Sensitivity analysis for nonlinear hyperbolic equations of conservation laws

Camilla Fiorini

NNT : 2018SACLV034

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université de Versailles Saint-Quentin-en-Yvelines

Laboratoire d'accueil : Laboratoire de mathématiques de Versailles, UMR 8100 CNRS

Spécialité de doctorat : Mathématiques appliquées

Camilla FIORINI

Sensitivity analysis for nonlinear hyperbolic systems of conservation laws

Date de soutenance : 11 juillet 2018

Après avis des rapporteurs : Frédéric LAGOUTIÈRE (Université Claude Bernard Lyon I)
Jeff BORGGAARD (Virginia Tech)

Jury de soutenance :

Edwige GODLEWSKI	(Sorbonne Université)	Présidente
Anca BELME	(Sorbonne Université)	Examinatrice
Christophe CHALONS	(UVSQ)	Directeur de thèse
Carole DELENNE	(Université Montpellier)	Examinatrice
Régis DUVIGNEAU	(Inria)	Directeur de thèse
Pierre GABRIEL	(UVSQ)	Examinateur
Jean-Marc HÉRARD	(EDF R&D)	Examinateur
Frédéric LAGOUTIÈRE	(UCBL)	Rapporteur

Remerciements

After a roller coaster of almost four years, here I am, writing these acknowledgements, which will be a mixture of different languages and a lot of emotions.

D'abord, je tiens à remercier mes deux directeurs de thèse, je n'aurais pas pu demander mieux. Un grand merci à tous les deux pour m'avoir acceptée en thèse à mi-parcours. Régis, merci parce que même à distance tu as été très présent et parce qu'à chaque visite à Sophia ma thèse a énormément avancé. Christophe, merci d'avoir toujours trouvé du temps pour moi, même quand je me présentais à ton bureau sans rendez-vous. Ça a été un plaisir de travailler avec vous deux.

I want to thank Jeff Borggaard for his careful reading of my manuscript and for all his suggestions, english- and maths-wise. Je veux aussi remercier Frédéric Lagoutière pour son rapport très détaillé et l'intérêt porté à mon travail. Merci à Anca Belme, Carole Delenne, Pierre Gabriel, Edwige Godlewski et Jean-Marc Hérard pour avoir accepté de faire partie du jury.

Je tiens à remercier tout le laboratoire de mathématiques de Versailles et l'équipe ACUMES de l'Inria pour m'avoir accueillie. En particulier, merci Vincent pour ton aide et tes conseils dans un moment compliqué comme le changement de directeur de thèse, sans ton soutien j'aurais probablement abandonné; un grand merci aussi à Catherine et Luc pour leur aide dans cette situation difficile. Merci Abdou pour ta disponibilité; Pierre, pour ta gentillesse et l'aide avec le MA100bio; Otared, pour toutes tes blagues, tu es la découverte du CANUM. Merci aussi à Nadège, Alexis, Paola et Laurent. Enfin, merci Laure pour ton efficacité, mais surtout pour nos discussions toujours très agréables.

Je veux remercier tous les doctorants, postdocs et ATER avec qui j'ai partagé un bout de chemin pendant ces quatre ans.

Merci Benjamin pour ton aide au tout début, Salim pour ta capacité d'expliquer simplement les choses plus difficiles, Hélène pour ta bonne humeur, Anne Charlotte pour tes connaissances au quizz même si on n'a jamais réussi à gagner, Thomas pour nos discussions au bureau et dans le train.

Grazie Mattia per essere il mio nerd di fiducia e quasi gemello di tesi. Continuo a stupirmi di quanti interessi (molti dei quali assurdi, ammettiamolo...) abbiamo in comune.

Merci Patricio Patricio pour ton humour, pour tous les "oulà, c'est compliqué", les

“soyons fous ”, les festivals du cinéma équatorien et plein d’autres choses qu’il est mieux de ne pas écrire dans des remerciements de thèse. Enfin, voilà quoi...

Merci Pierrick pour toutes nos bières et nos discussions, toujours après 1h du matin.

Merci Victor pour toutes les conférences passées ensemble, pour avoir toujours gardé mes secrets (enfin, j’espère !), pour être devenu un très bon ami même à distance. Je promets que je vais venir à Toulouse un jour ou l’autre !

Merci Maxime, parler avec toi pendant cette dernière année m’a permis de comprendre énormément de choses. Merci pour tous les discours sérieux et également pour toutes les blagues, mais surtout je pense que je ne te remercierai jamais assez de m’avoir invitée au ski ! Merci aussi à tous ceux qui sont venus à Corvara, à Nat en particulier pour ses bons conseils.

Merci à mes super colocs Ilaria et Sybille, pour nos longues discussions autour de la petite table dans la cuisine, pour les soirées organisées, pour les bons (et mauvais) conseils qu’on a échangés. Merci Sybille pour ta patience et ta capacité d’écoute. Grazie Ila per il sostegno, le risate e per essere l’unica a capire questo strano misto di italiano e francese in cui abbiamo iniziato a parlare. Merci PE parce que quand tu es là on est sûr de ne pas s’ennuyer.

Merci Florian et Seb d’avoir amené de la vie au labo, même si brièvement. Cette année ça m’a beaucoup manqué de venir dans votre bureau à chaque fois que quelque chose ne marchait pas. Florian, merci parce que tu es une des personnes les plus disponibles que je connais et tu comprends toujours tout sans avoir besoin de t’expliquer. Seb, en italien je dirais “se non ci fossi bisognerebbe inventarti”, si tu n’existais pas il faudrait t’inventer. Merci pour ton côté direct et ta positivité.

Antoine, j’ai tellement de choses pour lesquelles je voudrais te remercier que je ne sais pas d’où partir. Merci pour m’avoir appris le français, pour avoir rendu agréable le temps passé au bureau, pour tous les “ça veut dire ce que ça veut dire”, mais surtout merci d’avoir été là pendant les moments difficiles de ma thèse, toujours avec une petite blague pour me faire rigoler. I couldn’t ask for a better F·R·I·E·N·D (c’est de l’anglais !).

Un grazie agli amici di sempre, soprattutto Maui e Robi perché farsi Milano-Versailles in giornata non è da tutti e ve ne sono immensamente grata. Grazie anche a chi non ha potuto essere qui oggi: Giò, Marti, Eli e Gue. Grazie a tutti gli amici riminesi e sammarinesi, in particolare al Conte, che è venuto a trovarmi fino a Parigi, e a Elia perché anche se ci si vede una volta all’anno quando siamo insieme è come se ci fossimo visti tutti i giorni. Soniet, grazie, per tutti i messaggi, le chiamate su Skype, i sempre troppo corti weekend a Milano e per essere sempre presente nella mia vita.

Il ringraziamento più grande per questi quattro anni va senza dubbio a Stefano, che ha assecondato ogni mio dubbio e ogni sbalzo d’umore senza mai farlo pesare. Grazie per esser stato il mio punto fermo nei momenti più duri e in quelli più belli e per aver dato sempre buoni consigli.

Un grande ringraziamento alla mia famiglia. Grazie ai miei zii, lontani ma presenti. Grazie nonni per tutto quello che fate per me, sempre senza esitazioni, non so come farei senza di voi.

Mamma e papà, so che questo non era quello che vi aspettavate quando mi avete spinta a fare un’esperienza all’estero ormai quattro anni fa. Grazie perché siete sempre

una certezza, in due modi completamente diversi, e so che su di voi potrò sempre contare. Spero che sappiate che è reciproco, anche se sono lontana. Tutto quello che ho raggiunto l'ho ottenuto solo grazie a voi e a come mi avete cresciuta.

Julien, merci pour t'être occupé de moi pendant la rédaction, pour avoir toujours réussi à me faire rire, pour m'avoir supportée en cette période de stress et conférences sans jamais te plaindre. Je ne sais pas comment j'aurais survécu sans toi ces derniers mois.

Young man, in mathematics
you don't understand things.
You just get used to them.
— John Von Neumann.

Reply to a physicist friend who had said:
"I'm afraid I don't understand
the method of characteristics."

Contents

1 Introduction

1.1 Context and positioning

This PhD thesis deals with sensitivity analysis (SA) for nonlinear hyperbolic partial differential equations (PDEs). These two topics separately are very well known and have been studied thoroughly: however, when considered together, many problems arise and the literature on this subject is far from being comprehensive.

1.1.1 Sensitivity analysis

SA is the study of how changes in the inputs of a model affect the outputs. The sensitivity itself is defined as the derivative of the state, i.e. the solution to the PDE model considered, with respect to a parameter of interest. SA is obviously a valuable tool for engineering applications, since it allows the quantification of changes in the physical response of a system to any change of parameter values: one of its straightforward application in this direction is uncertainty quantification (UQ). SA is an efficient and deterministic method to estimate expectation and standard deviation of the state variables, as an uncertainty propagation technique. Another application of SA is optimization: sensitivities can be used to compute the gradient of a cost functional. Finally, SA methods can likewise be employed to monitor and explore interactively neighbouring solutions for a negligible computational expense and provide an answer to the question "what if...".

There are two main classes of methods to compute the sensitivities: the *discretise-then-differentiate* approach and the *differentiate-then-discretise* one. The first approach consists of, as the name says, first discretising the state PDE system and then differentiating it to get a numerical approximation of the sensitivities. In the second approach, one differentiates the state system obtaining in this way the sensitivity system, which can then be discretised. The approximated sensitivities obtained with these approaches are different, because the differentiation step and the discretisation step do not commute in general. A popular strategy that falls into the discretise-then-differentiate category is automatic differentiation [HMB98]: a first advantage of this is the simplification of the code that has to be written for the sensitivity; a second one is the fact that the resulting sensitivity is consistent with the discrete state solution, even if a coarser grid is used for the resolution. However, an important drawback of all the discretise-then-differentiate approaches is the need to differentiate all the computational facilitators, such as, for instance, slope limiters in a finite volume framework or even MPI communication in parallel computing. In [MD10] for instance, they quantify the error due to automatic differentiation applied to classical slope limiters for the 3D Euler system. On the other

hand, the differentiate-then-discretise approach provides a sensitivity system, for which computational facilitators can be used, if needed. Both strategies are valid and they are suitable for different applications. A detailed comparison between the two for optimization problems is performed in [Gun03]. In this work, we will mainly focus on the differentiate-then-discretise approach.

An alternative to SA which is worth mentioning, especially for optimization, is the adjoint equation method [Jam88, MP01, Pir74], which introduces additional adjoint variables to compute the derivative of any functional output with respect to all input parameters. Note that for the adjoint method, too, one needs to choose whether to use a differentiate-then-discretise or a discretise-then-differentiate approach. The adjoint equation is independent of the input parameters, thus this approach is very efficient for optimization problems involving a large number of design parameters, as opposed to the SA approach, which requires the solution of a different sensitivity system for each parameter. However, if the PDEs considered are time-dependent, the adjoint equation should be solved backwards in time, which could lead to practical difficulties. Moreover, SA allows the computation of the derivative of the whole state and it is not restricted to functionals. Finally, we remark that the sensitivity systems are all independent of each other and therefore can be solved in parallel.

Therefore, in this work we focused our attention on the continuous sensitivity equation (CSE) method [BB97, DPB06, DP06, HEPB04], which allows to compute the derivative of the PDE solution itself, at any location and time, with respect to a single input parameter: this is done by formally differentiating the state system with respect to the parameter of interest, and then exchanging the derivatives with respect to the parameter with those in space and time, obtaining in such a way a new system of PDEs, the sensitivity system. Of course, this can be done for as many parameters as needed. This approach is a differentiate-then-discretise approach and it relies on a forward time integration. However, a certain regularity of the state solution is required, a condition that is not always met in the hyperbolic framework.

1.1.2 Hyperbolic equations

A one-dimensional conservation law is an equation of the following form:

$$\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = 0,$$

where ∂_t and ∂_x indicate the partial derivatives with respect to time and space, respectively, $\mathbf{U}$ is the conserved variable, or state, and $\mathbf{F}$ is the flux. It expresses the fact that, on an arbitrary domain, the time variation of the state variable $\mathbf{U}$ are equal to the flux of $\mathbf{F}$ through the boundary of the domain considered. These kinds of equations often result from the modelling of continuum physics in cases where the dissipation effects are negligible. A conservation law is said to be *strictly hyperbolic* if the Jacobian matrix of the flux $\mathbf{A}(\mathbf{U}) = \dfrac{\partial \mathbf{F}}{\partial \mathbf{U}}$ is $\mathbb{R}-$diagonalisable, whilst if the matrix $\mathbf{A}(\mathbf{U})$ has real eigenvalues but it is not $\mathbb{R}-$diagonalisable the system is *weakly hyperbolic*, and this distinction will be important in the next chapters.

Hyperbolic equations have been studied for a very long time, both from theoretical and numerical points of view and many books have been written on the subject [Tor09,

GR96]. For the scalar conservation laws, a technique named the method of characteristics can be used to compute an analytical solution to the equation, although it provides often an implicit solution, depending on the initial data. For systems, analytical solutions are known only for certain specific initial conditions, for instance for what it is usually called the Riemann problem: an initial value problem with a piecewise constant data presenting only one discontinuity. These problems play a crucial role in the numerical approximation of hyperbolic equations with a finite volume (FV) approach, which is, as of today, the most suitable approach for these kinds of equations thanks to its conservation property. In particular, the first step of Godunov-type method is the solution of a Riemann problem: this can be done exactly, for the problems for which the analytical solution is known, or using a so-called approximate Riemann solver. Many different approximate Riemann solvers have been developed, some based on a simplification of the structure of the solution, such as for instance the HLL solver, some based on a linearisation of the system, like for example the Roe solver.

Depending on the initial and boundary conditions, hyperbolic PDEs can present discontinuities, such as shock waves or contact discontinuities. This common feature in the hyperbolic framework is the main reason why SA in the hyperbolic equations framework is not straightforward. Indeed, such discontinuities in the solution lead to specific issues regarding SA, because they correspond to the presence of Dirac delta functions in the sensitivity fields: the CSE method briefly introduced in the previous section makes sense only under certain assumptions of regularity of the state solution, which may not be verified in the hyperbolic case. This question has been explored in [BP02, MP01] with a theoretical viewpoint, and more recently in [Gui09, GDC09] with a numerical viewpoint, where a modification of the sensitivity system was proposed, to "remove" the spikes from the numerical sensitivity solution. More specifically, a modification of HLL Riemann solver used to evaluate fluxes in a FV method was proposed in the context of Saint-Venant equations. The correction is based on the Rankine-Hugoniot conditions, which govern the state across a shock.

1.2 Thesis synopsis

In this work we deal with a hierarchy of models of increasing complexity. We start in Chapter 2 from the simplest example of nonlinear hyperbolic equation, the inviscid Burgers' equation, before focusing on one of the most known and important examples of nonlinear hyperbolic PDE system: the Euler system. The Euler system models the dynamics of a compressible material, which can be a liquid or a gas, and three physical variables are considered to describe the flow: the density of the fluid, its velocity and its pressure. At first, in Chapter 3, we deal with a version of this system, known as the $p-$system, which is simplified under certain physical and mathematical hypothesis: the fluid is assumed to be in barotropic conditions and Lagrangian coordinates are used; then, in Chapter 4 and Chapter 5 we deal with the complete Euler system: in particular, in Chapter 4 the Sod shock tube is studied, along with the corresponding problem for the sensitivity, and in Chapter 5 a non trivial topography and stationary solutions are considered.

As mentioned in the previous section, if standard techniques of SA, such as the CSE method, are used everywhere, even in presence of discontinuities, they provide a

sensitivity which contains Dirac delta functions. Since a Dirac delta function cannot be seized numerically, this leads to a spike which is spread in the neighbourhood of the shock and which deteriorates the solution. Moreover, the spikes can change with the numerical discretisation. The use of such sensitivities is impractical for many of the applications mentioned above. Therefore, along the same lines as what is done in [Gui09, GDC09], we modify the sensitivity equation in order to obtain a system which is valid also in the case of discontinuous solutions. In this work, we adopt a similar, though slightly different, point of view: we suggest to add a correction in the form of a source term to balance out the spikes. The source term is non-zero only in case of shock or contact discontinuity and has an amplitude which is proportional to the jump of the state across the discontinuity considered. The computation of the amplitude of the correction term is carried out in details in the next chapters and it is based on an integration of the equations over a control volume containing a discontinuity. Regarding the discretisation of such a source term, it is not straightforward: we remark that a shock detector is needed, in order for the term to be consistent and to avoid numerical overcorrection.

Concerning the discretisation of the conservation law, the state problem is well-known and all sorts of numerical schemes have been developed throughout the years. It is not the case for the sensitivity equations. Moreover, considering the system as a whole does not bring an advantage since the system composed by the state and the sensitivity together is only weakly hyperbolic in most cases. For these reasons, our strategy from a numerical point of view is the following: first, the state is solved using a classical FV scheme; secondly, the source term, which depends on the state, is computed; finally an adapted FV numerical scheme is applied to the sensitivity problem. This approach works well if the state equation is a scalar equation, however numerical results show that for systems some additional precaution is necessary: in particular, numerical diffusion plays a fundamental role and damages the convergence to the exact solution for the sensitivity. Hence, in this work we adapt an anti-diffusive (AD) Godunov-type numerical scheme, first introduced in [CG08], and we propose AD versions of different FV schemes, of different order in time and in space.

Finally, we explore two different applications of SA: optimization and uncertainty quantification. These applications deal with several input parameters. In Chapter 4 we perform a UQ analysis for a Riemann problem on the complete Euler system: different SA approaches are compared to the more expensive Monte Carlo one. In Chapter 5 an optimization problem is considered: in particular, we deal with a pressure matching problem, where the optimization parameters considered directly modify the topography. The aim of these applications is to understand the importance of the numerical diffusion and of the correction term in more realistic situations and to derive practical guidelines for SA.

During my first year of PhD I worked on the modelling and optimization of running strategies, under the supervision of Amandine Aftalion, who was my advisor. This work resulted in a published paper [Fio17], which is reported in Appendix A.

Introduction

1.3 Contexte et positionnement

Cette thèse porte sur l'analyse de sensibilité (AS) pour les équations aux dérivées partielles (EDP) hyperboliques non-linéaires. Les deux sujets séparément sont très bien connus et ont fait l'objet de plusieurs études approfondies : cependant, lorsqu'ils sont considérés ensemble, de nombreux problèmes se posent et la littérature sur ce sujet n'est pas importante.

1.3.1 Analyse de sensibilité

L'AS est l'étude de la façon dont les changements dans les entrées d'un modèle affectent la sortie. La sensibilité elle-même est définie comme la dérivée de l'état, c'est-à-dire la solution du modèle d'EDP considéré, par rapport à un paramètre d'intérêt. L'AS est évidemment un outil important pour des applications d'ingénierie, car elle permet de quantifier la réponse physique d'un système aux changements de valeurs des paramètres : une application directe dans ce sens est la quantification d'incertitude (UQ). L'AS est une méthode efficace et déterministe pour estimer l'espérance et l'écart-type des variables d'état, en tant que technique de propagation d'incertitude. Une autre application de l'AS est l'optimisation : les sensibilités peuvent être utilisées pour calculer le gradient d'une fonctionnel coût. Enfin, les méthodes d'AS peuvent également être utilisées pour l'exploration interactive des solutions voisines avec un coût de calcul négligeable et fournir une réponse à la question "*que se passe-t-il si ...*".

