DEPARTAMENT MATEMÀTICA APLICADA

HIGH RESOLUTION SCHEMES FOR CONSERVATION LAWS WITH SOURCE TERMS

ANNA MARTÍNEZ GAVARA

UNIVERSITAT DE VALENCIA
Servei de Publicacions
2009

Aquesta Tesi Doctoral va ser presentada a València el dia 24 d'octubre de 2008 davant un tribunal format per:

- D. Tomás Chacón Rebollo
- Dª. Mª Inmaculada Higueras Sanz
- D. Guillaume Chiavassa
- D. Jacques Liandrat
- D. Pep Mulet Mestre

Va ser dirigida per:
Dª. Rosa M. Donat Beneito

Depòsit legal:
I.S.B.N.:978-84-370-7428-3
D.L.:V-1287-2009
Edita: Universitat de València
Servei de Publicacions
C/ Artes Gráficas, 13 bajo
46010 València
Spain
Telèfon: 963864115

PhD Thesis

High Resolution Schemes for Hyperbolic Conservation Laws with Source Terms

Anna Martínez Gavara

Advisor: Rosa Donat Beneito

Universitat de València
València, 2008.

High Resolution Schemes for Hyperbolic Conservation Laws with Source Terms

Memòria presentada per Anna Martínez Gavara, Llicenciada en Ciències Matemàtiques; realitzada al departament de Matemàtica Aplicada de la Universitat de València baix la Direcció de Rosa Donat Beneito, Catedràtica del mencionat Departament, amb l'objectiu d'aspirar al Grau de Doctora en Matemàtiques.

València, 24 d' Octubre de 2008

Rosa Donat Beneito
Directora de la Memòria

Anna Martínez Gavara
Aspirant al grau de Doctor

Departament de Matemàtica Aplicada
Facultat de Matemàtiques
Universitat de València

Agraïments

I ara,
em trobe ací . . .
asseguda vora la mar
escoltant son suau remor.
I em pose a pensar,
i al cap i a la fi,
la mar i jo no som tant diferents.
Després d'un temps revolicada,
la trobes tranquil·la,
on el silenci
sols és trencat per ses petites ones.
Ara és temps de tranquil·litat,
i també de reflexió
de tots els anys que han passat.

Si he de començar, no puc fer-ho sinó amb qui ha fet possible aquest projecte, ha confiat en mi i m'ha portat de la mà per aquest difícil camí de la investigació, Rosa Donat.

Els meus agraïments s'estenen als membres del grup d'investigació ANIMS, en especial a Pep Mulet qui m'ha donat consell i ajuda sempre que l'he necessitat.

També reconeixement al treball del professor Chiavassa, qui ha guiat una part de la investigació atenent tots els meus dubtes, i no dubtes, de manera magistral, per tot li done les gràcies.

Igualment a Inmaculada Higueras, professora de la Universitat Pública de Navarra, a qui mostre la meua gratitud per la seua contribució, des

de la distància i en les curtes estances en les que hem treballat juntes, a la present tesi.

I also want to thank the evaluators of this thesis, for their suggestions and advices.

A Toni, Jose i Ana per estar sempre quan els necessite, i per l'amistat que ens uneix.

Als membres del Departament de Matemàtica Aplicada pel seu tracte durant aquests anys.

Als meus amics i amigues, companys de doctorat i de dansa per tots els bons i grans moments que m'han donat.

Finalment, i com no, a les persones més importants de la meua vida, la meua família i Òscar.

I amb la mar en lluna plena,
em despedisc amb un gràcies.

València, 2008 *Anna*

A la meua família,

Potser si tanques fort els ulls veuràs
que la foscor se t'il·lumina.
Peus feixucs de vell recorren passadissos.
La casa és freda i els records també.
No t'assenyalis fites: fes camí
i espera que la barca toqui port
per escriure amb els dits, damunt la sorra,
el lluminós senyal del teu retorn.

Miquel Martí i Pol

Contents

Resum

Les lleis que estableixen la conservació de la massa, el moment i l'energia en un sistema físic es tradueixen en un sistema ben definit d'equacions en derivades parcials. En aquestes equacions, els efectes de les entrades, eixides, reaccions químiques i altres fenòmens d'interès es modelen per la inclusió de termes addicionals, generalment es refereixen com a *termes font*.

Aquesta memòria es dedica a l'estudi del tractament numèric dels termes font en lleis de conservació hiperbòliques i sistemes. En particular, estudiem dos tipus de situacions que són més delicades des del punt de vista de la seua aproximació numèrica: El cas de lleis de balanç, amb el sistema d'aigües poc fondes ("shallow water") com a principal exemple, i el cas d'equacions hiperbòliques amb terme font rígid ("stiff").

Actualment, n'hi han moltes tècniques que produeixen solucions numèriques amb precisió de lleis de conservació homogènies i sistemes. És ben conegut que les solucions poden ser discontinues, encara que la condició inicial siga perfectament suau. Tècniques estàndard com diferències finites, volums finits o elements finits tendeixen a produir aproximacions numèriques oscil·latòries quan l'ordre d'exactitud de l'esquema és major que u. Una gran quantitat d'investigacions en les últimes dècades ha desenvolupat en una tecnologia ben establida per a construir esquemes d'*alt d'ordre de captura dels xocs* (*High Resolution Shock Capturing*, HRSC). Aquests esquemes condueixen a resultats amb exactitut lluny de les discontinuïtats, així com, perfils monòtons, molt empinats en aquells llocs on es produeixen discontinuïtats en la solució.

D'altra banda, la solució numèrica d'equacions diferencials ordinàries (ODE) és una disciplina ben establida que ha produït una varietat de tècniques numèriques que són útils en moltes aplicacions.

Per tots aquests fets, una aproximació estàndard per a resoldre lleis

de conservació hiperbòliques és aplicar l'anomenada tècnica de passos fraccionats ("fractional steps"). Aquest procediment alterna entre resoldre una llei de conservació homogènia i resoldre una ODE que conté sols el terme font. No obstant, hi ha situacions on la proposta de passos fraccionats no condueix a aproximacions numèriques acceptables.

Quan calculem aproximacions numèriques a lleis de balanç, com les equacions d'aigües poc fondes, en situacions d'estat estacionari o quasi estacionari, les solucions numèriques de l'equació en derivades parcials (PDE) homogènia i l'ODE han d'equilibrar-se exactament. Aquest balanç exacte no es respecta amb la utilització del procediment de passos fraccionats, i poden ocórrer ones espúries de naturalesa numèrica.

Per a termes font rígids, la utilització d'un resolvedor d'ODE rígid amb un mètode HRSC pot conduir a solucions numèriques que pareixen raonables però són completament errònies. Aquest fenomen es va observar a principis de 1986 per Colella, Majda i Royburd en [19], en un problema model de combustió que involucra les equacions d'Euler de Dinàmica de Gasos junt amb una variable química que representa la porció de massa d'un gas no cremat en una ona de detonació. L'estructura de les ones de detonació obtinguda és ben coneguda, i s'observa que la solució numèrica que s'obté és qualitativament incorrecta quan es calcula amb malles grosses. Els termes font rígids poden descriure també models hidrodinàmics per a semiconductors elàstics amb memòria, ones d'aigua, circulació de tràfic, etc.

En aquests darrers anys hi ha aparegut una gran quantitat de literatura dedicada a problemes numèrics que poden ocórrer en aquests dos tipus de situacions.

Per a les equacions aigües poc fondes, molts autors estenen el clàssic resolvedor de Riemann de Roe per a problemes no homogenis relacionats amb lleis de balanç [5], [7], [17], [58], [104]. En aquests treballs, la forma discreta dels termes font és construeix en una forma similar a aquella utilitzada per a la construcció dels fluxos numèrics, buscant l'equilibri que existeix per a la construcció d'estats estacionaris de lleis de conservació amb termes font. La idea de "source-term upwinding" (terme font en la direcció del vent) permeteix a Bermúdez i Vázquez-Cendón [5] a formular l'anomenada propietat-C (propietat de Conservació) per a un esquema numèric, la qual prevé la propagació d'ones paràsits en fluids estacionaris i quasi estacionaris. Independentment, Greenberg i Leroux tragueren el terme ben-balancejat ("well-balance") per a esquemes que mantenen els estats estacionaris a nivell discret. Aquestes idees han sigut explorades i desenvolupades per a aigües poc fondes en la recent literatura [4], [24], [42], [61], [71], [78], [106] ...

En aquest treball, seguim l'estratègia descrita per Gascón i Corberán en [38] i Donat, Caselles i Haro en [10]. En [38], els autors proposen escriure el terme font en la forma de la divergència, i per això el problema no homogeni es 'transforma' en 'forma homogènia' a través de la nova definició de una nova funció flux. Aquest canvi busca mantindre el balanç del terme font i el flux en els estats estacionaris d'una forma quasi automàtica, i suggereix una manera d'aplicar esquemes per a lleis de conservació homogenis coneguts al cas no homogeni. No obstant, com ells també observen, l'aplicació per a mètodes numèrics per al cas homogeni no és immediata i requereix d'una formalització adequada.

En [10], la idea d'equilibri entre el flux gradient i els termes font en [38] s'incorpora en l'esquema numèrica desenvolupat per Donat i Marquina en [26], llavors s'estén efectivament aquest esquema a lleis de balanç.

En aquest treball, ens centren en els fonaments teòrics d'esquemes d'alta resolució amb variació total decreixent ("Total Variation Diminishing", TVD), per a lleis de conservació escalars homogènies, fermament establides a través del treball de Harten [50], Sweby [95], i Roe [80] i analitzar les propietats del segon ordre, versió de l'esquema de Lax-Wendroff amb flux-limitat que evita les oscil·lacions al voltant de les discontinuïtats, mentre que manté els estats estacionaris [38]. Quan s'aplica a les lleis de conservació homogènies, els esquemes TVD prevenen un increment en la variació total de la solució numèrica, mentre que garanteix l'absència de generació d'oscil·lacions numèriques. Aquests estan implementats amb èxit en la forma de limitador de fluxos o pendent limitada per a lleis de conservació escalar i sistemes. La nostra tècnica es basa en un procediment de limitador de flux aplicat sols a aquells termes relacionats en la derivada/Jacobià del fluid físic.

Respecte del tractament numèric dels termes font rígids, seguim a LeVeque i Yee en [73]. Agafant el senzill problema model considerat en [73], estudiem les propietats de la solució numèrica obtinguda amb diferents tècniques. Som capaços d'identificar el *factor de retard* ("*delay factor*"), que és responsable de la velocitat anòmala de propagació de la solució numèrica en malles grosses. El retard es deu a la introducció de valors que no estan en equilibri a través de dissipació numèrica, i sols poden controlar-se amb una adequada reducció de la resolució espacial de la simulació. Els esquemes explícits pateixen de la mateixa patologia numèrica, inclòs després de reduir el pas temporal i estar satisfets els requeriments d'estabilitat imposats per les escales més ràpides. Estudiem el comportament de tècniques numèriques Implícites-Explícites (IMEX), com a ferramenta per a obtindre simulacions d'alta resolució que

incorporen els termes font rígid en una forma implícita, sistemàtica. La tècnica IMEX s'ha aplicat amb èxit a sistemes hiperbòlics amb relaxació (veure [9], [110], [73]).

Normalment, quan utilitzem malles uniformes i fines, trobem que el temps computacional és el major entrebanc en la simulació numèrica. Per a alguns esquemes d'alta resolució de captura de xocs, simulacions en malles fines en dos dimensions estan fora de l'abast simplement perquè costen massa. Les avaluacions del flux numèric són massa cares. No obstant, el flux computacional és necessari sols perquè es desenvolupen espontàniament en la solució d'un sistema hiperbòlic de lleis de conservació, estructures no homogènies que evolucionen en temps, la qual cosa condueix a crear tècniques de reducció del cost computacional associat a aquestes simulacions. Harten en [49] proposa un esquema basat en reduir el cost computacional utilitzant la informació de suavitat de les dades, i reemplaçant el car flux numèric amb una barata interpolació polinòmica en les regions de suavitat. La clau és l'ús d'una estratègia de diferents multi-nivells per a reduir l'esforç computacional associat a l'esquema de HRSC. En les regions suaus, Harten en [49] proposa avaluar la funció del flux numèric sols en la malla més fina utilitzant un barat procediment d'interpolació polinòmica en forma multi-nivell. Ací, estenem la tècnica desenvolupada en [16] per a lleis de conservació amb termes font i apliquem la tècnica multi-nivell a un sistema hiperbòlic d'aigües poc fondes.

Esquema de la Tesi

Aquesta tesi està organitzada de la següent forma:

Capítol 1: Donem una introducció a les lleis de conservació i els problemes que presenten les aproximacions numèriques. S'introdueixen els conceptes de conservació, solucions entròpiques, monotonia i variació total entre d'altres, així com una selecció de mètodes numèrics clàssics.

Capítol 2: Repassem un dels mètodes bàsics per a resoldre una llei de conservació no homogènia.

Capítol 3: Estudiem la estabilitat i la velocitat de l'ona per a esquemes explícits, implícits i semi-implícits per a lleis de conservació hiperbòlics amb terme font rígid.

Capítol 4: Proposem una tècnica basada en el procediment de limitació de flux aplicada sols a aquells termes relacionats amb la derivada/Jacobià del flux físic, que evita oscil·lacions al voltant de les discontinuïtats, mantenint l'estat estacionari.

Capítol 5: Apliquem la tècnica multi-nivell al sistema d'aigües someres, on la base de l'esquema és el mètode presentat en el capítol anterior i l'esquema de [10].

Apèndix A: Revisem alguns aspectes teòrics per a les equacions diferencials ordinàries i esquemes Runge-Kutta.

Apèndix B: Repassem aspectes teòrics de les equacions d'aigües poc fones i les hipòtesi de la seua derivació a partir del model de les equacions de Navier-Stokes. Aleshores, recordem el problema de Riemann per a dos estats estacionaris adjacents així com les seues possibles solucions: xocs i rarefaccions.

Capítol 1
Formulació general de lleis de conservació

La forma general d'un sistema de lleis de conservació, incloent els termes font, que considerarem en aquesta memòria és de la forma

$$\mathbf{u}_t + \nabla \cdot \mathbf{f}(\mathbf{u}) = \mathbf{s}(\mathbf{x}, \mathbf{u}), \tag{1}$$

on $\mathbf{u} \in \mathbb{R}^m$ representa el vector d'incògnites (les *variables d'estat*), $\mathbf{f}(\mathbf{u}) \in \mathbb{R}^m$ és el vector flux, $\mathbf{x} \in \mathbb{R}^n$ i $t \in \mathbb{R}^+$ són les variables independents i $\mathbf{s} \in \mathbb{R}^m$ és la funció del terme font.

En el capítol 1 preliminar, revisem alguns aspectes teòrics i numèrics de les lleis de conservació homogènies. Els resultats teòrics més rellevants recollits en aquest capítol es troben en els treballs [36], [68], [69], [39], [72], [70], [90], [100], [108]. Per a temes relacionats amb tècniques numèriques, les principals fonts de referència són els treballs de [50], [82], [94], [95].

Les lleis de conservació homogènies són un important subconjunt de (1). Ens concentrarem majoritàriament, en aquesta memòria, en el cas unidimensional, $n = 1$. La forma general del cas homogeni d'una dimensió és, doncs

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0. \tag{2}$$

El sistema (2) és hiperbòlic si la matriu Jacobiana

$$A(\mathbf{u}) = \frac{\partial \mathbf{f}(\mathbf{u})}{\partial \mathbf{u}} \tag{3}$$

té m valors propis reals i m vectors propis linealment independents.

Per a aquests sistemes, estudiem el *problema de Cauchy*, o el problema de valors inicials: Trobar una funció $\mathbf{u} : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^m$ que és solució de (2) satisfent la condició inicial

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x), \tag{4}$$

on $\mathbf{u}_0$ és una funció donada. Quan $\mathbf{u}_0$ té la següent forma particular,

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_l, & x < 0 \\ \mathbf{u}_r, & x > 0, \end{cases} \tag{5}$$

el problema de Cauchy s'anomena *problema de Riemann*.

Per a sistemes homogenis de lleis de conservació, com (2), si formalment integrem respecte de la variable d'espai en $\mathbb{R}$ i assumim que els valors del vector d'estat a $\pm\infty$, els denotem com $\mathbf{u}|_{\pm\infty}$, estan ben definits, obtenim

$$\frac{d}{dt} \int_{-\infty}^{\infty} \mathbf{u}_i(x, t) dx = - \left(f_i(\mathbf{u}|_{+\infty}) - f_i(\mathbf{u}|_{-\infty}) \right) \tag{6}$$

llavors, quan $f(\mathbf{u}|_{+\infty}) = f(\mathbf{u}|_{-\infty})$ la integral de la funció de densitat de cada variable d'estat és constant respecte del temps, és a dir, es *conserva*, encara que la distribució espacial de $\mathbf{u}_i$ és lliure d'evolucionar en temps. És aquesta evolució que volem modelitzar utilitzant tècniques discretes.

Per a resoldre numèricament les equacions diferencials que provenen d'una llei de conservació, és necessari reemplaçar el problema continuu a un problema discret en una malla fina. Normalment es pot realitzar, entre d'altres, de dues formes diferents. Per a una aproximació en *diferències finites*, els valors de les quantitats conservades es calculen com a valors puntuals en les interseccions de la malla, utilitzant aproximacions de la forma diferencial de la llei de conservació. En un procediment de *volums finits* s'utilitza la forma integral de la llei de conservació i les quantitats són mitges en cel·la. La formulació en diferències finites és la que ens interessa ací.

La solució numèrica d'equacions no lineals de la forma (2), afegeix problemes addicionals que són, en general, molt més durs d'analitzar que en problemes lineals. En molts casos, és necessari linealitzar el problema abans de realitzar qualsevol anàlisi útil. No obstant, poden ocórrer inestabilitats en la solució, inclòs quan la versió linealitzada és

estable. Si l'esquema convergeix, pot convergir a una solució feble que viola l'entropia, o pitjor encara, pot convergir a una funció que no és una solució feble de l'equació diferencial original [70]. Per a dissenyar esquemes robusts per a equacions no lineals s'han de tractar cadascun dels temes exposats en el capítol 1.

En la secció 1.5, revisem els esquemes bàsics per a resoldre lleis de conservació en una dimensió, prestant especial atenció a aquells que s'utilitzaran després en aquesta tesi. Per a una més completa descripció, us adressem a [72], [70] i [100].

En general, els esquemes es construeixen com segueix: Discretitzem el pla $x-t$ amb una malla de grossària $\triangle x$ i un pas de temps $\triangle t$, i definim els punts discrets en la malla (x_i, t_n) per

$$\begin{aligned} x_i &= i\triangle x, \quad i = \ldots, -1, 0, 1, \ldots \\ t_n &= n\triangle t, \quad n = 0, 1, 2, \ldots \end{aligned}$$

Els mètodes en diferències finites produeixen aproximacions U_i^n a la solució $u(x_i, t_n)$ en els punts discrets de la malla. Denotem per $U^n = (U_i^n)_i$. Considerem un esquema de dos nivells que es pot escriure en forma general com

$$U_i^{n+1} = H(U_{i-p}^n, \ldots, U_{i+q}^n) \tag{7}$$

on H és l'operador de la solució discret, i p, q són constants positives.

Capítol 2

Esquemes numèrics per a lleis de conservació no homogènies

En aquest capítol estem interessats en la solució numèrica d'equacions hiperbòliques, en particular equacions de la forma

$$u_t + f(u)_x = s(x, u) \tag{8}$$

amb condició inicial $u(x, 0) = u_0(x)$.

El terme font s és una funció de x i $u(x, t)$. Comparant amb el cas homogeni presentat en el primer capítol, apareixen dues dificultats. La solució, u, no és necessàriament constant al llarg de les característiques de l'equació i a més, la pendent de les característiques canvia. De fet, $u(x, t)$ satisfà

$$\frac{du}{dt} = s(x, u), \tag{9}$$

al llarg dels camins

$$\frac{dx}{dt} = \frac{df}{du}(u) \qquad x(0) = x_0. \tag{10}$$

La pendent dels camins depèn d'u (veure (10)) i no és necessàriament constant, ja què u no és constant al llarg de les característiques, (9).

Se sap que la solució feble no és necessàriament única, i que la solució física es caracteritza per la següent condició d'entropia [65]:

$$\int_{-\infty}^{\infty}\int_{0}^{T} \left(\eta(u)\phi_t + F(u)\phi_x\right) dxdt \geq -\int_{-\infty}^{\infty} \eta'(u)s(x,u)\phi(x,t)dxdt, \tag{11}$$

on $\phi \in \mathcal{C}^1(\mathbb{R}\times(0,T))$ és una funció test positiva amb suport compacte en $\mathbb{R}\times(0,T)$, i $\eta \in \mathcal{C}^2(\mathbb{R})$ és una funció d'entropia estrictament convexa, amb la corresponent funció flux d'entropia F, que és

$$\eta'(u)f'(u) = F'(u) \quad \forall u \in \mathbb{R}. \tag{12}$$

Kružkov en [65] prova els primers resultats d'existència i unicitat per al cas no homogeni (8). Molts autors han estudiat situacions particulars. Recordarem ací un resultat particular, estret de [43], i el qual és rellevant per al problema model considerat en el capítol 3, on $s(x,u) = s(u)$.

Suposem que es donen les següents condicions

1. f i s són funcions suaus en $\mathcal{C}^1(\mathbb{R})$,

2. $s(0) = 0$,

3. per a evitar el fenòmen d'amplitud d'explosió ("blow-up"), assumim que

$$\exists M \in \mathbb{R}^+ \text{ tal que } |u| \geq M \Longrightarrow u \cdot s(u) \leq 0.$$

Theorem 1. *([65]) Considerem el problema de Cauchy per a*

$$u_t + f_x = s(u), \qquad x \in \mathbb{R}, t > 0,$$

on es verifiquen les condicions 1-3. $\gamma = \max(s'(u))$ *és un nombre finit. Aleshores, per a* $u_0 \in \mathcal{L}^1(\mathbb{R}) \cap \mathcal{BV}(\mathbb{R})$ *existeix una única solució entròpica del problema de Cauchy amb condició inicial* $u(\cdot, 0) = u_0$ *satisfent que*

1. $||u||_{\mathcal{L}^\infty} \leq \max(||u_0||_{\mathcal{L}^\infty}, M)$
2. $TV_x(u(\cdot, t)) \leq e^{\gamma t} TV_x(u_0)$
3. *Donats els valors inicials* u_0, v_0 *tal que* $u_0 \leq v_0$*, les corresponents solucions entròpiques* $u(x,t)$ *i* $v(x,t)$ *satisfan*

$$u(x,t) \leq v(x,t). \tag{13}$$

4. *Donats els valors inicials* u_0, v_0 *aleshores les solucions corresponents satisfan*

$$\|u(x,t) - v(x,t)\|_{\mathcal{L}^1(\mathbb{R})} \leq e^{\gamma t} \|u_0 - v_0\|_{\mathcal{L}^1(\mathbb{R})}. \tag{14}$$

Més detalls en l'existència, unicitat i algunes propietats de les solucions es poden trobar en [2] ,[15], [48], [65], [74], [75].

En aquest capítol, tractem alguns dels mètodes numèrics que es poden utilitzar per a aproximar numèricament l'equació (8). Hi han varies formes de tractar el terme font, que es classifiquen en dos categories:

- Mètodes no-separats ("unsplit"), en els quals una única fórmula en diferències finites es desenvolupa per a avançar tota l'equació en un pas de temps.
- Mètodes de passos fraccionats, en els quals el problema es trenca en dues peces corresponents a diferents processos, i on s'utilitza un mètode numèric apropiat per a cadascuna de les peces independentment. Aquesta aproximació també s'utilitza sovint per a separar problemes multi-dimensionals en una seqüència de problemes unidimensionals. Hi ha situacions on aquesta tècnica porta a solucions numèriques espúries o inclús errònies. Com és el cas dels termes font rígids o problemes quasi estacionaris, com és veu en el capítol.

Capítol 3

Esquemes numèrics per a lleis de conservació escalar amb terme font rígid

Molts problemes físics estan governats per lleis de conservació hiperbòliques amb terme font rígid que no s'esvaeix. Aquests problemes descriuen l'efecte de relaxació com en la teoria cinètica de gasos, reaccions químiques, elasticitat amb memòria, ones d'aigua, circulació del tràfic, etc.

En alguns problemes els termes font depenen sols de la solució, és a dir, $s(x,u) = s(u)$ i encara que de forma natural la solució genera estructures on els termes font no són zero, i possiblement grans, sols en una petita regió de l'espai. Açò passa sovint si els termes font modelitzen reaccions químiques entre diferents espècies, en casos on les reaccions passen en escales de temps molt més ràpides que l'escala de temps del fluid dinàmic. Aleshores, les solucions poden desenvolupar en zones de reacció fines on es concentra l'activitat química-cinètica. Aquests problemes es solen anomenar que tenen termes font rígid, en analogia al cas clàssic de equacions diferencials ordinàries rígides.

Dificultats numèriques solen passar quan les reaccions ràpides són a prop de l'equilibri durant la majoria de la computació. Algunes escales de temps, típicament aquelles que condueixen a termes de reacció, són d'alguns ordres de magnitud més ràpides que les escales on la solució evoluciona i en la qual volem calcular. Amb molts mètodes numèrics, incloguen tots els mètodes explícits, agafant un pas temporal apropiat per a les escales més baixes d'interès pot resultar en una violenta inestabilitat numèrica, causada per les escales més ràpides.

L'estabilitat, comprés com l'absència d'un comportament oscil·latori violent, es pot abastir utilitzant mètodes implícits. Una gran varietat d'excel·lents mètodes implícits s'ha desenvolupat per a resoldre sistemes d'ODEs rígides, i moltes de les mateixes tècniques es poden aplicar quan els termes font són rígids per a obtindre resultats estables.

No obstant, es troben diferents dificultats numèriques en la simulació de PDEs hiperbòliques amb terme font rígid: l'aparició del front de propagació amb velocitats errònies [19].

L'anàlisi que es fa en aquest capítol s'interessa en el problema model introduït per LeVeque and Yee en [73], on un problema prototip de valors

inicials (IVP) de la forma

$$u_t + u_x = s(u) \qquad x \in \mathbb{R} \qquad t > 0, \tag{15}$$

$$u(x,0) = u_0(x) \qquad x \in \mathbb{R}. \tag{16}$$

amb un terme font amb un paràmetre dependent ($s(u) = -\mu u(u-1)(u-0.5)$), s'utilitza en l'estudi del comportament de mètodes numèrics respecte a aquest fenomen patològic. LeVeque i Yee realitzen un estudi numèric utilitzant dos tipus de tècniques discretes: Una extensió semi-implícita del mètode predictor-corrector de MacCormak, on el fluid dinàmic i el químic es manipulen simultàniament, i una aproximació de temps fraccionat, on s'alterna entre la solució de la llei de conservació i l'ODE representant la química. En ambdós casos, observen que, per a termes de reacció rígids, és possible obtindre solucions estables on el perfil de l'ona numèrica pareix raonable però viatja a una velocitat errònia. En [73], s'argumenta que aquesta patologia es degut a la introducció de valors que no estan en equilibri a través de la dissipació numèrica en el pas d'advecció.

Ahmad i Berzins en [1] també consideren el mateix problema model i utilitzen una discretització pel Mètode de Línies (MOL) utilitzant esquemes que preserven la monotonia per al pas d'advecció en un marc adaptatiu en l'espai en lloc de considerar un esquema numèric eficient.

El mètode de linies és un procediment de discretització estàndard quan es dissenyen esquemes d'alta resolució de captura de xocs per a problemes de convecció.

En aquest capítol, considerem una aproximació MOL en una malla amb tamany fixe, on utilitzem un esquema que preserva la monotonia per als termes d'advecció, i considerem esquemes explícits, totalment implícits i semi-implícits que marxen en temps.

Analitzem els esquemes numèrics construïts respecte la seua habilitat de produir perfils numèrics lliures d'oscil·lacions, i establim condicions en els paràmetres de discretització per a obtenir perfils numèrics no-oscil·latoris per al problema model per als diferents esquemes numèrics considerats.

El nostre estudi per als esquemes totalment explícits i implícits mostren que hi ha, en efecte, una relació directa entre la malla espacial i el retard numèric. Aquest retard és, efectivament, un producte de la discretització considerada que no pot ser evitat.

Els esquemes que marxen en temps semi-implícits considerats en aquest capítol són els esquemes Runge-Kutta Implícit-Explícit (IMEX), també considerats per a problemes de relaxació rígids per Pareschi i

Russo [79]. L'habilitat de tractar la part convectiva de forma explícita, mentre que es manté un tractament implícit dels termes font, dóna un avantatge clar quan es construeixen esquemes d'alt ordre, d'alta resolució numèrica. L'estudi de les propietats numèriques d'aquests esquemes, quan s'apliquen al problema model ens porta al concepte d'esquemes que *Preserven l'estabilitat feble* ("*Weak Stability Preserving*"). Aquests són esquemes que preserven una propietat no-oscil·latòria feble en la solució numèrica donat que la mateixa propietat es manté per als operadors de discretització temporal bàsics involucrats. Les propietats d'aquests esquemes amb respecte del retard numèric són absolutament similars per a esquemes de primer ordre.

En aquest capítol, s'observa que el retard o avançament del perfil discret de l'ona en una discretització MOL està controlada bàsicament per $\mu\Delta x$. Per al senzill, primer ordre, discretització temporal, s'obté un retard discret de la forma

$$\alpha(\mu\triangle x, U^n) = \mu\triangle x \sum_{i=1}^{N} U_i^n(U_i^n - 1)(U_i^n - \frac{1}{2}). \tag{17}$$

Notem que, en el cas estudiat en aquest capítol, el valor de

$$U_i^n(U_i^n - 1)(U_i^n - \frac{1}{2})$$

és quasi sempre zero, excepte per als punts que formen la discontinuïtat. Perquè per la forma del terme font d'aquest problema, són possibles valors negatius i positius, per això la suma,

$$\sum_{i=1}^{N} U_i^n(U_i^n - 1)(U_i^n - \frac{1}{2})$$

la qual és normalment petita, pot tindre qualsevol signe, i el perfil discret pot avançar, o retrassar-se, amb respecte del perfil real.

Els resultats que es mostren en aquest capítol indiquen que, baix condicions apropiades, els esquemes IMEX es poden utilitzar per a obtindre perfils de l'ona monòtons per al problema model. Aquests perfils es menegen a una velocitat que està també directament relacionada amb el paràmetre $\mu\Delta x$, pel que és necessari suficient refinament per a obtindre l'ona ben col·locada.

Encara que es mostren resultats numèrics per a una discretització de primer ordre "upwind" per al terme convectiu, el comportament obtingut

en aquest cas és típicament el mateix que s'obtindria per a discretitzacions conservatives més complicades per a la solució del problema de valors inicials model. L'avantatge d'una discretització "upwind" de primer ordre és que ens permet un anàlisi sistemàtic de la solució numèrica, la qual ens permet establir certes fites, per a propòsit pràctic, que asseguren la monotonia dels perfils numèrics.

Capítol 4

Esquema de segon ordre de flux-limitat per lleis de balanç

Els esquemes d'alta resolució, com aquells descrits en el capítol 1, s'ha provat que capturen les ones de xocs i les solucions discontinues de lleis de conservació escalar amb èxit. No obstant, com s'explica en el capítol 2, quan s'aplica a lleis de balanç com les equacions d'aigües poc fondes, s'ha d'anar en compte en la tècnica numèrica.

Per a lleis de balanç, és normalment necessari calcular amb precisió solucions en estat estacionari, o quasi estacionari, per a les quals el flux gradient no és zero però s'equilibra amb el terme font. S'observa en el capítol 2 que molts mètodes numèrics no respecten aquest equilibri i, per tant, no es poden utilitzar per a calcular amb precisió petites pertorbacions de solucions de l'estat estacionari.

Els mètodes numèrics que respecten aquest equilibri que ocorreix en fluids estacionaris s'anomenen *ben balancejats* ("*well balanced*"), en el treball de Leroux i col·laboradors [46], [47]. Independentment, Bermúdez and Vázquez-Cendón [5] introduiren el concepte de *propietat C* ("*C-property*") que garanteix que l'esquema és well-balanced. La falta d'un equilibri apropiat, a nivell discret, entre els efectes dels fluxos numèrics i els termes font condueix a un comportament oscil·latori.

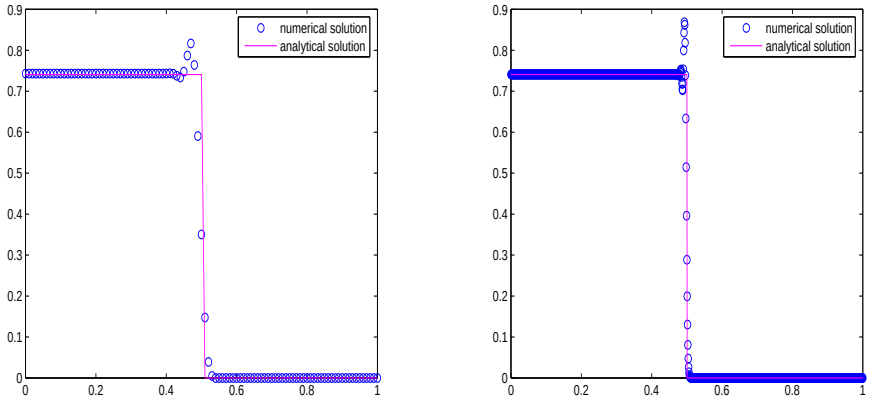
Quan intentem dissenyar esquemes d'alta resolució per a lleis de conservació no homogènies, el bon equilibri és una qüestió important que s'ha de tenir en compte. Les oscil·lacions també poden apareixer com a conseqüència de la utilització d'un esquema de segon ordre, la qual és una ben coneguda deficiència dels esquemes de segon ordre en el cas homogeni. La figura 1 mostra una simulació numèrica per al senzill cas model

$$u_t + u_x = -u \qquad u(x,0) = \begin{cases} 1, & x \leq 0.2 \\ 0, & x > 0.2, \end{cases}$$

obtés amb la següent extensió del mètode de MacCormack, modificada per a incloure els termes font en una forma explícita mentre que manté el segon ordre [73], [107]:

$$\begin{aligned}
\Delta U_i^{(1)} &= -\frac{\triangle t}{\triangle x}\left(f(U_i^n) - f(U_{i-1}^n)\right) + \triangle t s(U_i^n) \\
U_i^{(1)} &= U_i^n + \Delta U_i^{(1)} \\
\Delta U_i^{(2)} &= -\frac{\triangle t}{\triangle x}\left(f(U_i^{(1)}) - f(U_{i-1}^{(1)})\right) + \triangle t s(U_i^{(1)}) \\
U_i^{n+1} &= U_i^n + \frac{1}{2}\left(\Delta U_i^{(1)} + \Delta U_i^{(2)}\right).
\end{aligned}$$

En la figura 1, observem oscil·lacions espúries, típiques d'esquemes de dades independents de segon ordre, els quals no disminueixen amb un refinament de la malla. Notem que són del mateix tipus que s'observen en el cas homogeni.



$\triangle x = 0.01$ $\triangle x = 0.001$

Figure 1: *Second order scheme applied to* $u_t + u_x = -u$ *with CFL*$= 0.9$.

En [38] es fa una observació similar, per a un esquema de segon ordre, de dos pasos, que estén l'esquema de Lax-Wendroff per al cas no homogeni de lleis de conservació.

En [38], Gascón and Corberán busquen definir un esquema de segon ordre no oscil·latori per a lleis de conservació no homogènies, estudiant quines condicions han d'imposar-se per a obtindre esquemes TVD per a lleis de conservació hiperbòliques no homogènies. En [38], els autors presenten una tècnica basada en la transformació del problema no homogènia a *forma homogènia* a través de la definició d'un nou flux format

pel flux físic i la primitiva del terme font. Aquesta tècnica simplifica els requeriments per al ben-balancejat i suggereix una via per a aplicar l'existent tecnologia en esquemes TVD per al cas no homogeni.

Aquest és el punt inicial del capítol. Mentre els resultats presentats per Gascón i Corberán en [38], el qual tracta principalment en simulacions numèriques d'un fluid en un conducte amb variable àrea de la secció, mostren bones propietats numèriques per a l'esquema de segon ordre proposat, nosaltres creiem que hi han alguns punts bàsics en el disseny que impedeixen que aquest esquema es convertisca en una veritable tècnica general per a simulacions numèriques que involucren lleis de balanç.

Observem en el capítol 2, que l'ús d'esquemes TVD per a lleis de conservació escalar està justificat pel fet que la solució vertadera (entròpica) d'una llei de conservació escalar també satisfà la propietat TVD. No obstant, per a lleis de balanç escalars la propietat TVD de la solució no es manté gaire, veure el Teorema 1. Encara que certs termes font preserven la propietat TVD de la part homogènia, altres incrementaran la variació de la solució. Aquesta observació fa que posem atenció al cas escalar, on revisem, i modifiquem, el disseny bàsic principal de l'esquema de Gascón i Corberán.

En general, el concepte de TVD i les idees bàsiques són encara útils, donat que la restricció TVD ens du a solucions no oscil·latòries.

En aquest capítol, desenvolupem dos esquemes de segon ordre de flux limitat per a lleis de conservació no homogènies. La derivació d'aquests esquemes segueix la tècnica de limitador de flux que es veu en el capítol 1, i el seu disseny bàsic es basa en el que pensem són les flebeses essencials en el desenvolupament de Gascón i Corberán en [38]. El comportament de l'esquema és analitzat a través d'un conjunt d'experiments numèrics, en particular amb propietats de respectar un bon equilibri. L'esquema més robust s'estén a un sistema d'aigües poc fondes, a través d'una aproximació local característica.

És important remarcar que, com s'observa en [10], la nostra tècnica no es basa en convertir la llei de balanç en una llei de conservació homogènia. La definició de $b(x,t) = \int^x -s(y,u(y,t))dy$ permeteix *expressar* la llei de balanç en *forma homogènia*. No obstant, açò sols s'utilitza per a obtindre un tractament adequat del terme font. Com en [10], la nostra tècnica numèrica utilitza sols les velocitats característiques que provenen del terme convectiu i, com en [10], sempre obtenim la solució entròpica correcta en tots els nostres experiments.

Capítol 5

Un esquema multiescala per a sistemes de lleis de balanç

La simulació numèrica de problemes físics que es modelen per sistemes de lleis de conservació és difícil per la presència de discontinuïtats en la solució. Esquemes d'alt ordre de captura de xocs (HRSC) obtenen solucions numèriques amb alt ordre de precisió, normalment segon o tercer ordre en les regions suaus, mentre que mantenen perfils numèrics definits, sense oscil·lacions a les discontinuïtats. El poder dels esquemes HRSC rau normalment en el càlcul de la funció de flux numèric que normalment és car. Aquest és, de fet, el major desavantatge d'aquests esquemes, especialment en càlculs multidimensionals.

És ben conegut, no obstant, que el car càlcul del flux numèric sols és necessària al voltant de les singularitats, així, Harten en [49] proposa diversos esquemes multiescala basats en reduir aquest cost utilitzant la informació de suavitat d'una transformació de multiresolució de les dades. L'objectiu és guanyar temps en l'avaluació de les funcions de flux numèric, mentre que mantindre l'exactitud d'alta resolució de l'esquema. Aquest s'abasta reemplaçant les cares avaluacions del flux numèric per una barata interpolació polinòmica en les regions suaus.

L'estratègia multinivell original de Harten, va ser desenvolupada per a esquemes de volums finits, on les dades numèriques són tractades de forma natural com aproximacions a les mitges en cel·la de la solució vertadera. Chiavassa i Donat en [16] apliquen la mateixa estratègia de reducció del cost per a esquemes de diferències finites, on les dades són interpretades com a valors puntuals.

En aquest capítol, estenem la tècnica desenvolupada en [16] per a sistemes de lleis de conservació no homogenis. En particular, investiguem les propietats de l'esquema estès, en termes d'eficiència i qualitat, per una sèrie d'experiments numèrics en el sistema d'aigües poc fondes.

Les equacions d'aigües poc fondes en una i dos dimensions s'utilitzen per a modelitzar situacions de la vida real. En moltes ocasions, el fluid es troba en estat estacionari o quasi-estacionari, i es necessita molt d'esforç per a dissenyar tècniques numèriques que siguen capaces de preservar aquests estats a nivell discret això com calcular amb precisió l'evolució de petites pertorbacions d'aquests estats estacionaris.

En aquest capítol ens centrem en els següents esquemes

- L'esquema TVDB presentat en el capítol 4, que utilitza una tècnica de limitació de fluxos.
- L'esquema HRSC proposat en [10].

Aquest dos esquemes tenen formules de flux numèric que són significativament diferents en termes d'esforç computacional. L'esquema TVDB utilitza una linealització de Roe i involucra una avaluació d'un Jacobià (1J) per cel·la, mentre que l'esquema proposat en [10] combina la utilització d'avaluacions de dos Jacobians per cel·la (1J-2J), essent aquesta última més cara. Aquest fet, s'utilitza per a investigar les propietats de la tècnica multinivell.

En primer lloc, es resumeixen els passos principals de l'algorisme d'1D, i es realitzen alguns experiments numèrics en ben coneguts problemes test. Veiem que l'esquema multinivell preserva les propietats fonamentals de l'esquema base, com mantindre la propietat C. Finalment, presentem l'algorisme de 2D i tests preliminars.

Els resultats que s'observen, mostren que l'estratègia multinivell condueix a una ferramenta eficient per a la simulació numèrica de sistemes de lleis de balanç, especialment aquelles situacions on es necessita alta qualitat i alta resolució en un cost raonable.

Introduction

The laws establishing the conservation of mass, momentum and energy in a physical system translate into a well defined system of partial differential equations. In these equations, the effect of sources, sinks, chemical reactions and other phenomena of interest are modeled by the inclusion of additional terms, which are generically referred to as *source terms*.

This memoir is devoted to the study of the numerical treatment of source terms in hyperbolic conservation laws and systems. In particular, we study two types of situations that are particularly delicate from the point of view of their numerical approximation: The case of balance laws, with the shallow water system as the main example, and the case of hyperbolic equations with stiff source terms.

There are nowadays many techniques that can produce accurate numerical solutions of *homogeneous* conservation laws and systems. It is well known that these solutions can be discontinuous, even when the initial data in a Cauchy problem is perfectly smooth. Standard (linear, data-independent) finite-difference, finite-volume or finite-element techniques tend to produce oscillatory numerical approximations when the order of accuracy of the scheme is larger than one. A large amount of research in the last decades has resulted in a well established technology to construct *High Resolution Shock Capturing* (HRSC henceforth) schemes. These schemes lead to accurate results away from discontinuities, as well as well defined, very steep, monotone profiles at those locations where discontinuities in the solution do occur.

On the other hand, the numerical solution of Ordinary Differential Equations (ODE) is a well established discipline that has produced a variety of numerical techniques that have proved to be useful in many applications.

Because of these facts, a standard approach to solve hyperbolic conservation laws with source terms is to apply the so-called *fractional step approach*. This technique alternates between solving a homogeneous conservation law and solving an ODE that contains only the source term. However, there are situations where the fractional-step approach does not lead to acceptable numerical approximations.

When computing numerical approximations to balance laws, such as the shallow water equations, in steady-state or quasi-steady-state situations, the numerical solutions of the homogeneous PDE and the ODE are required to balance exactly. This exact balance is not likely to be respected by the fractional splitting procedure, and parasitic waves of a purely numerical nature can occur.

For stiff source terms, the usage of a stiff ODE solver combined with a HRSC method can lead to numerical solutions that *look reasonable* but are completely wrong. This phenomenon was observed as early as 1986 by Colella, Majda and Roytburd in [19] on a model combustion problem that involved the Euler equations of Gas Dynamics coupled with a single chemistry variable representing the mass fraction of unburnt gas in a detonation wave. The structure of the detonation waves obtained was well understood, and it was observed that the numerical solution obtained was qualitatively incorrect when computing on coarse grids. Stiff source terms could describe also the effect of relaxation as in the kinetic theory of rarefied gases, hydrodynamical models for semiconductors elasticity with memory, water waves, traffic flows, etc.

There has been a large amount of literature in recent years devoted to the numerical problems that occur in these two types of situations.

For shallow water equations, several authors extended the classical Riemann solver of Roe to nonhomogeneous problems related to balance laws [5], [7], [17], [58], [104]. In these works, the discrete form of the source terms is constructed in a way similar to that employed for the construction of the numerical fluxes, seeking an equilibria that exists in a steady-state conservation law with source terms. The idea of *source-term upwinding* lead Bermúdez and Vázquez-Cendón [5] to formulate the so-called C-property (for Conservation property) for a numerical scheme, which prevents the propagation of parasitic waves in steady and quasi-steady flows. Independently, Greenberg and Leroux [46] coined the term *well-balanced* for schemes that preserve steady states at the discrete level. These ideas have been explored and developed for shallow water flows in the recent literature [4], [24], [42], [61], [71], [78], [106] ...

In this work, we follow the strategy described by Gascón and Corberán in [38] and Donat, Caselles and Haro in [10]. In [38], the authors

propose to *formally* write the source term in divergence form so that the nonhomogeneous problem can be 'transformed' into a 'homogeneous form' through the definition of a new flux function. This change seeks to preserve the balance of the source and flux terms at steady states in an almost automatic manner, and suggests a way to apply well known schemes for homogeneous conservation laws to the non-homogenous case. However, as they readily observe, the application of the numerical methods for the homogeneous case is not immediate and adequate formalizations are required.

In [10], the idea of flux gradient and source term balancing in [38] was incorporated into the numerical scheme developed by Donat and Marquina in [26], thus effectively extending this scheme to balance laws.

In this work, we concentrate on the theoretical foundations of high-resolution total variation diminishing (TVD) schemes for homogeneous scalar conservation laws, firmly established through the work of Harten [50], Sweby [95], and Roe [80] and analyze the properties of a second order, flux-limited version of the Lax-Wendroff scheme which avoids oscillations around discontinuities, while preserving steady states [38]. When applied to homogeneous conservation laws, TVD schemes prevent an increase in the total variation of the numerical solution, hence guaranteeing the absence of numerically generated oscillations. They are successfully implemented in the form of flux-limiters or slope limiters for scalar conservation laws and systems. Our technique is based on a flux limiting procedure applied only to those terms related to the physical flow derivative/Jacobian.

With respect to the numerical treatment of stiff source terms, we follow Leveque and Yee in [73]. Taking the simple model problem considered in [73], we study the properties of the numerical solution obtained with different numerical techniques. We are able to identify the *delay factor*, which is responsible for the anomalous speed of propagation of the numerical solution on coarse grids. The delay is due to the introduction of non-equilibrium values through numerical dissipation, and can only be controlled by adequately reducing the spatial resolution of the simulation. Explicit schemes suffer from the same numerical pathology, even after reducing the time step so that the stability requirements imposed by the fastest scales are satisfied. We study the behavior of Implicit-Explicit (IMEX) numerical techniques, as a tool to obtain high resolution simulations that incorporate the stiff source term in an implicit, systematic, manner. The IMEX framework has been also successfully applied to hyperbolic systems with relaxation (see [9], [110], [73]).

Usually, when using very fine uniform grids, we find that the compu-

tational time becomes the main drawback in the numerical simulation. For some high resolution shock capturing schemes (HRSC), fine mesh simulations in two dimensions are out of reach simply because they cost too much. The numerical flux evaluations are too expensive. However, the flux computations are needed only because nonsmooth structures may develop spontaneously in the solution of a hyperbolic system of conservation laws and evolve in time, which lead to develop techniques that reduce the computational effort associated to these simulations. Harten in [49] proposed a scheme based on reducing the computational cost using the smoothness information of the data, and replacing the expensive numerical flux with a cheap polynomial interpolation in the smooth regions. The key is the use of different multilevel strategy to reduce the computational effort associated to HRSC scheme. In smooth regions, Harten in [49] proposes to evaluate the numerical flux function of the HRSC only on a coarse grid and to use these values to compute the fluxes on the finest grid using an inexpensive polynomial interpolation process in a multilevel fashion. Here, we extend the technique developed in [16] to hyperbolic conservation laws with source terms and apply the multilevel technique to the shallow water system.

Outline of Dissertation

This thesis is organized as follows:

Chapter 1: An introduction to conservation laws and the problems they present to numerical approximation is given in this chapter. The important concepts of conservation, entropy-satisfying solutions, monotonicity and total variation are described along with a selection of classical numerical methods.

Chapter 2: We discuss the basic methods used to solve inhomogeneous conservation laws, unsplit or splitting methods, and the main features of splitting.

Chapter 3: We study the stability and the wave speed for explicit, implicit and semi-implicit schemes for hyperbolic systems of conservation laws with stiff source terms.

Chapter 4: We propose a technique which is based on a flux limiting procedure applied only to those terms related to the physical flow derivative/Jacobian, which avoids oscillations around discontinuities, while preserving steady states.

Chapter 5: We apply the multilevel technique to the shallow water system, where the basic underlying scheme is the method presented in the previous chapter and the scheme of [10].

Appendix A: We revise some theoretical aspects for Ordinary Differential Equations and the Runge-kutta schemes.

Appendix B: We revise some theoretical aspects of the shallow water equations and the hypothesis underlying the derivation of the model from the Navier-Stokes equations. Then, we recall the Riemann problem of two adjacent states as well as their possible solutions:rarefaction waves and shock waves.

1

General formulation of conservation laws

The general form of a system of conservation laws, including source terms, that we shall consider in this memoir is as follows

$$\mathbf{u}_t + \nabla \cdot \mathbf{f}(\mathbf{u}) = \mathbf{s}(\mathbf{x}, \mathbf{u}), \tag{1.1}$$

where $\mathbf{u} \in \mathbb{R}^m$ represents the vector of unknowns (the *state variables*) $\mathbf{f}(\mathbf{u}) \in \mathbb{R}^m$ is the flux vector, $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}^+$ are the independent variables and $\mathbf{s} \in \mathbb{R}^m$ is the source term function.

In this preliminary chapter we shall revise some theoretical and numerical aspects of homogeneous conservation laws. The relevant theoretical results recalled in this chapter may be found in the works [36],

[39], [68], [69], [70], [72], [90], [100], [108]. For issues related to numerical techniques, the main sources of reference have been the works [50], [82], [94], [95].

Homogeneous conservation laws are an important subset of (1.1). We shall concentrate on the one-dimensional case, $n = 1$, in most of this memoir. The general form in this case is then

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0. \tag{1.2}$$

The system (1.2) is hyperbolic if the Jacobian matrix

$$A(\mathbf{u}) = \frac{\partial \mathbf{f}(\mathbf{u})}{\partial \mathbf{u}} \tag{1.3}$$

has m real eigenvalues and m linearly independent eigenvectors.

For such systems, we shall study the *Cauchy problem*, or initial value problem: Find a function $\mathbf{u} : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^m$ that is a solution of (1.2) satisfying the initial condition

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x), \tag{1.4}$$

where $\mathbf{u}_0$ is a given function. When $\mathbf{u}_0$ has the following particular form,

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_l, & x < 0, \\ \mathbf{u}_r, & x > 0, \end{cases} \tag{1.5}$$

this Cauchy problem is called the *Riemann problem*.

For a homogeneous system of conservation laws, such as (1.2), if we formally integrate with respect to the space variable on $\mathbb{R}$ and assume that the values of the state vector at $\pm\infty$, denoted as $\mathbf{u}|_{\pm\infty}$, are well defined, we obtain

$$\frac{d}{dt} \int_{-\infty}^{\infty} \mathbf{u}_i(x, t) dx = - \left(f_i(\mathbf{u}|_{+\infty}) - f_i(\mathbf{u}|_{-\infty}) \right) \tag{1.6}$$

hence, when $f(\mathbf{u}|_{+\infty}) = f(\mathbf{u}|_{-\infty})$ the integral of the density function of each state variable is constant with respect to time, i.e. it is *conserved*, although the spatial distribution of $\mathbf{u}_i$ is free to evolve with time. It is this evolution that we wish to model using discrete techniques.

1.1 The linear advection equation

Many of the numerical difficulties encountered with systems of conservation laws are already encountered in the scalar case, i.e. $m = 1$. In addition, it is often the case that numerical techniques used for systems were developed first having in mind the special structure of the solutions of scalar hyperbolic conservation laws. For this reason, and for simplicity in the description, we shall restrict ourselves mainly to the case of scalar conservation laws in one dimension, i.e. $n = 1$, in this introductory chapter.

A fundamental property of hyperbolic conservation laws is that they admit discontinuous solutions in a rather natural way. This fact can be most easily observed by considering the simplest of all conservation laws, the linear advection equation,

$$u_t + \lambda u_x = 0, \tag{1.7}$$

which we shall assume to be defined on the domain $-\infty < x < \infty, t \geq 0$, with initial conditions

$$u(x, 0) = u_0(x). \tag{1.8}$$

This equation is a model for the convective transport of a scalar quantity, u, moving in a flow of constant velocity, λ. There are no diffusive or dispersive effects so, as time evolves, the initial data propagates unchanged to the left ($\lambda < 0$), or to the right ($\lambda > 0$). It can be easily seen that the function

$$u(x, t) = u_0(x - \lambda t) \tag{1.9}$$

is a solution of the Cauchy problem, and it shows that the initial data moves, unchanged, to the right or to the left depending on the sign of λ. Clearly, if the initial data is discontinuous, the function defined as (1.9) would propagate the discontinuity of the initial data with speed λ, and it would make sense to consider this function as a *non classical* solution to the PDE, since it would not be differentiable.

In order to arrive at a proper definition of non classical, or *weak*, solutions, we shall start by noticing that the solution of (1.7) is constant along the straight lines $x - \lambda t = x_0$, which are known as the **characteristics** of the equation. The characteristic lines satisfy the ordinary differential equation $x'(t) = \lambda$, $x(0) = x_0$, and they are special curves along which we can obtain the value of the solution $u(x, t)$ by direct integration. Indeed,

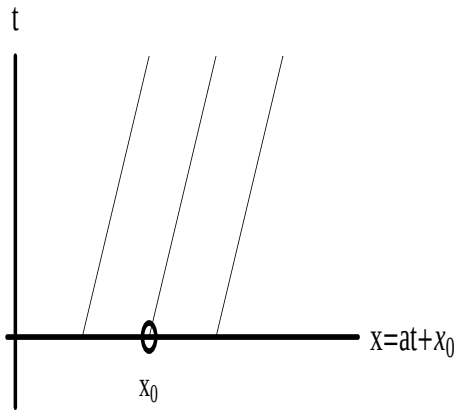


Figure 1.1: *Characteristics for the linear advection equation.*

if we differentiate the solution of (1.7) along one of these curves we find that

$$\begin{aligned} \frac{d}{dt}u(x(t),t) &= \frac{\partial}{\partial t}u(x(t),t) + \frac{\partial}{\partial x}u(x(t),t)x'(t) \\ &= u_t + \lambda u_x = 0. \end{aligned} \tag{1.10}$$

The rate of change of u along a characteristic line is zero, confirming that u is constant along these curves.

For linear, constant coefficient, hyperbolic conservation laws the characteristics are parallel lines of slope $1/\lambda$ in (x,t) space, where λ is the characteristic speed (see figure 1.1). The solution u at any point $(\bar{x},\bar{t})$ depends on the initial data u_0 only at a single point, namely the point $\bar{x_0} = \bar{x} - \lambda\bar{t}$ such that $(\bar{x},\bar{t})$ and $(\bar{x}_0,0)$ lie on the same characteristic line, see figure 1.1. The set $\mathcal{D}(\bar{x},\bar{t}) = \{\bar{x}_0\}$ is called the **domain of dependence** of the point $(\bar{x},\bar{t})$. Here this domain consist of a single point. For a system of equations this domain is typically an interval, but a fundamental fact about hyperbolic equations is that it is always a *bounded* interval. Conversely, the initial data at any given point x_0 can influence the solution only within some cone $\{x : |x-x_0| \leq \lambda t\}$ of the $x-t$ plane. This region is called the **range of influence** of the point x_0.

1.2 Nonlinear scalar equations

Let us now consider the nonlinear conservation law

$$u_t + f(u)_x = 0, \tag{1.11}$$

where $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a $\mathcal{C}^1$ function so that $f(u)$ is a nonlinear function of u.

Using the function $\lambda(u) = \frac{\partial f(u)}{\partial u}$, (1.11) can be written in quasi-linear form,

$$u_t + \lambda(u)u_x = 0, \tag{1.12}$$

allowing certain comparisons to be made with the linear advection equation (1.7). The curves $x = x(t)$ defined as the integral curves of the differential equation

$$\frac{dx}{dt} = \lambda(u), \quad x(0) = x_0, \tag{1.13}$$

satisfy that the rate of change of $u(x,t)$ along these curves is zero,

$$\frac{du}{dt} = u_t + \lambda(u)u_x = 0, \tag{1.14}$$

hence, the solution along these curves can be determined by integrating the ODE (1.14). These curves are also called characteristic curves. It follows from (1.13) that the characteristic curves are straight lines whose constant slopes depend on the initial data. The characteristic straight line passing through the point $(x_0, 0)$ is defined by the equation

$$x = x_0 + t\lambda(u_0(x_0)). \tag{1.15}$$

This important property gives a way to construct smooth solutions. One sets $u(x,t) = u_0(x_0)$, where x_0 is solution of (1.15). This is the so-called *method of characteristics*.

In the linear case the characteristics speed $\lambda(u) = \lambda$ is constant, so that the initial data is propagated unchanged (see figure 1.2). In the nonlinear case, however, the characteristic speed $\lambda(u)$ depends upon the solution u and so the initial data may deform on translation. As an example, consider the case of smooth initial data with a sine-like hump as shown in figure 1.3. The wave speed on the characteristics depends of the initial value $u_0(x_0)$ so the higher the value of u at x_0, the faster

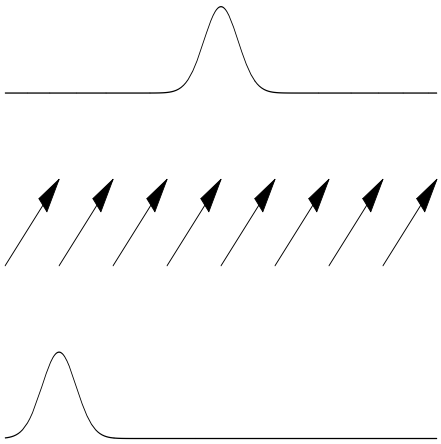


Figure 1.2: *Characteristics and solution for the linear equation* $u_t + \lambda u_x = 0, \lambda > 0$.

the characteristic speed and the shallower the slope of the characteristic in (x, t) space. As the solution evolves, the wave front will become increasingly steep until eventually some of the characteristics will cross. In general, let us assume that there exist two points $x_1 < x_2$ such that,

$$m_1 = \frac{1}{\lambda(u_0(x_1))} < m_2 = \frac{1}{\lambda(u_0(x_2))}.$$

Then, the characteristics C_1 and C_2 drawn from the points $(x_1, 0)$ and $(x_2, 0)$, respectively, have slopes m_1 and m_2 and intersect necessarily at some point P.

At this point P, the solution u should take both values $u_0(x_1)$ and $u_0(x_2)$, which is clearly impossible. Hence, the solution u cannot be continuous at the point P. Note that this phenomenon is independent of the smoothness of the functions u_0 and f. Indeed, using (1.15), we see that the two characteristics intersect at time t if

$$t(\lambda(u_0(x_1)) - \lambda(u_0(x_2))) = x_2 - x_1.$$

Thus, unless the function $x \to \lambda(u_0(x))$ is monotonically non-decreasing, in which case this equation has no positive solution t, we cannot define a classical solution u for all time $t > 0$ (see figure 1.3). Moreover, one can determine the critical time T_b up to which a classical solution exists and can be constructed by the method of characteristics; T_b is given by

$$T_b = -\frac{1}{\min\limits_{x \in \mathbb{R}} \left(\dfrac{d}{dx} \lambda(u_0(x)) \right)}.$$

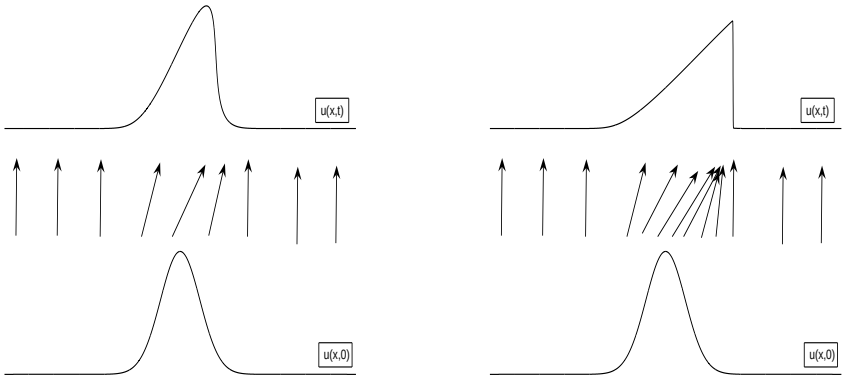


Figure 1.3: *Shock formation for nonlinear scalar equation.*

Since the solution is constant on each characteristic curve, it will become multi-valued at the crossing point. The classical solution ceases to exist, but a *weak* solution, a non-differentiable piecewise smooth function, can still be defined. At the breaking point, a discontinuity, or **shock** forms. This distinctly unphysical phenomenon is a result of inaccuracies in the original model (1.11), in particular the assumption of zero viscosity.

1.3 Weak solutions of Conservation Laws

Consider the initial-value problem for (1.2), (1.4), with $m = n = 1$ for simplicity in the description,

$$u_t + f(u)_x = 0, \quad u(x,0) = u_0(x). \tag{1.16}$$

A *classical solution* of (1.16) is a $\mathcal{C}^1$ function[1] $u : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ that satisfies the equations (1.16) pointwise.

As pointed out in the previous section, an essential feature of this problem is that there do not exist, in general, classical solutions of (1.16) beyond some finite time interval, even when the initial condition u_0 is

[1] it means that u is continuously differentiable in all variables

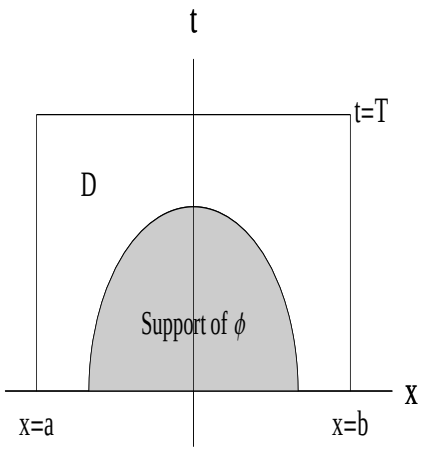


Figure 1.4

a very smooth function. A *variational* or *weak formulation* can then be applied in order to rewrite a differential equation in a form where less smoothness is required.

