \\ &= W_{i^*}^0 - \frac{\Delta t_0}{\Delta x}\mathcal{D}_0 \\ &= W_{i^*}^0 - \frac{\Delta t_0}{\Delta x}\sigma(W_r - W_l) \\ &= \left(d + \frac{\sigma\Delta t}{\Delta x}\right) W_l + \left(1 - d - \frac{\sigma\Delta t}{\Delta x}\right) W_r, \end{aligned}$$

where it has been used that

$$W_{i-1+2}^-(t^{1/2}) = W_{i-1+2}^+(t^{1/2}) = W_l,$$

$$W_{i+1+2}^-(t^{1/2}) = W_{i+1+2}^+(t^{1/2}) = W_r,$$

so that

$$\mathcal{D}_{-\frac{1}{2}}^+(t^{\frac{1}{2}}) = \mathcal{D}_{\frac{1}{2}}^-(t^{\frac{1}{2}}) = 0.$$

On the other hand, due to the time step restrictions one has

$$x_{i^*-1/2} \leq x_{i^*-1/2} + d\Delta x + \sigma\Delta t = \sigma\Delta t \leq x_{i^*+1/2}.$$

Thus, it can be easily checked that:

$$\frac{1}{\Delta x}\int_{x_{i^*-1/2}}^{x_{i^*+1/2}} W(x,t^1)\,dx = W_{i^*}^1,$$

and (5.2.17) has been proved for $n=1$.

If $d=1$ (resp. $d=0$) the only marked cell is $I_{i+1}$ (resp. $I_{i-1}$) and the proof is similar. The proof of the equality (5.2.17) for $n \geq 2$ is similar to the case $n=1$.

□

## 5.3 Numerical tests

The following methods are applied here to three nonconservative systems:

- O1_noDisRec: standard first-order path-conservative Roe (1.2.23) or Godunov (1.2.13)-(1.2.14) (if it is indicated between parentheses) methods.
- O1_DisRec: first-order path-conservative method with discontinuous reconstruction;
- O2_noDisRec: second-order extension standard of the first order path-conservative method based on the MUSCL-Hancock reconstruction;
- O2_DisRec: second-order path-conservative method that combines MUSCL-Hancock and discontinuous reconstruction.

### 5.3.1 Coupled Burgers system

Let us first consider the toy system

$$\left\{\begin{array}{rcl} \partial_t u + \partial_x\left(\dfrac{u^2}{2}\right) + u\partial_x v &=& 0,\\ \partial_t v + \partial_x\left(\dfrac{v^2}{2}\right) + v\partial_x u &=& 0,\end{array}\right. \qquad (x,t)\in\mathbb{R}\times\mathbb{R}^+, \tag{5.3.1}$$

introduced in [44], where $W=(u,v)^T$ belongs to the state space $\Omega=\{W\in\mathbb{R}^2, u+v>0\}$. This system can be written in the form (5.1.1) with

$$\mathcal{A}(W)=\begin{bmatrix} u & u\\ v & v\end{bmatrix}.$$

The system is strictly hyperbolic in $\Omega$ with eigenvalues

$$\lambda_1(W)=0, \quad \lambda_2(W)=u+v,$$

whose characteristic fields, given by the eigenvectors

$$R_1(W) = [1, -1]^T, \quad R_2(W) = [u, v]^T,$$

are respectively linearly degenerate and genuinely nonlinear.

The sum $u + v$ satisfies the standard Burgers equation

$$\partial_t(u+v) + \partial_x\left(\frac{1}{2}(u+v)^2\right) = 0,$$

and thus is conserved.

Once the family of paths has been chosen, the simple waves of this system are:

- Stationary contact discontinuities linking states $W_l$, $W_r$ such that

$$u_l + v_l = u_r + v_r.$$

- Rarefactions waves joining states $W_l$, $W_r$ such that

$$u_l + v_l < u_r + v_r, \quad \frac{u_l}{v_l} = \frac{u_r}{v_r}.$$

- Shock waves joining states $W_l$ and $W_r$ such that

$$u_l + v_l > u_r + v_r$$

that satisfy the jump condition:

$$\begin{aligned}
\sigma[u] &= \left[\frac{u^2}{2}\right] + \int_0^1 \phi_u(s; W_l, W_r)\partial_s\phi_v(s; W_l, W_r)\, ds, \\
\sigma[v] &= \left[\frac{v^2}{2}\right] + \int_0^1 \phi_v(s; W_l, W_r)\partial_s\phi_u(s; W_l, W_r)\, ds.
\end{aligned}$$

As usual, for any variable $\phi$, $[\phi]$ stands for the jump on the variable $\phi_r - \phi_l$. Remark that this leads, independently of the choice of the family of paths, to

$$\sigma = \frac{u_l + v_l + u_r + v_r}{2}.$$

If, for instance, the family of straight segments is chosen

$$\phi_u(s; W_l, W_r) = u_l + s(u_r - u_l); \quad \phi_v(s; W_l, W_r) = v_l + s(v_r - v_l), \tag{5.3.2}$$

the jump conditions reduce to:

$$\begin{aligned}
\sigma[u] &= \left(\frac{u_l + u_r}{2}\right)(u_r - u_l + v_r - v_l), \\
\sigma[v] &= \left(\frac{v_l + v_r}{2}\right)(u_r - u_l + v_r - v_l),
\end{aligned}$$

and two states can be joined by an admissible shock if

$$u_l + v_l > u_r + v_r, \quad \frac{u_l}{v_l} = \frac{u_r}{v_r}.$$

A Roe matrix is given in this case by:

$$\mathcal{A}(W_l, W_r) = \begin{bmatrix} 0.5(u_l + u_r) & 0.5(u_l + u_r) \\ 0.5(v_l + v_r) & 0.5(v_l + v_r) \end{bmatrix}. \tag{5.3.3}$$

As it will be seen in Test 1, the corresponding Roe method captures correctly the discontinuities of the weak solutions, what puts on evidence that being path-conservative is not in itself a barrier to the convergence to the right solutions. Nevertheless this is not true for other choices of family of paths. Let us consider, for instance, the family of paths given by the viscous profiles of the regularized system:

$$\begin{cases} \partial_t u + \partial_x \left(\dfrac{u^2}{2}\right) + u\partial_x v &= \epsilon u_{xx}, \\ \partial_t v + \partial_x \left(\dfrac{v^2}{2}\right) + v\partial_x u &= \epsilon v_{xx}, \end{cases} \quad (x,t) \in \mathbb{R} \times \mathbb{R}^+, \tag{5.3.4}$$

introduced in [15]: see this reference for the expression of the corresponding family of paths.

It will be seen in Test 2 that Godunov method does not converge to the right weak solutions. In [51] the in-cell discontinuous reconstruction technique has been used to correct this issue with good results. To apply this technique, a cell is marked if

$$u_{i-1}^n + v_{i-1}^n > u_{i+1}^n + v_{i+1}^n.$$

Strategy 1 (based on the exact solutions of the Riemann problems) is followed here to select the discontinuous reconstruction (see Subsection 5.2.2). More precisely, in a marked cell the left and right states are chosen as follows:

$$\sigma_i^n = \frac{1}{2}(u_{i-1}^n + v_{i-1}^n + u_{i+1}^n + v_{i+1}^n), \quad W_{i,l}^n = W^*(W_{i-1}^n, W_{i+1}^n), \quad W_{i,r}^n = W_{i+1}^n,$$

where $W^*(W_{i-1}^n, W_{i+1}^n)$ represents the state at the left of the shock wave appearing in the solution of the Riemann problem. Finally, the conserved variable $u + v$ is chosen to determine $d_i^n$, i.e.

$$d_i^n(u_{i,l}^n + v_{i,l}^n) + (1 - d_i^n)(u_{i,r}^n + v_{i,r}^n) = (u_i^n + v_i^n).$$

This method is extended here to second order following Section 5.2.

The stationary solutions of System (5.3.1) satisfy

$$u + v = constant.$$

In order to preserve these stationary solutions, the MUSCL-Hancock reconstruction is computed as follows:

- First the reconstruction operator is applied to the variable $u$ to obtain $\mathbb{P}^n_{u,i}(x,t)$.
- Then, the reconstruction operator is applied to $u+v$ to obtain $\mathbb{P}^n_{u+v,i}(x,t)$.
- Next, we define
$$\mathbb{P}_{v,i}(x,t) = \mathbb{P}^n_{u+v,i}(x,t) - \mathbb{P}^n_{u,i}(x,t).$$
- Finally, once these reconstructions have been computed, we define
$$\mathbb{P}^n_i(x,t) = \begin{pmatrix} \mathbb{P}_{u,i}(x,t) \\ \mathbb{P}_{v,i}(x,t) \end{pmatrix}.$$

It can be easily chequed that using this definition of $\mathbb{P}^n_i(x,t)$ in (5.2.10) with (5.2.11):

$$D^{\pm}_{i+\frac{1}{2}}(t^{n+\frac{1}{2}}) = 0, \qquad \mathcal{A}(\mathbb{P}^n_i(x,t^{n+1/2}))\partial_x \mathbb{P}^n_i(x,t^{n+1/2}) = 0,$$

what implies the well-balancedness of the method.

### Test 1: Coupled Burgers' equation with straight segment paths

In this test case we consider the definition of weak solution related to the family of straight segments (5.3.2) and the corresponding Roe matrix (5.3.3). Let us consider the following initial condition

$$W_0(x) = (u,v)_0(x) = \begin{cases} (2.0, 2.0) & \text{if } x < 0.5, \\ (1.0, 1.0) & \text{otherwise.} \end{cases}$$

The solution of the Riemann problem in this case consists of a shock wave joining the left and right states.

Figures 5.1 and 5.2 compare the exact solution with the numerical approximations at time $t = 0.1$ obtained with O$p$_noDisRec and O$p$_DisRec, $p = 1, 2$ using a 1000-cell mesh and CFL=0.5: as it can be seen, in spite of the numerical diffusion added by the CFL parameter's choice, the four methods capture correctly the exact solution. The same comparison has been done for a number of different Riemann problems and, in all cases, the numerical solutions converge to the weak solution. As we said before this test puts on evidence that being path-conservative is not in itself a barrier to the convergence to the right solutions.

### Test 2: Isolated shock wave

From now on, the family of paths given by the viscous profiles of the regularized equation (5.3.4) is considered. Let us consider the following initial condition taken from [39]

$$W_0(x) = (u,v)_0(x) = \begin{cases} (7.99, 11.01) & if\, x < 0.5, \\ (0.25, 0.75) & \text{othewise.} \end{cases}$$

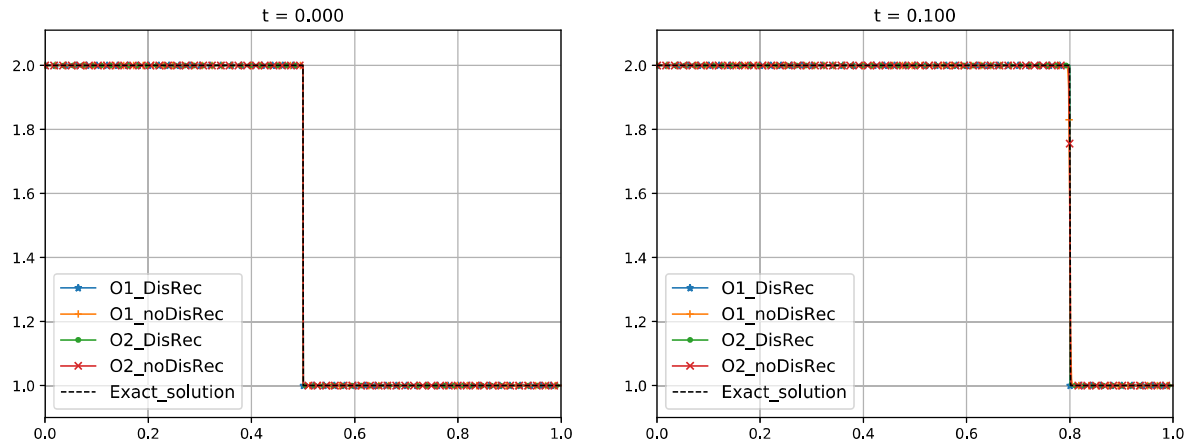



Figure 5.1: Coupled Burgers system. Test 1: variable $u$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.1$ with 1000 cells.

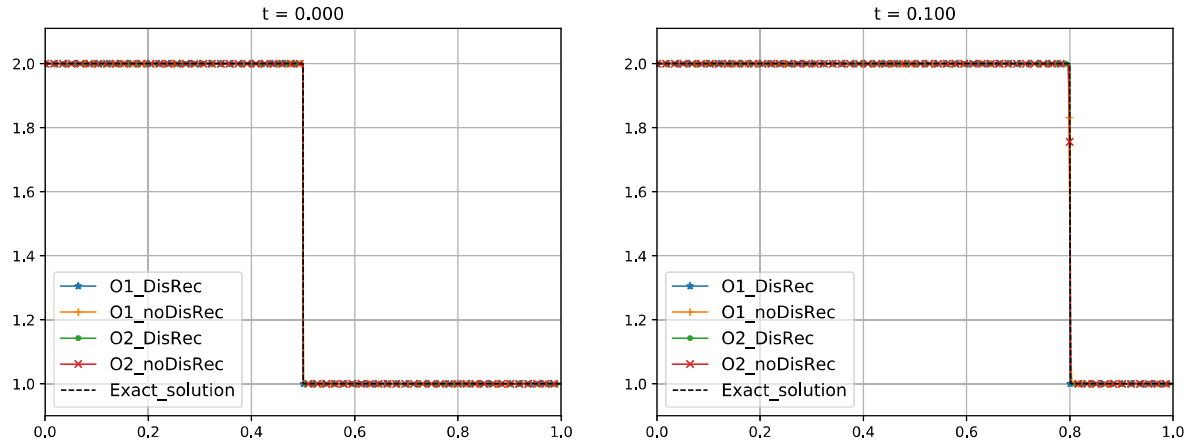



Figure 5.2: Coupled Burgers system. Test 1: variable $v$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.1$ with 1000 cells.

The solution of the Riemann problem consists of a shock wave joining the left and right states.

Figures 5.3 and 5.4 compare the exact solution with the numerical approximations at time $t = 0.03$ obtained with O$p$_noDisRec(Godunov) and O$p$_DisRec(Godunov), $p = 1, 2$ using a 100-cell mesh: as it can be seen Godunov method and its second order extension do not capture the discontinuity properly what is not the case for the methods based on the discontinuous reconstruction. In Figures 5.5 and 5.6 we compare the exact solution

with the numerical approximations obtained with the same methods but using a 1000-cell mesh and again Godunov method and its second order extension are not able to capture the right solution. A CFL = 0.5 has been considered.

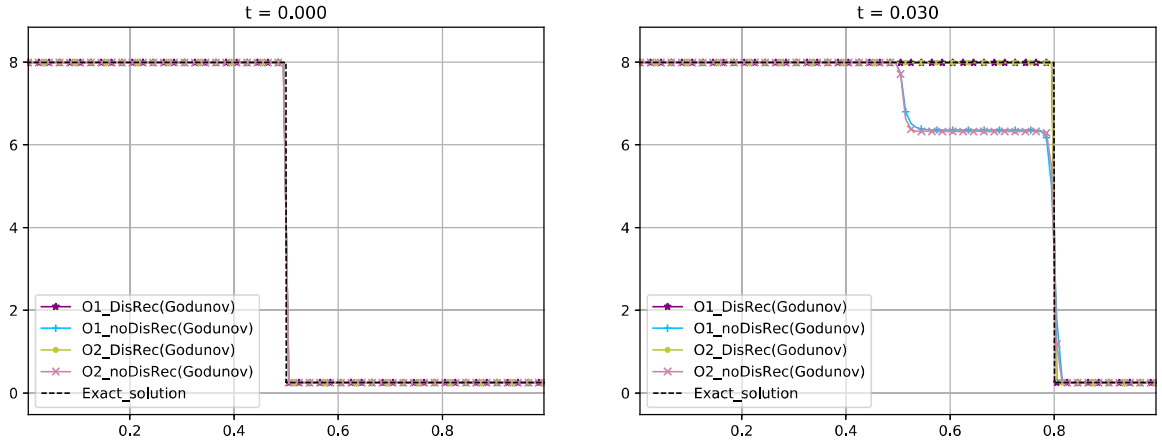



Figure 5.3: Coupled Burgers system. Test 2: variable $u$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.03$ with 100 cells.

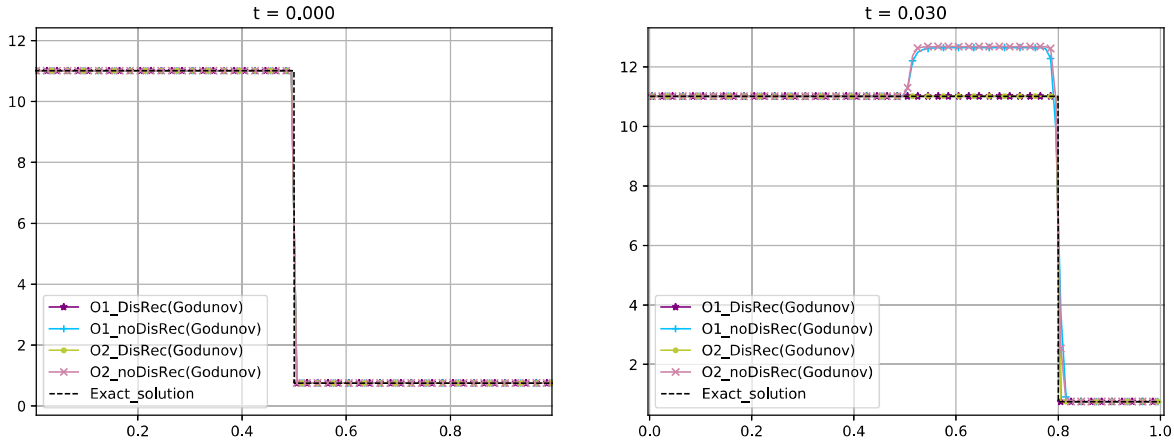



Figure 5.4: Coupled Burgers system. Test 2: variable $v$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.03$ with 100 cells.