Il y a deux classes principales de méthodes pour calculer les sensibilités : l'approche *discrétiser puis différencier* et l'approche *différencier puis discrétiser*. La première approche consiste, comme son nom l'indique, à discrétiser d'abord le système d'EDP d'état, puis à le différencier pour obtenir une approximation numérique des sensibilités. Dans la seconde approche, on différencie le système d'état obtenant ainsi le système de sensibilité, qui peut ensuite être discrétisé. Les sensibilités approchées obtenues avec ces deux approches sont différentes, car l'étape de différenciation et celle de discrétisation ne commutent pas en général. La différenciation automatique [HMB98] est une stratégie couramment employée qui entre dans la catégorie discrétiser puis différencier : un premier avantage est la simplification du code qui doit être écrit pour la sensibilité ; un second est le fait que la sensibilité obtenue est consistante avec la solution discrète de l'état, même si un maillage plus grossier est utilisé pour la résolution. Cependant, un inconvénient important de toutes les approches discrétiser puis différencier est la nécessité de différencier tous les facilitateurs de calcul, comme, par exemple, les limiteurs de pente

dans un cadre volumes finis ou même la communication MPI en calcul parallèle. Dans [MD10] par exemple, ils quantifient l'erreur due à la différentiation automatique appliquée à des limiteurs de pente classiques dans le cadre des équations d'Euler 3D. D'autre part, l'approche différencier puis discrétiser fournit un système de sensibilité pour lequel on peut, si nécessaire, utiliser des facilitateurs de calcul spécifiques. Les deux stratégies sont valables et adaptées à différentes applications. Une comparaison détaillée des deux pour les problèmes d'optimisation se trouve dans [Gun03]. Dans ce travail, nous nous concentrerons principalement sur l'approche différencier puis discrétiser.

Une alternative à l'AS qui mérite d'être mentionnée, spécialement pour l'optimisation, est la méthode de l'équation adjointe [Jam88, MP01, Pir74], qui introduit des variables supplémentaires, dites adjointes, afin de calculer la dérivée d'une fonctionnelle par rapport à tous les paramètres d'entrée. On note que pour la méthode adjointe aussi il existe les deux approches : différencier puis discrétiser ou discrétiser puis différencier. L'équation adjointe est indépendante du nombre de paramètres d'entrée, ce qui rend cette approche très efficace pour les problèmes d'optimisation avec beaucoup de paramètres, plutôt que l'approche AS, qui nécessite de résoudre un système de sensibilité différent pour chaque paramètre. Cependant, si les EDP considérées dépendent du temps, l'équation adjointe doit être résolue à rebours en temps, ce qui pourrait poser des difficultés pratiques. De plus, l'AS permet de calculer la dérivée de l'état et n'est pas limitée aux fonctionnelles. Enfin, nous remarquons que les systèmes de sensibilité sont tous indépendants entre eux et peuvent donc être résolus en parallèle.

Par conséquent, dans ce travail nous nous sommes concentrés sur la méthode de léquation de sensibilité continue [BB97, DPB06, DP06, HEPB04], qui permet de calculer la dérivée de la solution de l'EDP considérée, à n'importe quel endroit et temps, par rapport à un paramètre d'entrée : cela se fait en différenciant formellement le système d'état par rapport au paramètre d'intérêt, puis en échangeant les dérivées par rapport au paramètre avec celles en espace et en temps, obtenant ainsi un nouveau système d'EDP, le système de sensibilité. Bien sûr, cela peut être fait pour autant de paramètres que nécessaire. Cette approche est une approche différencier puis discrétiser et est basée sur une intégration avec avance en temps. Cependant, une certaine régularité de la solution d'état est nécessaire, condition qui n'est pas toujours vérifiée dans le cadre hyperbolique.

1.3.2 Équations hyperboliques

Une lois de conservation 1D est une équation de la forme suivante :

$$\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = 0,$$

où ∂_t et ∂_x désignent les dérivées partielles par rapport au temps et à l'espace, respectivement, $\mathbf{U}$ est la variable conservée, ou état, et $\mathbf{F}$ est le flux. Elle exprime le fait que, sur un domaine arbitraire, la variation temporelle de la variable d'état $\mathbf{U}$ est égale au flux de $\mathbf{F}$ à travers la frontière du domaine considéré. Ces équations proviennent souvent de la modélisation de la physique des milieux continus dans les cas où les effets dissipatifs sont négligeables. Une loi de conservation se dit *strictement hyperbolique* si la matrice jacobienne du flux $\mathbf{A}(\mathbf{U}) = \dfrac{\partial \mathbf{F}}{\partial \mathbf{U}}$ est $\mathbb{R}-$ diagonalisable, alors que si la matrice $\mathbf{A}(\mathbf{U})$ a des

valeurs propres réelles mais qu'elle n'est pas $\mathbb{R}-$ diagonalisable, le système est *faiblement hyperbolique*, et cette distinction sera importante dans la suite.

Les équations hyperboliques ont été étudiées depuis très longtemps, à la fois d'un point de vue théorique et numérique et de nombreux livres ont été écrits sur le sujet [Tor09, GR96]. Pour les lois de conservation scalaires, une technique appelée méthode des caractéristiques peut être utilisée pour calculer une solution analytique de l'équation, bien qu'elle ne fournisse souvent qu'une solution implicite, selon les données initiales. Pour les systèmes, les solutions analytiques ne sont connues que pour certaines conditions initiales spécifiques, comme par exemple pour ce que l'on appelle un problème de Riemann : un problème aux valeurs initiales avec une donnée constante par morceaux qui présente une seule discontinuité. Ces problèmes jouent un rôle fondamental dans l'approximation numérique des équations hyperboliques avec une approche aux volumes finis, qui reste, à ce jour, l'approche la plus appropriée pour ces équations grâce à ses propriétés de conservation. En particulier, la première étape de la méthode de Godunov est la résolution d'un problème de Riemann : cela peut être fait exactement, pour les problèmes pour lesquels la solution analytique est connue, ou en utilisant un solveur de Riemann approché. De nombreux solveurs de Riemann approchés ont été développés, certains basés sur une simplification de la structure de la solution, comme par exemple le solveur HLL, d'autres basés sur une linéarisation du système, comme par exemple le solveur de Roe.

Selon les conditions initiales et aux bords, les EDP hyperboliques peuvent présenter des discontinuités, comme des chocs ou des discontinuités de contact. Cette caractéristique, très commune dans le cadre hyperbolique, est la raison principale pour laquelle l'AS dans le cadre des équations hyperboliques n'est pas simple. En effet, des discontinuités dans l'état causent des problématiques spécifiques en AS, car elles correspondent à la présence de distributions de Dirac dans les sensibilités : la méthode de léquation de sensibilité introduite brièvement ci-dessus ne fonctionne que sous certaines hypothèses de régularité de l'état, qui peuvent ne pas être vérifiées dans le cadre hyperbolique. Cette question a été explorée dans [BP02, MP01] d'un point de vue théorique, et plus récemment dans [Gui09, GDC09] d'un point de vue numérique, où une modification du système de sensibilité a été proposée, pour "supprimer" les pics dans la sensibilité numérique. Plus spécifiquement, une modification du solveur de Riemann HLL utilisé pour évaluer les flux dans une méthode volumes finis a été proposée dans le contexte des équations de Saint-Venant. La correction est basée sur les conditions de Rankine-Hugoniot, qui gouvernent l'état à travers un choc.

1.4 Présentation des travaux de thèse

Dans cette thèse, nous étudions une hiérarchie de modèles de complexité croissante. Nous commençons dans le chapitre 2 par l'exemple le plus simple d'équation hyperbolique non linéaire, l'équation de Burgers, avant de nous intéresser à l'un des exemples les plus connus et les plus importants des systèmes d'EDP hyperboliques non linéaires : le système d'Euler. Le système d'Euler modélise la dynamique d'un matériau compressible, qui peut être un liquide ou un gaz. Pour décrire le flux, trois variables physiques sont considérées : la densité du fluide, sa vitesse et sa pression. Dans un premier temps, dans le chapitre 3, nous traitons une version de ce système, connue sous le nom de $p-$système, qui est

simplifiée sous certaines hypothèses physiques et mathématiques : le fluide est supposé être dans des conditions barotropes et il est décrit en coordonnées lagrangiennes ; puis, dans le chapitre 4 et le chapitre 5 nous traitons le système d'Euler complet : en particulier, dans le chapitre 4 nous étudions le tube à choc de Sod et le problème correspondant pour la sensibilité, et dans le chapitre 5 nous considérons une topographie non triviale et des solutions stationnaires.

Comme mentionné dans la section précédente, si des techniques standard d'AS, telles que la méthode de léquation de sensibilité continue, sont utilisées tout le temps, même en présence de discontinuités, elles fournissent une sensibilité qui contient des Dirac. Comme un Dirac ne peut pas être évalué numériquement, cela conduit à un pic qui s'étale dans le voisinage du choc et qui donc détériore la solution. De plus, les pics peuvent changer avec la discrétisation numérique. Ces sensibilités sont inadaptées à plusieurs des applications mentionnées ci-dessus. Par conséquent, dans le même ordre d'idées de ce qui est fait dans [Gui09, GDC09], nous modifions l'équation de sensibilité afin d'obtenir un système valable même dans le cas de solutions discontinues. Dans ce travail, nous adoptons un point de vue similaire, quoique légèrement différent : nous suggérons d'ajouter une correction sous la forme d'un terme source pour équilibrer les pics. Le terme source est non nul uniquement en cas de choc ou de discontinuité de contact et a une amplitude proportionnelle au saut de l'état à travers la discontinuité considérée. Le calcul de l'amplitude de cette correction est effectué en détail dans les chapitres suivants et il est basé sur une intégration des équations sur un volume de contrôle contenant une discontinuité. En ce qui concerne la discrétisation de ce terme source, nous remarquons qu'un détecteur de choc est nécessaire pour que le terme soit cohérent et pour éviter une surcorrection numérique.

En ce qui concerne la discrétisation de la loi de conservation, le problème d'état est bien connu et de nombreux de schémas numériques ont été développés au cours des années, mais ce n'est pas le cas pour les équations de sensibilité. De plus, considérer le système dans son ensemble n'apporte pas d'avantage puisque le système composé par l'état et la sensibilité n'est que faiblement hyperbolique en général. Pour ces raisons, notre stratégie d'un point de vue numérique est la suivante : d'abord, nous résolvons l'état en utilisant un schéma volumes finis classique ; ensuite, nous calculons le terme source, qui dépend de l'état ; enfin, un schéma numérique volumes finis adapté est appliqué au problème de sensibilité. Cette approche fonctionne bien si l'équation d'état est une équation scalaire, mais les résultats numériques montrent que pour les systèmes certaines précautions supplémentaires sont nécessaires : en particulier, la diffusion numérique joue un rôle fondamental et dégrade la convergence vers la solution exacte de la sensibilité. Par conséquent, dans ce travail nous adaptons un schéma numérique de type Godunov anti-diffusif (AD), introduit dans [CG08], et nous proposons des versions AD de différents schémas volumes finis, d'ordre différent en temps et en espace.

Enfin, nous explorons deux applications différentes de l'AS : l'optimisation et la quantification d'incertitude. Dans les deux applications nous considérons plusieurs paramètres d'entrée. Dans le chapitre 4, nous effectuons une analyse de quantification d'incertitude pour un problème de Riemann sur le système Euler complet : différentes approches d'AS sont comparées à la méthode de Monte Carlo plus coûteuse. Dans le chapitre 5 nous nous attaquons à un problème d'optimisation : en particulier, nous traitons un problème inverse pour la pression, où les paramètres d'optimisation considérés modifient directement

la topographie. Le but de ces applications est de comprendre l'importance de la diffusion numérique et du terme de correction dans des situations plus réalistes et de définir des directives pratiques pour l'utilisation de l'AS.

Au cours de ma première année de thèse j'ai travaillé sur la modélisation et l'optimisation de stratégies de course, sous la direction d'Amandine Aftalion. Ce travail a abouti à un papier publié [Fio17], qui est reproduit dans l'annexe A.

2 | Scalar case

This chapter deals with the nonlinear scalar case. In this simple framework we are able to introduce and tackle some of the main problems arising in the coupling of sensitivity analysis and hyperbolic equations: in particular, the definition of the source term and of a shock detector.

2.1 Analytical solution of the state equation

A scalar one-dimensional conservation law can be written in the following form:

$$\begin{cases} \partial_t u(x,t) + \partial_x f(u(x,t)) = 0 & x \in \mathbb{R},\ t > 0 \\ u(x,0) = g(x) & x \in \mathbb{R}, \end{cases} \tag{2.1}$$

where $u : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ is the conserved variable, $f : \mathbb{R} \to \mathbb{R}$ is the flux function and $g : \mathbb{R} \to \mathbb{R}$ is the initial condition. The method of characteristics can be used to solve analytically (2.1): the characteristics are curves in the plane (x,t) along which the PDE becomes an ordinary differential equation (ODE). In this case, we look for a set of curves along which the solution u is constant. Let $x_c(t)$ be the parametrisation of the curves, then one has:

$$0 = \frac{d}{dt} u(x_c(t),t) = \partial_x u(x_c(t),t) \frac{dx_c}{dt} + \partial_t u(x_c(t),t). \tag{2.2}$$

Considering that $\partial_x f(u(x,t)) = f'(u(x,t))\partial_x u(x,t)$, where $f' = \frac{df}{du}$, and comparing the equation in (2.1) with (2.2), one can write:

$$\frac{dx_c}{dt} = f'(u(x_c(t),t)) = f'(u(x_c(0),0)) \Rightarrow \frac{dx_c}{dt} = f'(g(x_0)), \tag{2.3}$$

where $x_0 = x_c(0)$. Therefore, the characteristics are straight lines and their slope is $f'(g(x_0))$. We can write the solution implicitly as follows:

$$u(x,t) = g(x - t f'(u(x,t))). \tag{2.4}$$

This method is valid as long as there is no intersection among the characteristics: in case of intersection a shock is generated and the solution $u(x,t)$ becomes discontinuous. Let us now examine whether or not it is possible for two characteristic to intersect. First, we deal with the case where g is a regular function. For this purpose, we consider two characteristics, $x_{c,1}(t)$ and $x_{c,2}(t)$, such that $x_{c,i}(0) = x_i$ and $x_1 < x_2$. If the two intersect, then one has:

$$x_1 + f'(g(x_1))t = x_2 + f'(g(x_2))t,$$

therefore:

$$t = -\frac{x_2 - x_1}{f'(g(x_2)) - f'(g(x_1))},$$

and if $f' \circ g =: h \in C^1$ for the mean value theorem $\exists\, \xi \in (x_1, x_2)$ such that:

$$t = -\frac{1}{h'(\xi)}.$$

Let us observe that if h is increasing the characteristics do not intersect for any time $t > 0$, therefore there is no shock. Otherwise, we can define the *breaking time* t_s as the smallest t for which the characteristics intersect:

$$t_s = -\frac{1}{\min h'(x)} = -\frac{1}{\min \frac{d}{dx} f'(g(x))}. \tag{2.5}$$

The point $x_{s,0}$ from which the shock originates is known too:

$$x_{s,0} = \bar{x} + f'(g(\bar{x}))t_s, \text{ where } \bar{x} := \operatorname{argmin} h'(x).$$

If g is discontinuous in a point x_d, two scenarios are possible: if $h(x_d^+) < h(x_d^-)$, the initial discontinuity is transported for all $t > 0$ ($t_s = 0$); otherwise the initial discontinuity is smoothed out and a rarefaction wave is generated.

Once the shock is generated, it moves along a curve $x_s(t)$ such that $\frac{dx_s}{dt} = \sigma(t)$, which separates the plane (x, t) into two parts, and in both of them the method of characteristics is valid. We denote with the superscript $+$ (respectively $-$) the quantities in the right (respectively left) part of the plane, i.e. $x > x_s(t)$ (respectively $x < x_s(t)$). The speed of the shock $\sigma(t)$ can be computed using the Rankine-Hugoniot conditions:

$$\sigma(t) = \frac{f(u^+(x_s(t), t)) - f(u^-(x_s(t), t))}{u^+(x_s(t), t) - u^-(x_s(t), t)}. \tag{2.6}$$

The position of the shock $x_s(t)$ can be then determined by solving the following ODE:

$$\begin{cases} \dfrac{dx_s}{dt} = \sigma(t), \\ x_s(t_s) = x_{s,0}. \end{cases} \tag{2.7}$$

This, along with some numerical methods to solve explicitly (2.4) and (2.7), is all we need to compute a reference solution. Finally, the solution of (2.1) can be written in the following compact form:

$$u(x, t) = u^+(x, t)H(x - x_s(t)) + u^-(x, t)H(x_s(t) - x), \tag{2.8}$$

where u^+ and u^- are obtained separately from (2.4) and H is the Heaviside function.

2.2 Sensitivity equation

2.2.1 Derivation of the sensitivity equation

Sensitivity analysis is the study of how variations in the output of a model can be attributed to different sources of perturbation in the model input. In this case, the model

is a PDE such as (2.1), the input is the vector parameter **a** on which the initial condition g and the flux function f depend, and the output is the state u. In the following, we will consider a scalar parameter a: this is to simplify the notation, without any loss of generality from a theoretical point of view.

Assuming that u is continuous and differentiable on its domain, we define its sensitivity with respect to the parameter of interest a as the derivative of u with respect to a and we use the notation u_a:

$$u_a = \frac{\partial u}{\partial a}. \tag{2.9}$$

We now apply the CSE method and we differentiate the system (2.1) with respect to the parameter a, obtaining:

$$\begin{cases} \partial_a(\partial_t u(x,t)) + \partial_a(\partial_x f(u(x,t))) = 0 & x \in \mathbb{R},\ t > 0 \\ \partial_a u(x,0) = \partial_a g(x) & x \in \mathbb{R}, \end{cases}$$

and exchanging the derivatives in space and time with the ones with respect to a one obtains the following equation and initial condition for the sensitivity:

$$\begin{cases} \partial_t u_a + \partial_x(f_a(u,u_a)) = 0 & x \in \mathbb{R},\ t > 0 \\ u_a(x,0) = g_a(x) & x \in \mathbb{R}, \end{cases} \tag{2.10}$$

where $g_a := \partial_a g$, $f_a(u,u_a) := f'(u)u_a + \partial_a f(u)$ and we dropped the time and space dependence in the equation for simplicity. We remark that the exchange of the derivatives can be done only under hypothesis of regularity of $\mathbf{U}$, more precisely the equality of mixed partials is guaranteed only if the second partial derivatives of $\mathbf{U}$ are continuous.