The basic idea is to take the PDE, multiply by a smooth "test function", integrate one or more times over an appropriate domain, and then use Green's theorem (or integration by parts) to move derivatives off the function u and onto the smooth test function. This results in a weak formulation that involves fewer derivatives on u, hence requiring less smoothness.

For (1.16) the test functions used are $\phi \in \mathcal{C}_0^1(\mathbb{R} \times \mathbb{R}^+)$. Here $\mathcal{C}_0^1$ is the space of functions that are continuously differentiable with "compact support". The latter requirement means that ϕ is identically zero outside of some bounded set, and so the support of the function lies in a compact set in $t \geq 0$, i.e., $(\text{supp}\phi) \cap (t \geq 0) \subseteq D$, where D is the rectangle $0 \leq t \leq T$, $a \leq x \leq b$, chosen so that $\phi = 0$ outside of D, and on the lines $t = T$, $x = a$, and $x = b$ (see figure 1.4).

If we multiply $u_t + f(u)_x = 0$ by $\phi(x,t) \in \mathcal{C}_0^1(\mathbb{R} \times \mathbb{R}^+)$ and then integrate over space and time, we obtain

$$\int_0^\infty \int_{-\infty}^\infty (\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x})\phi dx dt = 0, \tag{1.17}$$

which, after a formal integration by parts, yields

$$\begin{aligned}\int_0^\infty \int_{-\infty}^\infty (u\frac{\partial \phi}{\partial t} + f(u)\frac{\partial \phi}{\partial x})dxdt &= -\int_{-\infty}^\infty u(x,0)\phi(x,0)dx \\ &= -\int_{-\infty}^\infty u_0(x)\phi(x,0)dx.\end{aligned} \tag{1.18}$$

We remark that (1.18) makes sense if $u, u_0 \in \mathcal{L}^\infty_{loc}(\mathbb{R}\times\mathbb{R}^+)$, where $\mathcal{L}^\infty_{loc}$ is the space of locally bounded measurable functions. Hence, we find the following definition in [39].

Definition 1.1. *Assume that $u_0 \in \mathcal{L}^\infty_{loc}(\mathbb{R})$. A function $u \in \mathcal{L}^\infty_{loc}(\mathbb{R}\times\mathbb{R}^+)$ is called a weak solution of the Cauchy problem (1.16) if $u(x,t)$ satisfies (1.18) for all functions $\phi \in \mathcal{C}^1_0(\mathbb{R}\times\mathbb{R}^+)$.*

The concept of solution given by Definition 1.1 is a true generalization of the classical notion of solution. In fact, if, u is a $\mathcal{C}^1$ function so that (1.18) holds for all $\phi \in \mathcal{C}^1_0$, then we may integrate (1.18) by parts and get that for all such ϕ with support in $t > 0$

$$\int_0^\infty \int_{-\infty}^\infty \{\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x}\}\phi dxdt = 0,$$

by 'shrinking the support to a point' we get that $u_t + f(u)_x = 0$. Hence u is a classical solution of the PDE for $t > 0$. Notice that (1.18) reduces then to

$$\int_{-\infty}^\infty (u(x,0) - u_0(x))\phi(x,0)dx = 0,$$

and since u_0 is continuous, the arbitrariness of ϕ gives also that $u(x,0) = u_0(x)$ and we obtain that u is indeed a classical solution of the IVP (1.16).

The weak formulation (1.18) places severe restrictions on the curves across which there is a discontinuity of u. Let Γ be a smooth curve across which u has a jump discontinuity, that is, u has well-defined limits on both sides of Γ and u is smooth away from Γ. Let P be any point on Γ, and let D be a small ball centered at P. We assume that in D, Γ is given by $(x(t),t)$. Let D_1 and D_2 be the components of D which are determined by Γ,see figure 1.5. Let $\phi \in \mathcal{C}^1_0(D)$; then from (1.18),

$$\begin{aligned}0 &= \iint_D (u\phi_t + f(u)\phi_x)dxdt \\ &= \iint_{D_1} (u\phi_t + f(u)\phi_x)dxdt + \iint_{D_2} (u\phi_t + f(u)\phi_x)dxdt.\end{aligned}$$

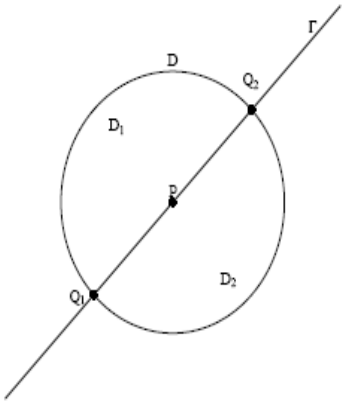


Figure 1.5

Now using the fact that U is $\mathcal{C}^1$ in D_i, the divergence theorem gives

$$\iint_{D_i}(u\phi_t + f(u)\phi_x)dxdt = \iint_{D_i}((u\phi)_t + (f(u)\phi)_x)dxdt = \oint_{\partial D_i}\phi(-ux'(t) + f(u))dt.$$

Since $\phi = 0$ on ∂D, these line integrals are nonzero only along Γ. Thus, if $u_l = u(x(t) - 0, t)$, and $u_r = u(x(t) + 0, t)$, then we have

$$\begin{aligned}\oint_{\partial D_1}\phi(-ux'(t) + f(u))dt &= \int_{Q_1}^{Q_2}\phi(-u_l x'(t) + f(u_l))dt \\ \oint_{\partial D_2}\phi(-ux'(t) + f(u))dt &= \int_{Q_2}^{Q_1}\phi(-u_r x'(t) + f(u_r))dt.\end{aligned}$$

Therefore,

$$0 = \int_\Gamma \phi(-[u]x'(t) + [f(u)])dt,$$

where $[u] = u_l - u_r$, the jump across Γ, and similarly, $[f(u)] = f(u_l) - f(u_r)$. Since ϕ was arbitrary, we conclude that

$$\xi[u] = [f(u)] \tag{1.19}$$

at each point on Γ, where $\xi = x'(t) = dx/dt$, is the *speed* of the discontinuity. Relation (1.19) is called the **Rankine-Hugoniot jump condition**. It holds, virtually unchanged, for systems (see e.g. [39] for details).

1.4 Entropy solutions

Weak solutions provide an adequate generalization of the concept of classical solution for hyperbolic conservation laws. However, weak solutions are not unique, and examples of this well known fact can be found in e.g. [90], [70], [39]. An additional condition is required to pick out the physically relevant solution.

For scalar conservation laws, the correct physical behavior can be determined by adopting the *vanishing viscosity approach*: The condition which defines the physically admissible solution is that it should be the limiting solution of the viscous equations

$$\frac{\partial}{\partial t}u^\varepsilon(x,t) + \frac{\partial}{\partial x}f(u^\varepsilon) = \varepsilon\frac{\partial^2}{\partial x^2}u^\varepsilon(x,t), \tag{1.20}$$

as $\varepsilon \to 0$. The physically relevant solution is often referred to as the *vanishing viscosity solution*.

However, this condition is hard to work with, and a variety of other conditions have been developed instead that can be applied directly in order to check whether or not a weak solution is physically admissible. Such conditions are sometimes called *admissibility conditions* or, more often, *entropy conditions*.

The name comes from gas dynamics, where the second law of thermodynamics demands that the entropy of a system must be a nondecreasing function. The entropy at each point can be computed as a simple function of the pressure and density, that should stay unchanged along particle paths and jump to a higher value only when the gas crosses a shock. The behavior of this function can be used to test a weak solution for admissibility.

For general systems of conservation laws it is sometimes possible to define a function $\eta(\mathbf{u})$, called an *entropy function*, which has similar diagnostic properties.

Given a classical solution $\mathbf{u}$ of (1.2) we wonder whether $\mathbf{u}$ would satisfy an additional conservation law of the form

$$\eta(\mathbf{u})_t + \psi(\mathbf{u})_x = 0 \tag{1.21}$$

where η and ψ are sufficiently smooth functions from Ω, an open subset of $\mathbb{R}^m$, into $\mathbb{R}$. This is indeed the case if

$$\psi'(\mathbf{u}) = \eta'(\mathbf{u})\mathbf{f}'(\mathbf{u}) \tag{1.22}$$

where, for ease of notation, we identify $\eta'(\mathbf{u}), \psi'(\mathbf{u}) : \mathbb{R}^m \to \mathbb{R}$ with the corresponding row vectors

$$\begin{aligned} \eta'(\mathbf{u}) &= \nabla\eta(\mathbf{u})^T = (\frac{\partial\eta}{\partial u_1}, \ldots, \frac{\partial\eta}{\partial u_m}) \\ \psi'(\mathbf{u}) &= \nabla\psi(\mathbf{u})^T = (\frac{\partial\psi}{\partial u_1}, \ldots, \frac{\partial\psi}{\partial u_m}) \end{aligned}$$

and $\mathbf{f}' : \mathbb{R}^m \to \mathbb{R}^m$ with the Jacobian matrix (1.3), $\mathbf{f}'(\mathbf{u}) = A(\mathbf{u})$

In fact, assuming that $\mathbf{u}$ is a classical solution of (1.2) and carrying out the differentiation in (1.21), we obtain

$$\eta'(\mathbf{u})\left(\mathbf{u}_t + \mathbf{f}'(\mathbf{u})\mathbf{u}_x\right) = 0 \tag{1.23}$$

and by (1.22)

$$\eta'(\mathbf{u})\mathbf{u}_t + \psi'(\mathbf{u})\mathbf{u}_x = 0$$

so that (1.21) follows.

Definition 1.2. *Assume that Ω is convex. Then, a convex function $\eta : \Omega \to \mathbb{R}$ is called an entropy for the system of conservation laws (1.2) if there exists a function $\psi : \Omega \to \mathbb{R}$, called entropy flux such that the relation (1.22) holds.*

The entropy $\eta(\mathbf{u})$ is conserved for smooth flows, by its definition, however, we do not expect this to hold for non-classical solutions since the manipulations performed above are not valid for non-classical solutions. As it turns out, the physically relevant solution is identified as the only one satisfying the following inequality (to be interpreted in the weak sense)

$$\eta(\mathbf{u})_t + \psi(\mathbf{u})_x \leq 0 \tag{1.24}$$

for all convex entropy functions (see below) and corresponding entropy fluxes.

For a scalar conservation law the equation (1.21) admits many *entropy pairs* $\eta(u), \psi(u)$. One trivial choice is $\eta(u) = u$ and $\psi(u) = f(u)$. For this choice η is conserved even across discontinuities in a weak solution and this would not help in defining an admissibility criterion. The requirement that $\eta(u)$ be a convex function of u with $\eta''(u) > 0$ for all u is important in the admissibility criterion (1.24) (see also next subsection).

For a system of equations, (1.22) is, in general, a system of m equations for the two variables η and ψ. Moreover, we also required that $\eta(\mathbf{u})$ be convex, which for a system requires that the Hessian matrix $\eta''(\mathbf{u})$ be positive definite. For $m > 2$ this may have no solutions.

Many physical systems do have entropy functions, however, including of course the Euler equations of gas dynamics, where the negative of the physical entropy can be used. For symmetric systems there is always an entropy function $\eta(\mathbf{u}) = \mathbf{u}^T\mathbf{u}$, as observed by Godunov ([41]). Conversely, if a system has a convex entropy, then its Hessian matrix $\eta''(\mathbf{u})$ symmetrizes the system ([36], [96]).

1.4.1 The vanishing viscosity method

For scalar conservation laws, it is relatively simple to show that the concept of entropy function described above will enable us to select among the weak solutions of (1.2), (1.4) the physically relevant solution. In what follows we recall a derivation, extracted from [70], that shows that the weak solution constructed as the limit when $\varepsilon \to 0$ of solutions to (1.20) satisfies the entropy condition (1.24).

In order to see how $\eta(u^\varepsilon)$ behaves for solutions u^ε to (1.20) we multiply (1.20) by $\eta'(u^\varepsilon)$. Taking into account that the solutions of (1.20) are $\mathcal{C}^1$, we obtain

$$\eta(u^\varepsilon)_t + \psi(u^\varepsilon)_x = \varepsilon\eta'(u^\varepsilon)u^\varepsilon_{xx} = \varepsilon(\eta'(u^\varepsilon)u^\varepsilon_x)_x - \varepsilon\eta''(u^\varepsilon)(u^\varepsilon_x)^2. \tag{1.25}$$

Integrating over the rectangle $[x_1, x_2] \times [t_1, t_2]$ gives

$$\begin{aligned}\int_{x_1}^{x_2} \eta(u^\varepsilon(x,t_2))dx \;=\; & \int_{x_1}^{x_2} \eta(u^\varepsilon(x,t_1))dx \\ & - \left(\int_{t_1}^{t_2} \psi(u^\varepsilon(x_2,t))dt - \int_{t_1}^{t_2} \psi(u^\varepsilon(x_1,t))dt\right) \\ & + \varepsilon \int_{t_1}^{t_2} \left(\eta'(u^\varepsilon(x_2,t))u^\varepsilon_x(x_2,t) - \eta'(u^\varepsilon(x_1,t))u^\varepsilon_x(x_1,t)\right) dt \\ & - \varepsilon \int_{t_1}^{t_2}\int_{x_1}^{x_2} \eta''(u^\varepsilon)(u^\varepsilon_x)^2 dxdt.\end{aligned}$$

In addition to the flux differences, the total entropy is modified by two terms involving ε. As $\varepsilon \to 0$, the first of these terms vanishes. This is clearly true if the limiting function $u(x,t)$ is smooth at x_1 and x_2, and can be shown more generally . The other term, however, involves integrating $(u^\varepsilon_x)^2$ over the rectangle $[x_1, x_2] \times [t_1, t_2]$. If the limiting weak solution is discontinuous along a curve in this rectangle, then this term will not vanish in the limit. However, since $\varepsilon > 0$, $(u^\varepsilon_x)^2$, and $\eta'' > 0$ (by the

convexity assumption), we can conclude that this term is nonpositive in the limit and hence the vanishing-viscosity weak solution satisfies

$$\int_{x_1}^{x_2} \eta(u(x,t_2))dx \leq \int_{x_1}^{x_2} \eta(u(x,t_1))dx + \int_{t_1}^{t_2} \psi(u(x_1,t))dt - \int_{t_1}^{t_2} \psi(u(x_2,t))dt. \tag{1.26}$$

Consequently the total integral of η is not necessarily conserved, but can only decrease (in gas dynamics η is the negative of the physical entropy).

The inequality above is one realization of (1.24). More often, the entropy inequality is formulated by integrating against smooth test functions ϕ, now required to be nonnegative, to respect the sign in the inequality.

Definition 1.3. *A weak solution u of (1.2), (1.4) is called an entropy solution if u satisfies, for all convex entropy functions η and corresponding entropy flux ψ, and for all test functions $\phi \in \mathcal{C}_0^1(\mathbb{R} \times \mathbb{R}^+)$, $\phi \geq 0$,*

$$\int_0^\infty \int_{-\infty}^\infty \left(\phi_t \eta(u) + \phi_x \psi(u)\right) dxdt + \int_{-\infty}^\infty \phi(x,0)\eta(u(x,0))dx \geq 0. \tag{1.27}$$

1.4.2 The entropy solution in the scalar case

For scalar equations, other admissibility conditions of a more geometrical nature have been also developed. We recall them here since these are very often the easiest ones to apply.

In the scalar case, it is obvious that a shock should have characteristics going into the shock, as time advances, while a propagating discontinuity with characteristics coming out of it (an *expansion shock*) would be unstable to perturbations. Either smearing out the initial profile a little, or adding some viscosity to the system, will cause this to be replaced by a rarefaction fan of characteristics. This gives the entropy conditions:

Entropy Condition 1. ***(Lax)***. *For a convex scalar conservation law, a discontinuity propagating with speed ξ given by (1.19) satisfies the Lax entropy condition if*

$$f'(u_l) > \xi > f'(u_r). \tag{1.28}$$

Note that $f'(u)$ is the characteristic speed. For convex or concave f, the Rankine-Hugoniot speed ξ from (1.19) must lie between $f'(u_l)$ and $f'(u_r)$, so (1.28) simply reduces to the requirement that $f'(u_l) > f'(u_r)$.

Another form of the entropy condition is based on the spreading of characteristics in a rarefaction fan. For the convex case with $f''(u) > 0$, if $u(x,t)$ is an increasing function of x in some region, then the characteristics spread out at a rate which can be quantified. This gives the following condition.

Entropy Condition 2. ***(Oleinik).*** *$u(x,t)$ is the entropy solution to a scalar conservation law $u_t + f(u)_x = 0$ with $f''(u) > 0$ if there is a constant $E > 0$ such that for all $a > 0$,$t > 0$, and $x \in \mathbb{R}$,*

$$\frac{u(x+a,t) - u(x,t)}{a} < \frac{E}{t}. \tag{1.29}$$

Note that for a discontinuity propagating with constant left and right states u_l and u_r, this can be satisfied only if $u_r - u_l \leq 0$, so this agrees with (1.28). The form of (1.29) also gives information on the rate of spreading of rarefaction waves as well as on the form of allowable jump discontinuities, and is easier to apply in some contexts. In particular, this formulation has advantages in studying numerical methods, and is related to stability concepts.

1.5 Numerical Methods

In order to solve numerically the differential equations arising from the conservation laws it is necessary to replace the continuous problem with a discrete problem on a finite mesh. This is usually done in one of two ways. For a *finite difference* approximation, the values of the conserved quantities are calculated as point values at the intersections of the mesh, using approximations of the differential form of the conservation law. In a *finite volume* approach use is made of the integral form of the conservation law and the quantities are averaged over the cells. It is the finite difference formulation which interests us here.

The numerical solution of nonlinear equations of the form (1.2), poses many additional problems which are, in general, much harder to analyze

than linear problems. In many cases it is necessary to linearize the problem before any useful analysis can be performed. However, nonlinear instabilities can still occur, triggered by oscillations in the solution, even when the linearized version is stable. If the schemes do converge they may do so to an entropy-violating weak solution, or worse still, may converge to a function which is not at all a weak solution of the original differential equation [70]. For the design of robust numerical schemes for nonlinear equations each of these issues must be addressed.

In this section we revise the basic schemes applied solve 1-D conservation laws, paying special attention to those that will be employed later on in this thesis. We refer to the reader to [72], [70] and [100] for a more complete description.

The general setup is as follows: The $x-t$ plane is discretized by choosing a mesh width $\triangle x$ and a time step $\triangle t$, and define the discrete mesh points (x_i, t_n) by

$$\begin{aligned} x_i &= i\triangle x, \quad i = \ldots, -1, 0, 1, \ldots \\ t_n &= n\triangle t, \quad n = 0, 1, 2, \ldots \end{aligned}$$

Finite difference methods produce approximations U_i^n to the solution $u(x_i, t_n)$ at the discrete grid points. We shall denote $U^n = (U_i^n)_i$. We shall consider two-level schemes that can be written in the general form

$$U_i^{n+1} = H(U_{i-p}^n, \ldots, U_{i+q}^n) \tag{1.30}$$

where H is the discrete solution operator, and p, q are positive constants.

1.5.1 The CFL condition

One of the first papers on finite difference methods for PDEs was written in 1928 by Courant, Friedrichs and Lewy [20]. They used finite difference methods as an analytic tool for proving existence of solutions of certain PDEs. The idea is to define a sequence of approximate solutions (via finite difference equations), prove that they converge as the grid is refined, and then show that the limit function must satisfy the PDE, giving the existence of a solution.

In the course of proving convergence of this sequence, they recognized that a necessary stability condition, not sufficient, for any numerical method is that the domain of dependence of the finite difference method should include the domain of dependence of the PDE, at least in the limit as the grid is refined. It simply states that the method must be used in

such a way that the information has a chance to propagate at the correct physical speeds, as determined by the eigenvalues of the flux Jacobian $\mathbf{f}'(\mathbf{u})$. This condition is known as the CFL condition.

For example, if we consider the linear advection equation (1.7), for positive wave speed $\lambda = \frac{dx}{dt} > 0$ the characteristic $x = \lambda t$ always lies outside the domain of dependence of the forward difference scheme, defined by

$$U_i^{n+1} = U_i^n - \frac{\lambda \triangle t}{\triangle x}\left(U_{i+1}^n - U_i^n\right), \tag{1.31}$$

therefore the scheme will be unstable. The backwards difference scheme defined by

$$U_i^{n+1} = U_i^n - \frac{\lambda \triangle t}{\triangle x}\left(U_i^n - U_{i-1}^n\right), \tag{1.32}$$

will only be stable if the characteristic crosses the n^{th} time level between x_{i-1} and x_i. Since the time-step is $\triangle t$, the characteristic crosses t_n at $x = x_i - \lambda \triangle t = i\triangle x - \lambda \triangle t$, and so we require that

$$(i-1)\triangle x \leq i\triangle x - \lambda \triangle t \leq i \triangle x$$

or,

$$0 \leq \frac{\lambda \triangle t}{\triangle x} \leq 1. \tag{1.33}$$

Since λ and $\triangle x$ are either fixed or defined by the problem, the CFL condition imposes a limit on the size of the time-step. If the wave direction changes so that $\lambda < 0$, (1.33) will be no longer satisfied and the backwards difference scheme becomes unstable. The forward difference scheme, on the other hand, will become stable under the CFL condition

$$-1 \leq \frac{\lambda \triangle t}{\triangle x} \leq 0.$$

The above examples illustrate the importance of considering the physics of the flow, such as the wave direction, when choosing the numerical scheme.

1.5.2 Conservation Form

It is possible to ensure convergence to weak solutions for numerical schemes in 'conservation form'

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x}[F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}], \tag{1.34}$$

where F is a numerical flux function of the form

$$F^n_{i+\frac{1}{2}} = F(U^n_{i-p}, \ldots, U^n_{i+q}). \tag{1.35}$$

In the simplest case, $p = 0$ and $q = 1$, so that F is a function of only two variables and (1.35) becomes

$$F^n_{i+\frac{1}{2}} = F(U^n_i, U^n_{i+1}). \tag{1.36}$$

The method (1.34) is **consistent** with the original conservation law if the numerical flux function F reduces to the true flux f for the case of constant flow. If $u(x,t) \equiv \bar{u}$, then we expect that

$$F(\bar{u}, \bar{u}) = f(\bar{u}) \quad \forall \bar{u} \in \mathbb{R}. \tag{1.37}$$

Some smoothness is also required, so that as the two arguments of F approach some common value $\bar{u}$, the value of F approaches $f(\bar{u})$ smoothly. For consistency it suffices to have F a **Lipschitz continuous** function of each variable. We say that F is Lipschitz at $\bar{u}$ if there is a constant $K \geq 0$ (which may depend on $\bar{u}$) such that

$$|F(v,w) - f(\bar{u})| \leq K \max(|v - \bar{u}|, |w - \bar{u}|) \tag{1.38}$$

for all $v, w \in \mathbb{R}$ with $|v - \bar{u}|$ and $|w - \bar{u}|$ sufficiently small. We say that F is a Lipschitz continuous function if it is Lipschitz at every point. Numerical schemes written in this form are called *conservative schemes*.

The Lax-Wendroff Theorem [69] states that if a numerical scheme in conservation form converges to $\bar{u}$ as $\triangle x \to 0$, with $\triangle t / \triangle x$ fixed, then $\bar{u}$ will be a weak solution of the conservation law. This result, whilst not guaranteeing convergence, does preclude convergence to functions which are not weak solutions.

However, the Lax-Wendroff theorem does not ensure that, upon convergence, the scheme does so to the correct weak solution. Provided that the convergence can be ensured in some way, it can be shown that the weak solution obtained in the limit, $\bar{u}$, satisfies an entropy inequality such as (1.27) for a suitable entropy pair (η, ψ) by showing that a discrete version of the entropy inequality (1.24) holds,

$$\eta(U^{n+1}_i) \leq \eta(U^n_i) - \frac{\triangle t}{\triangle x}(\Psi^n_{i+\frac{1}{2}} - \Psi^n_{i-\frac{1}{2}}), \tag{1.39}$$

where $\Psi^n_{i+\frac{1}{2}} = \Psi(U^n_i, U^n_{i+1})$, and $\Psi(u_l, u_r)$ is some numerical entropy flux function that must be consistent with ψ in the same manner that F

is required to be consistent with f. The proof mimics that of the Lax Wendroff theorem; we refer the reader to [70] and references therein for further details.

In general, the above conditions are far from easy to test for individual schemes; however, there do exist classes of schemes which are known to possess this entropy-satisfying property.

1.5.3 Nonlinear Stability

The Lax-Wendroff theorem (see [69], [72]) does not say anything about the convergence of a numerical scheme in conservation form, only that if a sequence of approximations converges, then the limit is a weak solution.

For linear difference methods applied to linear PDEs, the Lax Equivalence theorem establishes that, for a *consistent*, method, *stability* is necessary and sufficient for *convergence*. However, for nonlinear problems (PDEs or numerical schemes), the Lax equivalence theorem no longer holds (the proof relies heavily on linearity) and the primary tool to prove convergence is *compactness*.

Here we follow [72] (also in [70]) and describe the concept of **Total Variation stability**, one form of nonlinear stability that allows to prove convergence for a wide class of practical methods. So far, this approach has been completely successful only for scalar problems.

When dealing with solutions of hyperbolic conservation laws, a natural function space to consider is that of functions which are integrable in absolute value, as functions of x, for each value of the time variable, i.e. $u(\cdot, t) \in \mathcal{L}^1(\mathbb{R})$, $\forall t$. As observed in e.g. [70], the key point is that the sets of functions with bounded Total Variation are compact in $\mathcal{L}^1$. The Total Variation, defined as

$$TV(v) = \lim_{\epsilon \to 0} \int_{-\infty}^{\infty} \frac{|v(x+\epsilon) - v(x)|}{\epsilon} dx \tag{1.40}$$

becomes, hence, a key concept in proving stability. In fact, the set

$$\{v \in \mathcal{L}^1 : TV(v) \leq R \text{ and } v(x) = 0 \text{ for} |x| > M\}$$

is compact in $\mathcal{L}^1(\mathbb{R})$, so that any sequence of functions with uniformly bounded total variation and support must contain convergent subsequences.

In order to properly discuss the convergence of a numerical scheme under grid refinement, it is customary to resort to the definition of the auxiliary piecewise constant function, $U^{\triangle t}(x,t)$, defined as follows

$$U^{\triangle t}(x,t) = U_i^n \text{ for } (x,t) \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times [t_n, t_{n+1}),$$

where the discrete values U_i^n have been obtained with the given numerical scheme for a given time step Δt on a mesh with stepsize Δx. The function is identified by the time step, and it is assumed that Δt and Δx are related in a fixed way (normally by a fixed CFL condition) so that the choice of Δt defines a unique mesh.

Since $U^{\triangle t}(x,t)$ is a piecewise constant function, we have

$$TV(U^{\Delta t}(\cdot,t)) = \sum_i |U_{i+1}^n - U_i^n|, \qquad \forall t \in [t_n, t_{n+1}). \tag{1.41}$$

The Total Variation of the numerical solution at time n is defined as the quantity above

$$TV(U^n) := TV(U^{\Delta t}(\cdot,t)) = \sum_i |U_{i+1}^n - U_i^n|. \tag{1.42}$$

The main theorem relating these definitions with the convergence of a numerical scheme is stated below (see e.g.[70] for further information).

Theorem 1.1. *([72]) Suppose $U^{\triangle t}$ is generated by a numerical method in conservation form with Lipschitz continuous numerical flux, consistent with some scalar conservation law. If the method is TV-stable, i.e., if $TV(U^n)$ is uniformly bounded for all n, $\triangle t$ with $\triangle t < \triangle t_0$, $n\triangle t \leq T$, then the method is convergent.*

It should be mentioned that for the entropy solution, the Total Variation is non-increasing with respect to time, which makes the concept of TV-stability much more relevant. We recall the following theorem from ([39])

Theorem 1.2. *Assume that the function u_0 belongs to $\mathcal{L}^\infty(\mathbb{R})$. Then, the problem (1.16) has a unique entropy solution $u \in \mathcal{L}^\infty(\mathbb{R} \times (0,T))$. This solution satisfies for almost all $t \geq 0$*

$$\|u(\cdot,t)\|_{\mathcal{L}^\infty(\mathbb{R})} \leq \|u_0\|_{\mathcal{L}^\infty(\mathbb{R})}.$$

Moreover, if u and v are the entropy solutions of (1.16) associated with the initial conditions u_0 and v_0, respectively, we have

$$u_0 \geq v_0 \Longrightarrow u(\cdot,t) \geq v(\cdot,t) \quad a.e.$$

Finally, if u_0 belongs to $\mathcal{L}^\infty(\mathbb{R}) \cap \mathcal{BV}(\mathbb{R})$, then $u(\cdot,t)$ belongs to $\mathcal{BV}(\mathbb{R})$ with

$$TV(u(\cdot,t)) \leq TV(u_0).$$

Remark $\mathcal{BV}(\mathbb{R})$ is the space of functions with bounded variation.

For a continuously differentiable function, it is easy to see that the definition in (1.40) leads to $TV(v) = \int |v_x| dx$.

Total Variation Diminishing schemes

An important class of TV-stable schemes (hence convergent) are the so-called Total Variation Diminishing (TVD) schemes.

Definition 1.4. *A numerical method is called Total Variation Diminishing (TVD) if, for any set of data U^n, the values U^{n+1} computed by the method satisfy*

$$TV(U^{n+1}) \leq TV(U^n), \tag{1.43}$$

with $TV(U^n)$ as defined in (1.41).

Harten [50] introduced the use of this tool in developing and analyzing numerical methods for conservation laws. While this seems a natural requirement, at least for the scalar case, according to theorem 1.2, an important feature of these schemes are that numerical solutions obtained with TVD schemes will not have spurious oscillations. Indeed, if the numerical method introduces oscillations, then we would expect the total variation of U^n to increase with time.

Harten gave a useful characterization of TVD schemes. We recall here the following theorem from [50]

Theorem 1.3. *Consider a general method of the form*

$$U_i^{n+1} = U_i^n - C_{i-1}^n(U_i^n - U_{i-1}^n) + D_i^n(U_{i+1}^n - U_i^n) \tag{1.44}$$

over one time step, where the coefficients C_{i-1}^n and D_i^n are arbitrary values (which in particular may depend on values of U^n in some way, i.e., the method may be nonlinear). Then $TV(U^{n+1}) \leq TV(U^n)$ provided the following conditions are satisfied:

$$\begin{aligned} C_{i-1}^n &\geq 0 \quad \forall i, \\ D_i^n &\geq 0 \quad \forall i, \\ C_i^n + D_i^n &\leq 1 \quad \forall i. \end{aligned} \tag{1.45}$$

Monotone methods

It is also possible to ensure convergence in the nonlinear case if the numerical method is contractive in some norm. In particular, this is true for the class of monotone methods. These are methods which can be written in the form

$$U_j^{n+1} = H(U_{j-p}^n, \ldots, U_{j+q}^n), \tag{1.46}$$

where H is monotonically non-decreasing in each of its arguments, i.e.

$$\frac{\partial H}{\partial U_i} \geq 0 \quad \forall j - p \leq i \leq j + q. \tag{1.47}$$

Monotone schemes produce non-oscillatory solutions. Furthermore it is proven in [52] that the converged solutions of monotone schemes always correspond to physically acceptable states, ruling out entropy-violating shocks. It can be proven (see e.g. [70]) that the numerical solutions of scalar conservation laws computed with monotone schemes converge to the entropy solution as $\Delta t \to 0$.

The limitations of such schemes were highlighted by Godunov [40] who demonstrated that all monotone linear schemes are at most first-order accurate. Monotone schemes will therefore suffer the same practical limitations as other first-order schemes in being too diffusive and leading to a smearing of discontinuities.

Other less stringent conditions which may be brought to bear on numerical schemes to prevent the generation of spurious oscillations in-

clude the local maximum principle,

$$\min(U_{i-p}^n, \ldots, U_{i+q}^n) \leq U_i^{n+1} \leq \max(U_{i-p}^n, \ldots, U_{i+q}^n), \tag{1.48}$$

and monotonicity preservation, which requires that if a solution is monotone at time $n\triangle t$ then it will remain monotone at time $(n+1)\triangle t$. Unfortunately neither of these criteria can be easily shown to be satisfied by a particular scheme with arbitrary data.

The notion of Total Variation Stability is much more useful because it is possible to derive schemes that satisfy this property (hence convergent) and have order of accuracy greater than one.

1.5.4 First order schemes for scalar equations

Some of the most important first order schemes to solve conservation laws are explained in this section.

Lax-Friedrichs

The classical Lax-Friedrichs method has the form

$$U_i^{n+1} = \frac{1}{2}(U_{i-1}^n + U_{i+1}^n) - \frac{\triangle t}{2\triangle x}\left(f(U_{i+1}^n) - f(U_{i-1}^n)\right). \tag{1.49}$$

For a linear hyperbolic this method is stable provided CFL ≤ 1, where the Courant number CFL is defined by

$$\text{CFL} = \frac{\triangle t}{\triangle x}|\lambda| \tag{1.50}$$

where λ is the wave speed.

This scheme can be put in conservation form (1.34) by defining the numerical flux as

$$F_{i+\frac{1}{2}}^n = \frac{1}{2}\left(f(U_{i+1}^n) + f(U_i^n) - \frac{\triangle x}{\triangle t}(U_{i+1}^n - U_i^n)\right). \tag{1.51}$$

The Lax-Friedrichs method exhibits a curious stair-step pattern in which alternates $U_{2i}^n = U_{2i+1}^n \Rightarrow U_{2i-1}^{n+1} = U_{2i}^{n+1}$ for each value of i. This results from the fact that the formula (1.49) for U_i^{n+1} involves only U_{i-1}^n and U_{i+1}^n, so there is a decoupling of even and odd grid points.

An improvement to the Lax-Friedrichs method is obtained by replacing the value $\triangle x/\triangle t$ in (1.51) by a locally determined value,

$$F^n_{i+\frac{1}{2}} = \frac{1}{2}\left(f(U^n_{i+1}) + f(U^n_i) - \nu^n_{i+\frac{1}{2}}(U^n_{i+1} - U^n_i)\right), \tag{1.52}$$

where

$$\nu^n_{i+\frac{1}{2}} = \max\left(\left|f'(u)\right|\right) \qquad \text{over all } u \text{ between } U^n_{i-1} \text{ and } U^n_i.$$

For a convex function this reduces to

$$\nu^n_{i+\frac{1}{2}} = \max\left(\left|f'(U^n_i)\right|, \left|f'(U^n_{i+1})\right|\right).$$

This resulting method is often called the *local Lax-Friedrichs (LLF) method* because it has the same form as the Lax-Friedrichs method but the viscosity coefficient is chosen locally at each Riemann problem. Note that if the CFL condition is satisfied (which is a necessary condition for stability), then $|f'(u)|\triangle t/\triangle x \leq 1$ for each value of u arising in the whole problem, and so

$$\left|f'(u)\right| \leq \frac{\triangle x}{\triangle t}.$$

Hence using $\frac{\triangle x}{\triangle t}$ in the standard Lax-Friedrichs method amounts to taking a uniform viscosity that is sufficient everywhere, at the expense of too much smearing in most cases.

For a linear hyperbolic system this method is stable provided CFL ≤ 1, where the Courant number CFL is defined by

$$\text{CFL} = \frac{\triangle t}{\triangle x}\max_p |\lambda_p|, \tag{1.53}$$

where $\lambda_1, \ldots, \lambda_m$ are a set of m wave speeds for the system of equations.

E-schemes

Osher [77] defined a class of schemes, denoted **E-schemes**, which guarantee satisfaction of the entropy condition. An E-scheme is any scheme in conservative form, with numerical flux $F^n_{i+\frac{1}{2}}$, which satisfies

$$\text{sgn}(U^n_{i+1} - U^n_i)(F^n_{i+\frac{1}{2}} - f(u)) \leq 0 \tag{1.54}$$

for all u between U^n_i and U^n_{i+1}.

E-schemes are in fact monotone schemes, hence they converge to the entropy-satisfying solution and do not produce spurious oscillations.

An example is the first-order scheme of Engquist-Osher [29] which, when written in conservative form (1.34), has a numerical flux function defined by

$$F_{i+\frac{1}{2}} = f_i^+ + f_{i+1}^-, \tag{1.55}$$

where

$$f_i^+ = \int_{\bar{U}}^{\bar{U}_i} \chi(s) f'(s) ds, \qquad f_i^- = \int_{\bar{U}}^{\bar{U}_i} (1 - \chi(s)) f'(s) ds, \tag{1.56}$$

$\bar{U}$ is the sonic point satisfying $f'(\bar{U}) = 0$, and $\chi(s)$ is the switching function

$$\chi(s) = \begin{cases} 1, & f'(s) > 0, \\ 0, & f'(s) \leq 0. \end{cases}$$

For convex flux functions, such as in Burgers equation ($f(u) = \frac{1}{2}u^2$), the split fluxes become

$$f_i^+ = f(\max(U_i^n, \bar{U})) \qquad f_i^- = f(\min(U_i^n, \bar{U})). \tag{1.57}$$

1.5.5 Upwind Schemes: Godunov's method

For the scalar advection equation (1.7) with $\lambda > 0$, the one-sided method (1.32) can be applied and is stable provided (1.33) is satisfied. This method is usually called the **first order upwind method**, since the one-sided stencil points in the upwind or upstream direction, the correct direction from which characteristic information propagates. If we think of the advection equation as modeling the advection of a concentration profile in a fluid stream, then this is literally the upstream direction. Similarly, the method (1.31) is the upwind method for the advection equation with $\lambda < 0$.

When computing discontinuous solutions, upwind differencing turns out to be an important tool, even for indefinite systems with eigenvalues of mixed sign. The appropriate application of upwind methods requires some sort of decomposition into characteristic fields. The fundamental method of this type is *Godunov's method*.

Godunov [40] proposed a way to make use of the characteristic information within the framework of a conservative method. Rather than

attempting to follow characteristics backwards in time, Godunov suggested solving Riemann problems forward in time. Solutions to Riemann problems are relatively easy to compute, give substantial information about characteristic structure, and lead to conservative methods since they are themselves exact solutions of the conservation laws and hence conservative. The structure of the Godunov's method is reconstruct-evolve-average. Let us see the algorithm:

1. *Reconstruct* a piecewise polynomial function $\tilde{u}^n(x, t_n)$ defined for all x, from the cell averages U_i^n. In the simplest case this is a piecewise constant function that takes the value U_i^n in the ith grid cell, i.e.,

$$\tilde{u}^n(x, t_n) = U_i^n \text{ for all } x \in (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}). \tag{1.58}$$

2. *Evolve* the hyperbolic equation exactly (or approximately) with this initial data to obtain $\tilde{u}^n(x, t_{n+1})$ one timestep $\triangle t$ later.

3. *Average* this function over each grid cell to obtain new cell averages

$$U_i^{n+1} = \frac{1}{\triangle x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{u}^n(x, t_{n+1}) dx. \tag{1.59}$$

This whole process is then repeated in the next time step. In practice this algorithm is considerably simplified by observing that the cell average (1.59) can be easily computed using the integral form of the conservation law. Since $\tilde{u}^n$ is assumed to be an exact weak solution, we know that

$$\begin{aligned} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{u}^n(x, t_n + 1) dx &= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{u}^n(x, t_n) dx \\ &\quad + \int_{t_n}^{t_{n+1}} f(\tilde{u}^n(x_{i-\frac{1}{2}}, t)) dt - \int_{t_n}^{t_{n+1}} f(\tilde{u}^n(x_{i+\frac{1}{2}}, t)) dt. \end{aligned}$$

Dividing by $\triangle x$, using (1.59), and noting that $\tilde{u}^n(x, t_n) \equiv U_i^n$ over the cell $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, this equation reduces to

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x} \left(F(U_i^n, U_{i+1}^n) - F(U_{i-1}^n, U_i^n) \right), \tag{1.60}$$

where the numerical flux function F is given by

$$F(U_i^n, U_{i+1}^n) = \frac{1}{\triangle t} \int_{t_n}^{t_{n+1}} f(\tilde{u}^n(x_{i+\frac{1}{2}}, t)) dt. \tag{1.61}$$

This shows that Godunov's method can be written in conservation form. Moreover, note that the integral we need to compute in (1.61) is trivial because $\tilde{u}^n$ is constant at the point $x_{i+\frac{1}{2}}$ over the time interval (t_n, t_{n+1}). This follows from the fact that the solution of the Riemann problem at $x_{i+\frac{1}{2}}$ is a similarity solution, constant along each ray $(x - x_{i+\frac{1}{2}})/t$.

The constant value of $\tilde{u}^n$ along the line $x = x_{i+\frac{1}{2}}$ depends only on the data U_i^n and U_{i+1}^n for this Riemann problem. If we denote this value by $u^*(U_i^n, U_{i+1}^n)$, then the flux (1.61) reduces to

$$F(U_i^n, U_{i+1}^n) = f(u^*(U_i^n, U_{i+1}^n)). \tag{1.62}$$

For large t, of course, the solution may not remain constant at $x_{i+\frac{1}{2}}$ because of the effect of waves arising from neighboring Riemann problems, $\tilde{u}^n(x_{i+\frac{1}{2}}, t)$ will be constant over $[t_n, t_{n+1}]$ provided $\triangle t$ is sufficiently small

$$\left| \frac{\triangle t}{\triangle x} f'(U_i^n) \right| \leq 1. \tag{1.63}$$

Consider the constant coefficient, linear hyperbolic system

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \quad \mathbf{f}(\mathbf{u}) = A\mathbf{u}. \tag{1.64}$$

The Godunov first-order upwind method utilizes the conservative formula (1.34) and requires the solution of the local Riemann problem for (1.64) to compute the intercell numerical flux. As in the scalar case, for large t, of course, the solution may not remain constant at $x_{i+\frac{1}{2}}$ because of the effect of waves arising from neighboring Riemann problems. However, since the wave speeds are bounded by the eigenvalues of $\mathbf{f}'(\mathbf{u})$ and the neighboring Riemann problems are at a distance $\triangle x$ away, $\tilde{u}^n(x_{i+\frac{1}{2}}, t)$ will be constant over $[t_n, t_{n+1}]$ provided $\triangle t$ is sufficiently small. It is required that

$$\left| \frac{\triangle t}{\triangle x} \lambda_p(U_i^n) \right| \leq 1 \tag{1.65}$$

for all eigenvalues λ_p at each U_i^n. The maximum of this quantity over the values of $\mathbf{u}$ arising in a particular problem is the Courant number. Note that (1.65) allows the interaction of waves from neighboring Riemann problems during the time step, provided the interaction is entirely contained within a mesh cell.

Roe's method

Godunov's method requires the solution of Riemann problems at every cell boundary in each time step. Although in theory these Riemann problems can be solved, in practice doing so is expensive, and typically requires some iteration for nonlinear equations. In the Godunov's method the exact solution is averaged over each grid cell, introducing large numerical errors. This suggests that it is not worthwhile calculating the Riemann solutions exactly and that we may be able to obtain equally good numerical results with an approximate Riemann solution obtained by some less expensive means.

One of the most popular approximate Riemann solvers currently in use is due to Roe [82]. The idea is to determine $\hat{\mathbf{u}}(x,t)$, an approximation of the function $\mathbf{u}^*(\mathbf{u}_l, \mathbf{u}_r)$, by solving a constant coefficient linear system of conservation laws instead of the original nonlinear system, i.e., we solve a modified conservation law as described above with flux $\hat{\mathbf{f}}(\mathbf{u}) = \hat{A}\mathbf{u}$. Of course, the coefficient matrix used to define this linear system must depend on $\mathbf{u}_l$ and $\mathbf{u}_r$ in order that

$$\hat{\mathbf{f}}(\mathbf{u}_r) - \hat{\mathbf{f}}(\mathbf{u}_l) = \mathbf{f}(\mathbf{u}_r) - \mathbf{f}(\mathbf{u}_l) \tag{1.66}$$

is satisfied. We could write the linear system for $\hat{\mathbf{u}}$ as

$$\hat{\mathbf{u}}_t + \hat{A}(\mathbf{u}_l, \mathbf{u}_r)\hat{\mathbf{u}}_x = 0. \tag{1.67}$$

This Riemann problem is relatively easy to solve. If $\hat{A}$ has eigenvalues $\hat{\lambda}_i$ and eigenvectors $\hat{\mathbf{r}}_i$, and if we decompose

$$\mathbf{u}_r - \mathbf{u}_l = \sum_p \alpha_p \hat{\mathbf{r}}_p, \tag{1.68}$$

then the approximate Riemann solution $\hat{\mathbf{u}}(x,t) = \hat{\mathbf{w}}(x/t)$ is

$$\hat{\mathbf{w}}(\psi) = \mathbf{u}_l + \sum_{\hat{\lambda}_p < \psi} \alpha_p \hat{\mathbf{r}}_p, \tag{1.69}$$

where the sum is over all p for which $\hat{\lambda}_p < \psi$. Equivalently,

$$\hat{\mathbf{w}}(\psi) = \mathbf{u}_r - \sum_{\hat{\lambda}_p > \psi} \alpha_p \hat{\mathbf{r}}_p. \tag{1.70}$$

Roe suggested that the following conditions should be imposed on $\hat{A}$ for determining it in a reasonable way:

1. $\hat{A}(\mathbf{u}_l, \mathbf{u}_r)(\mathbf{u}_r - \mathbf{u}_l) = \mathbf{f}(\mathbf{u}_r) - \mathbf{f}(\mathbf{u}_l)$. On the one hand, this condition ensures that (1.66) is satisfied, which is necessary to obtain a conservative scheme. On the other hand, in the special case where $\mathbf{u}_l, \mathbf{u}_r$ are the left and right states of a discontinuity, the approximate Riemann solution agrees with the exact Riemann solution. This follows from the fact that the Rankine-Hugoniot condition is satisfied for $\mathbf{u}_l$ and $\mathbf{u}_r$ in this case, so

$$\mathbf{f}(\mathbf{u}_r) - \mathbf{f}(\mathbf{u}_l) = \xi(\mathbf{u}_r - \mathbf{u}_l)$$

   for some ξ, speed of the shock. This condition shows that $\mathbf{u}_r - \mathbf{u}_l$ must, in this situation, be an eigenvector of $\hat{A}$ with eigenvalue ξ, and so, the approximate solution $\hat{u}(x,t)$ also consists of this single jump $\mathbf{u}_r - \mathbf{u}_l$ propagating with speed ξ.

2. $\hat{A}(\mathbf{u}_l, \mathbf{u}_r)$ is diagonalizable with real eigenvalues. In this case, the problem is hyperbolic and solvable.

3. $\hat{A}(\mathbf{u}_l, \mathbf{u}_r) \to \mathbf{f}'(\mathbf{u})$ smoothly as $\mathbf{u}_l, \mathbf{u}_r \to u$. This condition guarantees that the method behaves reasonably on smooth solutions.

One way to guarantee that the last two conditions are satisfied is to take

$$\hat{A}(\mathbf{u}_l, \mathbf{u}_r) = \mathbf{f}'(\mathbf{u}_{ave})$$

for some average value of $\mathbf{u}$, e.g., $\mathbf{u}_{ave} = \frac{1}{2}(\mathbf{u}_l + \mathbf{u}_r)$. Unfortunately, this simple choice of $\mathbf{u}_{ave}$ will not give an $\hat{A}$ that satisfies the first condition in general. Harten and Lax [53] show that for a general system with an entropy function, a more complicated averaging of the Jacobian matrix in state space can be used. This shows that such linearizations exist, but are too complicated to use in practice. Fortunately, for special systems of equations it is possible to derive suitable $\hat{A}$ matrices that are efficient to use relative to the exact Riemann solution. Roe [82] showed how to do this for the Euler equations.

For a scalar conservation law the first condition determines $\hat{a} = \hat{A}(u_l, u_r)$ uniquely as

$$\hat{a} = \frac{f(u_r) - f(u_l)}{u_r - u_l}.$$

The linearized problem is the scalar advection equation $\hat{u}_t + \hat{a}\hat{u}_x = 0$ and the approximate Riemann solution consists of the jump $u_r - u_l$ propagating with speed $\hat{a}$.

1.5.6 High resolution schemes for homogeneous conservation laws

In the relevant literature on numerical schemes for hyperbolic conservation laws, the term High Resolution schemes is applied to those schemes that can produce an *accurate* approximation away from discontinuities, that is the scheme is at least second order accurate on smooth regions, while, at the same time, producing sharp and non-oscillatory discontinuity profiles.

There has been a large body of literature concerning high resolution schemes for conservation laws during the last twenty years. The main issue in most of the existing literature concerns the analysis and development of high resolution schemes for homogeneous conservation laws. Here, we shall mention the ENO and WENO schemes developed initially by Harten, Osher, Shu and collaborators [51] ,[86], [88], which have become a robust and popular option.

In this section, we shall cover only a slightly older alternative to obtain high resolution schemes for homogeneous conservation laws: The flux limited schemes, see e.g. [95], or [70]. In this approach, a judicious hybridization between a high order numerical flux function, typically a version of Lax-Wendroff scheme, and a lower order flux produces the expected high resolution results. This is the type of scheme that we shall use later on in this memoir.

Lax-Wendroff

This is perhaps the best known second order scheme. Its derivation is based on the Taylor expansion with respect to a temporal, rather than spatial, perturbation.

Given the linear equation $u_t + \lambda u_x = 0$ it is clear that

$$u_t = -\lambda u_x \tag{1.71}$$

and, on differentiating with respect to time,

$$u_{tt} = -\lambda u_{xt} = -\lambda u_{tx} = -\lambda(-\lambda u_x)_x = \lambda^2 u_{xx}. \tag{1.72}$$

Now, the Taylor expansion for $u(x, t + \triangle t)$ is

$$u(x, t + \triangle t) = u + \triangle t u_t + \triangle t^2 u_{tt} + \mathcal{O}(\triangle t^3). \tag{1.73}$$

Replacing the time derivatives by (1.71) and (1.72) leads to

$$u(x,t+\triangle t) = u - \lambda \triangle t u_x + \frac{1}{2}\lambda^2 \triangle t^2 u_{xx} + \mathcal{O}(\triangle t^3). \tag{1.74}$$

This is discretized by approximating the space derivatives with central differences and ignoring terms of $\mathcal{O}(\triangle t^3)$ to give

$$U_i^{n+1} = U_i^n - \frac{1}{2}\nu\left(U_{i+1}^n - U_{i-1}^n\right) + \frac{1}{2}\nu^2\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right), \tag{1.75}$$

where $\nu = \lambda\frac{\triangle t}{\triangle x}$. The resulting scheme is second-order accurate in both space and time. In fact it can be shown that the Lax-Wendroff scheme is the unique second-order accurate, spatially centered scheme with three-point support for the linear advection equation.

For nonlinear equations, a derivation of the Lax-Wendroff scheme in conservative form can be made as follows: Assume that the numerical scheme has the following form,

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x}\left(\hat{f}_{i+1/2}^{n+1/2} - \hat{f}_{i-1/2}^{n+1/2}\right). \tag{1.76}$$

The scheme is second order accurate if

$$\hat{f}_{i+1/2}^{n+1/2} = f|_{i+1/2}^n + \frac{\triangle t}{2} f_t|_{i+1/2}^n \tag{1.77}$$

thus, considering

$$\hat{f}_{i+1/2}^{n+1/2} = \frac{1}{2}(f_i + f_{i+1}) + \frac{\triangle t}{2} f_u|_{i+1/2}^n \frac{U_{i+1}^n - U_i^n}{\triangle x}, \tag{1.78}$$

also provides a second order accurate scheme. This derivation provides the form of the Lax Wendroff scheme for scalar conservation laws that we shall use in this memoir

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x}\left(f_{i+1/2}^{n+1/2} - f_{i-1/2}^{n+1/2}\right), \tag{1.79}$$

with

$$f_{i+1/2}^{n+1/2} = \frac{1}{2}(f_i + f_{i+1}) + \frac{\triangle t}{2} f_u|_{i+1/2}^n \frac{U_{i+1}^n - U_i^n}{\triangle x}. \tag{1.80}$$

There are various ways that this can be extended to give a second order method for nonlinear system of conservation laws. If we let $\lambda \equiv$

$A(u) = \mathbf{f}'(\mathbf{u})$ be the Jacobian matrix, then a conservative form of Lax-Wendroff is

$$\begin{aligned} U_i^{n+1} &= U_i^n - \frac{\triangle t}{2\triangle x}\left(\mathbf{f}(U_{i+1}^n) - \mathbf{f}(U_{i-1}^n)\right) \\ &+ \frac{\triangle t^2}{2\triangle x^2}\left(A_{i+\frac{1}{2}}\left(\mathbf{f}(U_{i+1}^n) - \mathbf{f}(U_i^n)\right) - A_{i-\frac{1}{2}}\left(\mathbf{f}(U_i^n) - \mathbf{f}(U_{i-1}^n)\right)\right), \end{aligned} \tag{1.81}$$

where $A_{i\pm\frac{1}{2}}$ is the Jacobian matrix evaluated at $\frac{1}{2}(U_i^n + U_{i\pm1}^n)$. The difficulty with this form is that it requires evaluating the Jacobian matrix, and is more expensive to use than other forms that only use the function $\mathbf{f}(\mathbf{u})$.

One way to avoid using A is to use a two-step procedure. This was first proposed by Richtmyer, and the **Richtmyer two-step Lax-Wendroff method** is

$$\begin{aligned} U_{i+\frac{1}{2}}^{n+\frac{1}{2}} &= \frac{1}{2}(U_i^n + U_{i+1}^n) - \frac{\triangle t}{2\triangle x}\left(\mathbf{f}(U_{i+1}^n) - \mathbf{f}(U_i^n)\right) \\ U_i^{n+1} &= U_i^n - \frac{\triangle t}{\triangle x}\left(\mathbf{f}(U_{i+\frac{1}{2}}^{n+\frac{1}{2}}) - \mathbf{f}(U_{i-\frac{1}{2}}^{n+\frac{1}{2}})\right). \end{aligned} \tag{1.82}$$

Another method of this type is the **MacCormack's method** which uses first forward differencing and then backward differencing to achieve second order accuracy:

$$\begin{aligned} U_i^* &= U_i^n - \frac{\triangle t}{\triangle x}\left(\mathbf{f}(U_{i+1}^n) - \mathbf{f}(U_i^n)\right) \\ U_i^{n+1} &= \frac{1}{2}(U_i^n + U_i^*) - \frac{\triangle t}{2\triangle x}\left(\mathbf{f}(U_i^*) - \mathbf{f}(U_{i-1}^*)\right). \end{aligned} \tag{1.83}$$

Alternatively, we could use backward differencing in the first step and the forward differencing in the second step.

Flux limiter methods

Second order accurate methods such as the Lax-Wendroff scheme give much better accuracy on smooth solutions than the upwind method, but fail near discontinuities, where oscillations are generated. Upwind methods have the advantage of keeping the solution monotonically varying in regions where the solution should be monotone, even though the

accuracy is not very good. The idea behind the *high resolution* Flux-limited schemes is to combine the best features of both methods. Second-order accuracy is obtained where possible, but we do not insist on it in regions where the solution is not behaving smoothly (and the Taylor series expansion is not even valid).

Flux-limiter methods [95], [70] construct a numerical flux of the form

$$F^{TVD}_{i+\frac{1}{2}} = F^{LO}_{i+\frac{1}{2}} + \phi_{i+\frac{1}{2}}(F^{HI}_{i+\frac{1}{2}} - F^{LO}_{i+\frac{1}{2}}) \tag{1.84}$$

where $F^{HI}_{i+\frac{1}{2}}$ is a high-order numerical flux, $F^{LO}_{i+\frac{1}{2}}$ is a low order flux associated with a first-order scheme and $\phi_{i+\frac{1}{2}} = \phi(r_{i+\frac{1}{2}})$. The function $\phi(r)$ is a flux limiter function, whose value depends on the smoothness. Setting $\phi(r) \equiv 1$ for all r gives the Lax-Wendroff method, while setting $\phi(r) \equiv 0$ gives upwind. B. Van Leer [102] proposed the upwind method to measure the smoothness of the data by looking at the ratio of consecutive gradients. In the linear case it amounts to

$$r^n_{i+\frac{1}{2}} = \frac{U^n_{I+1} - U^n_I}{U^n_{i+1} - U^n_i}, \tag{1.85}$$

where $I = i + \mathrm{sgn}(\lambda^n_{i+\frac{1}{2}})$. Here, λ is the wave speed and if $r^n_{i+\frac{1}{2}}$ is close to 1, then the data is smooth, but near a discontinuity we expect that $r^n_{i+\frac{1}{2}}$ is far from 1. We find various methods by choosing various flux-limiter functions as we see in table 1.1 (a), and some high-resolution limiters in table 1.1 (b).

Following the example of Sweby [95], consider the scalar advection equation

$$u_t + \lambda u_x = 0 \tag{1.86}$$

with positive wave speed, $\lambda > 0$. When applied to this equation the Lax-Wendroff scheme could be rewritten in conservation form (1.34), where the numerical flux can be considered as a first-order upwind scheme plus an anti-diffusive correction

$$F^n_{i+\frac{1}{2}} = \lambda u^n_i + \frac{1}{2}\lambda(1-\nu)(u^n_{i+1} - u^n_i), \tag{1.87}$$

where $\nu = \lambda\frac{\triangle t}{\triangle x}$. A flux-limiter method is obtained through the application of the limiter function, $\phi_{i+\frac{1}{2}}$, to the correction term. The numerical flux is clearly of the form (1.84)

$$F^n_{i+\frac{1}{2}} = \lambda u^n_i + \frac{1}{2}\phi(r^n_{i+\frac{1}{2}})\lambda(1-\nu)(u^n_{i+1} - u^n_i). \tag{1.88}$$

Method	$\phi(r)$
Upwind	0
Lax-Wendroff	1
Beam-Warming	r
Fromm	$\frac{1}{2}(1+r)$

(a)

Flux-Limiter	$\phi(r)$
Minmod (ϕ_{mm})	$\max(0, \min(r, 1))$
Superbee (ϕ_{sb})	$\max(0, \min(2r, 1), \min(r, 2))$
Van Leer (ϕ_{vl})	$\frac{\lvert r\rvert + r}{1+\lvert r\rvert}$
MC (ϕ_{mc})	$\max(0, \min((1+r)/2, 2, 2r))$

(b)

Table 1.1: *(a) Linear methods. (b) High-resolution Limiters*

It is easy to check that the limited Lax-Wendroff scheme (1.88) can be put in the form (1.44) by taking

$$C_i^n = \nu + \frac{1}{2}\nu(1-\nu)\left(\frac{\phi(r^n_{i+\frac{1}{2}})}{r^n_{i+\frac{1}{2}}} - \phi(r^n_{i-\frac{1}{2}})\right), \tag{1.89}$$

$$D_i^n = 0. \tag{1.90}$$

The conditions (1.45) are satisfied if

$$0 \le C_i^n \le 1.$$

This in turn holds provided that the CFL condition $0 \le \nu \le 1$ holds, along with the bound

$$\left|\frac{\phi(r^n_{i+\frac{1}{2}})}{r^n_{i+\frac{1}{2}}} - \phi(r^n_{i-\frac{1}{2}})\right| \le 2. \tag{1.91}$$

Since $r^n_{i+\frac{1}{2}}$ and $r^n_{i-\frac{1}{2}}$ in (1.91) are independent, it is required that

$$0 \le \frac{\phi(r)}{r} \le 2 \quad \text{and} \quad 0 \le \phi(r) \le 2 \tag{1.92}$$

for all values of $r \ge 0$ in order to guarantee that the condition (1.91) is satisfied (along with $\phi(r) = 0$ for $r < 0$).

This defines the TVD region (Sweby [95]) in $r - \phi$ plane: the curve $\phi(r)$ must lie in the shaded region in figure 1.6. This figure also shows the

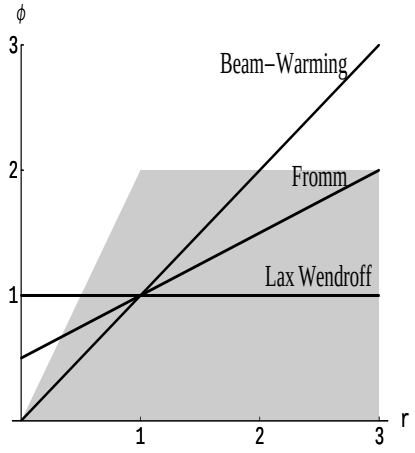


Figure 1.6: *Limiter functions $\phi(r)$. The shaded regions shows where the function must lie for the method to be TVD. The second order linear methods have functions $\phi(r)$ that leave this region.*

functions $\phi(r)$ from table 1.1 (a) for the Lax-Wendroff, Beam-Warming and Fromm methods. All of these functions lie outside the shaded region for some values of r, and indeed these methods are not TVD. Note that for any second order accurate method $\phi(1) = 1$. Sweby found, moreover, that it is best to take ϕ to be a convex combination of $\phi(r) = 1$ (Lax-Wendroff) and $\phi(r) = r$ (Beam-Warming). Other choices apparently produce a lot of compression, and smooth data such as elliptic wave tends to turn into a square wave as time evolves, as is already seen to happen with the superbee limiter (figure 1.8). Imposing this restriction gives the second order TVD region of Sweby, which is shown in figure 1.7.

The high resolution limiter functions from table 1.1 (b) satisfy (1.92), so these limiters all give TVD methods. The functions ϕ are graphed in figure 1.7.