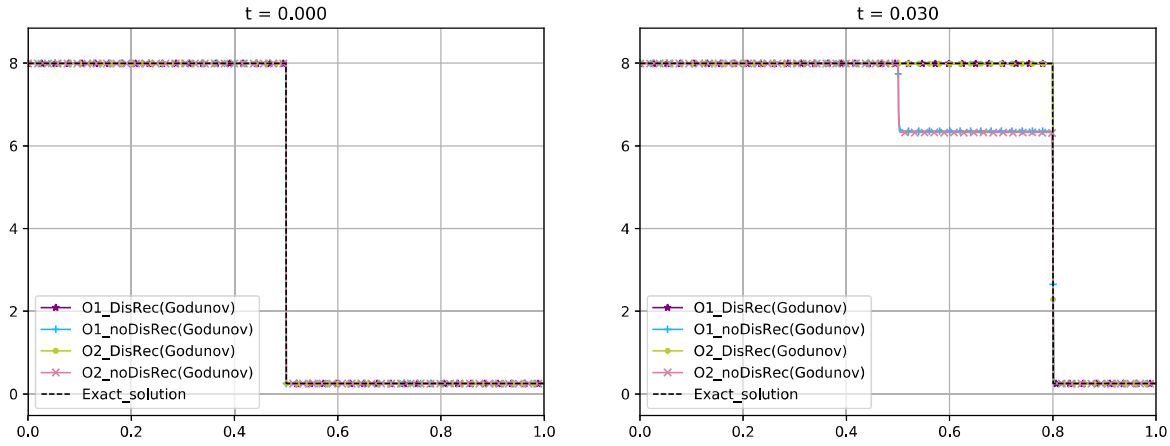



Figure 5.5: Coupled Burgers system. Test 2: variable $u$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.03$ with 1000 cells.

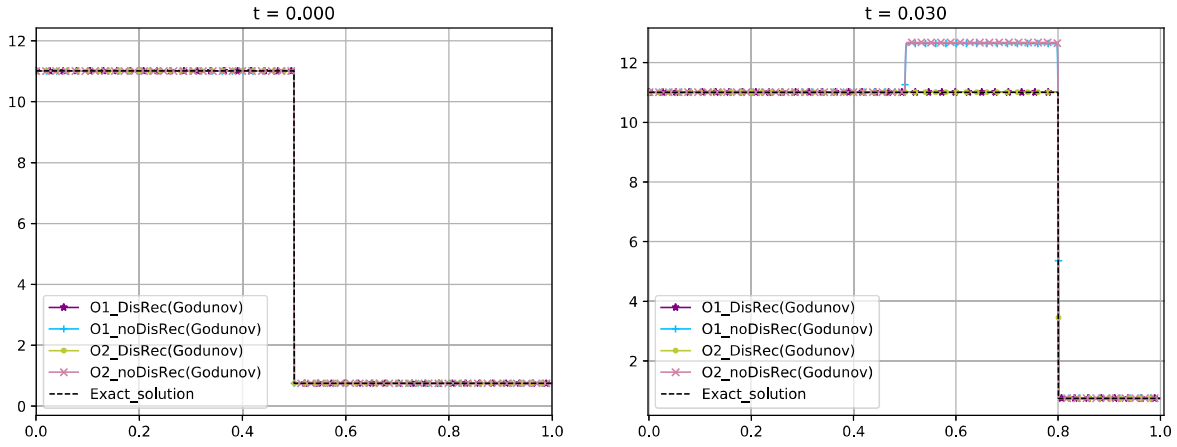



Figure 5.6: Coupled Burgers system. Test 2: variable $v$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.03$ with 1000 cells.

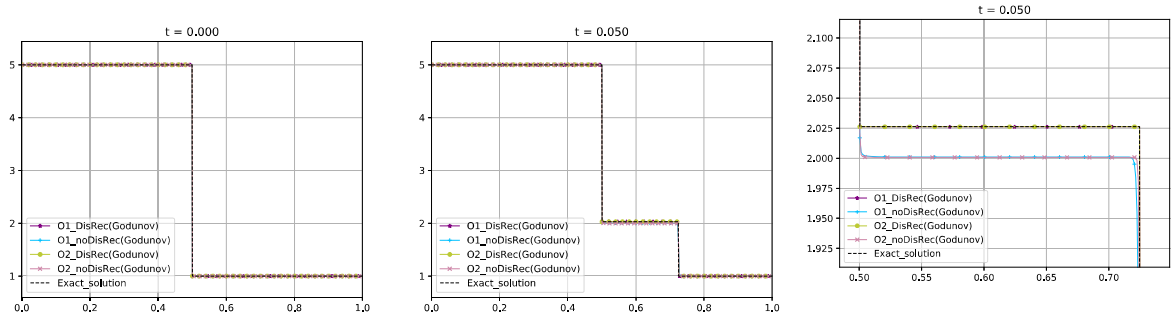
### Test 3: Contact discontinuity + shock wave

We consider now the initial condition

$$W_0(x) = (u, v)_0(x) = \begin{cases} (5, 1) & \text{if } x < 0.5, \\ (1, 2) & \text{otherwise.} \end{cases}$$

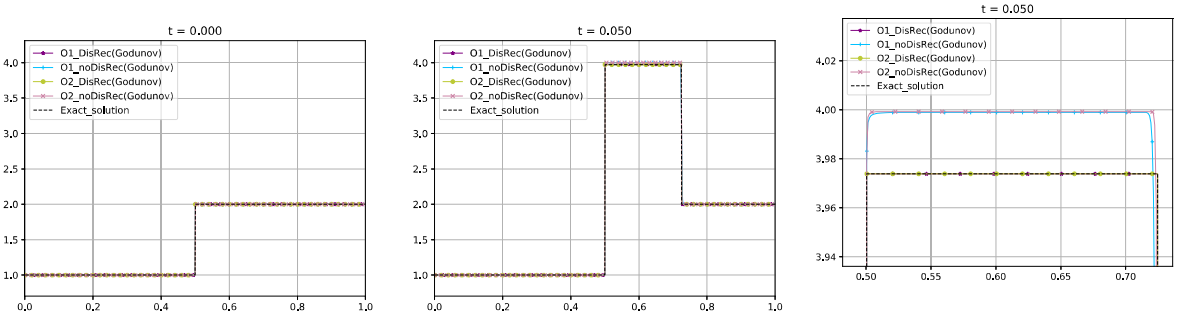
The solution of the corresponding Riemann problems consists of a stationary contact discontinuity followed by a shock. Figures 5.7 and 5.8 show the exact and the numerical

solutions at time $t = 0.05$ using a 1000-cell mesh and CFL $= 0.5$. The conclusions are the same: the in-cell discontinuous reconstruction methods of order 1 and 2 get the exact solution while the standard Godunov methods do not.




Zoom

Figure 5.7: Coupled Burgers system. Test 3: variable $u$. Left: initial condition. Center: exact solution and numerical solutions obtained at time $t = 0.05$ with 1000 cells. Right: zoom.




Zoom

Figure 5.8: Coupled Burgers system. Test 3: variable $u$. Left: initial condition. Center: exact solution and numerical solutions obtained at time $t = 0.05$ with 1000 cells. Right: zoom.

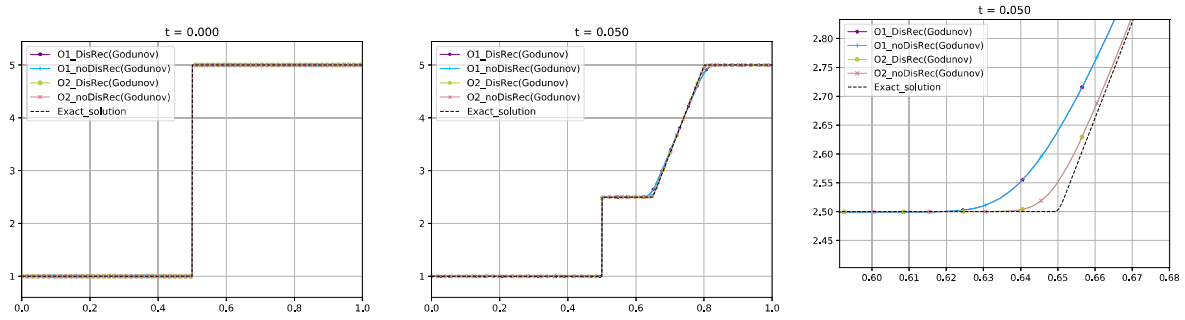
### Test 4: Contact discontinuity + rarefaction

We consider the initial condition

$$W_0(x) = (u, v)_0(x) = \begin{cases} (1, 2) & \text{if } x < 0.5, \\ (5, 1) & \text{otherwise.} \end{cases}$$

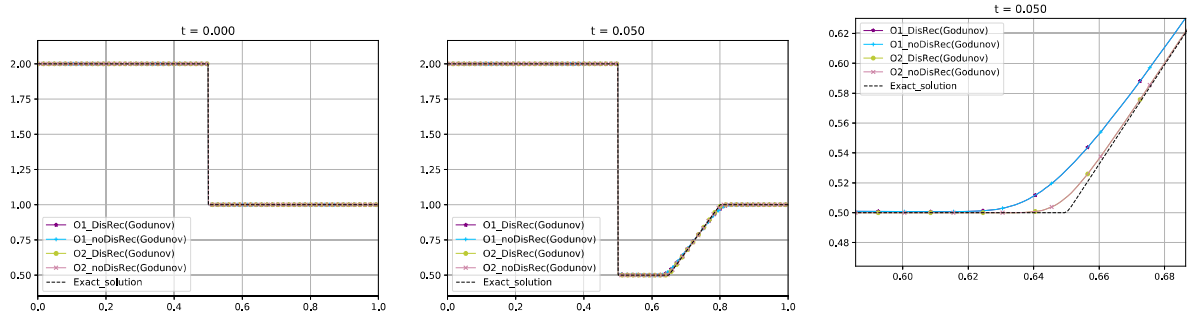
The solution of the corresponding Riemann problem consists of a stationary contact discontinuity followed by a rarefaction.

Figures 5.9 and 5.10 show the exact and the numerical solutions at time $t = 0.05$ using a 1000-cell mesh. In this case all the methods converge to the exact solution but the second order one captures better the solution, as expected.




Zoom

Figure 5.9: Coupled Burgers system. Test 4: variable $u$. Left: initial condition. Center: exact solution and numerical solutions obtained at time $t = 0.05$ with 1000 cells. Right: zoom.




Zoom

Figure 5.10: Coupled Burgers system. Test 4: variable $v$. Left: initial condition. Center: exact solution and numerical solutions obtained at time $t = 0.05$ with 1000 cells. Right: zoom.

### Test 5: Stationary solution

We consider finally the initial condition

$$W_0(x) = (u, v)_0(x) = (\sin(x), 1 - \sin(x)), \tag{5.3.5}$$

that is a stationary solution of the system (5.3.1). We show in Figure 5.11 the numerical solution obtained with the first and second order discontinuous in-cell reconstruction using a 1000-mesh. The results in Figure 5.12 and Table 5.1 show that the both schemes are well-balanced.

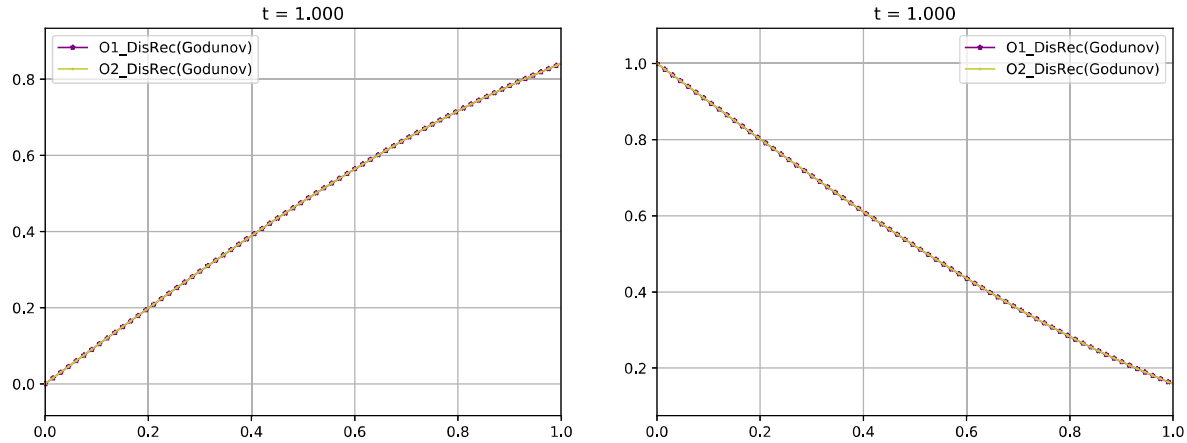



Figure 5.11: Coupled Burgers system. Test 5: numerical solution of (5.3.5) at time $t = 1.00$ with 1000 cells. Left: variable $u$. Right: variable $v$.

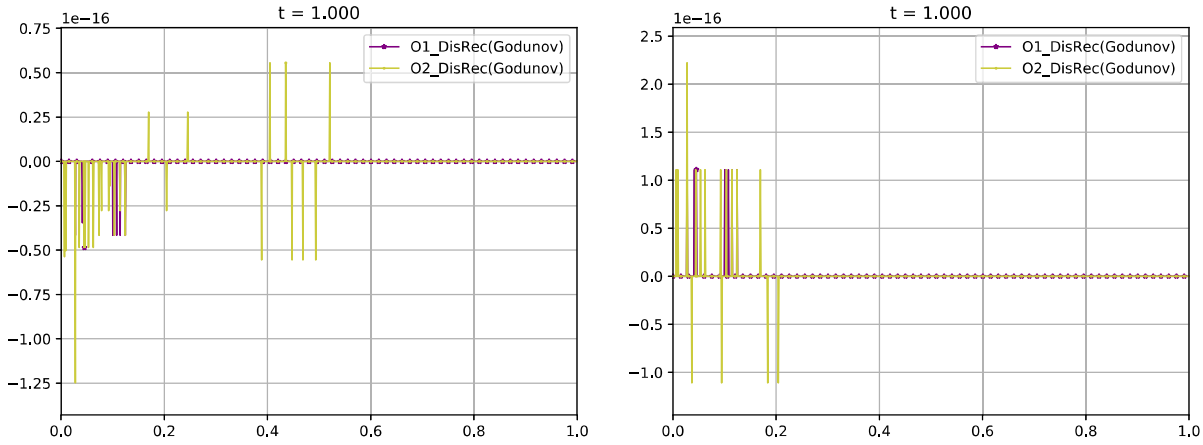



Figure 5.12: Coupled Burgers system. Test 5: difference between the numerical solution at $t = 1.00$ and the stationary solution. Left: variable $u$. Right: variable $v$.

| $\|\|\Delta u\|\|_1$ (1st) | $\|\|\Delta v\|\|_1$ (1st) | $\|\|\Delta u\|\|_1$ (2nd) | $\|\|\Delta v\|\|_1$ (2nd) |
|---|---|---|---|
| 3.40e-19 | 8.88e-19 | 1.17e-18 | 1.78e-18 |

Table 5.1: $L^1$ errors $||\Delta \cdot ||_1$ at time $t = 1$ for the Coupled Burgers model with initial conditions (5.3.5).

**Test 6: Perturbed stationary solution**

We consider finally the initial condition

$$W_0(x) = (u, v)_0(x) = (\sin(x) + 0.2e^{-2000(r-0.5)^2}, 1 - \sin(x)), \tag{5.3.6}$$

that is the stationary solution (5.3.5) with a perturbation in the variable $u$. Figures 5.13 and 5.14 show the numerical solutions obtained at time $t = 0.2$ and $t = 1$ using a 1000-cell mesh together with a reference solution obtained with the first order in-cell discontinuous reconstruction Godunov scheme using a 10000-cell mesh. As it can be seen the second order methods capture better the smooth parts of the solution and the ones with the in-cell reconstruction capture better the shock appearing in the perturbation. Observe that, in this case, the stationary solution (5.3.5) is not restored: a different equilibrium with a stationary bump placed at the initial location of the perturbation is obtained once the waves generated by the perturbation leaves the computational domain.