2.2.2 Analytical solution of the sensitivity equation

The analytical solution of (2.10) can be computed in the regular zones starting from (2.4), where the definition (2.9) is valid. By differentiating (2.4) with respect to a, one finds:

$$u_a(x,t) = g'(x - tf'(u(x,t)))(-tf''(u(x,t))u_a) + g_a(x - tf'(u(x,t))),$$

therefore one can obtain an explicit expression for the sensitivity:

$$\begin{aligned} u_a(x,t) = \frac{g_a(x - tf'(u(x,t)))}{1 + tf''(u(x,t))g'(x - tf'(u(x,t)))} = \\ = \frac{g_a(x - tf'(u(x,t)))}{1 + th'(x - tf'(u(x,t)))}. \end{aligned} \tag{2.11}$$

Let us observe that the denominator $1 + th'(x - tf'(u(x,t)))$ is zero if and only if there is an intersection between two characteristics, therefore only along the shock.

If the state u is discontinuous, (2.11) is still valid on both sides of the shock. Along the shock the state u is not differentiable in the classical sense; however, it admits a Dirac distribution as weak derivative. Therefore, differentiating the compact expression (2.8) one obtains:

$$\begin{aligned} u_a(x,t) = u_a^+(x,t)H(x - x_s(t)) + u_a^-(x,t)H(x_s(t) - x) \\ +(u^- - u^+)\partial_a x_s(t)\delta(x_s(t) - x), \end{aligned} \tag{2.12}$$

where δ is the Dirac delta function.

2.3 Correction term

In this section, we aim at proposing a new sensitivity system, whose solution does not exhibit spikes. To do that, we add a source term which should compensate the final term of (2.12), obtaining this new equation:

$$\partial_t u_a + \partial_x f_a(u, u_a) = S(u) \quad x \in \mathbb{R},\ t > 0. \tag{2.13}$$

Since we do not want to change the equation if the state u is regular, the source term has the following form:

$$S(u) = \alpha(t)\delta(x_s(t) - x),$$

where α is the amplitude of the correction and it is computed by integrating the equation (2.13) over a control volume $(x_1, x_2) \times (t_1, t_2)$, containing a discontinuity which moves at speed σ . We obtain:

$$\sigma(t_2 - t_1)u_a^- - \sigma(t_2 - t_1)u_a^+ + (t_2 - t_1)(f_a(u^+, u_a^+) - f_a(u^-, u_a^-)) = \int_{t_1}^{t_2} \alpha(t)dt,$$

and dividing this by $(t_2 - t_1)$ and as the control volume goes to zero one has:

$$\alpha(t) = \sigma(u_a^- - u_a^+) + f_a(u^+, u_a^+) - f_a(u^-, u_a^-). \tag{2.14}$$

If we differentiate the Rankine-Hugoniot conditions (2.6) with respect to a we obtain:

$$\begin{aligned}\sigma(u_a^+ - u_a^-) + \partial_a\sigma(u^+ - u^-) + \sigma_k(\partial_x u^+ - \partial_x u^-)\partial_a x_s(t) = \\ = f_a(u^+, u_a^+) - f_a(u^-, u_a^-) + (f'(u^+)\partial_x u^+ - f'(u^-)\partial_x u^-)\partial_a x_s(t),\end{aligned}$$

where the terms with $\partial_a x_s(t)$ come from the fact that the Rankine-Hugoniot are valid only if evaluated in the position of the shock. This terms are very difficult to estimate, however they are all zero in a first order finite volume framework, where the solution u is a piecewise constant, and therefore $\partial_x u^+ = 0$ and $\partial_x u^- = 0$. Neglecting this terms leads to a much simpler formulation:

$$\sigma(u_a^+ - u_a^-) + \partial_a\sigma(u^+ - u^-) = f_a(u^+, u_a^+) - f_a(u^-, u_a^-).$$

Comparing the latter to (2.14) we have the definition of the amplitude:

$$\alpha(t) = \partial_a\sigma(u^+ - u^-). \tag{2.15}$$

2.4 Global system

It is possible to write (2.1) and (2.10) as a system, by defining the following vectors:

$$\mathbf{U} = \begin{bmatrix} u \\ u_a \end{bmatrix},\ \mathbf{F}(\mathbf{U}) = \begin{bmatrix} f(u) \\ f'(u)u_a + \partial_a f(u) \end{bmatrix},\ \mathbf{G} = \begin{bmatrix} g \\ g_a \end{bmatrix}.$$

Therefore, the global system in the conservative form is the following:

$$\begin{cases} \partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = 0 & x \in \mathbb{R},\ t > 0 \\ \mathbf{U}(x, 0) = \mathbf{G}(x) & x \in \mathbb{R}. \end{cases} \tag{2.16}$$

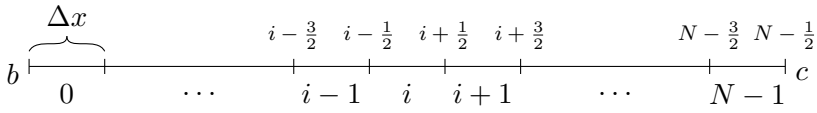


Figure 2.1 – Spatial discretisation.

We can also write (2.16) in the non conservative form:

$$\begin{cases} \partial_t \mathbf{U} + A(\mathbf{U})\partial_x \mathbf{U} = 0 & x \in \mathbb{R},\ t > 0 \\ \mathbf{U}(x,0) = \mathbf{G}(x) & x \in \mathbb{R}, \end{cases} \tag{2.17}$$

where the matrix A is the Jacobian of the system and can be written:

$$A(\mathbf{U}) = \frac{\partial F}{\partial \mathbf{U}} = \begin{bmatrix} f'(u) & 0 \\ f''(u)u_a + \partial_a f'(u) & f'(u) \end{bmatrix}. \tag{2.18}$$

Let us observe that A has two repeated eigenvalues, therefore it is not diagonalisable, unless $f''(u)u_a + \partial_a f'(u) \equiv 0$. This means that the global system (2.16) is only weakly hyperbolic. Weakly hyperbolic systems are known to develop Dirac distributions in their solution. We remark that the fact that A is non-diagonalisable is not straightforward to prove in the case of systems.

2.5 Numerical methods

The aim of this section is to present the numerical schemes that can be used to solve (2.16). From now on, we will consider the equations on an interval (b, c), which is divided into N cells, all of the same length Δx. The subscript i indicates the $i-$th cell, while $i \pm \frac{1}{2}$ refers to the interfaces (see Figure 2.1). Concerning the time discretisation, we will use a variable time step Δt, chosen such that the CFL-condition is respected. The superscript n indicates the $n-$th time step.

As already mentioned, we treat the state and the sensitivity system separately. For the state there is no particular issue: all the classical numerical schemes can be used. However, for the sensitivity a modified scheme is used. The suggested strategy for solving (2.16) consists of the following steps:

(i) solution of the state using numerical methods for hyperbolic equations;

(ii) computation of the correction term;

(iii) solution of the sensitivity with a corrected numerical scheme.

2.5.1 Numerical schemes for the state

The method chosen for the state equation is the approximate Riemann solver of Roe: the reason of this choice is that it provides the exact solution in the case of an isolated

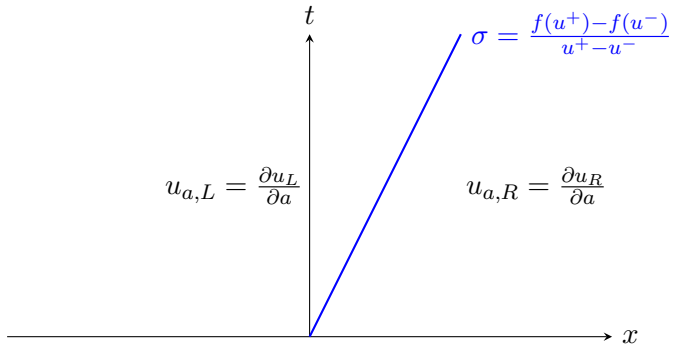


Figure 2.2 – Scheme of the solver of Roe for the sensitivity.

shock. For the sake of simplicity we consider the case $f'(u) > 0$. The scheme in the internal cells (i.e. $i \neq 0$) is the following:

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}(f(u_{i-1}^n) - f(u_i^n)), \tag{2.19}$$

while for the first cell we impose $u_0^n = g(b)\ \forall n$.

2.5.2 Numerical schemes for the sensitivity

For the sensitivity equation (2.10), i.e. the one without the source term, the same scheme used for the state can be applied:

$$u_{a,i}^{n+1} = u_{a,i}^n + \frac{\Delta t}{\Delta x}(f_a(u_{i-1}^n, u_{a,i-1}^n) - f_a(u_i^n, u_{a,i}^n)),$$

however, this scheme provides sensitivity with spikes. We here propose a second scheme, which comes from a quite natural discretisation of the source term:

$$u_{a,i}^{n+1} = u_{a,i}^n + \frac{\Delta t}{\Delta x}(f_a(u_{i-1}^n, u_{a,i-1}^n) - f_a(u_i^n, u_{a,i}^n)) + \sigma_{a,i}^n(u_i^n - u_{i-1}^n)\delta_i^n, \tag{2.20}$$

where $\delta_i^n = 1$ if there is a shock in the $i-$th cell at the $n-$th time step, $\delta_i^n = 0$ otherwise. One of the main difficulties of this problem is to find an efficient shock detector.

We remark that the same scheme (2.20) can be found if we consider that the solution of equation (2.13) is the derivative of the state in the regular zones without any spike along the shock. Therefore it is easy to define a Roe Riemann solver for the sensitivity (cf. Figure 2.2).

Once the solver is defined, one can compute the average on each cell using a general Godunov method (cf. Figure 2.3), which results in the following scheme:

$$\begin{aligned} u_{a,i}^{n+1} =& \frac{1}{\Delta x}\left[\Delta t \frac{f(u_i^n) - f(u_{i-1}^n)}{u_i^n - u_{i-1}^n} u_{a,i-1}^n + \left(\Delta x - \Delta t \frac{f(u_i^n) - f(u_{i-1}^n)}{u_i^n - u_{i-1}^n}\right) u_{a,i}^n\right] = \\ =& u_{a,i}^n + \frac{\Delta t}{\Delta x}\frac{f(u_i^n) - f(u_{i-1}^n)}{u_i^n - u_{i-1}^n}(u_{a,i-1}^n - u_{a,i}^n), \end{aligned}$$

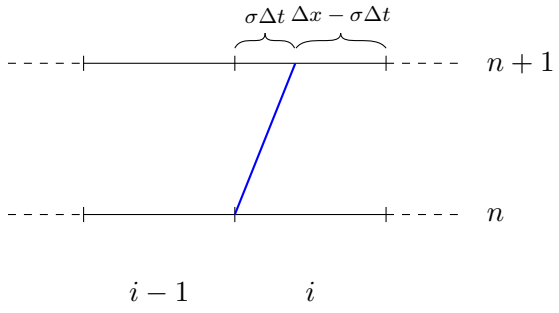


Figure 2.3 – Godunov scheme: σ is the slope of the blue line, i.e. the speed at which the information travels.

which can be rewritten as follows:

$$u_{a,i}^{n+1} = u_{a,i}^n + \frac{\Delta t}{\Delta x}(f_a(u_{i-1}^n, u_{a,i-1}^n) - f_a(u_i^n, u_{a,i}^n)) + \sigma_{a,i}^n(u_i^n - u_{i-1}^n), \tag{2.21}$$

and one can find the correction term by difference:

$$\sigma_{a,i}^n(u_i^n - u_{i-1}^n) = (f_a(u_i^n, u_{a,i}^n) - f_a(u_{i-1}^n, u_{a,i-1}^n)) - \frac{f(u_i^n) - f(u_{i-1}^n)}{u_i^n - u_{i-1}^n}(u_{a,i}^n - u_{a,i-1}^n),$$

and using once again the Rankine-Hugoniot conditions we have the identity:

$$\sigma_{a,i}^n(u_i^n - u_{i-1}^n) = f_a(u_i^n, u_{a,i}^n) - f_a(u_{i-1}^n, u_{a,i-1}^n) - \sigma_i^n(u_{a,i}^n - u_{a,i-1}^n).$$

2.6 Numerical results

In this section, we show the numerical results obtained with the schemes presented in section 2.5 and we compare them with the reference solution computed as in sections 2.1 and 2.2. For this purpose, we consider the Burgers' equation, therefore the flux function is the following:

$$f(u) = \frac{u^2}{2},$$

with different initial data $g(x)$. We remark that $f'(u) = u$, it follows that $h = g$. For the sensitivity flux, we have:

$$f_a(u, u_a) = uu_a.$$

2.6.1 Riemann problem

We start with the Riemann problem, i.e.:

$$g(x) = \begin{cases} u_L & x < x_c, \\ u_R & x > x_c, \end{cases} \tag{2.22}$$

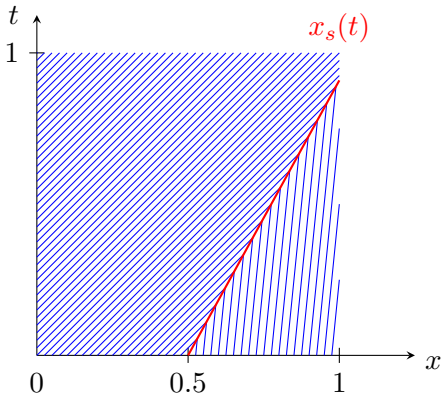


Figure 2.4 – Characteristic curves for the Burger's equation with initial condition (2.22).

where $x_c \in (b, c)$. Let us observe that if g is increasing (i.e. $u_R > u_L$), the initial discontinuity is smoothed out, otherwise it is transported at a speed which is known explicitly in this simple case:

$$\sigma = \frac{u_R + u_L}{2}. \tag{2.23}$$

Being σ constant, the shock propagates along a straight line of equation:

$$x_s(t) = \sigma t + x_c. \tag{2.24}$$

Since we are interested in the case of shocks, we consider $u_L > u_R$, and in particular $u_L = 1$ and $u_R = 0.1$. The domain is the interval $(0, 1)$, and $x_c = 0.5$. The parameter of interest a is u_L, therefore the initial condition for the sensitivity g_{u_L} is the following:

$$g_{u_L}(x) = \begin{cases} 1 & x < 0.5, \\ 0 & x > 0.5. \end{cases} \tag{2.25}$$

The characteristics and the shock curve for this problem are plotted in Figure 2.4. The analytical solution for the state is the following:

$$u(x,t) = \begin{cases} u_L & x < \sigma t + x_c, \\ u_R & x > \sigma t + x_c. \end{cases} \tag{2.26}$$

Differentiating (2.26) with respect to u_L one obtains the analytical solution for the sensitivity, which is:

$$u_{u_L}(x,t) = \begin{cases} 1 & x < \sigma t + x_c, \\ 0 & x > \sigma t + x_c. \end{cases} \tag{2.27}$$

In Figure 2.6 we show the numerical solution for the state obtained with $\Delta x = 10^{-3}$ compared to the analytical one, for different times: the difference between the two is due to numerical diffusion, and it gets smaller as Δx is reduced. Knowing the analytical solution, which will be denoted by the subscript ex in the following, one can compute the error as follows:

$$err(t^n) = \|u_{ex}(x_i, t^n) - u_i^n\|_{L^1(0,1)}, \tag{2.28}$$

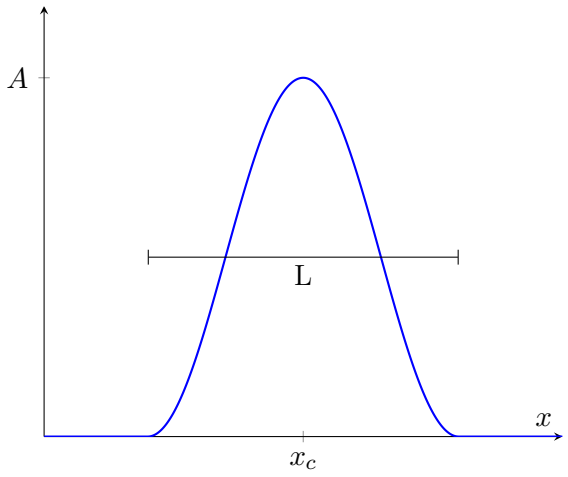


Figure 2.5 – Continuous initial condition.

where $u_{ex}(x_i, t^n)$ is the exact solution in the center of the i-th cell at the n-th time step. Figure 2.8 shows the sensitivity solution without the correction term (i.e. $\delta_i = 0\ \forall i$): as one can see, where the state is discontinuous the numerical solution of the sensitivity has a peak, which gets bigger with time, as expected, since the coefficient of the Dirac delta function in (2.12) can be easily computed in this case and it is linear in t.

In order to apply the corrected scheme (2.20), we need to define a shock detector. In this simple case, it is easy to define a good one:

$$\delta_i^n = \begin{cases} 0 & \text{if } u_{i-1}^n = u_i^n, \\ 1 & \text{otherwise.} \end{cases} \tag{2.29}$$

Figure 2.10 shows in blue the cells in which a shock is detected using (2.29) and in red the real position of the shock: the region is quite large due to the numerical diffusion.

The results obtained with the correction are shown in Figure 2.7.

Finally, in Figure 2.9 we show the L^1 norm of the error at final time $T = 0.2$ for the state and for the sensitivity. For reference, $\|u(x,T)\|_{L^1(0,1)} = 0.649$ and $\|u_{u_L}(x,T)\|_{L^1(0,1)} = 0.61$. We remark that for this purpose we compared the numerical sensitivity obtained with the correction to the analytical sensitivity in (2.12) but without the Dirac term.

2.6.2 Continuous initial condition

Now we present the case of a continuous initial condition, which is plotted in Figure 2.5:

$$g(x) = \begin{cases} A\sin^2(\frac{\pi}{L}(x - x_c) + \frac{\pi}{2}) & \text{if } x_c - \frac{L}{2} < x < x_c + \frac{L}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{2.30}$$

Let us observe that, for $x > x_c$, g is decreasing, therefore a shock will occur. The breaking time can be computed using (2.5) and using the fact that for the Burger's equation $h = g$:

$$t_s = -\frac{1}{\min g'(x)} = -\frac{1}{\min\{-\frac{A\pi}{L} sin(\frac{2\pi}{L}(x - x_c))\}} = \frac{L}{A\pi}.$$

The point from which the shock originates is:

$$x_{s,0} = x_c + \frac{L}{4}\left(1 + \frac{2}{\pi}\right).$$

In this case, the curve $x_s(t)$ it is not known analytically because the values u^+ and u^- are not (we recall that the solution for the state is known only in the implicit form (2.4)). However, it is possible to compute numerically $x_s(t)$ and to plot the exact solution $u(x,t)$ as follows. One can define the following matrix:

$$X[n,i] = x_i + t^n g(x_i).$$

To compute a numerical approximation of $x_s(t^n)$, the strategy is the following:

(a) let $\bar{n}$ be the closest time step to the breaking time t_s;

(b) find $\bar{k}$ such that $X[\bar{n},\bar{k}] \leq x_{s,0}$ and $X[\bar{n},\bar{k}+1] \geq x_{s,0}$;

(c) $u^+ = g(x_{\bar{k}+1})$ and $u^- = g(x_{\bar{k}}) \quad \Rightarrow \quad \sigma = \frac{u^+ + u^-}{2}$;

(d) $x_s(t^{n+1}) = x_s(t^n) + \sigma(t^{n+1} - t^n)$;

(e) erase all the characteristics that cross the shock between time $t^{\bar{n}}$ and $t^{\bar{n}+1}$

(f) repeat (b)-(e) until the final time T.