Linear advection results

The upwind TVD scheme (1.89), with a variety of different limiter functions, is applied to a linear advection problem with $\lambda = 1$ and semi-elliptical and square wave initial data. Although the problem is essentially a very simple one, the choice of initial data provides a good test of the ability of the various limiters to capture sharp discontinuities while maintaining smooth profiles where needed, the semi-ellipse being specially difficult to capture numerically due to its steep sides and rounded tip. Figure 1.8 shows the results at time $t = 2$ where the region $0 \leq x \leq 1$

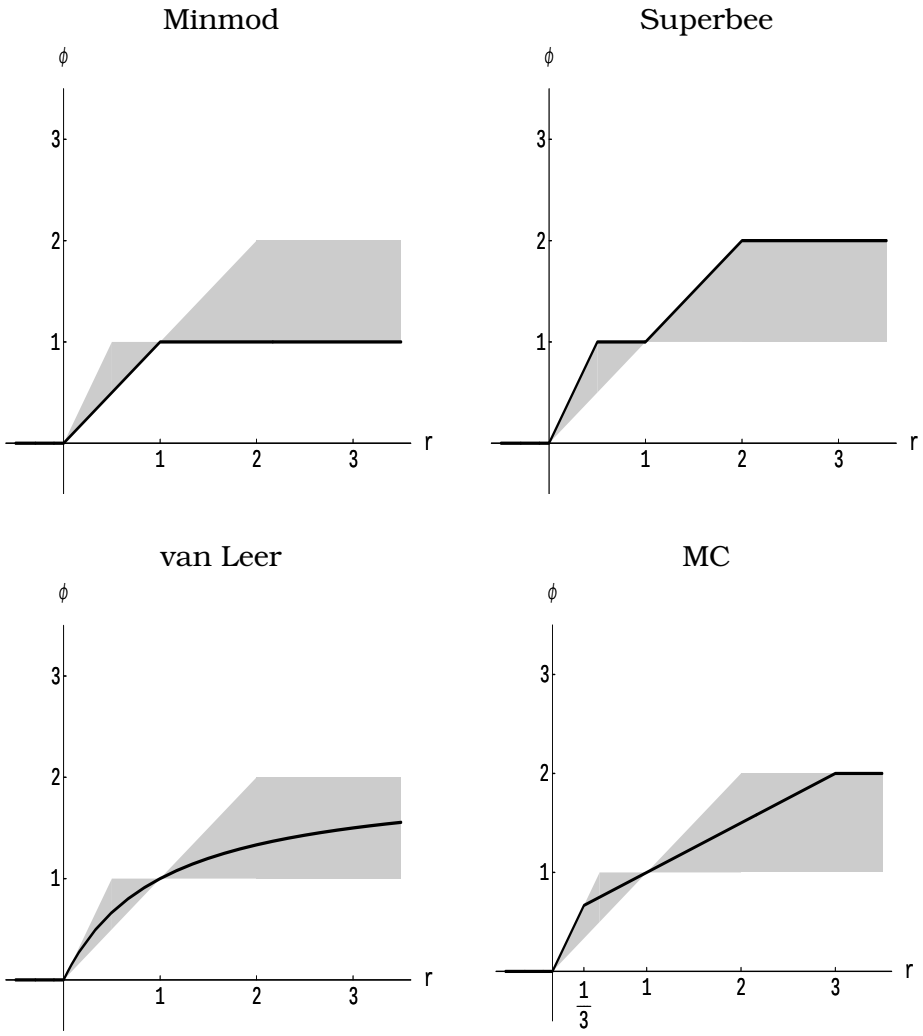


Figure 1.7: *Limiter functions $\phi(r)$: high-resolution TVD limiters. The shaded region is the Sweby region of second-order TVD methods.*

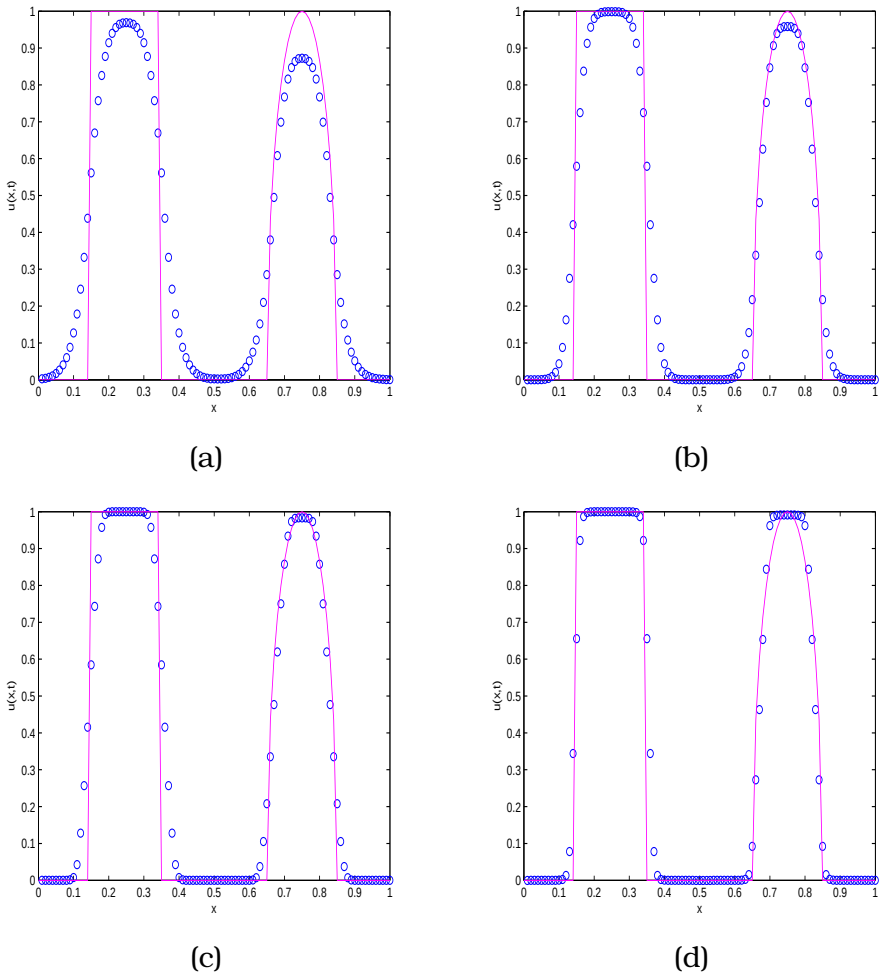


Figure 1.8: *Linear advection test problem (a) Minmod. (b) van Leer. (c) Monotonized central-difference (MC). (d) Superbee.*

is discretized using $\triangle x = 0.01$. For the square wave it is seen that superbee limiter gives remarkable resolution, followed by MC, Van Leer and minmod. The minmod limiter is the most diffusive with a loss in height and a spreading of the profiles, and it is particularly bad at maintaining the sharp discontinuities of the square wave. At the other end of the scale, the superbee limiter proves to be too compressive for the semi-elliptic wave, "squaring off" the smooth data. The van Leer and MC limiters represent something of a trade-off between the two cases previous , the latter being the slightly more compressive.

Slope limiter methods

We can also use a geometric approach to obtain a high-resolution scheme. LeVeque [70] discussed slope-limiter methods in detail. The basic idea is to generalize Godunov's method by replacing the piecewise constant representation of the solution by some more accurate representation, say piecewise linear. Recall that Godunov's method can be viewed as consisting of three steps. To generalize this procedure, we replace the first step by a more accurate reconstruction, taking for example the piecewise linear function

$$\tilde{u}(x, t_n) = U_i^n + \sigma_i^n (x - x_i) \text{ on the cell } [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]. \tag{1.93}$$

Here σ_i^n is a slope on the ith cell which is based on the data U^n. For a system of equations, $\sigma_i^n \in \mathbb{R}^m$ is a vector of slopes for each component of u. Note that taking $\sigma_i^n = 0$ for all i and n recovers Godunov's method. The cell average of $\tilde{u}^n(x, t_n)$ from (1.93) over $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ is equal to U_i^n for any choice of σ_i^n. Since steps 2 and 3 are also conservative, the overall method is conservative for any choice of σ_i^n.

For nonlinear problems we will generally not be able to perform step 2 exactly. The construction of the exact solution $\tilde{u}^n(x, t_n)$ based on solving Riemann problems no longer works when $\tilde{u}^n(x, t_n)$ is piecewise linear. However, it is possible to approximate it in a suitable way.

Let us consider the following example, as discussed by LeVeque [70], for the scalar linear advection equation (1.7), whose exact solution is (1.9). By constructing the piecewise linear function of the first step from the data, we obtain the exact solution

$$\tilde{u}(x, t_{n+1}) = \tilde{u}(x - \lambda \triangle t, t_n), \tag{1.94}$$

and by integrating the exact solution (for $\lambda > 0$) as described in step 3,

we obtain

$$U_i^{n+1} = U_i^n - \frac{\lambda\triangle t}{\triangle x}\left(U_i^n - U_{i-1}^n\right) - \frac{1}{2}\lambda\triangle t\left(1 - \frac{\lambda\triangle t}{\triangle x}\right)\left(\sigma_i^n - \sigma_{i-1}^n\right). \tag{1.95}$$

Now, we must choose the slopes, σ_i^n, such that the scheme (1.95) is second order accurate and satisfies the TVD property. If we choose $\sigma_i^n = 0$, we obtain the first order upwind scheme and by choosing $\sigma_i^n = \frac{1}{\triangle x}(U_{i+1}^n - U_i^n)$, we obtain the classic Lax-Wendroff scheme. If the upwind slopes are used, then overshoots may occur in the linear representation, which results in an increase in Total Variation. Thus, we can view these as being a poor choice of slopes. A better choice of slopes, which makes the scheme second order accurate and satisfy the TVD property, is to use the minmod limiter,

$$\sigma_i^n = \frac{1}{\triangle x}\text{minmod}\left(U_{i+1}^n - U_i^n, U_i^n - U_{i-1}^n\right), \tag{1.96}$$

where

$$\text{minmod}(a,b) = \frac{1}{2}\left(\text{sgn}(a) + \text{sgn}(b)\right)\min\left(|a|,|b|\right). \tag{1.97}$$

It is also interesting to note that by setting

$$\sigma_i^n = \frac{1}{\triangle x}\left(U_{i+1}^n - U_i^n\right)\phi_{i+\frac{1}{2}}^n, \tag{1.98}$$

the scheme reverts back to a flux-limited scheme, where $\phi_{i+\frac{1}{2}}^n$ is a flux-limiter, which was discussed in the previous section. Thus the minmod limiter can be also used as a slope-limiter.

1.6

Characteristic-based schemes for systems of Hyperbolic Conservation Laws

High resolution shock capturing schemes for scalar conservation laws can be extended to systems of homogeneous conservation laws by a so-called *characteristic based* approach. The characteristic based approach has been extensively used in the design of finite-difference Essentially Non Oscillatory (ENO) schemes [87]. In what follows, we explain in some detail the basic mechanisms underlying this technique. Here we shall

explain the basic guidelines; the interested reader can find more information on [31], [27].

Let us consider a system of m convective conservation laws in one spatial dimension,

$$\mathbf{u}_t + [\mathbf{f}(\mathbf{u})]_x = 0. \tag{1.99}$$

In a smooth region of the flow, we can get a better understanding of the structure of the system by expanding out the derivative term as

$$\mathbf{u}_t + J\mathbf{u}_x = 0,$$

where $J = \frac{\partial \mathbf{f}}{\partial \mathbf{u}}$ is the Jacobian matrix of the system. In a hyperbolic system, this matrix is diagonalizable. If L is the matrix whose rows are the left eigenvectors of J and R is the matrix whose columns are the right eigenvectors of J we have

$$LJR = \mathrm{diag}(\lambda^p),$$

and the eigenvectors λ^p are all real.

Let us now fix a state U_0 and consider the linear system

$$\mathbf{u}_t + R_0 J_0 L_0 \mathbf{u}_x = 0, \tag{1.100}$$

where $L_0 = L(\mathbf{u}_0)$, $R_0 = R(\mathbf{u}_0)$, $J_0 = J(\mathbf{u}_0)$. System (1.100) can be equivalently written as follows

$$\mathbf{w}^0_t + J_0 \mathbf{w}^0_x = 0, \tag{1.101}$$

where $\mathbf{w}^0 = L_0\mathbf{u}$. This is a diagonal system, each equation being of the form

$$w_t + \lambda w_x = 0,$$

and we can discretize each scalar equation independently in a λ-upwind biased fashion.

Clearly, when $\mathbf{u} \approx \mathbf{u}_0$, the systems (1.99) and (1.101) are very *close*, hence the local propagation of information mechanisms in (1.99) can be conveniently approximated by those of (1.101). The local upwind directions at the cell boundary $x_{i+1/2}$ could, thus, be obtained by analyzing the Jacobian matrix at an appropriately selected *interface state*.

Let us assume that $\mathbf{u}_{i+1/2}$ is the interface state at the cell boundary $x_{i+1/2}$. The rationale behind the flux computation for $F_{i+1/2}$ put forward in finite-difference Essentially Non Oscillatory (ENO) schemes [87] goes as follows: multiply the entire system by the *constant* left eigenvector matrix $L_{i+1/2} = L(\mathbf{u}_{i+1/2})$ to obtain

$$[L_{i+1/2}\mathbf{u}]_t + [L_{i+1/2}\mathbf{f}(\mathbf{u})]_x = 0. \tag{1.102}$$

According to the local linearization (1.100), (1.101), it is approximately true that the p-th component of this system. i.e. *p-th local characteristic field*, rigidly translates in space at the corresponding *characteristic velocity* $\lambda^p_{i+1/2}$. Hence, we proceed to discretize the $p = 1, \ldots, m$ scalar components of this system independently, using upwind biased differencing with the upwind direction for the p-th equation determined by the sign of $\lambda^p_{i+1/2}$. The corresponding discretization expressed in the original variables is obtained by pre-multiplying the resulting spatially discretized system of equations by $R_{i+1/2} = R(\mathbf{u}_{i+1/2})$:

$$\mathbf{u}_t + R_{i+1/2}\Delta(L_{i+1/2}\mathbf{f}(\mathbf{u})) = 0,$$

where Δ stands for the upwind biased discretization operator.

Thus, if $w^p_s = L^p_{i+1/2}\mathbf{u}_s$, and $\mathcal{F}^p_s = L^p_{i+1/2}\mathbf{f}(\mathbf{u}_s)$, $p = 1, \ldots, m$, $s = i - r, \ldots, i + r + 1$, are the *characteristic variables* and *characteristic fluxes* at the $x_{i+1/2}$ cell boundary, the numerical flux function at this location is obtained as

$$F_{i+1/2} = \sum_{p=1}^{m} F^p_{i+1/2} R^p_{i+1/2}, \tag{1.103}$$

where the *characteristic numerical fluxes* $F^p_{i+1/2}$ are obtained from appropriate upwind discretizations of the components of (1.102).

1.7 Conclusions

In this preliminary chapter, we have revised those theoretical and numerical aspects of conservation laws that are used in this work, whose main objective is to analyze the behavior of certain numerical schemes applied to inhomogeneous conservation laws.

2

Numerical schemes for inhomogeneous conservation laws

In this chapter we are concerned with the numerical solution of hyperbolic equations involving source terms, in particular equations of the form

$$u_t + f(u)_x = s(x,u), \tag{2.1}$$

with initial condition $u(x,0) = u_0(x)$. The source term s is a function of x and $u(x,t)$. Comparing with the homogeneous case presented in the first chapter, two main difficulties arise. The solution, u, needs no longer

be constant along the characteristic of the equation and the slope of the characteristics changes as well. In fact, $u(x,t)$ satisfies

$$\frac{du}{dt} = s(x,u), \tag{2.2}$$

along paths

$$\frac{dx}{dt} = \frac{df}{du}(u), \qquad x(0) = x_0. \tag{2.3}$$

The slope of the paths depends on u (see (2.3)) and needs not be constant, since u is not constant along the characteristics (2.2).

It is known that weak solutions are not necessary unique, and the physical one is characterized by the following entropy condition [65]

$$\int_{-\infty}^{\infty}\int_{0}^{T} \left(\eta(u)\phi_t + F(u)\phi_x\right) dxdt \geq -\int_{-\infty}^{\infty} \eta'(u)s(x,u)\phi(x,t)dxdt, \tag{2.4}$$

where $\phi \in \mathcal{C}^1(\mathbb{R}\times(0,T))$ is any positive test function with compact support in $\mathbb{R}\times(0,T)$, and $\eta \in \mathcal{C}^2(\mathbb{R})$ is a strictly convex entropy function, with corresponding entropy flux function F, that is

$$\eta'(u)f'(u) = F'(u) \quad \forall u \in \mathbb{R}. \tag{2.5}$$

Existence and uniqueness results for the inhomogeneous case (2.1) were first provided by Kružkov in [65]. Many authors have studied particular situations. We shall recall here one particular result, extracted from [43], that is relevant for the model problem considered in chapter 3, where $s(x,u) = s(u)$.

We shall consider that the following assumptions hold,

1. f and s are smooth $\in \mathcal{C}^1(\mathbb{R})$, functions,

2. $s(0) = 0$,

3. in order to avoid any amplitude blow-up phenomena, we assume

$$\exists M \in \mathbb{R}^+ \text{ such that } |u| \geq M \Longrightarrow u \cdot s(u) \leq 0.$$

Theorem 2.1. *([65]) Let us consider the Cauchy problem for*

$$u_t + f_x = s(u), \qquad x \in \mathbb{R}, t > 0,$$

where assumptions 1-3 above hold. $\gamma = \max(s'(u))$ *is a finite number. Then, for* $u_0 \in \mathcal{L}^1(\mathbb{R}) \cap \mathcal{BV}(\mathbb{R})$ *there exists a unique* $u(x,t)$ *entropy solution of the Cauchy problem with initial condition* $u(\cdot,0) = u_0$ *satisfying*

1. $||u||_{\mathcal{L}^\infty} \leq \max(||u_0||_{\mathcal{L}^\infty}, M)$.
2. $TV_x(u(\cdot,t)) \leq e^{\gamma t} TV_x(u_0)$.
3. *Given initial values* u_0, v_0 *such that* $u_0 \leq v_0$*, the corresponding entropy solutions* $u(x,t)$ *and* $v(x,t)$ *satisfy*

$$u(x,t) \leq v(x,t). \tag{2.6}$$

4. *Given initial values* u_0, v_0 *then the corresponding solutions satisfy*

$$\|u(x,t) - v(x,t)\|_{\mathcal{L}^1(\mathbb{R})} \leq e^{\gamma t} \|u_0 - v_0\|_{\mathcal{L}^1(\mathbb{R})}. \tag{2.7}$$

More details on existence, uniqueness and some properties of the solution can be found in [2], [15], [48], [65], [74], [75].

We discuss next a variety of numerical schemes that can be used to numerically approximate equation (2.1). There are various ways to handle the source terms, which fall into two basic categories:

- Unsplit methods, in which a single finite-difference formula is developed to advance the full equation over one time step.

- Fractional step (splitting) methods, in which the problem is broken down into pieces corresponding to the different processes, and a numerical method appropriate for each separate piece is applied independently. This approach is also often used to split multi-dimensional problems into a sequence of one-dimensional problems.

2.1 Fractional step methods

A popular method for treating inhomogeneous hyperbolic equations of the form (2.1) is to use a *fractional-step* or *operator-splitting* technique, in which we somehow alternate between solving the simpler problems

$$\text{Problem A: } u_t + f(u)_x = 0, \tag{2.8}$$

and

$$\text{Problem B: } u_t = s(x,u), \tag{2.9}$$

in order to approximate the solution of the full problem (2.1). This approach, which is described next, is quite simple and it allows us to use high-resolution methods for (2.8) without change, coupling these methods with standard ODE solvers for (2.9), see [72], [73], [100] and [110].

2.1.1 General formulation

To compute the numerical solution of a general scalar problem of the type (2.1), we first find the numerical solution, $\bar{u}^{n+1}$, of (2.8) with initial data $u(x,t_n) = u^n$, then use a numerical ODE solver to obtain u^{n+1} from (2.9) with initial data $u = \bar{u}^{n+1}$. In operator notation this can be written

$$u^{n+1} = \mathcal{B}^{\triangle t}\mathcal{A}^{\triangle t}u^n, \tag{2.10}$$

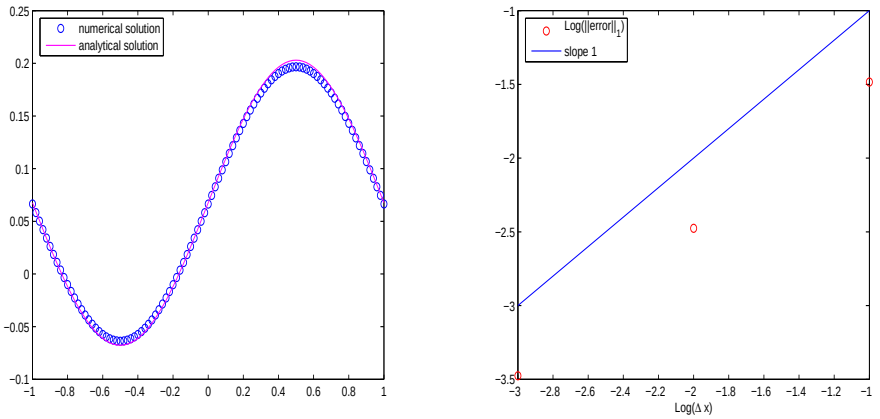
where $\mathcal{A}^{\triangle t}$ represents the numerical solution operator for the homogeneous conservation law over the time step $\triangle t$, and $\mathcal{B}^{\triangle t}$ represents the numerical solution operator for the ODE. For some convergence results, see for example [67], [99], [98].

Although benefitting from simplicity, the above method suffers from being only first-order in time, regardless of the accuracy of the solvers [72], as we can see in the following example. Considering the advection equation with source term

$$u_t + u_x = -u, \tag{2.11}$$

with initial data

$$u(x,0) = \frac{1}{2} + \sin(\pi x) \tag{2.12}$$



First order Order of accuracy

Figure 2.1: *First order fractional step method for* $u_t + u_x = -u$*, with initial condition* $u(x,0) = 1/2 + \sin(\pi x)$*.*

with $x \in [-1,1]$, the first order accuracy of the scheme is clearly seen in figure 2.1, where a splitting technique that uses the Lax-Wendroff second order scheme for the homogeneous part and a second order Runge-Kutta ODE solver for the ODE part has been used.

A slight modification of the splitting idea will yield second-order accuracy quite generally (assuming each subproblem is solved with a method of at least this accuracy). The idea is to solve the second subproblem (2.9) over only a half time step of length $\triangle t/2$. Then we use the result as data for a full time step on the first subproblem (2.8), and finally take another half time step on the second subproblem. We can equally reverse the roles of the subproblems. This approach is often called Strang splitting, as it was popularized in a paper by Strang [94] on solving multidimensional problems. It can be summarized as follows,

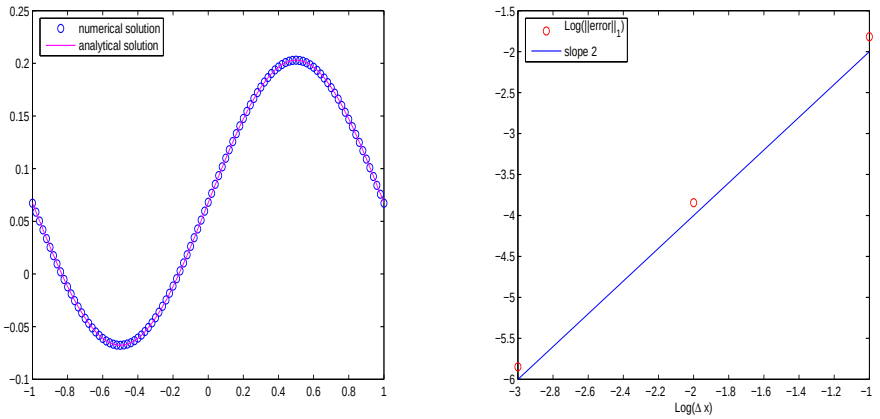
$$u^{n+1} = \mathcal{B}^{\triangle t/2}\mathcal{A}^{\triangle t}\mathcal{B}^{\triangle t/2}u^n, \tag{2.13}$$

which is second-order accurate when $\mathcal{A}$ and $\mathcal{B}$ are at least second-order accurate operators [72], [94].

When several time-steps are taken together the operators can be combined so that (2.13) becomes

$$u^{n+1} = \mathcal{B}^{\triangle t/2}\mathcal{A}^{\triangle t}(\mathcal{B}^{\triangle t}\mathcal{A}^{\triangle t})^{n-1}\mathcal{B}^{\triangle t/2}u^0. \tag{2.14}$$

The resulting scheme becomes almost as efficient to implement as (2.10).



Second order Order of accuracy

Figure 2.2: *Strang splitting method for* $u_t + u_x = -u$*, with initial condition* $u(x,0) = 1/2 + \sin(\pi x)$*.*

As in the first order case, we can check the second order accuracy of Strang's splitting scheme using the same example (see figure 2.2).

Despite their advantages, splitting schemes need to be implemented with caution, especially in the choice of operators. The time evolution of the ODE is dictated by the time-step of the overall scheme, and since an ODE needs to be solved at every time step as well as at every grid point, it is important to choose a method that will remain stable.

The advantages of such an approach are clear since numerical schemes for both (2.8) and (2.9) are well developed and can be chosen to optimal effect. However, there are some situations where this technique does lead to spurious results, or even to wrong numerical solutions.

2.1.2 Special situations

There are situations where a fractional-step method is not adequate, and in this subsection we recall two special situations where the splitting procedure may introduce errors which might be relevant and cannot be ignored.

Stiff source terms

Source terms sometimes model phenomena which occur on much faster time scales than we are attempting to resolve. In this case the source terms are said to be **stiff**, by analogy with the stiff problems of ordinary differential equations. One such example is combustion, where chemical reactions (or nuclear reactions in stars) may occur on much faster time scales than the gas flow, much faster even than high speed detonation waves.

Stiff source terms that are not treated carefully can lead to serious numerical difficulties. Computations may produce waves that look reasonable at first glance and yet are propagating at nonphysical speeds due to purely numerical artifacts. This was observed in a simple model combustion problem by Colella, Majda, and Roytburd [19]. The difficulty of solving such problems was illustrated by LeVeque and Yee [73] who showed that spurious numerical solution phenomena, such as incorrect wave speeds may occur when insufficient spatial and temporal resolutions are used.

In what follows we apply a splitting technique to the model problem proposed by LeVeque and Yee in [73],

$$u_t + u_x = -\mu u(u-1)(u-\frac{1}{2}), \tag{2.15}$$

with initial data

$$u(x,0) = \begin{cases} 1, & x < x_d, \\ 0, & x > x_d, \end{cases} \tag{2.16}$$

where $x \in [0,1]$ and $x_d = 0.3$.

We reproduce below some of the results obtained in [73] by using a second order accurate splitting technique of the form

$$U^{n+1} = \mathcal{B}^{\triangle t/2}\mathcal{A}^{\triangle t}\mathcal{B}^{\triangle t/2}U^n,$$

that alternates between solving the conservation law with no source term (2.8) and then solving the ordinary differential equation (2.9), using the Strang splitting and second order operators for each subproblem.

The splitting operators are defined as follows (see [73]),

$$
\begin{aligned}
\mathcal{B}^{\triangle t/2}: \quad & \left(1-\frac{1}{4}\triangle t s'(U_i^n)\right)\triangle U_i^* = \frac{\triangle t}{2}s(U_i^n) \\
& U_i^* = U_i^n + \triangle U_i^* \\
\mathcal{A}^{\triangle t} \quad & \triangle U_i^{(1)} = -\frac{\triangle t}{\triangle x}\left(f(U_i^*) - f(U_{i-1}^*)\right) \\
& U_i^{(1)} = U_i^* + \triangle U_i^{(1)} \\
& \triangle U_i^{(2)} = -\frac{\triangle t}{\triangle x}\left(f(U_{i+1}^{(1)}) - f(U_i^{(1)})\right) \\
& U_i^{(2)} = U_i^* + \frac{1}{2}\left(\triangle U_i^{(1)} + \triangle U_i^{(2)}\right) \\
& U_i^{**} = U_i^{(2)} + \left(\phi_{i+\frac{1}{2}}^* - \phi_{i-\frac{1}{2}}^*\right) \\
\mathcal{B}^{\triangle t/2} \quad & \left(1-\frac{1}{4}\triangle t s'(U_i^{**})\right)\triangle U_i^{**} = \frac{\triangle t}{2}s(U_i^{**}) \\
& U_i^{n+1} = U_i^{**} + \triangle U_i^{**}.
\end{aligned}
\tag{2.17}
$$

where ϕ^* involves limited fluxes, in order to avoid oscillations and maintain second order accuracy. Their definition can be based either on the intermediate value U^*, as shown below, or on $U^{(2)}$ (see figure 2.3),

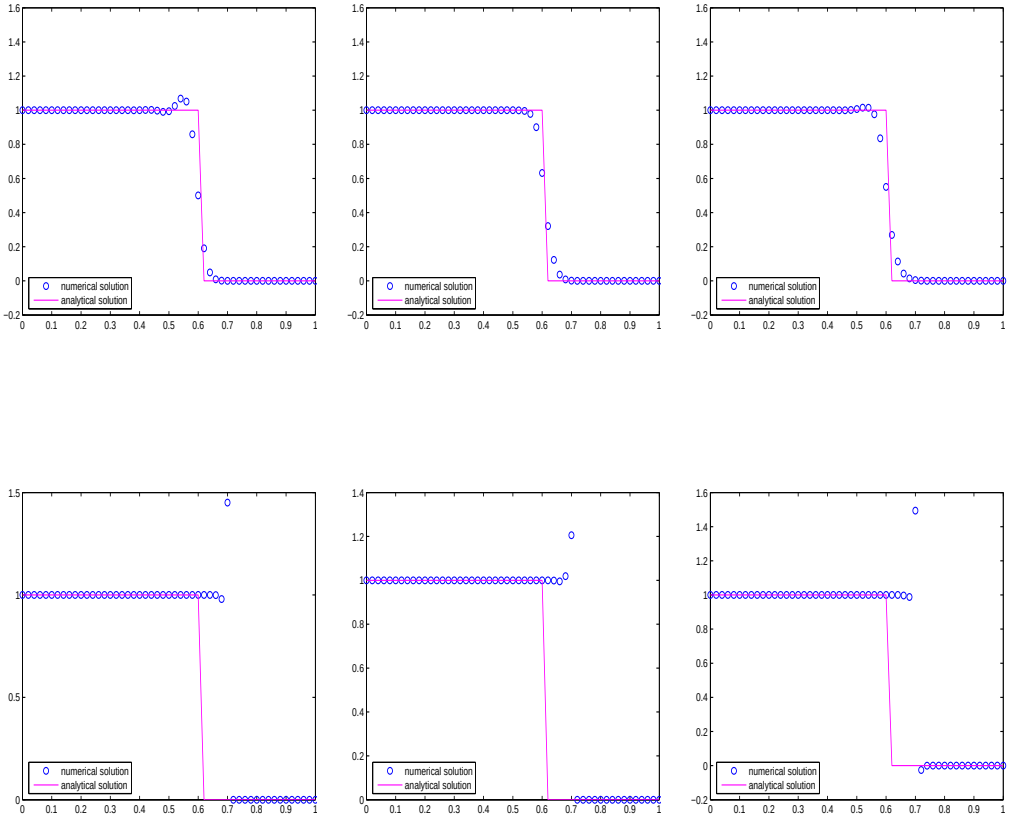
$$
\phi_{i+\frac{1}{2}}^* = \frac{1}{2}\left(\left|\nu_{i+\frac{1}{2}}\right| - \nu_{i+\frac{1}{2}}^2\right)\left(U_{i+1}^* - U_i^* - Q_{i+\frac{1}{2}}\right),
$$

with $Q_{i+\frac{1}{2}}$ chosen as

$$
Q_{i+\frac{1}{2}} \quad = \quad \text{minmod}\left(\triangle_{i-\frac{1}{2}}, \triangle_{i+\frac{1}{2}}, \triangle_{i+\frac{3}{2}}\right) \tag{2.18}
$$

where $\triangle_{i+\frac{1}{2}} = U_{i+1}^n - U_i^n$.

Numerical results are shown in figure 2.3. Notice that, for small $\triangle t\mu$, non-oscillatory results can be obtained when implementing a high resolution TVD scheme for the homogeneous conservation law subproblem. For larger $\triangle t\mu$ (for example, $\triangle t\mu = 15$), a large overshoot appears (with either version of the limiter), which must originate within the ODE solver step, since the TVD scheme considered does not increment the total variation of the data to which it is applied. As noted in [73], the overshoots can be avoided by switching to a different kind of ODE solver, however the discrepancy between the location of the discontinuous profile in the numerical solution and the true solution will persist.



no limiter limit based on U^* limit based on $U^{(2)}$

Figure 2.3: *LeVeque and Yee method for stiff source terms. Top:* $\triangle t\mu = 0.15$. *Bottom:* $\triangle t\mu = 15$

The observed 'delay' in the computed solution is a phenomenon that is intrinsically related to the discretization of the PDE by finite difference techniques. It is simpler to analyze this phenomenon for unsplit techniques and we will do so in chapter 3.

Quasisteady problems

There are some other potential pitfalls in using a fractional-step method to handle source terms in (2.1). This approach performs very poorly in those situations where u_t is small relative to the other two terms, in particular when steady or quasi-steady solutions are being sought. For such solutions, highly accurate numerical simulations can only be obtained from numerical methods that respect the balance that occurs between the flux gradient and the source term when u_t is small.

It is known [71] that this balance is not likely to be respected when using a fractional step approach. In some cases, the fractional-step method may not even converge, oscillating in time near the correct solution. This can happen if a high resolution method, involving limiter functions, is used for the hyperbolic part, since the limiter depends on the solution and effectively switches between different methods based on the behavior of the solution.

For the sake of illustration, let us consider the non-linear balance law presented in [46]

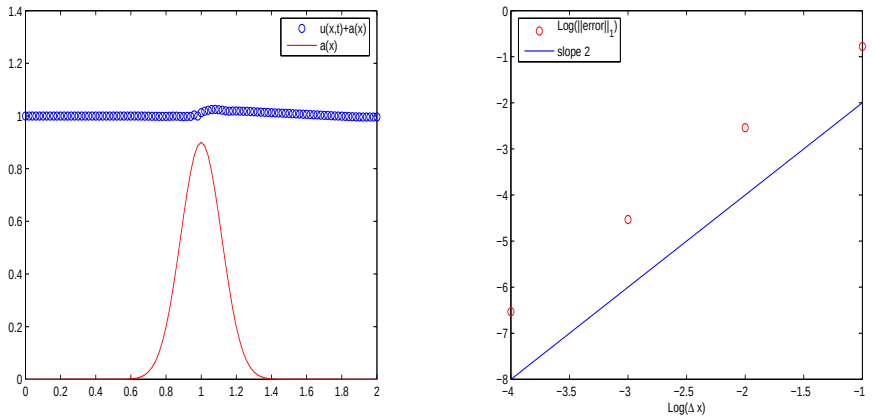
$$u_t + (\frac{u^2}{2})_x = -a_x(x)u. \tag{2.19}$$

where

$$a(x) = 0.9 \begin{cases} 0, & x < 0; \\ (\cos(\pi \frac{x-1}{2}))^{30}, & 0 \leq x \leq 2; \\ 0, & 2 < x. \end{cases} \tag{2.20}$$

with the initial condition $u(x,0) = 1 - a(x)$. This scalar equation is a model for certain types of source terms that are balanced by internal forces, such as those described by the shallow water equations over a nonuniform ocean bottom.

We approximate the solution of this IVP by using Strang Splitting, where the Lax-Wendroff scheme is used for the homogeneous conservation law and second order Runge-Kutta scheme is used for the ODE involving the source term. The results are shown in figure 2.4 (left), where we can observe the occurrence of spurious waves. In fact, the resulting scheme does not preserve the steady states exactly, even though it is second order accurate, see figure 2.4 (right).



Numerical solution Order of Accuracy

Figure 2.4: *Strang splitting method for* $u_t + (u^2/2)_x = -a_x u$ *with initial condition* $u(x,0) = 1 - a(x)$.

Even when a fractional step approach does converge, the numerical steady state obtained might depend on the time step used. This is rather unsatisfying, since the steady solution depends only on x and so we would like the numerical solution generated by a particular method to depend only on $\triangle x$. By contrast, unsplit methods can often be developed in which the steady state is independent of $\triangle t$. We illustrate these facts by considering the following IVP

$$u_t + u_x = -\mu u$$

with $\mu > 0$ and $u(x,0) = e^{-\mu x}$. The exact solution is given by $u(x,t) = u(x-t,0)e^{-\mu t} = e^{-\mu x}$. Let us consider the unsplit method

$$U_i^{n+1} = U_i^n - \frac{\lambda \triangle t}{\triangle x}\left(U_i^n - U_{i-1}^n\right) - \triangle t \mu U_i^n,$$

and suppose we have reached a numerical steady state, so that $U_i^{n+1} = U_i^n$ for all i. This numerical steady-state solution satisfies

$$U_i = \frac{U_{i-1}}{1 + \mu \triangle x}.$$

On the other hand, for the simplest fractional-step method

$$\text{Problem A:} \quad U_i^* = U_i^n - \frac{\triangle t}{\triangle x}\left(U_i^n - U_{i-1}^n\right),$$
$$\text{Problem B: } U_i^{n+1} = U_i^* - \triangle t \mu U_i^*,$$

a numerical steady-state solution would satisfy

$$U_i = \frac{U_{i-1}}{1 + (\mu \triangle t)/(\text{CFL}(1 - \mu \triangle t))},$$

where the CFL$= \frac{\triangle t}{\triangle x}$.

Well Balanced Schemes

The search for numerical schemes that respect the equilibrium that exists between the flux and the source terms in steady state solutions to balance laws such as (2.1) has been an active field of research in the last decade. As we have shown in the previous section, if this balance is not respected, parasitic waves do occur. These have a purely numerical nature, and might be of the same order of the waves one would like to compute. LeVeque in [71] provided examples of quasi-steady flows where this occurs.

The idea of *source-term upwinding* lead Bermúdez and Vázquez-Cendón [5] to formulate the so-called C-property (for Conservation property), which prevents the propagation of parasitic waves in steady and quasi-steady flows. Independently, Greenberg and Leroux [46] coined the term *well-balanced* for schemes that preserve steady states at the discrete level.

From these seminal papers, Well Balanced schemes have been explored and developed in various scenarios, mainly related to shallow water flows, in the recent literature, see e.g. [4], [42], [78], [106] and chapter 4.

In chapter 4, we aim at avoiding spurious oscillations in balance laws by following a strategy described by Gascón and Corberán in [38] and Donat, Caselles and Haro in [10]. In [38], the authors propose to write the source term in divergence form so that it can be incorporated into the flux vector of the homogeneous system to be later discretized in an upwind manner. We shall propose a flux-limiting procedure that avoids spurious oscillations and preserves exactly the steady states in some cases.

3

Numerical schemes for scalar conservation laws with a stiff source term

Many physical problems are governed by hyperbolic conservation laws with non vanishing stiff source terms. These problems could describe the effect of relaxation as in the kinetic theory of gases, chemical reactions, elasticity with memory, water waves, traffic flows, etc.

In some problems the source terms depend only on the solution, i.e. $s(x,u) = s(u)$ and yet the solution naturally develops structures in which the source terms are nonzero, and possibly large, only over very small regions in space. This often happens if the source terms model chemical

reactions between different species (reacting flow) in cases where the reactions happen on time scales much faster than the fluid dynamic time scales. Then, solutions can develop thin reaction zones where the chemical-kinetics activity is concentrated. Such problems are said to have stiff source terms, in analogy with the classical case of stiff ordinary differential equations (ODEs).

Numerical difficulties often appear when the fast reactions are in near-equilibrium during most of the computation. Some time scales, typically those driving the reaction terms, are several orders of magnitude faster than the scale on which the solution is evolving and on which one would like to compute. With many numerical methods, including all explicit methods, taking a time step appropriate for the slower scale of interest can result in violent numerical instability, caused by the fast scales.

Stability, meaning the absence of violent oscillatory behavior, can be achieved by using implicit methods. A variety of excellent implicit methods have been developed for solving stiff systems of ODEs, and many of the same techniques can be applied when the source terms are stiff in order to obtain stable results.

A different type of numerical difficulty is, however, also encountered in numerical simulations concerning hyperbolic PDEs with stiff source terms: the occurrence of fronts propagating at the wrong speeds. This phenomenon was first reported by Colella, Majda and Roytburd [19], as early as 1986, for the numerical simulation of stiff detonation waves. On coarse grids, they obtained a numerical solution which was qualitatively incorrect, with reaction waves traveling at the speed on one mesh cell per time step, which is totally nonphysical.

The analysis carried out in this chapter concerns the model problem stated by LeVeque and Yee in [73], where a prototype initial value problem (IVP) of the form

$$u_t + u_x = s(u) \qquad x \in \mathbb{R} \qquad t > 0, \tag{3.1}$$

$$u(x,0) = u_0(x) \qquad x \in \mathbb{R}, \tag{3.2}$$

with a parameter dependent source term was used to study the behavior of numerical methods with respect to this pathological phenomenon. LeVeque and Yee carry out a numerical study using two types of discrete techniques: A semi-implicit extension of MacCormak's predictor-corrector method, where the fluid dynamics and the chemistry are handled simultaneously, and a time-splitting approach, where one alternates between the solution of the conservation law and the ODE representing

the chemistry. In both cases, they also observe that, for stiff reaction terms, it is possible to obtain stable solutions where the numerical wave profile looks reasonable but is traveling at the wrong speed. In [73], it is argued that this pathological behavior is due to the introduction of nonequilibrium values through numerical dissipation in the advection step.

Ahmad and Berzins in [1] also consider the same model problem and use a Method of Lines (MOL) discretization using monotonicity preserving schemes for the advection step with space/time error balancing in a space-adaptive framework in order to provide an efficient numerical scheme.

Method of Lines discretizations have become a standard procedure when designing high resolution shock capturing schemes for convection-dominated problems. In this chapter, we shall consider the MOL approach on a mesh with a fixed mesh-size, where we use a monotonicity preserving scheme for the advection terms, and we consider explicit, fully implicit and semi-implicit time marching schemes.

The numerical schemes constructed are analyzed with respect to their ability to produce oscillation-free numerical profiles, and we establish conditions on the discretization parameters to obtain non-oscillatory numerical profiles for the model problem for the different numerical schemes considered.

Our study for the fully explicit and fully implicit time marching schemes shows that there is, in fact, a rather direct relation between the mesh spacing and the numerical delay. The delay is, indeed, a by-product of the spatial discretization considered which cannot be avoided.

The semi-implicit time marching schemes considered here are Runge-Kutta Implicit-Explicit (IMEX) schemes, also considered for stiff relaxation problems by Pareschi and Russo [79]. The ability to treat the convective part in an explicit fashion, while still maintaining an implicit handling of the source terms gives a distinct advantage when constructing high order, high resolution numerical schemes. The study of the numerical properties of these schemes, when applied to the model problem has led us to the concept of *Weak Stability Preserving* schemes. These are schemes that preserve a weak non-oscillatory property on the numerical solution *provided* the same property holds for appropriate first order time-discretizations of the basic operators involved. The properties of these schemes with respect to the numerical delay are absolutely similar to that of the first order schemes.

3.1 Model Problem

It is observed in [73] that the essential numerical difficulties to be encountered in the numerical approximation of convection-reaction problems with stiff reaction terms can be identified and studied most easily by looking at the equation

$$u_t + u_x = -\mu u(u-1)(u-\frac{1}{2}) \qquad 0 < x < 1 \qquad t > 0, \tag{3.3}$$

where the parameter $\mu > 0$ controls the stiffness of the problem.

Along the characteristic $x = x_0 + t$, the solution to (3.3) evolves according to the ODE

$$\frac{d}{dt}u(x_0 + t, t) = s(u(x_0 + t, t)), \tag{3.4}$$

where $s(u) = -\mu u(u-1)(u-\frac{1}{2})$. For $\mu > 0$, this equation has a stable equilibria at $u = 0$ and $u = 1$, and an unstable equilibrium at $u = \frac{1}{2}$. Consequently, the solution $u(x,t)$ with an arbitrary initial data $u(x,0)$ rapidly approaches a piecewise constant traveling wave solution $w(x-t)$, where

$$w(x) = \begin{cases} 1, & \text{if } u(x,0) < \frac{1}{2}, \\ \frac{1}{2}, & \text{if } u(x,0) = \frac{1}{2}, \\ 0, & \text{if } u(x,0) > \frac{1}{2}. \end{cases} \tag{3.5}$$

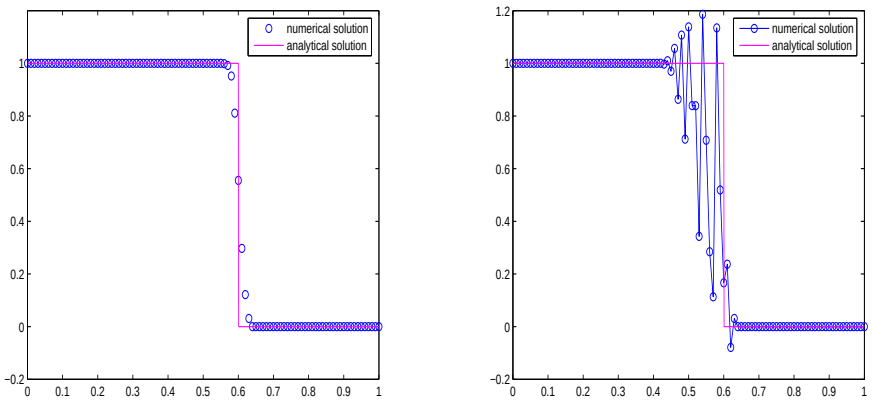
In particular, the solution with piecewise constant initial data

$$u(x,0) = \begin{cases} 1, & \text{if } x < x_d, \\ 0, & \text{if } x > x_d, \end{cases} \tag{3.6}$$

is simply $u(x,t) = u(x-t,0)$. The ODE solution is in equilibrium on each side of the discontinuity, and it theoretically behaves as it would if the source term were not present, and we solved $u_t + u_x = 0$.

All explicit methods taking a time step appropriate for the slower scale of interest can result in violent numerical instability caused by the faster scales. Typically, the computations tend to become very inefficient because the time-step sizes dictated by the stability requirements are much smaller than those required by accuracy considerations for the slowly varying solution.

Let us illustrate these well known facts with a few numerical simulations. Figure 3.1-(a) shows the numerical solution for the simplest first



(a) $\mu = 10$ (b) $\mu = 100$

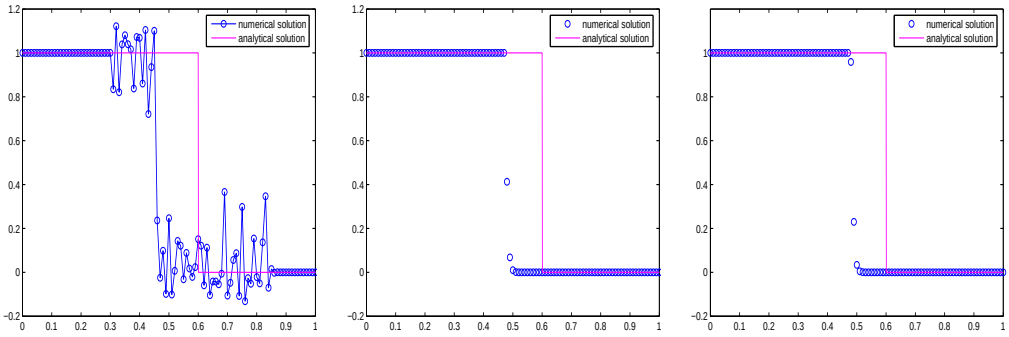
Figure 3.1: *Numerical solution for* $u_t + u_x = -\mu u(u-1)(u-1/2)$ *using the first order explicit scheme.* $\Delta x = 0.01$, $\mathrm{CFL} = 0.9$, $t = 0.3$.

order explicit scheme,

$$U_j^{n+1} = U_j^n - \Delta t \frac{U_j^n - U_{j-1}^n}{\Delta x} + \Delta t s(U_j^n), \tag{3.7}$$

using a mesh size $\triangle x = 0.01$ and CFL$= \frac{\triangle t}{\triangle x} = 0.9$ for a moderate value of μ ($\mu = 10$). When increasing the stiffness of the model, oscillations in the numerical solution are bound to appear. They can clearly be observed in figure 3.1-(b), which corresponds to a simulation with the same parameters (Δx and CFL) and $\mu = 100$. These oscillations become larger when μ increases, to the point of rendering the numerical solution useless. In order to obtain a numerical solution free of unwanted oscillations, the time step has to be reduced (on the same mesh). However, the reduction of Δt only guarantees an effective control over the oscillations developed in the numerical solution, but the approximate solution obtained might still be qualitatively wrong.

In figure 3.2, we display results for the numerical simulation with $\mu = 1000$ on a mesh with $\Delta x = 0.01$. In figure 3.2-(a) we show a typically oscillatory behavior, which is obtained for $\mathrm{CFL} = \Delta t/\Delta x = 0.4$ (i.e. $\mu\Delta t = 4$, for larger values of $\mu\Delta t$ more violent oscillations are observed). In figure 3.2-(b) we have considered $\mathrm{CFL} = 0.2$ (i.e. $\mu\Delta t = 2$). Here we see that the solution looks 'reasonable', however the discontinuous profile is 'delayed' with respect to that of the true solution. This delay persists when we reduce the time step even further, as can be observed in figure

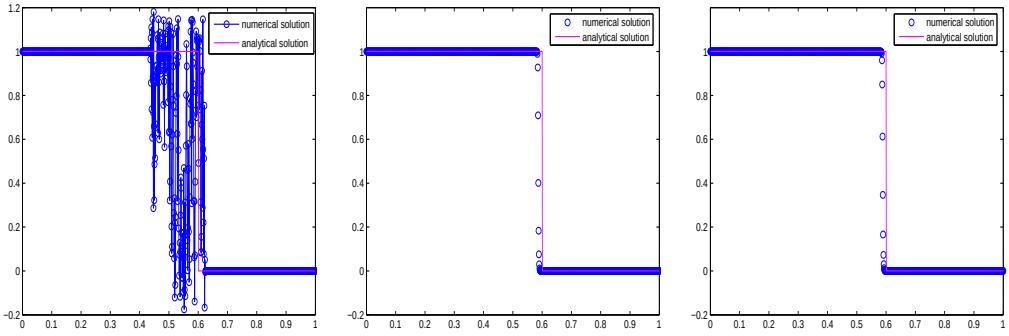


(a) CFL $= 0.4$ (b) CFL $= 0.2$ (c) CFL $= 0.01$

Figure 3.2: *Numerical solution for model problem at* $t = 0.3$ *for* $\mu = 1000$ *using the first order explicit scheme.* $\Delta x = 0.01$.

3.2-(c), where we show the numerical profile for CFL $= 0.01$ ($\mu\Delta t = 0.1$).

A numerical profile moving at the right speed can be obtained after a sufficient level of refinement on the spatial mesh is attained. In figure 3.3 we show relevant results for a finer mesh, where $\Delta x = 0.001$ ($\mu\Delta x = 1$).



(a) CFL $= 0.9$ (b) CFL $= 0.4$ (c) CFL $= 0.2$

Figure 3.3: *Numerical solution for model problem at* $t = 0.3$ *for* $\mu = 1000$ *using the first order explicit scheme.* $\Delta x = 0.001$.

For the numerical treatment of stiff source terms, it is usual to resort to an implicit treatment of, at least, the source terms. Our analysis shows that implicit or semi-implicit treatments of the stiff source terms can control the numerical oscillations in a straightforward manner, but will not solve the problem of the numerical delay in the shock profiles,

which still demands an adequate spatial resolution.

3.1.1 A Fully implicit scheme

In section 3.1 we have considered an Euler discretization of the model problem which is explicit in time. In the numerical treatment of stiff ODEs, it is customary to apply implicit techniques in order to bypass the strict requirements on the time step imposed by the stiffness of the problem. In this section we shall analyze a fully implicit time discretization for the model equation. This technique was proposed by Ahmad and Berzins in [1], as a first step in order to study the effect of neglecting various terms in the nonlinear solvers involved in higher order implicit discretizations of (3.3).

Let us consider the following implicit Euler discretization of (3.3)

$$U_j^{n+1} = U_j^n - \Delta t \frac{U_j^{n+1} - U_{j-1}^{n+1}}{\Delta x} + \Delta t s(U_j^{n+1}), \qquad j = 1, \ldots, N. \tag{3.8}$$

The vector of unknowns $U^{n+1} = (U_1^{n+1}, U_2^{n+1}, \ldots, U_N^{n+1})$ can be computed from the known solution at time t_n, $U^n = (U_1^n, \ldots, U_N^n)$ by applying a Newton procedure on the system

$$U^{n+1} = U^n + \triangle t \; F(U^{n+1}), \tag{3.9}$$

where $F(U)$ in (3.9) is defined as

$$F(U) = L(U) + S(U) \qquad L(U)_j = -\frac{U_j - U_{j-1}}{\Delta x}, \quad S(U)_j = s(U_j). \tag{3.10}$$

We follow Ahmad and Berzins in [1] and consider the vector function

$$G(V) := V - U^n - \triangle t F(V). \tag{3.11}$$

Since U^{n+1} in (3.9) is a root of $G(V)$, they propose to approximate this root the following iterative procedure

$$\begin{aligned} V^{(0)} &= U^n, \\ V^{(m+1)} &= V^{(m)} - (JG(V^{(m)}))^{-1} G(V^{(m)}), \end{aligned} \tag{3.12}$$

where $JG(V) = \dfrac{\partial G(V)}{\partial V}$ is the Jacobian matrix of the system, i.e.,

$$JG(V) = I - \triangle t \frac{\partial L(V)}{\partial V} - \triangle t \frac{\partial S(V)}{\partial V}. \tag{3.13}$$

The expression of $JL := \dfrac{\partial L}{\partial V}$ depends on the advective terms. Clearly, it might be very complicated or even impossible to compute if a nonlinear high order scheme is used (see [1] for a discussion of alternatives in this case), but for the simplest first order spatial discretization operator in (3.20) we easily obtain that

$$JL = -\frac{1}{\triangle x} A_f, \tag{3.14}$$

where A_f is the $N \times N$ matrix

$$A_f = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}, \tag{3.15}$$

where the first row takes into account inflow boundary conditions: $U_0^n = 1 \; \forall n$. Given the form of the solution, since $U_1^{n+1} = U_1^n$, the first row of A_f can be taken as all zero, for simplicity.

The Jacobian $JS := \dfrac{\partial S}{\partial V}$ depends on the source term discretization. For $S(U)$ in (3.10) and $s(u) = -\mu u(u-1)(u-0.5)$, we get that JS is a diagonal matrix with:

$$(JS)_{jj} = -\mu(3U_j^2 - 3U_j + 0.5) = -\mu P(U_j), \tag{3.16}$$

where $P(u) = 3u^2 - 3u + 0.5$. Therefore, the Jacobian matrix

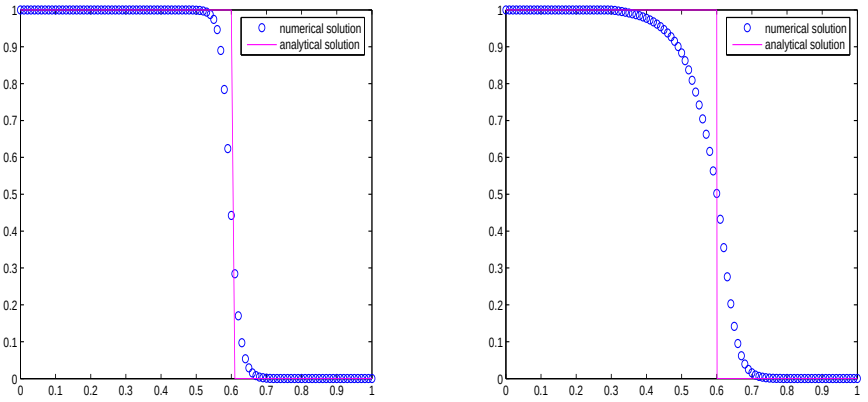
$$JG(U) = I + \frac{\triangle t}{\triangle x} A_f - \triangle t JS, \tag{3.17}$$

is a bi-diagonal matrix.

The use of a fully implicit scheme allows for the use of larger time steps because the stability requirements imposed by the CFL condition of the advective part are no longer necessary.

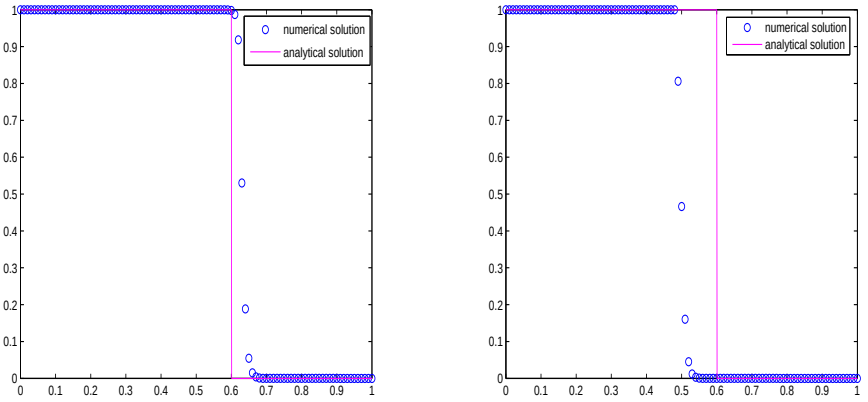
We show in figure 3.4-(a) a numerical simulation obtained with this method for the same parameters used in 3.1-(b). As expected, the numerical solution is non-oscillatory. In figure 3.4-(b), the time step is increased so that $\frac{\triangle t}{\triangle x} = 8$, the solution smoothes out as a result of the larger Δt considered and the fact that it is a first order scheme.

When the stiffness of the model is increased even further, the occurrence of fronts moving at the wrong speeds is observed again. In figure



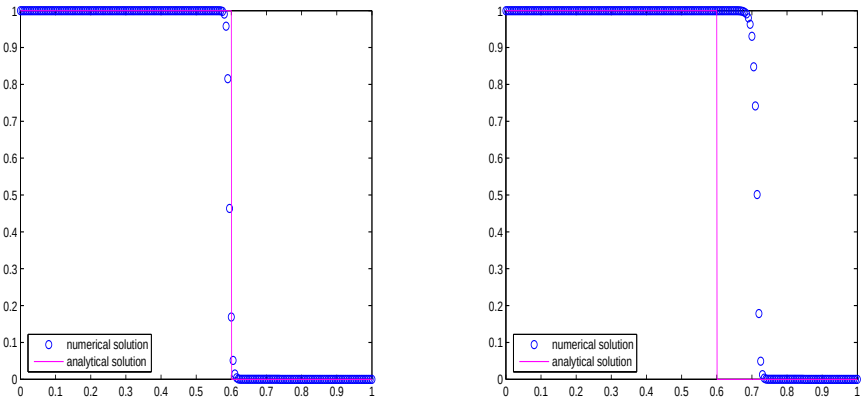
(a) CFL= 0.9 (b) CFL= 8

Figure 3.4: *Numerical solution for model problem with* $\mu = 100$*, at* $t = 0.3$*, using the first order fully implicit scheme.* $\Delta x = 0.01$*. CFL=* $\frac{\triangle t}{\triangle x}$ *.*



(a) CFL= 0.9 (b) CFL= 1.4

Figure 3.5: *Numerical solution for model problem with* $\mu = 1000$ *at* $t = 0.3$*, using the first order fully implicit scheme.* $\Delta x = 0.01$*. CFL=* $\frac{\triangle t}{\triangle x}$ *.*



(a) CFL= 0.9 (b) CFL= 2

Figure 3.6: *Numerical solution for model problem with* $\mu = 1000$ *at* $t = 0.3$*, using the first order fully implicit scheme on a mesh with* $\Delta x = 0.005$*.* CFL= $\frac{\triangle t}{\triangle x}$.

3.5, a front moving at the wrong speed is displayed, due to lack of proper spatial resolution. In both cases, $\triangle x = 0.01$ is used and $\mathrm{CFL} = 0.9$ for the left plot while $\mathrm{CFL} = 1.4$ for the right plot.

We should mention here that the increased stiffness of the case $\mu = 1000$ makes it harder to find, numerically, the solution of the Newton-like procedure in (3.12). For CFL≥ 1.5 the iterative procedure does not seem to converge. For CFL$= 1.4$ we need around sixty (60) iterations to lower the error between consecutive iterations below $1e-10$. For CFL$= 0.9$, we need around thirty (30) iterations instead. On the other hand, the simulations reported in figure 3.4, with $\mu = 100$, only require 3 to 5 iterations for the same tolerance.

Finally, we show in figure 3.6 a simulation with $\mu = 1000$ and $\Delta x = 0.005$. After refining the spatial mesh by a factor of two, we observe that we can increase the CFL number (the value CFL$= 2$ is shown in figure 3.6-(b)). However, we do observe a 'negative delay' in the computed numerical profile.

3.2

A Method-Of-Lines discretization

The application of the method of lines to the model problem of the

previous section reduces the PDE to an initial value problem for a system of ordinary differential equations (ODEs),

$$\frac{\partial U}{\partial t} = F(U,t), \qquad U(0) = (u(x_1,0), u(x_2,0), \ldots, u(x_N,0))^T, \tag{3.18}$$

for the vector $U(t) = (U_1(t), U_2(t), \ldots, U_N(t))^T$ with components $U_i(t) \approx u(x_i,t)$. Due to the nature of the problem, different operators are assigned to the convective derivative and the source term, so that, in general, one has for the model problem

$$F(U,t) = F(U(t)) = L(U(t)) + S(U(t)), \tag{3.19}$$

where

- $L(U)$ is the spatial discretization operator for the advective derivative. Unless specifically stated, we shall consider the simplest, first order operator (also considered in [1])

$$L(U)_j = -\frac{U_j - U_{j-1}}{\triangle x}. \tag{3.20}$$

- $S(U)$ represents the discrete approximation of the source term, which will always be defined in this chapter as

$$S(U)_j = s(U_j). \tag{3.21}$$

For the model problem under examination, the use of the first order *monotone* spatial discretization (3.20) for the convective term, u_x has several nice features. In particular, it can be proved that the solution of this system satisfies certain inequalities that lead to specific bounds which will be relevant in our study of the speed at which the discrete monotone profile moves.

The results in the next section follow from the theory of monotone dynamical systems [57]. We refer the reader to [57], [59] for full details on the underlying theory, and simply include here those definitions and theorems which are necessary in order to proceed with our derivations.

3.2.1 Properties of the solution to the MOL discretization

Let us consider a system of ODEs

$$U' = F(U,t) \tag{3.22}$$

where $F : D \times J \to \mathbb{R}^N$ is a locally Lipschitz vector-valued function, $D \subset \mathbb{R}^N$ an open set and $J \subset \mathbb{R}$ a nontrivial open interval[1]. Let us denote by $\Phi = \{\Phi_t\}_{t\geq}0$ the semiflow that describes the evolution of states in positive time, i.e. the solution of (3.22) with initial value U_0 is given by $U(t) = \Phi_t(U_0)$.