### 5.3.2 Gas dynamics equations in Lagrangian coordinates

Let us consider the gas dynamics equations in Lagrangian coordinates:

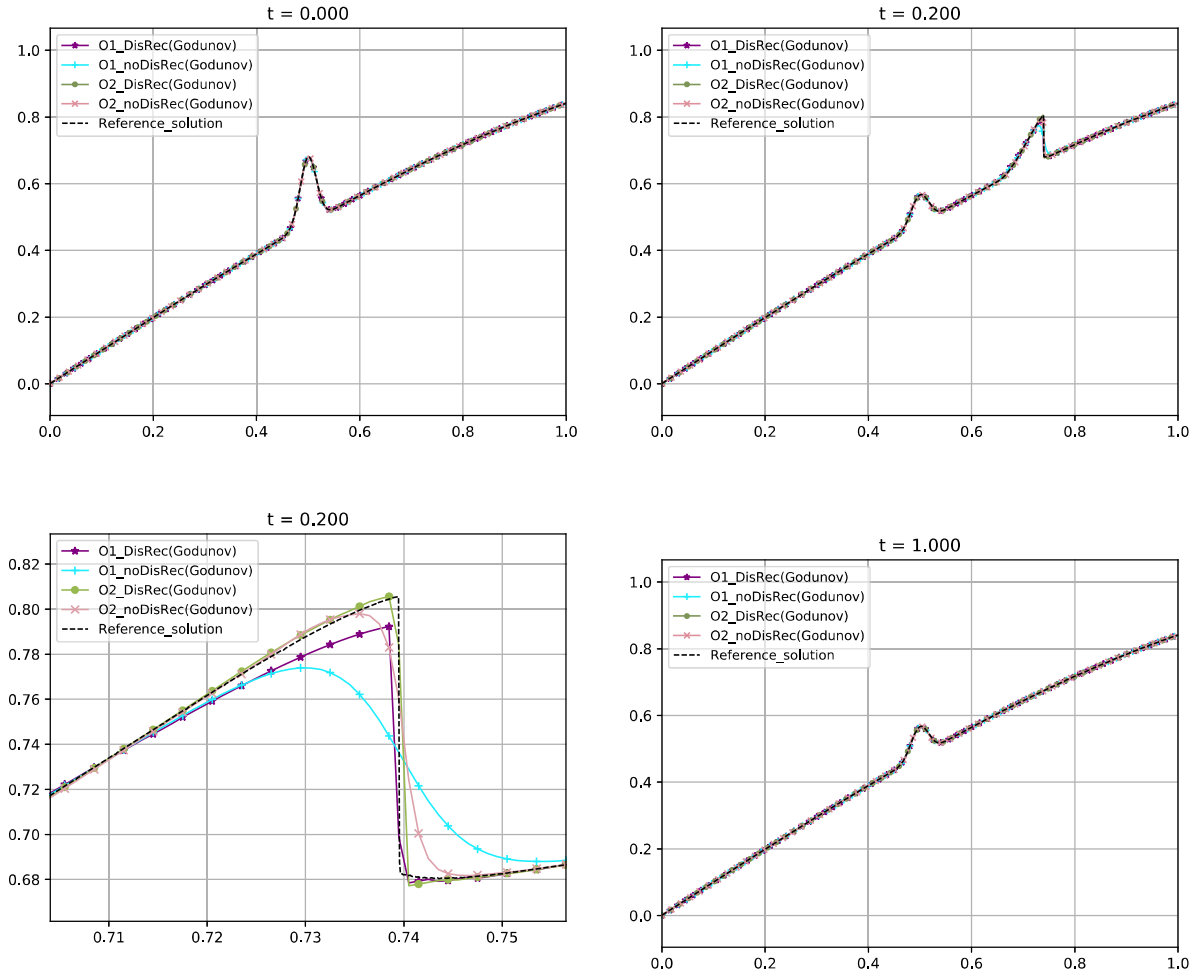
$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x p = 0, \\ \partial_t E + \partial_x (pu) = 0, \end{cases} \tag{5.3.7}$$

where $\tau > 0$ represents the inverse of the density, $u$ is the velocity, $p = p(\tau, e) > 0$ is the pressure, $e$ is the internal energy, and $E = e + u^2/2$ the total energy. For the sake of simplicity, we consider a perfect gas equation of state $p(\tau, e) = (\gamma - 1)e/\tau$ where $\gamma > 1$. System (5.3.7) can be rewritten in nonconservative form as follows

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x p = 0, \\ \partial_t e + p\partial_x u = 0, \end{cases} \tag{5.3.8}$$

that can be written in the form (5.1.1) with

$$W = \begin{pmatrix} \tau \\ u \\ e \end{pmatrix}, \quad \mathcal{A}(W) = \begin{pmatrix} 0 & -1 & 0 \\ -\dfrac{(\gamma-1)e}{\tau^2} & 0 & \dfrac{\gamma-1}{\tau} \\ 0 & \dfrac{(\gamma-1)e}{\tau} & 0 \end{pmatrix}.$$




(a) zoom

Figure 5.13: Coupled Burgers system. Test 6: variable $u$. Top: initial condition (left), reference and numerical solutions obtained at time $t = 0.2$ with 1000 cells (right). Down: zoom of the perturbation area at time $t = 0.2$ (left), reference and numerical solutions obtained at time $t = 1$ (right).

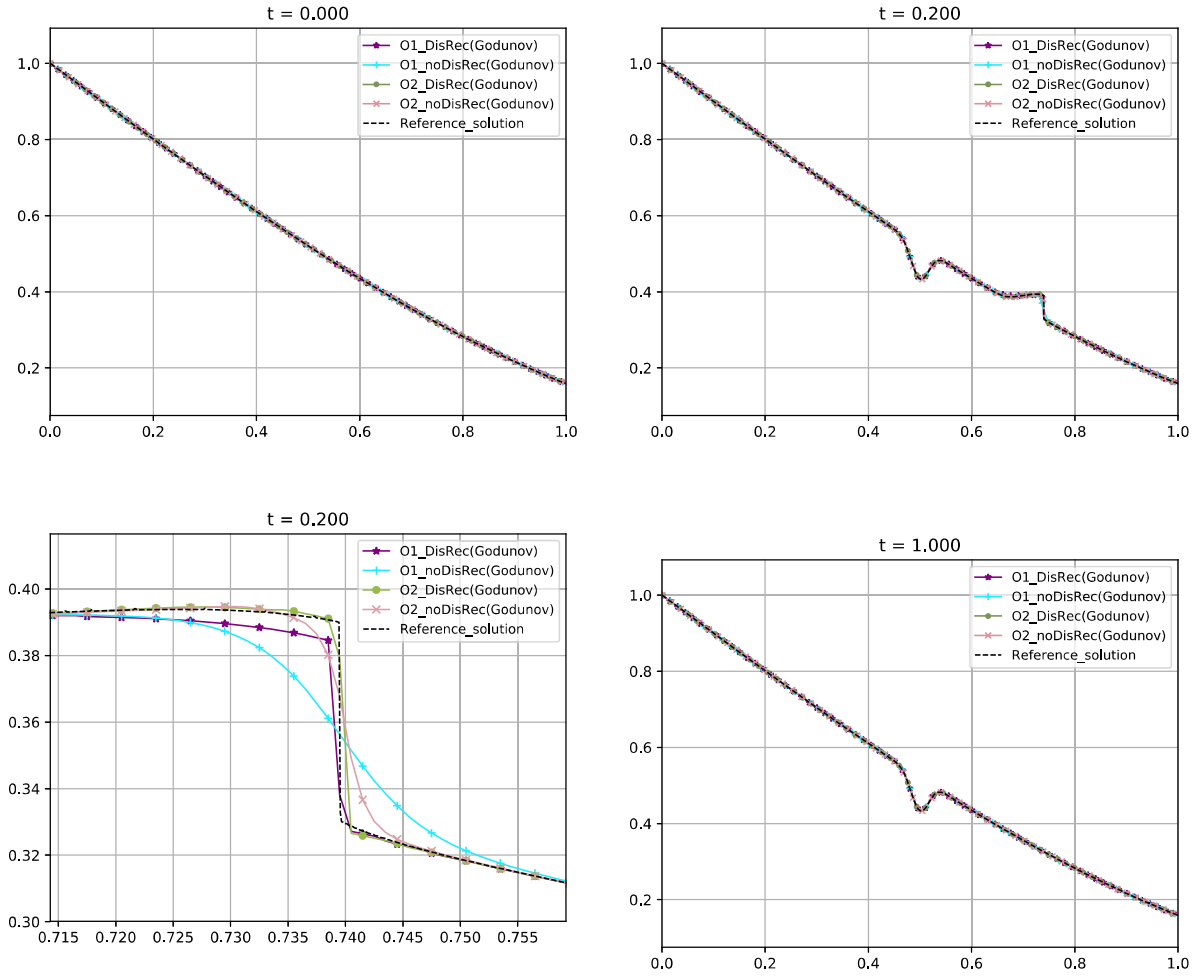
The system is strictly hyperbolic with eigenvalues

$$\lambda_1(W) = -\sqrt{\gamma p/\tau}, \quad \lambda_2(W) = 0, \quad \lambda_3(W) = \sqrt{\gamma p/\tau},$$

whose characteristic fields are given by the eigenvectors

$$R_1(W) = [1, \sqrt{\gamma p/\tau}, -p]^T, \quad R_2(W) = [1, 0, p/(\gamma - 1)], \quad R_3(W) = [1, -\sqrt{\gamma p/\tau}, -p]^T.$$

$R_2(W)$ is linearly degenerate and $R_i(W)$, $i = 1, 3$ genuinely nonlinear: see [85]. On the other hand, the admissible solutions of (5.3.7) are selected by Lax entropy inequalities,




(a) zoom

Figure 5.14: Coupled Burgers system. Test 6: variable $v$. Top: initial condition (left), reference and numerical solutions obtained at time $t = 0.2$ with 1000 cells (right). Down: zoom of the perturbation area at time $t = 0.2$ (left), reference and numerical solutions obtained at time $t = 1$ (right).

which here are equivalent to:

$$\sigma(\tau_+ - \tau_-) \geq 0, \tag{5.3.9}$$

where $\tau_-$ and $\tau_+$ are the values of $\tau$ at both sides of the discontinuity and $\sigma$ its speed of propagation.

Once the family of paths has been chosen, the simple waves of this system are:

- Stationary contact discontinuities linking states $W_l$, $W_r$ such that

$$u_l = u_r.$$

- Rarefactions waves joining states $W_l$, $W_r$ such that

$$u_l < u_r,$$

  and the relations given by the Riemann invariants:

  – 1-rarefaction waves:

$$2\sqrt{\frac{\gamma e_l}{\gamma-1}} + u_l = 2\sqrt{\frac{\gamma e_r}{\gamma-1}} + u_r, \quad \frac{e_l}{\tau_l^{\gamma-1}} = \frac{e_r}{\tau_r^{\gamma-1}}.$$

  – 2-rarefaction waves:

$$2\sqrt{\frac{\gamma e_l}{\gamma-1}} - u_l = 2\sqrt{\frac{\gamma e_r}{\gamma-1}} - u_r, \quad \frac{e_l}{\tau_l^{\gamma-1}} = \frac{e_r}{\tau_r^{\gamma-1}}.$$

- Shock waves joining states $W_l$ and $W_r$ such that

$$u_l > u_r$$

  that satisfy the jump conditions:

$$\begin{aligned} \sigma[\tau] &= -[u], \\ \sigma[u] &= [p], \\ \sigma[e] &= \int_0^1 \phi_p(s; W_l, W_r)\partial_s\phi_u(s; W_l, W_r)\, ds. \end{aligned}$$

If, for instance, the family of straight segments is chosen for the variables $\tau, u, p$

$$\phi_\tau(s; W_l, W_r) = \tau_l + s(\tau_r - \tau_l); \quad \phi_u(s; W_l, W_r) = u_l + s(u_r - u_l); \quad \phi_p(s; W_l, W_r) = p_l + s(p_r - p_l),$$

the jump conditions reduce to:

$$\begin{aligned} \sigma[\tau] &= (u_l - u_r), \\ \sigma[u] &= p_r - p_l, \\ \sigma[e] &= \frac{1}{2}(p_r + p_l)(u_r - u_l). \end{aligned}$$

It can be easily checked that these jump conditions are equivalent to the standard Rankine-Hugoniot conditions corresponding to the conservative formulation (5.3.7) and thus, the weak solutions are the same.

A Roe matrix is given in this case by:

$$\mathcal{A}(W_l, W_r) = \mathcal{A}(\bar{W}), \quad \bar{W}(W_l, W_r) = (\bar{\tau}, \bar{u}, \bar{p}),$$

with

$$\bar{\tau} = \frac{\tau_l + \tau_r}{2}, \quad \bar{u} = \frac{u_l + u_r}{2}, \quad \bar{e} = \frac{\bar{p}\bar{\tau}}{\gamma - 1}, \quad \bar{p} = \frac{p_l + p_r}{2},$$

see [130].

In [51] the in-cell discontinuous reconstruction technique has been used to correct the results that are obtained with the standard Roe path-conservative scheme. To apply this technique, a cell is marked if

$$u_{i-1}^n \geq u_{i+1}^n.$$

The second strategy to select the speed, and the left and right states of the discontinuous reconstruction based on the Roe matrix is used here (see Subsection 5.2.2). More precisely:

- If $u_{i-1}^n = u_{i+1}^n$ then

$$\sigma_i^n = 0, \quad W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_{i+1}^n.$$

- If $u_{i-1}^n > u_{i+1}^n$ and $\tau_{i+1}^n - \tau_{i-1}^n < 0$ then

$$\sigma_i^n = -\sqrt{\gamma \bar{p}/\bar{\tau}}, \quad W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_{i-1}^n + \alpha_1 R_1(W_{i-1}^n, W_{i+1}^n).$$

- If $u_{i-1}^n > u_{i+1}^n$ and $\tau_{i+1}^n - \tau_{i-1}^n > 0$ then

$$\sigma_i^n = \sqrt{\gamma \bar{p}/\bar{\tau}}, \quad W_{i,l}^n = W_{i+1}^n - \alpha_3 R_3(W_{i-1}^n, W_{i+1}^n), \quad W_{i,r}^n = W_{i+1}.$$

Here $\bar{p}$ and $\bar{\tau}$ represent the Roe intermediate values of $p$ and $\tau$, and $\alpha_k$, $k = 1, 2, 3$ the coordinates of $W_{i+1}^n - W_{i-1}^n$ in the basis of eigenvectors of the Roe matrix, i.e. $W_{i+1}^n - W_{i-1}^n = \sum_{k=1}^3 \alpha_k R_k(W_{i-1}^n, W_{i+1}^n)$. This method is extended here to second order by following the procedure described in Section 5.2.

**Test 1: Isolated 1-shock**

Let us consider the following initial condition taken from [51]

$$(\tau, u, p)_0(x) = \begin{cases} (2.09836065573770281, 2.3046638387921279, 1) & \text{if } x < 0.5, \\ (8, 0, 0.1) & \text{otherwise.} \end{cases}$$

The solution of the Riemann problem consists of a 1-shock wave joining the left and right states. Figures 5.15, 5.16, and 5.17 compare the exact solution with the numerical

approximations at time $t = 0.5$ obtained with Roe method, its second order extension based on the standard MUSCL reconstruction, and the first and second order discontinuous in-cell reconstruction schemes using 300-cell mesh and CFL $= 0.5$: as it can be seen Roe methods does not capture the discontinuities properly what is not the case for the two other methods.

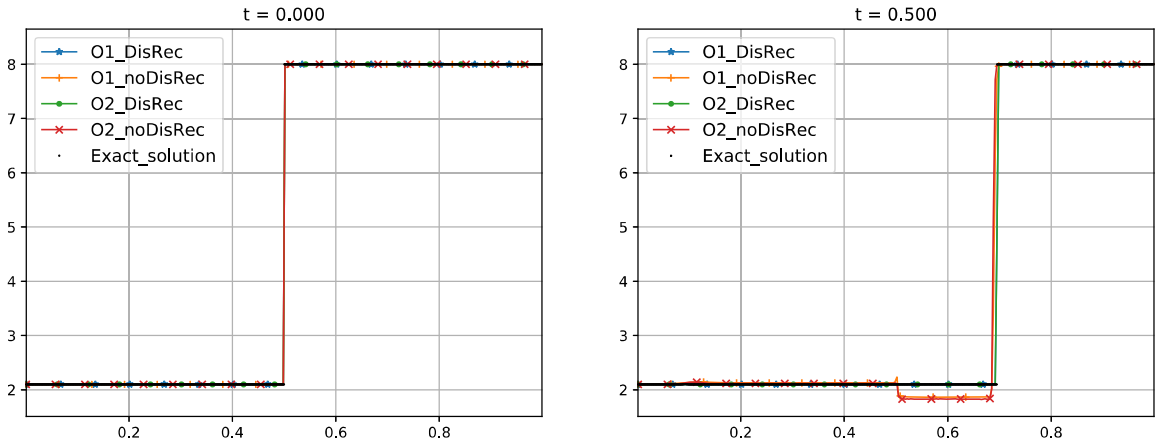



Figure 5.15: Gas dynamics equations in Lagrangian coordinates. Test 1: variable $\tau$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells.

### Test 2: 1-shock + contact discontinuity + 3-shock

Let us consider the following initial condition taken from [51]

$$(\tau, u, p)_0(x) = \begin{cases} (5, 3.323013993227, 0.481481481481) & \text{if } x < 0.5, \\ (8, 0, 0.1) & \text{otherwise.} \end{cases}$$

The solution of the Riemann problem consists of a 1-shock wave with negative speed, a stationary contact discontinuity, and a 3-shock that coincides with the one in the first test problem. Figures 5.18, 5.19, and 5.20 show the numerical solutions at time $t = 0.5$ using a mesh of 300 cells and CFL $= 0.5$ and the conclusions are the same: the in-cell discontinuous reconstruction methods of order 1 and 2 get the exact solution while Roe method and its second-order extension based on the standard MUSCL reconstruction do not.

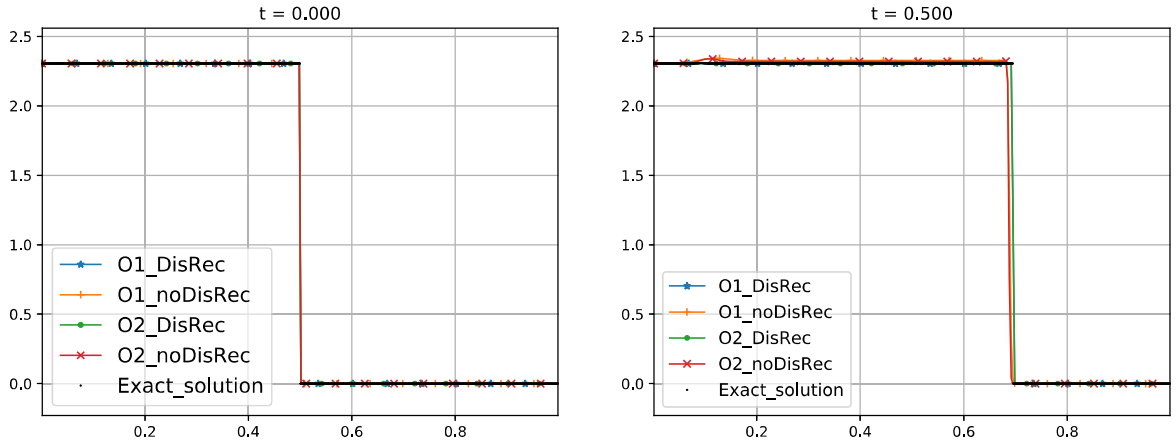



Figure 5.16: Gas dynamics equations in Lagrangian coordinates. Test 1: variable $u$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells.

#### Test 3: 1-rarefaction + contact discontinuity + 3-shock

Let us consider now the initial condition

$$(\tau, u, p)_0(x) = \begin{cases} (2.09836065573770281, 3.323013993227, 1) & \text{if } x < 0.5, \\ (8, 4, 0.1) & \text{otherwise.} \end{cases}$$

The solution of the Riemann problem consists of a 1-rarefaction wave whose head and tail have negative speeds, a stationary contact discontinuity, and a 3-shock with positive speed. Figures 5.21, 5.22, and 5.23 show the numerical solutions at time $t = 0.5$ using a mesh of 300 cells and CFL $= 0.5$. Although all the methods capture correctly the rarefaction wave, second order methods do it better, as expected; concerning the stationary contact discontinuity and the shock wave, only the first and second order in-cell discontinuous reconstruction methods capture the exact solution.