In Figure 2.11 we show the characteristics of the Burger's equation with initial condition (2.30) and the shock curve computed as explained above.

It is now possible to plot u for all time steps t^n in the points $X[n,i]$ ($u(X[n,i],t^n) = g(x_i)$), which is exact and is a good reference solution if the initial mesh is fine enough. The same can be done for the sensitivity: $u_a(X[n,i],t^n) = \dfrac{g_a(X[n,i] - t^n g(x_i))}{1 + t^n g'(X[n,i] - t^n g(x_i))}$.

In Figure 2.12 we show the numerical and the reference solution for the state, obtained with $\Delta x = 10^{-3}$.

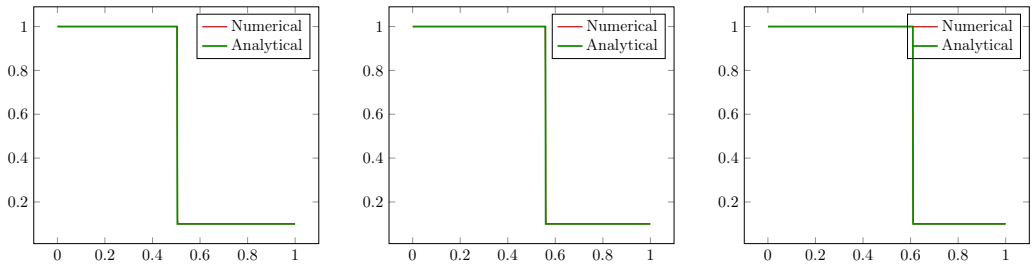
In this case a more complex shock detector, based on the second derivative, has been used. Let D be the maximum of the second derivative of the initial condition, then δ_i is defined as follows:

$$\delta_i^n = \begin{cases} 1 & \text{if } \dfrac{|u_{i-1}^n - 2u_i^n + u_{i+1}^n|}{\Delta x^2} > kD, \\ 0 & \text{otherwise,} \end{cases}$$

where k is a proper constant. In Figure 2.17 we show the results of this shock detector with $k = 30$.

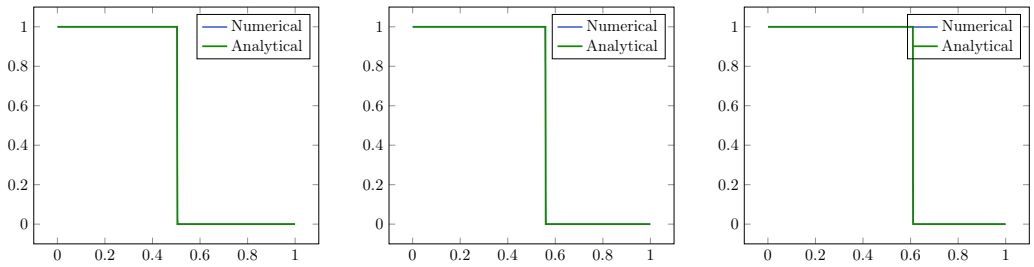
In Figures 2.16 and 2.15 we show the solution of the sensitivity with respect to the parameter x_c, with and without the correction. The figures in the middle correspond to $t = t_s$: as one can see, before the breaking time the solutions are the same, whilst after the breaking time the solution given by the corrected scheme does not present any peak.

Finally, in Figures 2.14 and 2.13 we show the solution of the sensitivity with respect to the parameter A, with and without the correction: in this case, the over-correction due to the fact that a shock is detected even before the breaking time, is slightly more evident.



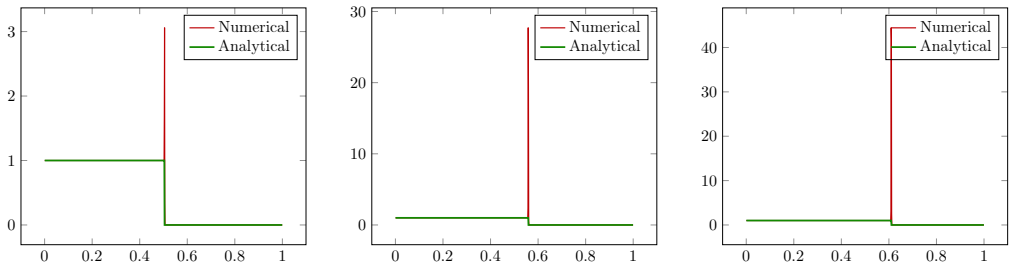
(a) $t = 0.0081$ (b) $t = 0.1071$ (c) $t = 0.2$

Figure 2.6 – State solution for the Riemann problem.



(a) $t = 0.0081$ (b) $t = 0.1071$ (c) $t = 0.2$

Figure 2.7 – Sensitivity solution for the Riemann problem.



(a) $t = 0.0081$ (b) $t = 0.1071$ (c) $t = 0.2$

Figure 2.8 – Sensitivity solution for the Riemann problem without correction.

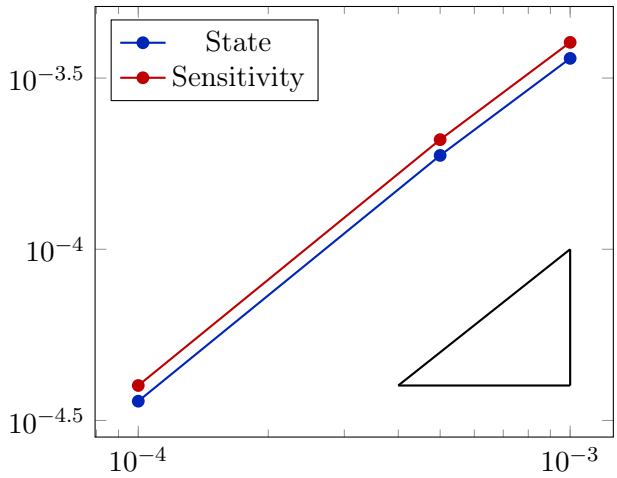


Figure 2.9 – L^1 norm of the error at final time $T = 0.2$.

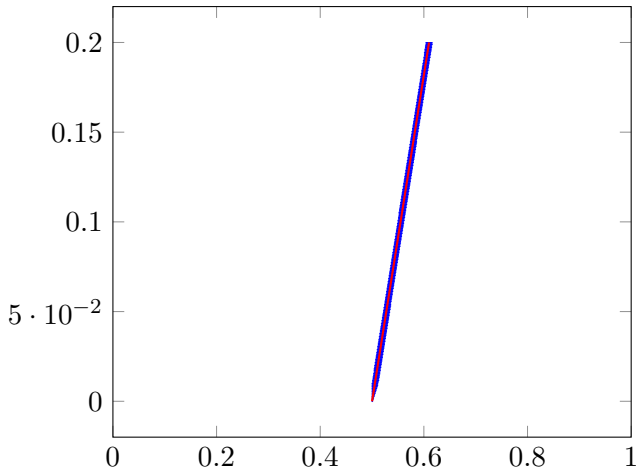


Figure 2.10 – Shock detector (2.29) for the Riemann problem.

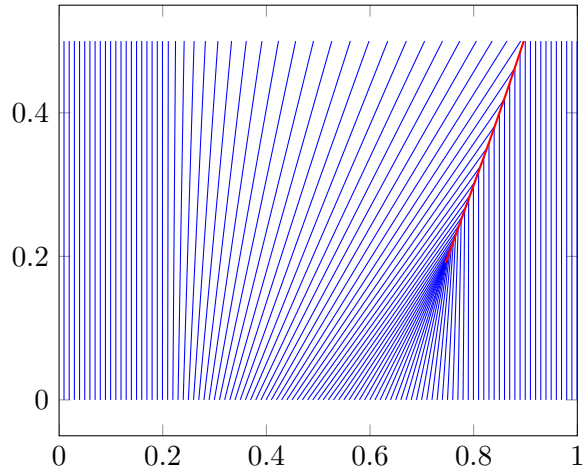
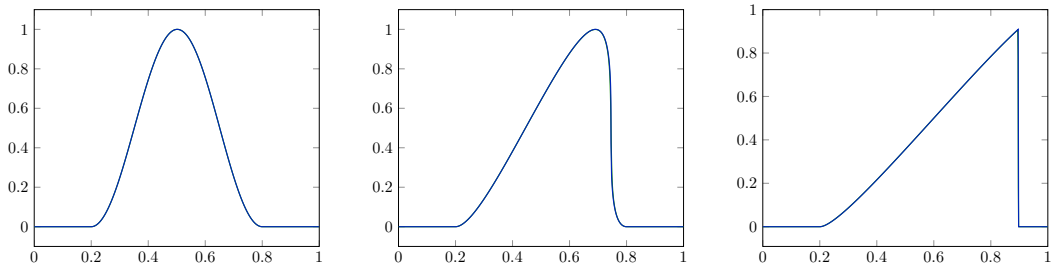
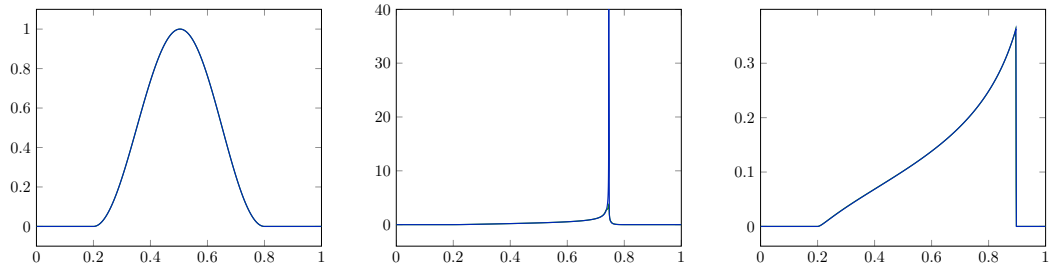


Figure 2.11 – Characteristics for the Burger's equation with initial condition (2.30).



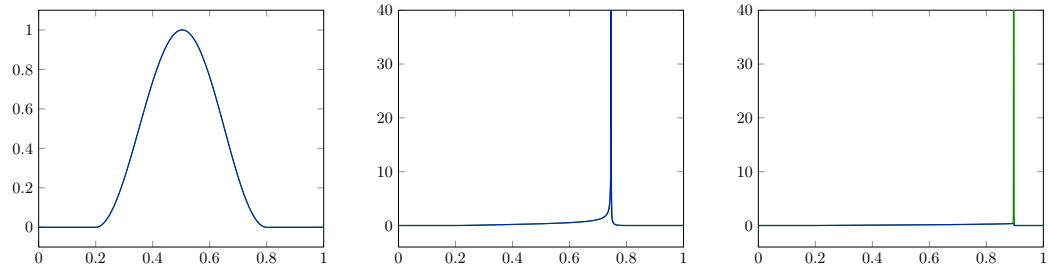
(a) $t = 0.0018$ (b) $t = 0.1908$ (c) $t = 0.5$

Figure 2.12 – State solution for the problem with continuous i.c.



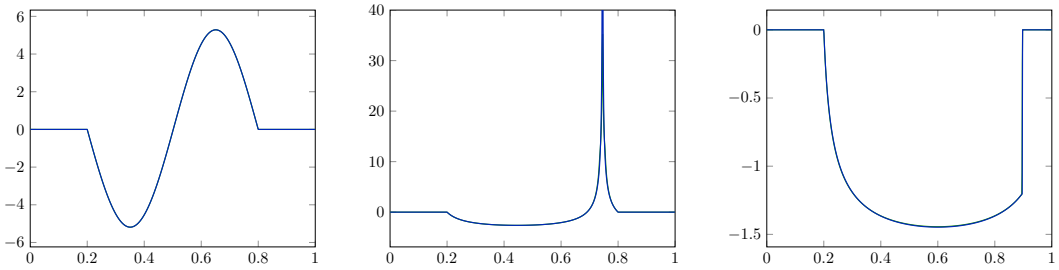
(a) $t = 0.0018$ (b) $t = 0.1908$ (c) $t = 0.5$

Figure 2.13 – Sensitivity solution for the problem with continuous i.c. with correction, $a = A$.



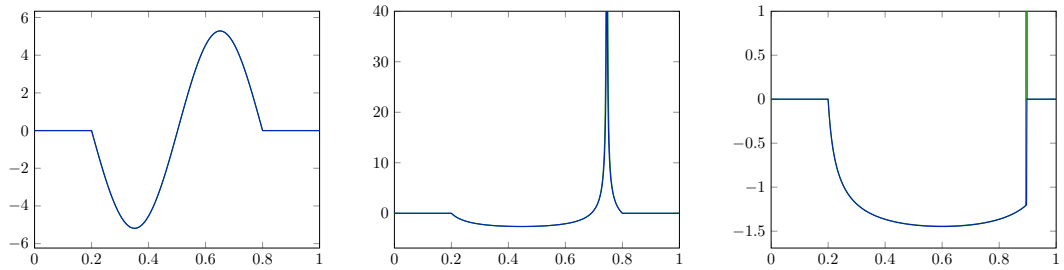
(a) $t = 0.0018$ (b) $t = 0.1908$ (c) $t = 0.5$

Figure 2.14 – Sensitivity solution for the problem with continuous i.c. without correction, $a = A$.



(a) $t = 0.0018$ (b) $t = 0.1908$ (c) $t = 0.5$

Figure 2.15 – Sensitivity solution for the problem with continuous i.c. with correction, $a = x_c$.



(a) $t = 0.0018$ (b) $t = 0.1908$ (c) $t = 0.5$

Figure 2.16 – Sensitivity solution for the problem with continuous i.c. without correction, $a = x_c$.

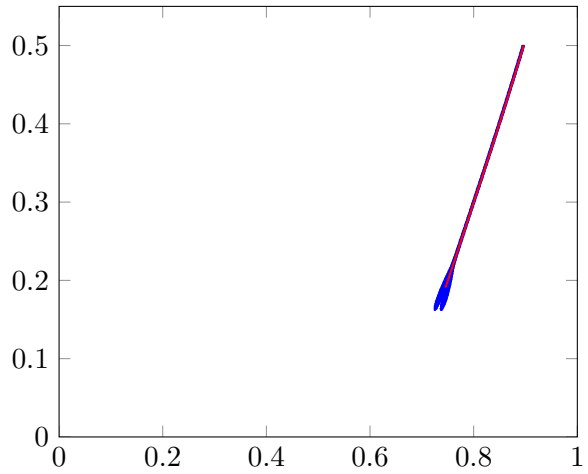


Figure 2.17 – Shock detector based on the second derivative.

3 | The $p-$system

In this chapter, which is an adaptation of two works [CDF17b, CDF17a], we deal with the Euler system in barotropic conditions and in Lagrangian coordinates.

3.1 Problem description

As already mentioned in the previous chapters, standard SA methods can be used only if the solution $\mathbf{U}$ is regular enough [BP02], which is usually not the case for hyperbolic systems of the general form

$$\begin{cases} \partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = 0, \quad x \in \mathbb{R}, \quad t > 0, \\ \mathbf{U}(x, 0) = \mathbf{U}_0(x). \end{cases}$$

In fact, it is well known systems of this type, as well as the scalar conservation laws introduced in Chapter 2, can have discontinuous solutions, regardless of the regularity of the initial condition $\mathbf{U}_0$. If the state $\mathbf{U}$ is discontinuous, the sensitivity $\mathbf{U}_a = \partial_a \mathbf{U}$ will exhibit Dirac delta functions.

In this Chapter, we consider the barotropic Euler equations in Lagrangian coordinates, i.e. the p-system; however, everything can be extended to any hyperbolic system: for instance, the complete Euler system and the quasi 1D Euler system will be the subject of the following chapters. The choice of starting with the $p-$system is motivated by the fact that, although quite simple, it presents all the main features of hyperbolic systems: this allows us to solve the state problem easily and to focus on the sensitivity problem.

The system is written:

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x p(\tau) = 0, \end{cases} \tag{3.1}$$

where $\tau > 0$ is the co-volume (i.e. $\tau = \frac{1}{\rho}$, a ρ is the density of the fluid), u is the Lagrangian velocity and the pressure $p(\tau)$ is a function only of τ. We assume $p'(\tau) < 0$ and $p''(\tau) > 0$. The Jacobian matrix of the system is the following:

$$\mathbf{M}(\tau, u) = \begin{bmatrix} 0 & -1 \\ p'(\tau) & 0 \end{bmatrix} \tag{3.2}$$

and its eigenvalues are real and distinct $\lambda_\pm = \pm c$, where $c = \sqrt{-p'(\tau)}$ is the Lagrangian sound speed. Therefore $\mathbf{M}$ is $\mathbb{R}-$diagonalisable, and (3.1) is strictly hyperbolic. In this work we will consider $p(\tau) = \tau^{-\gamma}$, where $\gamma = 1.4$ is the heat capacity ratio.