Definition 3.1. *The semiflow Φ is monotone if the maps Φ_t preserve the vector ordering $u \leq v \leftrightarrow u_i \leq v_i, \forall i$. That is if*

$$U_0 \leq V_0 \quad \Longrightarrow \quad U(t) = \Phi_t(U_0) \leq V(t) = \Phi_t(V_0), \qquad \forall t \in J.$$

The system (3.22) is called monotone if the corresponding solution semiflow is monotone.

We shall prove that the system (3.18)-(3.20)-(3.21) is monotone, hence it preserves vector orderings. For this we need the following characterization.

Definition 3.2. *The time-dependent vector field $F : D \times J \to \mathbb{R}^n$, is said to satisfy the quasimonotone condition in D if the following condition is satisfied:*

$$\forall (U,t),(V,t) \in D\times J, \text{ with } U \leq V \text{ and } U_i = V_i \text{ for some } i \ \Rightarrow F_i(U,t) \leq F_i(V,t).$$

This condition is known as the Kamke-Müller condition, *see [62], [76], [57], [59] .*

Theorem 3.1. *([57], theorem 3.2) Assume that F satisfies the quasimonotone condition in D. Then if*

$$U_0 \leq V_0 \quad \Rightarrow \quad U(t) = \Phi_t(U_0) \leq V(t) = \Phi_t(V_0), \quad t \geq t_0, t \in J. \tag{3.23}$$

Hence, (3.22) is monotone. Conversely, if (3.22) is monotone then F satisfies the quasimonotone condition.

This characterization allows us to prove the monotonicity of the system of ODEs that results from the basic MOL discretization described before.

[1] The system (3.22) with the initial value $U(t_0) = u_0$ has a unique continuable solution.

Theorem 3.2. *The system of ODEs given by (3.18)-(3.21) is monotone.*

Proof. Because of theorem 3.1, it is sufficient to see that $F(U)$ satisfies the Kamke-Müller condition. To this aim, consider $U = (U_j)_{j=1}^N$, $V = (V_j)_{j=1}^N$, such that $U \leq V$ and $U_i = V_i$. Then,

$$F_i(U) = \frac{U_{i-1} - U_i}{\triangle x} + s(U_i) = \frac{U_{i-1} - V_i}{\triangle x} + s(V_i) \tag{3.24}$$

$$\leq \frac{V_{i-1} - V_i}{\triangle x} + s(V_i) = F_i(V). \tag{3.25}$$

■

Corollary 3.1. *The solution of the system of ODEs given by (3.18)-(3.21) satisfies the following property*

$$0 \leq U(0) \leq e \quad \Rightarrow \quad 0 \leq U(t) \leq e, \tag{3.26}$$

where $e = (1, \ldots, 1)^T \in \mathbb{R}^N$

Proof. Notice that if $U_i(0) = 0$, $\forall i$ then $U_i(t) = 0$, $\forall i$, $\forall t$. Analogously, if $U_i(0) = 1$, $\forall i$ then $U_i(t) = 1$, $\forall i$, $\forall t$. The result follows immediately from the monotonicity of the system of ODEs, granted in theorem (3.2). ■

It should be noted that these results do not imply, by themselves, the absence of an oscillatory behavior in the vector $(U_i(t))_{i=1}^N$. However, all the numerical evidence gathered so far points out that, when oscillations appear in the numerical solution, they lead to numerical values that lie outside of the unit interval $[0, 1]$. Ensuring that the solutions lie in $[0, 1]$ is a first step towards ensuring a non-oscillatory behavior in the computed solution. We seek to ensure that the numerical solution satisfies a discrete analogy of property (3.26). Numerical schemes that satisfy this property will be referred later on as *weakly stable* schemes.

3.2.2 Wave speed analysis

The numerical results obtained in the previous sections indicate that it is possible to obtain perfectly reasonable results that are stable and free of oscillations and yet are completely incorrect. The analysis below, which is based on simple considerations on the wave speed of the front, shows that the incorrect speed of propagation of the numerical front is, in fact, a by-product of the spatial discretization, that will always be present as long as there are values of the unknown that lie strictly between 0 and 1.

For the model problem (3.3) with initial condition (3.6) we have that

$$\int_0^1 u_t(x,t)dx + \int_0^1 u_x(x,t)dx = \int_0^1 s(u(x,t))dx \tag{3.27}$$

hence, the area under the discontinuous solution

$$\phi(t) := \int_0^1 u(x,t)dx \tag{3.28}$$

satisfies

$$\begin{aligned} \frac{d}{dt}\phi(t) &= \int_0^1 \frac{d}{dt}u(x,t)dx \\ &= -\int_0^1 u_x(x,t)dx + \int_0^1 s(u(x,t))dx \\ &= -\left(u(1,t) - u(0,t)\right) = 1, \end{aligned} \tag{3.29}$$

for $s(u) = -\mu u(u-1)(u-0.5)$. Hence $\phi'(t) = 1$, which is the speed of propagation of the true solution.

On the other hand, if the solution to the system of ODEs (3.18)-(3.21) is a monotone profile of the type shown in the previous sections, if is natural to define the area under the discrete profile, at time t, as

$$\phi_{\triangle x}(t) := \triangle x \sum_{i=1}^{N} U_i(t). \tag{3.30}$$

For a MOL discretization, as established in (3.18), the time variation of

this quantity can also be easily computed:

$$\begin{aligned}\frac{d}{dt}\phi_{\triangle x}(t) &= \triangle x \sum_{i=1}^{N} \frac{d}{dt} U_i(t) \\ &= \triangle x \sum_{i=1}^{N} \left(\frac{-1}{\triangle x}(U_i - U_{i-1}) - \mu U_i (U_i - 1)(U_i - \frac{1}{2}) \right) \\ &= 1 - \mu \triangle x \sum_{i=1}^{N} U_i (U_i - 1)(U_i - \frac{1}{2}),\end{aligned} \tag{3.31}$$

Clearly, $\frac{d}{dt}\phi_{\triangle x}(t)$ represents the velocity of propagation of the discrete front, hence the relation above implies that the discrete profile can move at a speed that can be quite different from that of the true profile. In fact, the discrepancy is equal to the *delay factor*

$$\alpha(\mu \triangle x, U) = \mu \triangle x \sum_{i=1}^{N} U_i (U_i - 1)(U_i - \frac{1}{2}). \tag{3.32}$$

The function $\tilde{s}(u) = u(u-1)(u-0.5)$ satisfies $|\tilde{s}(u)| \leq 5 \cdot 10^{-2}$ for $u \in [0,1]$ and in a typical simulation there are only a small number of points that contribute to the sum, precisely the points at the discrete discontinuous profile, hence it is to be expected that $\alpha(\mu \Delta x, U) = \mathcal{O}(\mu \Delta x 10^{-2})$, but non-zero. Thus, the simplest way to ensure a *correct* speed of propagation for the front is to ensure that $\mu \triangle x$ is below a security threshold.

The need of proper spatial resolution when attempting numerical simulations of hyperbolic PDEs with stiff reaction terms is well known. In [73], the parameter $\mu \Delta t$ was identified as the key parameter for the control of incorrect propagation speeds. The numerical simulations there showed that $\mu \Delta t \approx 1$ was necessary in order to obtain fronts propagating at the correct speeds. However, the need of sufficient spatial resolution for a given stiffness parameter was also recognized, independently of the value of Δt. This derivation shows clearly the origin of the phenomenon and its relation with the mesh spacing.

It is worth noticing that our derivation can be carried out in a similar manner for a convective term of the general form $f(u)_x$, and its discrete equivalent in *conservation form* $(F_{\cdot+1/2} - F_{\cdot-1/2})/(\triangle x)$. In this case we have

$$\frac{d}{dt}\phi(t) = f(u(1,t)) - f(u(0,t)) + \int_0^1 s(u(x,t))dx. \tag{3.33}$$

Hence, if $\int_0^1 s(u(x,t))dx = 0$, the resulting wavefront moves with constant speed, as if the source term was not present.

The equivalent derivation for the solution of the system of ODE's obtained from the MOL discretization

$$\frac{dU}{dt} = L(U) + S(U), \tag{3.34}$$

with

$$(LU)_i = -\frac{F_{i+1/2} - F_{i-1/2}}{\triangle x}, \tag{3.35}$$

would lead to

$$\begin{aligned} \frac{d}{dt}\phi_{\triangle x}(t) &= \triangle x \sum_{i=1}^{N} \left(\frac{-1}{\triangle x}(F_{i+1/2} - F_{i-1/2}) + s(U_i) \right) \\ &= F_{N+1/2} - F_{1/2} - \mu \triangle x \sum_{i=1}^{N} U_i(U_i - 1)(U_i - \frac{1}{2}). \end{aligned}$$

Hence, when the discrete profile is monotone, there is an expected delay in the speed of the numerical wavefront, which is analogous to that of the upwind discretization (3.20).

3.3 Stability Properties of First order MOL Discretizations

In order to carry out a similar analysis for the wave speed of the numerical profile obtained by a fully discrete numerical scheme, we shall analyze first the conditions that ensure the occurrence of a non-oscillatory fully-discrete wave profile.

In general, when the system of ODEs is solved numerically, it is natural to require that the numerical solution satisfies as many qualitative properties of the analytical solution as possible. Stability requirements stem from the desire to have numerical schemes that preserve, at the discrete level, certain properties of the analytic solution of the problem to be solved.

An important class of problems are those whose solutions $U(t)$ satisfy a monotonicity property of the form

$$||U(t)|| \leq ||U(t_0)||, \qquad \forall t \geq t_0 \tag{3.36}$$

for a given norm $||\cdot||$ or semi-norm. For solutions satisfying (3.36), it is natural to require

$$||U^{n+1}|| \leq ||U^n|| \qquad \forall n \geq 0 \tag{3.37}$$

on the numerical solution as well. Methods satisfying this property are called *monotone* or *strongly stable* in the specialized literature (see Appendix A).

Monotonicity, or strong stability, is studied for systems of ODEs, $U' = F(U)$ in a particular class. Usually, it is required that $F(U)$ satisfies an inequality of the type (see e.g. [55])

$$||\rho U + F(U)|| \leq \rho ||U||, \qquad \forall U \tag{3.38}$$

for some fixed $\rho > 0$. This class of problems is denoted by $\mathcal{F}(\rho)$. It is easily seen (see [55], section 1) that this condition implies

$$||U + \tau F(U)|| \leq ||U||, \quad 0 \leq \tau \leq \frac{1}{\rho}, \qquad \forall U, \tag{3.39}$$

and that this also leads to (3.36) for the true solution of the system.

However, the source term considered in the model problem does not allow us to expect *strong stability* (or stability in norm) results, since inequalities like

$$|u + \tau s(u)| \leq |u|, \qquad \forall u \tag{3.40}$$

do not hold for any $\tau > 0$, for the source term in the model problem.

As stated in the previous section, a main issue in the study of the numerical delay is that of ensuring discrete numerical profiles without numerical oscillations. In this section we shall study the conditions that lead us to expect that this property holds.

To this end, we shall introduce a weaker form of stability that seeks to preserve property (3.26), which holds for the true solution of the model, namely that the values of the solution always belong to the interval $[0, 1]$. In this memoir, we refer to this property as *weak stability* (**WS**), as opposed to the *strong stability* (**SS**) just mentioned, which prevents growth in a given norm (or semi-norm).

A numerical scheme is then termed *weakly stable* if it satisfies

$$0 \leq U^0 \leq e \quad \Rightarrow \quad 0 \leq U^n \leq e, \quad \forall n \geq 0. \tag{3.41}$$

3.3.1 Stepsize restrictions for Weak Stability

We start by stating and proving the following lemmas:

Lemma 3.1. *Let $L(U)$ be the operator defined in (3.20). Then,*

$$0 \leq U \leq e \quad \Rightarrow \quad 0 \leq U + \tau L(U) \leq e \tag{3.42}$$

for all $0 \leq \tau \leq \tau_L = \Delta x$.

Proof. For $0 \leq \tau \leq \tau_L = \triangle x$, $0 \leq \tau/\Delta x \leq 1$, hence

$$\begin{aligned} (U + \tau L(U))_i &= U_i - \frac{\tau}{\triangle x}(U_i - U_{i-1}) \\ &= \left(1 - \frac{\tau}{\triangle x}\right) U_i + \frac{\tau}{\triangle x} U_{i-1}, \end{aligned}$$

which is a convex combination of $0 \leq U_i^n, U_{i-1}^n \leq 1$. ■

Lemma 3.2. *Let $s(u) = -\mu u(u-1)(u-0.5)$ with $\mu > 0$. If $0 \leq u \leq 1$ then we have that*

$$0 \leq u + \tau s(u) \leq 1 \qquad 0 \leq \tau \leq \tau_{\mu_+} = \frac{2}{\mu}, \tag{3.43}$$

$$0 \leq u - \tau s(u) \leq 1 \qquad 0 \leq \tau \leq \tau_{\mu_-} = \frac{16}{\mu}. \tag{3.44}$$

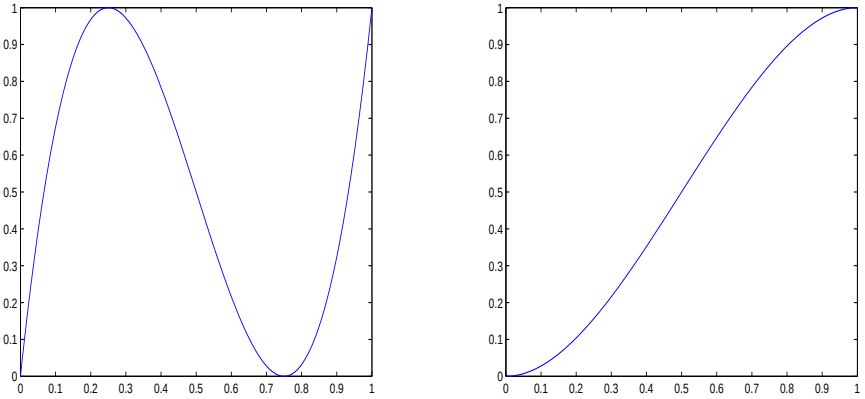
Proof. Let us define the bivariate function

$$g(u, \alpha) := u - \alpha u(u-1)(u-0.5)$$

and notice that $u \pm \tau s(u) = g(u, \pm\tau\mu)$

It is easy to check that

$$\frac{-1}{16} \leq (u-1)(u-\frac{1}{2}) \leq \frac{1}{2} \qquad 0 \leq u \leq 1.$$



(a) $g(u,-16)$ (b) $g(u,2)$

Figure 3.7: *The function $g(u,\alpha)$ for $\alpha=-16$ (left) and $\alpha=2$ (right)*

Then, straightforward manipulations lead to

$$\begin{cases} g(u,\alpha) \geq u(1-\frac{\alpha}{2}), & \alpha \geq 0, \\ g(u,\alpha) \geq u(1+\frac{\alpha}{16}), & \alpha \leq 0. \end{cases}$$

Thus,

$$g(u,\alpha) \geq 0, \text{ for } 0 \leq u \leq 1 \text{ and } -16 \leq \alpha \leq 2. \tag{3.45}$$

Finally, we check that $g(u,\alpha) \leq 1$ for $u \in [0,1]$ and $-16 \leq \alpha \leq 2$. To this aim, notice that

$$\frac{\partial g}{\partial \alpha}(u,\alpha) = -u(u-1)(u-0.5) \begin{cases} \leq 0, & \text{if } u \leq \frac{1}{2}, \\ \geq 0, & \text{if } u \geq \frac{1}{2}. \end{cases}$$

Thus, let $u \in [0,1]$,

- If $u \leq \frac{1}{2}$, then $g(u,\alpha) \leq g_1(u,-16) \leq 1$ (see figure 3.7-(a)).
- If $u \geq \frac{1}{2}$, then $g(u,\alpha) \leq g(u,2) \leq 1$ (see figure 3.7-(b)).

It then follows that

$$0 \leq g(u,\tau\mu) = u + \tau s(u) \leq 1 \quad \text{for} \quad 0 \leq \tau \leq \frac{2}{\mu} = \tau_{\mu_+}, \tag{3.46}$$

$$0 \leq g(u,-\tau\mu) = u - \tau s(u) \leq 1 \quad \text{for} \quad 0 \leq \tau \leq \frac{16}{\mu} = \tau_{\mu_-}. \tag{3.47}$$

■

With these two lemmas we can prove a *weak stability* result for the numerical solution obtained with the explicit Euler scheme used in section 3.1.

Theorem 3.3. *If $0 \leq U^0 \leq e$ and U^n is computed with an explicit Euler discretization of the MOL system (3.18)-(3.20)-(3.21), then*

$$0 \leq \tau \leq \frac{2\Delta x}{2 + \mu \Delta x} \quad \Rightarrow 0 \leq U^n \leq e, \quad \forall n \geq 0. \tag{3.48}$$

Proof. Let $\alpha, \beta \in \mathbb{R}$ be such that $0 < \alpha, \beta < 1$ and $\alpha + \beta = 1$. We write

$$\begin{aligned} U^{n+1} &= U^n + \tau F(U^n) = U^n + \tau L(U^n) + \tau S(U^n) \\ &= \alpha \left(U^n + \frac{\tau}{\alpha} L(U^n) \right) + \beta \left(U^n + \frac{\tau}{\beta} s(U^n) \right) \end{aligned}$$

Given any $0 < \alpha < 1$, the two terms in the sum above remain in $[0, 1]$ provided that

$$0 \leq \frac{\tau}{\alpha} \leq \tau_L = \Delta x, \qquad 0 < \frac{\tau}{1 - \alpha} \leq \tau_{\mu_+} = \frac{2}{\mu}$$

since in this case lemmas 3.1 and 3.2 apply.

Hence, U^{n+1} is a convex combination of two vectors whose components are between 0 and 1 provided

$$\tau \leq \min\{\alpha \Delta x, (1 - \alpha) \frac{2}{\mu}\}$$

for any given α in $(0, 1)$. For

$$\alpha = \frac{\tau_{\mu_+}}{\tau_L + \tau_{\mu_+}} = \frac{\frac{2}{\mu}}{\Delta x + \frac{2}{\mu}} \tag{3.49}$$

we have $\tau_L \alpha = (1 - \alpha)\tau_{\mu_+} = \dfrac{2\Delta x}{2 + \mu \Delta x}$, which proves the result.

■

This theorem states only sufficient conditions in order to get a discrete solution whose values are always between zero and one. It does not

guarantee the absence of oscillations in the numerical solution. However, the numerical results in section 3.1, show that the occurrence of oscillations always comes associated to values on the solution that exceed 0 and/or 1. Thus, ensuring that the discrete solution values lie in $[0, 1]$ strongly indicates the absence of oscillatory behavior.

Notice that condition (3.48) can be translated into a CFL condition for weak stability.

Corollary 3.2. *If $0 \leq U^0 \leq e$ and U^n is computed with an explicit Euler discretization of the MOL system (3.18)-(3.20)-(3.21), then*

$$0 \leq \text{CFL} \leq \frac{2}{2 + \mu\Delta x}, \quad \Rightarrow 0 \leq U^n \leq e, \quad \forall n \geq 0 \tag{3.50}$$

where $\text{CFL} = \Delta t/\Delta x$.

For $\mu\Delta x = 1$, the corollary above ensures 'non-oscillatory' results for $\text{CFL} \leq 2/3 \approx 0.66$. Notice the oscillatory profiles obtained in figures 3.1-(b) and 3.3-(a), where the CFL considered well exceeds this value.

In figure 3.2, we have $\mu\Delta x = 10$, so that the theorem ensures a 'non-oscillatory' discrete profile for $\text{CFL} \leq 0.16$. In 3.2-(a) the CFL is far from this bound and we get an oscillatory profile. However, it should be noted that the derivation only ensures that condition (3.50) is a sufficient condition. A non-oscillatory profile can be seen in 3.2-(b), obtained for $\text{CFL} = 0.2$ which is slightly larger than the CFL condition in (3.50) for this simulation.

We examine next the stepsize restrictions for the fully implicit Euler discretization of the MOL system of section 3.1.1. We start by proving analogous weak stability results for the convective and source-term discrete operators.

Lemma 3.3. *Let $L(U)$ be the operator defined in (3.20) and define*

$$U^{n+1} = U^n + \tau L(U^{n+1})$$

implemented with inflow boundary conditions at the left boundary so that $U_0^n = 1$, $\forall n \geq 0$. Then for all $\tau > 0$

$$0 \leq U^0 \leq e \quad \Rightarrow \quad 0 \leq U^n \leq e. \tag{3.51}$$

Proof. Let us assume that $0 \leq U_j^n \leq 1$, $\forall j$, but $\exists j_0$ such that $U_{j_0}^{n+1} < 0$. Then, since

$$U_{j_0}^n = U_{j_0}^{n+1} - \tau L(U^{n+1})_{j_0} = U_{j_0}^{n+1} + \frac{\tau}{\Delta x}(U_{j_0}^{n+1} - U_{j_0-1}^{n+1})$$

and $U_{j_0}^n > 0$, we must have $(U_{j_0}^{n+1} - U_{j_0-1}^{n+1}) > 0$, that is $U_{j_0}^{n+1} > U_{j_0-1}^{n+1}$. Hence $U_{j_0-1}^{n+1} < 0$ and, proceeding recursively, we would arrive at $U_0^{n+1} < 0$, which contradicts the inflow boundary conditions.

A similar argument shows that it is not possible to have $U_{j_0}^{n+1} > 1$ either.

■

Lemma 3.4. *Let us consider the ODE*

$$u'(t) = s(u)$$

with $s(u) = -\mu u(u-1)(u-0.5)$ *and* $\mu > 0$. *If* $0 \leq u^n \leq 1$ *then the implicit Euler method*

$$u^{n+1} = u^n + \tau s(u^{n+1}), \tag{3.52}$$

satisfies that $0 \leq u^{n+1} \leq 1$ *for any step size* $\tau \geq 0$, *provided* u^{n+1} *exists.*

Proof. We write

$$u^n = u^{n+1} - \tau s(u^{n+1}) = u^{n+1} + \tau\mu u^{n+1}\left(u^{n+1} - 1\right)\left(u^{n+1} - \frac{1}{2}\right).$$

Suppose that $u^{n+1} < 0$, then $\tau\mu u^{n+1}\left(u^{n+1} - 1\right)\left(u^{n+1} - \frac{1}{2}\right) \leq 0$, and so

$$u^n = u^{n+1} + \tau\mu u^{n+1}\left(u^{n+1} - 1\right)\left(u^{n+1} - \frac{1}{2}\right) < 0,$$

which is a contradiction because $0 \leq u^n \leq 1$. In a similar way it is possible to prove that $u^{n+1} \leq 1$. ■

We prove next an unconditional weak stability result for the numerical profiles obtained with the fully implicit scheme.

Theorem 3.4. *If $0 \leq U^0 \leq e$ and U^n is computed with an implicit Euler discretization of the MOL system (3.18)-(3.20)-(3.21), implemented with inflow boundary conditions at the left boundary so that $U_0^n = 1, \forall n \geq 0$. Then*

$$0 \leq U^n \leq e, \quad \forall n \geq 0, \quad \forall \tau \tag{3.53}$$

provided U^n can be computed.

Proof. We write

$$U^{n+1} = U^n + \tau \left(L(U^{n+1}) + S(U^{n+1}) \right)$$

as

$$U^n = U^{n+1} - \tau S(U^{n+1}) - \tau L(U^{n+1}).$$

Let us assume that $0 \leq U_j^n \leq 1 \ \forall j$, but $\exists j_0$ such that $U_{j_0}^{n+1} < 0$. Then

$$U_{j_0}^{n+1} - \tau S(U^{n+1})_{j_0} = U_{j_0}^{n+1} + \tau \mu U_{j_0}^{n+1}(U_{j_0}^{n+1} - 1)(U_{j_0}^{n+1} - 0.5) < 0.$$

Since

$$U_{j_0}^n = U_{j_0}^{n+1} - \tau S(U^{n+1})_{j_0} - \tau L(U^{n+1})_{j_0}$$

and $U_{j_0}^n > 0$, we must have $0 < U_{j_0}^n < -\tau L(U^{n+1})_{j_0}$, hence $0 < -U_{j_0-1}^{n+1} + U_{j_0}^{n+1}$, and $U_{j_0-1}^{n+1} < 0$. Applying this recursively, we would have $U_0^{n+1} < 0$, which contradicts the inflow boundary conditions.

A similar argument shows that it is not possible to have $U_{j_0}^{n+1} > 1$ either.

■

Thus, we are led to expect monotone profiles for numerical solutions to the model problem obtained with the fully implicit first order scheme *provided* these solutions can be computed. These results are absolutely consistent with the numerical simulations shown in section 3.1.1. In practice, as we have observed in section 3.1.1, an increase in the CFL number seems to lead to convergence problems in the Newton-like iterative process.

3.3.2 A numerical study of the discrete wave speed

LeVeque and Yee in [73] define the *discrete wave speed* at time t_{n+1}, $ws(n+1)$ as

$$ws(n) = \frac{\Delta x}{\Delta t}\left(\sum_j U_j^{n+1} - \sum_j U_j^n\right), \tag{3.54}$$

which is, in fact, a discrete realization of (3.31).

For the IVP problem studied in this chapter, and the methods considered so far, a straightforward calculation shows that for the explicit Euler method we have

$$ws(n+1)^{\mathrm{EE}} = 1 - \alpha(\mu\Delta x, U^n),$$

while for the (fully) implicit Euler method we have

$$ws(n+1)^{\mathrm{IE}} = 1 - \alpha(\mu\Delta x, U^{n+1}),$$

where $\alpha(\mu\Delta x, U)$ is as defined in (3.32).

Hence, provided we have a non-oscillatory numerical profile, we expect a numerical delay in the computed solution of order $\mathcal{O}(10^{-2}\mu\Delta x)$. For $\mu\Delta x = 10$ we expect, thus, a noticeable numerical delay, of $\mathcal{O}(10^{-1})$, while for $\mu\Delta x \approx 1$, we expect the delay to be $\mathcal{O}(10^{-2})$ and, hence, barely visible in the displayed figures. These observations are in agreement with the numerical results shown in the previous section, both for the explicit and for the implicit schemes.

In figure 3.8, we show a plot of $ws(n)^{\mathrm{IE}}$ for $\mu = 1000$ and $\mu\Delta x = 10$ and $\mu\Delta x = 2.5$ for $\mathrm{CFL} = 0.9$. Here, it can be clearly appreciated that the delay is proportional to $\mu\Delta x$.

We notice also that in 'borderline cases', like in figure 3.6 where $\mu\Delta x = 5$ and we expect the delay to be proportional to $5{\cdot}10^{-2}$, the influence of the term $\sum_i U_i(U_i-1)(U_i-0.5)$ might also be relevant, and that this influence seems to depend on the CFL considered. In figure 3.6-(a) the CFL is 0.9 and the delay is barely noticeable, while in 3.6-(b) the CFL=2 and the numerical solutions is ahead of the true solution.

We also observe that the function $\alpha(\mu\Delta x, U)$ may vary in a regular, but sometimes oscillatory, manner in the numerical simulations. In figure 3.9, we show results for a smaller CFL number. Here $\mathrm{CFL} = 0.6$ and we observe a regular but oscillatory pattern in $1 - \alpha(\mu\Delta x, U^n)$, which is

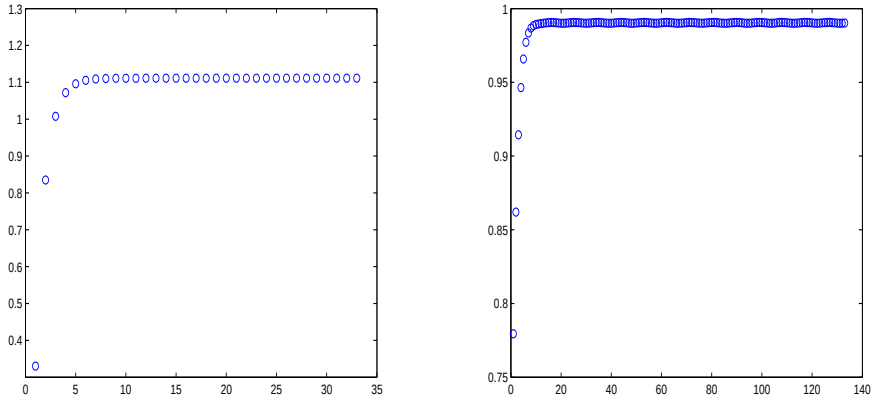


Figure 3.8: $1 - \alpha(\mu\Delta x, U^n)$ *as a function of* n. *Left:* $\mu\Delta x = 10$. *Right:* $\mu\Delta x = 10/4$. $\mathrm{CFL} = 0.9$

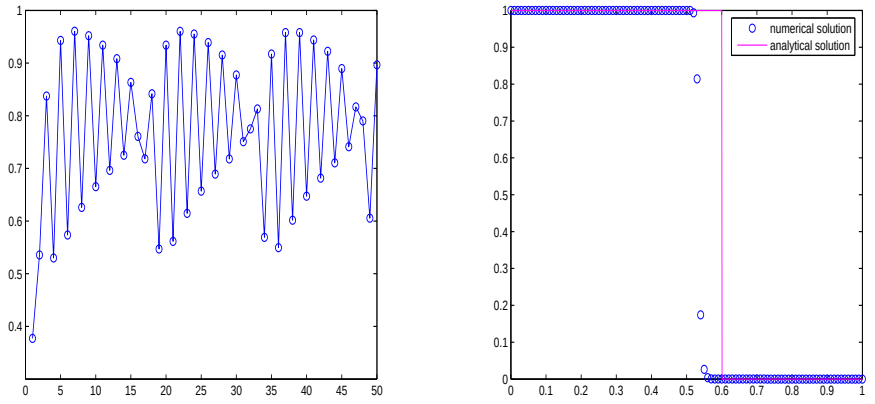
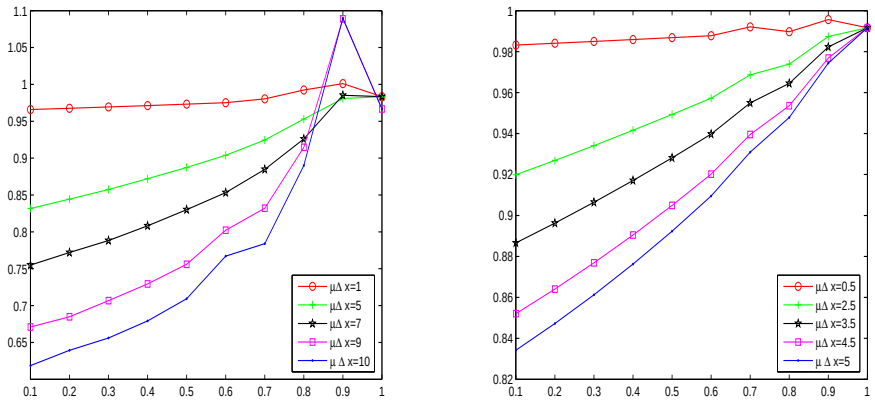


Figure 3.9: $\mu\Delta x = 10$, $\mathrm{CFL} = 0.6$. *Left:* $1 - \alpha(\mu\Delta x, U^n)$ *as a function of* n. *Right numerical wave profile*



(a) $\Delta x = 0.01$ (b) $\Delta x = 0.005$

Figure 3.10: *Average speed for the fully implicit scheme as a function of CFL for different values of* $\mu = 100, 500, 700, 900, 1000$. $t = 0.3$.

due to the influence of the points in the discrete profile that lie strictly between 0 and 1.

In order to compare the numerical delay for different values of the parameters involved, LeVeque and Yee [73] define an *average speed*, as the average of the function $ws(n)$over a fixed time interval $[t_0, t_n]$

$$\text{average speed } = \frac{\Delta x}{(t_n - t_0)} \left(\sum_j U_j^n - \sum_j U_j^0 \right). \tag{3.55}$$

In figure 3.10 we show the average speed for various values of $\mu\Delta x$ as a function of the CFL number, on two different grids. It can be readily observed that there is always a delay with respect to 1, the value of the speed for the model problem. For a given CFL number the delay increases for increasing values of $\mu\Delta x$. The dependence on the number $\mu\Delta x$ is confirmed by looking at the line corresponding to $\mu\Delta x = 5$ in figures 3.10-(a) and (b), where we see that the behavior is similar for the different meshes considered.

Also, on average, for each value of $\mu\Delta x$, the delay is seen to be of $\mathcal{O}(\mu\Delta x 10^{-2})$. Notice that, for a given $\mu\Delta x$, decreasing the value of the CFL number does not improve the numerical delay.

The plots in figure 3.10 only cover the range $\text{CFL} \in [0.1, 1]$. We have observed that, even though the CFL number can be increased beyond one for the fully implicit scheme, in doing so we often run into conver-

gence problems with respect to the Newton-like iterative procedure used to compute U^{n+1}. We have already mentioned this problem in section 3.1.1.

From a practical point of view, it can be inferred that a fully implicit treatment of all terms present in the equation does not represent an advantage over semi-implicit treatments, as those proposed in [73]. Clearly, when dealing with higher order, nonlinear discretizations of the convective derivative, the Newton like procedure has to be modified, since the Jacobian JL in (3.14) cannot be computed. Ahmad and Berzins [1] propose to use a monotonicity preserving advection scheme, combined with space/time error balancing in an adaptive mesh framework and a Gauss-Seidel iteration, in order to provide an efficient solver.

Instead of constructing 'ad-hoc' techniques to increase the accuracy and/or the resolution power, we are interested in exploring the use Implicit-Explicit (IMEX) Runge-Kutta schemes, as a general procedure to treat stiff reaction terms. In the following section we study the properties of these schemes, and the numerical wave profiles that can be obtained when using them for the model problem.

3.4 IMEX-RK: Implicit-Explicit Runge-Kutta schemes

In the finite difference context pursued here, MOL discretizations have the advantage of decoupling the issues of spatial and temporal accuracy, which can be handled separately. High accuracy in space is obtained by assuming that the discretization of the convective term is of the general conservative form $L(U)_i = -(F_{i+1/2} - F_{i-1/2})/\triangle x$, where $F_{i+1/2}$ is an appropriately defined numerical flux at the $i + 1/2$ cell interface. Higher order accuracy in time is often achieved by integrating the MOL system with a high order accurate ODE solver, and Runge-Kutta schemes are often the preferred choice.

This general technique is more amenable than those proposed in [73] for the model problem. We remark that ad-hoc second order schemes, such as the semi-implicit generalization of MacCormack's scheme to the non-homogeneous case proposed in [73] are hard to generalize to more complicated situations.

IMEX (IMplicit-EXplicit) Runge Kutta methods allow for an explicit

treatment of the convective terms, so that high resolution conservative discretizations can be used in a straightforward manner, while still maintaining the implicit treatment of the stiff source terms necessary for stability. In this sense, they provide, also, a more general framework for the numerical treatment of these problems than that used in [1] where a TVD-like discretization of the convective term is used together with a backward Euler time integrator in a fully-implicit mode. In this work, some modifications need to be made in the Newton procedure involved in the computation of the numerical solution at each time step, since the Jacobian JL cannot be computed for nonlinear discretizations of the convective term.

IMEX schemes were proposed and analyzed as far back as the late 1970's [103], [22]. Instances of these methods have been successfully applied to the incompressible Navier-Stokes equations [63] and in environmental modeling studies [105]. A systematic, comparative study for PDEs of convection-diffusion type was carried out in [3], and a corresponding study for reaction-diffusion problems arising in morphology is reported in [84]. The IMEX framework has been also successfully applied to hyperbolic systems with stiff relaxation terms (see [9], [110], [79]).

In this section we describe the application of IMEX-RK techniques to the system of ODEs (3.18)-(3.19), derived from a MOL discretization of the model problem. In addition, we shall also analyze their behavior with respect to the concept of weak stability previously defined, as well as with respect to the phenomenon of incorrect propagation speeds.

As in [79], simplicity and efficiency in solving the algebraic equations corresponding to the implicit terms at each stage of the IMEX-RK process is of paramount importance. This strongly suggests to consider only *Diagonally Implicit Runge-Kutta* (**DIRK**) schemes for the source term, and we shall restrict our presentation to this class of IMEX schemes. The general formulation of an s-stage IMEX-RK scheme in this class, applied to the model problem, is as follows [79],

$$U^{(i)} = U^n + \triangle t \sum_{j=1}^{i-1} a_{ij} L(U^{(j)}) + \triangle t \sum_{j=1}^{i} \tilde{a}_{ij} S(U^{(j)}), \quad 1 \leq i \leq s \tag{3.56}$$

$$U^{n+1} = U^n + \triangle t \sum_{i=1}^{s} b_i L(U^{(i)}) + \triangle t \sum_{i=1}^{s} \tilde{b}_i S(U^{(i)}), \tag{3.57}$$

where $U^{(i)}$ represent the internal stages of the method. When $c_i = \sum_{j=1}^{s} a_{ij} = \tilde{c}_i = \sum_{j=1}^{s} \tilde{a}_{ij}$, the internal stages approximate the solution at time $t_n + c_i \Delta t$.

The general technique to derive order conditions for Runge-Kutta schemes is based on Taylor expansions. The explicit form of the order conditions for IMEX Runge-Kutta schemes up to order $p = 3$, obtained by imposing that the true solution and the numerical solution agree up to order $(\Delta t)^{p+1}$, can be found in [79].

An IMEX method allows an explicit treatment of the convective derivative, which can be performed by a state of the art high resolution conservative technique. We notice that if $F_{i+1/2} = U_i$, then we obtain the first order discretization (3.20).

We remark here that IMEX-RK schemes fall within the wider class of *Additive Runge Kutta* methods. We refer the reader to [56] (see also Appendix A, section A.4) for a detailed description of ARK schemes and some of their properties.

RK methods are often represented in a so-called compact form by using their *coefficient scheme*. We refer the reader to Appendix A for a brief review of the basic notation, as well as relevant results. Here, we shall follow the compact form notation employed in [56] for ARK schemes.

The coefficients of the IMEX scheme define the matrices

$$\mathbb{A} = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix} \qquad \tilde{\mathbb{A}} = \begin{pmatrix} \tilde{\mathcal{A}} & 0 \\ \tilde{b}^t & 0 \end{pmatrix},$$

where $(\mathcal{A}, b) = ((a_{ij}), (b_i))$, represents the explicit scheme, and $(\tilde{\mathcal{A}}, \tilde{b}) = ((\tilde{a}_{ij}), \tilde{b}_i)$, contains the information related to the implicit scheme.

The IMEX scheme (3.56)-(3.57) can be written in compact form (see Appendix A) as follows,

$$\mathcal{U} = e \otimes U^n + \triangle t(\mathbb{A} \otimes I)\mathcal{L}(\mathcal{U}) + \triangle t(\tilde{\mathbb{A}} \otimes I)\mathcal{S}(\mathcal{U}), \tag{3.58}$$

where we have denoted $e = (1, \ldots, 1)^T \in \mathbb{R}^{s+1}$, $\mathcal{U} = ({U^{(1)}}^T, \ldots, {U^{(s)}}^T, (U^{n+1})^T)^T \in \mathbb{R}^{(s+1)N}$, and $\mathcal{L}(\mathcal{U}) = (L(U^{(1)})^T, \ldots, L(U^{(s)})^T, 0^T)^T \in \mathbb{R}^{(s+1)N}$, with analogous notation for $\mathcal{S}(\mathcal{U})$. The symbol $\otimes$ denotes the Kronecker product ([23] and Appendix A).

In the following section we shall investigate the properties of the numerical solution obtained with IMEX numerical schemes applied to the model problem, in particular those related to the weak stability of the numerical solution obtained for the model IVP, just as it was investigated in the previous sections for the first order schemes.

3.4.1 Weak Stability Preservation

The main issue to be considered in this section is that of ensuring discrete numerical profiles without numerical oscillations.

The use of RK schemes with special properties for the time discretization of MOL systems coming from (homogeneous) hyperbolic conservation laws was already proposed by Shu and Osher in the late eighties. In [87], they proposed a class of explicit RK schemes that do not increase the total variation of the numerical solution with time, under an appropriate CFL restriction, provided the first order explicit Euler discretization also does so. In the context of homogeneous hyperbolic conservation laws, it is rather natural to seek preservation of stability in the Total Variation semi-norm, since the true (entropy) solution in the scalar case also has this property.

The class of RK schemes developed in [87] were termed TVD-RK schemes as short notation for TVD-preserving schemes. In fact, the main idea in the original Shu-Osher derivation of [87] was to *assume* that the first order forward Euler discretization of the system of ODEs is TVD, when the time step is suitably restricted, and then to try to find a higher order time discretization that maintains this property, maybe under a different time-step restriction. Since the proofs relied on the ability to write a RK scheme as a convex combination of Forward Euler steps, it was soon realized that the same preservation results held for any norm or semi-norm. Later on [44], these were referred to as *Strong Stability Preserving* (SSP henceforth) schemes, since Strong Stability in the ODE literature was the term used for monotonicity in a given norm, or semi-norm. We refer the reader to Appendix A (and references therein) for more details about monotonicity properties of RK schemes and the relation with SSP schemes.

In recent years, much effort has been made in order to obtain optimal Strong Stability Preserving (SSP) RK schemes ([85], [92]). IMEX schemes where the explicit part is an SSP-RK scheme have been used with success in relaxation problems by Pareschi and Russo in [79]. In this case, the main issue is to obtain a numerical scheme that can deal with stiff source terms while at the same time enforcing an SSP method for the limiting conservation law.

For the model problem, we cannot expect, in general, to derive TVD schemes, since the solution will not have this property (see theorem 2.1).

In addition, it is not possible to ensure an inequality of the type

$$|v + \tau s(v)| \leq |v|, \qquad \forall v \in [0,1], \tag{3.59}$$

for any $\tau > 0$, so that it is not straightforward to apply known monotonicity results developed for Additive Runge-Kutta (ARK) methods in [56].

Nevertheless, we will be able to ensure certain bounds in the numerical solution that, as in the first order case analyzed in previous sections, lead us to expect a monotone profile for the numerical solution.

In theorem 3.5, we shall see that, under certain stepsize restrictions, the numerical solution time $t = t_n$ satisfies $0 \leq U^n \leq e$ provided that the numerical solution at time t_0 also satisfies this property. This result is based on the fact that the basic *weak stability* property, $0 \leq U^0 \leq 1 \quad \Rightarrow \quad 0 \leq U^n \leq 1$, holds for first order discretizations of both the convective derivative discretization and the source term, and can, thus, be interpreted as a *Weak Stability Preservation* property.

The proof of the corresponding bounds is closely related to the proof of theorems A.4 and A.5 in Appendix A, and relies on an appropriate partitioning of the coefficient matrix that allows us to rewrite the given scheme using only linear combinations, with positive coefficients, of terms for which the WS property is known to hold, under the appropriate stepsize restrictions.

We state and prove first the following proposition, valid for general splittings $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ and real triplets (r_1, r_2, r_3).

Proposition 3.1. *Let us consider any splitting of the matrix $\tilde{\mathbb{A}}$ in an IMEX-RK scheme as $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$. Let $r_1, r_2, r_3 \in \mathbb{R}$, be nonzero numbers such that the matrix*

$$\mathbb{B} := r_1 \mathbb{A} + r_2 \tilde{\mathbb{A}}_+ + r_3 \tilde{\mathbb{A}}_- \tag{3.60}$$

satisfies that $(I + \mathbb{B})$ is invertible. Then the scheme (3.58) can be rewritten as

$$\begin{aligned} \mathcal{U} &= (I+\mathbb{B})^{-1} e \otimes U^n + r_1 \left((I+\mathbb{B})^{-1} \mathbb{A} \otimes I \right) \left(\mathcal{U} + \frac{\triangle t}{r_1} \mathcal{L}(\mathcal{U}) \right) \\ &\quad + r_2 \left((I+\mathbb{B})^{-1} \tilde{\mathbb{A}}_+ \otimes I \right) \left(\mathcal{U} + \frac{\triangle t}{r_2} \mathcal{S}(\mathcal{U}) \right) \\ &\quad + r_3 \left((I+\tilde{\mathbb{B}})^{-1} \mathbb{A}_- \otimes I \right) \left(\mathcal{U} - \frac{\triangle t}{r_3} \mathcal{S}(\mathcal{U}) \right). \end{aligned} \tag{3.61}$$

Proof. As $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, we can write (3.58)

$$\begin{aligned} \mathcal{U} &= e \otimes U^n + \triangle t(\mathbb{A} \otimes I)\mathcal{L}(\mathcal{U}) \\ &\quad + \triangle t \left(\tilde{\mathbb{A}}_+ \otimes I\right) \mathcal{S}(\mathcal{U}) - \triangle t \left(\tilde{\mathbb{A}}_- \otimes I\right) \mathcal{S}(\mathcal{U}), \end{aligned} \tag{3.62}$$

so that if we add $(\mathbb{B}\otimes I)\mathcal{U} = (r_1\mathbb{A} + r_2\tilde{\mathbb{A}}_+ + r_3\tilde{\mathbb{A}}_- \otimes I)\mathcal{U}$ to both sides of (3.62), we obtain

$$\begin{aligned} \mathcal{U} + (\mathbb{B} \otimes I)\,\mathcal{U} &= e \otimes U^n + r_1(\mathbb{A} \otimes I) \left(\mathcal{U} + \frac{\triangle t}{r_1}\mathcal{L}(\mathcal{U})\right) \\ &\quad + r_2 \left(\tilde{\mathbb{A}}_+ \otimes I\right) \left(\mathcal{U} + \frac{\triangle t}{r_2}\mathcal{S}(\mathcal{U})\right) \\ &\quad + r_3 \left(\tilde{\mathbb{A}}_- \otimes I\right) \left(\mathcal{U} - \frac{\triangle t}{r_3}\mathcal{S}(\mathcal{U})\right) \end{aligned}$$

Using the properties of the Kronecker product, and the hypothesis on $\mathbb{B}$ we readily obtain

$$\begin{aligned} \mathcal{U} &= (I + \mathbb{B})^{-1}\, e \otimes U^n + r_1 \left((I + \mathbb{B})^{-1}\, \mathbb{A} \otimes I\right) \left(\mathcal{U} + \frac{\triangle t}{r_1}\mathcal{L}(\mathcal{U})\right) \\ &\quad + r_2 \left((I + \mathbb{B})^{-1}\, \tilde{\mathbb{A}}_+ \otimes I\right) \left(\mathcal{U} + \frac{\triangle t}{r_2}\mathcal{S}(\mathcal{U})\right) \\ &\quad + r_3 \left((I + \mathbb{B})^{-1}\, \tilde{\mathbb{A}}_- \otimes I\right) \left(\mathcal{U} - \frac{\triangle t}{r_3}\mathcal{S}(\mathcal{U}\right), \end{aligned}$$

which proves the result. ■

Our main result requires the splitting above to satisfy certain properties, in addition to the invertibility of the matrix $(I + \mathbb{B})$. In particular we shall require that

- $(I + \mathbb{B})^{-1}\mathbb{A} \geq 0$ and a strictly lower triangular matrix
- $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_+ \geq 0$ and a lower triangular matrix
- $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_- \geq 0$ and a strictly lower triangular matrix

In what follows we shall denote

$$\begin{aligned} (I + \mathbb{B})^{-1} e &= (\alpha_i), & (I + \mathbb{B})^{-1}\mathbb{A} &= (\beta_{i,j}), \\ (I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_+ &= (\tilde{\beta}^+_{ij}), & (I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_- &= (\tilde{\beta}^-_{ij}). \end{aligned} \tag{3.63}$$

In an IMEX scheme, $\mathbb{A}$ is the matrix associated to the explicit scheme, hence $\mathbb{A}$ is always a strictly lower triangular matrix. We will always assume that, in addition, $\mathbb{A} \geq 0$. The matrix $\tilde{\mathbb{A}}$ is only a lower triangular matrix, but we shall assume in what follows that $\tilde{a}_{ii} \geq 0$. We shall seek splittings $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ such that $\tilde{\mathbb{A}}_\pm \geq 0$ and $\tilde{\mathbb{A}}_-$ is a strictly lower triangular matrix. In this case, the matrix $(I + \mathbb{B})$ is lower triangular, and $(I + \mathbb{B})_{ii} = 1 + r_2\tilde{a}_{ii} > 1$ when $r_2 > 0$, hence $(I + \mathbb{B})$ is invertible for $r_2 > 0$.

In addition, we also have

- $(I + \mathbb{B})^{-1}\mathbb{A}$ is a strictly lower triangular matrix.
- $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_+$ is a lower triangular matrix with diagonal elements

$$\tilde{\beta}^+_{ii} = \frac{\tilde{a}_{ii}}{(1 + r_2\tilde{a}_{ii})}. \tag{3.64}$$

- $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_-$ is a strictly lower triangular matrix.

Notice that

$$\begin{aligned} &e = (I + \mathbb{B})^{-1}(I + \mathbb{B})e = \\ &(I + \mathbb{B})^{-1}e + r_1(I + \mathbb{B})^{-1}\mathbb{A}e + r_2(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_+ e + r_3(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_- e. \end{aligned} \tag{3.65}$$

Hence, the $i-th$ component above leads to the following relation between the coefficients in (3.63):

$$\alpha_i + r_1 \sum_{j=1}^{i-1} \beta_{ij} + r_2 \sum_{j=1}^{i} \tilde{\beta}^+_{ij} + r_3 \sum_{j=1}^{i-1} \tilde{\beta}^-_{ij} = 1. \tag{3.66}$$

A splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ which is appropriate for our purposes can be constructed by using a technique developed in [56], which allows to obtain a Shu-Osher representation of a RK scheme given by a coefficient matrix $\mathbb{A}$, provided $R(\mathbb{A}) > 0$, where $R(\mathbb{A})$ is the radius of absolute stability of $\mathbb{A}$ (see Appendix A and e.g. [56]). Applying theorem A.3 to $\mathbb{A}$, we can find $\Lambda \geq 0, \Gamma \geq 0$ such that $\Lambda - r\Gamma \geq 0$ and $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$. Multiplying (3.58) by $(I - \Lambda)$, we get

$$\begin{aligned} \mathcal{U} &= \alpha \otimes U^n + (\Lambda \otimes I)\mathcal{U} \\ &\quad + \triangle t(\Gamma \otimes I)\mathcal{L}(\mathcal{U}) + \triangle t\left((I - \Lambda)\,\tilde{\mathbb{A}} \otimes I\right)\mathcal{S}(\mathcal{U}). \end{aligned} \tag{3.67}$$

Let us define $\tilde{\Gamma} := (I - \Lambda)\,\tilde{\mathbb{A}}$. Notice that $\tilde{\Gamma}$ has the same nonnegative diagonal entries as $\tilde{\mathbb{A}}$. However, since there might be negative off-diagonal entries, we decompose $\tilde{\Gamma}$ as

$$\tilde{\Gamma} = \tilde{\Gamma}_+ - \tilde{\Gamma}_-, \text{ with } \tilde{\Gamma}_+, \tilde{\Gamma}_- \geq 0.$$

Hence, we can rewrite $\tilde{\mathbb{A}}$ as

$$\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-,$$

where the matrices $\tilde{\mathbb{A}}_+ = (I-\Lambda)^{-1}\tilde{\Gamma}_+$, $\tilde{\mathbb{A}}_- = (I-\Lambda)^{-1}\tilde{\Gamma}_-$ satisfy the desired properties. In particular, since $(I-\Lambda)^{-1} \geq 0$ (see [55] lemma 3.8), we also have $\tilde{\mathbb{A}}_+ \geq 0$, $\tilde{\mathbb{A}}_- \geq 0$.

Let us assume that there exist $r_1, r_2, r_3 \geq 0$ such that

$$(I+\mathbb{B})^{-1}e \geq 0, \quad (I+\mathbb{B})^{-1}\mathbb{A} \geq 0, \quad (I+\mathbb{B})^{-1}\tilde{\mathbb{A}}_\pm \geq 0, \tag{3.68}$$

Notice that, in this case, we readily deduce that $0 \leq \alpha_i \leq 1$ when $r_i \geq 0$, $i = 1,2,3$ from (3.66). This property, along with the lemmas 3.1 and 3.2 are key ingredients in the proof of the following theorem,

Theorem 3.5. *Let us consider the triplet* $(r_1, r_2, r_3) = (r_1, r_1\frac{\mu\triangle x}{2}, r_1\frac{\mu\triangle x}{16})$. *Let us assume that we have constructed a partition* $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ *so that* $\tilde{\mathbb{A}}_+ \geq 0$, $\tilde{\mathbb{A}}_- \geq 0$ *with* $\tilde{\mathbb{A}}_-$ *strictly lower triangular.*
Let us also assume that there exists $r_1 > 0$ *(which can depend on* $\mu\Delta x$*) so that the inequalities in (3.68) hold. Then*

$$0 \leq U^n \leq e \qquad \Longrightarrow \qquad 0 \leq U^{n+1} \leq e$$

under the restriction

$$\frac{\triangle t}{\triangle x} \leq r_1(\mu\triangle x)$$

provided U^{n+1} *can be computed.*

Proof. Using proposition 3.1, we rewrite (3.58) as (3.61). According to the observations made before, due to the prosperities of $\mathbb{A}$, $\tilde{\mathbb{A}}$ and $\tilde{\mathbb{A}}_\pm$, each component of (3.61) has the form

$$\begin{aligned} U^{(i)} = \alpha_i U^n &+ r_1 \sum_{j=1}^{i-1} \beta_{ij} \left(U^{(j)} + \frac{\triangle t}{r_1} L(U^{(j)}) \right) \\ &+ r_2 \sum_{j=1}^{i-1} \tilde{\beta}^+_{ij} \left(U^{(j)} + \frac{\triangle t}{r_2} S(U^{(j)}) \right) \\ &+ r_2 \tilde{\beta}^+_{ii} U^{(i)} + \triangle t \tilde{\beta}^+_{ii} S(U^{(i)}) \\ &+ r_3 \sum_{j=1}^{i-1} \tilde{\beta}^-_{ij} \left(U^{(j)} - \frac{\triangle t}{r_3} S(U^{(j)}) \right). \end{aligned} \tag{3.69}$$

which we can rewrite as follows

$$(1 - r_2\tilde{\beta}^+_{ii})U^{(i)} = \widehat{U}^{(i)} + \Delta t\tilde{\beta}^+_{ii}S(U^{(i)}) \tag{3.70}$$

where

$$\begin{aligned}\widehat{U}^{(i)} = \alpha_i U^n \quad + \quad & r_1\sum_{j=1}^{i-1}\beta_{ij}\left(U^{(j)} + \frac{\triangle t}{r_1}L(U^{(j)})\right) \\ & + r_2\sum_{j=1}^{i-1}\tilde{\beta}^+_{ij}\left(U^{(j)} + \frac{\triangle t}{r_2}S(U^{(j)})\right) \\ & + r_3\sum_{j=1}^{i-1}\tilde{\beta}^-_{ij}\left(U^{(j)} - \frac{\triangle t}{r_3}S(U^{(j)})\right).\end{aligned} \tag{3.71}$$

Recall that, under the hypothesis of the theorem, we also have (3.64), i.e.

$$\tilde{\beta}^+_{ii} = \tilde{a}_{ii}/(1 + r_2\,\tilde{a}_{ii}). \tag{3.72}$$

This implies the following two relations

$$1 - r_2\tilde{\beta}^+_{ii} = (1 + r_2\tilde{a}_{ii})^{-1} > 0, \qquad \frac{\tilde{\beta}^+_{ii}}{1 - r_2\tilde{\beta}^+_{ii}} = \tilde{a}_{ii}. \tag{3.73}$$

Thus, if we define

$$\bar{U}^{(i)} = \frac{1}{1 - r_2\tilde{\beta}^+_{ii}}\,\widehat{U}^{(i)} \tag{3.74}$$

then, for each $i = 1, \ldots, s$ we can write (3.69) as

$$U^{(i)} = \bar{U}^{(i)} + \triangle t\tilde{a}_{ii}S(U^{(i)}). \tag{3.75}$$

We shall prove the result by an induction process over the internal stages.

For the first-stage we have

$$U^{(1)} = \bar{U}^{(1)} + \triangle t\tilde{a}_{11}S(U^{(1)}),$$

with

$$\bar{U}^{(1)} = \frac{\alpha_1}{1 - r_2\tilde{\beta}^+_{ii}}U^n = U^n$$

thanks to (3.66). Then, since $\tilde{a}_{ii} \geq 0$, lemma (3.4) readily implies that $0 \leq U^{(1)} \leq e$.

Suppose now that $0 \leq U^{(j)} \leq e$, $j = 1, \ldots, i-1$. Then using lemmas 3.1 and 3.2 we have

$$\begin{aligned} 0 \leq U^{(j)} + \frac{\triangle t}{r_1} L(U^{(j)}) \leq e, \qquad & \frac{\triangle t}{r_1} \leq \tau_L = \Delta x \\ 0 \leq U^{(j)} + \frac{\triangle t}{r_2} s(U^{(j)}) \leq e, \qquad & \frac{\triangle t}{r_2} \leq \tau_{\mu_+} = \frac{2}{\mu} \\ 0 \leq U^{(j)} - \frac{\triangle t}{r_3} s(U^{(j)}) \leq e, \qquad & \frac{\triangle t}{r_3} \leq \tau_{\mu_-} = \frac{16}{\mu}. \end{aligned}$$

Then, under these restrictions for Δt, (3.71) and (3.65) imply that

$$\begin{aligned} 0 \leq \widehat{U}^{(i)} &\leq (\alpha_i + r_1 \sum_{j=1}^{i-1} \beta_{ij} + r_2 \sum_{j=1}^{i-1} \tilde{\beta}_{ij}^{+} + r_3 \sum_{j=1}^{i-1} \tilde{\beta}_{ij}^{-})\, e \\ &= (1 - r_2\, \tilde{\beta}_{ii}^{+})\, e \end{aligned}$$

Therefore, we deduce from (3.74) that

$$0 \leq \bar{U}^{(i)} \leq 1$$

and from lemma 3.4 and (3.75) we get

$$0 \leq U^{(i)} \leq 1$$

Taking into account that $\tau_L = \triangle x$, $\tilde{\tau}_{\mu_+} = 2/\mu$, and $\tilde{\tau}_{\mu_-} = 16/\mu$, by defining

$$r_2 = r_1 \frac{\mu \triangle x}{2}, \qquad r_3 = r_1 \frac{\mu \triangle x}{16},$$

the stepsize restrictions

$$\frac{\triangle t}{r_1} \leq \tau_L, \quad \frac{\triangle t}{r_2} \leq \tau_{\mu_+}, \quad \frac{\triangle t}{r_3} \leq \tau_{\mu_-}$$

can be equivalently written as

$$\frac{\triangle t}{\triangle x} \leq r_1,$$

which concludes the proof. ■

The theorem above provides a CFL-like condition for *weak stability preservation*. We remark again that the result does not prevent the occurrence of oscillations, 'per se'. However, all the numerical evidence

points out that when the oscillations do occur, the values on the numerical wave profile do not lie in $[0,1]$. Hence for all practical purposes this result 'guarantees' that the IMEX scheme, when applied to the model problem, produces a monotone discrete profile.

For practical purposes, we need to find r_1 such that (3.68) is satisfied, where $(r_1, r_2, r_3) = (r_1, r_1\frac{\mu\triangle x}{2}, r_1\frac{\mu\triangle x}{16})$. In what follows, we shall work out an example based on the IMEX SSP2(3,3,2) scheme, proposed in [79], just to see how this number can be computed.

In [79], Pareschi and Russo study IMEX schemes for hyperbolic systems of conservation laws with stiff relaxation terms. The explicit part is treated by a SSP scheme and in the notation SSP$k(s,\sigma,p)$ s is the number of stages of the implicit scheme, σ the number of stages in the explicit scheme and k is the order of the explicit scheme and p is the order of the IMEX scheme.

The IMEX SSP2(3,3,2) scheme studied in [79] is given by the following double tableau

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \mathbb{A} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline \tilde{\mathbb{A}} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \tag{3.76}$$

We shall find an appropriate splitting, as required by theorem 3.5. For this, we proceed as explained after proposition 3.1. It is known (see e.g. [56] that $R(\mathbb{A}) = 2$. We then apply theorem A.3 and find the matrices Λ and Γ that convert the explicit part $\mathbb{A}$ into a Shu-Osher form (see also the example after theorem A.3),

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & 0 \end{pmatrix} \qquad \Gamma = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \end{pmatrix} \qquad \alpha = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix}. \tag{3.77}$$

We recall that, by multiplying (3.58) by $(I - \Lambda)$, we get

$$\begin{aligned} \mathcal{U} &= \alpha \otimes U^n + (\Lambda \otimes I)\mathcal{U} \\ &\quad + \triangle t(\Gamma \otimes I)\mathcal{L}(\mathcal{U}) + \triangle t\left((I - \Lambda)\,\tilde{\mathbb{A}} \otimes I\right)\mathcal{S}(\mathcal{U}). \end{aligned}$$

We define $\tilde{\Gamma} = (I - \Lambda)\tilde{\mathbb{A}}$, which becomes in this case,

$$\tilde{\Gamma} = \begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{3} & \frac{1}{12} & \frac{1}{3} & 0 \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 \end{pmatrix}.$$

Notice that the element (2,1) is negative. We then consider the splitting

$$\tilde{\Gamma} = \tilde{\Gamma}_{+} - \tilde{\Gamma}_{-},$$

so that $\tilde{\Gamma}_{+}, \tilde{\Gamma}_{-} \geq 0$. This splitting allows us to write also

$$\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_{+} - \tilde{\mathbb{A}}_{-}$$

with $\tilde{\mathbb{A}}_{+} = (I - \Lambda)^{-1}\tilde{\Gamma}_{+}$, $\tilde{\mathbb{A}}_{-} = (I - \Lambda)^{-1}\tilde{\Gamma}_{+}$.

Then, these matrices become

$$\tilde{\mathbb{A}}_{+} = \begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{7}{12} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}}_{-} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{6} & 0 & 0 & 0 \end{pmatrix}.$$

A formal computer language, such as Mathematica, can be used to compute $r_1(\mu\Delta x)$ as the maximum real number so that (3.68) is satisfied, when $(r_1, r_2, r_3) = (r_1, r_1\frac{\mu\triangle x}{2}, r_1\frac{\mu\triangle x}{16})$. In doing so for the above example, one finds that ($y := \mu\triangle x$)

$$r_1(y) = \begin{cases} -\frac{4(43y+192)}{31y^2-44y-384} - 4\sqrt{\frac{4825y^2+12288y}{(31y^2-44y-384)^2}} & \text{if } 0 \leq y < \frac{2}{31}\left(11+\sqrt{3097}\right), \\ \frac{192}{43y+192} & \text{if } y = \frac{2}{31}\left(11+\sqrt{3097}\right), \\ 4\sqrt{\frac{4825y^2+12288y}{(31y^2-44y-384)^2}} - \frac{4(43y+192)}{31y^2-44y-384} & \text{if } y > \frac{2}{31}\left(11+\sqrt{3097}\right). \end{cases}$$

The plot of $r_1(\mu\Delta x)$ is shown in the left-most display of figure 3.11. Notice that $r_1(1) = 1.04971$.