### 5.3.3 Modified shallow water system

Let us consider the modified shallow water system introduced in [42]:

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left( \dfrac{q^2}{h} \right) + qh\partial_x h = 0, \end{cases} \tag{5.3.10}$$

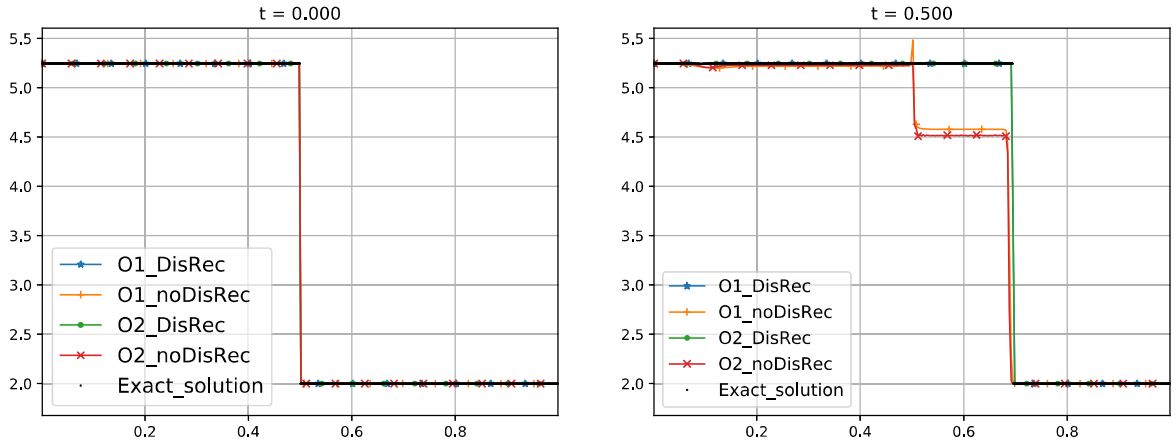



Figure 5.17: Gas dynamics equations in Lagrangian coordinates. Test 1: variable $e$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells.

where $W = (h, q)^t$ belongs to $\Omega = \{W \in \mathbb{R}^2 | \quad 0 < q,\ 0 < h < (16q)^{1/3}\}$. This system can be written in the form (5.1.1) with

$$\mathcal{A}(W) = \begin{bmatrix} 0 & 1 \\ -u^2 + uh^2 & 2u \end{bmatrix},$$

being $u = q/h$. The system is strictly hyperbolic $\Omega$ with eigenvalues

$$\lambda_1(W) = u - h\sqrt{u}, \quad \lambda_2(W) = u + h\sqrt{u},$$

whose characteristic fields, given by the eigenvectors

$$R_1(W) = [1, u - h\sqrt{u}]^T, \quad R_2(W) = [1, u + h\sqrt{u}]^T,$$

are genuinely nonlinear. Once the family of paths has been chosen, the simple waves of this system are:

- 1-rarefaction waves joining states $W_l$, $W_r$ such that

$$h_r < h_l, \quad \sqrt{u_l} + h_l/2 = \sqrt{u_r} + h_r/2,$$

  and 2-rarefaction waves joining states $W_l$, $W_r$ such that

$$h_l < h_r, \quad \sqrt{u_l} - h_l/2 = \sqrt{u_r} - h_r/2.$$

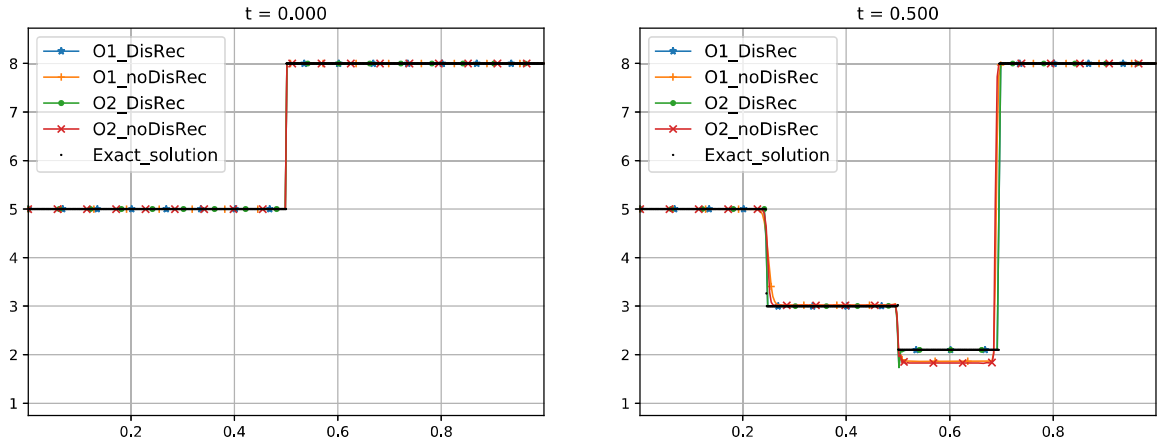



Figure 5.18: Gas dynamics equations in Lagrangian coordinates. Test 2: variable $\tau$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells.

- 1-shock and 2-shock waves joining states $W_l$ and $W_r$ such that $h_l < h_r$ or $h_r < h_l$ respectively, that satisfy the jump conditions:

$$\begin{aligned} \sigma[h] &= [q], \\ \sigma[q] &= \left[\frac{q^2}{h}\right] + \int_0^1 \phi_q(s; W_l, W_r)\phi_h(s; W_l, W_r)\partial_s\phi_h(s; W_l, W_r)\, ds. \end{aligned}$$

If, for instance, the following family of path is chosen:

$$\phi(s; W_l, W_r) = \begin{bmatrix} \phi_h(s; W_l, W_r) \\ \phi_q(s; W_l, W_r) \end{bmatrix} = \begin{cases} \begin{bmatrix} h_l + 2s(h_r - h_l) \\ q_l \end{bmatrix} & \text{if } 0 \le s \le \frac{1}{2}, \\ \begin{bmatrix} h_r \\ q_l + (2s-1)(q_r - q_l) \end{bmatrix} & \text{if } \frac{1}{2} \le s \le 1, \end{cases}$$

the jump conditions reduce to:

$$\begin{aligned} \sigma[h] &= [q], \\ \sigma[q] &= \left[\frac{q^2}{h}\right] + q_l \left[\frac{h^2}{2}\right]. \end{aligned}$$

If this family of paths has been selected and Lax's entropy criterion is used, the simple waves of the system are as follows:

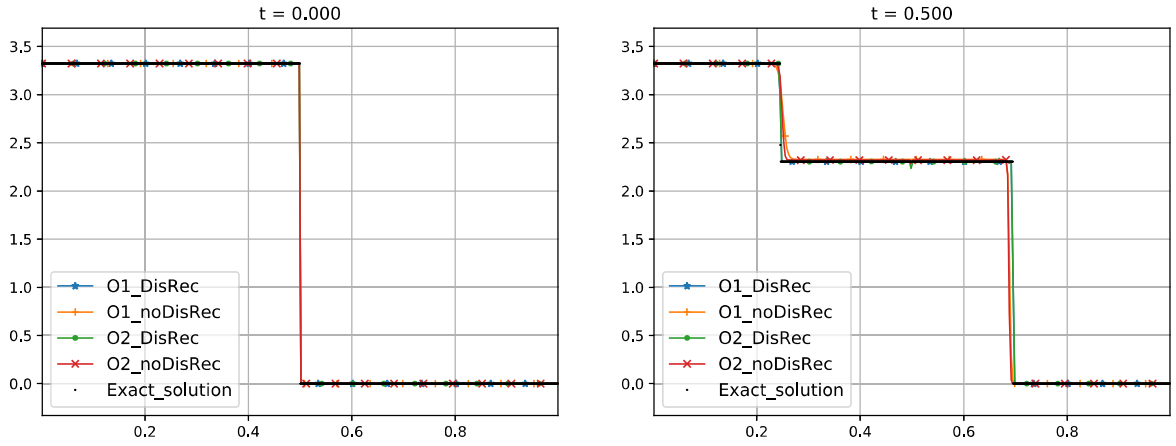



Figure 5.19: Gas dynamics equations in Lagrangian coordinates. Test 2: variable $u$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells.

- Given a left-hand state $W_l$, the 1-shock $\mathcal{S}_1(W_l)$ and the 2-shock $\mathcal{S}_2(W_l)$ curves consisting of all the right-hand states that can be connected with $W_l$ through a 1-shock and a 2-shock wave respectively, are:

$$\mathcal{S}_1(W_l) : u = u_l - \sqrt{\frac{u_l(h+h_l)}{2h}}(h-h_l), \quad h > h_l, \tag{5.3.11}$$

$$\mathcal{S}_2(W_l) : u = u_l - \sqrt{\frac{u_l(h+h_l)}{2h}}(h-h_l), \quad h < h_l. \tag{5.3.12}$$

  Moreover, given two states $W_l$ and $W_r$ connected by a 1-shock wave or a 2-shock wave, the speed of the shock is given by:

$$\sigma_1(W_l, W_r) = u_l - \sqrt{h_r u_l \frac{h_l + h_r}{2}}, \tag{5.3.13}$$

$$\sigma_2(W_l, W_r) = u_l + \sqrt{h_r u_l \frac{h_l + h_r}{2}}, \tag{5.3.14}$$

  respectively.

- Given a left-hand state $W_l$, the 1-rarefaction $\mathcal{R}_1(W_l)$ and the 2-rarefaction $\mathcal{R}_2(W_l)$ consisting of all the right-hand states that can be connected with $W_l$ through a 1-rarefaction and a 2-rarefaction wave, respectively, are:

$$\mathcal{R}_1(W_l) : u = \left(\frac{h_l - h}{2} + \sqrt{u_l}\right)^2, \quad h < h_l, \tag{5.3.15}$$

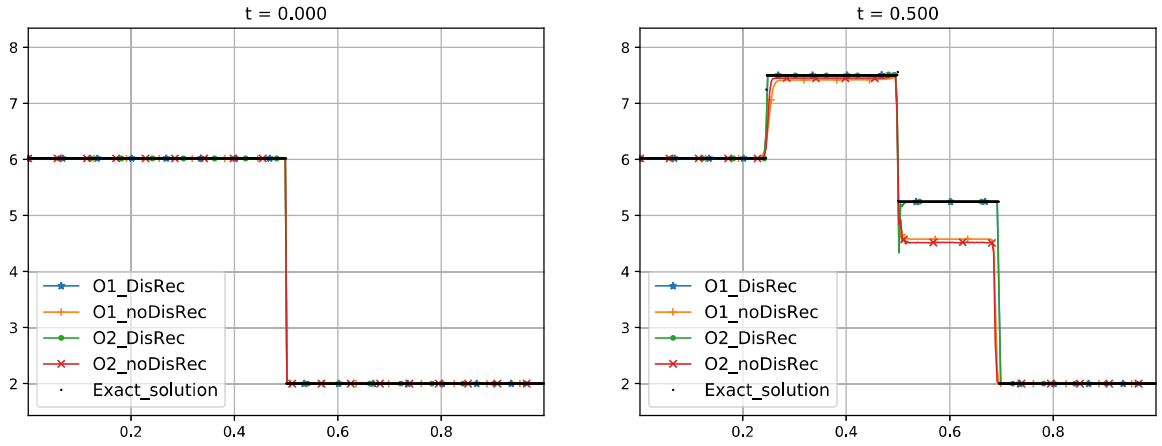



Figure 5.20: Gas dynamics equations in Lagrangian coordinates. Test 2: variable $e$. Left: initial condition. Right: exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells.

$$\mathcal{R}_2(W_l) : u = \left(\frac{h - h_l}{2} + \sqrt{u_l}\right)^2, \quad h > h_l. \tag{5.3.16}$$

The criterion to mark the cells is the following:

1. If $h^n_{i+1} > h^n_{i-1}$ and

$$u^n_{i-1} - \sqrt{\frac{u^n_{i-1}(h^n_{i+1} + h^n_{i-1})}{2h^n_{i+1}}}(h^n_{i+1} - h^n_{i-1}) < u^n_{i+1} < \left(\frac{h^n_{i+1} - h^n_{i-1}}{2} + \sqrt{u^n_{i-1}}\right)^2,$$

the solution of the Riemann problem consists of a 1-shock and a 2-rarefaction waves: the cell is marked.

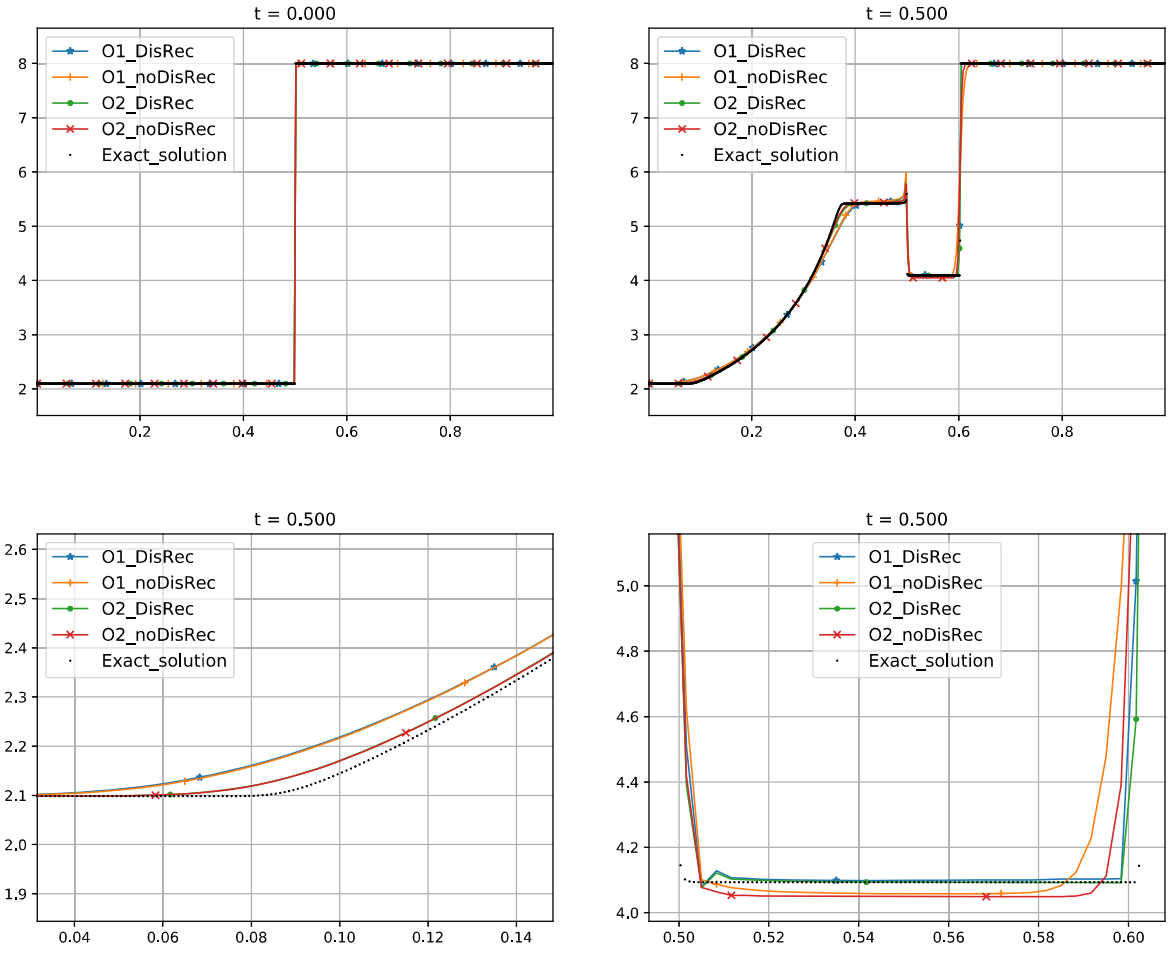
2. If $h^n_{i+1} < h^n_{i-1}$ and

$$u^n_{i-1} + \sqrt{\frac{u^n_{i-1}(h^n_{i+1} + h^n_{i-1})}{2h^n_{i+1}}}(h^n_{i+1} - h^n_{i-1}) < u^n_{i+1} < \left(\frac{h^n_{i-1} - h^n_{i+1}}{2} + \sqrt{u^n_{i-1}}\right)^2,$$

the solution of the Riemann problem consists of a 1-rarefaction and a 2-shock waves: the cell is marked.