If we consider *smooth* solutions of (3.1), we can apply the CSE method, differentiate (3.1) with respect to a and obtain the following sensitivity equations:

$$\begin{cases} \partial_t \tau_a - \partial_x u_a = 0, \\ \partial_t u_a + \partial_x (p'(\tau)\tau_a) = 0. \end{cases} \tag{3.3}$$

We remark that if the pressure law depends directly on the parameter a (i.e. $p = p(\tau; a)$), there is an additional term in the second equation: $\partial_x p_a(\tau; a)$, where $\partial_a p(\tau; a) = p_a(\tau; a)$. In the following, we will not consider this case. In order to introduce a more compact notation, we define the state and sensitivity vectors and their fluxes:

$$\mathbf{U} = \begin{bmatrix} \tau \\ u \end{bmatrix}, \qquad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} -u \\ p(\tau) \end{bmatrix}, \qquad \mathbf{U}_a = \begin{bmatrix} \tau_a \\ u_a \end{bmatrix}, \qquad \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = \begin{bmatrix} -u_a \\ p'(\tau)\tau_a \end{bmatrix},$$

and rewrite the systems (3.1) and (3.3) in a vectorial form:

$$\begin{cases} \partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = 0, \\ \partial_t \mathbf{U}_a + \partial_x \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = 0. \end{cases} \tag{3.4}$$

The Jacobian matrix of the global system (3.4) is calculated as:

$$\mathbf{A}(\mathbf{V}) = \frac{\partial \mathbf{G}}{\partial \mathbf{V}} = \left[\begin{array}{cc:cc} 0 & -1 & 0 & 0 \\ p'(\tau) & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & -1 \\ p''(\tau)\tau_a & 0 & p'(\tau) & 0 \end{array}\right], \text{ with } \mathbf{V} = \begin{bmatrix} \mathbf{U} \\ \mathbf{U}_a \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{F} \\ \mathbf{F}_a \end{bmatrix}.$$

One can remark that $\mathbf{A}(\mathbf{V})$ is a lower triangular block matrix whose diagonal blocks are identical to each other and to the state system's Jacobian matrix. We observe that the global system (3.4) has two repeated eigenvalues, the same $\lambda_\pm$ as the original system (3.1), and that the matrix $\mathbf{A}(\mathbf{V})$ is not $\mathbb{R}-$diagonalisable as soon as $\tau_a \neq 0$. This means that the global system (3.4) is only weakly hyperbolic. Therefore, the system (3.4) as it is will provide us, in case of discontinuous state $\mathbf{U}$, with a sensitivity $\mathbf{U}_a$ presenting Dirac delta functions, in addition to the usual discontinuity, so that these solutions have to be interpreted in the sense of measures. We refer for instance the reader to the following papers, and the references therein: [BJL03, CKM12, FLF92, Jos93, LeF90, YZ12]. However, sensitivities with Dirac delta functions are unusable for many applications. For this reason, we add a correction term to the sensitivity equations, as done in [Gui09]. The definition of a proper correction term is the subject of the next section.

3.2 Source term

In this section, we aim at proposing a new version of (3.4) which is also valid for discontinuous solutions of the state variable $\mathbf{U}$. Recall indeed that (3.4) has been derived assuming formally that the solution is smooth whilst hyperbolic equations are well known to develop discontinuities in finite time even for smooth initial data $\mathbf{U}(x, t = 0) = \mathbf{U}_0(x)$.

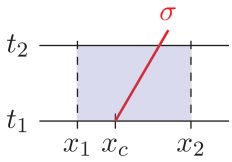


Figure 3.1 – Control volume in light blue.

In order to compensate for the Dirac delta functions that appear in the solutions $\mathbf{U}_a$ of (3.4) when $\mathbf{U}$ is discontinuous, we add to (3.3) a source term $\mathbf{S}$ of the following form:

$$\mathbf{S} = \sum_{k=1}^{N_s} \delta_k \boldsymbol{\rho}_k, \tag{3.5}$$

where N_s is the number of discontinuities in the state solution $\mathbf{U}$, $\boldsymbol{\rho}_k$ is the amplitude of the k-th correction (to be computed), and δ_k is the Dirac delta function $\delta_k = \delta(x - x_{s,k})$, where $x_{s,k}$ is the position of the k-th discontinuity. The new version of (3.4) we are going to consider can thus be written as:

$$\begin{cases} \partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = 0, \\ \partial_t \mathbf{U}_a + \partial_x \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = \mathbf{S}. \end{cases} \tag{3.6}$$

Let us motivate our choice and define $\boldsymbol{\rho}_k$ by considering a control volume $(x_1, x_2) \times (t_1, t_2)$ as in Figure 3.1, which contains only the k-th discontinuity, propagating at speed σ_k. We integrate the second equation of (3.6) over the control volume:

$$\int_{x_1}^{x_2} \int_{t_1}^{t_2} \partial_t \mathbf{U}_a + \partial_x \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) \mathrm{d}x\mathrm{d}t = \int_{x_1}^{x_2} \int_{t_1}^{t_2} \delta(x - x_{s,k}) \boldsymbol{\rho}_k(t) \mathrm{d}x\mathrm{d}t,$$

and we obtain

$$\int_{x_1}^{x_2} \mathbf{U}_a(x, t_2) - \mathbf{U}_a(x, t_1) \mathrm{d}x + \int_{t_1}^{t_2} \mathbf{F}_a(x_2, t) - \mathbf{F}_a(x_1, t) \mathrm{d}t = \int_{t_1}^{t_2} \boldsymbol{\rho}_k(t) \mathrm{d}t, \tag{3.7}$$

where we used the simplified notation $\mathbf{F}_a(x, t) = \mathbf{F}_a(\mathbf{U}(x, t), \mathbf{U}_a(x, t))$. We divide (3.7) by $(t_2 - t_1)$ and as the size of the control volume tends to zero we have:

$$\boldsymbol{\rho}_k(t) = (\mathbf{U}_a^- - \mathbf{U}_a^+)\sigma_k + \mathbf{F}_a^+ - \mathbf{F}_a^-, \tag{3.8}$$

where the plus (respectively minus) stands for the value of the variables to the right (respectively left) of the discontinuity. The relation (3.8) gives a natural meaning of $\boldsymbol{\rho}_k$ in terms of a defect measure of the Rankine-Hugoniot conditions for (3.3). Now, we want to define $\boldsymbol{\rho}_k$ so that the new model with the source term is valid also in the case of a discontinuous state. We start from the Rankine-Hugoniot conditions for the state variable $\mathbf{U}$ across a discontinuity:

$$(\mathbf{U}^- - \mathbf{U}^+)\sigma_k = \mathbf{F}^- - \mathbf{F}^+,$$

and we suggest to differentiate them with respect to the parameter a. As we do that, we should consider the fact that the Rankine-Hugoniot conditions are valid only at the discontinuity location $x_{k,s}(t)$, which depends on the parameter a. Therefore, we obtain

$$\begin{aligned}(\mathbf{U}_a^- - \mathbf{U}_a^+)\sigma_k + (\mathbf{U}^- - \mathbf{U}^+)\sigma_{k,a} + \sigma_k(\nabla\mathbf{U}^+ - \nabla\mathbf{U}^-)\partial_a x_{k,s}(t) = \\ = \mathbf{F}_a^- - \mathbf{F}_a^+ + \left(\frac{\partial\mathbf{F}(\mathbf{U}^+)}{\partial\mathbf{U}}\nabla\mathbf{U}^+ - \frac{\partial\mathbf{F}(\mathbf{U}^-)}{\partial\mathbf{U}}\nabla\mathbf{U}^-\right)\partial_a x_{k,s}(t).\end{aligned}$$

The terms depending on $\partial_a x_{k,s}(t)$ are very difficult to estimate. However we remark that, thanks to the presence of the gradients, they are zero if we consider that the solution $\mathbf{U}$ is constant in the left and right neighbourhoods of the shock. This is verified in a standard first order finite volume approach. We obtain therefore a simpler formula:

$$(\mathbf{U}_a^- - \mathbf{U}_a^+)\sigma_k + (\mathbf{U}^- - \mathbf{U}^+)\sigma_{k,a} = \mathbf{F}_a^- - \mathbf{F}_a^+, \tag{3.9}$$

with $\sigma_{k,a} = \partial_a\sigma_k$. Comparing the latter with (3.8), one is thus led to set

$$\boldsymbol{\rho}_k(t) = \sigma_{k,a}(\mathbf{U}^+ - \mathbf{U}^-). \tag{3.10}$$

Our choice is of course valid for each $k-$discontinuity of the state solution, leading us to definition (3.5) where the sum is taken over the number of discontinuities.

Note also that by construction, if a triple $(\mathbf{U}^-, \mathbf{U}^+, \sigma)$ is associated with an admissible discontinuity with a left (respectively right) state $\mathbf{U}^-$ (respectively $\mathbf{U}^+$) and σ is the speed of propagation, then the triple $(\mathbf{U}_a^-, \mathbf{U}_a^+, \sigma)$ with $\mathbf{U}_a^- = \partial_a\mathbf{U}^-$ and $\mathbf{U}_a^+ = \partial_a\mathbf{U}^+$ is also admissible in the sense that it satisfies the generalised Rankine-Hugoniot relations imposed by $\mathbf{S}$. In other words, the sensitivity solution of (3.6) is obtained by differentiating the state solution with respect to a when the solution is smooth or discontinuous with constant left and right states. As far as the initial condition is concerned, we have:

$$\begin{pmatrix}\mathbf{U}(x, t=0)\\ \mathbf{U}_a(x, t=0)\end{pmatrix} = \begin{pmatrix}\mathbf{U}_0(x)\\ \partial_a\mathbf{U}_0(x)\end{pmatrix}.$$

3.3 Exact solution of the Riemann problem

In this section, we present the exact resolution of the Riemann problem for the state and the sensitivity, associated with the initial data:

$$\mathbf{U}(x,0) = \begin{cases}\mathbf{U}_L & x < x_c,\\ \mathbf{U}_R & x > x_c,\end{cases} \qquad \mathbf{U}_a(x,0) = \begin{cases}\mathbf{U}_{a,L} & x < x_c,\\ \mathbf{U}_{a,R} & x > x_c,\end{cases}$$

for a given x_c. First, we compute the solution of the state system (3.1), which is well-known but necessary to solve (3.4). Then we differentiate it with respect to the parameter of interest a to obtain the sensitivity. As we will see, the sensitivity exhibits interesting and non trivial behaviours.

3.3.1 The state variable

We recall that the eigenvalues of the Jacobian matrix of the state system are:

$$\lambda_1(\mathbf{U}) = -\sqrt{-p'(\tau)} \quad \text{and} \quad \lambda_2(\mathbf{U}) = \sqrt{-p'(\tau)},$$

and the eigenvectors:

$$\mathbf{r}_1(\mathbf{U}) = \begin{bmatrix} \dfrac{2\sqrt{-p'(\tau)}}{p''(\tau)} \\ -\dfrac{2p'(\tau)}{p''(\tau)} \end{bmatrix} \quad \text{and} \quad \mathbf{r}_2(\mathbf{U}) = \begin{bmatrix} -\dfrac{2\sqrt{-p'(\tau)}}{p''(\tau)} \\ -\dfrac{2p'(\tau)}{p''(\tau)} \end{bmatrix},$$

which are chosen in such a way that $\nabla\lambda_i \cdot \mathbf{r}_i = 1$. Since the pairs $(\lambda_i, \mathbf{r}_i)$ are both genuinely non linear, the waves associated can be either shocks or rarefaction waves. The structure of the analytical solution of the state is presented in Figure 3.4 and it consists of two waves, whose speeds can be computed exactly.

In order to give more details on this structure, which will be necessary to explain the structure of the sensitivity, let us consider the plane (τ, u) and the points $\mathbf{U}_L$ and $\mathbf{U}_R$: starting from $\mathbf{U}_L$ we need to reach $\mathbf{U}_R$ passing from an intermediate state $\mathbf{U}^*$ using shocks and rarefaction waves, see Figure 3.4 for the notations. First, we compute which points $\mathbf{U}$ are reachable through a shock of speed σ from $\mathbf{U}_L$. Across a shock, the Rankine-Hugoniot conditions are valid, therefore:

$$\begin{cases} -u + u_L = \sigma(\tau - \tau_L) \\ p(\tau) - p(\tau_L) = \sigma(u - u_L). \end{cases} \tag{3.11}$$

Finding σ from the first equation and replacing it in the second one, one has:

$$(u - u_L)^2 = -(p(\tau) - p(\tau_L))(\tau - \tau_L),$$

and we observe that the right-hand side is always positive because $p'(\tau) < 0$, therefore:

$$u = u_L \pm \sqrt{-(p(\tau) - p(\tau_L))(\tau - \tau_L)}. \tag{3.12}$$

In order to chose the sign in (3.12), we use the Lax conditions:

$$\lambda_1(\mathbf{U}) < \sigma < \lambda_1(\mathbf{U}_L),$$

which implies that σ is negative (because λ_1 is negative), and:

$$-\sqrt{-p'(\tau)} < -\sqrt{-p'(\tau_L)} \Rightarrow \tau < \tau_L,$$

where we used the hypothesis $p''(\tau) > 0$. Both σ and $(\tau - \tau_L)$ are negative, therefore their product is positive and from the first equation of (3.11) we can say that $u_L > u$. We can conclude that the sign in (3.12) is a minus and therefore the points reachable from $\mathbf{U}_L$ through a shock are on the curve of equation:

$$u = u_L - \sqrt{-(p(\tau) - p(\tau_L))(\tau - \tau_L)}. \tag{3.13}$$

Repeating everything for the 2−wave one finds that the states $\mathbf{U}$ reachable from $\mathbf{U}_R$ through a shock are those on the curve of equation:

$$u = u_R + \sqrt{-(p(\tau) - p(\tau_R))(\tau - \tau_R)}. \tag{3.14}$$

We now repeat the same procedure to compute the set of points reachable through a rarefaction wave from $\mathbf{U}_L$ and from $\mathbf{U}_R$ as we did for the shock.

The self-similar solution $\xi \mapsto \widetilde{\mathbf{U}}_1(\xi)$ in a 1−rarefaction wave starting from $\mathbf{U}_L$ respects the following equation:

$$\begin{cases} \widetilde{\mathbf{U}}_1'(\xi) = \mathbf{r}_1(\widetilde{\mathbf{U}}_1(\xi)), \\ \widetilde{\mathbf{U}}_1(\xi_0) = \mathbf{U}_L, \end{cases} \tag{3.15}$$

with $\xi = \lambda_1(\mathbf{U}(\xi))$ for all ξ, which gives the following system:

$$\begin{cases} \tilde{\tau}_1'(\xi) = \dfrac{2\sqrt{-p'(\tilde{\tau}_1)}}{p''(\tilde{\tau}_1)}, & \tilde{\tau}_1(\xi_0) = \tau_L, \\ \tilde{u}_1'(\xi) = -\dfrac{2p'(\tilde{\tau}_1)}{p''(\tilde{\tau}_1)}, & \tilde{u}_1(\xi_0) = u_L. \end{cases} \tag{3.16}$$

The first equation can be rewritten as

$$(p'(\tilde{\tau}_1))' = 2\sqrt{-p'(\tilde{\tau}_1)},$$

and its solution is $p'(\tilde{\tau}_1) = -\xi^2$, therefore:

$$\tilde{\tau}_1 = (p')^{-1}(-\xi^2).$$

After a change of variable, the second equation is written as:

$$\frac{d\tilde{u}_1}{d\tilde{\tau}_1} = \sqrt{-p'(\tilde{\tau}_1)},$$

hence, one has the curve of points reachable through a 1−rarefaction starting from $\mathbf{U}_L$:

$$\tilde{u}_1 = u_L + \int_{\tau_L}^{\tilde{\tau}_1} \sqrt{-p'(\tau)}\, d\tau.$$

If $p(\tau) = \tau^{-\gamma}$, one would find:

$$\tilde{\tau}_1 = \left(\frac{\gamma t^2}{(x - x_c)^2}\right)^{\frac{1}{\gamma+1}}, \qquad \tilde{u}_1 = u_L + \frac{2\sqrt{\gamma}}{1-\gamma}\left(\tilde{\tau}_1^{\frac{1-\gamma}{2}} - \tau_L^{\frac{1-\gamma}{2}}\right). \tag{3.17}$$

Repeating exactly the same procedure, one can find the points reachable through a 2-rarefaction starting from $\mathbf{U}_R$:

$$\tilde{u}_2 = u_R + \int_{\tilde{\tau}_2}^{\tau_R} \sqrt{-p'(\tau)}\, d\tau,$$

hence for our choice of p:

$$\tilde{u}_2 = u_R + \frac{2\sqrt{\gamma}}{1-\gamma}\left(\tau_R^{\frac{1-\gamma}{2}} - \tilde{\tau}_2^{\frac{1-\gamma}{2}}\right), \quad \tilde{\tau}_2 = \left(\frac{\gamma t^2}{(x - x_c)^2}\right)^{\frac{1}{\gamma+1}}. \tag{3.18}$$

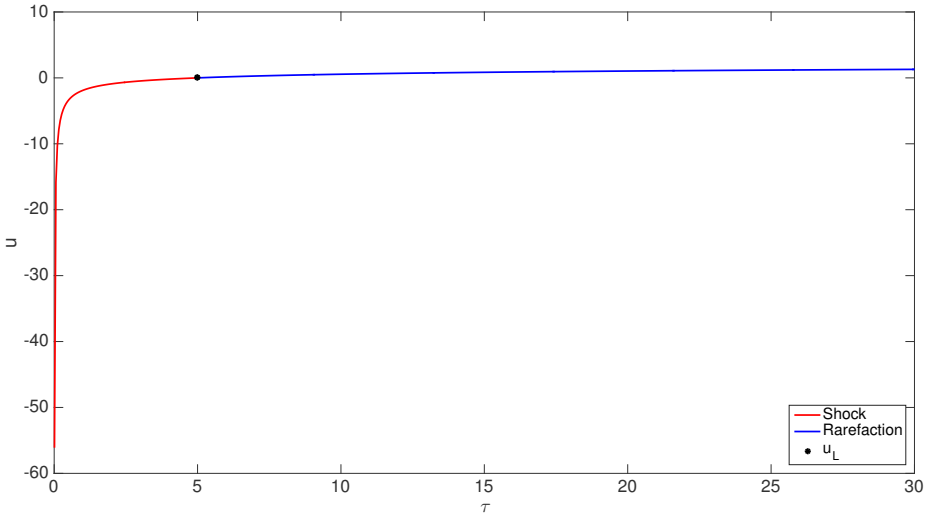


Figure 3.2 – Set of points reachable from u_L through a shock (in red) or through a rarefaction (in blue).