In [56], Higueras obtains that the *curve of absolute monotonicity*, $\partial R(\mathbb{A}, \tilde{\mathbb{A}})$ (see Appendix A, section A.4) of the IMEX scheme SSP2(3,3,2) above is reduced to the point $(0,0)$. She then considers a modification of the SSP(3,3,2) that has the same coefficient scheme for the explicit part, i.e.

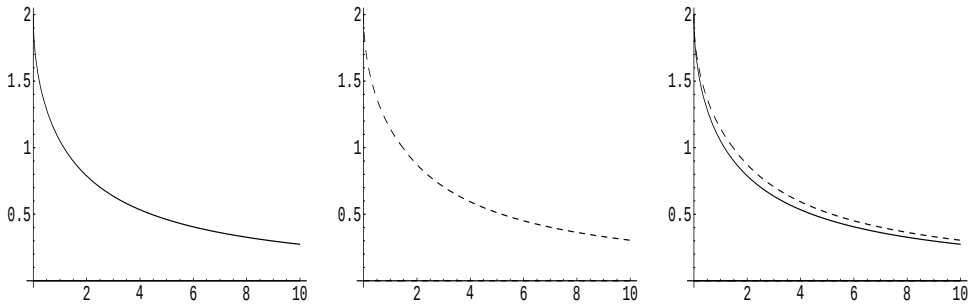


Figure 3.11: $r_1(\mu\triangle x)$. *Left: SSP2(3,3,2) from [79]. Center: IMEXH in [56]. Right: superposition of both*

same $\mathbb{A}$, but where the implicit part is represented by the matrix $\tilde{\mathbb{A}}$ below. The coefficient matrices for the scheme in [56], which shall be referred as IMEXH

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \mathbb{A} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{5} & \frac{1}{5} & 0 & 0 \\ \frac{3}{10} & \frac{1}{10} & \frac{1}{5} & 0 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline \tilde{\mathbb{A}} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} . \tag{3.78}$$

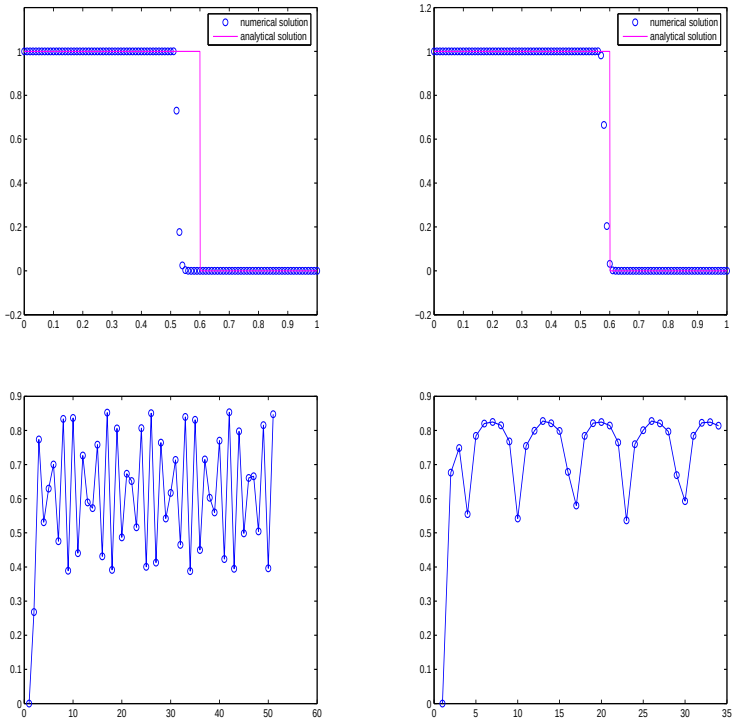
It is shown in [56] that this scheme has better monotonicity properties than the SSP2(3,3,2) scheme considered above. In particular, the curve of absolute monotonicity $\partial R(\mathbb{A}, \tilde{\mathbb{A}}) \neq (0, 0)$.

The same computations as above can be carried out for the new scheme. The function $r_1(\mu\Delta x)$ obtained in this case is shown in the middle display in figure 3.11. In this case $r_1(1) = 1.14274$. A slight improvement is observed with respect to the IMEX SSP2(3,3,2) scheme. The two graphs are superimposed and displayed in the rightmost plot of 3.11, which allows us to conclude that only a slight improvement on the range of $y = \mu\Delta x$ shown in the picture is in fact obtained.

We remark that the conditions found in theorem 3.5 are sufficient conditions, and that these are going to depend on the decomposition used in $\mathbb{A}$, $\tilde{\mathbb{A}}_+$ and $\tilde{\mathbb{A}}_-$. Therefore, it might be possible to obtain larger bounds.

3.4.2

A numerical analysis of the discrete Wave speed

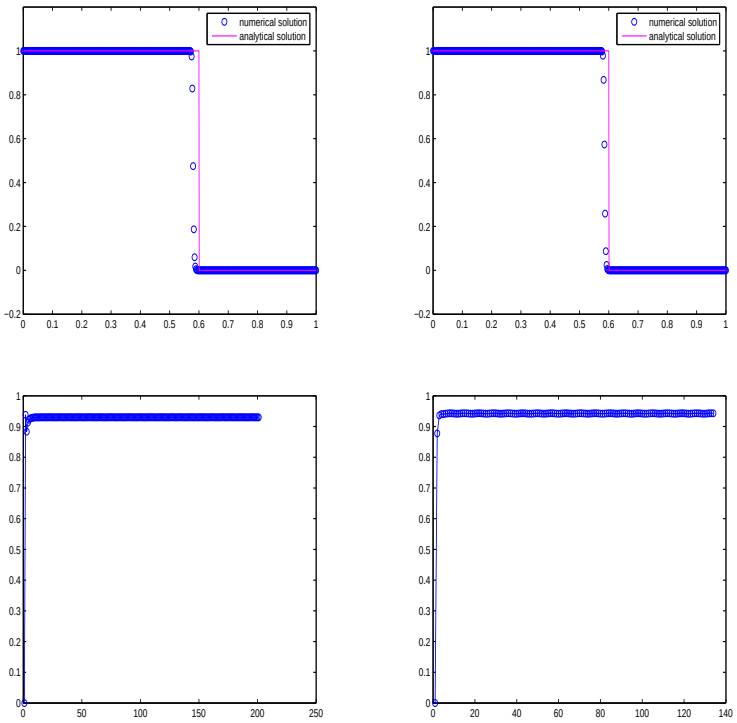


(a) $\mu\triangle x = 10$, CFL$= 0.6$ (b) $\mu\triangle x = 10$, CFL$= 0.9$

Figure 3.12: *IMEXH scheme. Top: Numerical wave profile. Bottom:* $1 - \alpha(\mu\triangle x, U^n)$ *as a function of* n*.*

In this section we shall show some sample plots of the behavior obtained when using the IMEXH scheme, which is very similar to that obtained with the SSP(3,3,2) used by Pareschi and Russo in [79].

In figure 3.12 we show the numerical solution and the delay factor for $\mu\Delta x = 10$. We can see in figure 3.11 that $r_1(10) < 0.5$. However, the good stability properties of the implicit schemes seem to prevail and we observe that nonoscillatory numerical profiles are obtained, even though the conditions on the theorem are not satisfied. In figure 3.13 we display the results when $\mu\Delta x = 10/4$. In this case, the delay is, as expected, much smaller for both CFL numbers.

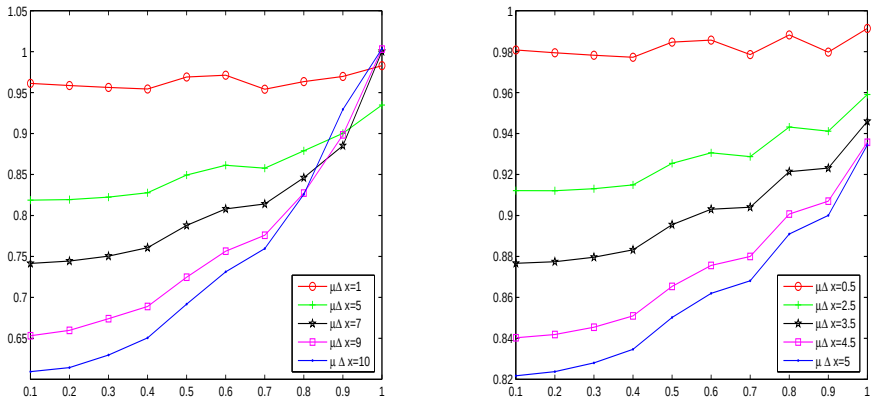


(a) $\mu\triangle x = 10/4$, CFL$= 0.6$ (b) $\mu\triangle x = 10/4$, CFL$= 0.9$

Figure 3.13: *IMEXH scheme. Top: Numerical wave profile. Bottom:* $1 - \alpha(\mu\triangle x, U^n)$ *as a function of* n.

In general, for $\mu\Delta x = 10$ the delay is clearly visible. In figure 3.14 we have also plotted the average speed (3.55) with respect to the CFL number, obtained from various simulations with the IMEXH scheme. We observe that the results obtained confirm that, as in the first order case, the delay is, on average, of order $\mathcal{O}(\mu\Delta x 10^{-2})$.

The plots in figure 3.14 show that the qualitative behavior of the average speed with respect to the CFL number is the same for the IMEX scheme and the first order implicit scheme.



(a) $\Delta x = 0.01$ (b) $\Delta x = 0.005$

Figure 3.14: *IMEXH scheme. Average speed as a function of CFL for different values of* $\mu = 100, 500, 700, 900, 1000$.

3.4.3 Conclusion

A MOL discretization of a hyperbolic conservation law with stiff source terms will lead to waves propagating at nonphysical speeds. This is an artifact produced by the numerical technique, due to the fact that nonequilibrium points are introduced in a discontinuous profile by the numerical discretization of the convective derivative. We have analyzed this phenomenon for MOL schemes obtained from first order upwind spatial discretization of the convective derivative in a model problem proposed by LeVeque and Yee in [73]. The use of the simple upwind discretization allows us to carry out a detailed study of the *delay factor*. It is seen in section 3.2.2 and 3.3.2, that the delay or advance of the discrete wave profile in a MOL discretization is controlled mainly by $\mu\Delta x$. For simple, first order, time discretizations, a discrete delay of the form

$$\alpha(\mu\triangle x, U^n) = \mu\triangle x \sum_{i=1}^{N} U_i^n (U_i^n - 1)(U_i^n - \frac{1}{2}), \tag{3.79}$$

is obtained. Notice that, for the IVP studied in this chapter, the value of

$$U_i^n (U_i^n - 1)(U_i^n - \frac{1}{2}),$$

is almost always zero except for the points that conform the discontinuity profile. Because of the form of the source term for this problem, negative and positive values are possible, so the sum,

$$\sum_{i=1}^{N} U_i^n (U_i^n - 1)(U_i^n - \frac{1}{2}),$$

which is usually small, can have either sign, and the discrete profile can be advanced, or retarded, with respect to the true profile.

We have also studied stepsize restrictions that would lead to expect a non-oscillatory discrete profile, arriving at the concept of *weak stability*.

The results shown in this chapter indicate that, under appropriate conditions which can be made precise, IMEX schemes can be used to obtain monotone wave profiles for the model problem. These profiles move at a speed that is also directly related to the parameter $\mu\Delta x$, so that sufficient refinement on the underlying mesh is necessary in order to ensure that the discontinuity is correctly located.

Even though we only showed numerical results for the first order upwind discretization of the convective term, the behavior obtained in this case is typical of what would be obtained with more sophisticated, state of the art conservative discretizations for the solution of the model IVP. The advantage of the first order upwind discretization is that it allows for a rather systematic analysis of the numerical solution, which allowed us to establish certain bounds that, for all practical purposes, ensure monotonicity of the numerical profiles.

4

Flux-limited second order schemes for balance laws

High resolution schemes, such as those described in chapter 1, have proved to be particularly successful at capturing shock waves and discontinuous solutions for homogeneous scalar conservation laws. However, and as explained in chapter 2, when applied to balance laws such as shallow water equations, additional care must be placed on the numerical technique.

For balance laws, it is often necessary to be able to compute accurately steady state, or nearly steady state, solutions for which the flux

gradients are nonzero but are exactly balanced by the source terms. It has been observed in chapter 2 that many numerical methods do not respect this balance and, hence, cannot be used to compute accurately small perturbations of steady state solutions.

Numerical schemes that respect the balance that occurs in steady flow are called *well balanced*, after the work of Leroux and collaborators [46], [47]. Independently, Bermudez and Váquez-Cendón [5] stated a property of the numerical scheme that guarantees the absence of parasitic waves in the form or spurious oscillations near equilibrium states. The so called *C-property* [5] ensures also that the scheme is well balanced. The lack of a proper balance, at the discrete level, between the effects of the convective fluxes and the source terms leads almost invariably to spurious oscillatory behavior.

When trying to design high order schemes for inhomogeneous conservation laws, well balancing is one important issue that must be taken into account. Oscillations do occur also as a consequence of the plain use of data-independent second order schemes, which is a well known deficiency of data-independent second order schemes in the homogeneous case. Figure 4.1 shows a numerical simulation for the very simple model case

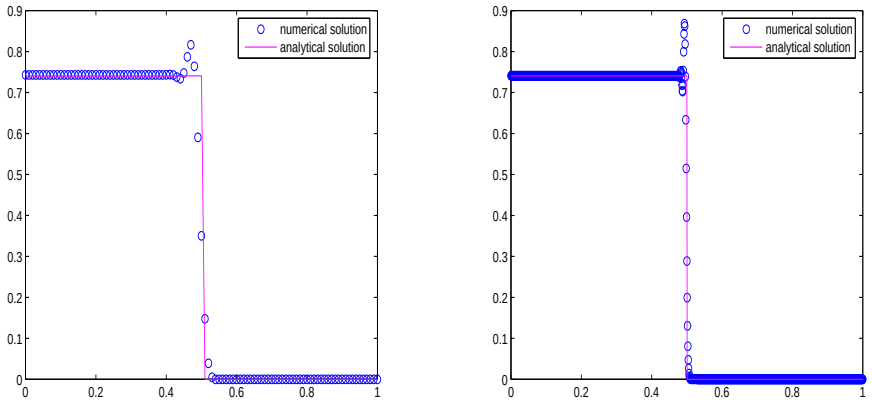
$$u_t + u_x = -u \qquad u(x,0) = \begin{cases} 1, & x \leq 0.2 \\ 0, & x > 0.2, \end{cases}$$

obtained with the following extension of MacCormack's method, modified to include source terms in an explicit manner while maintaining second-order accuracy [73], [107]:

$$\begin{aligned} \Delta U_i^{(1)} &= -\frac{\triangle t}{\triangle x}\left(f(U_i^n) - f(U_{i-1}^n)\right) + \triangle t s(U_i^n) \\ U_i^{(1)} &= U_i^n + \Delta U_i^{(1)} \\ \Delta U_i^{(2)} &= -\frac{\triangle t}{\triangle x}\left(f(U_i^{(1)}) - f(U_{i-1}^{(1)})\right) + \triangle t s(U_i^{(1)}) \\ U_i^{n+1} &= U_i^n + \frac{1}{2}\left(\Delta U_i^{(1)} + \Delta U_i^{(2)}\right). \end{aligned}$$

In figure 4.1, we observe spurious overshoots typical of data-independent second order schemes, which do not diminish with mesh refinement. Notice that they are of the same type as those observed for the analogous computation in the homogeneous case.

A similar observation is made in [38], for a more involved two-step, three-point stencil, second order scheme which extends the Lax-Wendroff scheme to non homogeneous conservation laws. When used on Embid's



$\triangle x = 0.01$ $\triangle x = 0.001$

Figure 4.1: *Second order scheme applied to* $u_t + u_x = -u$ *with CFL*$= 0.9$.

problem, which is a scalar model for gas flow through a duct with variable cross-section, the typical oscillations around discontinuous profiles, characteristic of the Lax-Wendroff scheme, are obtained (see [38]).

In [38], Gascón and Corberán seek to define a non-oscillatory second order scheme for non homogeneous conservation laws by studying what conditions should be imposed in order to obtain a TVD scheme for non-homogeneous hyperbolic conservation laws. In [38], the authors present a technique based on the transformation of the nonhomogeneous problem to a *homogeneous form* through the definition of a new flux formed by the physical flux and the primitive of the source term. This technique simplifies the requirements for well-balancing and suggests a way to apply the existing technology on TVD schemes to the nonhomogeneous case.

This is the starting point of the work carried out in this chapter. While the results presented by Gascón and Corberán in [38], which deal mainly with numerical simulations of duct flow with variable cross-sectional area, indicate good numerical properties for the proposed second order scheme, we believe that there are some aspects in the basic design that prevent this scheme from becoming a truly general technique for numerical simulations involving balance laws.

We have observed in chapter 2 that the use of TVD schemes for scalar conservation laws is justified by the fact that the true (entropy) solution of a scalar conservation law also satisfies a TVD property. However, for scalar balance laws the TVD property of the solution might no longer

hold, see Theorem 2.1. Although certain source terms may preserve the TVD property of the homogeneous part, others will actively increase the variation in the solution. This observation makes us turn our attention to the scalar case, where we shall revise, and modify, the basic design principles of Gascón and Corberán's scheme.

In general, the TVD concepts and basic ideas are still useful, since TVD constraints lead to non-oscillatory solutions. TVD schemes succeed in avoiding numerically produced oscillations, and this has been the key concept in promoting their extension to systems of hyperbolic conservation laws, where the TVD property of the solution might no longer hold.

In this chapter we shall develop two flux-limited second-order schemes for inhomogeneous conservation laws. The derivation of these schemes follows the flux-limiter technology covered in chapter 1, and their basic design stems from what we believe to be essential weaknesses in Gascón and Corberán developments in [38]. The behavior of the proposed schemes is analyzed through an extensive set of numerical experiments, in particular with respect to their well balancing properties. The preferred scheme is then extended to the shallow water system, via a local characteristic approach.

4.1 Gascón and Corberán TVD scheme

We shall describe in this section the main steps leading to the second order TVD scheme proposed by Gascón an Corberán in [38]. The derivation below follows step by step that of [38], in particular section 2 in their paper.

Gascón and Corberán propose to *convert* a non-homogeneous conservation law of the form

$$u_t + f(u)_x = s(x,u), \tag{4.1}$$

to a *homogeneous form*

$$u_t + g_x = 0, \tag{4.2}$$

where the new flux function, $g = g(x,u)$ in their notation, is defined as

$$g(x,u) = f(u) - \int_0^x s(y,u(y,t))dy. \tag{4.3}$$

Notice that g satisfies $g_x = f_x - s(x,u)$, ensuring the equivalence between (4.1) and (4.2). The choice of $x = 0$ for the definition of the primitive of

the source term implicitly assumes that $x = 0$ is the starting point of the computational domain.

In [38], the authors seek to design first a second order scheme following a Lax-Wendroff type Taylor expansion for (4.2) of the form

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}(g_{i+\frac{1}{2}}^{n+\frac{1}{2}} - g_{i-\frac{1}{2}}^{n+\frac{1}{2}}) \tag{4.4}$$

with

$$g_{i+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}\left(g_i^n + g_{i+1}^n - \frac{\Delta t}{\Delta x}\frac{\partial g}{\partial u}\bigg|_{i+\frac{1}{2}}^{n}(g_{i+1}^n - g_i^n)\right). \tag{4.5}$$

In the formula above

$$g_i^n = f(U_i^n) - \int_0^{x_i} s(y, u(y, t_n))dy, \tag{4.6}$$

and

$$\frac{\partial g}{\partial u} = \frac{\partial f}{\partial u} - \frac{\partial}{\partial u}\left(\int_0^x s(y, u(y, t))dy\right). \tag{4.7}$$

Notice that the computation of g_i^n demands some sort of numerical integration, unless $s(x, u) = s(x)$ so that the primitive can be computed explicitly. Taking into account that the fluxes appear always in difference form, Gascón and Corberán introduce the following notation

$$b_{i,k}^n = -\int_{x_i}^{x_k} s(y, u(y, t_n))dy, \tag{4.8}$$

and arrive at a specific second order scheme (formulas (16) and (17) in [38]) which is written in terms of the following discrete values: f_k^n, $f_u|_{k+1/2}^n$, $s_u|_{k+1/2}^n$, $b_{k,l}^n$, $b_{k\pm1/2,l}^n$, $b_{k,l\pm1/2}^n$, where

$$x_{i+\frac{1}{2}} = x_i + \frac{\triangle x}{2}, \quad t_{n+\frac{1}{2}} = t_n + \frac{\triangle t}{2},$$

and

$$f_i^n = f(U_i^n) \qquad s_i^n = s(x_i, U_i^n).$$

All of these values can be computed from the discrete data, either directly or by a convenient approximation, producing a fully discrete second order scheme.

When this scheme is used in a numerical simulation involving Embid's problem (see section 4.5.2 in this chapter and figure 2 in [38]), one can observe the typical oscillatory behavior encountered in numerical

simulations involving the Lax Wendroff scheme for homogeneous conservation laws.

It is observed in [38] that these oscillations are not related to the well-balancing property. Indeed, the scheme is well balanced when

$$g_{i+1} - g_i = 0, \qquad \forall i, \tag{4.9}$$

which amounts to

$$f^n_{i+1} - f^n_i + b^n_{i,i+1} = 0,$$

which can easily be enforced for Embid's problem by an appropriate integration rule in the numerical computation of $b^n_{i,i+1}$.

The derivation of a TVD scheme for the inhomogeneous case in [38] is inspired by the work of Harten in [50], which we recall below (see [50] for further details).

Let us consider a three point finite difference scheme in conservation form for the homogeneous conservation law, with a numerical flux $\bar{f}$ of the form

$$\bar{f}(U_i, U_{i+1}) = \frac{1}{2}\left(f(U_i) + f(U_{i+1}) - \frac{\Delta x}{\Delta t} Q(\tilde{a}^n_{i+\frac{1}{2}})(U_{i+1} - U_i)\right) \tag{4.10}$$

where

$$\tilde{a}^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{\triangle t}{\triangle x}\dfrac{f^n_{i+1} - f^n_i}{U^n_{i+1} - U^n_i}, & \text{if } U^n_{i+1} - U^n_i \neq 0 \\ \dfrac{\triangle t}{\triangle x}\left.\dfrac{\partial f}{\partial u}\right|^n_i, & \text{if } U^n_{i+1} - U^n_i = 0, \end{cases} \tag{4.11}$$

and $Q(x)$ is some function, which is often referred to as the coefficient of numerical viscosity. Notice that $Q(x) = x$ gives the Lax-Wendroff scheme for the scalar conservation law.

Harten states the following lemma.

Lemma 4.1. *([50], Lemma 3.1) Let $Q(x)$ in (4.10) satisfy the inequalities*

$$|x| \leq Q(x) \leq 1, \qquad 0 \leq |x| \leq \mu \leq 1 \tag{4.12}$$

then, the finite difference scheme (4.10) with (4.11) is Total Variation Non Increasing (TVNI) under the CFL-like restriction

$$\max_i |\tilde{a}^n_{i+\frac{1}{2}}| \leq \mu. \tag{4.13}$$

Taking into account (4.5), (4.7) and the above results, Gascón and Corberán propose a finite difference scheme in conservation form

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x}(\tilde{G}^n_{i+\frac{1}{2}} - \tilde{G}^n_{i-\frac{1}{2}}), \tag{4.14}$$

with a numerical flux $\tilde{G}^n_{i+\frac{1}{2}}$ defined as follows

$$\tilde{G}^n_{i+\frac{1}{2}} := \frac{1}{2}(g_i^n + g_{i+1}^n - h(\tilde{a}^n_{i+\frac{1}{2}} + \tilde{b}^n_{i+\frac{1}{2}})(g_{i+1}^n - g_i^n)). \tag{4.15}$$

The function h is named as the coefficient of numerical viscosity in the homogeneous case, where $\tilde{a}^n_{i+\frac{1}{2}}$ is defined by

$$\tilde{a}^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{\triangle t}{\triangle x}\dfrac{f^n_{i+1} - f^n_i}{U^n_{i+1} - U^n_i}, & \text{if } U^n_{i+1} - U^n_i \neq 0 \\ \\ \dfrac{\triangle t}{\triangle x}\dfrac{\partial f}{\partial u}\Big|^n_i, & \text{if } U^n_{i+1} - U^n_i = 0, \end{cases} \tag{4.16}$$

and analogously, the authors denote

$$\tilde{b}^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{\triangle t}{\triangle x}\dfrac{b^n_{i+1} - b^n_i}{U^n_{i+1} - U^n_i}, & \text{if } U^n_{i+1} - U^n_i \neq 0 \\ \\ 0, & \text{if } U^n_{i+1} - U^n_i = 0. \end{cases} \tag{4.17}$$

Remark The definition of $\tilde{b}^n_{i+\frac{1}{2}}$ above is written in terms of the quantities b^n_k, defined in a natural way as

$$b^n_k = -\int_{x_0}^{x_k} s(y, u(y, t_n))dy$$

so that the expression $b^n_{i+1} - b^n_i = b^n_{i,i+1}$. However, we note that this interpretation of $b = b(u)$ is rather unnatural, since the behavior of b with respect to u is not easy to determine. In fact, the definition $\tilde{b}^n_{i+\frac{1}{2}} = 0$ when $U^n_{i+1} - U^n_i = 0$ given in [38] and reproduced above seems rather arbitrary, specially when compared to the analogous formula for $\tilde{a}^n_{i+\frac{1}{2}}$ in (4.20). Our derivation of a second order scheme will bypass this anomaly.

The following proposition, stated in [38], collects the main properties of the scheme. The proof follows the guidelines of Harten's theorem [50] (Theorem 1.3 in chapter 1) in order to ensure the TVD property.

Proposition 4.1. *If $h(x)$ in (4.15) satisfies the following inequalities*

$$\begin{aligned} 1 \leq h(x) \leq \frac{1}{x}, \qquad & 0 < x \leq 1 \\ \frac{1}{x} \leq h(x) \leq -1, \qquad & -1 \leq x < 0 \end{aligned} \tag{4.18}$$

then the scheme (4.14) with the flux defined by (4.15) is TVD under the CFL restriction

$$\max_i |\tilde{a}^n_{i+\frac{1}{2}} + \tilde{b}^n_{i+\frac{1}{2}}| \leq 1. \tag{4.19}$$

As observed in [38], the requirements placed on h in the proposition above imply that the resulting scheme is only first order accurate. In addition, as we shall see shortly, heavy restrictions on the time step might also result from the requirement (4.19) in practical numerical simulations.

In order to convert a first order TVD scheme, of the form specified in the proposition above, into a second order accurate TVD scheme Gascón and Corberán in [38] exploit the *homogeneous form* (4.2) and use a technique developed in [50] for the homogeneous case. The basic idea is to apply the TVD first order scheme to the modified equation of the first order scheme. For the case under study, a first order accurate TVD scheme is applied to

$$u_t + f_x = s + \frac{1}{2}\left(h(\tilde{a}+\tilde{b}) - (\tilde{a}+\tilde{b})\right)(\triangle x)\left(f_{xx} - s_x\right),$$

after rewriting it in the form

$$u_t + (g+\psi)_x = 0$$

where

$$\psi \approx \frac{1}{2}\left(h(\tilde{a}+\tilde{b}) - (\tilde{a}+\tilde{b})\right)(\triangle x) g_x.$$

As mentioned in the introduction of this chapter, it is not clear that one would want to design a TVD scheme for non-homogeneous conservation laws, since the TVD property might not be preserved for solutions to the scalar equation.

In the following section, we propose to implement a flux limiter alternative in order to *partially* impose TVD requirements. Our objective is to obtain the benefits of the non-oscillatory character of TVD schemes

for homogeneous conservation laws, without imposing conditions on the scheme for the non-homogeneous case that fully ensure the TVD property.

Before we start our derivation, we shall point out another *anomaly* that we have observed when applying the scheme proposed by Gascón and Corberán, even in its simplest first order version.

We consider the first order scheme (4.15), with $h(x) =$sign(x), which is one of the proposed choices in [38] satisfying the requirements in proposition 4.1. We have applied it to solve the model problem

$$\begin{aligned} u_t + u_x &= -u \\ u_0(x) &= u(x,0) = \begin{cases} 1, & x \le x_d, \\ 0, & x > x_d, \end{cases} \end{aligned}$$

with $x_d = 0.2$, whose solution is

$$u(x,t) = u_0(x-t)e^{-t}.$$

We set $\triangle x = 10^{-2}$, and a CFL restriction in (4.19) equal to 0.9, where we implement (4.20) and (4.17) in the following form

$$\tilde{a}^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{\triangle t}{\triangle x}\dfrac{f^n_{i+1} - f^n_i}{U^n_{i+1} - U^n_i}, & \text{if } |U^n_{i+1} - U^n_i| > \varepsilon, \\ \dfrac{\triangle t}{\triangle x}\dfrac{\partial f}{\partial u}\Big|^n_i, & \text{otherwise,} \end{cases} \tag{4.20}$$

$$\tilde{b}^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{\triangle t}{\triangle x}\dfrac{b^n_{i+1} - b^n_i}{U^n_{i+1} - U^n_i}, & \text{if } |U^n_{i+1} - U^n_i| > \varepsilon, \\ 0, & \text{otherwise,} \end{cases} \tag{4.21}$$

with $\varepsilon = 10^{-5}$. The numerical solution at time $t = 0.3$ is shown in figure 4.2 left. In our numerical simulation, 8344 time-steps were needed to obtain the numerical solution. Upon close examination, we find that the CFL-like restriction (4.19) leads to an effective restriction on the time step so that $\triangle t \simeq \mathcal{O}(\varepsilon)$. However, if we set $\tilde{b}^n_{i+\frac{1}{2}} = 0$ for all i in 4.19, only 34 time steps were needed to obtain figure 4.2 right, since in this case $\triangle t \simeq \mathcal{O}(\triangle x)$.

The condition (4.19), which ensures the TVD property of the numerical scheme, is indeed a strong restriction in terms of the temporal step size required. Moreover, even in the non-homogeneous case, the upwind

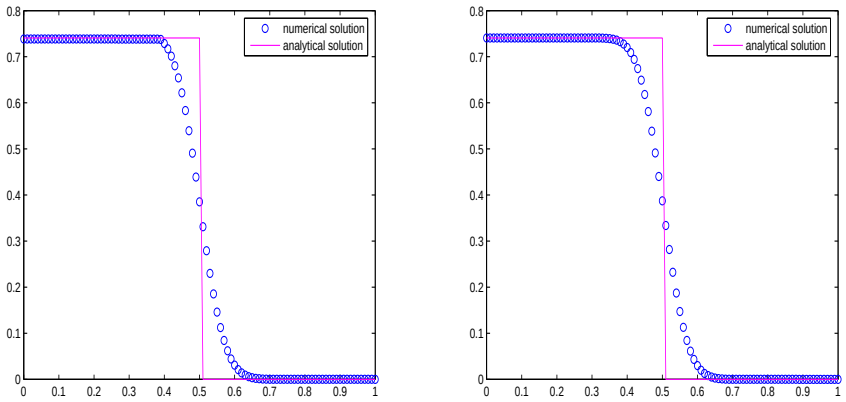


Figure 4.2: *Numerical solution for* $u_t + u_x = -u$ *at* $t = 0.3$*, obtained with the first order TVD scheme of Gascón and Corberán, with* $h(x) =$*sign*(x) *and a CFL*$= 0.9$*. Left: CFL*$= \max_i |\tilde{a}^n_{i+\frac{1}{2}} + \tilde{b}^n_{i+\frac{1}{2}}|$*. Right: CFL*$= \max_i |\tilde{a}^n_{i+\frac{1}{2}}|$*.*

direction is determined by the sign of $f'(u)$ in $[x_i, x_{i+1}]$. This is clearly seen by examining the particular case $f(u) = \lambda u$, $s(x,u) = k(x)u$ [10], [81]

$$u(x,t)_t + \lambda u(x,t)_x = k(x)u(x,t),$$

whose the solution is

$$u(x,t) = u(x - \lambda t, 0) + \int_0^t k(x - \lambda(t-\xi))u(x - \lambda(t-\xi), \xi)d\xi.$$

Notice that each term on the RHS exhibits an upwind domain of dependence, which is determined by the wind direction $f'(u) = \lambda$.

Other numerical schemes for non-homogeneous conservation laws [5], [71], [81] take into account only f_u in order to determine the *upwind direction* in the numerical scheme. However, the derivation in [38], is based on (4.5) and (4.15), that is, it uses a discrete equivalent of $\partial g/\partial u$ so that the quantities that determine the upwind directions are computed from g.

The schemes proposed in the next section will also take this specific point into account.

4.2

A Well Balanced second order scheme

Let us first provide a *consistent* derivation of the extension of the Lax-Wendroff scheme to scalar balance laws recalled in the previous section. By *consistent*, we mean that any use of the dependence of the *combined flux* g in terms of u will be avoided. For this, we observe, as in [10], that the most convenient way to deal with the modified flux function $g = f + b$ is to consider

$$g = g(x,t) = f(u(x,t)) + b(x,t), \tag{4.22}$$

with

$$b(x,t) = -\int_0^x s(y,u(y,t))dy. \tag{4.23}$$

From these, we easily obtain (using that $u_t = -g_x$ when necessary)

$$\begin{aligned} g_x(x,t) &= f_u(u(x,t))u_x(x,t) - s(x,u(x,t)) && (4.24)\\ g_t(x,t) &= f_u(u(x,t))u_t(x,t) - \int_0^x s_u(y,u(y,t))u_t(y,t)dy \\ &= -f_u(u(x,t))g_x(x,t) + \int_0^x s_u(y,u(y,t))g_x(y,t)dy. && (4.25) \end{aligned}$$

These relations will allow us to carry out the derivation of a second order scheme in an straightforward manner, without using $\partial b/\partial u$, or any discrete equivalent.

We first observe that a second order method is obtained by considering a scheme on the form

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x}(\hat{g}_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \hat{g}_{i-\frac{1}{2}}^{n+\frac{1}{2}}), \tag{4.26}$$

where $\hat{g}_{i+\frac{1}{2}}^{n+\frac{1}{2}}$ is defined as

$$\hat{g}_{i+\frac{1}{2}}^{n+\frac{1}{2}} := \hat{g}_{i+\frac{1}{2}}^{n} + \frac{\triangle t}{2}\hat{g}_t\big|_{i+\frac{1}{2}}^{n}, \tag{4.27}$$

as long as the quantities $\hat{g}_{i+1/2}^n$ and $\hat{g}_t|_{i+1/2}^n$ satisfy appropriate approximation properties:

Proposition 4.2. *The scheme (4.26)-(4.27) is a second order scheme for (4.1), provided that the conditions:*

$$\text{C1.} \qquad \frac{1}{\triangle x}\left(\hat{g}^n_{i+\frac{1}{2}} - \hat{g}^n_{i-\frac{1}{2}}\right) = (g_x)^n_i + \mathcal{O}\left(\triangle x^2\right),$$

$$\text{C2.} \qquad \frac{1}{\triangle x}\left(\hat{g}_t\big|^n_{i+\frac{1}{2}} - \hat{g}_t\big|^n_{i-\frac{1}{2}}\right) = (g_{tx})^n_i + \mathcal{O}(\triangle x^2)$$

are satisfied.

Proof. Let us compute formally the local truncation error (LTE). Assume that $U^n_i = u(x_i, t_n)$ where $u(x,t)$ is a smooth solution of (4.1). Then

$$\frac{U^{n+1}_i - U^n_i}{\triangle t} = \left(u_t + \frac{\triangle t}{2}u_{tt}\right)^n_i + \mathcal{O}(\triangle t^2). \tag{4.28}$$

Conditions C1 and C2 lead to

$$\frac{1}{\triangle x}\left(\hat{g}^{n+\frac{1}{2}}_{i+\frac{1}{2}} - \hat{g}^{n+\frac{1}{2}}_{i-\frac{1}{2}}\right) = \frac{1}{\triangle x}\left(\hat{g}^n_{i+\frac{1}{2}} - \hat{g}^n_{i-\frac{1}{2}}\right) + \frac{\triangle t}{2\triangle x}\left(\hat{g}_t\big|^n_{i+\frac{1}{2}} - \hat{g}_t\big|^n_{i-\frac{1}{2}}\right) \tag{4.29}$$

$$= g_x\big|^n_i + \frac{\triangle t}{2}g_{xt}\big|^n_i + \mathcal{O}(\triangle x^2). \tag{4.30}$$

Hence,

$$\text{LTE}^n_i = \left(u_t + \frac{\triangle t}{2}u_{tt} + g_x + \frac{\triangle t}{2}g_{xt}\right)^n_i + \mathcal{O}(\triangle t^2 + \triangle x^2) = \mathcal{O}(\triangle t^2 + \triangle x^2), \tag{4.31}$$

since u (smooth) satisfies $u_t + g_x = 0$ and also $u_{tt} + g_{xt} = 0$. ■

We also have the following proposition:

Proposition 4.3. *Let us define*

$$\hat{g}^n_{i+\frac{1}{2}} := \frac{g^n_{i+1} + g^n_i}{2}, \tag{4.32}$$

$$\hat{g}_t\big|^n_{i+\frac{1}{2}} := -f_u\big|^n_{i+\frac{1}{2}}\frac{g^n_{i+1} - g^n_i}{\triangle x} + b_t\big|^n_{i+\frac{1}{2}}, \tag{4.33}$$

where

$$b_t\big|^n_{i+\frac{1}{2}} = \int_0^{x_{i+\frac{1}{2}}} \frac{\partial s}{\partial u}(y, u(y,t_n))g_y(y,t_n)dy, \tag{4.34}$$

Then, the scheme (4.26)-(4.27) is second order accurate.

Proof. We shall prove that conditions C1 and C2 in proposition 4.2 are satisfied.

For the definition in (4.32) we have

$$\frac{1}{\triangle x}\left(\hat{g}^n_{i+\frac{1}{2}} - \hat{g}^n_{i-\frac{1}{2}}\right) = \frac{g^n_{i+1} - g^n_{i-1}}{2\triangle x} = g_x\big|^n_i + \mathcal{O}(\triangle x^2), \tag{4.35}$$

as it is readily deduced from the Taylor developments

$$g^n_{i+1} = g^n_i + \triangle x g_x\big|^n_i + \frac{\triangle x^2}{2} g_{xx}\big|^n_i + \mathcal{O}(\triangle x^3), \tag{4.36}$$

$$g^n_{i-1} = g^n_i - \triangle x g_x\big|^n_i + \frac{\triangle x^2}{2} g_{xx}\big|^n_i + \mathcal{O}(\triangle x^3). \tag{4.37}$$

Hence condition C1 is satisfied.

An analogous procedure is used in order to check the C2 condition. A Taylor development easily leads to

$$\frac{g^n_{i+1} - g^n_i}{\triangle x} = g_x(x_{i+1/2}, t_n) + \mathcal{O}(\triangle x^2)$$

hence

$$f_u|^n_{i+1/2}\ \frac{g^n_{i+1} - g^n_i}{\triangle x} = (f_u g_x)|^n_{x_i+1/2} + \mathcal{O}(\triangle x^2)$$

provided that $f_u(u)$ is a bounded function in the region of interest. Then, taking into account that

$$b_t\big|^n_{i+\frac{1}{2}} = \frac{\partial}{\partial t}\int_0^{x_{i+\frac{1}{2}}} -s(y, u(y, t_n))dy = \int_0^{x_{i+\frac{1}{2}}} \frac{\partial s}{\partial u}(y, u(y, t_n)) g_y(y, t_n)dy,$$

and (4.25), we get that $\hat{g}_t$ in (4.33) satisfies

$$\hat{g}_t|^n_{i+1/2} = (f_t + b_t)(x_{i+1/2}, t_n) + \mathcal{O}(\triangle x^2) = g_t(x_{i+1/2}, t_n) + \mathcal{O}(\triangle x^2).$$

It readily follows by Taylor expansions that

$$\frac{1}{\triangle x}\left(g_t\big|^n_{i+\frac{1}{2}} - g_t\big|^n_{i-\frac{1}{2}}\right) = (g_{tx})^n_i + \mathcal{O}(\triangle x^2)$$

which ensures C2.

■

In order to write the final form of the numerical scheme (4.26)-(4.27), with the definitions in (4.32) and (4.33), we rewrite first

$$\hat{g}^{n+1/2}_{i+1/2} = \widehat{\mathcal{G}}^n_{i+1/2} + \frac{\Delta t}{2} b_t|^n_{i+1/2},$$

with

$$\widehat{\mathcal{G}}^n_{i+\frac{1}{2}} = \frac{1}{2}\left(g^n_{i+1} + g^n_i - \frac{\triangle t}{\triangle x}\frac{\partial f}{\partial u}\Big|^n_{i+\frac{1}{2}}(g^n_{i+1} - g^n_i)\right). \tag{4.38}$$

Notice that the conservative form of the scheme will lead to the computation of differences of the form

$$b^n_{k,l} := b^n_k - b^n_l = \int_{x_k}^{x_l} s(y, u(y, t_n))dy \tag{4.39}$$

and

$$b_t\big|^n_{i+\frac{1}{2}} - b_t\big|^n_{i-\frac{1}{2}} = \int_{x_{i-1/2}}^{x_{i+1/2}} s_u(y, u(y, t_n))g_x(y, t_n)dy \tag{4.40}$$

which need to be computed by a quadrature rule, in general. By following the steps of the proof, it is easy to see that C1 and C2 will also hold for any quadrature rule so that the error is, at least, $\mathcal{O}(\triangle x^2)$.

In particular, we apply the trapezoidal rule to the expression in (4.40) and obtain

$$\begin{aligned}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} s_u g_x dx &= \left(s_u g_x\Big|^n_{i+\frac{1}{2}} + s_u g_x\Big|^n_{i-\frac{1}{2}}\right)\frac{\triangle x}{2} + \mathcal{O}(\triangle x^2) \\ &= \left(s_u\big|^n_{i+\frac{1}{2}}\frac{g^n_{i+1} - g^n_i}{\triangle x} + s_u\big|^n_{i-\frac{1}{2}}\frac{g^n_i - g^n_{i-1}}{\triangle x}\right)\frac{\triangle x}{2} + \mathcal{O}(\triangle x^2).\end{aligned} \tag{4.41}$$

With these considerations, it is clear that the following scheme

$$\begin{aligned}U^{n+1}_i &= U^n_i - \frac{\triangle t}{\triangle x}(\widehat{\mathcal{G}}^n_{i+\frac{1}{2}} - \widehat{\mathcal{G}}^n_{i-\frac{1}{2}}) \\ &- \frac{\triangle t^2}{4\triangle x}\left(\frac{\partial s}{\partial u}\Big|^n_{i+\frac{1}{2}}(g^n_{i+1} - g^n_i) + \frac{\partial s}{\partial u}\Big|^n_{i-\frac{1}{2}}(g^n_i - g^n_{i-1})\right),\end{aligned} \tag{4.42}$$

is second order accurate when computing smooth solutions of (4.2).

We remark that, in the scalar case, any second order approximation of $f_u|^n_{i+1/2}$ and $s_u|^n_{i+1/2}$ will also maintain the second order accuracy of the scheme on smooth solutions. We, hence, introduce the coefficients

$$\alpha_{i+1/2} \approx \frac{\triangle t}{\triangle x} f_u|^n_{i+1/2}, \qquad \beta_{i+1/2} \approx \frac{\triangle t}{2}\, s_u|^n_{i+1/2} \tag{4.43}$$

and write the previous scheme in the more general form

$$U^{n+1}_i = U^n_i - \frac{\triangle t}{\triangle x}(\mathcal{G}^n_{i+\frac{1}{2}} - \mathcal{G}^n_{i-\frac{1}{2}}) - \frac{\triangle t}{\triangle x}\mathcal{S}^n_i, \tag{4.44}$$

with

$$\mathcal{G}^n_{i+\frac{1}{2}} = \frac{1}{2}\left(g^n_{i+1} + g^n_i - \alpha^n_{i+\frac{1}{2}}(g^n_{i+1} - g^n_i)\right), \tag{4.45}$$

$$\mathcal{S}^n_i = \frac{1}{2}\left(\beta^n_{i+\frac{1}{2}}(g^n_{i+1} - g^n_i) + \beta^n_{i-\frac{1}{2}}(g^n_i - g^n_{i-1})\right). \tag{4.46}$$

In the scalar case, we can take

$$\alpha^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{\triangle t}{\triangle x}\dfrac{f^n_{i+1} - f^n_i}{U^n_{i+1} - U^n_i}, & \text{if } U^n_{i+1} - U^n_i \neq 0 \\ \dfrac{\triangle t}{\triangle x}\dfrac{\partial f}{\partial u}\Big|^n_i, & \text{if } U^n_{i+1} - U^n_i = 0, \end{cases} \tag{4.47}$$

and,

$$\beta^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{\triangle t}{2}\dfrac{s^n_{i+1} - s^n_i}{U^n_{i+1} - U^n_i}, & \text{if } U^n_{i+1} - U^n_i \neq 0 \\ \dfrac{\triangle t}{2}\dfrac{\partial s}{\partial u}\Big|^n_i, & \text{if } U^n_{i+1} - U^n_i = 0, \end{cases} \tag{4.48}$$

which provide convenient second order approximations to $f_u|^n_{i+1/2}$ and $s_u|^n_{i+1/2}$.

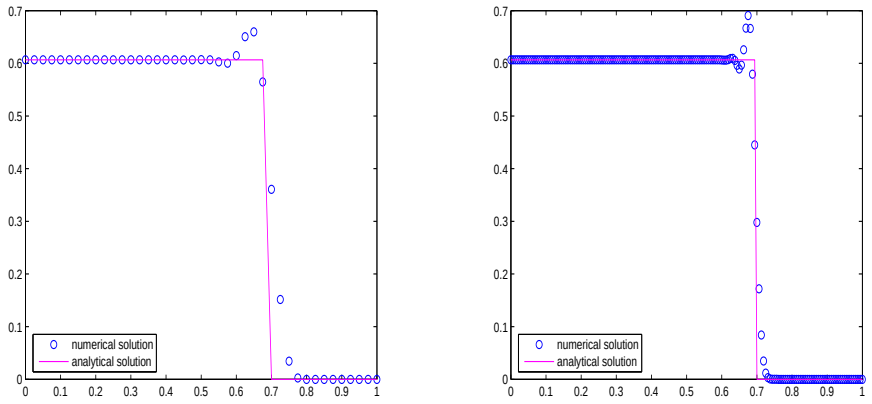
Remark We notice that, by following the technique of Gascón and Corberán and writing the non-homogeneous equation (4.1) in *homogeneous form* (4.2), the analysis of the well balancing properties of the scheme is greatly simplified. In fact, any stationary solution of (4.1) will satisfy $g_x = 0$, that is $f_x = s$, hence

$$f_{i+1} - f_i = \int_{x_i}^{x_{i+1}} f_x(u(x))dx = \int_{x_i}^{x_{i+1}} s(u,x)dx = -b_{i,i+1}$$

which implies $g_{i+1} - g_i = 0$. This is a discrete equivalent of $g_x = 0$ that ensures $U^{n+1}_i = U^n_i$ for all n, i in (4.44).

Hence, the formal scheme, where the quantities $b^n_{i,k}$ are defined by an integral as in (4.39), is always well balanced. At the fully discrete level, where a quadrature rule of the appropriate accuracy is employed for the computation of $b^n_{i,k}$, exact balancing will obviously depend on the particular integration rule employed.

However, any quadrature rule that is at least second order accurate will lead to an approximate well-balanced scheme, or in the terminology of Bermudez and Váquez-Cendón [5], to a numerical scheme that satisfies the approximate C-property.



$\triangle x = 0.025$ $\triangle x = 0.00625$

Figure 4.3: *Second order scheme applied to* $u_t + u_x = -u$ *with CFL*$= 0.9$.

In figure 4.3, we show numerical results for the model problem in the previous section obtained with the proposed second order scheme, using the trapezoidal rule for the computation of the $b_{i,i+1}$ terms, although a different quadrature rule does not alter the numerical results significantly. Spurious oscillations, typical of the Lax Wendroff scheme in the homogeneous case, are generated near the discontinuity and can be clearly observed.

In order to construct a numerical scheme that maintains the second order accuracy away from discontinuities, while producing monotone (or nearly monotone) discrete profiles at discontinuities, we shall follow the "flux-limited" technology, instead of modified-equation approach of Gascón and Corberán.

4.3

A partially limited numerical flux: The TVDB scheme

We propose to limit only the part of the numerical flux function of the combined scheme, i.e. $\hat{g}_{i+1/2}^{n+1/2}$ in (4.26) that contains the *upwind direction* $f_u|_{i+1/2}^n$. Hence, we shall consider a limited version of $\mathcal{G}_{i+1/2}^n$ in (4.44) of

the form (1.84)

$$\bar{\mathcal{G}}^n_{i+\frac{1}{2}} = \mathcal{G}^{LO}_{i+\frac{1}{2}} + \phi^n_{i+\frac{1}{2}}(\mathcal{G}^{HI}_{i+\frac{1}{2}} - \mathcal{G}^{LO}_{i+\frac{1}{2}}), \tag{4.49}$$

where $\phi^n_{i+\frac{1}{2}}$ is an appropriate a flux limited function which will be defined below. We use (4.45) as the high order numerical flux, $\mathcal{G}^{HI}_{i+\frac{1}{2}}$. Our choice for the low order numerical flux is

$$\mathcal{G}^{LO}_{i+\frac{1}{2}} = \frac{1}{2}(g^n_{i+1} + g^n_i - \mathrm{sgn}(\alpha^n_{i+\frac{1}{2}})(g^n_{i+1} - g^n_i)). \tag{4.50}$$

The function $\mathrm{sgn}(x)$ is the signum function of a real number x, which is defined as follows:

$$\mathrm{sgn}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases}$$

Carrying out the algebra, we obtain the following limited numerical flux function,

$$\mathcal{G}^n_{i+\frac{1}{2}} = \mathcal{G}^{LO}_{i+\frac{1}{2}} + \phi^n_{i+\frac{1}{2}} \left(\mathrm{sgn}(\alpha^n_{i+\frac{1}{2}}) - \alpha^n_{i+\frac{1}{2}}\right) \left(g^n_{i+1} - g^n_i\right). \tag{4.51}$$

The limiter $\phi^n_{i+\frac{1}{2}}$ is defined by using a flux limiter function ϕ acting on a quantity that measures the ratio of the upwind change to the local change (see chapter 1). We remark that the limited numerical flux (4.51) is written in terms of the differences $g_{k+1} - g_k$. It then seems reasonable to define the ratio of the upwind change to the local change, as:

$$r^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{g^n_i - g^n_{i-1}}{g^n_{i+1} - g^n_i} & \mathrm{sgn}(\alpha^n_{i+\frac{1}{2}}) > 0; \\ \\ \dfrac{g^n_{i+2} - g^n_{i+1}}{g^n_{i+1} - g^n_i} & \mathrm{sgn}(\alpha^n_{i+\frac{1}{2}}) < 0. \end{cases} \tag{4.52}$$

As a flux limiter function $\phi^n_{i+\frac{1}{2}} = \phi(r^n_{i+\frac{1}{2}})$, we use the minmod limiter[1],

$$\phi^n_{i+\frac{1}{2}} = \max(0, \min(r^n_{i+\frac{1}{2}}, 1)). \tag{4.53}$$

The resulting method will be referred to as the **TVDB** scheme, for the remaining of this chapter. Notice that this method is expressed in terms of differences $g_{i+1} - g_i$, therefore it is well balanced in the same sense as the second order scheme described in the previous section.

[1] In general, no significant differences have been observed by using other limiter functions. Later on we show one example where the differences are more noticeable

In figure 4.4 left, we display the numerical results obtained after applying the TVDB scheme to $u_t + u_x = -u$. We can clearly observe a significant reduction in the oscillations present in the computed solution. In fact, a very mild oscillatory behavior can be observed after zooming into the post-discontinuity region, see figure 4.5. In figure 4.5, we ob-

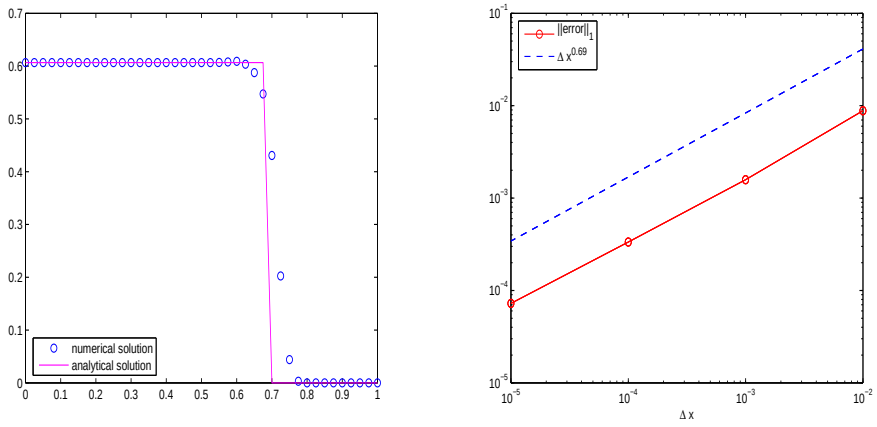


Figure 4.4: $u_t+u_x = -u$. *Left: Numerical solution obtained with the TVDB method using* 40 *points, a CFL*$= 0.9$ *and time* 0.5. *Right: Error for the TVDB method.*

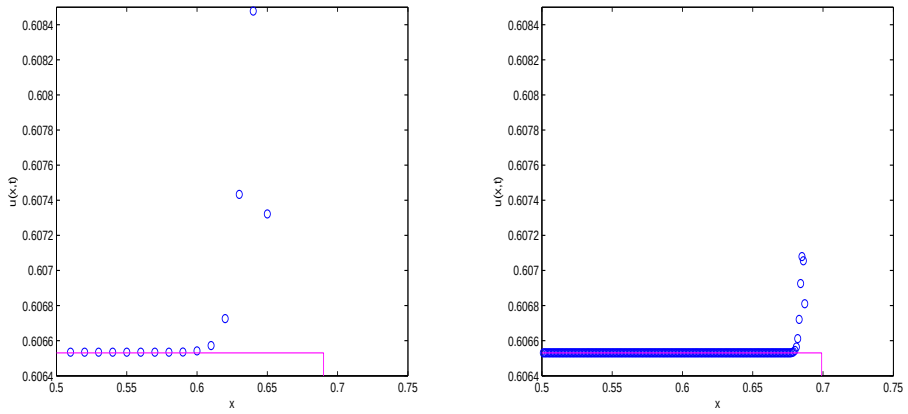
serve that the amplitude if the oscillations decreases with the mesh size, as opposed to the behavior of the oscillations produced by the original second order scheme.

We can argue that the slight oscillatory behavior displayed by the TVDB scheme might be due to the fact that the source term is still included in the limited flux, via de differences $g_{i+1} - g_i$. In the following section, we take our idea one step further and carry out the limiting process only on those terms that do not contain, in any form, a source term contribution.

4.4 The TVDF method

To construct a limited scheme where the limiting process does not affect any of the source term contributions, we start by recalling that

$$g^n_{i+1} - g^n_i = f^n_{i+1} - f^n_i + b^n_{i,i+1},$$



$\triangle x = 10^{-2}$ $\triangle x = 10^{-3}$

$\triangle x = 10^{-4}$ $\triangle x = 10^{-5}$

Figure 4.5: *Zoom of the numerical solution of* $u_t + u_x = -u$ *using TVDB method with different grid sizes* $\triangle x = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$.

where

$$b_{i,k}^n = \int_{x_i}^{x_k} -s(\xi, u(\xi, t_n))d\xi + \mathcal{O}(\triangle x^2),$$

then, we write $\mathcal{G}_{i+1/2}^n$ in (4.45) as

$$\mathcal{G}_{i+1/2}^n = \frac{1}{2}\left(f_i^n + f_{i+1}^n + b_i^n + b_{i+1}^n - \alpha_{i+1/2}^n(f_{i+1}^n - f_i^n + b_{i,i+1}^n)\right).$$

Taking into account that

$$b_{i+1}^n - b_{i-1}^n = \int_{x_{i-1}}^{x_{i+1}} -s(y, u(y, t_n))dy = b_{i-1,i}^n + b_{i,i+1}^n,$$

we can rewrite

$$\mathcal{G}_{i+1/2}^n - \mathcal{G}_{i-1/2}^n = \mathcal{F}_{i+\frac{1}{2}}^n - \mathcal{F}_{i-\frac{1}{2}}^n + (1 - \alpha_{i+1/2}^n)b_{i,i+1} + (1 + \alpha_{i-1/2})b_{i-1,i},$$

with

$$\mathcal{F}_{i+\frac{1}{2}} = \frac{1}{2}(f_{i+1}^n + f_i^n - \alpha_{i+\frac{1}{2}}^n(f_{i+1}^n - f_i^n)).$$

Then, (4.45) can be also expressed as

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x}(\mathcal{F}_{i+\frac{1}{2}}^n - \mathcal{F}_{i-\frac{1}{2}}^n) - \frac{\triangle t}{\triangle x}\tilde{\mathcal{S}}_i^n, \tag{4.54}$$

with

$$\begin{aligned}\tilde{\mathcal{S}}_i^n &= \frac{1}{2}((1 - \alpha_{i+\frac{1}{2}}^n)b_{i,i+1}^n + (1 + \alpha_{i-\frac{1}{2}}^n)b_{i-1,i}^n) \\ &\quad + \frac{1}{2}(\beta_{i+\frac{1}{2}}^n(f_{i+1}^n - f_i^n + b_{i,i+1}^n) + \beta_{i-\frac{1}{2}}^n(f_i^n - f_{i-1}^n + b_{i-1,i}^n)).\end{aligned}$$

We then consider a partially limited version of the second order scheme (4.54) by defining the limited flux

$$\mathcal{F}_{i+\frac{1}{2}}^n = \mathcal{F}_{i+\frac{1}{2}}^{LO} + \phi_{i+\frac{1}{2}}^n\left(\text{sgn}(\alpha_{i+\frac{1}{2}}^n) - \alpha_{i+\frac{1}{2}}^n\right)\left(f_{i+1}^n - f_i^n\right),$$

where $\mathcal{F}_{i+\frac{1}{2}}^n$ is (4.49) with:

$$\mathcal{F}_{i+\frac{1}{2}}^{HI} = \frac{1}{2}(f_{i+1}^n + f_i^n - \alpha_{i+\frac{1}{2}}^n(f_{i+1}^n - f_i^n)), \tag{4.55}$$

$$\mathcal{F}_{i+\frac{1}{2}}^{LO} = \frac{1}{2}(f_{i+1}^n + f_i^n - \text{sgn}(\alpha_{i+\frac{1}{2}}^n)(f_{i+1}^n - f_i^n)). \tag{4.56}$$

We shall refer to this new scheme as the **TVDF** scheme. Now, the ratio for measuring smoothness is defined as in the homogeneous case

$$r^n_{i+\frac{1}{2}} = \begin{cases} \dfrac{f^n_i - f^n_{i-1}}{f^n_{i+1} - f^n_i}, & \operatorname{sgn}(\alpha^n_{i+\frac{1}{2}}) > 0; \\ \dfrac{f^n_{i+2} - f^n_{i+1}}{f^n_{i+1} - f^n_i}, & \operatorname{sgn}(\alpha^n_{i+\frac{1}{2}}) < 0. \end{cases} \tag{4.57}$$

and the flux limiter function $\phi^n_{i+\frac{1}{2}}$ is also the minmod limiter (4.53).

We apply this scheme to the model problem $u_t + u_x = -u$ with discontinuous initial data and show the results in figure 4.6. We observe that the numerical solution obtained does not display any oscillations, even under zooming.

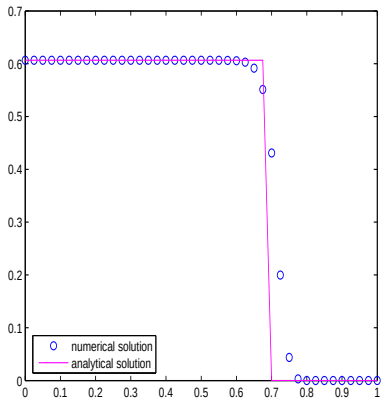


Figure 4.6: *Numerical solution of* $u_t + u_x = -u$ *using TVDF scheme with* 40 *points, a CFL*$= 0.9$ *and time* 0.5*.*

It should be noticed that this scheme is not written only in terms of $g_{i+1} - g_i$, hence there is no guarantee, a priori, that the scheme is well balanced. This difference with respect to the TVDB scheme will become more evident in the next section, where we test both schemes for nonlinear scalar conservation laws.

4.5 Nonlinear scalar balance laws

In this section, we shall apply the numerical schemes developed in the previous section to nonlinear scalar balance laws, paying special attention to the computation of steady-state solutions. The prototype homogeneous conservation law is Burger's equation, and we shall consider different source terms that have become standard test problems in the relevant literature (see e.g. [38], [28], [46], [47]).

4.5.1 Greenberg et al. tests

In [47], the authors consider the model balance law

$$u_t + (\frac{u^2}{2})_x = a_x(x) \tag{4.58}$$

with

$$a(x) = \begin{cases} 0, & x < -1; \\ \cos^2(\pi \frac{x}{2}), & -1 \leq x \leq 1; \\ 0, & 1 < x. \end{cases} \tag{4.59}$$

and different initial conditions. Since $s(x,u) = a_x(x)$, the flux-limited schemes described in the previous two sections can be carried out by performing a direct integration of the terms $b^n_{i,k}$. In addition, and since $s_u(x,u) \equiv 0$, we take $\beta^n_{i+1/2} \equiv 0$.