3. If $h^n_{i+1} > h^n_{i-1}$ and

$$u^n_{i+1} < u^n_{i-1} - \sqrt{\frac{u^n_{i-1}(h^n_{i+1} + h^n_{i-1})}{2h^n_{i+1}}}(h^n_{i+1} - h^n_{i-1}),$$



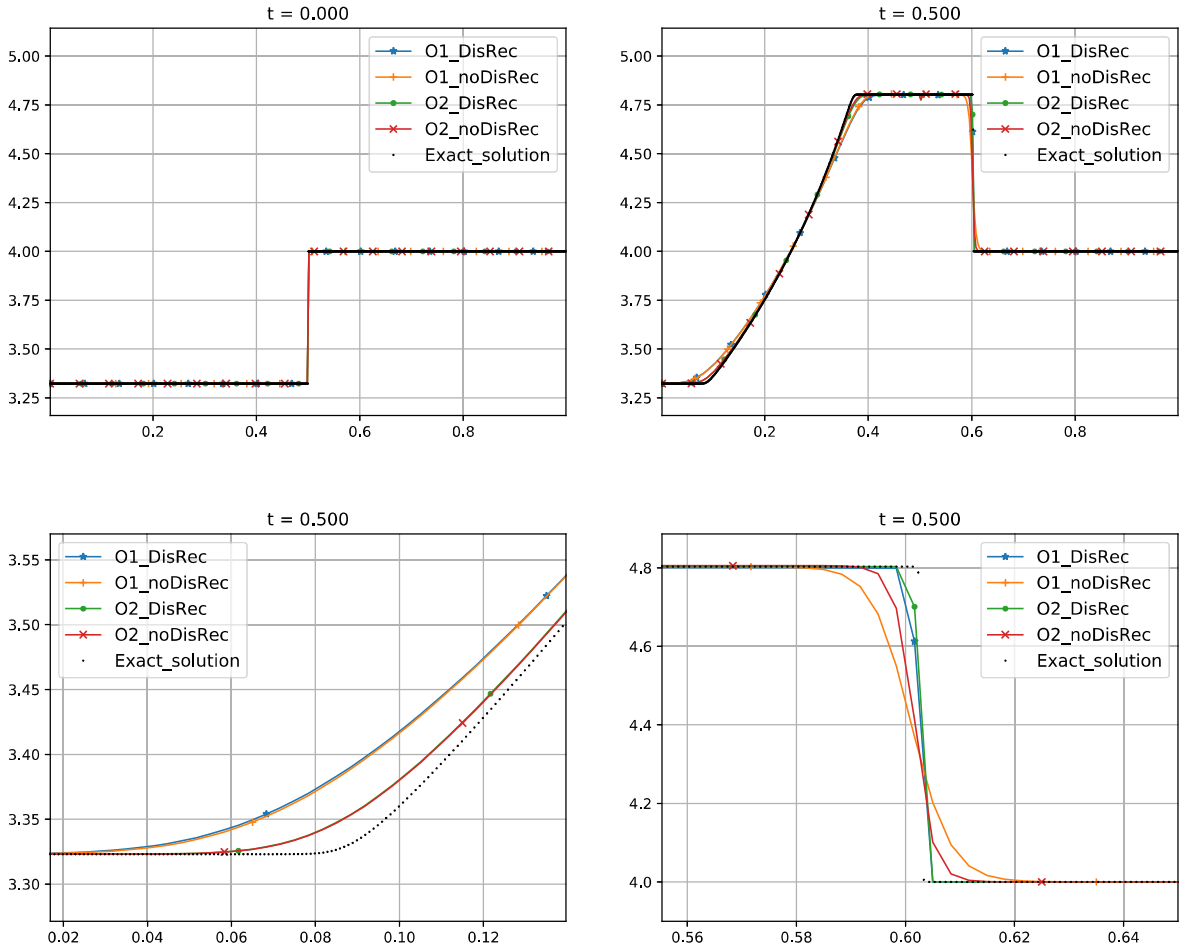

(a) zoom rarefaction (b) zoom shock

Figure 5.21: Gas dynamics equations in Lagrangian coordinates. Test 3: variable $\tau$. Top: initial condition (left), exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells (right). Down: zooms of the rarefaction (left) and the shock waves (right) at time $t = 0.5$.

or $h^n_{i+1} < h^n_{i-1}$ and

$$u^n_{i+1} < u^n_{i-1} + \sqrt{\frac{u^n_{i-1}(h^n_{i+1} + h^n_{i-1})}{2h^n_{i+1}}}(h^n_{i+1} - h^n_{i-1}),$$

the solution of the Riemann problem consists of a 1-shock and a 2-shock waves: the cell is marked.



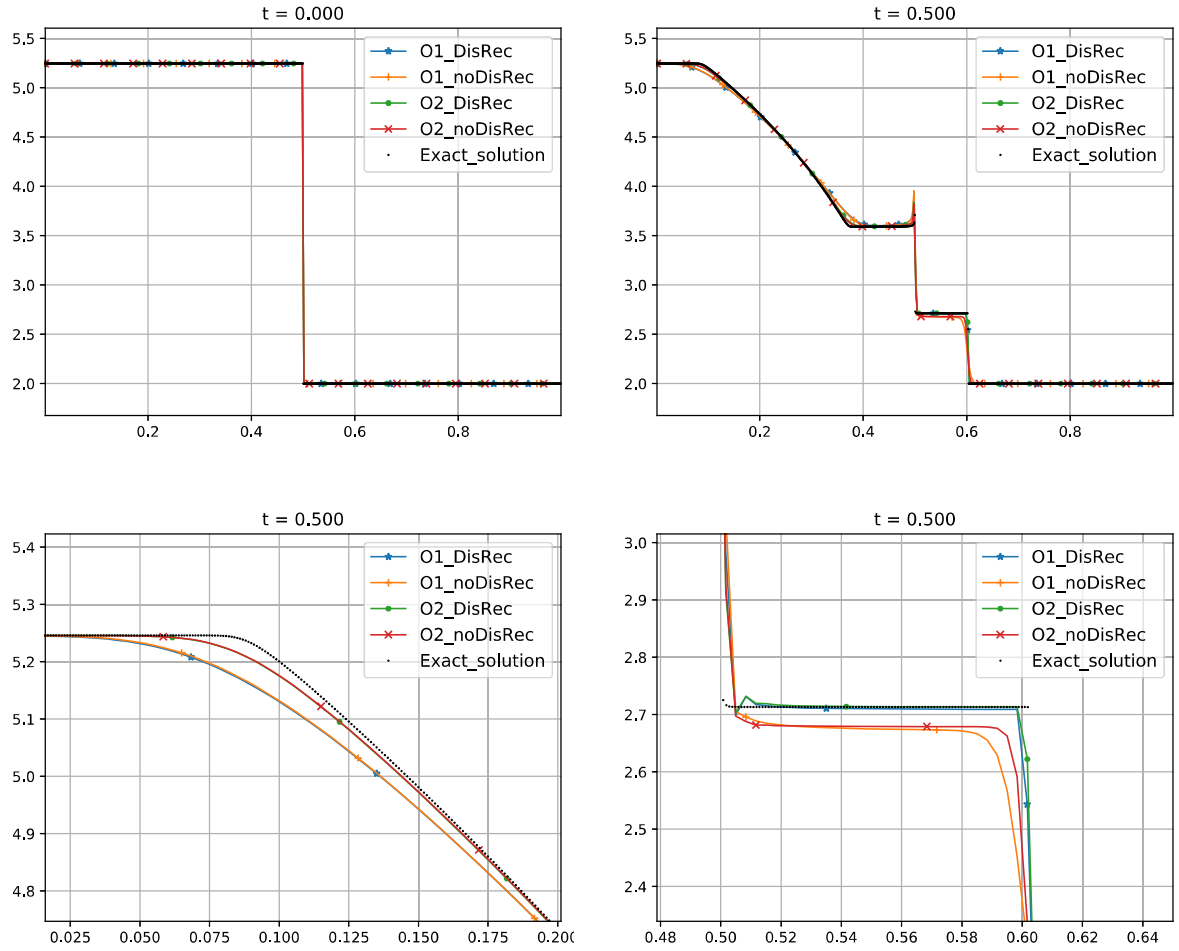

(a) zoom rarefaction (b) zoom shock

Figure 5.22: Gas dynamics equations in Lagrangian coordinates. Test 3: variable $u$. Top: initial condition (left), exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells (right). Down: zooms of the rarefaction (left) and the shock waves (right) at time $t = 0.5$.

4. Otherwise the solution of the Riemann problem consists of two rarefactions and the cell is not marked.

A Roe matrix is given in this case by

$$\mathcal{A}(W_l, W_r) = \begin{bmatrix} 0 & 1 \\ -\overline{u}^2 + q_l\overline{h} & 2\overline{u} \end{bmatrix},$$




(a) zoom rarefaction (b) zoom shock

Figure 5.23: Gas dynamics equations in Lagrangian coordinates. Test 3: variable $e$. Top: initial condition (left), exact solution and numerical solutions obtained at time $t = 0.5$ with 300 cells (right). Down: zooms of the rarefaction (left) and the shock waves (right) at time $t = 0.5$.

where

$$\bar{u} = \frac{\sqrt{h_l}u_l + \sqrt{h_r}u_r}{\sqrt{h_l} + \sqrt{h_r}}, \quad \bar{h} = \frac{h_l + h_r}{2}.$$

The following strategy based on the Roe matrix (see Subsection 5.2.2) is used to select the speed, and the left and right states of the discontinuous reconstruction:

- If the solution of the Riemann problem consists of a 1-shock and a 2-rarefaction

waves (case 1):

$$\sigma_i^n = \bar{u} - h_{i-1}^n\sqrt{\bar{u}}, \quad W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_{i-1}^n + \alpha_1 R_1(W_{i-1}^n, W_{i+1}^n),$$

where $\bar{u}$ is the Roe average of $u_{i-1}^n$ and $u_{i+1}^n$, and $\alpha_k$, $k = 1, 2$ represent the coordinates of $W_{i+1}^n - W_{i-1}^n$ in the basis of eigenvectors of the Roe matrix, i.e. $W_{i+1}^n - W_{i-1}^n = \sum_{k=1}^2 \alpha_k R_k(W_{i-1}^n, W_{i+1}^n)$.

- If the solution of the Riemann problem consists of a 1-rarefaction and a 2-shock waves (case 2):

$$\sigma_i^n = \bar{u} + h_{i-1}^n\sqrt{\bar{u}}, \quad W_{i,l}^n = W_{i+1}^n - \alpha_2 R_2(W_{i-1}^n, W_{i+1}^n), \quad W_{i,r}^n = W_{i+1}^n.$$

- If the solution of the Riemann problem consists of a 1-shock and a 2-shock waves (case 3) we select one of them depending on the amplitude of the $\alpha_1$ and $\alpha_2$ coefficients in order to choose the 'dominant' one:

  - If $|\alpha_1| \leq |\alpha_2|$ then:

$$\sigma_i^n = \bar{u} + h_{i-1}^n\sqrt{\bar{u}}, \quad W_{i,l}^n = W_{i+1}^n - \alpha_2 R_2(W_{i-1}^n, W_{i+1}^n), \quad W_{i,r}^n = W_{i+1}.$$

  - If $|\alpha_1| > |\alpha_2|$ then:

$$\sigma_i^n = \bar{u} - h_{i-1}^n\sqrt{\bar{u}}, \quad W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_{i-1}^n + \alpha_1 R_1(W_{i-1}, W_{i+1}).$$

The variable $h$ is selected in (5.2.4).

According to Theorem 5.2.1, the corresponding first and second-order in-cell discontinuous reconstruction methods capture correctly isolated shock waves and, as it will be also seen in Test 3, it also captures correctly the solution of Riemann problems consisting of two shock waves traveling in the same direction. Nevertheless, although it improves the results obtained with the standard methods and gets closer to the exact solution when the mesh is refined, it fails in capturing exactly the solution of Riemann problems involving two shocks traveling in opposite directions: the reason is that the intermediate state linking the two shocks is not exactly captured by Roe method.

Nevertheless, a more sophisticated strategy based on the exact solution of the Riemann problems (see Subsection 5.2.2) allows one to handle correctly with these situations. The key ingredients are:

- The solution of the Riemann problem with initial data $W_{i-1,r}^{n-1}$ and $W_{i+1,l}^{n-1}$ is used to mark the cells instead of the one corresponding to the initial data $W_{i-1}^n$ and $W_{i+1}^n$, where $W_{i-1,r}^{n-1}$ and $W_{i+1,l}^{n-1}$ are the states selected in the discontinuous reconstruction in the previous time step.

- The exact intermediate state is used when the solution of the Riemann problem involves two shock waves.

- If the solution of this Riemann problem involves two shock waves traveling in the same direction, a reconstruction with two discontinuities (one for each of the shock waves) is considered, so that the complete structure of the Riemann solution is imposed.

In order to avoid an excess of indices the following notation will be used:

$$W_{i-1,r}^{n-1} = W_L = [h_L, q_L]^T, \quad W_{i+1,l}^{n-1} = W_R = [h_R, q_R]^T.$$

The discontinuous reconstruction is then as follows:

- If the solution of the Riemann problem consists of 1-shock and a 2-rarefaction (case 1) then

$$\sigma_i^n = \sigma_1(W_l, W_*), \quad W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_*,$$

where $W_* = [h_*, q_*]^T$ is the intermediate state in the solution of the Riemann problem: $h_*$ is the root of the function:

$$f_{s,r}(h) = \left(\frac{h - h_r}{2} + \sqrt{u_r}\right)^2 - u_l + \sqrt{\frac{u_l(h + h_l)}{2h}}(h - h_l),$$

such that $h_l < h_* < h_r$. Once $h_*$ has been computed, $q_*$ is given by

$$q_* = h_* \left(\frac{h_* - h_r}{2} + \sqrt{u_r}\right)^2.$$

- If the solution of the Riemann problem consists of a 1-rarefaction and a 2-shock, then:

$$\sigma_i^n = \sigma_2(W_*, W_r), \quad W_{i,l}^n = W_*, \quad W_{i,r}^n = W_{i+1}^n,$$

where $W_* = [h_*, q_*]^T$ is the intermediate state: $h_*$ is the root of the function:

$$f_{r,s}(h) = \left(\frac{h_l - h}{2} + \sqrt{u_l}\right)\left(\frac{h_l - h}{2} + \sqrt{u_l} + \sqrt{\frac{h_r + h}{2h_r}}(h_r - h)\right) - u_r,$$

such that $h_r < h_* < h_l$. Once $h_*$ has been computed, $q_*$ is given by

$$q_* = h_* \left(\frac{h_l - h_*}{2} + \sqrt{u_l}\right)^2.$$

- If the solution of the Riemann problem consists of a 1-shock and a 2-shock, the intermediate state $W_* = [h_*, q_*]^T$ can be computed as follows: $h_*$ is the root of the function

$$f_{s,s}(h) = u_*(h) + \sqrt{\frac{u_*(h)(h+h_r)}{2h_r}}(h_r - h) - u_r,$$

where

$$u_*(h) = u_l - \sqrt{\frac{u_l(h+h_l)}{2h}}(h - h_l),$$

such that $h_* < h_l$ and $h_* < h_r$. Once $h_*$ has been computed, $q_*$ is obtained by:

$$q_* = h_* u_*(h_*).$$

Let us denote by $\sigma_1$ and $\sigma_2$ the speeds of the 1 and the 2 shock waves $\sigma_1(W_l, W_*)$ and $\sigma_2(W_*, W_r)$. The discontinuous reconstruction is then selected as follows:

  - If $\sigma_1 < 0 < \sigma_2$: let $d_1$ and $d_2$ be given by

$$d_1 = \frac{h_* - h_i^n}{h_* - h_l}, \quad d_2 = \frac{h_r - h_i^n}{h_r - h_*}.$$

    Then:

    * If $|\sigma_1| \leq |\sigma_2|$:
      · If $0 \leq d_2 \leq 1$, then

$$\sigma_i^n = \sigma_2, \quad W_{i,l}^n = W_*, \quad W_{i,r}^n = W_{i+1}^n.$$

      · Otherwise, if $0 \leq d_1 \leq 1$, then

$$\sigma_i^n = \sigma_1, \quad W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_*.$$

    * If $|\sigma_1| > |\sigma_2|$:
      · If $0 \leq d_1 \leq 1$, then

$$\sigma_i^n = \sigma_1, \quad W_{i,l}^n = W_{i-1}^n, \quad W_{i,r}^n = W_*.$$

      · Otherwise, if $0 \leq d_2 \leq 1$, then

$$\sigma_i^n = \sigma_2, \quad W_{i,l}^n = W_*, \quad W_{i,r}^n = W_{i+1}^n.$$

  - Otherwise (i.e. if $0 \leq \sigma_1 < \sigma_2$ or $\sigma_1 < \sigma_2 \leq 0$): let $d_1$ and $d_2$ be such that

$$\begin{cases} d_1 h_l + (d_2 - d_1)h_* + (1 - d_2)h_r = h_i^n, \\ d_1 q_l + (d_2 - d_1)q_* + (1 - d_2)q_r = q_i^n. \end{cases} \tag{5.3.17}$$

Then:

$$\mathbb{P}_i^n(x,t) = \begin{cases} W_l & \text{if } x \leq x_{i-1/2} + d_1\Delta x + \sigma_1(t-t_n), \\ W_* & \text{if } x_{i-1/2} + d_1\Delta x + \sigma_1(t-t_n) \leq x \leq x_{i-1/2} + d_2\Delta x + \sigma_2(t-t_n), \\ W_r & \text{otherwise.} \end{cases} \tag{5.3.18}$$

Observe this in-cell discontinuous reconstruction can only be done if $0 \leq d_1, d_2 \leq 1$, otherwise the cell is unmarked. Moreover, if $d_1 = d_2 = 1$ and the speeds of the shocks are positive (resp. if $d_1 = d_2 = 0$ and the speeds of the shocks are negative) the cell is unmarked and the cell $I_{i+1}$ (resp. the cell $I_{i-1}$) is marked if necessary.

Observe that, when the speeds of the shocks have the same sign, the discontinuous reconstruction coincides with the solution of the Riemann problem.

The numerical methods using the first strategy for the discontinuous reconstruction (based on the Roe matrix) will be labeled again by O$p$_DisRec and those using the second one (based on the exact solutions of the Riemann problems) by O$p$_ExactDisRec.

**Test 1: Isolated 1-shock**

Let us consider the following initial condition taken from [42]

$$(h,q)_0(x) = \begin{cases} (1,1) & \text{if } x < 0, \\ (1.8, 0.530039370688997) & \text{otherwise.} \end{cases}$$

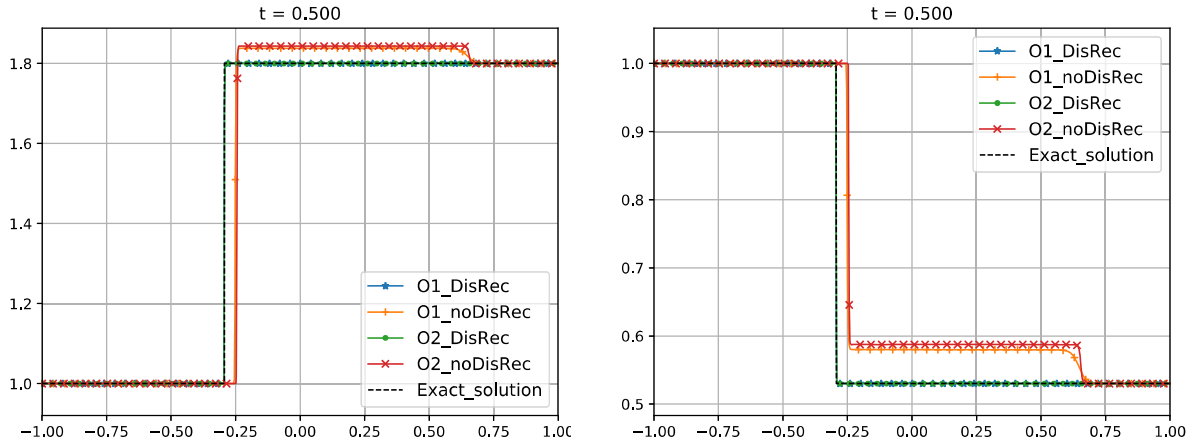
The solution of the Riemann problem consists of a 1-shock wave joining the left and right states. Figure 5.24 compares the exact solution and the numerical approximations at time $t = 0.15$ obtained with Roe method, its second order extension based on the standard MUSCL-Hancock reconstruction, and the first and second order discontinuous in-cell reconstruction schemes based on the Roe matrix using 1000-cell mesh and CFL $= 0.5$: as it can be seen the standard Roe methods does not capture the discontinuities properly what is not the case for the in-cell discontinuous reconstruction methods based on the Roe structure. The results obtained with O$p$_ExactDisRec are similar.