Finally, in order to compute $\mathbf{U}^*$ we define the two following 1- and 2-wave curves:

$$g_1(\tau;\mathbf{U}_L) = \begin{cases} u_L - \sqrt{-(\tau^{-\gamma} - \tau_L^{-\gamma})(\tau - \tau_L)} & \text{if } \tau \leq \tau_L, \\ u_L + \frac{2\sqrt{\gamma}}{1-\gamma}(\tau^{\frac{1-\gamma}{2}} - \tau_L^{\frac{1-\gamma}{2}}) & \text{if } \tau > \tau_L, \end{cases} \tag{3.19}$$

$$g_2(\tau;\mathbf{U}_R) = \begin{cases} u_R + \sqrt{-(\tau^{-\gamma} - \tau_R^{-\gamma})(\tau - \tau_R)} & \text{if } \tau \leq \tau_R, \\ u_R + \frac{2\sqrt{\gamma}}{1-\gamma}(\tau_R^{\frac{1-\gamma}{2}} - \tau^{\frac{1-\gamma}{2}}) & \text{if } \tau > \tau_R, \end{cases} \tag{3.20}$$

which are smooth functions whose derivatives with respect to τ will be denoted g_i'. The intermediate state τ^* is defined as the intersection between g_1 and g_2, and one has: $u^* = g_1(\tau^*;\mathbf{U}_L) = g_2(\tau^*;\mathbf{U}_R)$. Newton's method can be used to compute τ^*. We remark that there is no intersection between g_1 and g_2 under the following condition:

$$u_L - u_R < \frac{2\sqrt{\gamma}}{1-\gamma}\left(\tau_L^{\frac{1-\gamma}{2}} + \tau_R^{\frac{1-\gamma}{2}}\right).$$

3.3.2 The sensitivity variable

As already explained, to compute the sensitivity we differentiate with respect to a the state solution, this means that the initial data for the state and the sensitivity are linked by the following relation:

$$\mathbf{U}_{a,L} = \frac{\partial \mathbf{U}_L}{\partial a} \qquad \mathbf{U}_{a,R} = \frac{\partial \mathbf{U}_R}{\partial a}.$$

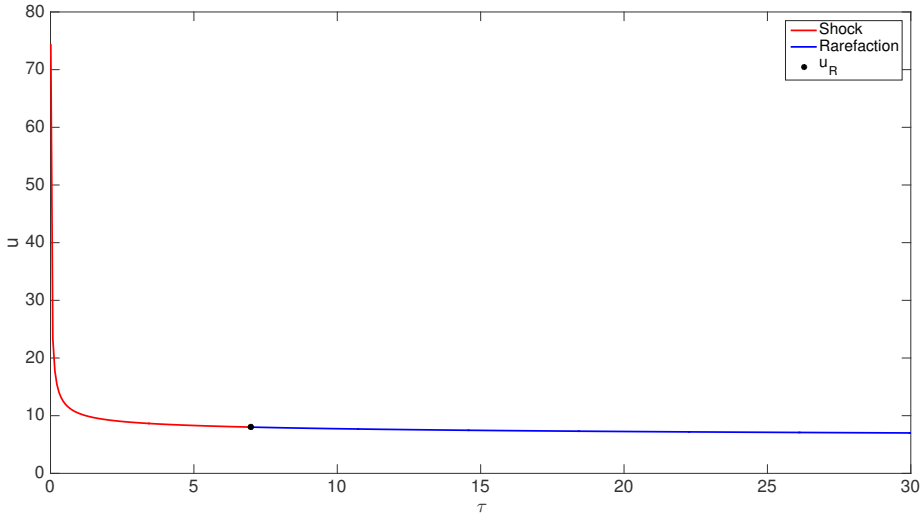


Figure 3.3 – Set of points reachable from u_R through a shock (in red) or through a rarefaction (in blue).

Furthermore, the sensitivity has the same two-wave structure as the state and the waves travel at the same speed as for the state. Therefore, we need to compute the derivative of $\mathbf{U}^*$ and $\widetilde{\mathbf{U}}$ and this concludes the computation of the analytical sensitivity. To compute τ_a^* and u_a^*, we differentiate, with respect to a, the following equality:

$$g_1(\tau^*; \mathbf{U}_L) = g_2(\tau^*; \mathbf{U}_R)$$

and we obtain (recall that $\mathbf{U}_L = (\tau_L, u_L)^T$ and $\mathbf{U}_R = (\tau_R, u_R)^T$):

$$\begin{aligned} u_a^* &= g_1'(\tau^*; \mathbf{U}_L)\tau_a^* + \frac{\partial g_1}{\partial \tau_L}(\tau^*; \mathbf{U}_L)\tau_{a,L} + \frac{\partial g_1}{\partial u_L}(\tau^*; \mathbf{U}_L)u_{a,L} = \\ &= g_2'(\tau^*; \mathbf{U}_R)\tau_a^* + \frac{\partial g_2}{\partial \tau_R}(\tau^*; \mathbf{U}_R)\tau_{a,R} + \frac{\partial g_2}{\partial u_R}(\tau^*; \mathbf{U}_R)u_{a,R}, \end{aligned}$$

which gives an explicit solution (although dependent on $\mathbf{U}^*$) for τ_a^*:

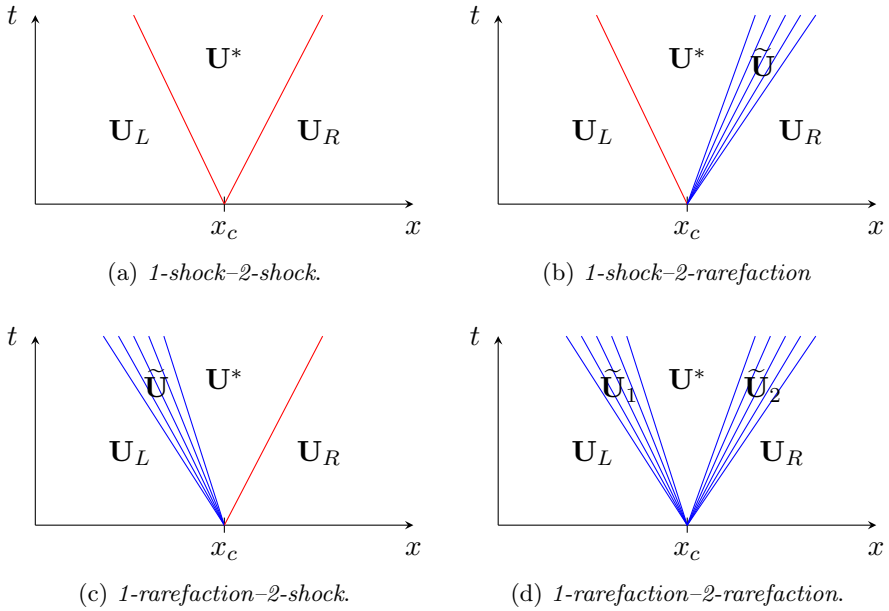
$$\tau_a^* = \frac{\frac{\partial g_2}{\partial \tau_R}(\tau^*; \mathbf{U}_R)\tau_{a,R} + \frac{\partial g_2}{\partial u_R}(\tau^*; \mathbf{U}_R)u_{a,R} - \frac{\partial g_1}{\partial \tau_L}(\tau^*; \mathbf{U}_L)\tau_{a,L} - \frac{\partial g_1}{\partial u_L}(\tau^*; \mathbf{U}_L)u_{a,L}}{g_1'(\tau^*; \mathbf{U}_L) - g_2'(\tau^*; \mathbf{U}_R)}.$$

Finally, we differentiate the state solution in the rarefaction $\widetilde{\mathbf{U}}$ given by (3.17)-(3.18).

3.3.3 Examples

In the numerical section of this work, we will consider as a parameter of interest the initial data, which means that a can either be τ_L, u_L, τ_R, u_R or a combination of them, and from (3.17) and (3.18) one can observe that $\tilde{\tau}_i$ does not depend on those parameters, therefore:

$$\tilde{\tau}_{a,i} = \frac{\partial \tilde{\tau}_i}{\partial a} = 0 \quad i = 1, 2.$$



(a) *1-shock–2-shock.* (b) *1-shock–2-rarefaction*

(c) *1-rarefaction–2-shock.* (d) *1-rarefaction–2-rarefaction.*

Figure 3.4 – Configurations for the state variable $\mathbf{U}$.

Concerning the sensitivity of the velocity, one obtains:

$$\tilde{u}_{a,1} = \frac{\partial \tilde{u}_1}{\partial a} = u_{a,L} - \sqrt{\gamma}\tau_L^{-\frac{1+\gamma}{2}}\tau_{a,L}, \qquad \tilde{u}_{a,2} = \frac{\partial \tilde{u}_2}{\partial a} = u_{a,R} + \sqrt{\gamma}\tau_R^{-\frac{1+\gamma}{2}}\tau_{a,R}.$$

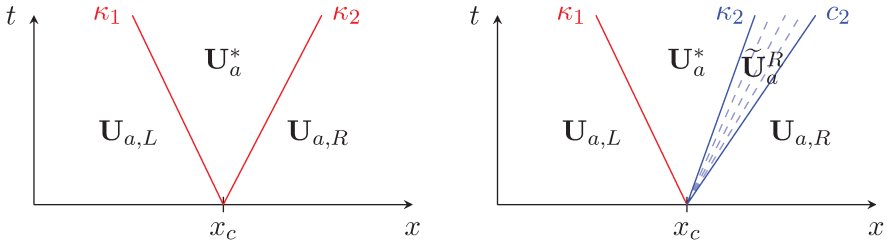
Interestingly, we remark that the sensitivity is constant in the rarefaction zone of the state variable, which means that for the sensitivity this zone corresponds to at most two discontinuities propagating with velocities given by the extreme left and right velocities of the rarefaction in the state variable, see Figure 3.5. This simplification is due to the fact that we are considering a reduced Euler system, under barotropic conditions (cf. [GLC07]). In particular, there are two cases:

(i) if the state presents a $1-$rarefaction (respectively a $2-$rarefaction) and the parameter of interest a is τ_L (respectively, τ_R), the wave associated with the rarefaction in the sensitivity splits in two discontinuities, as explained above (cf. Figure 3.6).

(ii) if the parameter of interest is u_L (or u_R) we have $\tilde{u}_{a,1} = u_{a,L}$ and $\tilde{u}_{a,2} = u_{a,R}$, therefore the wave associated with the rarefaction becomes a single discontinuity for the sensitivity, travelling at the more internal velocity of the state rarefaction wave (cf. Figure 3.7).

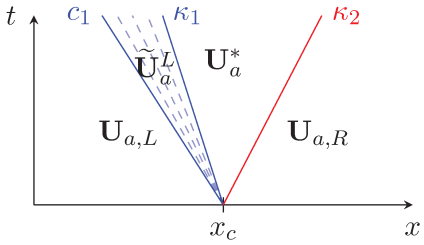
3.4 Classical numerical schemes

The aim of this section is to design relevant numerical schemes for (3.6). As we will see, this task is not easy and requires a nice discretisation of $\mathbf{S}$ in order to avoid Dirac delta functions and it is necessary to control numerical diffusion across the shocks where this term is active. Only under these conditions will we get a perfect agreement between exact

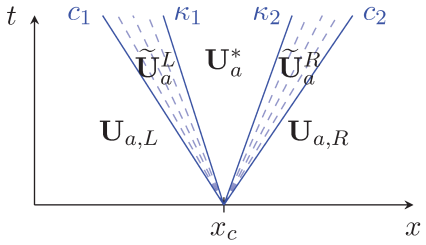


(a) *1-shock–2-shock.*

(b) *1-shock–2-rarefaction.*

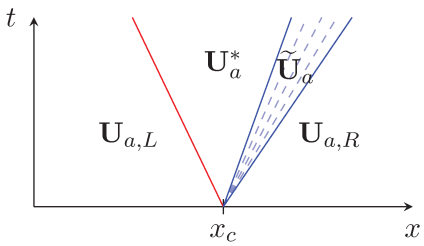


(c) *1-rarefaction–2-shock.*

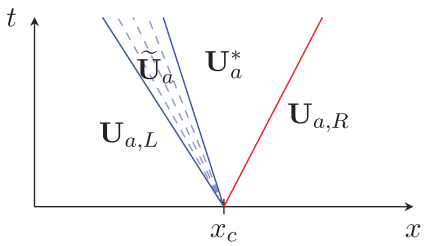


(d) *1-rarefaction–2-rarefaction.*

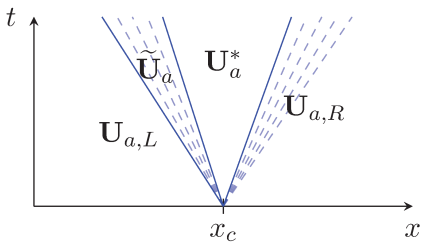
Figure 3.5 – Corresponding configurations for the sensitivity $\mathbf{U}_a$.



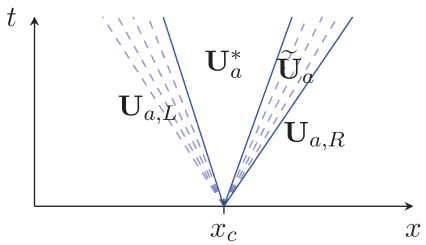
(a) *Corresponding configuration to state case (b)-Figure 3.4 if* $a = \tau_R$.



(b) *Corresponding configuration to state case (c)-Figure 3.4 if* $a = \tau_L$.

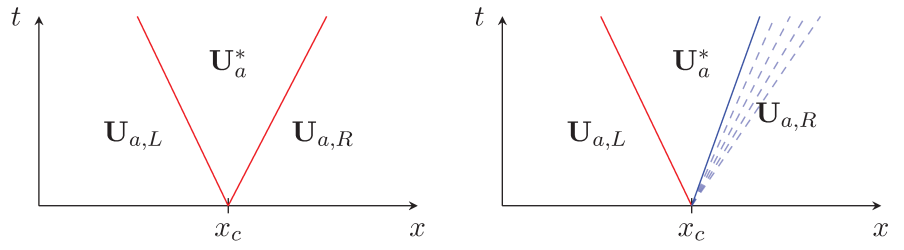


(c) *Corresponding configuration to state case (d)-Figure 3.4 if* $a = \tau_L$.



(d) *Corresponding configuration to state case (d)-Figure 3.4 if* $a = \tau_R$.

Figure 3.6 – Corresponding configurations for the sensitivity $\mathbf{U}_a$, example (i).



(a) *Corresponding configuration to state case (a)-Figure 3.4 if* $a = u_L$ *or* $a = u_R$.

(b) *Corresponding configuration to state case (b)-Figure 3.4 if* $a = u_L$ *or* $a = u_R$.

(c) *Corresponding configuration to state case (c)-Figure 3.4 if* $a = u_L$ *or* $a = u_R$.

(d) *Corresponding configuration to state case (d)-Figure 3.4 if* $a = u_L$ *or* $a = u_R$.

Figure 3.7 – Corresponding configurations for the sensitivity $\mathbf{U}_a$, example (ii).

and numerical solutions. Let us first introduce our notation, although quite classical: we use a constant space step Δx and a varying time step Δt^n. The mesh interfaces are denoted $x_{j+1/2} = j\Delta x$, the cells $C_j = [x_{j-1/2}, x_{j+1/2}]$, the cell centres x_j and the intermediate times $t^{n+1} = t^n + \Delta t^n$, where Δt^n is chosen according to the usual CFL condition. In the following subsections, we will briefly introduce two classical schemes for the state, the Godunov and Roe methods, and we will adapt them to the sensitivity.

3.4.1 The Godunov method

In this section, we present an exact Godunov-type method. Since the state equations (3.1) are conservative, the classic update formula can be used:

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}(\mathbf{F}(\mathbf{U}_{j+1/2}^*) - \mathbf{F}(\mathbf{U}_{j-1/2}^*)), \tag{3.21}$$

where $\mathbf{U}_{j-1/2}^*$ is the exact intermediate state known implicitly from (3.19)-(3.20), with $\mathbf{U}_L = \mathbf{U}_{j-1}$ and $\mathbf{U}_R = \mathbf{U}_j$ [Tor13].

The update formula (3.21) cannot be applied to the sensitivity variable, because of the source term. However, as explained in section 3.3.2, the structure of the sensitivity is made of discontinuities only. Therefore, we can directly compute the average on each cell, if the slopes of the red lines and the solid blue lines in Figure 3.5 are known at each interface $j - 1/2$. The slopes of the red lines are computed from the Rankine-Hugoniot conditions, while the ones of the blue lines are the eigenvalues evaluated in the correct

state. We obtain the following formulas (cf. Figure 3.5 for the notations):

$$\kappa_{1,j-1/2} = \begin{cases} \frac{u^*_{j-1/2}-u_{j-1}}{\tau_{j-1}-\tau^*_{j-1/2}} & \text{if 1-shock at interface } j-1/2, \\ \lambda_1(\mathbf{U}^*_{j-1/2}) & \text{if 1-rarefaction at interface } j-1/2, \end{cases}$$

$$\kappa_{2,j-1/2} = \begin{cases} \frac{u^*_{j-1/2}-u_{j}}{\tau_{j}-\tau^*_{j-1/2}} & \text{if 2-shock at interface } j-1/2, \\ \lambda_2(\mathbf{U}^*_{j-1/2}) & \text{if 2-rarefaction at interface } j-1/2, \end{cases}$$

$$c_{1,j-1/2} = \begin{cases} \kappa_{1,j-1/2} & \text{if 1-shock at interface } j-1/2, \\ \lambda_1(\mathbf{U}_{j-1}) & \text{if 2-rarefaction at interface } j-1/2, \end{cases}$$

$$c_{2,j-1/2} = \begin{cases} \kappa_{2,j-1/2} & \text{if 2-shock at interface } j-1/2, \\ \lambda_2(\mathbf{U}_{j}) & \text{if 2-rarefaction at interface } j-1/2. \end{cases}$$

Then, the update formula for the sensitivity becomes:

$$\begin{aligned} \mathbf{U}^{n+1}_{a,j} = \mathbf{U}^n_{a,j} + \frac{\Delta t}{\Delta x} \Big(& \kappa_{2,j-1/2}(\mathbf{U}^*_{a,j-1/2} - \widetilde{\mathbf{U}}^R_{a,j-1/2}) + c_{2,j-1/2}(\widetilde{\mathbf{U}}^R_{a,j-1/2} - \mathbf{U}^n_{a,j}) \\ & - \kappa_{1,j+1/2}(\mathbf{U}^*_{a,j+1/2} - \widetilde{\mathbf{U}}^L_{a,j+1/2}) - c_{1,j-1/2}(\widetilde{\mathbf{U}}^L_{a,j+1/2} - \mathbf{U}^n_{a,j}) \Big), \end{aligned} \tag{3.22}$$

where the intermediate states $\mathbf{U}^*_{a,j-1/2}$, $\mathbf{U}^*_{a,j+1/2}$, $\widetilde{\mathbf{U}}^R_{a,j-1/2}$ and $\widetilde{\mathbf{U}}^L_{a,j-1/2}$ are known analytically, from section 3.3.2. We remark that the source term is encompassed in (3.22), since (3.22) comes from the exact Riemann solver of (3.6).