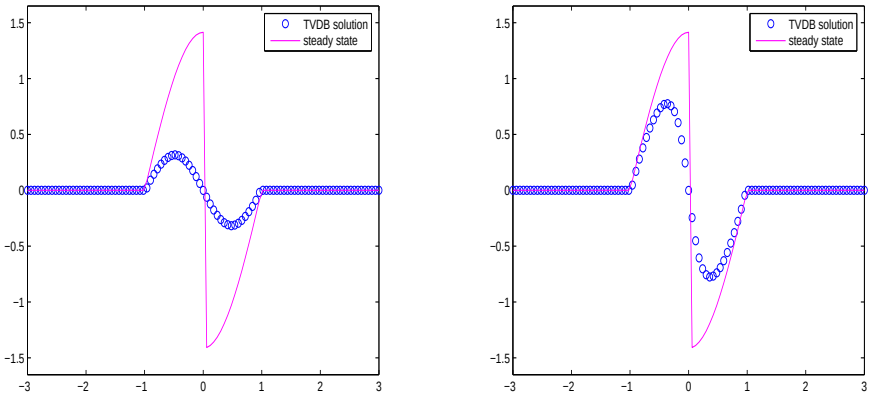
We shall apply the TVDB scheme to the test problems in [47], showing the numerical solution at the same times displayed in [47], i.e. at times $t = 0.2, 0.5, 1$ and 3, so that a direct comparison can be made. Details about the exact solution on each one of the tests presented in this section can be found in [47]. In each case, there is a transient solution that converges, as time evolves, to a steady state solution of the balance law (4.58).

Snapshots of the numerical solution corresponding to the initial condition $u(x,0) = 0$ are shown in figure 4.7. As the time advances, the

solution approaches the steady solution

$$u_1(x) = \begin{cases} 0, & x < -1; \\ \sqrt{2}\cos(\pi\frac{x}{2}), & -1 \le x \le 0; \\ \sqrt{2}\cos(\pi\frac{x}{2}), & 0 \le x \le 1; \\ 0, & 1 < x. \end{cases} \tag{4.60}$$

shown as a solid line in figure 4.7. The snapshots in figure 4.7 show the transient solution at different times. No oscillations are observed, as expected.



$t = 0.2$ $t = 0.5$

$t = 1$ $t = 3$

Figure 4.7: *Evolution of the numerical solution of* $u_t + (\frac{u^2}{2})_x = a_x(x)$ *with initial condition,* $u(x,0) = 0$, $\mathrm{CFL} = 0.9$, 100 *nodes.*

For the second experiment we consider the following initial condition,

$$u(x,0) = \begin{cases} 0, & x < -1; \\ 1, & -1 \le x \le 1; \\ 0, & 1 < x. \end{cases} \tag{4.61}$$

In figure (4.8) snapshots are shown at $t = 0.2, 0.5, 1$ and 3. In this case a shock emerges from $x = 1$ which is correctly captured, without spurious oscillatory behavior, as the transient solution evolves to the steady state solution below,

$$u_2(x) = \begin{cases} 0, & x < -1; \\ \sqrt{2}\cos(\pi\frac{x}{2}), & -1 \le x \le 1; \\ 0, & 1 < x. \end{cases} \tag{4.62}$$

For the third experiment (see figure 4.9) the initial data is

$$u(x,0) = \begin{cases} 0, & x < -1; \\ -1, & -1 \le x \le 1; \\ 0, & 1 < x. \end{cases} \tag{4.63}$$

Here, a shock emerges from $x = -1$. The solution in this case converges to the steady solution

$$u_3(x) = \begin{cases} 0, & x < -1; \\ -\sqrt{2}\cos(\pi\frac{x}{2}), & -1 \le x \le 1; \\ 0, & 1 < x. \end{cases} \tag{4.64}$$

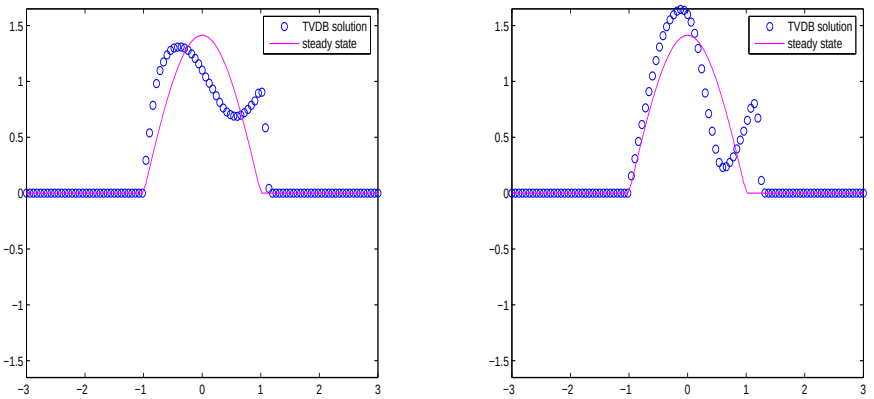
The fourth experiment (see figure 4.10) considers a different source term, given as

$$a(x) = \begin{cases} 0, & x < -1; \\ -\cos^2(\pi\frac{x}{2}), & -1 \le x \le 1; \\ 0, & 1 < x. \end{cases} \tag{4.65}$$

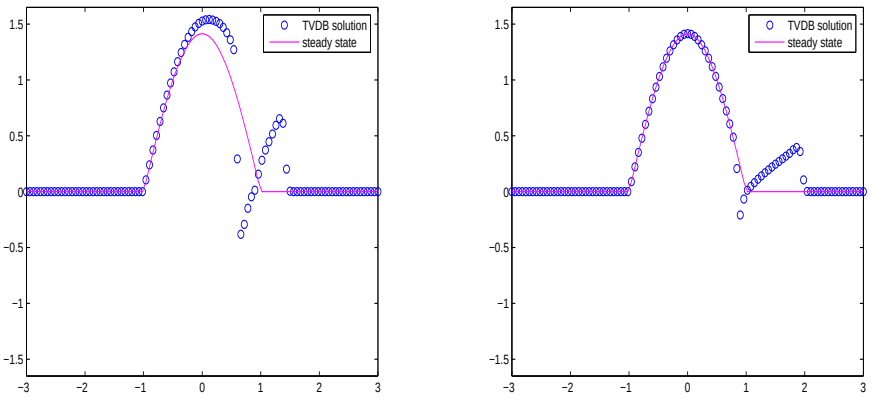
with initial data $u(x,0) = 0$.

Snapshots of the solution at time $t = 0.2, 0.5, 1$ and 3, are shown in figure 4.10. In this case, shocks are generated at $x = -1$ and $x = 1$. In the region bounded by the shocks the solution converges to

$$u_4(x) = \begin{cases} -\sqrt{2}, & x < -1; \\ -\sqrt{2}(1-\cos^2(\pi\frac{x}{2}))^{1/2}, & -1 \le x \le 0; \\ \sqrt{2}(1-\cos^2(\pi\frac{x}{2}))^{1/2}, & 0 \le x \le 1; \\ \sqrt{2}, & 1 < x. \end{cases} \tag{4.66}$$

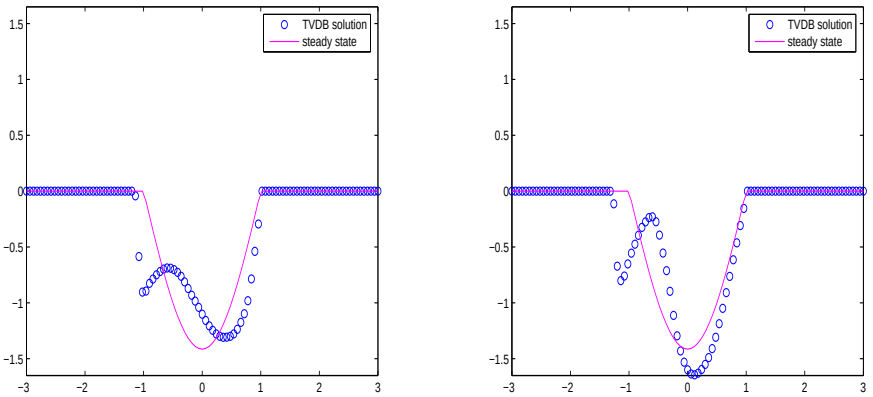


$t = 0.2$ $t = 0.5$

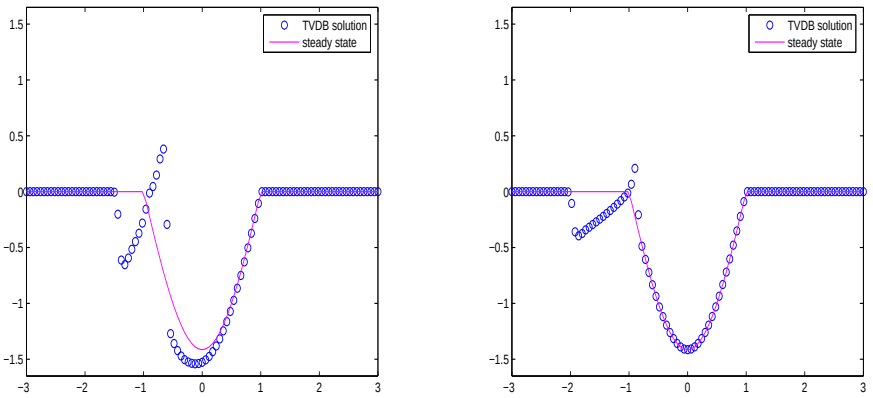


$t = 1$ $t = 3$

Figure 4.8: *Evolution of the numerical solution of* $u_t + (\frac{u^2}{2})_x = a_x(x)$ *with initial condition (4.61),* CFL $= 0.9$, 100 *nodes.*



$t = 0.2$ $t = 0.5$



$t = 1$ $t = 3$

Figure 4.9: *Evolution of the numerical solution of* $u_t + (\frac{u^2}{2})_x = a_x(x)$ *with initial condition (4.63),* $\mathrm{CFL} = 0.9$, 100 *nodes.*

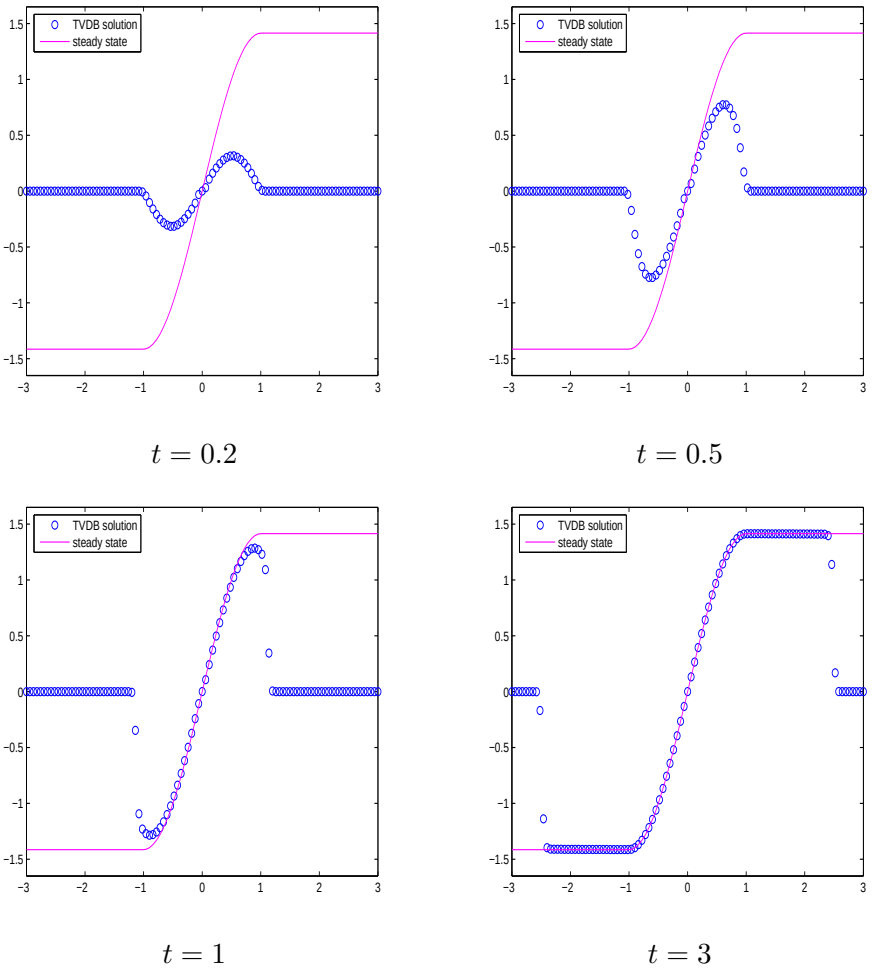


Figure 4.10: *Evolution of the numerical solution of* $u_t + (\frac{u^2}{2})_x = a_x(x)$ *with initial condition* $u(x,0) = 0$, CFL $= 0.9$, 100 *nodes.*

The same numerical tests have been performed for the TVDF scheme obtaining similar results (not shown). These experiments show that the flux-limited schemes perform in a non-oscillatory manner, as expected, in transient solutions leading to steady-state situations.

The model problem considered in the first four experiments allows us to test the schemes in the special case where no numerical integration formula is required. In addition, the scheme is simplified by the fact that $s_u \equiv 0$. The following test problem considers a more general situation, where no such simplifications are possible.

In [46] the authors consider the model problem

$$u_t + (\frac{u^2}{2})_x + a_x(x)u = 0, \tag{4.67}$$

where

$$a(x) = 0.9 \begin{cases} 0, & x < 0; \\ (\cos(\pi \frac{x-1}{2}))^{30}, & 0 \leq x \leq 2; \\ 0, & 2 < x. \end{cases} \tag{4.68}$$

Notice that in this case $s_u(x,u) = -a_x(x)$. This has also been used in the design of the scheme in order to avoid numerical integrations in the definition of $\mathcal{S}_i^n$. No further simplifications can be made in this case, and the terms $b_{i,k}^n$ have been computed by the trapezoidal rule.

As in [46], we shall perform two sets of numerical experiments, corresponding to two different initial conditions for the model balance law.

The initial conditions considered are

$$u + a = 1, \qquad -\infty < x < \infty \tag{4.69}$$

and

$$u + a = \begin{cases} 1.3, & x \leq 0.2; \\ 1, & x > 0.2. \end{cases} \tag{4.70}$$

Numerical solutions obtained with the TVDB scheme at time $t = 1.5$ are shown in figure 4.11 right. As in [46], we display the variable $u + a$, and the bottom graph in these figures is the function $a(\cdot)$.

We see in figure 4.11 left that the equilibrium $u+a \equiv 1$ is exactly maintained by the TVDB scheme. We have computed the l_1-error between the numerical solution and the true solution, and its value is $6.9044 \cdot 10^{-17}$. For the TVDF scheme l_1-error is $2.7065 \cdot 10^{-17}$ (see figure 4.12 left).

The solution corresponding to the TVDB scheme for the initial data in (4.70) is displayed in figure 4.11 right. We observe that the shock moving to the right is cleanly represented by the TVDB scheme and no spurious oscillatory behavior is observed, even under zooming. This should be

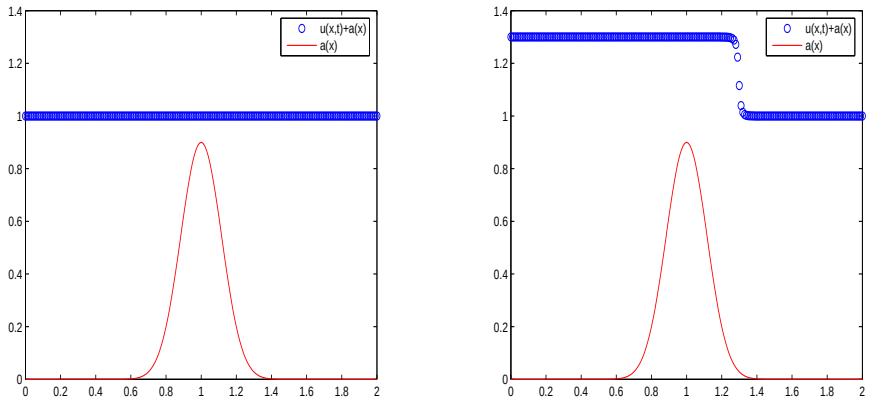


Figure 4.11: *Numerical solution using TVDB scheme of* $u_t + (\frac{u^2}{2})_x = -a_x(x)u$. *Left:* $u(x,0) = 1 - a(x)$. *Right:* $u(x,0) = 1.3 - a(x)$ *for* $x < 0.2$ *and* $u(x,0) = 1 - a(x)$ *otherwise.*

compared to the example shown in section 4.3, where a slightly oscillatory behavior was observed. In our opinion this is most likely due to the nature of the discontinuity that appears in the simulations, a shock in this case, versus a contact discontinuity in the model problem in section 4.3.

The simulation corresponding to the TVDF scheme is shown in figure 4.12 right. In this case, noticeable oscillations appear around the shock profile.

This experiments puts in evidence a flaw of the TVDF scheme, when used to compute quasi-steady flow. In the first test case for this model problem, $u = 1 - a(x)$, hence $u(x)$ is a smooth function and the TVDF scheme and TVDB scheme both equal the original second order scheme, which is well balanced. The results obtained simply reflect this fact.

In the second test case, u is no longer a smooth function. The TVDB scheme is well balanced, so it does not produce spurious oscillatory behavior. However, this is not the case for the TVDF scheme. In figure 4.13 we show numerical results obtained by considering only $\mathcal{F}^{LO}_{i+1/2}$ in (4.56) in (4.54). The numerical solution obtained in figure 4.13 clearly indicates that this first order scheme is already not well-balanced. In the TVDF scheme, formula (4.56) is activated near the discontinuous profile, leading to the oscillatory behavior observed.

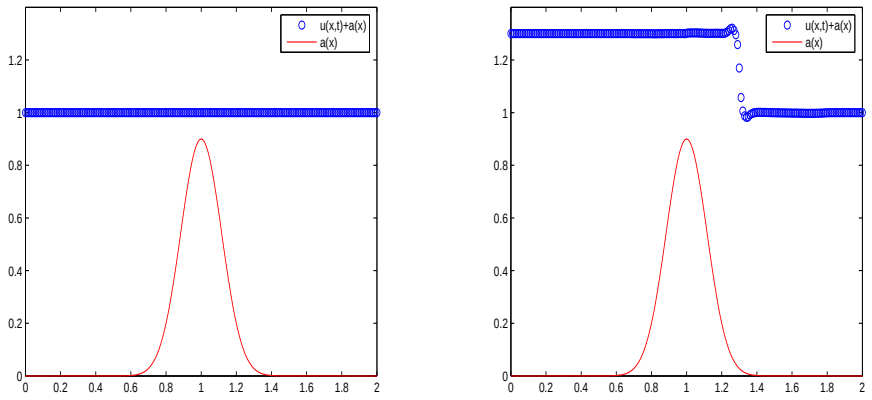


Figure 4.12: *Numerical solution using TVDF scheme of* $u_t + (\frac{u^2}{2})_x = -a_x(x)u$. *Left:* $u(x,0) = 1 - a(x)$. *Right:* $u(x,0) = 1.3 - a(x)$ *for* $x < 0.2$ *and* $u(x,0) = 1 - a(x)$ *otherwise.*

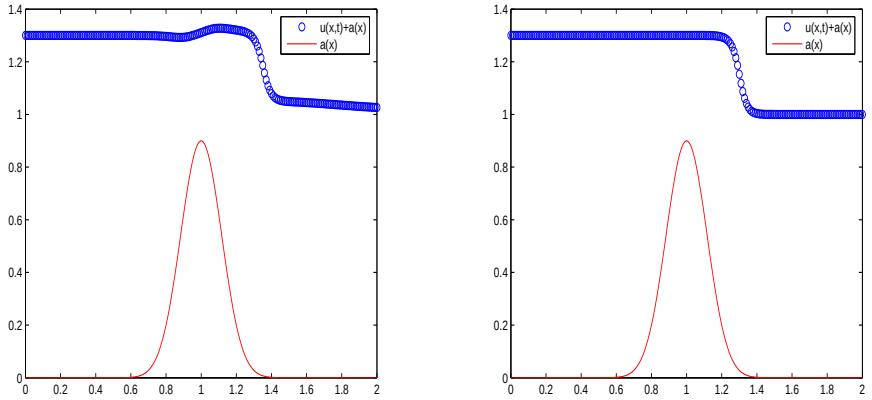


Figure 4.13: *First order scheme of* $u_t + (\frac{u^2}{2})_x = -a_x(x)u$ *with* $u(x,0) = 1.3 - a(x)$ *for* $x < 0.2$ *and* $u(x,0) = 1 - a(x)$ *otherwise. Left: TVDF scheme. Right: TVDB scheme.*

4.5.2 The Embid problem

This problem was presented in [28] as a simple scalar approximation to the 1-D equations that model the flow of a gas through a duct of variable cross-section.

$$\begin{cases} u_t + (\frac{u^2}{2})_x = (6x-3)u, & 0 < x < 1 \\ u(0,t) = 1, u(1,t) = -0.1. \end{cases} \tag{4.71}$$

There are two entropy satisfying steady solutions for the Embid problem. One is stable in time with a standing shock at $x_1 = 0.18$ and the other with an unstable standing shock at $x_2 = 0.82$. The steady-state solutions for the Embid problem are

$$u(x) = \begin{cases} 1 + 3x^2 - 3x, & x < x_i; \\ -0.1 + 3x^2 - 3x, & x > x_i, \end{cases} \tag{4.72}$$

for $i = 1, 2$. We run our numerical schemes by taking initial data with a jump at the stable location, using a CFL number equal to 0.8.

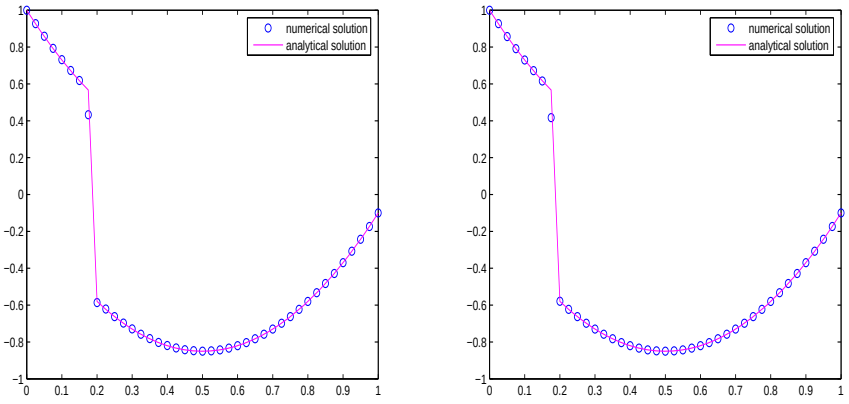


Figure 4.14: *Stationary solution for the Embid problem by marching the scheme until the difference between two consecutive iterations is less than* 10^{-10}. *Left: TVDF scheme. Right: TVDB scheme.*

The results with the TVDF and TVDB schemes are both very similar reproducing the exact steady solution, except for one internal shock point. Figure 4.15 shows the logarithm of residual errors with respect to the number of iterations for both schemes.

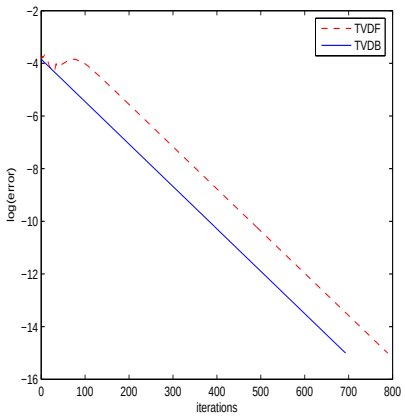


Figure 4.15: *Convergence histories of the TVDF and TVDB schemes for the Embid problem.*

As a result of this series of numerical tests for scalar conservation laws, we select the TVDB scheme as the most robust scheme to perform numerical simulations for shallow water flows, where it is often required to accurately simulate quasi-steady flows. As, we have seen, the TVDF scheme is able compute non-oscillatory solutions for a variety of transient flows that include discontinuities, however, it fails to perform properly near certain steady-state solutions.

4.6

Extension to systems: The Shallow water equations

The shallow water equations form a hyperbolic system of conservation laws that approximately describes various geophysical flows. We give a brief derivation of the equations, as well as some properties of the solution in Appendix B. We consider source terms due to topography, but we do not consider wind effects and Coriolis force. In this case, the one dimensional shallow water equations are as follows

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{s}(x, \mathbf{u}), \tag{4.73}$$

where

$$\mathbf{u} = \begin{pmatrix} h \\ q \end{pmatrix}, \quad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} q \\ \frac{q^2}{h} + \frac{g}{2}h^2 \end{pmatrix}, \quad \text{and} \quad \mathbf{s}(x,\mathbf{u}) = \begin{pmatrix} 0 \\ -ghz_x \end{pmatrix}.$$

In this section we propose an extension of the TVDB scheme to this system via the so-called characteristic-based approach, see [31], [27] and chapter 1. This technique makes use of the fact that all the eigenvalues of the Jacobian matrix of the convective flux vector are real, and the matrix is diagonalizable, i.e., there is a complete set of N linearly independent eigenvectors. This information is readily available for the shallow water system, see Appendix B.

For the non-homogeneous system above, we follow Gascón and Corberán's technique and rewrite the system as

$$\mathbf{u}_t + (\mathbf{f} + \mathbf{b})_x = 0 \tag{4.74}$$

where

$$\mathbf{b} = -\int_{\bar{x}}^{x} \mathbf{s}(y, \mathbf{u}(y,t))dy \tag{4.75}$$

The computation of the numerical flux function in characteristic-based schemes is based on the ability of the Jacobian of the convective fluxes to *reduce* the original system into a system of nearly independent scalar equations that can be *independently* discretized by a convenient scheme for scalar conservation laws.

Taking into account that, for scalar conservation laws, the TVDB scheme is of the form (4.44) and using the same arguments as in section 1.6, we arrive at the following extension of the TVDB scheme for the system case,

$$U_i^{n+1} = U_i^n - \frac{\triangle t}{\triangle x}\left(\mathcal{G}_{i+1/2}^n - \mathcal{G}_{i-1/2}^n\right) - \frac{\triangle t}{\triangle x}\mathcal{S}_i^n, \tag{4.76}$$

with

$$\mathcal{G}_{i+1/2}^n = \sum_{p=1}^{2} \mathcal{G}_{i+1/2}^{n,p} R_{i+1/2}^{n,p}. \tag{4.77}$$

Here $R_{i+1/2}^{n,p}$ are the right eigenvectors at the interface state, and the characteristic fluxes $\mathcal{G}_{i+1/2}^{n,p}$ are computed by pre-multiplying by the corresponding left eigenvectors, $L_{i+1/2}^{n,p}$, as specified in section 1.6. We shall assume that the interface state $U_{i+1/2}$ is given by the Roe mean of the states U_i, U_{i+1}. For the shallow water system we have

$$\hat{h}_{i+\frac{1}{2}} = \frac{1}{2}\left(h_i + h_{i+1}\right) \quad \text{and} \quad \hat{u}_{i+\frac{1}{2}} = \frac{\sqrt{h_i}u_i + \sqrt{h_{i+1}}u_{i+1}}{\sqrt{h_i} + \sqrt{h_{i+1}}}. \tag{4.78}$$

It should be noted that the source term $\mathcal{S}_i^n$ is not affected by the characteristic decomposition, due to the specific form of the TVDB scheme in the scalar case.

Numerical approximation of the source term contributions

For the shallow water system, the form of $\mathbf{b}(x,t)$ in (4.75) leads to a source term contribution that involves the terms

$$B_{i,i+1} = \begin{pmatrix} 0 \\ \int_{x_i}^{x_{i+1}} ghz_x dx \end{pmatrix}. \tag{4.79}$$

Hence, we need to define numerical approximation of the integrals

$$b_{i,i+1} = \int_{x_i}^{x_{i+1}} ghz_x dx. \tag{4.80}$$

These integrals are approximated as in [10], by applying the trapezoidal rule as follows,

$$b_{i,i+1} \approx \int_{x_i}^{x_{i+1}} ghz_x dx = g\left((hz_x)_i + (hz_x)_{i+1}\right)\frac{\triangle x}{2}. \tag{4.81}$$

If the topography and the flow are smooth, we can write

$$\begin{aligned} (hz_x)_i &= h_i\left(\frac{z_{i+1}-z_i}{\triangle x} - \frac{\triangle x}{2}(z_{xx})_i + \mathcal{O}(\triangle x^2)\right) \\ (hz_x)_{i+1} &= h_{i+1}\left(\frac{z_{i+1}-z_i}{\triangle x} + \frac{\triangle x}{2}(z_{xx})_{i+1} + \mathcal{O}(\triangle x^2)\right). \end{aligned}$$

Replacing these terms in (4.81) we obtain

$$b_{i,i+1} = \frac{g}{2}(z_{i+1}-z_i)(h_i+h_{i+1}) + \mathcal{O}(\triangle x^2).$$

Hence, as in [10], we use the following discrete realizations of the integral terms $b_{i,i+1}$

$$\hat{b}_{i,i+1} = \frac{g}{2}(z_{i+1}-z_i)(h_i+h_{i+1}) \tag{4.82}$$

in the computation of the numerical flux functions. For smooth flows, these approximations respect the second order accuracy of our scheme.

In addition, we also need to approximate derivatives of the form

$$\left.\frac{\partial S}{\partial U}\right|_{i+\frac{1}{2}} = \begin{pmatrix} 0 & 0 \\ -gz_x\big|_{i+\frac{1}{2}} & 0 \end{pmatrix}.$$

Here, $z_x\big|_{i+\frac{1}{2}}$ is approximated by using Taylor series as

$$z_x\big|_{i+\frac{1}{2}} \approx \frac{z_{i+1} - z_i}{\triangle x},$$

which maintains the second order accuracy for a smooth topography.

4.6.1 C-property

Bermúdez and Vázquez [5] and Vázquez-Cendón [104] discussed an approach for approximating source terms which is designed for quasi-steady and steady flow. Consider the shallow water equation for the quiescent flow case,

$$q(x,t) = 0 \quad \text{and} \quad h(x,t) + z(x,t) = D \quad \forall (x,t).$$

For this stationary flow, the numerical scheme satisfies:

- the approximate C-property, if the numerical scheme is accurate to the order $\mathcal{O}(\triangle x^2)$
- the exact C-property, if the numerical scheme is exact.

It is observed in [5] that if a numerical scheme does not satisfy the C-property (exact or approximate) then spurious waves may occur in the numerical results.

Notice that, by construction, all terms in the TVDB scheme are written in terms of differences of the form $(\mathbf{f}+\mathbf{b})_{i+1}-(\mathbf{f}+\mathbf{b})_i$. For steady-state flows this difference is zero if the integral terms are computed exactly. When using an integral rule, the exact balance might not be respected, and we might only get an approximate balance. However, the definition of the approximate integrals $\hat{b}_{i,i+1}$ in (4.82) allows us to prove the exact C-property for quiescent flow.

Indeed, as observed in [10], since $q = 0$ and $h + z = D$ =constant, we also get $\partial_x(h + z) = 0$, thus

$$\mathbf{f}(U) = \begin{pmatrix} 0 \\ \frac{1}{2}gh^2 \end{pmatrix} \quad \text{and} \quad B_{i,i+1} = \begin{pmatrix} 0 \\ -\int_{x_i}^{x_{i+1}} ghh_x \end{pmatrix}.$$

Since $h_{i+1} + z_{i+1} = h_i + z_i$,

$$\begin{pmatrix} 0 \\ -\int_{x_i}^{x_{i+1}} ghh_x \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{g}{2}\left(h_{i+1}^2 - h_i^2\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{g}{2}\left(z_{i+1} - z_i\right)\left(h_{i+1} + h_i\right) \end{pmatrix}.$$

Hence, for quiescent flows $b_{i,i+1} = \hat{b}_{i,i+1}$ and the scheme satisfies the exact C-property in a rather natural way.

This results will be numerically validated in the next section.

4.6.2 Numerical experiments

The following series of numerical experiments are standard in the literature. Firstly, and in order to give a numerical validation of the C-property, we consider the following steady flow cases.

Steady flow

Following [106], let consider a channel with a length of 20 m defined as

$$z(x) = 0.2e^{-\frac{2}{5}(x-10)^2}, \tag{4.83}$$

with the quiescent state $h + z = 2$m ($q = 0$) as initial condition. As expected, the TVDB scheme exactly preserves the steady state (see figure 4.16), because the $\mathcal{L}^1$ error of the numerical solution at time 50 s is $3.3267 \cdot 10^{-14}$, which is roundoff error.

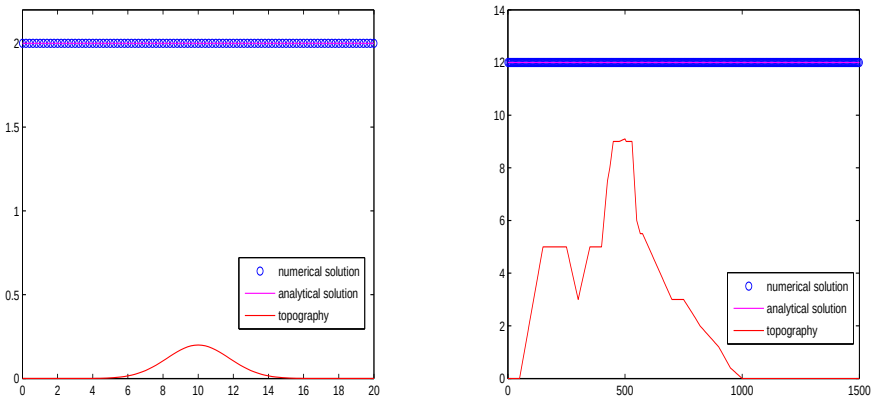


Figure 4.16: *Flow at rest. Left: Smooth topography. Right: Complex topography.*

Usually, the bottom topography is not smooth. With the aim of evaluating the performance of a numerical scheme in the presence of complex and possibly non-smooth geometry, the following experiment was proposed in the workshop on dam-break wave simulation [45]. The initial

data is the water at rest at a level of 12m. Numerical results obtained after a simulation of 200s are displayed in figure (4.16), we can observed that also in this test the $\mathcal{L}^1$ error, which is $1.4989 \cdot 10^{-15}$ is roundoff error.

The interest of the next three tests (extracted from [45]) is to study the convergence of this scheme towards a steady state. In these simulations a bottom topography of 25m length is defined as:

$$z(x) = \begin{cases} 0.2 - 0.05(x-10)^2, & 8\text{m} < x < 12\text{m} \\ 0, & \text{otherwise.} \end{cases} \tag{4.84}$$

In all cases, the initial data are $h + z$ =constant and $q = 0$. The analytical solution is computed with the Bernoulli equation

$$\frac{q^2}{2gh^2} + h + z = H_a,$$

where H_a is the upstream head, q is the steady discharge and h the water level.

For the initial conditions are $h + z = 2$m, and $q = 0$ and boundary conditions

- downstream: $h = 2$ m
- upstream: $q = 4.42$ m$^2/s$.

The resulting flow is a subcritical flow (4.17). If we use as boundary conditions

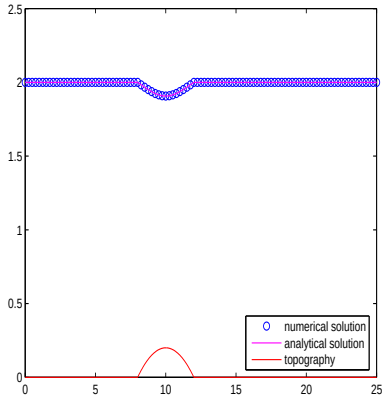


Figure 4.17: *Subcritical flow over a hump.*

- downstream: $h = 0.66$ m only when $F_r < 1$
- upstream: $q = 1.53$ m^2/s.

where $F_r = u/\sqrt{gh}$ is the Froude number, and $h + z = 0.66$ m and $q = 0$ as initial condition, then a transcritical flow without shocks is obtained, see figure 4.18 left. Transcritical flow, with a shock, is obtained if $h + z =$

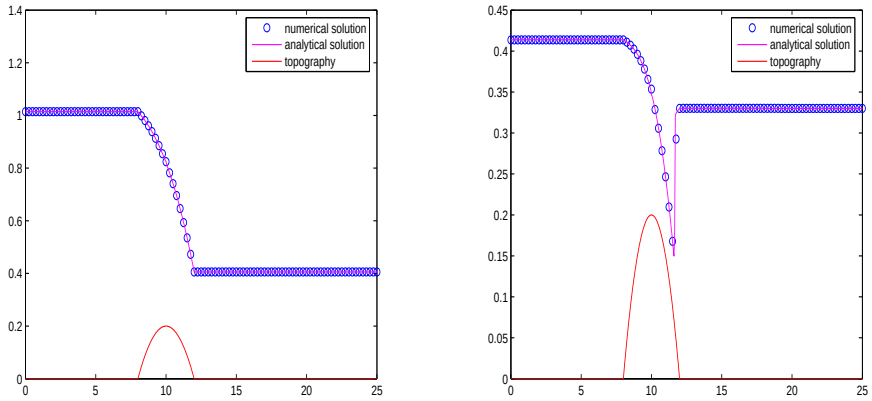


Figure 4.18: *Transcritical flow over a hump.Left: Without shock. Right: with shock.*

0.33 m and $q = 0$ is used as initial data, and in this case the boundary conditions are

- downstream: $h = 0.33$ m
- upstream: $q = 0.18$ m^2/s.

Quasi stationary flow

This last test were proposed in [71] by LeVeque in order to evaluate the capability of the scheme to accurately compute small perturbations of the water surface over a variable topography, in this case is given as

$$z(x) = \begin{cases} 0.25\left(\cos\left(\pi\dfrac{x-0.5}{0.1}\right)+1\right), & \text{if } |x-0.5| < 0.1 \\ 0, & \text{otherwise,} \end{cases} \tag{4.85}$$

on $0 < x < 1$and with $g = 1$. The initial condition is

$$\begin{aligned} h + z &= 1 + 0.001 \text{ for } 0.1 < x < 0.2 \\ q &= 0, \end{aligned}$$

which represents a small hump perturbation of the quiescent state $(h, q) = (1 - z, 0)$. LeVeque uses this test to show the disadvantages of schemes that do not preserve steady states. In figure 4.19, we show the numerical result at time 0.7, using different limiters. We could observe in this example the main features of both limiters, also explained in chapter 1.

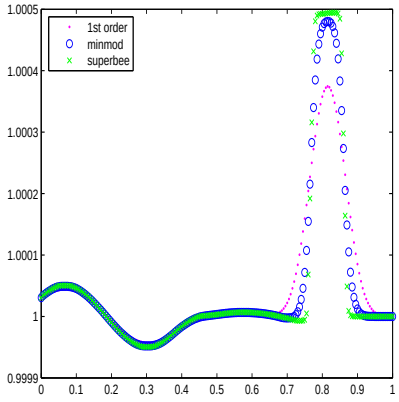


Figure 4.19: *Numerical solution for the quasi stationary flow, $t = 0.7$s, for the first order and second order using two different limiters.*

4.7 Conclusions

A general technique to construct numerical methods with a capacity to recognize steady solutions for hyperbolic conservation law systems with source terms has been presented. We follow the technique proposed by Gascón and Corberán in [38], where the nonhomogeneous conservation law is re-written in a *homogeneous form* by introducing a new flux function which is generated by adding the primitive of the source term to the physical flux.

In order to design a second order scheme that has the properties of TVD schemes for scalar hyperbolic conservation laws, we first derive a second order extension of the Lax-Wendroff scheme for the nonhomogeneous case. Then, we use the flux-limiting technology applied *partially* on the numerical flux function of the second order scheme. The motivation for the partial limiting is that the nature of the numerical os-

cillations obtained when applying the Lax-Wendroff scheme to the computation of discontinuous solutions is only due to the treatment of the convective derivative.

We explored two different partial limiting techniques, in particular with respect to their ability to compute accurately steady state solutions of balance laws. We found that TVDB scheme is the most robust for numerical simulations dealing with balance laws in quasi-steady flows. The extension to shallow water equations in one dimension has also been carried out. We have shown that the TVDB scheme satisfies the C-property for quiescent steady states and have performed a series of numerical tests that demonstrate the capabilities of the scheme.

It is important to remark that, as observed in [10], our technique does not amount to converting the balance law into a homogeneous conservation law. The definition of $b(x,t) = \int^x -s(y,u(y,t))dy$ allows to *express* the balance law in *homogeneous form*. However, this is only used to carry out an appropriate numerical treatment of the source term. As in [10], our numerical technique employs only the characteristic speeds coming from the convective term and, as in [10], we have always obtained the correct entropy solution in all our numerical experiments.

5

A multiscale scheme for systems of balance laws

The numerical simulation of physical problems modeled by systems of conservation laws is difficult due to the presence of discontinuities in the solution. High-resolution shock capturing (HRSC) schemes succeed in computing highly accurate numerical solutions, typically second-or third- order in smooth regions, while maintaining sharp, oscillation-free numerical profiles at discontinuities. The power of a HRSC scheme lies usually in the computation of the numerical flux function of the scheme which is often expensive. This is, in fact, the main drawback of these schemes, specially in multi-dimensional computations.

It is well known, however, that the costly numerical flux function of

a HRSC scheme is only necessary in a neighborhood of singularities, therefore, Harten in [49] proposed various multiscale schemes based on reducing the computational cost using the smoothness information obtained from a multiresolution transform of the data. The goal is to save time in the evaluation of numerical flux functions, while maintaining the overall accuracy of the high resolution scheme. This is achieved by replacing the expensive numerical flux evaluations with a cheap polynomial interpolation in the smooth regions.

Harten's original multilevel strategy was developed for finite volume schemes for homogeneous conservation laws [49], where the numerical data are naturally treated as approximations to the cell-averages of the true solution. Chiavassa and Donat in [16] applied the same cost-reduction strategy to finite-difference schemes, where the data are interpreted as point-values. This allows, in a rather natural way, to think of the basic ingredients of the smoothness analysis as interpolation errors, so that its relation to the smoothness of the underlying function is easy and well understood.

In this chapter, we apply a straightforward extension of the multilevel technique developed in [16] to non-homogeneous systems of conservation laws. In particular, we shall investigate the properties of the extended multilevel technique, in terms of efficiency and quality, by a series of numerical experiments on the shallow water system.

We also mention here that there has been some recent work on the inclusion of source terms in *fully adaptive* multilevel strategies [66]. Fully adaptive schemes are not only cost-efficient, but also significant memory gains can be achieved through the full exploitation of the multilevel structure of the scheme, see [18] for a detailed account of this technique in the homogeneous case, and [83], [25] for later developments on these *fully adaptive* multilevel schemes. These techniques do require special data structures in order to obtain the expected memory gains, and its incorporation into an existing code is not straightforward. The cost effective scheme of [16], which will be followed in this chapter, can be incorporated almost as an external routine, at each time step, and remains the easiest multilevel alternative to adapt to an existing code.

The shallow water equations in one and two dimensions are used to model real-life applications. In many cases, the flow regime is steady or quasi-steady, and much effort has been devoted to design numerical techniques that are capable to preserve steady states at the discrete level as well as to accurately compute the evolution of small dynamical perturbations of these steady states. The inclusion of the source term in a direct discretization of the system becomes a non-trivial issue, because

many schemes do not preserve stationary solutions. In this chapter we shall concentrate on the following two schemes

- The TVDB scheme presented in the chapter 4, which uses a flux-limited technique.
- The HRSC scheme proposed in [10].

In [10], the authors seek to obtain an extension of the numerical scheme developed by Donat and Marquina in [26], that avoids the use of averaged quantities in computing the numerical flux function at cell interfaces (1-Jacobian), for non-homogeneous conservation laws by incorporating the idea of flux gradient and source term balancing in [38]. However, the extension based on the use of two spectral decompositions at each computational interface (2-Jacobian) does not satisfy the exact C-property of [5], and a combined scheme is proposed (1-Jacobian/2-Jacobian).

Both schemes follow Gascón and Corberán's strategy of writing the non-homogeneous conservation law in *homogeneous form*. In our extension, the inclusion of the source term is, in a natural way, done directly through the numerical divergence operator.

These two schemes have numerical flux formulae which are significantly different in terms of computational effort. The TVDB scheme uses a Roe's linearization and involves one Jacobian (1J) evaluation per cell interface, while the scheme proposed in [10] combine the use of two Jacobian evaluations per cell interface (1J-2J), hence the latter is more expensive. This fact will be used in order to investigate the properties of our multilevel technique.

We shall first explain the main steps of the 1D algorithm, and will carry out some numerical experiments on well known test problems. We shall see that the multilevel strategy preserves the properties of the basic scheme with respect to well-balancing. Finally, we present the 2D algorithm and perform preliminary numerical tests in 2D.

5.1 The 1D multilevel algorithm

The multilevel strategy that we shall employ here relies on the smoothness analysis of the discrete data. The general setting for homogeneous

conservation laws has been described an analyzed in [16]. In what follows we recall the main steps and explain our extension to the non-homogeneous case.

5.1.1 Smoothness analysis

The most important step in the multilevel algorithm concerns the smoothness analysis of the data, and how this information is used. The different resolution levels are specified by a set of nested grids $\{\chi^l, \quad l = 0, 1, \ldots, L\}$, which in 1D are given as follows,

$$x_i \in \chi^l \iff x_{2^l i} \in \chi^0. \tag{5.1}$$

Here χ^0 is considered the finest grid.

Let us consider $(v_i^0)_i$, the point-values of a function v on χ^0. Due to the embedding of the grids, the representation of the function on the coarser grid χ^l, its point values on χ^l, is

$$v_i^l = v_{2^l i}^0, \quad i = 0, \ldots, N_x/2^l. \tag{5.2}$$

Notice that we have the following relation between the discrete data in χ^l and χ^{l-1}

$$v_i^l = v_{2i}^{l-1} \qquad i \in \chi^l. \tag{5.3}$$

To recover the representation of v on χ^{l-1} from the representation on χ^l (the next coarser grid), a set of predicted values is first computed, $\tilde{v}_i^{l-1}$, by polynomial interpolation on the data of the χ^l grid.

$$\begin{aligned} \tilde{v}_i^{l-1} &= v_{i/2}^l \qquad \text{if } x_i \in \chi^l \\ \tilde{v}_i^{l-1} &= \mathcal{I}\left[x_i; v^l\right] \quad \text{if } x_i \in \chi^{l-1} \backslash \chi^l. \end{aligned} \tag{5.4}$$

Next, we describe the interpolation operation $\mathcal{I}\left[x; v^l\right]$. To achieve maximal accuracy a centered interpolatory technique is used. A polynomial which interpolates the points $\left(v_{i-s}^{l-1}, \ldots, v_{i+s-1}^{l-1}\right)$, $r = 2s$, is constructed and evaluated at the appropriate locations on the χ^l grid. The interpolatory property gives $\tilde{v}_{2i}^{l-1} = v_{2i}^l$, while it is easy to obtain that

$$\tilde{v}_{2i-1}^{l-1} = \sum_{k=1}^{s} \beta_k \left(v_{i+k-1}^l + v_{i-k}^l\right). \tag{5.5}$$

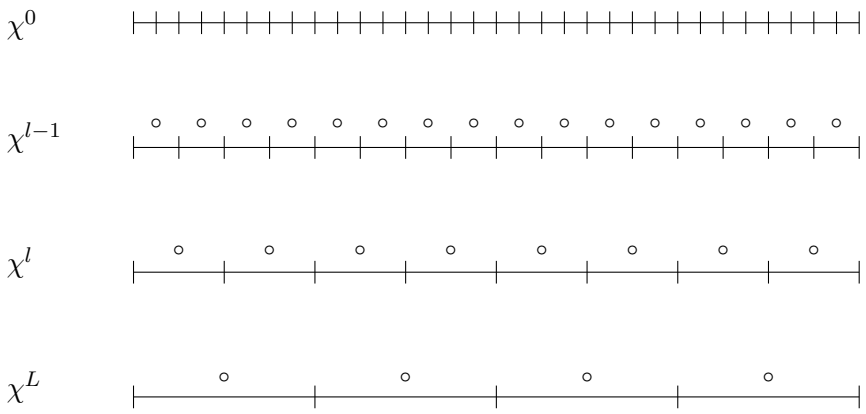


Figure 5.1: *Different resolution levels in one dimension. Grid points (+) and detail coefficients (∘)*

The coefficients $\beta_k, k = 1, \ldots, s$ come from standard interpolation results. In this work, we use a 4rd order interpolation procedure ($r = 4$), so we have $\beta_1 = \frac{9}{16}$, $\beta_2 = \frac{-1}{16}$.

The difference between the exact values (5.2), v_i^{l-1}, and the predicted values $\tilde{v}_i^{l-1}$ is represented by the details, or wavelet coefficients:

$$d_i^l = v_i^{l-1} - \tilde{v}_i^{l-1}, \quad x_i \in \chi^{l-1}. \tag{5.6}$$

Observe that the equation (5.4) implies that $d_i^l = 0$ for $i \in \chi^l$. We then have

$$\begin{aligned} v_i^{l-1} &= v_{i/2}^l, && \text{if } x_i \in \chi^l \\ v_i^{l-1} &= \tilde{v}_i^{l-1} + d_i^l && \text{if } x_i \in \chi^{l-1} \backslash \chi^l. \end{aligned} \tag{5.7}$$

Relations (5.3) and (5.7) show that the sets $\{v^l\}$ and $\{v^{l-1}, d^l\}$ are equivalent (notice that they have the same cardinality), in the sense that we can obtain one set from the other in a one-to-one way. When these transformations are carried out for all the levels involved, we have a multiresolution representation of the data on the finest grid $\{v^0\}$ as the data on the coarsest grid $\{v^L\}$ together with a sequence of details $\{d^1, \ldots, d^L\}$ representing the difference in information between the different resolution levels. In figure 5.1, we show the different resolution levels, the grid points and the details.

The detail coefficients are simply interpolation errors, which can then be used directly as "regularity sensors" to localize the nonsmooth behavior. When applying the point-value multiresolution transform to the

numerical solution, large values of the detail coefficients correspond to non-smooth zones of the solution, like shocks.

5.1.2 General framework

Let us consider the 1D system of balance laws:

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{s}(x, \mathbf{u}), \tag{5.8}$$

and rewrite in the form $\mathbf{u}_t + \mathbf{g}_x = 0$, with $\mathbf{g} = \mathbf{f}(\mathbf{u}) + \mathbf{b}$ where

$$\mathbf{b}_i(x,t) = -\int_{\bar{x}}^{x} \mathbf{s}_i(\xi, \mathbf{u}(\xi,t))d\xi.$$

We consider a semi-discrete formulation, where the spatial discretization of the system, on a Cartesian grid $\chi^0 = \{x_i = i\triangle x, \quad i = 0, \ldots, N_x\}$, can be expressed as follows:

$$\frac{dU_i}{dt} + \mathcal{D}_i = 0, \qquad i \in \chi^0. \tag{5.9}$$

When the numerical divergence $\mathcal{D}_i$ represents a spatial discretization of the combined flux $\mathbf{g}_x$, as it is the case for the two schemes that we shall consider here, the multilevel technique in [16] can be applied essentially unchanged. We recall below the essential steps for the sake of completeness.

The goal of the multilevel method is to reduce the CPU time associated to the underlying scheme by reducing the number of expensive flux evaluations. The basic mechanism is easily explained by considering the Forward Euler method applied to (5.9),

$$U_i^{n+1} = U_i^n - \triangle t \mathcal{D}_i^n. \tag{5.10}$$

If both U^n and U^{n+1} are smooth around x_i at time t^n, then (5.10) implies that the numerical divergence is also smooth and we can avoid using the costly numerical flux functions of the HRSC scheme in its computation. On the other hand, if a discontinuity appears during the time evolution, the numerical divergence needs to be computed with the HRSC scheme. So, the smoothness analysis of U^n to U^{n+1} and the computation of $\mathcal{D}$ using this information, are the most important parts of the algorithm.

The computation of the numerical divergence $\mathcal{D}$ on the finest grid is carried out in a sequence of steps. The numerical divergence is evaluated at all the points on the coarsest grid χ^L using the numerical flux function of the scheme, and for the finer grids, the divergence is evaluated recursively, either by the same procedure or with a cheap interpolation using the values obtained on the coarser grids.

The information about the regularity of the data contained in the multiresolution transform of the numerical solution is used to determine a flag vector $(b_i^l)_{l,i}$, whose value (0 or 1) will determine the choice of the procedure to evaluate the divergence. The detail coefficient $(d_i^l)_{l,i}$ computed in the smoothness analysis of the previous section can be interpreted as an interpolation error, hence its size is proportional to the local regularity of U. Given a tolerance parameter ε, the value of the flag vector is obtained by applying two tests to the detail coefficients (or wavelet coefficients):

$$\begin{aligned} \text{if } |d_i^l| &\geq \varepsilon \Rightarrow b_{i-k}^l = 1 \quad k,m = -2,\ldots,2 \\ \text{if } |d_i^l| &\geq 2^r\varepsilon \quad \text{and} \quad l > 0 \Rightarrow b_{2i-k}^{l-1} = 1 \quad k,m = -1,0,1. \end{aligned} \tag{5.11}$$

The determination of the flag vector above follows Harten's recipe [49]. It takes into account that large values of the detail coefficients correspond to non-smooth zones of the solution like shocks or contact discontinuities. In addition, compression regions leading to shock formation, exhibit a lack of regularity that can be estimated. This is also incorporated into the test that determines the flag vector.

The multilevel evaluation of the numerical divergence is carried out as follows: The divergence $\mathcal{D}_i^L$ is computed at the points of the coarsest grid χ^L using the full HRSC scheme. Then assuming that the divergence is known on χ^l, the values of $\mathcal{D}^{l-1}$ on χ^{l-1} are computed as specified below

- If $x_i \in \chi^l$, then $\mathcal{D}_i^{l-1} = \mathcal{D}_{i/2}^l$.
- If $x_i \in \chi^{l-1}\backslash\chi^l$, then $\mathcal{D}_i^{l-1}$ is computed using the boolean flag as follows:

  $$\begin{aligned} &\text{if } b_i^l = 1, \quad \text{compute } \mathcal{D}_i^{l-1} \text{directly with the scheme} \\ &\text{if } b_i^l = 0, \quad \text{compute } \mathcal{D}_i^{l-1} = \mathcal{I}\left[x_i; \mathcal{D}^l\right], \end{aligned}$$

  where $\mathcal{I}[x_i; \mathcal{D}^l]$ is a the polynomial interpolatory technique described in the previous section.

The process is repeated from $l = L,\ldots,1$ and, once it is completed, we obtain the values of $\mathcal{D}$ on the finest grid χ^0, which are needed by the ODE solver.

5.1.3 Quality and Efficiency

The multilevel technique described in the previous chapter is a *cost-effective* technique. Its objective is to compute the numerical divergence on the finest grid in a (hopefully) much cheaper way than the computation using the HRSC scheme at each point of the finest grid. This is the *reference simulation*. The target of the cost-effective multilevel technique is the reference simulation, and the difference between the values of the multilevel simulation and the reference simulation depend on the threshold parameter.

In order to evaluate the quality and efficiency of the algorithm, there are some parameters to be tested. The quality is analyzed by measuring the difference between the multilevel solution U^n and the reference one, U^n_{ref}, in some appropriate norm (we choose the discrete l_1-norm). In [16], the density is chosen as the representative variable for gas-dynamics simulations. In this chapter, we consider the shallow water equations with source terms due to the topography, which are defined in (4.73). For this system, the water height, h, retains all the possible nonsmooth structures of the flow; thus it seems appropriate to compute the l_1-error where h is the representative variable:

$$e_1^h = \frac{\|h^n - h^n_{ref}\|_{l_1}}{\|h^n_{ref}\|_{l_1}}, \tag{5.12}$$

where

$$\|h^n\|_{l_1} = \frac{1}{N_x}\sum_{i=0}^{N_x} |h_i^n|,$$

and $N_x + 1$ is the total number of points on the finest grid χ^0.

The efficiency of the multilevel algorithm with respect to the reference simulation is controlled by two parameters, on one side, the percentage of numerical divergences computed directly per time step, $\%f$, we show in the tables the maximum $\%f_{max}$ and the minimum $\%f_{min}$ of these values in the simulations. This is an important quantity, but a more concrete measure is given by θ_{iter}, the CPU gain for a given iteration, and θ, the gain for the global simulation. Introducing t^{iter}_{ref} and t^{iter}_{mr} as the CPU times at iteration $iter$ for the reference and the multilevel algorithm, respectively, θ_{iter} and θ are defined as

$$\theta_{iter} = \frac{t^{iter}_{ref}}{t^{iter}_{mr}} \quad \text{and} \quad \theta = \frac{\sum t^{iter}_{ref}}{\sum t^{iter}_{mr}}. \tag{5.13}$$

Numerical experiments are presented in the following two sections. In section 5.1.4 we use the TVDB scheme as the underlying HRSC scheme, while the 1J/2J scheme of [10] is used in section 5.1.5.

5.1.4
1D Numerical experiments for the TVDB-multilevel scheme

This section is devoted to the presentation of the results obtained with the multilevel algorithm applied to the flux-limited scheme presented in chapter 4. In this case, the spatial discretization $\mathcal{D}_i$, which represents the spatial discretization of the combined flux $\mathbf{g}_x$, is given as follows:

$$\mathcal{D}_i = \frac{1}{\triangle x}\left(\mathcal{G}_{i+\frac{1}{2}} - \mathcal{G}_{i-\frac{1}{2}}\right) + \frac{1}{\triangle x}\mathcal{S}_i.$$

Here $\mathcal{G}_{i+\frac{1}{2}}$ and $\mathcal{S}_i$ are defined in (4.77), see section 4.6.

The following two tests are meant to check whether the TVDB-multilevel scheme preserves the C-property. As in section 4.6.2, we consider as initial condition a quiescent state. The first one with smooth topography given by (4.83) and the second one with a complex geometry defined in [45].

The numerical results obtained with the TVDB-multilevel scheme with $\varepsilon = 10^{-2}$ and $L = 3$ are shown in figure 5.2. As in all test cases, we apply the multiresolution transform required by the smoothness analysis on the height h. No differences are observed with respect to 4.16, obtained directly with the TDVB scheme.

In table 5.1 and 5.2, respectively, we can see that the l_1-error is of the order of the roundoff error, which means that the scheme is exact for this test case. We also show the CPU gain and the measurements of the percentage of numerical divergences computed directly per time step (in these cases, the minimum ($\%f_{min}$) and maximum ($\%f_{max}$) percentage coincide).

We carry out next the experiments of section 4.6.2 involving subcritical flow, transcritical without shock or with shock. The results are shown in figure 5.3. Again, no noticeable differences can be appreciated with respect to the reference simulation. In table 5.3, we show the relevant parameters for quality and efficiency for the case of subcritical flow, the other results are similar.

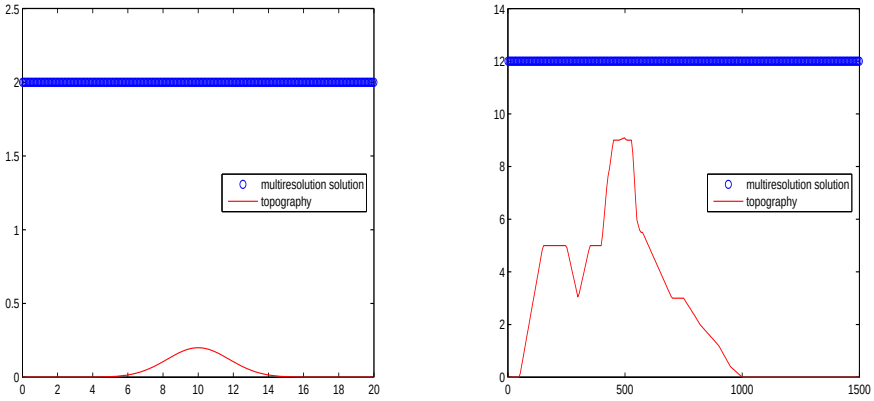
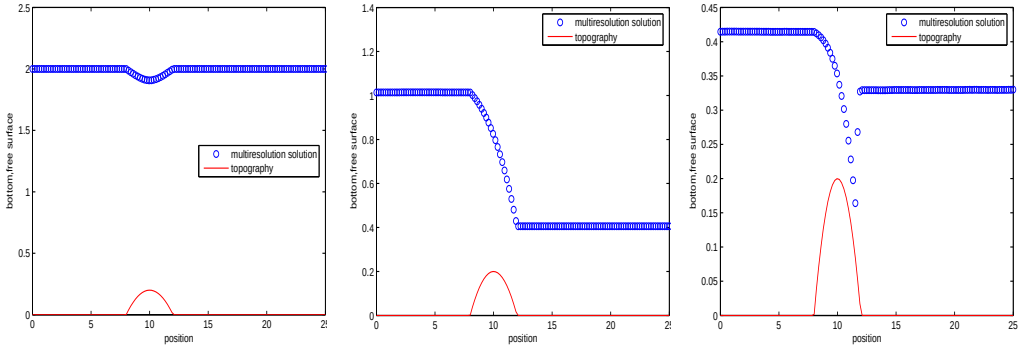


Figure 5.2: *Flow at rest using the multiscale flux-limited scheme $N_x = 256$, $L = 3$ levels and tolerance of $\varepsilon = 10^{-2}$.Left: Smooth topography (t=50s.). Right: Complex topography (t=200s.).*



(a) (b) (c)

Figure 5.3: *Convergence towards a steady state using the multiscale flux-limited scheme (128 nodes, t=200s). Using a tolerance parameter of $\varepsilon = 10^{-3}$ and $L = 3$ levels. (a) Subcritical flow. (b) Transcritical flow without shock. (c) Transcritical flow with shock.*

Grid size χ^0	$\%f$	CPU gain θ	l_1-error
128	39.5349	2.4100	$5.5087 \cdot 10^{-15}$
256	33.0739	2.9049	$1.0195 \cdot 10^{-11}$
512	32.1637	3.2286	$1.2860 \cdot 10^{-11}$

Table 5.1: *Quiescent state with smooth topography using the multiscale flux-limited scheme, $L = 3$ levels and tolerance of $\varepsilon = 10^{-2}$. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

Grid size χ^0	$\%f$	CPU gain θ	l_1-error
128	69.7674	1.3881	$5.1922 \cdot 10^{-17}$
256	52.9183	1.9069	$1.1129 \cdot 10^{-16}$
512	41.3255	2.7007	$4.0840 \cdot 10^{-15}$

Table 5.2: *Quiescent state with complex topography using the multiscale flux-limited scheme, $L = 3$ levels and tolerance of $\varepsilon = 10^{-2}$. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

Finally, we show the numerical solution for the quasi stationary test proposed by LeVeque in [71], see section 4.6.2. In this case, a tolerance parameter of $\varepsilon = 10^{-4}$ and $L = 3$ levels of multiresolution are used to obtain figure 5.4. As in the two previous cases, the parameters to evaluate the quality and efficiency of the algorithm are tested and showed in table 5.4.