**Test 2: left-moving 1-shock + right-moving 2-shock**

Let us consider the following initial condition

$$(h,q)_0(x) = \begin{cases} (1,1) & \text{if } x < 0, \\ (1.5, 0.1855893974385) & \text{otherwise.} \end{cases} \tag{5.3.19}$$

The solution of the Riemann problem consists of a 1-shock wave with negative speed and a 2-shock with positive speed with intermediate state $W_* = [1.8, 0.530039370688997]^T$.




(a) Variable $h$ (b) Variable $q$

Figure 5.24: Modified shallow water system. Test 1: Numerical solutions obtained withe the first and second-order methods with and without discontinuous reconstruction based on the Roe matrix at time $t = 0.5$ with 1000 cells. Left: variable $h$. Right: variable $q$

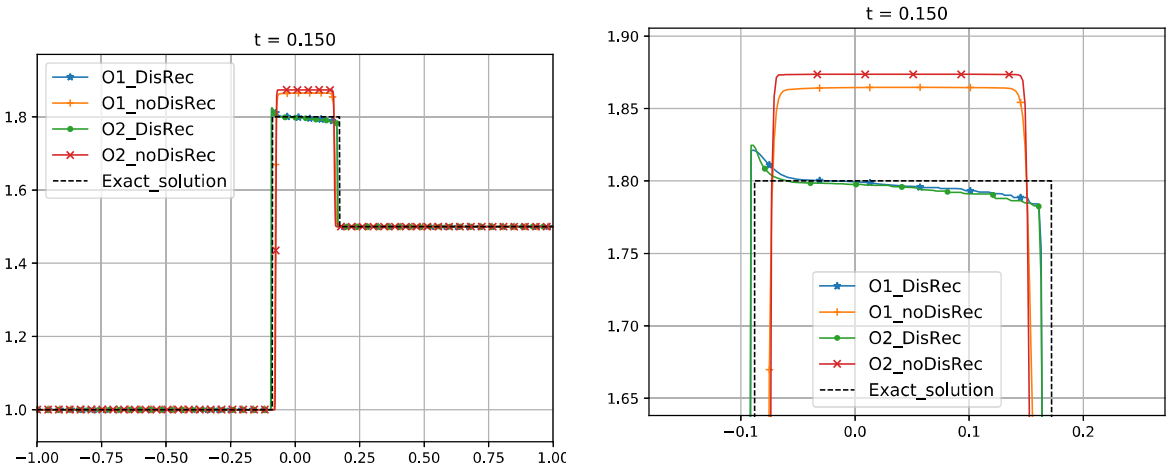
Figures 5.25 and 5.26 compare the exact solution with the numerical approximations at time $t = 0.15$ obtained with Roe method, its second order extension based on the standard MUSCL-Hancock reconstruction, and the first and second order discontinuous in-cell reconstruction schemes based on the Roe matrix using 1000-cell mesh and CFL = 0.5: as it can be seen none of them capture the discontinuities properly, although the ones with using in-cell discontinuous reconstruction do it better. Figures 5.27 and 5.28 show the numerical solutions obtained with the first-order method with discontinuous reconstruction based on the Roe matrix at time $t = 0.15$ using different cell meshes: as we can see the numerical solutions seem to converge to the exact solution when $\Delta x \to 0$. In Figure 5.29 the results given by the first and second order in-cell discontinuous schemes based in the exact solution of the Riemann problem are shown: we observe that both of them capture exactly the two shocks.

**Test 3: right-moving 1-shock + right-moving 2-shock**

Let us consider the following initial condition

$$(h, q)_0(x) = \begin{cases} (1, 1) & \text{if } x < 0, \\ (5, 2.86423084288) & \text{otherwise.} \end{cases} \tag{5.3.20}$$

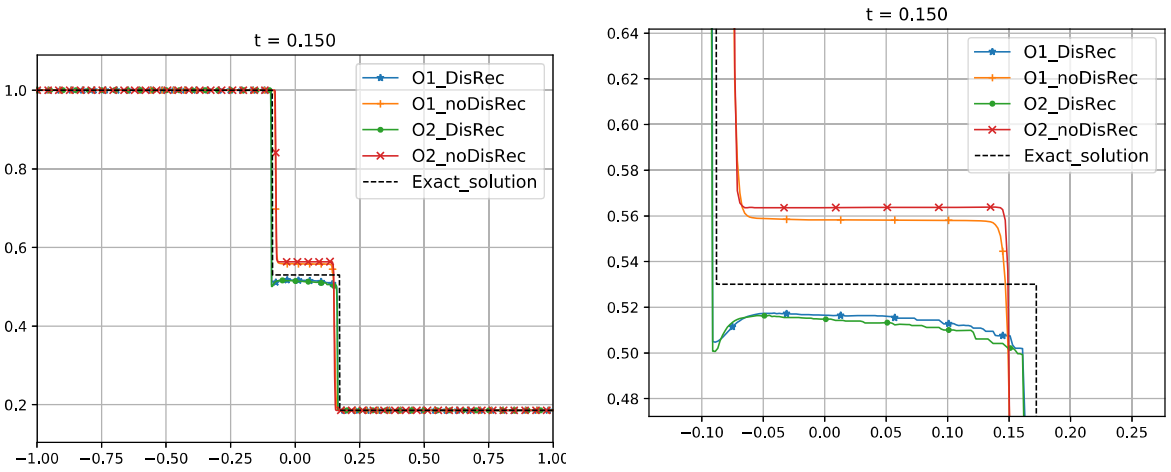
The solution of the Riemann problem consists of a 1-shock and a 2-shock waves with positive speed and intermediate state $W_* = [1.5, 5.96906891076]^T$. Figures 5.30 and 5.31




(a) Zoom

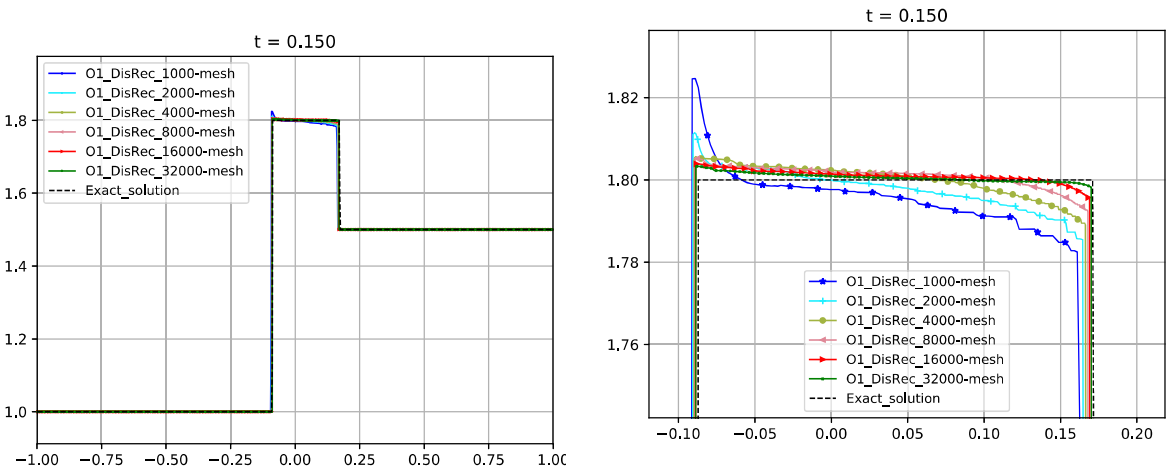
Figure 5.25: Modified shallow water system. Test 2: variable $h$. Left: Numerical solutions obtained with the first and second-order methods with and without discontinuous reconstruction based on the Roe matrix at time $t = 0.15$ with 1000 cells. Right: zoom

show the exact solution and the numerical approximations at time $t = 0.06$ obtained with Roe method, its second order extension based on the standard MUSCL-Hancock reconstruction, and the first and second order discontinuous in-cell reconstruction schemes based on the Roe structure using 1000-cell mesh and CFL = 0.5: as in the previous test case, the in-cell discontinuous reconstruction capture the shocks and intermediate state much better than the standard first and second order Roe methods. In Figure 5.32 the results given by the first and second order in-cell discontinuous schemes based in the exact solution of the Riemann problems are shown: both of them capture exactly the exact solution.



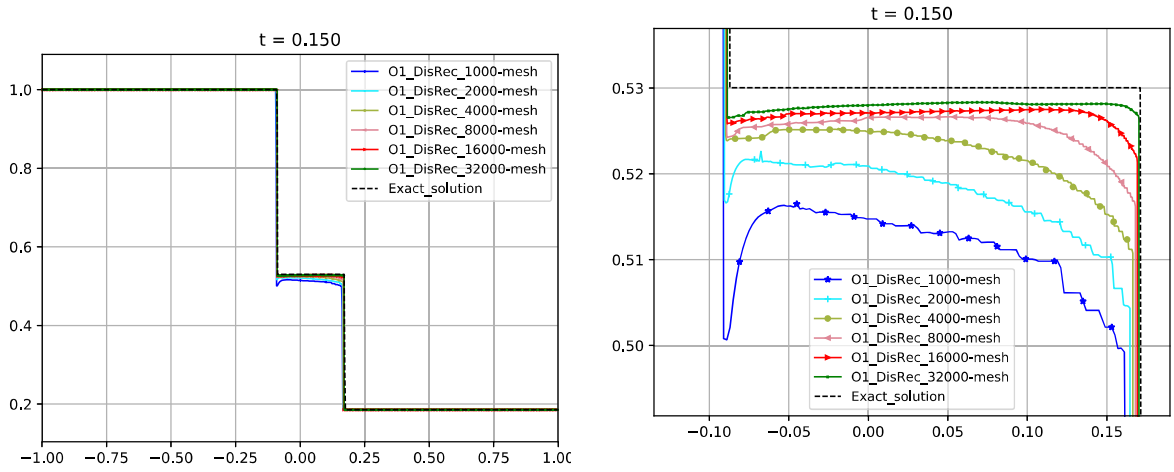

(a) Zoom

Figure 5.26: Modified shallow water system. Test 2: variable $q$. Left: Numerical solutions obtained with the first and second-order methods with and without discontinuous reconstruction based on the Roe matrix at time $t = 0.15$ with 1000 cells. Right: zoom



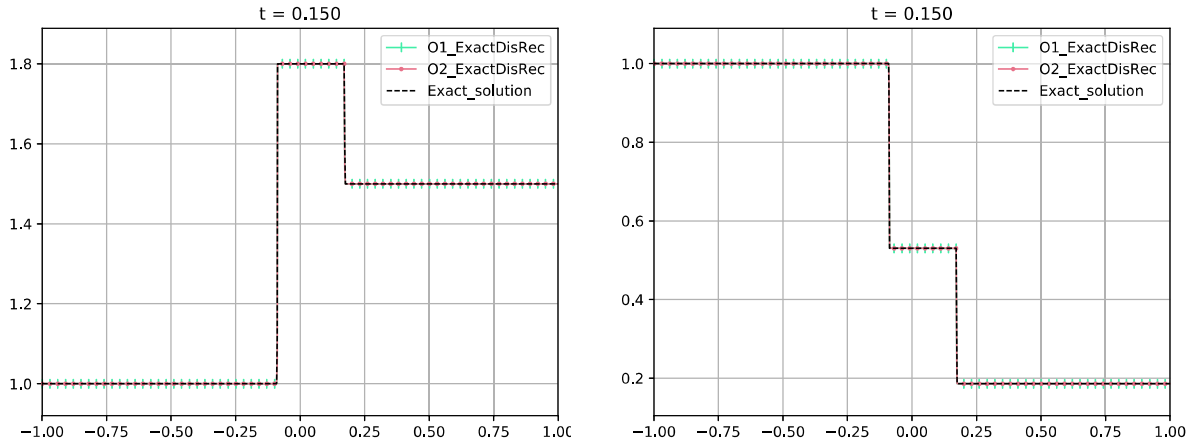

(a) Zoom

Figure 5.27: Modified shallow water system. Test 2: variable $h$. Left: Numerical solutions obtained with the first-order methods with discontinuous reconstruction based on the Roe matrix at time $t = 0.15$ with different cell meshes. Right: zoom.



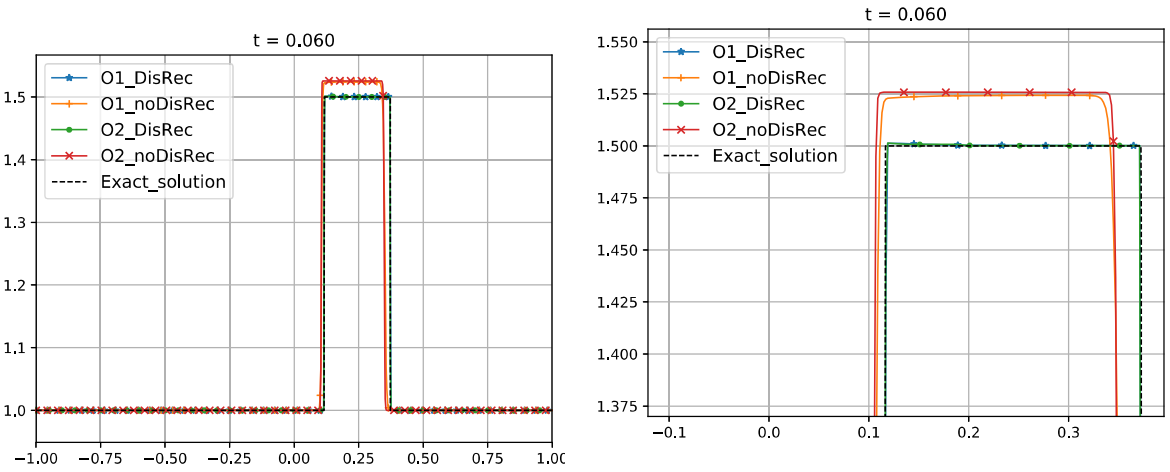

(a) Zoom

Figure 5.28: Modified shallow water system. Test 2: variable $q$. Left: Numerical solutions obtained with the first-order methods with discontinuous reconstruction based on the Roe matrix at time $t = 0.15$ with different cell meshes. Right: zoom.



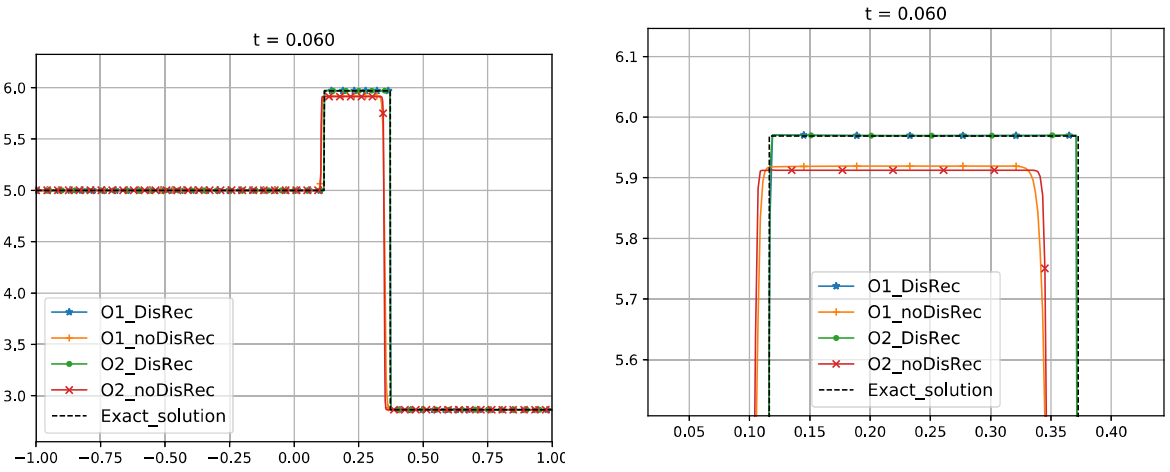

(a) Variable $h$ (b) Variable $q$

Figure 5.29: Modified shallow water system. Test 2: Numerical solutions obtained with the first and second-order methods with discontinuous reconstruction based on the exact solutions of the Riemann problems at time $t = 0.15$ with 1000 cells. Left : variable $h$. Right: variable $q$.



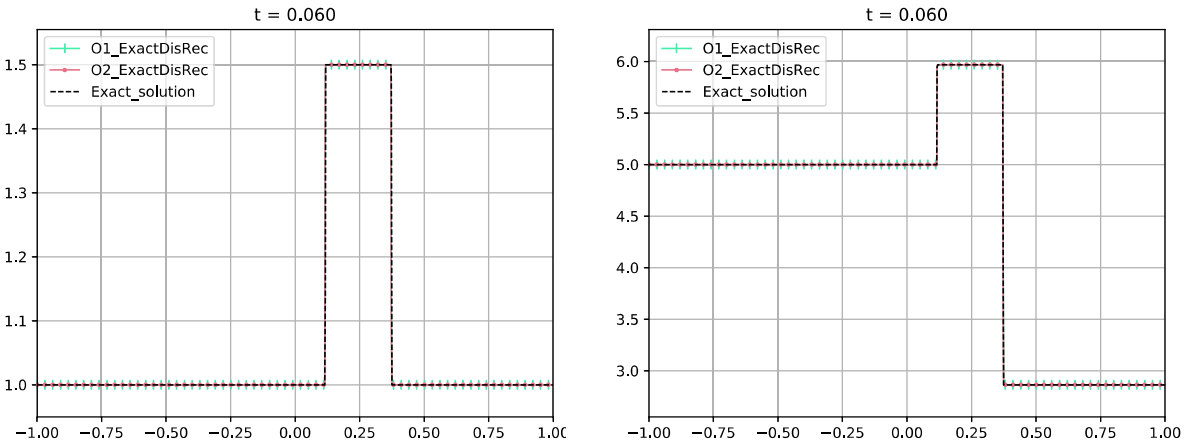

(a) Zoom

Figure 5.30: Modified shallow water system. Test 3: variable $h$. Left: Numerical solutions obtained with the first and second-order methods with and without discontinuous reconstruction based on the Roe matrix at time $t = 0.06$ with 1000 cells. Right: zoom.