3.4.2 A Roe-type method

First order

In this section we illustrate a Roe-type Riemann solver, consisting of three constant states (which we denote $\mathbf{U}_L$, $\mathbf{U}^*$ and $\mathbf{U}_R$ for the state, and $\mathbf{U}_{a,L}$, $\mathbf{U}^*_a$ and $\mathbf{U}_{a,R}$ for the sensitivity), connected by two discontinuities travelling at velocities

$$\lambda^{ROE}_{L,j-1/2} = -\sqrt{-\frac{p(\tau^n_{j-1}) - p(\tau^n_j)}{\tau^n_{j-1} - \tau^n_j}}, \qquad \lambda^{ROE}_{R,j-1/2} = \sqrt{-\frac{p(\tau^n_{j-1}) - p(\tau^n_j)}{\tau^n_{j-1} - \tau^n_j}}$$

if $\tau^n_{j-1} \neq \tau^n_j$ and $\mp\sqrt{-p'(\tau^n_j)}$ otherwise. In the following, we will use the notation $\lambda^{ROE}_{j-1/2} = \lambda^{ROE}_{R,j-1/2} = -\lambda^{ROE}_{L,j-1/2}$. The Harten, Lax and van Leer consistency relations [HLVL97] for the state at the interface $j-1/2$ are given as:

$$\mathbf{U}^*_{j-1/2} = \frac{1}{2}(\mathbf{U}^n_{j-1} + \mathbf{U}^n_j) - \frac{\mathbf{F}(\mathbf{U}^n_j) - \mathbf{F}(\mathbf{U}^n_{j-1})}{2\lambda^{ROE}_{j-1/2}}. \tag{3.23}$$

Since $\mathbf{U}^*_{j-1/2}$ and $\lambda^{ROE}_{j-1/2}$ are known at each interface, we can write the following update formula for the state:

$$\mathbf{U}^{n+1}_j = \mathbf{U}^n_j + \frac{\Delta t}{\Delta x}(\lambda^{ROE}_{j-1/2}(\mathbf{U}^*_{j-1/2} - \mathbf{U}^n_j) + \lambda^{ROE}_{j+1/2}(\mathbf{U}^*_{j+1/2} - \mathbf{U}^n_j)). \tag{3.24}$$

Writing the integral conditions for the sensitivity with the source term, one obtains:

$$\mathbf{F}_a(\mathbf{U}^n_j,\mathbf{U}^n_{a,j})-\mathbf{F}_a(\mathbf{U}^n_{j-1},\mathbf{U}^n_{a,j-1})-\Delta x\mathbf{S}^n_{j-1/2}=\lambda^{ROE}_{j-1/2}(\mathbf{U}^n_{a,j}+\mathbf{U}^n_{a,j-1})-2\lambda^{ROE}_{j-1/2}\mathbf{U}^*_{a,j-1/2},$$

from which we have the following form for $\mathbf{U}^*_{a,j-1/2}$:

$$\mathbf{U}^*_{a,j-1/2}=\frac{\mathbf{U}^n_{a,j-1}+\mathbf{U}^n_{a,j}}{2}-\frac{\mathbf{F}_a(\mathbf{U}^n_j,\mathbf{U}^n_{a,j})-\mathbf{F}_a(\mathbf{U}^n_{j-1},\mathbf{U}^n_{a,j})}{2\lambda^{ROE}_{j-1/2}}+\frac{\Delta x\mathbf{S}^n_{j-1/2}}{2\lambda^{ROE}_{j-1/2}}.$$

The source term is discretised as follows:

$$\mathbf{S}^n_{j-1/2}=\lambda^{ROE}_{a,j-1/2}\left((\mathbf{U}^n_{j-1}-\mathbf{U}^*_{j-1/2})\frac{d_{1,j-1}}{\Delta x}+(\mathbf{U}^n_j-\mathbf{U}^*_{j-1/2})\frac{d_{2,j}}{\Delta x}\right), \tag{3.25}$$

where $d_{\ell,j}$ is a shock detector which is equal to 1 if there is an $\ell-$shock in the $j-$th cell, it is zero elsewhere and $d_{\ell,j}/\Delta x$ approximates numerically the Dirac δ_k in the definition of the source term (3.5). In this work, we use a very simple shock detector: in section 3.3.1 we showed that the velocity u is decreasing across a shock, whilst the co-volume τ decreases across a $1-$shock, and it increases across a $2-$shock. Based on this, we set:

$$d_{1,j}=\begin{cases}1 & \text{if } u_j>u^*_{j+1/2} \text{ and } \tau_j>\tau^*_{j+1/2},\\ 0 & \text{otherwise,}\end{cases}$$

$$d_{2,j}=\begin{cases}1 & \text{if } u_j<u^*_{j-1/2} \text{ and } \tau_j>\tau^*_{j-1/2},\\ 0 & \text{otherwise.}\end{cases}$$

Finally, $\mathbf{U}^*_{a,j-1/2}$ is computed as follows:

$$\begin{aligned}\mathbf{U}^*_{a,j-1/2}=&\frac{1}{2}(\mathbf{U}^n_{a,j-1}+\mathbf{U}^n_{a,j})-\frac{\mathbf{F}_a(\mathbf{U}^n_j,\mathbf{U}^n_{a,j})-\mathbf{F}_a(\mathbf{U}^n_{j-1},\mathbf{U}^n_{a,j})}{2\lambda^{ROE}_{j-1/2}}\\&+\frac{\lambda^{ROE}_{a,j-1/2}}{2\lambda^{ROE}_{j-1/2}}\left((\mathbf{U}^n_{j-1}-\mathbf{U}^*_{j-1/2})d_{1,j-1}+(\mathbf{U}^n_j-\mathbf{U}^*_{j-1/2})d_{2,j}\right).\end{aligned} \tag{3.26}$$

We remark that the discretisation of the source term (3.25) is such that $\mathbf{U}^*_{a,j-1/2}=\partial_a\mathbf{U}^*_{j-1/2}$, in fact differentiating (3.23) with respect to a, one finds:

$$\partial_a\mathbf{U}^*_{j-1/2}=\frac{\mathbf{U}^n_{a,j-1}+\mathbf{U}^n_{a,j}}{2}-\frac{\mathbf{F}_a(\mathbf{U}^n_j,\mathbf{U}^n_{a,j})-\mathbf{F}_a(\mathbf{U}^n_{j-1},\mathbf{U}^n_{a,j})}{2\lambda^{ROE}_{j-1/2}}+\frac{\mathbf{F}(\mathbf{U}^n_j)-\mathbf{F}(\mathbf{U}^n_{j-1})}{2\lambda^{ROE}_{j-1/2}}\frac{\lambda^{ROE}_{a,j-1/2}}{\lambda^{ROE}_{j-1/2}}.$$

Using again (3.23) one has:

$$\mathbf{U}^*_a=\frac{\mathbf{U}^n_{a,j-1}+\mathbf{U}^n_{a,j}}{2}-\frac{\mathbf{F}_a(\mathbf{U}^n_j,\mathbf{U}^n_{a,j})-\mathbf{F}_a(\mathbf{U}^n_{j-1},\mathbf{U}^n_{a,j})}{2\lambda^{ROE}_{j-1/2}}+\frac{\lambda^{ROE}_{a,j-1/2}}{2\lambda^{ROE}_{j-1/2}}(\mathbf{U}^n_{j-1}+\mathbf{U}^n_j-2\mathbf{U}^*_{j-1/2}),$$

which is equal to (3.26), once the shock detectors are added. Furthermore, the definition (3.26) encompasses the source term, which means that we can use the update formula (3.24) for the sensitivity, too.

Finally, in order to prepare the second order extension in space, we define a residual as follows:

$$\mathbf{R}_j^I(\mathbf{U}^n) = \lambda_{j-1/2}^{ROE}(\mathbf{U}_{j-1/2}^* - \mathbf{U}_j^n) + \lambda_{j+1/2}^{ROE}(\mathbf{U}_{j+1/2}^* - \mathbf{U}_j^n), \tag{3.27}$$

where $\mathbf{R}_j^I(\mathbf{U}^n)$ is a more compact notation for $\mathbf{R}^I(\mathbf{U}_{j-1}, \mathbf{U}_j, \mathbf{U}_{j+1})$. This allows us to write the update formulas in the following way:

$$\begin{cases} \mathbf{U}_j^{n+1} = \mathbf{U}_j^n + \frac{\Delta t}{\Delta x}\mathbf{R}_j^I(\mathbf{U}^n), \\ \mathbf{U}_{a,j}^{n+1} = \mathbf{U}_{a,j}^n + \frac{\Delta t}{\Delta x}\mathbf{R}_j^I(\mathbf{U}_a^n). \end{cases}$$

Furthermore, it will be useful for the numerical schemes introduced hereafter.

Second order

We extend this scheme to the second order: for the time discretisation we use a two-step Runge-Kutta method, whilst in space we propose a MUSCL-type scheme with some minor modifications in order to have a second order discretisation of the source term. In particular, we remark that (3.10) is valid only if the solution $\mathbf{U}$ is locally constant to the left and to the right of the shock, which is true for a first order approximation but not for a second order, in which, classically, the numerical solution is a piecewise affine function. To overcome this problem, we suggest to consider the numerical solution to be a piecewise constant function on half of every cell (cf. [Bou04], section 2.8): the value in the left half (respectively right half) of the $j-$th cell is denoted $U_{j-1/4}$ (respectively $U_{j+1/4}$), as shown in Figure 3.8 and they are computed as in a classical MUSCL approach:

$$\mathbf{U}_{j\pm1/4}^n = \mathbf{U}_j \pm \Delta\mathbf{U}_j^n,$$

and a usual choice for $\Delta\mathbf{U}_j^n$ is given by a slope-limiter procedure. In this work we use the so-called minmod limiter:

$$\Delta\mathbf{U}^n = \frac{1}{2}\text{minmod}(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n, \mathbf{U}_j^n - \mathbf{U}_{j-1}^n),$$

where the function minmod is defined as follows:

$$\text{minmod}(a,b) = \begin{cases} \text{sgn}(a)\min(|a|,|b|) & \text{if } ab > 0, \\ 0 & \text{otherwise.} \end{cases}$$

This interpretation of the second order allows us to define the source term as we did for the first order, however we need to consider an additional Riemann problem for each cell. This leads to the following definition of the residual:

$$\mathbf{R}_j^{II}(\mathbf{U}^n) = (\lambda_{j-1/2}^{ROE}(\mathbf{U}_{j-1/2}^* - \mathbf{U}_{j-1/4}^n) + \lambda_{j+1/2}^{ROE}(\mathbf{U}_{j+1/2}^* - \mathbf{U}_{j+1/4}^n)) + \lambda_j^{ROE}(2\mathbf{U}_j^* - \mathbf{U}_{j-1/4} + \mathbf{U}_{j+1/4}),$$

where all the λ^{ROE} and the $\mathbf{U}^*$ are computed from the extrapolated values $\mathbf{U}_{j\pm1/4}^n$. Finally, the second order scheme is written as:

$$\begin{cases} \mathbf{U}_j^{n+1/2} = \mathbf{U}_j^n + \frac{\Delta t}{2\Delta x}\mathbf{R}^{II}(\mathbf{U}^n), \\ \mathbf{U}_j^{n+1} = \mathbf{U}_j^n + \frac{\Delta t}{\Delta x}\mathbf{R}^{II}(\mathbf{U}^{n+1/2}), \end{cases} \qquad \begin{cases} \mathbf{U}_{a,j}^{n+1/2} = \mathbf{U}_{a,j}^n + \frac{\Delta t}{2\Delta x}\mathbf{R}^{II}(\mathbf{U}_a^n), \\ \mathbf{U}_{a,j}^{n+1} = \mathbf{U}_{a,j}^n + \frac{\Delta t}{\Delta x}\mathbf{R}^{II}(\mathbf{U}_a^{n+1/2}). \end{cases}$$

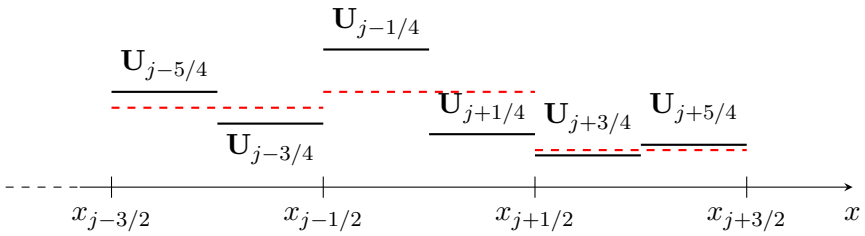


Figure 3.8 – Second order discretisation. In red, the corresponding first order discretisation.

3.5 Numerical results

We present some numerical results obtained with the schemes described in the previous section, on a spatial domain is $(0, 1)$, with final time $T = 0.03$. We consider Riemann problems with $x_c = 0.5$.

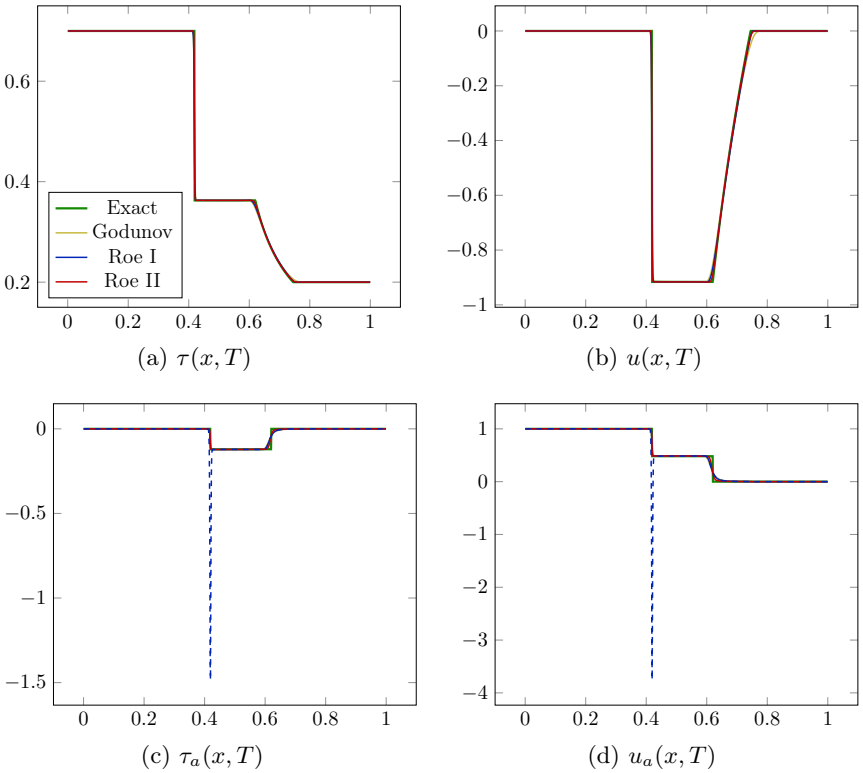
First, we consider a 1-shock–2-rarefaction case, with the following initial conditions for the state:

$$\mathbf{U}_L = \begin{pmatrix} 0.7 \\ 0 \end{pmatrix}, \quad \mathbf{U}_R = \begin{pmatrix} 0.2 \\ 0 \end{pmatrix}.$$

The parameter of interest is $a = u_L$, so that the initial conditions for the sensitivity are:

$$\mathbf{U}_{a,L} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{U}_{a,R} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Figure 3.9 shows the state variables u and τ and their sensitivities u_a and τ_a at the final time T. Since the state is a quite classical problem, it is not surprising that all the methods provide very similar solutions one to another. As for the sensitivity, we remark that the modified formulation is able to remove the peak which approximate the Dirac delta function, located at $x \approx 0.4$ and evident in the scheme without correction term, whose label is "$\mathbf{S} = \mathbf{0}$" in Figures 3.9-3.10. However, even with the addition of the source term, the sensitivity solution have two issues: first, the discontinuity associated with the state rarefaction is not well captured; secondly, the value of the plateau in the star zone is not the analytical one. Out of these two problems, the first is the less important one, for two reasons: the fact that the state rarefaction splits into two discontinuity for the sensitivity is typical to the PDEs system considered, it does not happen, for instance, in the case of the complete Euler system; furthermore, the numerical solution converges to the analytical one as Δx goes to 0, meaning that this issue can be solved by using a finer mesh or a higher-order scheme. The second problem is more critical and we believe that numerical diffusion is the cause of it. In Figure 3.11 we plot the convergence curves of the all the schemes for each variable: all the methods converge as expected for the state variable; however, for the sensitivity the error seems to be convergent for coarser meshes, but it reaches a plateau for finer ones. This can be explained if we split the error into two parts: the part concentrated in the rarefaction zone, which is the bigger one in the coarse meshes, converges; however when this part reaches the same order of magnitude as the error in the star zone, which is constant, the plateau is reached.



(a) $\tau(x,T)$ (b) $u(x,T)$

(c) $\tau_a(x,T)$ (d) $u_a(x,T)$

Figure 3.9 – Classical finite volume schemes.

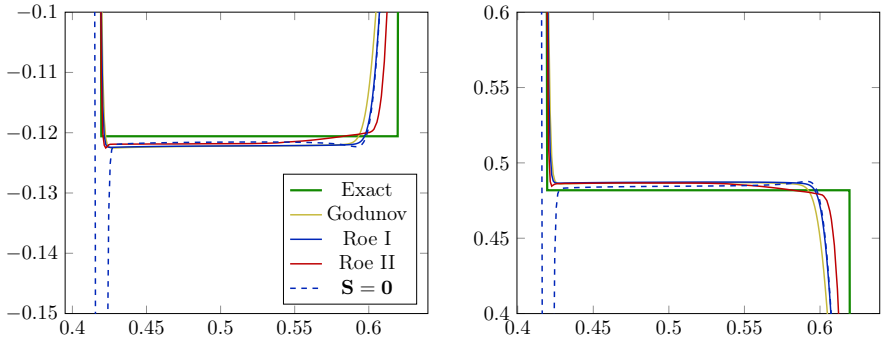
The second test case presented here is an isolated 2-shock for the state as well as for the sensitivity. In order to have an isolated shock we choose the following initial data:

$$\mathbf{U}_L = \begin{pmatrix} 0.2 \\ g_2(\tau_L; \mathbf{U}_R) \end{pmatrix} \simeq \begin{pmatrix} 0.2 \\ -1.56 \end{pmatrix}, \quad \mathbf{U}_R = \begin{pmatrix} 0.5 \\ -3 \end{pmatrix},$$

where g_2 is the $2-$wave curve defined in (3.20). As parameter of interest a we choose the arc length of the curve g_2, which yields the following initial data for the sensitivity:

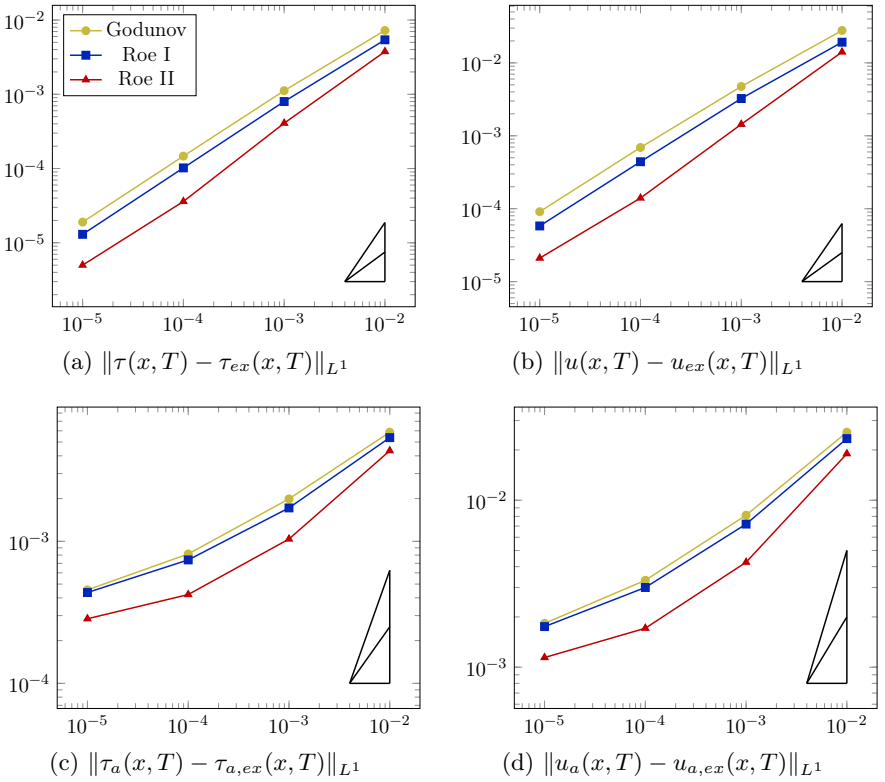
$$\mathbf{U}_{a,L} = \begin{pmatrix} 1 \\ g_2'(\tau_L; \mathbf{U}_R) \end{pmatrix} \simeq \begin{pmatrix} 1 \\ -9.35 \end{pmatrix}, \quad \mathbf{U}_{a,R} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Figure 3.12 shows the results for the state and the sensitivity obtained with a mesh $\Delta x = 10^{-3}$: one can notice a spurious wave in the state which does not affect the value in the star zone. However, in the sensitivity this spurious wave is amplified; moreover, the value in the star zone is not correct. Considering the fact that the approximate Riemann solver of Roe is exact in the case of an isolated shock (as well as the exact Godunov solver), the error is necessarily introduced in the average step of the numerical methods and therefore it is due to the numerical diffusion which comes along with the averaging operation. For this reason, in the next section we introduce a scheme without numerical diffusion in the shock.