5.1.5 1D Numerical experiments for the 1J-2J multilevel scheme

The numerical technique we consider in this section follows [10]. We first present the main steps of the scheme and then the numerical experiments.

The multiscale 1J-2J scheme

Let us consider the 1D system (4.73):

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{s}(x, \mathbf{u}) \tag{5.14}$$

Grid size χ^0	$\%f_{min}$	-	$\%f_{max}$	CPU gain θ	l_1-error
128	35.6589	-	89.9225	2.0277	$4.6968 \cdot 10^{-5}$
256	26.0700	-	82.4903	2.8718	$6.9315 \cdot 10^{-5}$
512	19.1033	-	84.0156	4.3757	$9.1074 \cdot 10^{-5}$

Table 5.3: *Subcritical flow over a hump using the multiscale flux-limited scheme, a tolerance parameter of $\varepsilon = 10^{-3}$ and $L = 3$ levels. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

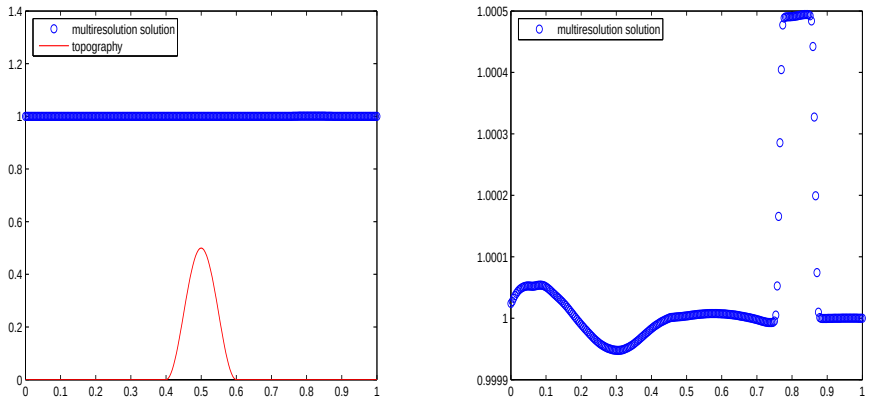


Figure 5.4: *Quasi stationary flow using multiscale flux-limited scheme, t=0.7s. With a tolerance parameter of $\varepsilon = 10^{-4}$ and $L = 3$ levels*

and rewrite in the form $\mathbf{u}_t + \mathbf{g}(x,t)_x = 0$, with $\mathbf{g}(x,t) = \mathbf{f}(\mathbf{u}) + \mathbf{b}(x,t)$ and $\mathbf{b}(x,t) = (0, \int_{\bar{x}}^{x} ghz_x ds)^T$. In [10], the authors arrive at a semi-discrete formulation of the type

$$U_t + \frac{G^+_{i+\frac{1}{2}} - G^-_{i-\frac{1}{2}}}{\triangle x} = 0. \tag{5.15}$$

The fully discrete technique uses method of lines approach in which the time integration is performed via a TVD-Runge-Kutta method (see [88]).

The multilevel procedure described in section 5.1.2 can be directly applied to

$$\mathcal{D}_i = \frac{G^+_{i+\frac{1}{2}} - G^-_{i-\frac{1}{2}}}{\triangle x}.$$

The computation of $G^{\pm}_{i+\frac{1}{2}}$ only involves integral terms over consecutive

Grid size χ^0	$\%f_{min}$	-	$\%f_{max}$	CPU gain θ	l_1-error
128	49.6124	-	66.6667	1.5556	$5.0819 \cdot 10^{-6}$
256	42.0233	-	59.5331	1.9000	$2.4874 \cdot 10^{-6}$
512	35.2827	-	52.2417	2.6709	$5.8699 \cdot 10^{-7}$

Table 5.4: *Quasi stationary flow using the multiscale flux-limited scheme, a tolerance parameter of $\varepsilon = 10^{-4}$ and $L = 3$ levels. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

cell centers and follows the basic design strategy in Marquina's flux formula: two states are computed at each side of a cell-interface, U^L and U^R, and the numerical flux functions are obtained by applying the scalar algorithm to "sided" local characteristic fluxes. The states U^L and U^R at each side of a given interface are obtained by ENO interpolation of the physical variables as specified in [31]. Unless specifically stated, the order of the interpolation used to compute these states is the same as the order of the scheme. Given $U^L = U^L_{i+\frac{1}{2}}$ and $U^R = U^R_{i+\frac{1}{2}}$, the left and right states at the $i+\frac{1}{2}$ cell-interface, the flux functions $G^{\pm}_{i+\frac{1}{2}}$ shall be defined as

$$G^{\pm}_{i+\frac{1}{2}} = \sum_{p=1}^{2} (\tilde{G}^{\pm}_{i+\frac{1}{2}})^{p,L} R^p(U^L) + (\tilde{G}^{\pm}_{i+\frac{1}{2}})^{p,R} R^p(U^R) \tag{5.16}$$

where $L^p(U^L), R^p(U^L)(L^p(U^R), R^p(U^R)), p = 1, 2$, are the left and right eigenvectors of the Jacobian matrix $J(U) = F'(U)$, associated to the eigenvalues $\lambda^p(U^L)(\lambda^p(U^R))$. $(\tilde{G}^{\pm,p}_{i+\frac{1}{2}})^{L,R}$ are the local modified characteristic fluxes, whose high order terms involve only quantities of the form

$$B_{i,i+1} = \mathbf{b}_{i+1} - \mathbf{b}_i = (0, \int_{x_i}^{x_{i+1}} g h z_x ds)^T \tag{5.17}$$

and the contribution of the source terms at first order depends only of the wind coming from the right (+) or left (–) at the interface (more details in [10]). For example, for a scheme of order r ($r = 1, 2$ or 3), we give next a precise description of the computation of these numerical fluxes for first order $2J$ numerical flux function, the index j below runs from $j = i - r, \ldots, i + r$.

For p=1,2

- If $\lambda^p(U^L_{j+\frac{1}{2}}) > 0$ and $\lambda^p(U^R_{j+\frac{1}{2}}) > 0$ then the wind come from the left and

$$(\tilde{G}^{p,\pm}_{j+\frac{1}{2}})^R = 0$$
$$(\tilde{G}^{p,+}_{j+\frac{1}{2}})^L = L^p(U^L)\cdot F_j + HOT^L_{j+\frac{1}{2}}$$
$$(\tilde{G}^{p,-}_{j+\frac{1}{2}})^L = L^p(U^L)\cdot (F_j - B_{j,j+1}) + HOT^L_{j+\frac{1}{2}}$$

- If $\lambda^p(U^L_{j+\frac{1}{2}}) < 0$ and $\lambda^p(U^R_{j+\frac{1}{2}}) < 0$ then the wind come from the right and
$$(\tilde{G}^{p,\pm}_{j+\frac{1}{2}})^L = 0$$
$$(\tilde{G}^{p,+}_{j+\frac{1}{2}})^R = L^p(U^R)\cdot (F_{j+1} + B_{j,j+1}) + HOT^R_{j+\frac{1}{2}}$$
$$(\tilde{G}^{p,-}_{j+\frac{1}{2}})^R = L^p(U^R)\cdot F_{j+1} + HOT^R_{j+\frac{1}{2}}$$

- If $\lambda^p(U^L_{j+\frac{1}{2}})\lambda^p(U^R_{j+\frac{1}{2}}) < 0$ then mixed wind (sonic point nearby), we define α as $\max(|\lambda^p(U^L_{j+\frac{1}{2}})|, |\lambda^p(U^R_{j+\frac{1}{2}})|)$, then
$$(\tilde{G}^{p,+}_{j+\frac{1}{2}})^L = \tfrac{1}{2}L^p(U^L)\cdot (F_j + \alpha U_j) + HOT^L_{j+\frac{1}{2}}$$
$$(\tilde{G}^{p,-}_{j+\frac{1}{2}})^L = \tfrac{1}{2}L^p(U^L)\cdot (F_j + \alpha U_j - B_{j,j+1}) + HOT^L_{j+\frac{1}{2}}$$
$$(\tilde{G}^{p,+}_{j+\frac{1}{2}})^R = \tfrac{1}{2}L^p(U^R)\cdot (F_{j+1} - \alpha U_{j+1} + B_{j,j+1}) + HOT^R_{j+\frac{1}{2}}$$
$$(\tilde{G}^{p,-}_{j+\frac{1}{2}})^R = \tfrac{1}{2}L^p(U^R)\cdot (F_{j+1} - \alpha U_{j+1}) + HOT^R_{j+\frac{1}{2}}$$

where $HOT_{j+\frac{1}{2}}$ are the higher order terms obtained from the ENO construction. The extension just described complies with the basic design principles of Donat and Marquina's flux formula [26], the superscript L refers to characteristic information carried by a left-wind, while R refers to right-wind driven information.

Moreover, in the fully discrete scheme, the integral in the second component of (5.17), is substituted by the following discrete expression. Assuming that the topography and the flow are smooth and applying the trapezoidal rule, we obtain

$$\int_{x_i}^{x_{i+1}} ghz_x = \frac{g}{2}\left(z_{i+1} - z_i\right)\left(h_{i+1} + h_i\right) + \mathcal{O}(\triangle x^3). \tag{5.18}$$

Finally, it is proven in [10] that if $U^L_{i+\frac{1}{2}} = U^R_{i+\frac{1}{2}}$ (e.g. $= \frac{U_i+U_{i+1}}{2}$) (**1J**), the scheme verifies the exact C-property, and if $U^L_{i+\frac{1}{2}} \neq U^R_{i+\frac{1}{2}}$ (**2J**), the scheme verifies the approximate C-property, provided the order of accuracy is at least 2. Hence, the preferred option is to combine both, the **1J-2J**

scheme, to get the benefits of both alternatives. This is the scheme of our choice in the next section.

1D Numerical experiments

We shall carry out the same set of experiments as in the previous section, and use the second order scheme.

The exact C-property is preserved for the 1J-2J shock capturing scheme presented by [10], and also for the multilevel version. As it is seen in table 5.5 for a steady flow with smooth topography and in 5.6 for a steady flow with complex bottom, the l_1-error is of the order of the roundoff error. In figure 5.5 and 5.6 are shown the numerical solutions for the reference scheme (left) and using the multiresolution method (right) for smooth and complex bottom with 256 grid points, respectively. Furthermore, we show the measurements of the percentage of numerical divergences computed directly per time step and the CPU gain, by using a tolerance parameter of $\varepsilon = 10^{-2}$ and $L = 3$ levels.

Grid size χ^0	$\%f_{min}$	-	$\%f_{max}$	CPU gain θ	l_1-error
128	39.5349	-	39.5349	2.5011	$7.5715 \cdot 10^{-15}$
256	33.0739	-	33.0739	3.1124	$3.0275 \cdot 10^{-13}$
512	32.1637	-	32.1637	3.0144	$6.0113 \cdot 10^{-13}$

Table 5.5: *Quiescent state with smooth topography using flux-limited multiscale scheme, with a tolerance parameter $\varepsilon = 10^{-2}$ and $L = 3$ levels. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

Grid size χ^0	$\%f$	CPU gain θ	l_1-error
128	69.7674	1.6837	$7.5709 \cdot 10^{-15}$
256	52.9183	2.0043	$6.2071 \cdot 10^{-14}$
512	41.3255	2.4069	$2.2738 \cdot 10^{-13}$

Table 5.6: *Quiescent state with complex topography using flux-limited multiscale scheme, with a tolerance parameter $\varepsilon = 10^{-2}$ and $L = 3$ levels. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

As in the previous section, in figure 5.7, we show the ability of scheme to converge towards a steady state. As in section 4.6.2, it depends on the initial and boundary conditions we are going to obtain a subcritical flow,

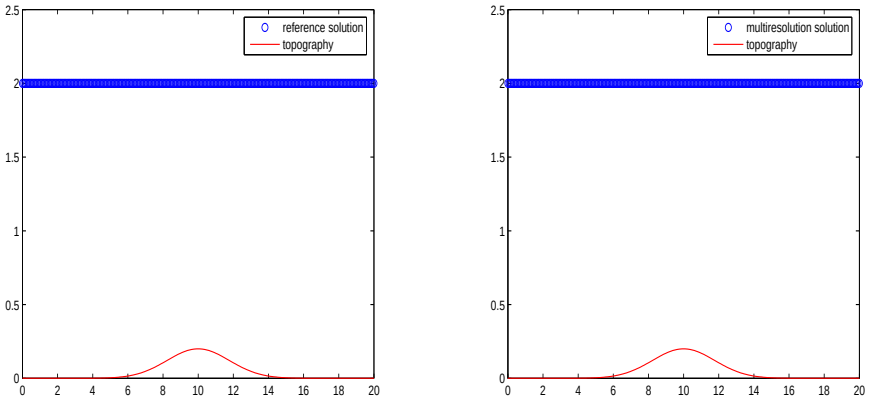


Figure 5.5: *Flow at rest using the 1J-2J scheme $N_x = 256$ with smooth topography, with $L = 3$ levels of multiresolution and a tolerance of 10^{-2}, $t = 50$ s. Left: Reference solution. Right: Multiresolution solution.*

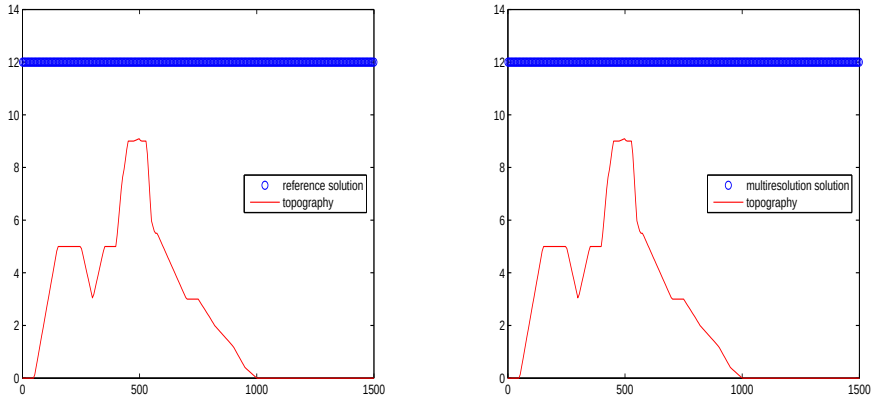
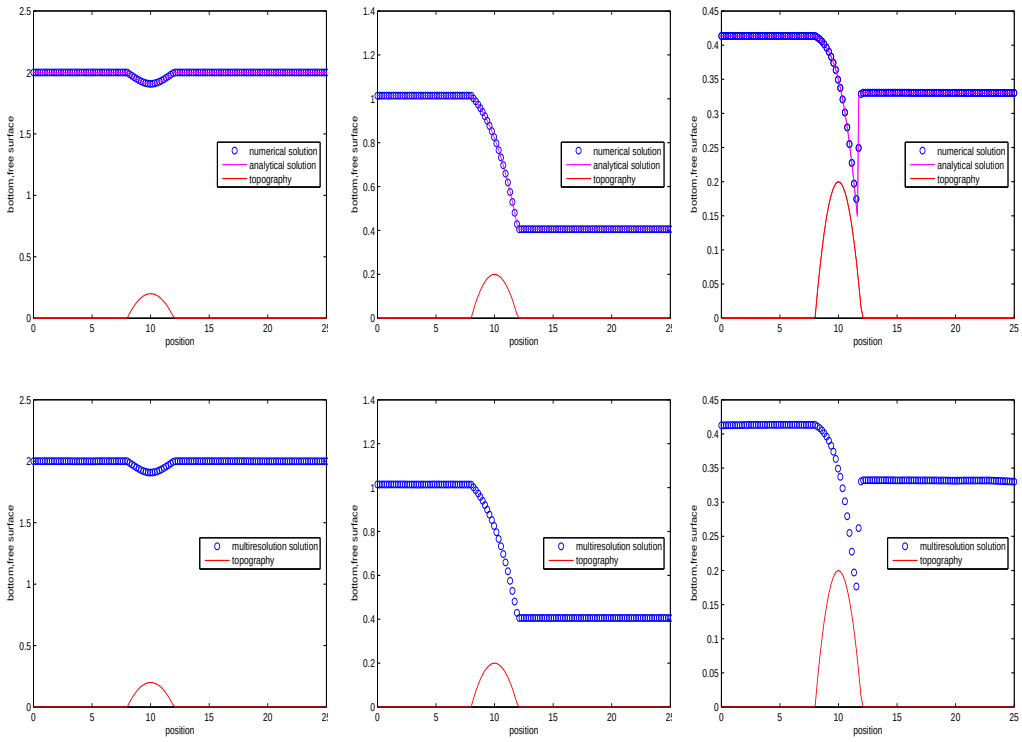


Figure 5.6: *Flow at rest using the 1J-2J scheme $N_x = 256$ with complex topography, with $L = 3$ levels of multiresolution and a tolerance of $\varepsilon = 10^{-2}$, $t = 200$ s. Left: Reference solution. Right: Multiresolution solution.*

transcritical without shock or with shock. In table 5.7, we show the results for the case of subcritical flow, the other results are similar.



(a) (b) (c)

Figure 5.7: *Convergence towards a steady state using the 1J-2J scheme (128 nodes, t=200s). Using a tolerance parameter of* $\varepsilon = 10^{-3}$ *and* $L = 3$ *levels. (a) Subcritical flow. (b) Transcritical flow without shock. (c) Transcritical flow with shock.*

Finally, we show the numerical solution for the quasi stationary test proposed by LeVeque in [71]. The parameters to evaluate the quality and efficiency of the algorithm are shown in table 5.8. In this case, a tolerance parameter of $\varepsilon = 10^{-4}$ and $L = 3$ level of multiresolution is used to obtain figure 5.8 with 512 grid points.

Grid size χ^0	$\%f_{min}$	-	$\%f_{max}$	CPU gain θ	l_1-error
128	35.6589	-	93.7984	2.0277	$4.6968 \cdot 10^{-5}$
256	26.0700	-	89.4942	2.8718	$6.9315 \cdot 10^{-5}$
512	19.1033	-	74.6589	4.3757	$9.1074 \cdot 10^{-5}$

Table 5.7: *Subcritical flow over a hump using the multiscale 1J-2J scheme, a tolerance parameter of $\varepsilon = 10^{-3}$ and $L = 3$ levels. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

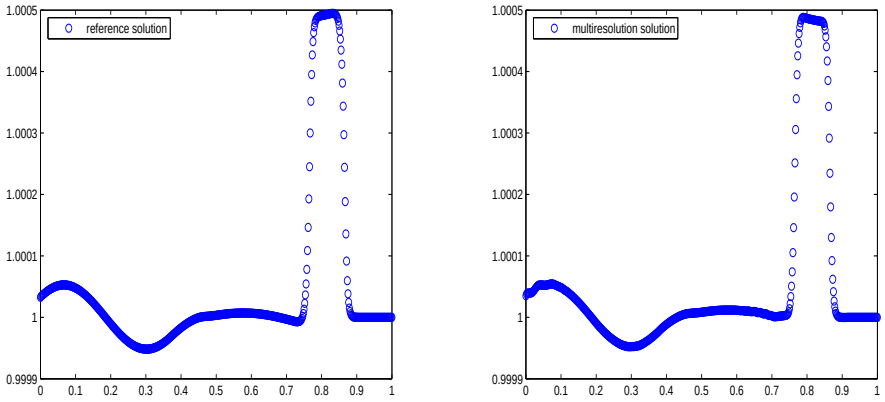


Figure 5.8: *Quasi stationary flow using the 1J-2J scheme $N_x = 512$ with smooth topography, with $L = 3$ levels of multiresolution and a tolerance of $\varepsilon = 10^{-4}$ at time $t = 0.7$. Left: Reference solution. Right: Multiresolution solution.*

5.2 The 2D multilevel algorithm

As observed in [6], the original idea of a multilevel computation of the numerical flux function (in one dimension) described by Harten in [49] cannot be used in a robust and general manner in two dimensions. The key point is then to observe that it is the numerical divergence the quantity that should be adapted to the multilevel computations. This is the approach that we have adopted already in our 1D multilevel schemes.

As in [16], the finite difference schemes considered here are extended to 2D in a dimension by dimension fashion [87], hence the changes between the 1D multilevel algorithm used in the previous sections and the

Grid size χ^0	$\%f_{min}$	-	$\%f_{max}$	CPU gain θ	l_1-error
128	51.1628	-	67.4419	1.7368	$5.0821 \cdot 10^{-6}$
256	37.7432	-	61.4786	2.2030	$2.5509 \cdot 10^{-6}$
512	34.1131	-	52.0468	2.3164	$1.17292 \cdot 10^{-6}$

Table 5.8: *Quasi stationary flow using the multiscale 1J-2J scheme, $L = 3$ levels of multiresolution and a tolerance of $\varepsilon = 10^{-4}$. l_1-error computed by (5.12), percentage of divergence computed and the CPU gain.*

2D multilevel algorithm are minimal. We outline below the relevant differences.

5.2.1 Smoothness analysis

As in the 1D case, the most important step concern the smoothness analysis of the data, and how this information is used. The different resolution levels are specified now by a set of nested grids $\{\chi^l, \quad l = 1, \ldots, L\}$ given as follows,

$$(x_i, y_j) \in \chi^l \iff (x_{2^l i}, y_{2^l j}) \in \chi^0, \tag{5.19}$$

where χ^0 is considered the finest grid.

The predicted values are computed as in (5.4), where, now, $\mathcal{I}\left[(x, y); v^l\right]$ denotes a 2D polynomial interpolatory technique of rth order on the $l-$th grid. The interpolated values will have different expressions for the different locations, explicit details can be found in [6]. We summarize the results below:

(odd,even): $(x_{2i-1}, y_{2j}) \in \chi^{l-1} \backslash \chi^l$, interpolation along i,

$$\tilde{v}^{l-1}_{2i-1\ 2j} = \mathcal{I}_1\left[(x_{2i-1}, y_{2j}); v^l\right] = \sum_{k=1}^{s} \beta_k \left(v^l_{i+k-1\ j} + v^l_{i-k\ j}\right), \tag{5.20}$$

(even,odd): $(x_{2i}, y_{2j-1}) \in \chi^{l-1} \backslash \chi^l$, interpolation along j,

$$\tilde{v}^{l-1}_{2i\ 2j-1} = \mathcal{I}_2\left[(x_{2i}, y_{2j-1}); v^l\right] = \sum_{m=1}^{s} \beta_m \left(v^l_{i\ j+m-1} + v^l_{i\ j-m}\right), \tag{5.21}$$

(odd,odd): $(x_{2i-1}, y_{2j-1}) \in \chi^{l-1} \backslash \chi^l$, interpolation along i and j,

$$\begin{aligned} \tilde{v}^{l-1}_{2i-1\ 2j-1} &= \mathcal{I}_3\left[(x_{2i-1}, y_{2j-1}); v^l\right] \\ &= \sum_{k=1}^{s} \beta_k \sum_{m=1}^{s} \beta_m (v^l_{i+k-1\ j+m-1} + v^l_{i-k\ j+m-1} \\ &\quad + v^l_{i+k-1\ j-m} + v^l_{i-k\ j-m}). \end{aligned} \tag{5.22}$$

In this work, we use $r = 4$ ($r = 2s$). Then, as in the 1D case the coefficients $\beta_1 = \frac{9}{16}$, $\beta_2 = \frac{-1}{16}$ are considered. Notice that the interpolatory property means that

$$\tilde{v}^{l-1}_{2i\ 2j} = v^l_{i\ j} \tag{5.23}$$

for all $i = 0, \ldots, N_x/2^{l-1}, \quad j = 0, \ldots, N_y/2^{l-1}$, where $i_0 = 0, \ldots, N_x/2^l, \quad j_0 = 0, \ldots, N_y/2^l$.

The wavelet coefficients are computed as in the 1D case, as the difference between v^l and $\tilde{v}^l$ for $l = 1, \ldots, L$.

5.2.2 General framework

Let us consider the 2D shallow water system

$$\begin{pmatrix} h \\ q_1 \\ q_2 \end{pmatrix} + \begin{pmatrix} q_1 \\ \frac{q_1^2}{h} + \frac{g}{2}h^2 \\ \frac{q_1 q_2}{h} \end{pmatrix}_x + \begin{pmatrix} q_2 \\ \frac{q_1 q_2}{h} \\ \frac{q_2^2}{h} + \frac{g}{2}h^2 \end{pmatrix}_y = \begin{pmatrix} 0 \\ -ghz_x \\ -ghz_y \end{pmatrix}$$

and rewrite in the form $\mathbf{u}_t + \mathbf{g}_1(x,t)_x + \mathbf{g}_2(y,t)_y = 0$, with

$$\mathbf{g}_1 = \begin{pmatrix} q_1 \\ \frac{q_1^2}{h} + \frac{g}{2}h^2 \\ \frac{q_1 q_2}{h} \end{pmatrix}_x + \begin{pmatrix} 0 \\ \int_{\bar{x}}^{x} ghz_x \\ 0 \end{pmatrix} \text{ and } \mathbf{g}_2 = \begin{pmatrix} q_2 \\ \frac{q_1 q_2}{h} \\ \frac{q_2^2}{h} + \frac{g}{2}h^2 \end{pmatrix}_y + \begin{pmatrix} 0 \\ 0 \\ \int_{\bar{y}}^{y} ghz_y \end{pmatrix}.$$

We consider discretizations of this system on a Cartesian grid $\chi^0 = \{(x_i = i\triangle x, y_j = j\triangle y), \quad i = 0, \ldots, N_x \quad j = 0, \ldots, N_y\}$ using the semi-discrete formulation:

$$\frac{dU_{ij}}{dt} + \mathcal{D}_{ij} = 0, \tag{5.24}$$

where $\mathcal{D}_{ij}$ represents a spatial discretization of the combined flux in a *dimension by dimension* fashion. Taking into account the computation of

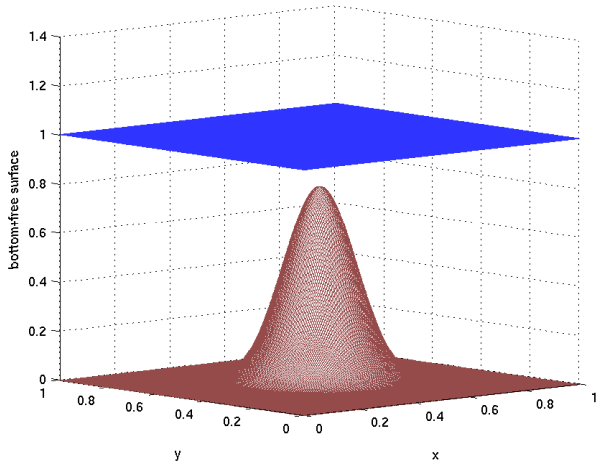


Figure 5.9: *Exact water surface and topography at steady state.*

the detail coefficients described in the previous section, the 2D multilevel technique follows the same steps as in the 1D case described before. We refer the reader to [16] for specific details.

5.2.3 2D Numerical experiments

In this section we shall present some numerical experiments for the 1J-2J multilevel scheme.

The first test pursues a numerical validation of the C-property for the 2D multilevel algorithm. We follow [109], and consider a smooth topography given by

$$z(x,y) = 0.8e^{-50((x-0.5)^2+(y-0.5)^2)} \tag{5.25}$$

with $(x,y) \in [0,1] \times [0,1]$. In a quiescent state ($q_1 = q_2 = 0$), the 2D multilevel scheme maintain the steady flow (see figure 5.9). In table 5.9, we show the measurements of the l_1-error, which for 2D equations is defined as

$$e_1^h = \frac{\|h_{ij}^n - h_{ref_{ij}}^n\|}{\|h_{ref}^n\|_{l_1}}, \tag{5.26}$$

where

$$\|h^n\|_{l_1} = \frac{1}{N_p}\sum_{i=0}^{N_x}\sum_{j=0}^{N_y}|h_{ij}^n|,$$

$N_p = (N_x+1)\times(N_y+1)$ the number of points, and also It is also shown the percentage of numerical divergences computed directly per time step with multiresolution. The l_1-error of the reference solution is of the order of the roundoff error. The C-property is, thus, preserved. In this case, $\%f_{min} = \%f_{max}$ as in the 1D case.

Grid size χ^0	$\%f_{min}$	-	$\%f_{max}$	CPU gain θ	l_1-error
128×128	8.6593	-	8.6593	5.4970	$1.215\cdot10^{-15}$
256×256	6.9539	-	6.9539	6.4846	$2.627\cdot10^{-15}$
512×512	6.3279	-	6.3279	6.9869	$7.690\cdot10^{-16}$

Table 5.9: *2D Steady state with smooth topography with $L=3$ levels, and a tolerance of $\varepsilon = 10^{-2}$. l_1-error computed by (5.26) and percentage of divergence computed*

Next, we consider a test containing shocks and rarefaction waves (LeVeque 2D test [109]). The bottom topography is given as

$$z(x,y) = 0.5e^{-50((x-0.5)^2+(y-0.5)^2)} \tag{5.27}$$

on $[0,1]\times[0,1]$ with $g=1$. The initial conditions are $q_1 = q_2 = 0$ and

$$h(x,y) = \begin{cases} 1-z(x,y)+\epsilon, & 0.1<x<0.2; \\ 1-z(x,y), & \text{otherwise} \end{cases} \tag{5.28}$$

where $\epsilon = 10^{-2}$ is a small perturbation of the data.

In figure 5.10, we display the level curves of the numerical solution obtained with and without the multilevel algorithm, and we can observe that the numerical simulation is of the same "quality" as the reference simulation. We also present two plots displaying only the points of χ^0 where the numerical divergence is computed directly with the 1J-2J scheme.

In figure 5.11 we show the l_1-error measured for variable $h(x,y)$ when applying the multilevel algorithm with $L=3$ levels of refinement, for different values of the tolerance ε, and different grid mesh. We can observe that, as in [16] the closeness to the reference simulation, can be controlled by adjusting the tolerance suitably. Furthermore, in table 5.10,

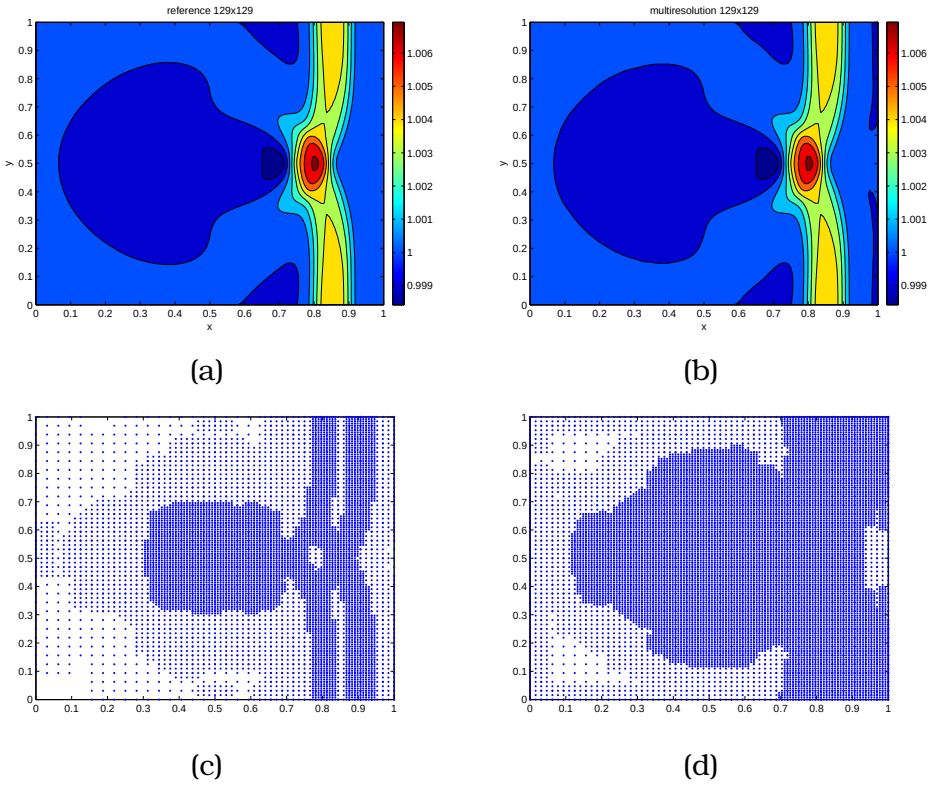
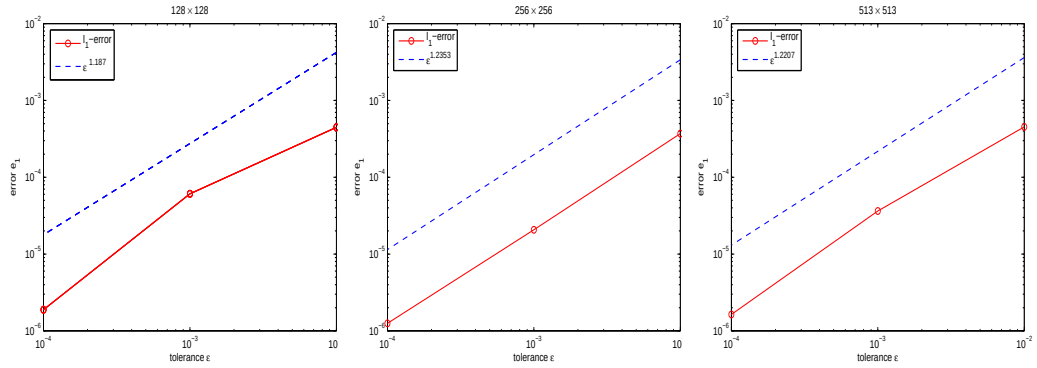


Figure 5.10: *LeVeque 2D test at time $t = 0.7$. (a) reference simulation. (b) multilevel simulation with $\varepsilon = 10^{-4}$. (c) Points of χ^0 where the numerical divergence is computed for $\varepsilon = 10^{-3}$. (d) For $\varepsilon = 10^{-4}$.*



128×128 256×256 512×512

Figure 5.11: *LeVeque 2D test at time $t = 0.7$. Error between the multilevel algorithm and the reference one for different values of ε, $L = 3$ levels.*

we show the l_1-error computed by (5.26) for a mesh of 512×512, with $L = 3$ levels of multiresolution, and the parameter p such that

$$e_1^h \leq C\varepsilon^p.$$

In all cases, $p \geq 1$ as well as Chiavassa and Donat observed in [16].

ε	l_1-error	p
10^{-2}	$4.514 \cdot 10^{-4}$	
10^{-3}	$3.630 \cdot 10^{-5}$	1.0951
10^{-4}	$1.600 \cdot 10^{-6}$	1.3464

Table 5.10: *ε versus l_1-error with $L = 3$ levels and a mesh of 512×512.*

Finally, we show in Table 5.11 the global gain for each simulation and the maximum and the minimum values for $\%f$ in the simulation. We can observe that the finer the grid, the smaller the percentage of direct flux evaluations. In Figure 5.12 we represent $\theta(t)$ and $\%f(t)$. We can observe that, there are few non-smooth structures in the flow, the gain is quite large for fine grids. The behavior of $\theta(t)$ is roughly inversely proportional to that of $\%f(t)$.

To end this section, we display in figure 5.13 the level curves of the numerical solution obtained with and without the multilevel algorithm (figure 5.13 middle and top, respectively), of two different numerical test. The first one (figure 5.13-(a)) is the same 2D test as before, where now

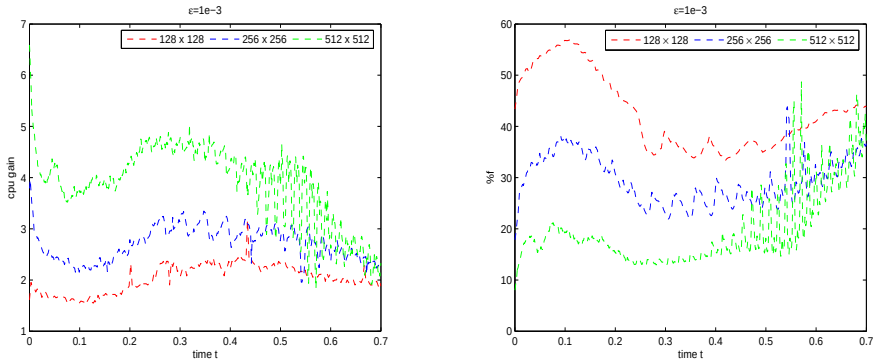


Figure 5.12: *LeVeque 2D test at time $t = 0.7$, $L = 3$ levels and $\varepsilon = 10^{-4}$. Time evolution of θ and $\%f$ for different initial grid.Left: CPU gain. Right: percentage flux*

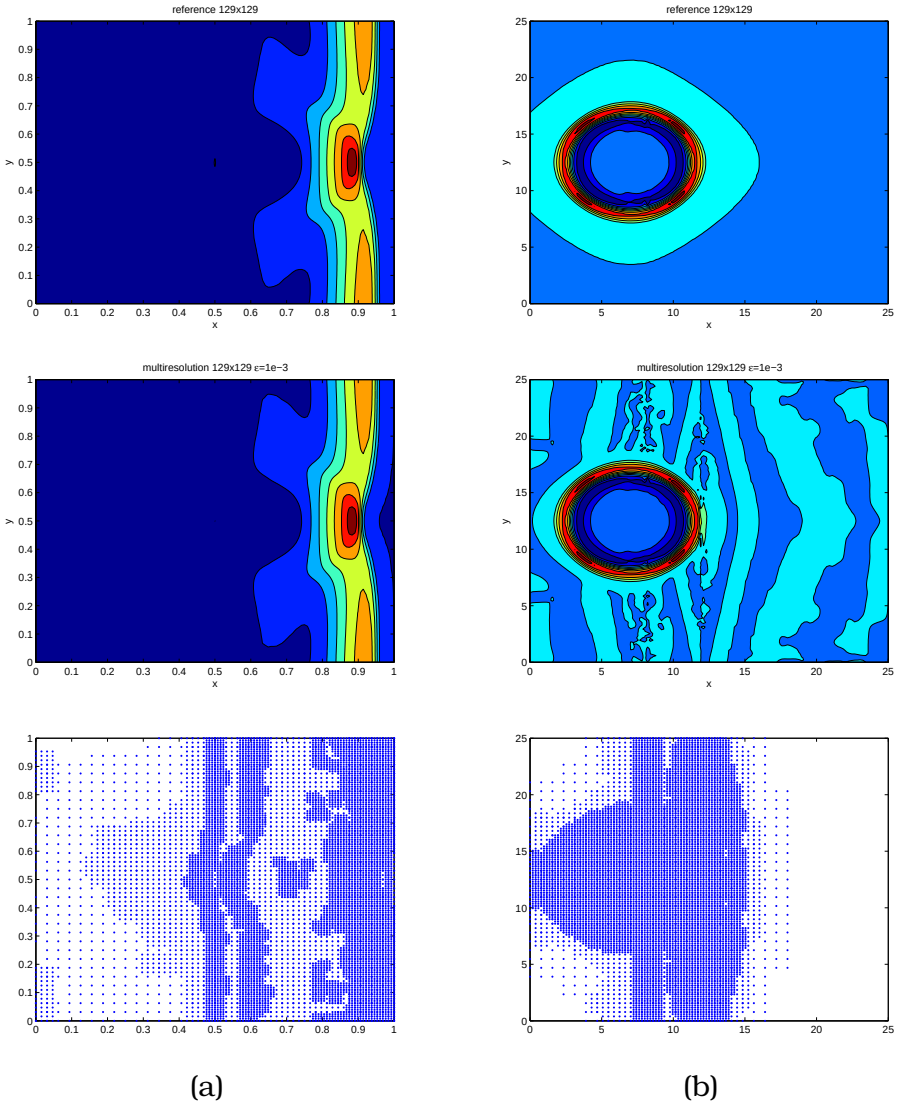
Grid size χ^0	$\%f_{min}$	-	$\%f_{max}$	CPU gain θ	l_1-error
128×128	54.5500	-	74.2804	1.4039	$1.900 \cdot 10^{-6}$
256×256	32.1872	-	75.7281	1.6431	$1.300 \cdot 10^{-6}$
513×513	11.5879	-	79.3850	2.2294	$1.600 \cdot 10^{-6}$

Table 5.11: *LeVeque 2D test at time $t = 0.7$, $L = 3$ levels and $\varepsilon = 10^{-4}$. Percentage of resolved flux and CPU gain*

ϵ in (5.28) is equals to 0.2. In the last test (figure 5.13-(b)), we solve the system in the rectangular domain $[0, 25] \times [0, 25]$. The bottom topography is given by

$$b(x, y) = \begin{cases} 0.2 - 0.05(x - 10)^2, & \text{if } 8 \leq x \leq 12; \\ 0, & \text{otherwise.} \end{cases}$$

These data correspond precisely to the one-dimensional subcritical steady state, and the cross section of the unperturbed solution can be seen on figure 5.3-(a). Our initial condition is given by a two-dimensional small perturbation of $h + z = 2$, which is perturbed upward by 0.05 in the box $6.5 \leq x \leq 7.5$, $12 \leq y \leq 13$. For each simulation, we also present a third plot displaying only the points of χ^0 where the numerical divergence is computed directly with the 1J-2J scheme (figure 5.13-bottom).



(a) (b)

Figure 5.13: *2D tests. Top: Reference solution. Middle: Multiresolution solution* $L = 3$ *and* $\varepsilon = 10^{-3}$*. Bottom: Numerical fluxes computed with the multiresolution. (a) Stationary state with a perturbation of* 0.2*. (b) 2D subcritical flow with a small 2D perturbation of* 0.05*.*

5.3 Conclusions

We presented numerical results concerning the application of the multilevel method proposed in [16] to non-homogeneous systems of conservation laws. The inclusion of the source term in the numerical divergence, allows us to construct easily the multilevel technique for balance laws.

Two multilevel schemes are applied to the 1D shallow water system. The underlying HRSC schemes have numerical flux functions that are significantly different in terms of computational effort. The numerical results obtained with the multilevel schemes are analyzed in terms of quality and efficiency. The two HRSC schemes considered satisfy the exact C-property for quiescent flow, and we have observed that this property remains valid for the corresponding multilevel strategy.

The computations presented in this chapter point out that there is a significant reduction of the computational time when using the multilevel algorithm. The more expensive the flux computation, the better the efficiency of the multilevel scheme with respect to the reference simulation, as was observed by Sjögreen in [89].

We carried out some numerical tests with the 1J-2J multilevel scheme in 2D. These preliminary computations are very encouraging and further experimentation with the multilevel strategy in more realistic situation is the subject of ongoing work.

The results reported here, show that the multilevel strategy leads to an efficient tool for the numerical simulation of systems of balance laws, specially in those situations where we need high quality and high resolution at an affordable cost.

A

Relevant results in Runge-Kutta methods for Ordinary Differential Equations

Let us consider an initial value problem for a system of ordinary differential equations (ODEs) of the form

$$\frac{d}{dt}U(t) = F(U(t)), \qquad t \geq t_0, \;\; U(t_0) = U_0. \tag{A.1}$$

We assume that $t_0 \in \mathbb{R}$, $U_0 \in \mathbb{R}^m$, $F : \mathbb{R}^m \to \mathbb{R}^m$ such that the problem

has unique solution $U(t) = U(t; t_0, u_0)$, for each t_0 and U_0.

A.1 Representations of Runge-Kutta methods

A.1.1 The standard form. Compact representations

An s-stage Runge-Kutta (RK) method is defined by an $s \times s$ real matrix $\mathcal{A} = (a_{i,j})$ and a vector $b \in \mathbb{R}^s$ as follows: From U^n, the numerical approximation of the solution at time t_n, we obtain U^{n+1}, the numerical approximation at time $t_{n+1} = t_n + \Delta t$ from

$$U^{n+1} = U^n + \triangle t \sum_{i=1}^{s} b_i F(U^{(i)}). \tag{A.2}$$

where the internal stages, $U^{(i)}$, $i = 1, \ldots, s$, are computed from

$$U^{(i)} = U^n + \triangle t \sum_{j=1}^{s} a_{ij} F(U^{(j)}), \quad 1 \leq i \leq s. \tag{A.3}$$

We will often refer to the method by its coefficient scheme $(\mathcal{A}, b)$. If the matrix A is strictly lower triangular, the method is explicit; otherwise the method is implicit.

The RK method above is also often represented by a so-called *Butcher tableau* as follows:

$$\begin{array}{c|c} c & A \\ \hline & b^t \end{array} \tag{A.4}$$

where $c_i = \sum_{j=1}^{s} a_{ij}$. It is well known that each internal stage $U^{(i)}$ approximates $U(t_n + c_i h)$.

Following [55], [56], we shall represent RK schemes such as (A.2)-(A.3) in *compact form* by considering the $(s+1) \times (s+1)$ matrix $\mathbb{A}$, defined as follows

$$\mathbb{A} = \begin{pmatrix} A & 0 \\ b^t & 0 \end{pmatrix}.$$

It is easy to see that (A.2)-(A.3) can be expressed also as

$$\mathcal{U} = e \otimes U^n + \triangle t(\mathbb{A} \otimes I)\mathcal{F}(\mathcal{U}), \tag{A.5}$$

where $e = (1, \ldots, 1) \in \mathbb{R}^{s+1}$, $\mathcal{U} = ({U^{(1)}}^T, \ldots, {U^{(s)}}^T, {U^{n+1}}^T)^T \in \mathbb{R}^{(s+1)m}$, and $\mathcal{F}(\mathcal{U}) = (F(U^{(1)})^T, \ldots, F(U^{(s)})^T, (0)^T) \in \mathbb{R}^{(s+1)m}$. The symbol $\otimes$ denotes the Kronecker product ([23]),

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{11}B & a_{22}B & \cdots & a_{2m}B \\ & & \cdots & \\ a_{m1}B & a_{m1}B & \cdots & a_{mm}B \end{pmatrix}. \tag{A.6}$$

The Kronecker product satisfies $(A \otimes B)(C \otimes D) = (AB \otimes CD)$, $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

A.1.2 The Shu-Osher form of a Runge-Kutta scheme

Systems of ODEs such as (A.1) appear naturally when applying a Method of Lines (MOL) discretization to a time-dependent partial differential equation (PDE).

For (homogeneous) hyperbolic conservation laws, Shu and Osher in [87] developed a class time discretization methods that has become common practice in the field. In [87], the authors develop a class a Runge-Kutta schemes that are termed TVD (Total Variation Diminishing) because they maintain a TVD property on the numerical solution *provided* the first order Forward Euler time-discretization does so.

An s-stage RK method, in the form considered by Shu and Osher in [87] has the following form

$$\begin{aligned} U^{(1)} &= U^n \\ U^{(i)} &= \sum_{k=0}^{i-1} \left(\alpha_{ik} U^{(k)} + \triangle t \beta_{ik} F(U^{(k)}) \right), \qquad i = 2, \ldots, s+1 \\ U^{n+1} &= U^{(s+1)}, \end{aligned} \tag{A.7}$$

where α_{ik} and β_{ik}, $i = 2, \ldots, s+1$, are real coefficients specifying the Runge-Kutta method such that $\alpha_{ik} \geq 0, \sum_k^{i-1} \alpha_{ik} = 1$, $i = 2, \ldots, s+1$.

For a RK method in Shu-Osher form (A.7), the intermediate stages, $U^{(i)}$, amount to convex combinations of forward Euler operators, with Δt replaced by $\beta_{ik}\Delta t/\alpha_{ik}$.

A.1.3 Representations of Runge-Kutta methods

A RK method in the Shu-Osher form (A.7) can be written in compact form by considering the matrices $\Lambda = (\alpha_{ij})$, $\Gamma = (\beta_{ij}) \in \mathbb{R}^{(s+1)\times(s+1)}$ with $\alpha_{1j} = \beta_{1j} = 0$, $j = 1, \ldots, s+1$. It is easily seen that (A.7) can be expressed as

$$\mathcal{U} = \alpha \otimes U^n + (\Lambda \otimes I)\,\mathcal{U} + \triangle t(\Gamma \otimes I)\mathcal{F}(\mathcal{U}), \tag{A.8}$$

where $\alpha = (1, 0, \ldots, 0)^T \in \mathbb{R}^{s+1}$, and $\mathcal{U}$, $\mathcal{F}(\mathcal{U})$ are as defined after (A.5). Observe that $\Lambda e + \alpha = e$, Λ and Γ are strictly lower triangular, the matrix $I - \Lambda$ is invertible, and the last column in Γ, Λ is zero.

Using the properties of the Kronecker product and the fact that $(I - \Lambda)e = \alpha$, it is straightforward to convert (A.8) to the general, compact, form (A.5), see e.g. [54], with $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$.

On the other hand, the conversion from the Butcher coefficients $\mathbb{A}$ to a Shu-Osher representation is not unique [92] ,[54]. If the RK coefficient matrix $\mathbb{A}$ can be factorized as $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$, it is straightforward to rewrite (A.5) in a *generalized* Shu-Osher form[1]. For this, we pre-multiply (A.5) by $(I - \Lambda) \otimes I$ to obtain

$$((I - \Lambda) \otimes I)\mathcal{U} = (I - \Lambda)e \otimes U^n + \Delta t\left(((I - \Lambda) \otimes I) \otimes (\mathbb{A} \otimes I)\right)\mathcal{F}(\mathcal{U}) \tag{A.9}$$

so that, using the properties of the Kronecker product, we get

$$\mathcal{U} - (\Lambda \otimes I)\mathcal{U} = \alpha \otimes U^n + \Delta t(\Gamma \otimes I)\mathcal{F}(\mathcal{U}) \tag{A.10}$$

where $\alpha = (I - \Lambda)e$.

As we shall see shortly, being able to express a given RK scheme in Shu-Osher form has certain advantages, provided we can find $\Lambda \geq 0$ and $\Gamma \geq 0$ and $\alpha \geq 0$.

A detailed study on representations of implicit and explicit schemes is done in [33], [55].

[1] Notice that such splittings are always possible, a trivial (and obviously uninteresting) one being $\Lambda = 0$ and $\Gamma = \mathbb{A}$

A.2
Strong Stability and monotonicity

A relevant question in the numerical solution of systems ODEs is that of stability. For problems with smooth solutions, usually linear stability concepts are adequate. However, for problems with discontinuous solutions, such as solutions to hyperbolic problems, and other nonlinear problems, a stronger measure of stability is usually required.

In general, when the system of ODEs is solved numerically, it is natural to require that the numerical solution satisfies as many qualitative properties of the analytical solution as possible. Stability requirements stem from the desire to have numerical schemes tat preserve, at the discrete level, certain properties of the analytic solution of the problem to be solved.

An important class of problems are those whose solutions $U(t)$ satisfy a monotonicity property of the form

$$||U(t)|| \leq ||U(t_0)||, \qquad \forall t \geq t_0 \tag{A.11}$$

for a given norm $||\cdot||$ or semi-norm. For solutions satisfying (A.11), it is natural to require

$$||U^{n+1}|| \leq ||U^n|| \qquad \forall n \geq 0 \tag{A.12}$$

on the numerical solution as well.

Runge-Kutta methods that satisfy (A.12) are called *monotone* (for the stepsize Δt, function F, and norm, or semi-norm, $||\cdot||$). As mentioned in [32], the use of the term "monotone" is nicely in agreement with earlier use of the term by [8], [23], [60], [91]. In other works [44], property (A.12) is referred to as *strong stability*.

Relevant questions that have been considered in the specialized literature are those concerning conditional and unconditional monotonicity.

Definition A.1. *A numerical scheme satisfying (A.12) for any $\Delta t > 0$ is called unconditionally monotone.*

Definition A.2. *A numerical scheme satisfying (A.12) for any $\Delta t \leq \tau_0$ is called conditionally monotone.*

Conditional monotonicity, or strong stability under stepsize restrictions, for RK schemes has been studied by various authors (see e.g. [32], [55], [54] and references therein). In the standard context described

above, the study of conditional monotonicity is performed for systems of ODEs in a particular class, those such that $F(U)$ satisfies an inequality of the type (see e.g. [55])

$$||\rho y + F(y)|| \leq \rho ||y||, \qquad \forall y \tag{A.13}$$

for some fixed $\rho > 0$. This class of problems is denoted by $\mathcal{F}(\rho)$. It is easily seen (see [55],section 1) that this condition implies

$$||y + \tau F(y)|| \leq ||y||, \quad 0 \leq \tau \leq \frac{1}{\rho}, \qquad \forall y, \tag{A.14}$$

and that this also leads to (A.11) for the true solution of system (A.1).

In the context of Runge Kutta methods in general form (A.2)-(A.3), the concept of radius of absolute monotonicity plays an important role [32], [64] for questions related to strong stability. We review the definitions of absolute monotonicity and radius of absolute monotonicity next.

Definition A.3. *([56], Definition 2.1). An s-stage Runge-Kutta method with matrix coefficient $\mathbb{A}$ is said to be absolutely monotonic (a.m.) at a given point $\xi \leq 0$ if the matrix $I - \xi\mathbb{A}$ is nonsingular and*

$$\begin{aligned} (I - \xi\mathbb{A})^{-1}\mathbb{A} &\geq 0, \\ (I - \xi\mathbb{A})^{-1}e &\geq 0, \end{aligned}$$

where $e = (1, 1, \cdots, 1)^t \in \mathbb{R}^{s+1}$, and the vector inequalities are understood componentwise. Further, the method is said to be a.m. on a given set $\Omega \subset \mathbb{R}$ if it is a.m. at each $\xi \in \Omega$. The radius of absolute monotonicity $R(\mathbb{A})$ is defined by

$$R(\mathbb{A}) = \sup\{r | r \geq 0 \text{ and } \mathbb{A} \text{ is a.m. on } [-r, 0]\}. \tag{A.15}$$

If there is no $r > 0$ such that $\mathbb{A}$ is a.m. on [-r,0], we set $R(\mathbb{A}) = 0$.

As representative results of the importance of the radius of a.m. of a RK in establishing monotonicity results (or strong stability), we collect the following two theorems from [54], stated for *irreducible* RK schemes. We recall (see e.g. [32] for precise definitions) that an s-stage RK method is irreducible if it cannot be made equivalent to a method with less than s stages.

Theorem A.1. *([54], Theorem 2.7) Let $(F, ||\cdot||) \in \mathcal{F}(\rho)$. For any irreducible RK method $\mathbb{A}$, the following conditions are equivalent*

1. *$R(\mathbb{A}) = \infty$;*
2. *$\mathbb{A}$ is monotone for all $\Delta t > 0$.*

Theorem A.2. *([54], Theorem 2.9) Let $(F, ||\cdot||) \in \mathcal{F}(\rho)$, and $H > 0$. For any irreducible RK method $\mathbb{A}$, the following conditions are equivalent*

1. *$R(\mathbb{A}) \geq \rho H$;*
2. *$\mathbb{A}$ is monotone for $\Delta t \leq H$.*

In [32], the authors give an algorithm to compute the radius of absolute monotonicity of an irreducible RK scheme.

A.2.1 The Shu-Osher form and SSP schemes

Recall that for a RK method in Shu-Osher form (A.7), the intermediate stages, $U^{(i)}$, amount to convex combinations of Forward Euler operators, with Δt replaced by $\beta_{ik}\Delta t/\alpha_{ik}$. This observation easily leads to the following lemma

Lemma A.1. *(See [87], [44]). If the forward Euler method is strongly stable under a time-step restriction i.e.*

$$\|U^n + \triangle t F(U^n)\| \leq \|u^n\| \quad \forall \triangle t \leq \tau_{FE}. \tag{A.16}$$

with respect to a suitable norm (or semi-norm) $||\cdot||$, then the Runge-Kutta method (A.7) with $\beta_{ik} \geq 0$ is SSP, $\|U^{n+1}\| \leq \|U^n\|$, provided the following time-step restriction is satisfied

$$\triangle t \leq c\tau_{FE}, \quad c = \min_{i,k} \frac{\alpha_{ik}}{\beta_{ik}}. \tag{A.17}$$

Since the proof relies on convexity arguments, the result holds for any convex function, and not just the TVD semi-norm advocated in [87]. Time discretization processes with the property specified in Lemma A.1 are nowadays referred to as *Strong Stability Preserving* (**SSP**) schemes. The main idea is to *assume* that the first order forward Euler discretization of the system of ODEs is strongly stable under certain norm, when the time step is suitably restricted, and then to try to find a higher order time discretization (Multistep techniques have also been considered in [44]) that maintains strong stability in the same norm, maybe under a different time-step restriction. Hence, the term SSP was judged more suitable in [44], and has been used since.

Although developed in different contexts, the concepts of conditional monotonicity and SSP for RK schemes are equivalent, and the connection between known results in both contexts has been explored in recent years [32], [34], [55], [54].

A.2.2 Optimal Representations of Runge-Kutta methods

It is known that the representation of RK methods in the form (A.7) is not unique [55], hence according to Lemma A.1, different representations give rise to different stepsize restrictions for stability. As noted by Ferracina and Spijker [32], the question arises as to what is the largest factor c, not necessarily defined via (A.17) such that the conclusion in Lemma A.1 is still valid for a given RK scheme. The answer is given in [32], where it is proved that if (A.16) holds, irreducible RK methods are monotone under the stepsize restriction $\Delta t \leq c\tau_{FE}$ if and only if $c \leq R(\mathbb{A})$. In other words, the optimal coefficient for conditional monotonicity is given by the radius of absolute monotonicity of the method.

These results lead naturally to the question of the connection between the different representations of a RK scheme, and in particular to whether it is possible to construct optimal Shu-Osher representations from the Butcher tableau, i.e. the coefficient scheme in the standard form, of a RK method.

From [55] we extract the following theorem, that provides the *optimal*[2] representation.

[2]The optimal Shu-Osher representation is that for which the step-size coefficient is $c = R(\mathbb{A})$, the largest possible value

Theorem A.3. *([55], Proposition 2.7). We consider an RK method $\mathbb{A}$. If $0 < r = \mathcal{R}(\mathbb{A}) < \infty$, then there exist matrices $\Lambda = (\alpha_{ij})$ and $\Gamma = (\beta_{ij})$ such that $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$ with $\Lambda \geq 0$, $\Gamma \geq 0$, $\Lambda \mathbf{e} \leq \mathbf{e}$, $I - (\Lambda - r\Gamma)$ invertible, and $\Lambda - r\Gamma \geq 0$.*

The proof of the result above is constructive. A representation can be constructed so that $\Lambda - r\Gamma = 0$ by defining

$$\Lambda = r\mathbb{A}(I + r\mathbb{A})^{-1}, \qquad \Gamma = \mathbb{A} - \Lambda\mathbb{A}. \tag{A.18}$$

In this case, the representation is optimal.

The following example, which shall be used in chapter 3, shows how to construct such matrices.

Example Let us consider the explicit RK scheme given by the following Butcher tableau

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \mathbb{A} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \tag{A.19}$$

It is known that $R(\mathbb{A}) = 2$. We can readily compute Λ and Γ in (A.18) and we obtain

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & 0 \end{pmatrix} \qquad \Gamma = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \end{pmatrix} \qquad \alpha = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix} \tag{A.20}$$

where, obviously, $\Lambda - 2\Gamma = 0$. Notice that, since $U^{(1)} = U^n$, we can add α_4 to $\Lambda_{4,1}$, so that we get an equivalent representation by considering

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \end{pmatrix} \qquad \alpha = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{A.21}$$

In this case, the representation satisfies $\Lambda - 2\Gamma \geq 0$. This process can be carried out for any explicit RK scheme, so that we can always assume that $\alpha = (1, 0, \ldots, 0)^T$ in this case.

A.3
Weak Stability and Weak Stability Preserving Schemes

In some cases, as for the model problem in chapter 3, the concept of strong stability is too demanding. A weaker form of stability has been introduced in chapter 3 that seeks to preserve an important property of the true solution of the model, namely that the values of the solution always belong to the interval $[0, 1]$. In this memoir, we refer to this property as *weak stability* (**WS**), as opposed to the *strong stability* (**SS**) just mentioned, which prevents growth in a given norm (or semi-norm).

A method is then termed *weakly stable* if it satisfies

$$0 \leq U^0 \leq 1 \quad \Rightarrow \quad 0 \leq U^n \leq 1, \quad \forall n \geq 0. \tag{A.22}$$

As in the SSP context, the goal is then to find higher order RK schemes that preserve the WS property, provided that the Forward Euler discretization is WS, perhaps under a different stepsize restriction. These schemes are then referred too as *Weak Stability Preserving* (**WSP**) in this memoir.

Convexity arguments can be used to prove that certain explicit RK schemes are WSP. The theory laid out in the previous section allow us to prove the following results for explicit RK schemes. In what follows, we assume that $F(U)$ is such that the Forward Euler discretization of system (A.1) is WS, under a certain stepsize restriction, i.e.

$$0 \leq U \leq 1 \quad \Rightarrow \quad U + \Delta t F(U) \leq 1 \qquad \forall \Delta t \leq \tau_{FE}. \tag{A.23}$$

Theorem A.4. *Let us consider an explicit RK scheme expressed as (A.10) satisfying*

$$\Lambda \geq 0, \quad \Gamma \geq 0, \quad \Lambda e \leq e, \qquad \alpha = (I - \Lambda)e \tag{A.24}$$

so that there exists $r > 0$ such that

$$\Lambda - r\Gamma \geq 0 \tag{A.25}$$

then, it is WSP under the stepsize restriction

$$0 \leq \Delta t \leq r\tau_{FE}. \tag{A.26}$$

Proof. Notice that the conditions (A.24) imply $\alpha_{ij} \geq 0$, $\beta_{ij} \geq 0$, $\sum_j \alpha_{ij} \leq 1$ and $0 \leq \alpha_i \leq 1$, while condition (A.25) implies that $\sum_j (\alpha_{ij} - r\beta_{ij}) \geq 0$.