(a) Zoom

Figure 5.31: Modified shallow water system. Test 3: variable $q$. Left: Numerical solutions obtained with the first and second-order methods with and without discontinuous reconstruction based on the Roe matrix at time $t = 0.06$ with 1000 cells. Right: zoom.




(a) Variable $h$ (b) Variable $q$

Figure 5.32: Modified shallow water system. Test 3: Numerical solutions obtained with the first and second-order methods with discontinuous reconstruction based on the exact solutions of the Riemann problems at time $t = 0.06$ with 1000 cells. Left: variable $h$. Right: variable $q$.

# Chapter 6

# Conclusions and future work

In this chapter we summarize the main contributions performed in this thesis and the possible future research that could arise from it.

## 6.1 Conclusions

Throughout this thesis we have considered a wide number of analytical and numerical aspects of conservative and nonconservative hyperbolic systems of Partial Differential Equations. Different problems that arise when dealing with this kind of systems have been studied and solved. Let us highlight the main contributions in each of the chapters.

In Chapter 2 the Riemann problem for the shallow water equations corresponding to a wet-dry front over a step is studied. Like in [112] a monotonicity criterion (MC) is imposed to select the admissible stationary discontinuities over the step. Depending on the state at the wet side, zero, one, or two solutions are found. In the case of non-existence of solution, the problem is interpreted as a partial Riemann problem for the homogeneous shallow water and a solution is proposed. Since the problem is nonconservative, all the difficulties concerning the numerical approximation of weak solutions are found. To check this, we compare the numerical solutions obtained with several standard numerical fluxes and treatment of the source terms. The main conclusions are the following:

- In cases where the Riemann problem does not have solutions, the numerical methods converge to the proposed solutions based on a reinterpretation of the problem.
- In order to capture correctly the stationary contact discontinuities it is necessary to have a numerical method that preserve them. Nevertheless, this is not sufficient to ensure the convergence to the proposed solutions: the numerical solution may converge to weak solutions containing stationary discontinuities over the step that do not satisfy the (MC) criterion.



- The combination of Godunov numerical flux and the Generalized Hydrostatic Reconstruction technique introduced in [51] seems to produce a numerical method that correctly captures all the proposed solutions.
- In cases where the Riemann problem has two solutions, the numerical methods may converge to one or to the other.

Besides the theoretical interest of this analysis, the results may be useful to design numerical methods and/or to produce reference solutions to compare different schemes.

In Chapter 3, an efficient implementation of PVM methods that are based on interpolation polynomials is presented: the Newton form of the polynomial is used to reduce the number of calculations. Next, the relation between SRS and PVM, already studied in [125], is revisited. In particular, it is shown that many SRS can be interpreted as PVM methods based on a Lagrange interpolation polynomial, what allows one to use the implementation based on the Newton form of the polynomial. In particular, Roe method can be interpreted in terms of a complete SRS and thus as a PVM method, what allows us to implement it using the Newton form of the polynomial. We compare numerically the efficiency of the standard implementation of Roe method and the new one for two different models: the two-layer shallow water equations and the Quadrature-Based Moment equations for rarefied gases. According to our results, a small speedup is obtained using the Newton Roe method compared to the standard one for the two-layer shallow water system, as the number of equations is not big enough. In the case of the QBME model the speedup increases with the number of moments: Newton Roe method is about 3.5 times faster than the standard Roe method for the 11 moment equations in primitive variables. In the case of the partially-conservative formulation of the QBME model the results are even better: Newton Roe method is 4.1 times faster than the standard one, due to the fact that the standard implementation requires the computation of the eigenvectors. Moreover, this factor increases with the number of moments. Therefore, we can conclude that the Newton Roe method yields an improvement of the standard Roe scheme for systems with a large number of equations.

In Chapter 4 the procedure introduced in [40] and recalled in [47] is extended to the relativistic fluid flows in the Schwarzschild background. More precisely, we develop first and higher order well-balanced schemes for the relativistic Burgers and Euler systems. Several numerical tests are used to validate the schemes and to highlight the relevance of the well-balanced property when dealing with these relativistic flows. We also use these schemes to perform a systematic numerical study of these two PDE systems in order to be able to extract general conclusions about the long time behavior of the flow. Such a study is expected to be a useful tool to direct the mathematical analysis of the models and the study of more complex relativistic models.



Finally, in Chapter 5, an extension to second-order accuracy of the in-cell discontinuous reconstruction methods introduced in [51] is presented: it is compared with the first-order one and with standard path-conservative numerical methods using several numerical tests. We observe, as expected, an improvement in the smooth regions of the solutions. The isolated shock-capturing property is enunciated, proved and tested. Two different strategies to design the in-cell discontinuous reconstructions are considered: the first one based on the solutions of the linearized Riemann problems when a Roe matrix is available, and the second one based on the exact solution of the Riemann problems when available. It is observed that the quality of the results provided by this technique for Riemann problems that involve more than one shock is worse although they seem to converge to the right solution. A more sophisticated definition of the in-cell discontinuous reconstruction is proposed for the modified shallow water system (5.3.10) that allows us to improve the quality of the solutions.

## 6.2 Future works

The study of the Riemann problems corresponding to dry-wet fronts in Chapter 2 will be extended to other shallow water systems such as the shallow water model with two velocities considered in [2].

The new efficient implementation of the interpolatory PVM methods based on the Newton form will be applied to different multilayer shallow water systems with or without dispersion effects. In the case of the multilayer system in [44] neither the eigenvalues nor the eigenvectors can be computed exactly so that a significant improvement of the computational cost is expected. The same happens with the multilayer systems introduced in [6, 7].

The following extensions of the study of the relativistic systems in the Schwarzschild background are expected:

- The design of well-balanced methods of order of accuracy bigger than two for the Euler-Schwarzschild model based on the numerical approximation of the stationary solutions using ODE system solvers (see [87]).
- Development of well-balanced high-order methods for multidimensional problems.
- Development of numerical methods for other relativistic models of greater complexity.

The new second order scheme from Chapter 5 based in the in-cell discontinuous reconstruction sets the basis for the following lines of research:

- Extension of this technique to arbitrary order of accuracy. The use of methods based on the Taylor expansion will be considered as in [29, 30].

- Design of numerical methods that capture correctly non isolated shocks. The use of a numerical solver to approximate the structure of the Riemann problem as it was done in [94] will be considered for problems for which the solution of the Riemann problem is not available.
- Application of the methods to more complex models.
- Development of Discontinuous Galerkin (DG) solvers based on discontinuous reconstructions.
- Extension to multidimensional problems.



# Bibliography


[1] R. Abgrall and S. Karni. A comment on the computation of non-conservative products. *Journal of Computational Physics*, 229(8):2759–2763, 2010.

[2] N. Aguillon, E. Audusse, E. Godlewski, and M. Parisot. Analysis of the Riemann problem for a shallow water model with two velocities. *SIAM Journal on Mathematical Analysis*, 50(5):4861–4888, 2018.

[3] F. Alcrudo and F. Benkhaldoun. Exact solutions to the Riemann problem of the shallow water equations with a bottom step. *Computers & Fluids*, 30(6):643–671, 2001.

[4] B. Audebert and F. Coquel. Hybrid Godunov-Glimm method for a nonconservative hyperbolic system with kinetic relations. In *Numerical Mathematics and Advanced Applications*, pages 646–653. Springer, 2006.

[5] 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 Journal on Scientific Computing*, 25(6):2050–2065, 2004.

[6] E. Audusse, M.-O. Bristeau, M. Pelanti, and J. Sainte-Marie. Approximation of the hydrostatic Navier–Stokes system for density stratified flows by a multilayer model: kinetic interpretation and numerical solution. *Journal of Computational Physics*, 230(9):3453–3478, 2011.

[7] E. Audusse, M.-O. Bristeau, B. Perthame, and J. Sainte-Marie. A multilayer Saint-Venant system with mass exchanges for shallow water flows. Derivation and numerical validation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 45(1):169–200, 2011.

[8] E. Audusse, C. Chalons, and P. Ung. A simple well-balanced and positive numerical scheme for the shallow-water system. *Communications in Mathematical Sciences*, 13(5):1317–1332, 2015.




[9] A. Baeza, R. Bürger, P. Mulet, and D. Zorío. On the efficient computation of smoothness indicators for a class of WENO reconstructions. *Journal of Scientific Computing*, 80(2):1240–1263, 2019.

[10] A. Beljadid, P. G. LeFloch, S. Mihsra, and C. Parés. Schemes with well-controlled dissipation. hyperbolic systems in nonconservative form. *Communications in Computational Physics*, 21(4):913–946, 2017.

[11] A. Beljadid, P. G. LeFloch, and A. Mohammadian. Late-time asymptotic behavior of solutions to hyperbolic conservation laws on the sphere. *Computer Methods in Applied Mechanics and Engineering*, 349:285–311, 2019.

[12] J. P. Berberich, P. Chandrashekar, and C. Klingenberg. High order well-balanced finite volume methods for multi-dimensional systems of hyperbolic balance laws. *Computers & Fluids*, 219:104858, 2021.

[13] A. Bermúdez and M. E. Vázquez-Cendón. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1049–1071, 1994.

[14] R. Bernetti, V. A. Titarev, and E. F. Toro. Exact solution of the Riemann problem for the shallow water equations with discontinuous bottom geometry. *Journal of Computational Physics*, 227(6):3212–3243, 2008.

[15] C. Berthon. Schéma nonlinéaire pour l'approximation numérique d'un système hyperbolique non conservatif. *Comptes Rendus Mathematique*, 335:1069–1072, 2002.

[16] C. Berthon and C. Chalons. A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. *Mathematics of Computation*, 85(299):1281–1307, 2016.

[17] C. Berthon and F. Coquel. Nonlinear projection methods for multi-entropies Navier-Stokes systems. In *Innovative Methods For Numerical Solution Of Partial Differential Equations*, pages 278–304. World Scientific, 2002.

[18] C. Berthon, F. Coquel, and P. G. LeFloch. Why many theories of shock waves are necessary: kinetic relations for non-conservative systems. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 142(1):1–37, 2012.

[19] C. Berthon, A. Duran, F. Foucher, K. Saleh, and J. D. D. Zabsonré. Improvement of the hydrostatic reconstruction scheme to get fully discrete entropy inequalities. *Journal of Scientific Computing*, 80(2):924–956, 2019.

[20] C. Berthon and F. Foucher. Efficient well-balanced hydrostatic upwind schemes for shallow-water equations. *Journal of Computational Physics*, 231(15):4993–5015, 2012.

[21] C. Berthon, M. M'Baye, M. Le, and D. Seck. A well-defined moving steady states capturing Godunov-type scheme for shallow-water model. *International Journal on Finite Volumes*, 2021.

[22] P. L. Bhatnagar, E. P. Gross, and M. Krook. A model for collision processes in gases. 1. Small amplitude processes in charged and neutral one-component systems. *Physical review*, 94(3):511–525, 1954.

[23] R. Borsche. A well-balanced solver for the Saint Venant equations with variable cross-section. *Journal of Numerical Mathematics*, 23(2):99–115, 2015.

[24] R. Borsche and A. Klar. Flooding in urban drainage systems: coupling hyperbolic conservation laws for sewer systems and surface flow. *International Journal for Numerical Methods in Fluids*, 76(11):789–810, 2014.

[25] F. Bouchut. *Nonlinear stability of finite Volume Methods for hyperbolic conservation laws: And Well-Balanced schemes for sources*. Springer Science & Business Media, 2004.

[26] Z. Cai, Y. Fan, and R. Li. Globally hyperbolic regularization of Grad's moment system in one dimensional space. *Communications in Mathematical Sciences*, 11(2):547–571, 2013.

[27] Z. Cai, Y. Fan, and R. Li. Globally hyperbolic regularization of Grad's moment system. *Communications on Pure and Applied Mathematics*, 67(3):464–518, 2014.

[28] A. Canestrelli, A. Siviglia, M. Dumbser, and E. F. Toro. Well-balanced high-order centred schemes for non-conservative hyperbolic systems. Applications to shallow water equations with fixed and mobile bed. *Advances in Water Resources*, 32(6):834–844, 2009.

[29] H. Carrillo and C. Parés. Compact approximate taylor methods for systems of conservation laws. *Journal of Scientific Computing*, 80(3):1832–1866, 2019.

[30] H. Carrillo, C. Parés, and D. Zorío. Lax-Wendroff approximate taylor methods with fast and optimized weighted essentially non-oscillatory reconstructions. *Journal of Scientific Computing*, 86(1):1–41, 2021.

[31] M. Castro, J. M. Gallardo, and C. Parés. High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. Applications to shallow-water systems. *Mathematics of Computation*, 75(255):1103–1134, 2006.

[32] M. J. Castro, T. Chacón Rebollo, E. D. Fernández-Nieto, and C. Parés. On well-balanced finite volume methods for nonconservative nonhomogeneous hyperbolic systems. *SIAM Journal on Scientific Computing*, 29(3):1093–1126, 2007.

[33] M. J. Castro, Y. Cheng, A. Chertock, and A. Kurganov. Solving two-mode shallow water equations using finite volume methods. *Communications in Computational Physics*, 16(5):1323–1354, 2014.

[34] M. J. Castro, T. Morales de Luna, and C. Parés. Well-balanced schemes and path-conservative numerical methods. In *Handbook of Numerical Analysis*, volume 18, pages 131–175. Elsevier, 2017.

[35] M. J. Castro and E. Fernández-Nieto. A class of computationally fast first order finite volume solvers: PVM methods. *SIAM Journal on Scientific Computing*, 34(4):A2173–A2196, 2012.

[36] M. J. Castro, E. D. Fernández-Nieto, and A. M. Ferreiro. Sediment transport models in shallow water equations and numerical approach by high order finite volume methods. *Computers & Fluids*, 37(3):299–316, 2008.

[37] M. J. Castro, E. D. Fernández-Nieto, T. Morales de Luna, G. Narbona-Reina, and C. Parés. A HLLC scheme for nonconservative hyperbolic problems. Application to turbidity currents with sediment transport. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 47(1):1–32, 2013.

[38] M. J. Castro, A. M. Ferreiro, J. A. García-Rodríguez, J. M. González-Vida, J. Macías, C. Parés, and M. E. Vázquez-Cendón. The numerical treatment of wet/dry fronts in shallow flows: application to one-layer and two-layer systems. *Mathematical and Computer Modelling*, 42(3-4):419–439, 2005.

[39] M. J. Castro, U. S. Fjordholm, S. Mishra, and C. Parés. Entropy conservative and entropy stable schemes for nonconservative hyperbolic systems. *SIAM Journal on Numerical Analysis*, 51(3):1371–1391, 2013.

[40] M. J. Castro, J. M. Gallardo, J. A. López-García, and C. Parés. Well-balanced high order extensions of Godunov's method for semilinear balance laws. *SIAM Journal on Numerical Analysis*, 46(2):1012–1039, 2008.

[41] M. J. Castro, J. M. González-Vida, and C. Parés. Numerical treatment of wet/dry fronts in shallow flows with a modified Roe scheme. *Mathematical Models and Methods in Applied Sciences*, 16(06):897–931, 2006.

[42] M. J. Castro, P. G. LeFloch, M. L. Muñoz-Ruiz, and C. Pares. Why many theories of shock waves are necessary: Convergence error in formally path-consistent schemes. *Journal of Computational Physics*, 227(17):8107–8129, 2008.

[43] M. J. Castro, J. A. López-García, and C. Parés. High order exactly well-balanced numerical methods for shallow water systems. *Journal of Computational Physics*, 246:242–264, 2013.

[44] M. J. Castro, J. Macías, and C. Parés. A $Q$-scheme for a class of systems of coupled conservation laws with source term. Application to a two-layer 1-D shallow water system. *ESAIM: Mathematical Modelling and Numerical Analysis*, 35(1):107–127, 2001.

[45] M. J. Castro, A. Pardo, C. Parés, and E. F. Toro. On some fast well-balanced first order solvers for nonconservative systems. *Mathematics of Computation*, 79(271):1427–1472, 2010.

[46] M. J. Castro, A. Pardo Milanés, and C. Parés. Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique. *Mathematical Models and Methods in Applied Sciences*, 17(12):2055–2113, 2007.

[47] M. J. Castro and C. Parés. Well-balanced high-order finite volume methods for systems of balance laws. *Journal of Scientific Computing*, 82(2):1–48, 2020.

[48] M. J. Castro, C. Parés, G. Puppo, and G. Russo. Central schemes for nonconservative hyperbolic systems. *SIAM Journal on Scientific Computing*, 34(5):B523–B558, 2012.

[49] J.-J. Cauret, J.-F. Colombeau, and A.Y. Le Roux. Discontinuous generalized solutions of nonlinear nonconservative hyperbolic equations. *Journal of mathematical analysis and applications*, 139(2):552–573, 1989.

[50] T. Chacón Rebollo, A. Dominguez Delgado, and E. D. Fernández-Nieto. A family of stable numerical solvers for the shallow water equations with source terms. *Computer Methods in Applied Mechanics and Engineering*, 192(1-2):203–225, 2003.

[51] C. Chalons. Path-conservative in-cell discontinuous reconstruction schemes for non conservative hyperbolic systems. *Communications in Mathematical Sciences*, 18(1):1–30, 2020.

[52] C. Chalons and F. Coquel. Navier-Stokes equations with several independent pressure laws and explicit predictor-corrector schemes. *Numerische Mathematik*, 101(3):451–478, 2005.



[53] C. Chalons and F. Coquel. A new comment on the computation of non-conservative products using Roe-type path conservative schemes. *Journal of Computational Physics*, 335:592–604, 2017.

[54] G. Chen and S. Noelle. A new hydrostatic reconstruction scheme based on subcell reconstructions. *SIAM Journal on Numerical Analysis*, 55(2):758–784, 2017.

[55] I. Cravero, G. Puppo, M. Semplice, and G. Visconti. CWENO: uniformly accurate reconstructions for balance laws. *Mathematics of Computation*, 87(312):1689–1719, 2018.

[56] I. Cravero and M. Semplice. On the accuracy of WENO and CWENO reconstructions of third order on nonuniform meshes. *Journal of Scientific Computing*, 67(3):1219–1246, 2016.

[57] G. Dal Maso, P. G. LeFloch, and F. Murat. Definition and weak stability of nonconservative products. *Journal de Mathématiques Pures et Appliquées*, 74(6):483–548, 1995.

[58] P. Degond, P.-F. Peyrard, G. Russo, and P. Villedieu. Polynomial upwind schemes for hyperbolic systems. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 328(6):479–483, 1999.

[59] O. Delestre, S. Cordier, F. Darboux, and F. James. A limitation of the hydrostatic reconstruction technique for shallow water equations. *Comptes Rendus Mathematique*, 350(13-14):677–681, 2012.

[60] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. A well-balanced scheme to capture non-explicit steady states in the Euler equations with gravity. *International Journal for Numerical Methods in Fluids*, 81(2):104–127, 2016.

[61] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. Well-balanced schemes to capture non-explicit steady states: Ripa model. *Mathematics of Computation*, 85(300):1571–1602, 2016.

[62] S. Dong and P. G. Lefloch. Convergence of the finite volume method on a Schwarzschild background. *ESAIM: Mathematical Modelling and Numerical Analysis*, 53(5):1459–1476, 2019.

[63] F. Dubois. Partial Riemann problem, boundary conditions, and gas dynamics. In *Absorbing boundaries and layers,domain decomposition methods*, pages 16–77. Nova Science Publishers, 2001.

[64] M. Dumbser, D. S. Balsara, E. F. Toro, and C.-D. Munz. A unified framework for the construction of one-step finite volume and discontinuous Galerkin schemes on unstructured meshes. *Journal of Computational Physics*, 227(18):8209–8253, 2008.

[65] M. Dumbser, M. J. Castro, C. Parés, and E. F. Toro. ADER schemes on unstructured meshes for nonconservative hyperbolic systems: Applications to geophysical flows. *Computers & Fluids*, 38(9):1731–1748, 2009.

[66] M. Dumbser, A. Hidalgo, M. J. Castro, C. Parés, and E. F. Toro. FORCE schemes on unstructured meshes II: Non-conservative hyperbolic systems. *Computer Methods in Applied Mechanics and Engineering*, 199(9-12):625–647, 2010.

[67] M. Dumbser and M. Käser. Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems. *Journal of Computational Physics*, 221(2):693–723, 2007.

[68] M. Dumbser, M. Käser, V. A. Titarev, and E. F. Toro. Quadrature-free non-oscillatory finite volume schemes on unstructured meshes for nonlinear hyperbolic systems. *Journal of Computational Physics*, 226(1):204–243, 2007.

[69] M. Dumbser and C.-D. Munz. Building blocks for arbitrary high order discontinuous Galerkin schemes. *Journal of Scientific Computing*, 27(1-3):215–230, 2006.

[70] G. Dziuk, D. Kroner, and T. Muller. Scalar conservation laws on moving hypersurfaces. *Interfaces and Free Boundaries*, 15(2):203–237, 2013.

[71] EDANYA. EDANYA web group. *https://www.uma.es/edanya*, 2021.

[72] Y. Fan, J. Koellermeier, J. Li, R. Li, and M. Torrilhon. Model reduction of kinetic equations by operator projection. *Journal of Statistical Physics*, 162(2):457–486, 2016.

[73] Y. Fan and R. Li. Globally hyperbolic moment system by generalized Hermite expansion. *Scientia Sinica Mathematica*, 45(10)(10):1635–1676, 2015.

[74] E. Fernández-Nieto, M. J. Castro, and C. Parés. On an intermediate field capturing Riemann solver based on a parabolic viscosity matrix for the two-layer shallow water system. *Journal of Scientific Computing*, 48(1-3):117–140, 2011.

[75] E. D. Fernández-Nieto, F. Bouchut, D. Bresch, M. J. Castro, and A. Mangeney. A new Savage–Hutter type model for submarine avalanches and generated tsunami. *Journal of Computational Physics*, 227(16):7720–7754, 2008.

[76] E. D. Fernández-Nieto, J. M. Gallardo, and P. Vigneaux. Efficient numerical schemes for viscoplastic avalanches. Part 1: the 1D case. *Journal of Computational Physics*, 264:55–90, 2014.

[77] U. S. Fjordholm and S. Mishra. Accurate numerical discretizations of non-conservative hyperbolic systems. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 46(1):187–206, 2012.

[78] E. Gaburro, M. J. Castro, and M. Dumbser. Well-balanced Arbitrary-Lagrangian-Eulerian finite volume schemes on moving nonconforming meshes for the euler equations of gas dynamics with gravity. *Monthly Notices of the Royal Astronomical Society*, 477(2):2251–2275, 2018.

[79] E. Gaburro, M. J. Castro, and M. Dumbser. A well balanced diffuse interface method for complex nonhydrostatic free surface flows. *Computers & Fluids*, 175:180–198, 2018.

[80] B. Ghitti, C. Berthon, M. H. Le, and E. F. Toro. A fully well-balanced scheme for the 1D blood flow equations with friction source term. *Journal of Computational Physics*, 421:109750, 2020.

[81] J. Giesselmann and P. G. LeFloch. Formulation and convergence of the finite volume method for conservation laws on spacetimes with boundary. *Numerische Mathematik*, pages 1–35, 2020.

[82] P. Glaister. Approximate Riemann solutions of the shallow water equations. *Journal of Hydraulic Research*, 26(3):293–306, 1988.

[83] J. Glimm. Solutions in the large for nonlinear hyperbolic systems of equations. *Communications on Pure and Applied Mathematics*, 18(4):697–715, 1965.

[84] P. Goatin and P. G. LeFloch. The Riemann problem for a class of resonant hyperbolic systems of balance laws. *Annales de l'Institut Henri Poincaré (C) Non Linear Analysis*, 21(6):881–902, 2004.

[85] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*. Springer, 1995.

[86] S. Godunov and I. Bohachevsky. Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Matematičeskij sbornik*, 47(3):271–306, 1959.

[87] I. Gómez-Bueno, M. J. Castro, and C. Parés. High-order well-balanced methods for systems of balance laws: a control-based approach. *Applied Mathematics and Computation*, 394:125820, 2021.

[88] S. Gottlieb and C.-W. Shu. Total variation diminishing Runge-Kutta schemes. *Mathematics of Computation*, 67(221):73–85, 1998.

[89] H. Grad. On the kinetic theory of rarefied gases. *Communications on Pure and Applied Mathematics*, 2(4):331–407, 1949.

[90] G. Guennebaud, B. Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[91] E. Han and G. Warnecke. Exact riemann solutions to shallow water equations. *Quarterly of Applied Mathematics*, 72(3):407–453, 2014.

[92] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes, III. In *Upwind and high-resolution schemes*, pages 218–290. Springer, 1987.

[93] A. Harten, P. D. Lax, and B. Van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM review*, 25(1):35–61, 1983.

[94] A. Hiltebrand, S. Mishra, and C. Parés. Entropy-stable space–time DG schemes for non-conservative hyperbolic systems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52(3):995–1022, 2018.

[95] T. Y. Hou and P. G. LeFloch. Why nonconservative schemes converge to wrong solutions: Error analysis. *Mathematics of Computation*, 62(206):497–530, April 1994.

[96] G.-S. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*, 126(1):202–228, 1996.

[97] C. Klingenberg, G. Puppo, and M. Semplice. Arbitrary order finite volume well-balanced schemes for the Euler equations with gravity. *SIAM Journal on Scientific Computing*, 41(2):A695–A721, 2019.

[98] J. Koellermeier. *Derivation and numerical solution of hyperbolic moment equations for rarefied gas flows.* dissertation, RWTH Aachen University, Aachen, 2017.

[99] J. Koellermeier, R. P. Schaerer, and M. Torrilhon. A framework for hyperbolic approximation of kinetic equations using quadrature-based projection methods. *Kinetic and Related Models*, 7(3):531–549, 2014.

[100] J. Koellermeier and M. Torrilhon. Simplified hyperbolic moment equations. In *Proceedings of the 16th International Conference on Hyperbolic Problems*, 2016.

[101] J. Koellermeier and M. Torrilhon. Numerical solution of hyperbolic moment models for the Boltzmann equation. *European Journal of Mechanics - B/Fluids*, 64:41–46, 2017.

[102] J. Koellermeier and M. Torrilhon. Numerical study of partially conservative moment equations in kinetic theory. *Communications in Computational Physics*, 21(04)(4):981–1011, 2017.

[103] N. Krvavica, M. Tuhtan, and G. Jelenić. Analytical implementation of Roe solver for two-layer shallow water equations with accurate treatment for loss of hyperbolicity. *Advances in Water Resources*, 122:187–205, 2018.

[104] P. D. Lax. Hyperbolic systems of conservation laws II. *Communications on Pure and Applied Mathematics*, 10(4):537–566, 1957.

[105] P. D. Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, volume 11. Society for Industrial and Applied Mathematics, 1973.

[106] P. G. LeFloch. *Hyperbolic Systems of Conservation Laws: The theory of classical and nonclassical shock waves*. Springer Science & Business Media, 2002.

[107] P. G. LeFloch. Graph solutions of nonlinear hyperbolic systems. *Journal of Hyperbolic Differential Equations*, 1(04):643–689, 2004.

[108] P. G. LeFloch and H. Makhlof. A geometry-preserving finite volume method for compressible fluids on Schwarzschild spacetime. *Communications in Computational Physics*, 15(3):827–852, 2014.

[109] P. G. LeFloch and S. Mishra. Numerical methods with controlled dissipation for small-scale dependent shocks. *Acta Numerica*, 23:743–816, 2014.

[110] P. G. Lefloch, C. Parés, and E. Pimentel-García. Well-balanced algorithms for relativistic fluids on a Schwarzschild background. *arXiv preprint arXiv:2011.07587*, 2020.

[111] P. G. Lefloch and M. D. Thanh. The Riemann problem for fluid flows in a nozzle with discontinuous cross-section. *Communications in Mathematical Sciences*, 1(4):763–797, 2003.

[112] P. G. LeFloch and M. D. Thanh. The Riemann problem for the shallow water equations with discontinuous topography. *Communications in Mathematical Sciences*, 5(4):865–885, 2007.

[113] P. G. LeFloch and M. D. Thanh. A Godunov-type method for the shallow water equations with discontinuous topography in the resonant regime. *Journal of Computational Physics*, 230(20):7631–7660, 2011.

[114] P. G. LeFloch and S. Xiang. A numerical study of the relativistic Burgers and Euler equations on a Schwarzschild black hole exterior. *Communications in Applied Mathematics and Computational Science*, 13(2):271–301, 2018.

[115] R. J. LeVeque. *Numerical methods for conservation laws*, volume 132. Springer, 1992.



[116] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.

[117] D. Levy, G. Puppo, and G. Russo. Central WENO schemes for hyperbolic systems of conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 33(3):547–571, 1999.

[118] D. Levy, G. Puppo, and G. Russo. Compact central WENO schemes for multidimensional conservation laws. *SIAM Journal on Scientific Computing*, 22(2):656–672, 2000.

[119] T.-P. Liu. The Riemann problem for general systems of conservation laws. *Journal of Differential Equations*, 18(1):218–234, 1975.

[120] T.-P. Liu. The deterministic version of the Glimm scheme. *Communications in Mathematical Physics*, 57(2):135–148, 1977.

[121] A. Marquina. Local piecewise hyperbolic reconstruction of numerical fluxes for nonlinear scalar conservation laws. *SIAM Journal on Scientific Computing*, 15(4):892–915, 1994.

[122] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography. *Computers & Mathematics with Applications*, 72(3):568–593, 2016.

[123] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography or Manning friction. *Journal of Computational Physics*, 335:115–154, 2017.

[124] T. Morales de Luna, M. J. Castro, and C. Parés. Reliability of first order numerical schemes for solving shallow water system over abrupt topography. *Applied Mathematics and Computation*, 219(17):9012–9032, 2013.

[125] T. Morales de Luna, M. J. Castro, and C. Parés. Relation between PVM schemes and simple Riemann solvers. *Numerical Methods for Partial Differential Equations*, 30(4):1315–1341, 2014.

[126] T. Morales de Luna, M J. Castro, C. Parés, and E. D. Fernández-Nieto. On a shallow water model for the simulation of turbidity currents. *Communications in Computational Physics*, 6(4):848–882, 2009.

[127] L. O. Müller, C. Parés, and E. F. Toro. Well-balanced high-order numerical schemes for one-dimensional blood flow in vessels with varying mechanical properties. *Journal of Computational Physics*, 242:53–85, 2013.

[128] S. T. Munkejord, S. Evje, and T. FlÅtten. A MUSTA scheme for a nonconservative two-fluid model. *SIAM Journal on Scientific Computing*, 31(4):2587–2622, 2009.

[129] M. L. Muñoz-Ruiz and C. Parés. Godunov method for nonconservative hyperbolic systems. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 41(1):169–185, 2007.

[130] C. D. Munz. On Godunov-type schemes for Lagrangian gas dynamics. *SIAM Journal on Numerical Analysis*, 31(1):17–42, 1994.

[131] S. Noelle, N. Pankratz, G. Puppo, and J. R. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *Journal of Computational Physics*, 213(2):474–499, 2006.

[132] C. Parés. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM Journal on Numerical Analysis*, 44(1):300–321, 2006.

[133] C. Parés and M. J. Castro. On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems. *ESAIM: M2AN*, 38(5):821–852, 2004.

[134] C. Parés and E. Pimentel-García. The Riemann problem for the shallow water equations with discontinuous topography: The wet–dry case. *Journal of Computational Physics*, 378:344–365, 2019.

[135] M. Pelanti, F. Bouchut, and A. Mangeney. A Roe-type scheme for two-phase shallow granular flows over variable topography. *ESAIM: Mathematical Modelling and Numerical Analysis*, 42(5):851–885, 2008.

[136] E. Pimentel-García, C. Parés, M. J. Castro, and J. Koellermeier. On the efficient implementation of PVM methods and simple Riemann solvers. Application to the Roe method for large hyperbolic systems. *Applied Mathematics and Computation*, 388:125544, 2021.

[137] E. Pimentel-García, M. J. Castro, C. Chalons, T. Morales de Luna, and C. Parés. In-cell discontinuous reconstruction path-conservative methods for non conservative hyperbolic systems – second-order extension. *arXiv preprint arXiv:2105.00424*, 2021.

[138] G. Puppo and M. Semplice. Well-balanced high order 1D schemes on non-uniform grids and entropy residuals. *Journal of Scientific Computing*, 66(3):1052–1076, 2016.

[139] J. Qiu and C.-W. Shu. Finite difference WENO schemes with Lax–Wendroff-type time discretizations. *SIAM Journal on Scientific Computing*, 24(6):2185–2198, 2003.

[140] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43(2):357 – 372, 1981.

[141] G. Rosatti and L. Begnudelli. The Riemann problem for the one-dimensional, free-surface shallow water equations with a bed step: Theoretical analysis and numerical simulations. *Journal of Computational Physics*, 229(3):760–787, 2010.

[142] J. A. Rossmanith, D. S. Bale, and R. J. LeVeque. A wave propagation algorithm for hyperbolic systems on curved manifolds. *Journal of Computational Physics*, 199(2):631–662, 2004.

[143] G. Russo. Central schemes for conservation laws with application to shallow water equations. In *Trends and Applications of Mathematics to Mechanics*, pages 225–246. Springer, 2005.

[144] G. Russo. High-order shock-capturing schemes for balance laws. In *Advanced Courses in Mathematics*, pages 59–147. Birkhäuser, 2009.

[145] G. Russo and A. Khe. High order well-balanced finite volume schemes for systems of balance laws. *Proceedings of Symposia in Applied Mathematics*, 2009.

[146] C. Sánchez-Linares, T. Morales de Luna, and M. J. Castro. A HLLC scheme for Ripa model. *Applied Mathematics and Computation*, 272:369–384, 2016.

[147] J. B. Schijf and J. C. Schonfled. Theoretical considerations on the motion of salt and fresh water. In *Proceedings Minnesota International Hydraulic Convention*, pages 321–333. IAHR, 1953.

[148] C.-W. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In *Advanced numerical approximation of nonlinear hyperbolic equations*, pages 325–432. Springer, 1998.

[149] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes, II. In *Upwind and High-Resolution Schemes*, pages 328–374. Springer, 1989.

[150] H. Struchtrup. *Macroscopic Transport Equations for Rarefied Gas Flows: Approximation Methods in Kinetic Theory*. Interaction of Mechanics and Mathematics. Springer Berlin Heidelberg, 2006.

[151] R. Temam. *Navier-Stokes equations: theory and numerical analysis*, volume 343. American Mathematical Society, 2001.

[152] V. A. Titarev and E. F. Toro. ADER schemes for three-dimensional non-linear hyperbolic systems. *Journal of Computational Physics*, 204(2):715–736, 2005.



[153] E. F. Toro. *Shock-capturing methods for free-surface shallow flows.* Wiley-Blackwell, 2001.

[154] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction.* Springer Science & Business Media, 2013.

[155] E. F. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the HLL-Riemann solver. *Shock waves*, 4(1):25–34, 1994.

[156] M. Torrilhon. Modeling nonequilibrium gas flow based on moment equations. *Annual Review of Fluid Mechanics*, 48(1):429–458, 2016.

[157] I. Toumi. A weak formulation of Roe's approximate Riemann solver. *Journal of Computational Physics*, 102(2):360–373, 1992.

[158] B. Van Leer. Towards the ultimate conservative difference scheme. II. Monotonicity and conservation combined in a second-order scheme. *Journal of Computational Physics*, 14(4):361–370, 1974.

[159] B. Van Leer. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. *Journal of Computational Physics*, 23(3):276–299, 1977.

[160] B. Van Leer. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. *Journal of Computational Physics*, 32(1):101–136, 1979.

[161] B. Van Leer. On the relation between the upwind-differencing schemes of Godunov, Engquist–Osher and Roe. *SIAM Journal on Scientific and Statistical Computing*, 5(1):1–20, 1984.

[162] M. E. Vázquez-Cendón. Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. *Journal of Computational Physics*, 148(2):497–526, 1999.

[163] A. I. Vol'pert. The spaces BV and quasilinear equations. *Matematicheskii Sbornik*, 115(2):255–302, 1967.

[164] X. Xia, Q. Liang, X. Ming, and J. Hou. An efficient and stable hydrodynamic model with novel source term discretization schemes for overland flow and flood simulations. *Water Resources Research*, 53(5):3730–3759, 2017.

[165] D. Zorío, A. Baeza, and P. Mulet. An approximate Lax–Wendroff-type procedure for high order accurate schemes for hyperbolic conservation laws. *Journal of Scientific Computing*, 71(1):246–273, 2017.