In order to prove the result, we rewrite first (A.10) by adding and subtracting $(r\Gamma \otimes I)\mathcal{U}$ to its right hand side (RHS). We easily get the following equivalent expression for the method:

$$\mathcal{U} = \alpha \otimes U^n + ((\Lambda - r\Gamma) \otimes I)\mathcal{U} + r(\Gamma \otimes I)\left(\mathcal{U} + \frac{\Delta t}{r}\mathcal{F}(\mathcal{U})\right). \tag{A.27}$$

The result can now be proven by induction over the internal stages of the method. Since the method is explicit, (A.27) is in fact

$$U^{(1)} = \alpha_1 U^n \tag{A.28}$$

$$U^{(i)} = \alpha_i U^n + \sum_{j=1}^{i-1}(\alpha_{ij} - r\beta_{ij})U^{(j)} + r\sum_{j=1}^{i-1}\beta_{ij}(U^{(j)} + \frac{\Delta t}{r}F(U^{(j)})),\ i = 1, \ldots, s \tag{A.29}$$

$$U^{n+1} = \alpha_{s+1} U^n + \sum_{j=1}^{s}(\alpha_{ij} - r\beta_{ij})U^{(j)} + r\sum_{j=1}^{s}\beta_{ij}(U^{(j)} + \frac{\Delta t}{r}F(U^{(j)})). \tag{A.30}$$

If $0 \leq U^n \leq 1$, since $0 \leq \alpha_1 \leq 1$, we get from (A.28) that $0 \leq U^{(1)} \leq 1$.

Let us assume that we have $0 \leq U^{(j)} \leq 1$ for $0 \leq j \leq j-1$. For $\Delta t \leq r\tau_{FE}$ we also have

$$0 \leq U^{(j)} + \frac{\Delta t}{r}F(U^{(j)}) \leq 1, \tag{A.31}$$

hence, from (A.29) we get

$$0 \leq U^{(i)} \leq \alpha_i + \sum_{j=1}^{i-1}(\alpha_{ij} - r\beta_{ij}) + r\sum_{j=1}^{i-1}\beta_{ij} = (e - \Lambda e)_i + (\Lambda e)_i = e_i = 1,$$

which completes the proof.

A similar study can be made directly with the RK method written in the standard form that comes directly from its Butcher tableau.

Theorem A.5. *Consider an explicit RK scheme given in (A.5) and let $r > 0$ be such that*

$$(I + r\mathbb{A})^{-1}e \geq 0, \qquad (I + r\mathbb{A})^{-1}\mathbb{A} \geq 0 \tag{A.32}$$

then, it is WSP under the stepsize restriction

$$0 \leq \Delta t \leq r\tau_{FE}. \tag{A.33}$$

Proof. We add $r(\mathbb{A}\otimes I)\mathcal{U}$ to both sides of (A.5) and combine terms to rewrite this relation as

$$\mathcal{U} = (I + r\mathbb{A})^{-1}e \otimes U^n + r((I + r\mathbb{A})^{-1}\mathbb{A}\otimes I)\left(\mathcal{U} + \frac{\Delta t}{r}\mathcal{F}(\mathcal{U})\right).$$

Then, conditions (A.32) allow us to carry out an induction process on the internal stages just as in the case of the previous theorem.

Notice that the largest positive number r such that (A.32) hold is $r_0 = R(\mathbb{A})$, provided $r_0 > 0$, as it can be deduced from definition A.3.

A.4
Additive Runge-Kutta methods. Stability properties

MOL discretizations of some time-dependent PDEs give rise to systems of ODEs that contain additive terms with different stiffness properties. A typical situation involves initial value problems of the form

$$\frac{d}{dt}U(t) = L(U(t)) + S(U(t)) \quad U(0) = U_0, \quad t \geq 0, \tag{A.34}$$

where L and S are continuous functions from $\mathbb{R}^m$ to $\mathbb{R}^m$ with different stiffness properties.

A common class of one step methods for solving the initial value problem (A.34) numerically is that of *Additive Runge-Kutta* (ARK) methods. The results collected in this section are taken mainly from [56].

An s-stage ARK method is defined by two $s \times s$ real matrices $\mathcal{A} = (a_{ij})$ and $\tilde{\mathcal{A}} = (\tilde{a}_{ij})$, and two real vectors $b, \tilde{b} \in \mathbb{R}^s$ such that

$$U^{(i)} = U^n + \triangle t \sum_{j=1}^{s} a_{ij} L(U^{(j)}) + \triangle t \sum_{j=1}^{s} \tilde{a}_{ij} S(U^{(j)}), \quad 1 \leq i \leq s \tag{A.35}$$

$$U^{n+1} = U^n + \triangle t \sum_{i=1}^{s} b_i L(U^{(i)}) + \triangle t \sum_{i=1}^{s} \tilde{b}_i S(U^{(i)}). \tag{A.36}$$

An ARK scheme is defined, in the usual Butcher notation, by a double *tableau*

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^t \end{array} \qquad \begin{array}{c|c} \tilde{c} & \tilde{\mathcal{A}} \\ \hline & \tilde{b}^t \end{array}$$

where the coefficients c and $\tilde{c}$ are given by the relations[3]

$$c_i = \sum_{j=1}^{s} a_{ij}, \qquad \tilde{c}_i = \sum_{j=1}^{s} \tilde{a}_{ij}. \tag{A.37}$$

As the standard RK schemes, an ARK scheme can be conveniently written in compact form by using the $(s+1)\times(s+1)$ matrices $\mathbb{A}$ and $\tilde{\mathbb{A}}$, defined as follows:

$$\mathbb{A} = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix} \qquad \tilde{\mathbb{A}} = \begin{pmatrix} \tilde{\mathcal{A}} & 0 \\ \tilde{b}^t & 0 \end{pmatrix}.$$

Then, (A.35) and (A.36) can be written as

$$\mathcal{U} = e \otimes U^n + \triangle t(\mathbb{A} \otimes I)\mathcal{L}(\mathcal{U}) + \triangle t(\tilde{\mathbb{A}} \otimes I)\mathcal{S}(\mathcal{U}), \tag{A.38}$$

where we have denoted $e = (1, \ldots, 1) \in \mathbb{R}^{s+1}$, $\mathcal{U} = ({U^{(1)}}^T, \ldots, {U^{(s)}}^T, {U^{n+1}}^T)^T \in \mathbb{R}^{(s+1)N}$, and $\mathcal{L}(\mathcal{U}) = (L(U^{(1)})^T, \ldots, L(U^{(s)})^T, (0)^T) \in \mathbb{R}^{(s+1)N}$, and an analogous definition for $S(\mathcal{U})$. The symbol $\otimes$ denotes the Kronecker product ([23]).

The concept of monotonicity has been extended to ARK methods. We recall the following definition from [56]. Monotonicity and SSP, properties for ARK methods are analyzed in [56].

Definition A.4. *([56], Definition 2.2, Definition 2.3) An s-stage ARK method* $(\mathbb{A}, \tilde{\mathbb{A}})$ *is said to be* absolutely monotonic *(a.m.) at a given point* $(\xi, \tilde{\xi})$*, with* $\xi, \tilde{\xi} \leq 0$*, if the matrix* $I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}}$ *is invertible and*

$$(I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\mathbb{A} \geq 0, \tag{A.39}$$

$$(I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\tilde{\mathbb{A}} \geq 0, \tag{A.40}$$

$$(I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}e \geq 0. \tag{A.41}$$

Further, the method is said to be a.m. on a given set $\Omega \in \mathbb{R}^2$ *if it is a.m. at each* $(\xi, \tilde{\xi}) \in \Omega$

The region of absolute monotonicity*, denoted by* $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$*, is defined by*

$$\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) = \left\{ r \geq 0, \tilde{r} \geq 0 \text{ and } (\mathbb{A}, \tilde{\mathbb{A}}) \text{ a.m. on } [-r, 0] \times [-\tilde{r}, 0] \right\}.$$

The curve of absolute monotonicity*, denoted by* $\partial\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$ *is the frontier of the set* $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$*, excluding the coordinate axis.*

If there is no $r > 0$*,* $\tilde{r} > 0$ *such that* $(\mathbb{A}, \tilde{\mathbb{A}})$ *is a.m on* $[-r, 0] \times [-\tilde{r}, 0]$*, the curve of a.m. is set to a point, i.e.* $\partial\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) = (0, 0)$*.*

[3] these coefficients are only used in the treatment of non autonomous systems

We recall here Theorem 3.1 in [56], that ensures monotonicity of the results obtained with the IMEX scheme under certain stepsize restrictions. We also remark that in order to obtain the results in [56] assume that that L and S satisfy

$$\|U + \tau_L L(U)\| \leq \|U\| \quad (\text{ for all } U \in \mathbb{R}^m), \tag{A.42}$$
$$\|U + \tau_S S(U)\| \leq \|U\| \quad (\text{ for all } U \in \mathbb{R}^m), \tag{A.43}$$

for some fixed τ_L and τ_S.

Theorem A.6. *(Monotonicity for ARK, Theorem* 3.1*, [56]). Assume that the ARK method ($\mathbb{A}$,$\tilde{\mathbb{A}}$) is absolute monotonic (a.m.) at ($-\sigma$,$-\tilde{\sigma}$). Then it holds that*

$$\| U^{(i)} \| \leq \| U^n \|, \qquad i = 1, \ldots, s, \tag{A.44}$$
$$\| U^{n+1} \| \leq \| U^n \|, \tag{A.45}$$

for

$$\triangle t \leq \sigma \tau_L, \qquad \triangle t \leq \tilde{\sigma} \tau_S, \tag{A.46}$$

where the parameters τ_L and τ_S are the maximal stepsizes for which relations (A.42) and (A.43) hold.

B

Shallow water equations

A wide variety of physical phenomena involving shallow water flow in oceanography and atmospheric sciences are conveniently modeled by considering the one-dimensional or two-dimensional shallow water equations, also called Saint-Venant equations, in honor to the French mathematician Adhémar Jean Claude Barré de Saint-Venant (1797-1886) who was the first one to deduce them. The resulting shallow water equations are a hyperbolic system of conservation laws that approximately describes various geophysical flows, such as tides in oceans, simulation of internal tides in the Strait of Gibraltar [11], [12], breaking of waves on shallow beaches, roll waves in open channels, flood waves in rivers, surges and dam-break modeling and also sediments transport [13],[21]. The shallow water approximation can also be applied to flows of hetero-

geneous mixtures and to the modeling of atmospheric flows.

B.1
Navier Stokes equations

The Navier Stokes equations describe the movement of a fluid. The equations were derived independently by G.G. Stokes, in England, and M. Navier, in France, in the early 1800's. The general conservation laws of mass and momentum written in differential conservation law form for an incompressible fluid, are

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, & \text{in } \tilde{D} \\ \dfrac{\partial \rho \mathbf{v}}{\partial t} + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v}) = \rho \mathbf{g} - \nabla p + \nu \triangle \mathbf{v}, & \text{in } \tilde{D} \end{cases} \tag{B.1}$$

where

$$\tilde{D} = \left\{ (\mathbf{x}, t) : \mathbf{x} \in \Omega_t \subset \mathbb{R}^3; t \in (0, T) \right\},$$

and Ω_t is the fluid generated volume at time t. The independent variables are t for time and $\mathbf{x} = (x, y, z)$ for space. The dependent variables are ρ for density, $\mathbf{v} = (v_1(\mathbf{x}, t), v_2(\mathbf{x}, t), v_3(\mathbf{x}, t))$ for velocity; p is the pressure; the vector g is a body force vector; the tensor involved is

$$\mathbf{v} \otimes \mathbf{v} \quad = \quad \begin{pmatrix} v_1^2 & v_1 v_2 & v_1 v_3 \\ v_2 v_1 & v_2^2 & v_2 v_3 \\ v_2 v_1 & v_3 v_2 & v_3^2 \end{pmatrix}.$$

Assuming the density of the fluid is known and constant

$$\rho(\mathbf{x}, t) = \rho_0,$$

we obtain

$$\begin{cases} \nabla \cdot \mathbf{v} = 0, & \text{in } \tilde{D} \\ \rho_0 \left(\dfrac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} \right) = \rho_0 \mathbf{g} - \nabla p + \nu \triangle \mathbf{v}, & \text{in } \tilde{D} \end{cases} \tag{B.2}$$

Let us assume that the body force vector is $\mathbf{g} = (0, 0, -g)$, where g is the acceleration due to the gravity, taken as $g = 9.8 m/s^2$, a constant. Then

we can rewrite the equation (B.2) as

$$\nabla \cdot \mathbf{v} = 0 \tag{B.3}$$

$$\frac{\partial v_1}{\partial t} + v_1\frac{\partial v_1}{\partial x} + v_2\frac{\partial v_1}{\partial y} + v_3\frac{\partial v_1}{\partial z} = -\frac{1}{\rho_0}\frac{\partial p}{\partial x} + \frac{\nu}{\rho_0}\triangle v_1 \tag{B.4}$$

$$\frac{\partial v_2}{\partial t} + v_1\frac{\partial v_2}{\partial x} + v_2\frac{\partial v_2}{\partial y} + v_3\frac{\partial v_2}{\partial z} = -\frac{1}{\rho_0}\frac{\partial p}{\partial y} + \frac{\nu}{\rho_0}\triangle v_2 \tag{B.5}$$

$$\frac{\partial v_3}{\partial t} + v_1\frac{\partial v_3}{\partial x} + v_2\frac{\partial v_3}{\partial y} + v_3\frac{\partial v_3}{\partial z} = -g - \frac{1}{\rho_0}\frac{\partial p}{\partial z} + \frac{\nu}{\rho_0}\triangle v_3 \tag{B.6}$$

See [37], [101] for more details.

Hydrostatic pressure

A key assumption made in the derivation of the approximate shallow water theory concerns the pressure distribution; this is given as in hydrostatics and results from assuming that the vertical acceleration of the water particles, given by

$$\frac{dv_3}{dt} = \frac{\partial v_3}{\partial t} + v_1\frac{\partial v_3}{\partial x} + v_2\frac{\partial v_3}{\partial y} + v_3\frac{\partial v_3}{\partial z}, \tag{B.7}$$

has a negligible effect on the pressure, as well as $\triangle v_3$. For this reason, equation (B.6) is reduced to

$$-g - \frac{1}{\rho_0}\frac{\partial p}{\partial z} = 0. \tag{B.8}$$

Finally, we integrate this equation between the free surface ($z = s$) and z

$$p = p_a + \rho_0 g(s - z) \tag{B.9}$$

where p_a is the atmospheric pressure.

B.2
Water Flow with a Free Surface

Consider the flow of water with a free surface under gravity in a three-dimensional domain, see figure B.1, where z defines the vertical direction, which is associated with the free-surface elevation. There is a new unknown variable, $h(x, y, t)$ the depth of water, the vertical distance

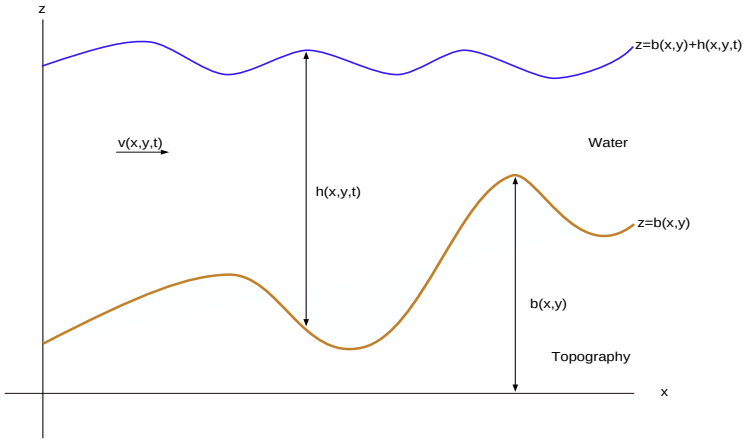


Figure B.1: *Flow with free surface under gravity, for a fixed section* y

between the bottom and the free-surface position. The classical shallow water models obtain h by integrating the conservation law equation $\nabla \cdot \mathbf{v} = 0$ between the bottom boundary, also called *bed*, and defined by a function

$$z = b(x, y), \tag{B.10}$$

and the free surface, defined by

$$z = s(x, y, t) = b(x, y) + h(x, y, t). \tag{B.11}$$

Following [37], [97] and [101], we have

$$\int_b^s \left(\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z} \right) dz = 0 \tag{B.12}$$

which applying Leibniz's formula to the first two terms leads to

$$\begin{aligned} \frac{\partial}{\partial x} \int_b^s v_1 dz - v_1|_{z=s} \frac{\partial s}{\partial x} + v_1|_{z=b} \frac{\partial b}{\partial x} \\ + \frac{\partial}{\partial y} \int_b^s v_2 dz - v_2|_{z=s} \frac{\partial s}{\partial y} + v_2|_{z=b} \frac{\partial b}{\partial y} \\ + v_3|_{z=s} - v_3|_{z=b} \quad = \quad 0. \end{aligned} \tag{B.13}$$

Assume that a boundary is given by the surface

$$\gamma(x, y, z, t) = 0. \tag{B.14}$$

For instance, in figure B.1 the height of the free surface is specified as $z = s(x, y, t)$ and an appropriate function $\gamma(x, y, z, t)$ would be given by

$$\gamma(x, y, z, t) \equiv z - s(x, y, t) = 0, \tag{B.15}$$

and for the bottom boundary

$$\gamma(x, y, z, t) \equiv z - b(x, y) = 0. \tag{B.16}$$

Fluid particles on the free surface always remain part of the free surface, therefore we must have

$$\frac{d}{dt}\gamma(x, y, z, t) = \gamma_t + v_1\gamma_x + v_2\gamma_y + v_3\gamma_z = 0. \tag{B.17}$$

This is the kinematic boundary condition. For surface whose position is described in the form (B.15), the kinematic boundary condition becomes

$$v_3|_{z=s} = \frac{\partial s}{\partial t} + v_1|_{z=s}\frac{\partial s}{\partial x} + v_2|_{z=s}\frac{\partial s}{\partial y}, \tag{B.18}$$

and for the bottom boundary $b(x, y)$, condition (B.17) also applies, with γ given by (B.16), we obtain

$$v_3|_{z=b} = v_1|_{z=b}\frac{\partial b}{\partial x} + v_2|_{z=b}\frac{\partial b}{\partial y}. \tag{B.19}$$

Substitution of (B.18) and (B.19) into (B.13) gives

$$\frac{\partial s}{\partial t} + \frac{\partial}{\partial x}\int_b^s v_1 dz + \frac{\partial}{\partial y}\int_b^s v_2 dz = 0. \tag{B.20}$$

Let us denote the horizontal average velocity by

$$\bar{v}_1 = \frac{1}{h}\int_b^s v_1 dz, \qquad \bar{v}_2 = \frac{1}{h}\int_b^s v_2 dz, \tag{B.21}$$

so that equation (B.20) becomes

$$\frac{\partial s}{\partial t} + \frac{\partial(h\bar{v}_1)}{\partial x} + \frac{\partial(h\bar{v}_2)}{\partial y} = 0. \tag{B.22}$$

As $s = b + h$ and $b_t = 0$, (B.22) simplifies to

$$\frac{\partial h}{\partial t} + \frac{\partial q_1}{\partial x} + \frac{\partial q_2}{\partial y} = 0, \tag{B.23}$$

where $\mathbf{q} = (q_1, q_2) = (h\bar{v}_1, h\bar{v}_2)$ is the discharge. The equation (B.23) is the law of **conservation of mass** for shallow water equations, and is written in differential conservation form.

In order to express the momentum equations in conservation form, we add equation (B.3) pre-multiplied by v_1 to equation (B.4) and using (B.9), to obtain

$$\frac{\partial v_1}{\partial t} + \frac{\partial v_1^2}{\partial x} + \frac{\partial v_1 v_2}{\partial y} + \frac{\partial v_1 v_3}{\partial z} = -\frac{1}{\rho_0}\frac{\partial p_a}{\partial x} - g\frac{\partial s}{\partial x} + \frac{\nu}{\rho_0}\triangle v_1. \tag{B.24}$$

Similarly for the (B.5) equation, we obtain

$$\frac{\partial v_2}{\partial t} + \frac{\partial v_1 v_2}{\partial x} + \frac{\partial v_2^2}{\partial y} + \frac{\partial v_2 v_3}{\partial z} = -\frac{1}{\rho_0}\frac{\partial p_a}{\partial y} - g\frac{\partial s}{\partial y} + \frac{\nu}{\rho_0}\triangle v_2. \tag{B.25}$$

Now, we integrate vertically, between the bed $b(x, y)$ and the free surface $s(x, y, t)$, the equation (B.24) and (B.25)

$$\begin{aligned}\int_b^s \left(\frac{\partial v_1}{\partial t} + \frac{\partial v_1^2}{\partial x} + \frac{\partial v_1 v_2}{\partial y} + \frac{\partial v_1 v_3}{\partial z}\right) dz &= -h\left(\frac{1}{\rho_0}\frac{\partial p_a}{\partial x} + g\frac{\partial s}{\partial x}\right) \\ &\quad + \frac{\nu}{\rho_0}\int_b^s \triangle v_1 dz\end{aligned} \tag{B.26}$$

$$\begin{aligned}\int_b^s \left(\frac{\partial v_2}{\partial t} + \frac{\partial v_1 v_2}{\partial x} + \frac{\partial v_2^2}{\partial y} + \frac{\partial v_2 v_3}{\partial z}\right) dz &= -h\left(\frac{1}{\rho_0}\frac{\partial p_a}{\partial y} + g\frac{\partial s}{\partial y}\right) \\ &\quad + \frac{\nu}{\rho_0}\int_b^s \triangle v_2 dz.\end{aligned} \tag{B.27}$$

In order to determine the left hand side in the above equations, we use Leibniz's formula to obtain

$$\begin{aligned}&\int_b^s \left(\frac{\partial v_1}{\partial t} + \frac{\partial v_1^2}{\partial x} + \frac{\partial v_1 v_2}{\partial y} + \frac{\partial v_1 v_3}{\partial z}\right) dz = \frac{\partial}{\partial t}\int_b^s v_1 dz - v_1|_{z=s}\frac{\partial s}{\partial t} \\ &+ \frac{\partial}{\partial x}\int_b^s v_1^2 dz - v_1^2|_{z=s}\frac{\partial s}{\partial x} + v_1^2|_{z=b}\frac{\partial b}{\partial x} + \frac{\partial}{\partial y}\int_b^s (v_1 v_2) dz \\ &- (v_1 v_2)|_{z=s}\frac{\partial s}{\partial y} + (v_1 v_2)|_{z=b}\frac{\partial b}{\partial y} + (v_1 v_3)|_{z=s} - (v_1 v_3)|_{z=b}.\end{aligned} \tag{B.28}$$

We apply kinematic boundary conditions (B.18) and (B.19) to (B.28)

$$\int_b^s \left(\frac{\partial v_1}{\partial t} + \frac{\partial v_1^2}{\partial x} + \frac{\partial v_1 v_2}{\partial y} + \frac{\partial v_1 v_3}{\partial z}\right) dz = \frac{\partial}{\partial t}\int_b^s v_1 dz + \frac{\partial}{\partial x}\int_b^s v_1^2 dz + \frac{\partial}{\partial y}\int_b^s (v_1 v_2) dz, \tag{B.29}$$

finally, using the average horizontal velocity (B.21), we could rewrite (B.29) as

$$\int_b^s \left(\frac{\partial v_1}{\partial t} + \frac{\partial v_1^2}{\partial x} + \frac{\partial v_1 v_2}{\partial y} + \frac{\partial v_1 v_3}{\partial z} \right) dz \simeq \frac{\partial \left(\bar{v}_1 h \right)}{\partial t} + \frac{\partial \left(\bar{v}_1^2 h \right)}{\partial x} + \frac{\partial \left(\bar{v}_1 \bar{v}_2 h \right)}{\partial y}. \quad \text{(B.30)}$$

Similarly for the left hand side of the equation (B.25) to obtain

$$\int_b^s \left(\frac{\partial v_2}{\partial t} + \frac{\partial v_1 v_2}{\partial x} + \frac{\partial v_2^2}{\partial y} + \frac{\partial v_2 v_3}{\partial z} \right) dz \simeq \frac{\partial \left(\bar{v}_2 h \right)}{\partial t} + \frac{\partial \left(\bar{v}_1 \bar{v}_2 h \right)}{\partial x} + \frac{\partial \left(\bar{v}_2^2 h \right)}{\partial y}. \quad \text{(B.31)}$$

Let us see how to rewrite the right hand side of the equations (B.26) and (B.27):

$$\int_b^s \triangle v_1 dz = \int_b^s \triangle_{xy} v_1 dz + \int_b^s \frac{\partial^2 v_1}{\partial z^2} dz \quad \text{(B.32)}$$

$$\int_b^s \triangle v_2 dz = \int_b^s \triangle_{xy} v_2 dz + \int_b^s \frac{\partial^2 v_2}{\partial z^2} dz, \quad \text{(B.33)}$$

where $\triangle_{xy} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. We make the assumption that the flow velocity is independent of depth, so that $\frac{\partial^2 v_1}{\partial z^2} = \frac{\partial^2 v_2}{\partial z^2} = 0$ and we have

$$\int_b^s \triangle v_1 dz = \int_b^s \triangle_{xy} v_1 dz \quad \text{(B.34)}$$

$$\int_b^s \triangle v_2 dz = \int_b^s \triangle_{xy} v_2 dz. \quad \text{(B.35)}$$

Now, if we use the horizontal average velocity (B.21), we could approximate (B.34) and (B.35) as

$$\int_b^s \triangle v_1 dz \simeq h \triangle_{xy} \bar{v}_1 \quad \text{(B.36)}$$

$$\int_b^s \triangle v_2 dz \simeq h \triangle_{xy} \bar{v}_2. \quad \text{(B.37)}$$

Then, the equations (B.26) and (B.27) are rewritten by

$$\frac{\partial \left(\bar{v}_1 h \right)}{\partial t} + \frac{\partial \left(\bar{v}_1^2 h \right)}{\partial x} + \frac{\partial \left(\bar{v}_1 \bar{v}_2 h \right)}{\partial y} = -h \left(\frac{1}{\rho_0} \frac{\partial p_a}{\partial x} + g \frac{\partial s}{\partial x} \right) + \frac{\nu}{\rho_0} h \triangle_{xy} \bar{v}_1 \quad \text{(B.38)}$$

$$\frac{\partial \left(\bar{v}_2 h \right)}{\partial t} + \frac{\partial \left(\bar{v}_1 \bar{v}_2 h \right)}{\partial x} + \frac{\partial \left(\bar{v}_2^2 h \right)}{\partial y} = -h \left(\frac{1}{\rho_0} \frac{\partial p_a}{\partial y} + g \frac{\partial s}{\partial y} \right) + \frac{\nu}{\rho_0} h \triangle_{xy} \bar{v}_2 \quad \text{(B.39)}$$

Finally, for convenience, p_a, the atmospheric pressure is taken to be identically zero. We also make use of (B.11) and assume differentiability

of the water depth h, as well as rewrite the equations above in terms of the discharge $\mathbf{q}$ in order to obtain the **shallow water** equations

$$\frac{\partial h}{\partial t} + \frac{\partial q_1}{\partial x} + \frac{\partial q_2}{\partial y} = 0 \tag{B.40}$$

$$\frac{\partial q_1}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q_1^2}{h} + \frac{g}{2}h^2\right) + \frac{\partial}{\partial y}\left(\frac{q_1 q_2}{h}\right) = -gh\frac{\partial b}{\partial x} + \frac{\nu}{\rho_0}h\triangle_{xy}\bar{v}_1 \tag{B.41}$$

$$\frac{\partial q_2}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q_1 q_2}{h}\right) + \frac{\partial}{\partial y}\left(\frac{q_2^2}{h} + \frac{g}{2}h^2\right) = -gh\frac{\partial b}{\partial y} + \frac{\nu}{\rho_0}h\triangle_{xy}\bar{v}_2. \tag{B.42}$$

The system that we are going to consider is the shallow water equations neglecting viscosity terms

$$\begin{aligned}
\frac{\partial h}{\partial t} + \frac{\partial q_1}{\partial x} + \frac{\partial q_2}{\partial y} &= 0 \\
\frac{\partial q_1}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q_1^2}{h} + \frac{g}{2}h^2\right) + \frac{\partial}{\partial y}\left(\frac{q_1 q_2}{h}\right) &= -gh\frac{\partial b}{\partial x} \\
\frac{\partial q_2}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q_1 q_2}{h}\right) + \frac{\partial}{\partial y}\left(\frac{q_2^2}{h} + \frac{g}{2}h^2\right) &= -gh\frac{\partial b}{\partial y}.
\end{aligned} \tag{B.43}$$

Through the derivation of the depth averaged shallow water equations several assumptions have been done. It is very important to have in mind the approximations made in each different term, in order to know the limitations of the equations, when they can be applied, and to understand and interpret the results obtained from them. The different hypotheses made are summarized below

- The water is of uniform density ρ_0 and the layer of water has thickness h.
- The slope of the water surface is small compared to unity and the horizontal scale of flow features is large compared to the depth of the water.
- Friction with the bottom surface is neglected.
- The water within the layer is in hydrostatic balance. The pressure at the upper surface is zero. (This is trivially extendable to the case of constant pressure at the surface).

There is a lot of literature around the derivation of the shallow water equations from Navier-Stokes equations, we would like to mention some of them, like Stoker in [93], Friedrichs in [35], Whitham in [108], Toro in [101], and the Phd thesis of García [37], Cea [14] and Fe [30].

B.3 Wave formation

This section is intended to revise some theoretical aspects in the formation of different types of waves as a solution of the Riemann problem of two adjacent states. A detailed and rigorous study on this topic can be found in [39], [70], [90] and a direct study on the shallow water equations can be found in [101].

The system of shallow water equations (B.43) can be written in the two dimensional case as follows

$$\mathbf{u}_t + f_1(\mathbf{u})_x + f_2(\mathbf{u})_y = \mathbf{s}(\mathbf{x}, \mathbf{u}) \tag{B.44}$$

$$\begin{pmatrix} h \\ q_1 \\ q_2 \end{pmatrix} + \begin{pmatrix} q_1 \\ \frac{q_1^2}{h} + \frac{g}{2}h^2 \\ \frac{q_1 q_2}{h} \end{pmatrix}_x + \begin{pmatrix} q_2 \\ \frac{q_1 q_2}{h} \\ \frac{q_2^2}{h} + \frac{g}{2}h^2 \end{pmatrix}_y = \begin{pmatrix} 0 \\ -ghz_x \\ -ghz_y \end{pmatrix}$$

where as we mentioned in the previous section, h is the water depth, q_1 and q_2 are the two components of the discharge (momentum), and z denotes now the bottom topography. The corresponding eigenvalues (*characteristic velocities*) of the Jacobian matrices of the flux components f_1 and f_2 are:

$$\begin{array}{lll} \lambda_1^{(1)} = v_1 - c & \lambda_2^{(1)} = v_1 & \lambda_3^{(1)} = v_1 + c \\ \lambda_1^{(2)} = v_2 - c & \lambda_2^{(2)} = v_2 & \lambda_3^{(2)} = v_2 + c \end{array}$$

where $v_i = q_i/h$ for $i = 1, 2$ are the components of the fluid velocity and $c = \sqrt{gh}$ is the sound velocity. The superscripts (1) and (2) refer to the component of the flux vector. The *characteristic variables* are defined as $R^{-1}\mathbf{u}$, where the matrices of right ($R^{(1)}, R^{(2)}$ with components $r_i(\mathbf{u})$) and left ($L^{(1)}, L^{(2)}$) eigenvectors are:

$$R^{(1)} = \begin{pmatrix} 1 & 0 & 1 \\ \lambda_1^{(1)} & 0 & \lambda_3^{(1)} \\ v_2 & 1 & v_2 \end{pmatrix} \quad L^{(1)} = \begin{pmatrix} \lambda_3^{(1)}/(2c) & -1/(2c) & 0 \\ -v_2 & 0 & 1 \\ -\lambda_1^{(1)}/(2c) & 1/(2c) & 0 \end{pmatrix}$$

$$R^{(2)} = \begin{pmatrix} 1 & 0 & 1 \\ v_1 & 1 & v_1 \\ \lambda_1^{(2)} & 0 & \lambda_3^{(2)} \end{pmatrix} \quad L^{(2)} = \begin{pmatrix} \lambda_3^{(2)}/(2c) & 0 & -1/(2c) \\ -v_1 & 1 & 0 \\ -\lambda_1^{(2)}/(2c) & 0 & 1/(2c) \end{pmatrix}$$

We now study the solution of the Riemann problem formed by two adjacent states, so we consider the one dimension shallow water equations with flat topography:

$$\begin{aligned} &h_t + q_x = 0 \\ &q_t + \left(\tfrac{q^2}{h} + \tfrac{1}{2}gh^2\right)_x = 0 \end{aligned} \tag{B.45}$$

with the initial-value data of the form

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_l, & x < 0 \\ \mathbf{u}_r, & x > 0, \end{cases} \tag{B.46}$$

B.3.1 Elementary wave solutions

There are four possible patterns that may occur in the solution of the Riemann problem (B.45), (B.46). These are illustrated in figure B.2. Case (a) is where the left wave is a rarefaction wave and the right wave is a shock wave; case (b) is where the left wave is a shock and the right wave is a rarefaction; case (c) is where both left and right waves are rarefactions, and case (d) is where both left and right waves are shock waves. The solution is a similarity solution $\mathbf{u}(x/t)$, that is $\mathbf{u}$ depends on the ratio x/t.

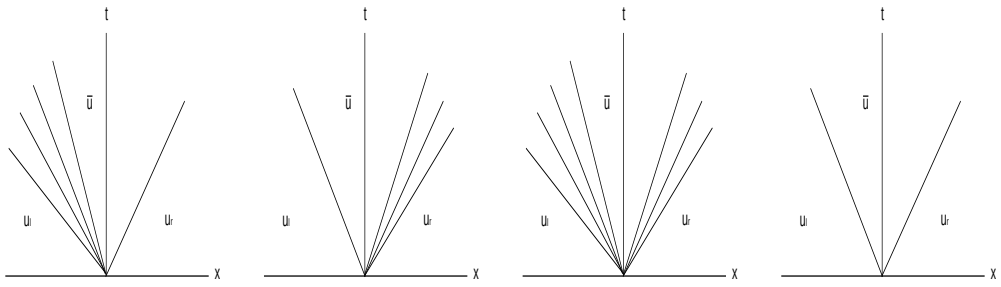


Figure B.2: *Possible wave patterns in the solution of the Riemann problem for the one-dimensional shallow water equations.*

In the rest of the subsection we study the much simpler case in which the initial data states for the Riemann problem are connected by a single wave, that is the solution of the Riemann problem consists of a single non-trivial wave.

Shock waves

Here we assumed that the solution of the Riemann problem consist of an isolated shock wave of speed s_i. The two constant data states $\mathbf{u}_l$ and $\mathbf{u}_r$ are connected through a single jump discontinuity in a genuinely non linear field i. The two states of the discontinuity must satisfy , from one side the *Rankine Hugoniot* conditions

$$\mathbf{f}(\mathbf{u}_r) - \mathbf{f}(\mathbf{u}_l) = s_i \left(\mathbf{u}_r - \mathbf{u}_l\right), \tag{B.47}$$

and the *entropy condition*

$$\lambda_i(\mathbf{u}_l) > s_i > \lambda_i(\mathbf{u}_r). \tag{B.48}$$

In figure B.3 we depict a shock wave of speed s_i. The characteristic direc-

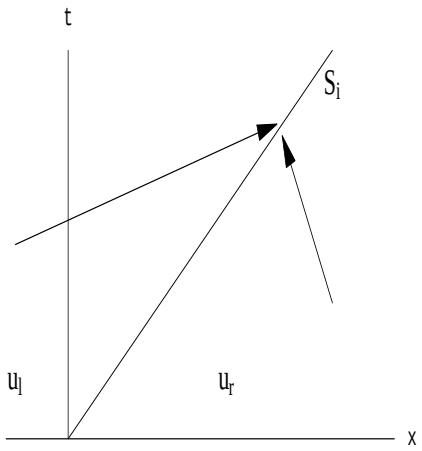


Figure B.3: *Shock wave solution of Riemann problem, in accordance with the entropy condition.*

tions $\frac{dx}{dt} = \lambda_i$ on both sides of the wave show the compressive character of the shock wave. Characteristics from both sides run into the shock wave, which is consistent with the physical condition (B.48). Hyperbolic conservation laws admit rarefaction shocks and compressive shocks as weak solutions. In our field of application it is the latter shocks which are physically acceptable, characteristics ahead and behind the shock wave run into the shock path (see [90] for more details).

Let us fix the state $\mathbf{u}_l = (h_l, q_l)$ and compute the possible states $\mathbf{u} = (h, q)$ that can be connected to $\mathbf{u}_l$ by a shock wave. In such a case both states must satisfy the Rankine-Hugoniot condition (B.47). So, we can

define the 1-shock curve (associated with the eigenvalue $\lambda_1 = v - \sqrt{gh}$) as:

$$v - v_l = -(h - h_l)\sqrt{\frac{g}{2}\frac{(h + h_l)}{hh_l}} \equiv S_1(h; \mathbf{u}_l). \tag{B.49}$$

By analogy, the 2-shock curve which is associated to the eigenvalue $\lambda_2 = v + \sqrt{gh}$, is defined as:

$$v - v_l = (h - h_l)\sqrt{\frac{g}{2}\frac{(h + h_l)}{hh_l}} \equiv S_2(h; \mathbf{u}_l). \tag{B.50}$$

In addition shocks must satisfy the entropy condition (B.48), based on this (see [90], [70] for more details), we obtain the second condition that must verify the right state $\mathbf{u}_r$ on the 1-shock wave (B.49):

$$h_r < h_l \tag{B.51}$$
$$v_r < v_l. \tag{B.52}$$

the same condition applied on the 2-shock wave (B.50) gives:

$$h_r > h_l \tag{B.53}$$
$$v_r < v_l. \tag{B.54}$$

Rarefaction waves

In this case, the two data states are connected through a smooth transition in a genuinely non-linear field i, say, via a rarefaction wave. In general, a centered rarefaction wave has a fan like structure, it is a smooth wave, all flow quantities vary continuously across the wave, at any fixed time. As in the shock waves, there are two families of rarefaction waves, each one of them corresponds to the characteristic family of the i eigenvector. In addition, an i-rarefaction has the property that i-Riemann invariants remain constant across the wave. A i-Riemann invariant is defined as a smooth function $w : \mathbb{R}^N \to \mathbb{R}$ that satisfies

$$\langle r_i(\mathbf{u}), \nabla w(\mathbf{u})\rangle = 0 \qquad \text{for any } \mathbf{u} \in \mathbb{R}^N. \tag{B.55}$$

The 1-Riemann invariant is then $v + 2\sqrt{gh}$ and the 2-Riemann invariant is $v - 2\sqrt{gh}$ which are constant in rarefaction waves. Moreover, the rarefactions waves satisfy the divergence of the characteristics

$$\lambda_i(\mathbf{u}_l) < \lambda_i(\mathbf{u}_r). \tag{B.56}$$

This condition says that the corresponding eigenvalue increases monotonically as the wave is crossed from left to right. Based on the two conditions above mentioned, constancy of generalized Riemann invariants across the wave and divergence of characteristics, the one parameter family of 1-rarefaction waves gives:

$$\begin{aligned} \sqrt{gh} &= \frac{1}{3}\left(2\sqrt{gh_l} + v_l - \frac{x}{t}\right) \\ v &= \frac{1}{3}\left(2\sqrt{gh_l} + v_l + 2\frac{x}{t}\right) \end{aligned} \tag{B.57}$$

thus, the set of states which can be connected to the right of $\mathbf{u}_l$ by a 1-rarefaction lie in the curve

$$v - v_l = 2\left(\sqrt{gh_l} - \sqrt{gh}\right) \equiv R_1(h; \mathbf{u}). \tag{B.58}$$

In the same manner, the family of 2-rarefaction waves is

$$\begin{aligned} \sqrt{gh} &= \frac{1}{3}\left(2\sqrt{gh_l} - v_l + \frac{x}{t}\right) \\ v &= \frac{1}{3}\left(-2\sqrt{gh_l} + v_l + 2\frac{x}{t}\right), \end{aligned} \tag{B.59}$$

now, the curve that defines the states which can be connected to $\mathbf{u}_l$ by a 2-rarefaction on the right is

$$v - v_l = -2\left(\sqrt{gh_l} - \sqrt{gh}\right) \equiv R_2(h; \mathbf{u}). \tag{B.60}$$

As well as i-rarefaction must satisfy (B.56), then if $\mathbf{u}_r$ is connected to $\mathbf{u}_l$ by a 1-rarefaction, it must verify

$$h_r < h_l \tag{B.61}$$

$$v_r > v_l. \tag{B.62}$$

In a similar way a 2-rarefaction must satisfy

$$h_r > h_l \tag{B.63}$$

$$v_r > v_l. \tag{B.64}$$

B.3.2
Solution of the Riemann problem

Let us consider the Riemann problem of the one dimensional shallow water equations. Let $(\mathbf{u}_l, \mathbf{u}_r)$ be two initial constant states. We can put both i-shocks and i-rarefaction curves together in the h-v plane, and observe that the plane is divided into four disjoint regions, as in figure (B.4). As it is proved by Smoller in [90], the curves R_1 and S_1 have a second-order contact at $\mathbf{u}_l$, as well as R_2 and S_2 have.

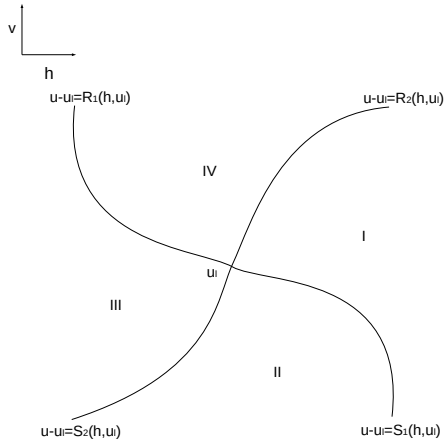


Figure B.4: *Integral curves for the state* $\mathbf{u}$.

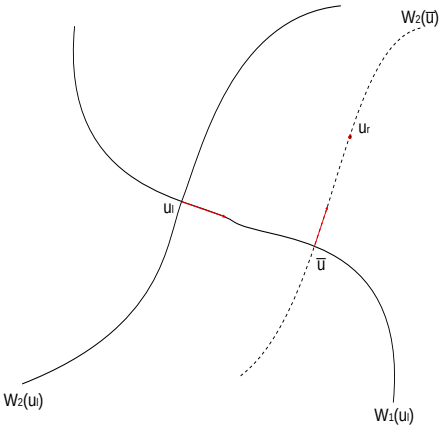


Figure B.5: *Sketch of the Riemann problem if* $\mathbf{u}_r$ *lies in region I.*

Let us consider the point $\mathbf{u}_l$ as fixed, and allow $\mathbf{u}_r$ to vary. If $\mathbf{u}$ lies on any of the four studied curves (B.49), (B.50), (B.58) or (B.60) then the solution is defined by one of these curves. On the other hand, $\mathbf{u}_r$ belongs to any of the open regions I, II, III or IV. Let us define, for $\bar{\mathbf{u}} \in \mathbb{R}^2$ the curves:

$$\bar{S}_i(\bar{\mathbf{u}}) = \{(h,v) : v = v_l + S_i(h,\bar{\mathbf{u}})\}, \qquad i = 1,2 \tag{B.65}$$

$$\bar{R}_i(\bar{\mathbf{u}}) = \{(h,v) : v = v_l + R_i(h,\bar{\mathbf{u}})\}, \qquad i = 1,2 \tag{B.66}$$

$$W_i(\bar{\mathbf{u}}) = \bar{S}_i(\bar{\mathbf{u}}) \cup \bar{R}_i(\bar{\mathbf{u}}). \tag{B.67}$$

For a fixed $\mathbf{u}_l \in \mathbb{R}^2$, we consider the family of curves

$$\mathfrak{F} = \{W_2(\bar{\mathbf{u}}) : \bar{\mathbf{u}} \in W_1(\mathbf{u_l})\}. \tag{B.68}$$

It is proved in [90] that the $h - v$ plane is covered univalently by the family of curves $\mathfrak{F}$, i.e., through each point $\mathbf{u}_r$ there passes exactly one curve of $\mathfrak{F}$. For the regions I, II and III the proof follows the same arguments as is done in [90]. However, in the region IV an additional restriction must be imposed in order to be covered by curves in $\mathfrak{F}$, this is called the *depth positivity condition* [101]:

$$v_r - v_l < 2\left(\sqrt{gh_l} + \sqrt{gh_r}\right).$$

Then the solution to the Riemann problem (B.45), (B.46) is given as follows: we connect $\bar{\mathbf{u}}$ to $\mathbf{u}_l$ on the right by a 1-shock or 1-rarefaction wave (backward wave), and $\mathbf{u}_r$ to $\bar{\mathbf{u}}$ on the right by a 2-shock or 2-rarefaction wave (forward wave).

For example, if $\mathbf{u}_r$ lies in the region I, there is a unique point $\bar{\mathbf{u}}$ for which the curve $W_2(\bar{\mathbf{u}})$ is in $\mathfrak{F}$ and passes through $\mathbf{u}_r$. Since $\bar{\mathbf{u}} \in \bar{S}_1(\mathbf{u}_l)$, $\bar{\mathbf{u}}$ is connected to $\mathbf{u}_l$ on the left by a back shock. Since $\mathbf{u}_r \in \bar{R}_2(\bar{\mathbf{u}})$, $\mathbf{u}_r$ is connected to $\bar{\mathbf{u}}$ on the right by a front rarefaction wave, see figure B.5.

Bibliography

[1] I. Ahmad and M. Berzins. MOL solvers for hyperbolic PDEs with source terms. *Math. Comput. Simulation*, 56(2):115–125, 2001. Method of lines (Athens, GA, 1999).

[2] D. Amadori, L. Gosse, and G. Guerra. Global BV entropy solutions and uniqueness for hyperbolic systems of balance laws. *Arch. Ration. Mech. Anal.*, 162(4):327–366, 2002.

[3] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25(2-3):151–167, 1997. Special issue on time integration (Amsterdam, 1996).

[4] E. Audusse, F. Bouchut, M. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004.

[5] A. Bermúdez and M. E. Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23:1049–1071, 1994.

[6] B. L. Bihari and A. Harten. Multiresolution schemes for the numerical solution of 2-D conservation laws. I. *SIAM J. Sci. Comput.*, 18(2):315–354, 1997.

[7] J. Burguete and P. García-Navarro. Efficient construction of high-resolution TVD conservative schemes for equations with source terms: application to shallow water flows. *Internat. J. Numer. Methods Fluids*, 37(2):209–248, 2001.

[8] K. Burrage and J.C. Butcher. Nonlinear stability of a general class of differential equation methods. *BIT*, 20:185–203, 1980.

[9] T. R. A. Bussing and E. M. Murman. Finite-volume method for the calculation of compressible chemically reacting flows. *AIAA J.*, 26(9):1070–1078, 1988.

[10] V. Caselles, R. Donat, and G. Haro. Flux gradient and source term balancing for certain high resolution shock capturing schemes. Submitted.

[11] M. J. Castro, J. A. García-Rodríguez, J. Macías, C. Parés, and M. E. Vázquez-Cendón. A two-layer numerical model for flows through channels with irregular geometry: application to the water exchange through the Strait of Gibraltar. In *Finite volumes for complex applications, III (Porquerolles, 2002)*, pages 457–464. Hermes Sci. Publ., Paris, 2002.

[12] M. J. Castro, J. M. González-Vida, J. Macias, M. L. Muñoz, C. Parés, J. García-Rodríguez, and M. E. Vázquez-Cendón. Numerical simulation of internal tides in the Strait of Gibraltar. *RACSAM Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat.*, 96(3):321–341, 2002. Mathematics and environment (Spanish) (Paris, 2002).

[13] M. J. Castro-Díaz, E. D. Fernández-Nieto, and A. M. Ferreiro. Some well-balanced shallow water-sediment transport models. In *Numerical mathematics and advanced applications*, pages 190–197. Springer, Berlin, 2006.

[14] L. Cea. *An unstructured finite volume model for unsteady turbulent shallow water flow with wet-dry fronts: Numerical solver and experimental validation*. PhD thesis, University of A Coruña, 2005.

[15] A. Chalabi. On convergence of numerical schemes for hyperbolic conservation laws with stiff source terms. *Math. Comput.*, 66(218):527–545, 1997.

[16] G. Chiavassa and R. Donat. Point value multiscale algorithms for 2D compressible flows. *SIAM J. Sci. Comput.*, 23(3):805–823, 2001.

[17] B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor. Advanced numerical approximation of nonlinear hyperbolic equations. *Lecture Notes in Mathematics*, 1697, 1998. Springer-Verlag, Berlin.

[18] A. Cohen, S. M. Kaber, S. Müller, and M. Postel. Fully adaptive multiresolution finite volume schemes for conservation laws. *Math. Comp.*, 72(241):183–225, 2003.

[19] P. Colella, A. Majda, and V. Roytburd. Theoretical and numerical structure for reacting shock waves. *SIAM J. Sci. Statist. Comput.*, 7(4):1059–1080, 1986.

[20] R. Courant, K. Friedrichs, and H. Lewy. On the partial difference equations of mathematical physics. *IBM J. Res. Develop.*, 11:215–234, 1967.

[21] N. Crnjaric-Zic, S. Vukovic, and L. Sopta. Extension of ENO and WENO schemes to one-dimensional sediment transport equations. *Computers & Fluids*, 26:31–56, 2004.

[22] M. Crouzeix. Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques. *Numer. Math.*, 35:257–276, 1980.

[23] K. Dekker and J. G. Verwer. Stability of Runge-Kutta methods for stiff nonlinear differential equations. *CWI Monographs, North-Holland*, 1984.

[24] A. I. Delis and Th. Katsaounis. Relaxation schemes for the shallow water equations. *Internat. J. Numer. Methods Fluids*, 41(7):695–719, 2003.

[25] M.O. Domingues, O. Roussel, and K. Schneider. Global time step control in adaptive multiresolution methods for pdes. *submitted to Internat. J. Numer. Meth. Engin.*, (2), 2007.

[26] R. Donat and A. Marquina. Capturing shock reflections: An improved flux formula. *J. Comput. Phys.*, 125:42–58, 1996.

[27] R. Donat and P. Mulet. The two-jacobian scheme for systems of conservation laws. In *Analysis and Simulation of Fluid Dynamics*, pages 89–108. Birkhäuser, 2007.

[28] P. Embid, J. Goodman, and A. Majda. Multiple steady states for 1-D transonic flow. *SIAM J. Sci. Statist. Comput.*, 5:21–41, 1984.

[29] B. Engquist and S. Osher. Stable and entropy satisfying approximations for transonic flow calculations. *Math. Comp.*, 34(149):45–75, 1980.

[30] J. M. Fe. *Aplicación del método de volúmenes finitos a la resolución numérica de las ecuaciones de aguas someras con incorporación de los esfuerzos debidos a la turbulencia*. PhD thesis, University of A Coruña, 2005.

[31] R. P. Fedkiw, B. Merriman, R. Donat, and S. Osher. The penultimate scheme for systems of conservation laws: finite difference ENO with Marquina's flux splitting. In *Innovative methods for numerical solutions of partial differential equations (Arcachon, 1998)*, pages 49–85. World Sci. Publ., River Edge, NJ, 2002.

[32] L. Ferracina and M. N. Spijker. Stepsize restrictions for the Total-Variation-Diminishing property in general Runge–Kutta methods. *SIAM J. Numer. Anal.*, 42(3):1073–1093, 2004.

[33] L. Ferracina and M. N. Spijker. An extension and analysis of the Shu-Osher representation of Runge-Kutta methods. *Math. Comp.*, 74(249):201–219, 2005.

[34] L. Ferracina and MN Spijker. Strong stability of singly-diagonally-implicit Runge–Kutta methods. *Applied Numerical Mathematics*, 2007.

[35] K. O. Friedrichs. On the derivation of the shallow water theory. Appendix to The formation of breakers and bores by J. J. Stoker. *Communications of Pure and Applied Mathematics*, 1:81–85, 1948.

[36] K. O. Friedrichs and P. D. Lax. Systems of conservation equations with a convex extension. *Proc Natl Acad Sci U S A*, 68(8):1686–1688, 1971.

[37] J. A. García Rodríguez. *Paralelización de esquemas de volúmenes finitos: aplicación a la resolución de sistemas de tipo aguas someras*. PhD thesis, University of Málaga, 2005.

[38] Ll. Gascón and J. M. Corberán. Construction of second-order TVD schemes for nonhomogeneous hyperbolic conservation laws. *J. Comput. Phys.*, 172(1):261–297, 2001.

[39] E. Godlewski and P. A. Raviart. *Numerical Approximation of Hyperbolic System of Conservation Laws*. Springer, 1996.

[40] S. K. Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb. (N.S.)*, 47 (89):271–306, 1959.

[41] S. K. Godunov. The problem of a generalized solution in the theory of quasi-linear equations and in gas dynamics. *Russ. Math. Surv.*, 17(3):145–156, 1962.

[42] L. Gosse. A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. *Comput. Math. Appl.*, 39(9-10):135–159, 2000.

[43] Laurent Gosse. A priori error estimate for a well-balanced scheme designed for inhomogeneous scalar conservation laws. *C. R. Acad. Sci. Paris Sér. I Math.*, 327(5):467–472, 1998.

[44] S. Gottlieb, W. C.Shu, and E. Tadmor. Strong stability preserving high-order time discretization methods. *SIAM Rev.*, 43(1)(TR-2000-15):89–112, 2001.

[45] N. Goutal and F. Maurel, editors. *Proceedings of the 2nd Workshop on Dam Break Wave Simulation.* EDF-DER Report HE-43/97/016/B, 42,45, 1997.

[46] J. M. Greenberg and A. Y. Leroux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.*, 33(1):1–16, 1996.

[47] J. M. Greenberg, A. Y. Leroux, R. Baraille, and A. Noussair. Analysis and approximation of conservation laws with source terms. *SIAM J. Numer. Anal.*, 34(5):1980–2007, 1997.

[48] G. Guerra. Well-posedness for a scalar conservation law with singular nonconservative source. *J. Differential Equations*, 206(2):438–469, 2004.

[49] A. Harten. Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. *Comm. Pure Appl. Math.*, 48(12):1305–1342, 1995.

[50] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 135(2):260–278, 1997.

[51] A. Harten, B. Engquist, S. Osher, and S. R. Chakrabarthy. Uniformly high order accurate essentially non-oscillatory schemes, iii. *J. Comput. Phys.*, 71:231–303, 1987.

[52] A. Harten, J. M. Hyman, and P. D. Lax. On finite-difference approximations and entropy conditions for shocks. *Comm. Pure Appl. Math.*, 29(3):297–322, 1976. With an appendix by B. Keyfitz.

[53] A. Harten and P. D. Lax. A random choice finite difference scheme for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 18(2):289–315, 1981.

[54] I. Higueras. On strong stability preserving time discretization methods. *Journal of Scientific Computing*, 21(2):193–223, 2004.

[55] I. Higueras. Representations of Runge–Kutta methods and strong stability preserving methods. *SIAM J. Numer. Anal.*, 43(3):924–948, 2005.

[56] I. Higueras. Strong stability for additive Runge-Kutta methods. *SIAM J. Numer. Anal.*, 44(4):1735–1758 (electronic), 2006.

[57] M. W. Hirsch and H. Smith. Monotone Dynamical Systems. In *Handbook of differential equations: ordinary differential equations. Vol. II*, pages 239–357. Elsevier B. V., Amsterdam, 2005.

[58] M. E. Hubbard and P. Garcia-Navarro. Flux difference splitting and the balancing of source terms and flux gradients. *J. Comput. Phys.*, 165(1):89–125, 2000.

[59] W. Hundsdorfer and J. G. Verwer. *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer, 2003.

[60] W. Hundsdorfer, J. G. Verwer, and R. J. Spiteri. Monotonicity-preserving linear multistep methods. *SIAM J. Numer. Anal.*, 41:605–623, 2003.

[61] P. Jenny and B. Müller. Rankine-Hugoniot-Riemann solver considering source terms and multidimensional effects. *J. Comput. Phys.*, 145(2):575–610, 1998.

[62] E. Kamke. Zur Theorie der Systeme gewöhnlicher Differentialgleichungen. II. *Acta Math.*, 58(1):57–85, 1932.

[63] G. E. Karniadakis, M. Isracli, and S. A. Orszag. High-order splitting methods for the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 97:414–443, 1991.

[64] J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta methods. *BIT*, 31(3):482–528, 1991.

[65] S. N. Kruzkov. First order quasi-linear equations in several independent variables. *Math. USSR-Sb.*, 10:217–243, 1970.

[66] P. Lamby, S. Müller, and Y. Stiriba. Solution of shallow water equations using fully adaptive multiscale schemes. *Internat. J. Numer. Methods Fluids*, 49(4):417–437, 2005.

[67] J. O. Langseth, A. Tveito, and R. Winther. On the convergence of operator splitting applied to conservation laws with source terms. *SIAM J. Numer. Anal.*, 33(3):843–863, 1996.

[68] P. D. Lax. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, volume 11 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, U.S.A., 1973.

[69] P. D. Lax and B. Wendroff. Systems of conservation laws. *Commun. Pure Appl. Math.*, 13:217–237, 1960.

[70] R. J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser, 1992.

[71] R. J. LeVeque. Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *J. Comput. Phys.*, 146(1):346–365, 1998.

[72] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[73] R. J. LeVeque and H. C. Yee. A study of numerical methods for hyperbolic conservation laws with stiff source terms. *J. Comput. Phys.*, 86(1):187–210, 1990.

[74] C. Mascia and C. Sinestrari. The perturbed Riemann problem for a balance law. *Adv. Differential Equations*, 2(5):779–810, 1997.

[75] C. Mascia and A. Terracina. Large-time behavior for conservation laws with source in a bounded domain. *J. Differential Equations*, 159(2):485–514, 1999.

[76] M. Müller. Über das Fundamentaltheorem in der Theorie der gewöhnlichen Differentialgleichungen. *Math. Z.*, 26(1):619–645, 1927.

[77] S. Osher. Riemann solvers, the entropy condition, and difference approximations. *SIAM J. Numer. Anal.*, 21(2):217–235, 1984.

[78] C. Parés and M. Castro. On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems. *M2AN Math. Model. Numer. Anal.*, 38(5):821–852, 2004.

[79] L. Pareschi and G. Russo. Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25(1-2):129–155, 2005.

[80] P. L. Roe. Generalized formulation of TVD Lax-Wendroff schemes. *ICASE*, 84:53, 1984.

[81] P. L. Roe. Upwind differencing schemes for hyperbolic conservation laws with source terms. In *Nonlinear hyperbolic problems (St. Etienne, 1986)*, volume 1270 of *Lecture Notes in Math.*, pages 41–51. Springer, Berlin, 1987.

[82] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 135(2):250–258, 1997.

[83] O. Roussel, K. Schneider, A. Tsigulin, and H. Bockhorn. A conservative fully adaptive multiresolution algorithm for parabolic PDEs. *J. Comput. Phys.*, 188(2):493–523, 2003.

[84] S. Ruuth. Implicit-explicit methods for resaction-diffusion problems in pattern formation. *J. Math. Biology*, 34(2):148–176, 1995.

[85] S. J. Ruuth and R. J. Spiteri. Two barriers on strong-stability-preserving time discretization methods. *J. Sci. Comput.*, 17(1):211–220, 2002.

[86] C.W. Shu. Total-Variation-Diminishing time discretizations. *SIAM J. Sci. Statist. Comput.*, 9:1073–1084, 1988.

[87] C.W. Shu and S. Osher. Efficient implementation of essentially nonoscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439–471, 1988.

[88] C.W. Shu and S. Osher. Efficient implementation of essentially nonoscillatory shock-capturing schemes. II. *J. Comput. Phys.*, 83(1):32–78, 1989.

[89] B. Sjögreen. Numerical experiments with the multiresolution scheme for the compressible Euler equations. *J. Comput. Phys.*, 117:251–261, 1995.

[90] J. Smoller. *Shock Waves and Reaction-Diffusion Equations*. Springer-Verlag, 1994.

[91] M. N. Spijker. Monotonicity and boundedness in implicit Runge–Kutta methods. *Numer. Math.*, 50:97–109, 1986.

[92] R. J. Spiteri and S. J. Ruuth. A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM J. Numer. Anal.*, 40(2):469–491, 2002.

[93] J. J. Stoker. The formation of breakers and bores (the theory of nonlinear wave propagation in shallow water and open channels). *Comm. Pure Appl. Math.*, 1:1–87, 1948.

[94] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[95] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21(5):995–1011, 1984.

[96] E. Tadmor. Entropy functions for symmetric systems of conservation laws. *J. Math. Anal. Appl.*, 122:355–359, 1987.

[97] W. Y. Tan. *Shallow Water Hydrodynamics*. Water and Power Press, Beijing, 1992.

[98] T. Tang. Convergence analysis for operator-splitting methods applied to conservation laws with stiff source terms. *SIAM J. Numer. Anal.*, 35(5):1939–1968 (electronic), 1998.

[99] T. Tang and Z. H. Teng. Error bounds for fractional step methods for conservation laws with source terms. *SIAM J. Numer. Anal.*, 32(1):110–127, 1995.

[100] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics – a practical introduction*. Springer, Berlin, Germany, 1997.

[101] E. F. Toro. *Shock-Capturing Methods for Free-Surface Shallow Flows*. Wiley and Sons, LTD, 2001.

[102] B. van Leer. Towards the ultimate conservative difference scheme v. a second-order sequel to Godunov's method. *J. Comput. Phys.*, 135(2):229–248, 1997.

[103] J. M. Varah. Stability restrictions on second order, three level finite difference schemes for parabolic equations. *SIAM J. Numer. Anal.*, 17(2):300–309, 1980.

[104] M. E. Vázquez-Cendón. Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. *J. Comput. Phys.*, 148:497–526, 1999.

[105] J. G. Verwer, J. G. Blom, and W. Hundsdorfer. An implicit-explicit approach for atmospheric transport-chemistry problems. *Appl. Numer. Math.*, 20(1-2):191–209, 1996. Workshop on the method of lines for time-dependent problems (Lexington, KY, 1995).

[106] S. Vukovic and L. Sopta. ENO and WENO schemes with the exact conservation property for one-dimensional shallow water equations. *J. Comput. Phys.*, 179(2):593–621, 2002.

[107] R. F. Warming, P. Kutler, and H. Lomax. Second- and third-order noncentered difference schemes for nonlinear hyperbolic equations. *AIAA J.*, 11:189–196, 1973.

[108] G. B. Whitham. *Linear and Nonlinear Waves*. John Wiley & Sons, New York, 1974.

[109] Y. Xing and C.W. Shu. High order finite difference WENO schemes with the exact conservation property for the shallow water equations. *J. Comput. Phys.*, 208(1):206–227, 2005.

[110] H. C. Yee and J. Shinn. Semi-implicit and fully implicit shcok-capturing methods for hyperbolic conservation laws with stiff source terms. *AIAA*, 87-1116, 1987.