(a) $\tau_a(x,T)$ (b) $u_a(x,T)$

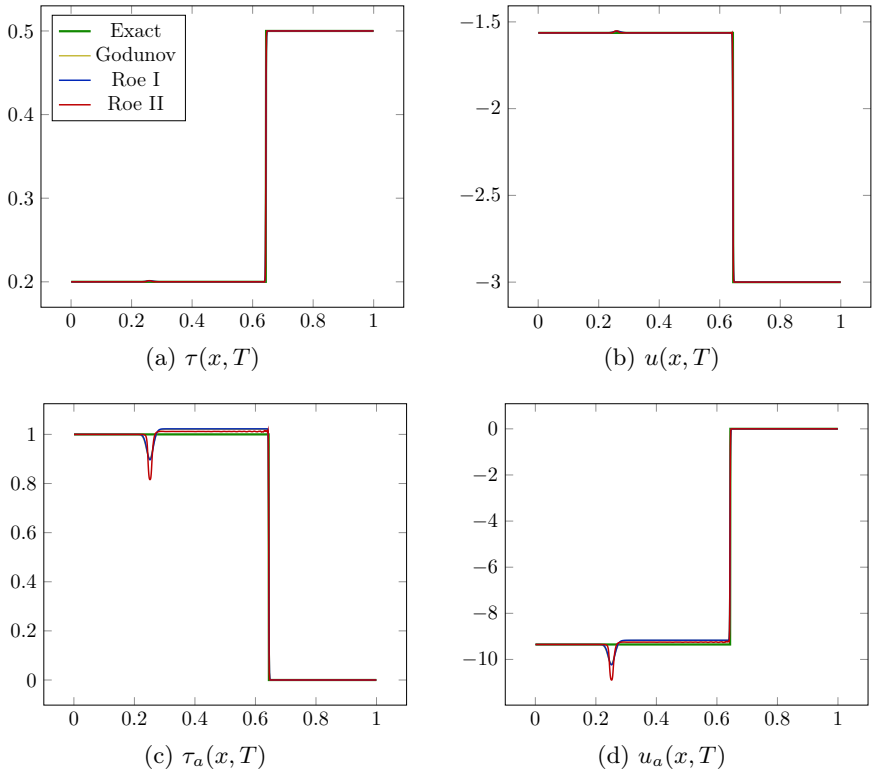
Figure 3.10 – Classical finite volume schemes for sensitivities - zoom.



(a) $\|\tau(x,T) - \tau_{ex}(x,T)\|_{L^1}$ (b) $\|u(x,T) - u_{ex}(x,T)\|_{L^1}$

(c) $\|\tau_a(x,T) - \tau_{a,ex}(x,T)\|_{L^1}$ (d) $\|u_a(x,T) - u_{a,ex}(x,T)\|_{L^1}$

Figure 3.11 – Convergence of the classical finite volume schemes.



(a) $\tau(x,T)$ (b) $u(x,T)$

(c) $\tau_a(x,T)$ (d) $u_a(x,T)$

Figure 3.12 – Test case: isolated shock.

3.6 An anti-diffusive Roe-type numerical scheme

Since we believe that the failure of the previous schemes is caused by the numerical diffusion in the shock, we present a scheme that does not have any numerical diffusion in the shock. The scheme was first introduced in [CG08] and here we adapt it to the sensitivity problem. It is a modified Godunov method and it can be coupled with any Riemann solver, in this work we couple it with the Roe-type method proposed in the previous section. In fact, the first step is to solve a Riemann problem at each interface, as for a standard Godunov method. The difference between the two methods is in the average step: instead of averaging on the cells $[x_{j-1/2}, x_{j+1/2}]$, a new temporary mesh is defined, whose $j-$th cell is denoted $[\overline{x}^n_{j-1/2}, \overline{x}^n_{j+1/2}]$, and the average is performed on this mesh. The new mesh is non uniform and it is defined as follows:

$$\overline{x}^n_{j-1/2} = x_{j-1/2} + \sigma^n_{j-1/2}\Delta t^n,$$

where $\sigma^n_{j-1/2}$ is a proper speed and it depends on the problem. The average operation on the modified mesh provides us with a piecewise constant solution on the new mesh, which we denote $\overline{\mathbf{U}}^{n+1}_j$. The final step of this method is to go back to the initial mesh, i.e. compute $\mathbf{U}^{n+1}_j$ starting from $\overline{\mathbf{U}}^{n+1}_j$, and this is done using a sampling technique: the value of the solution on the $j-$th cell at time t^{n+1}, $\mathbf{U}^{n+1}_j$, is chosen randomly among $\overline{\mathbf{U}}^{n+1}_{j-1}$, $\overline{\mathbf{U}}^{n+1}_j$, and $\overline{\mathbf{U}}^{n+1}_{j+1}$, in agreement with their rate of presence in the cell. More

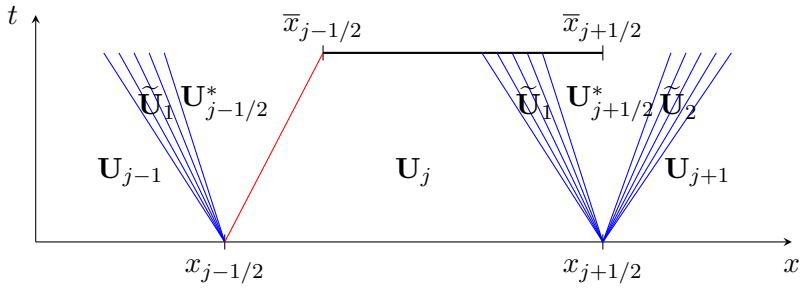


Figure 3.13 – Definition of the temporary staggered mesh.

precisely, given a random sequence (α_n) varying in $(0,1)$, the choice is the following:

$$\mathbf{U}_j^{n+1} = \begin{cases} \overline{\mathbf{U}}_{j-1}^{n+1} & \text{if } \alpha_{n+1} \in \left(0, \frac{\Delta t}{\Delta x}\max(\sigma_{j-1/2}^n, 0)\right), \\ \overline{\mathbf{U}}_j^{n+1} & \text{if } \alpha_{n+1} \in \left[\frac{\Delta t}{\Delta x}\max(\sigma_{j-1/2}^n, 0), 1 + \frac{\Delta t}{\Delta x}\min(\sigma_{j+1/2}^n, 0)\right), \\ \overline{\mathbf{U}}_{j+1}^{n+1} & \text{if } \alpha_{n+1} \in \left[1 + \frac{\Delta t}{\Delta x}\min(\sigma_{j+1/2}^n, 0), 1\right). \end{cases} \tag{3.28}$$

The sampling technique mimics the classical averaging if (α_n) is a well distributed random sequence, for instance $\alpha_n \sim \mathcal{U}(0,1)$, or if it is a deterministic low discrepancy sequence, such as the van der Corput sequence (cf. [CG08]):

$$\alpha_n = \sum_{k=0}^{m} i_k 2^{-(k+1)}, \quad n = \sum_{k=0}^{m} i_k 2^k,$$

where $i_k = 0, 1$ is the binary expansion of the integers.

Our choice for $\sigma_{j+1/2}^n$ is the following:

$$\sigma_{j+1/2}^n = \begin{cases} \lambda_{j+1/2}^n & u_j > u_{j+1} \text{ and } \tau_j < \tau_{j+1}, \\ -\lambda_{j+1/2}^n & u_j > u_{j+1} \text{ and } \tau_j > \tau_{j+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.29}$$

If u is increasing, which means that a rarefaction is expected, the mesh is not modified, whilst for the case of an expected shock the mesh follows it: in this way one never performs the average across a shock and therefore there is no numerical diffusion (cf. Figure 3.13).

Remark. Considering only the initial (non moving) mesh, we remark that this method can also be understood as solving the following two-step problem:

$$\begin{cases} \partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) - \sigma \partial_x \mathbf{U} = 0, \\ \partial_t \mathbf{U} + \sigma \partial_x \mathbf{U} = 0. \end{cases} \tag{3.30}$$

The first step is equivalent to solving the Riemann problems at each interface and performing the average on the initial uniform mesh, whilst the second step is equivalent to the sampling (3.28).

First order formulas

Here, as already said, we couple this anti-diffusive approach with the same Roe-type approximate Riemann solver we presented in the previous section, so we define the following residual:

$$\widetilde{\mathbf{R}}_j^I(\mathbf{U}^n) = \mathbf{R}_j^I(\mathbf{U}^n) - \sigma_{j-1/2}^n \mathbf{U}_{j-1/2}^* + \sigma_{j+1/2}^n \mathbf{U}_{j+1/2}^*,$$

where in the definition of $\sigma_{j-1/2}$ we use $\lambda_{j-1/2} = \lambda_{j-1/2}^{ROE}$. Then, the scheme is written:

$$\begin{cases} \overline{\mathbf{U}}_j^{n+1} = \frac{\Delta x}{\Delta x_j}\mathbf{U}_j^n + \frac{\Delta t}{\Delta x_j}\widetilde{\mathbf{R}}_j^I(\mathbf{U}^n), \\ +(3.28). \end{cases} \tag{3.31}$$

Second order formulas

One can also couple this approach with the second order Roe-type scheme from the previous section. As for the first order, we define the new residual as follows:

$$\widetilde{\mathbf{R}}_j^{II}(\mathbf{U}^n) = \mathbf{R}_j^{II}(\mathbf{U}^n) - \sigma_{j-1/2}^n \mathbf{U}_{j-1/2}^* + \sigma_{j+1/2}^n \mathbf{U}_{j+1/2}^*.$$

Then the scheme writes:

$$\begin{cases} \mathbf{U}_j^{n+1/2} = \mathbf{U}_j^n + \frac{\Delta x - \Delta x_j}{\Delta x_j}\mathbf{U}_j^n + \frac{\Delta t}{2\Delta x_j}\widetilde{\mathbf{R}}_j^{II}(\mathbf{U}^n), \\ \overline{\mathbf{U}}_j^{n+1} = \mathbf{U}_j^n + \frac{\Delta x - \Delta x_j}{\Delta x_j}\mathbf{U}_j^{n+1/2} + \frac{\Delta t}{\Delta x_j}\widetilde{\mathbf{R}}_j^{II}(\mathbf{U}^{n+1/2}), \\ +(3.28). \end{cases} \tag{3.32}$$

From the two-step problem (3.30) point of view, the discretisation (3.32) is a second order discretisation of the first step followed by the second step, i.e. the sampling technique, which remains unvaried.

3.7 Numerical results of the anti-diffusive method

The results of the anti-diffusive method are shown in Figures 3.14-3.15. As one can see, removing the numerical diffusion in the shock for the state variables allows us to be more precise in the definition of the source term which, in turns, provides us with better solution for the sensitivity: the plateau in the star zone is correct. Furthermore, we show in Figure 3.16 the convergence results of the classical Roe-type schemes with diffusion compared to the same schemes without diffusion: the latter show a good convergence rate even for the sensitivity variables. For reference, $\|\tau(x,T)\|_{L^1(0,1)} \simeq 0.4502$, $\|u(x,T)\|_{L^1(0,1)} \simeq 0.2357$, $\|\tau_{u_L}(x,T)\|_{L^1(0,1)} \simeq 0.0242$, and $\|u_{u_L}(x,T)\|_{L^1(0,1)} \simeq 0.5158$.

We now present another test case with initial data:

$$\mathbf{U}_L = \begin{pmatrix} 0.7 \\ 0 \end{pmatrix}, \quad \mathbf{U}_R = \begin{pmatrix} 0.2 \\ 0 \end{pmatrix}, \quad \mathbf{U}_{a,L} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{U}_{a,R} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

therefore the parameter of interest a is in this case τ_R. The initial data for the state is the same as in the previous test case, meaning that we are in configuration (b) of Figure 3.4 and, since $a = \tau_R$, the rarefaction wave splits into two discontinuities for the sensitivity as shown in Figure 3.7-(a). For this test case we chose a bigger final time

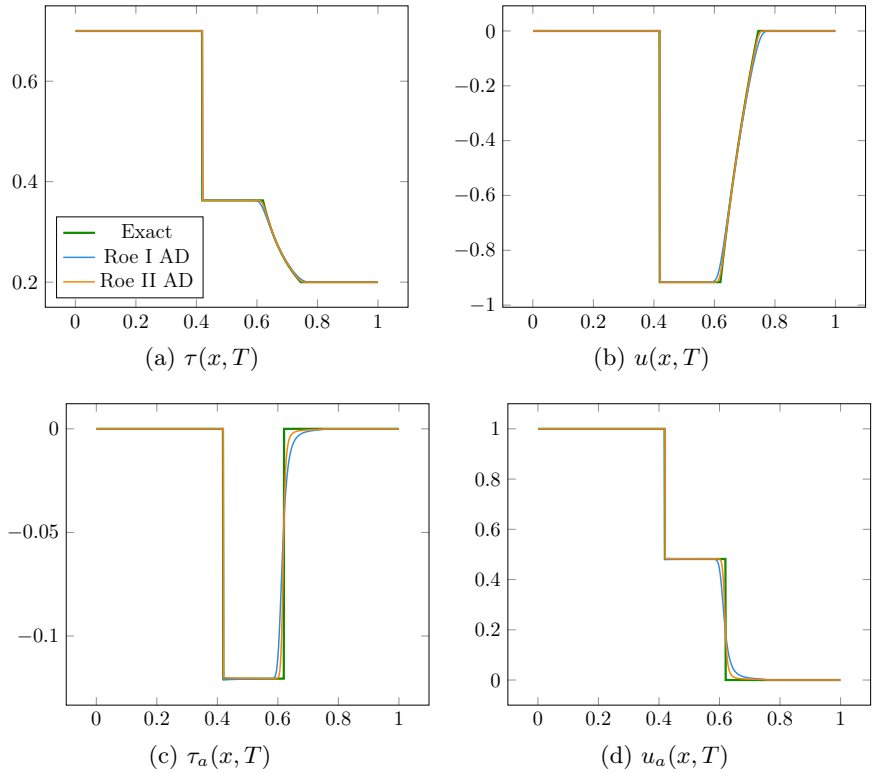
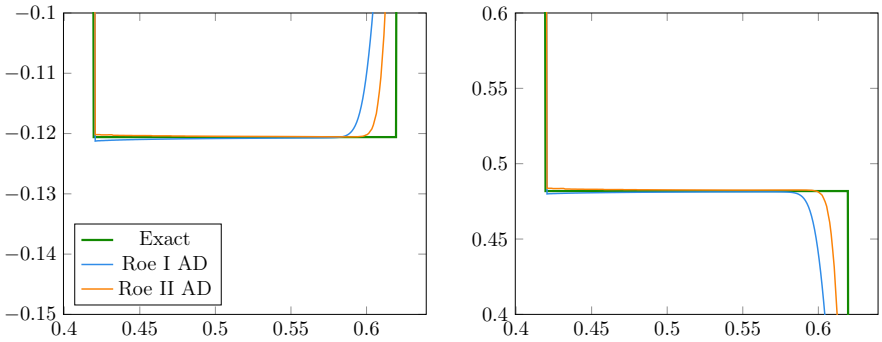


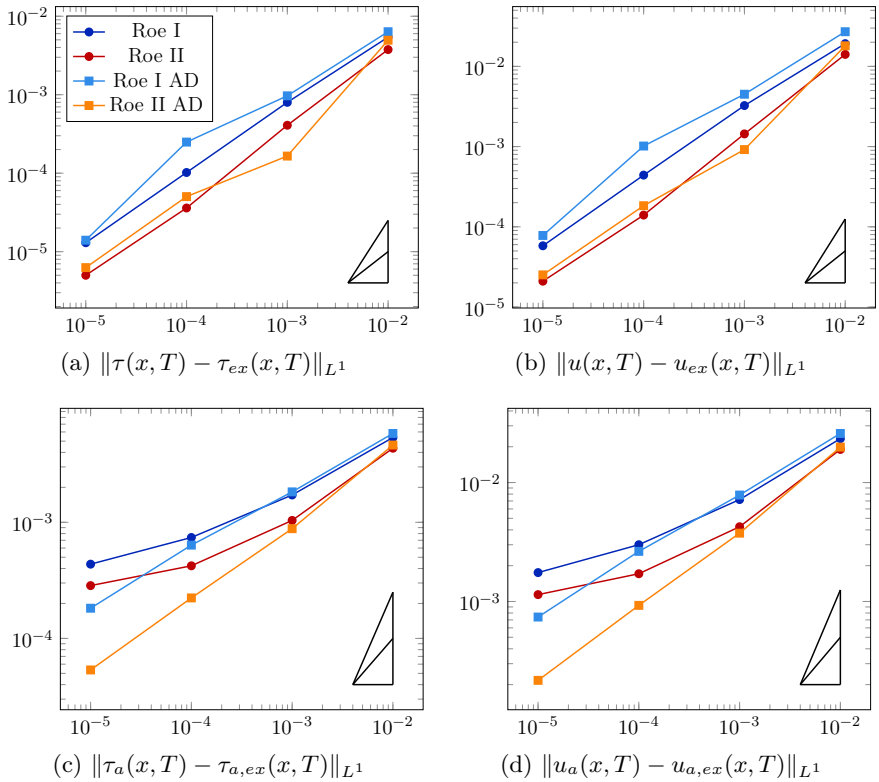
Figure 3.14 – Roe-type schemes without numerical diffusion.

($T = 0.07$) so that the two extremes of the rarefaction wave could be well separated, in order to attenuate the effect of the numerical diffusion in the middle. We also changed the starting point of the discontinuity ($x_c = 0.3$) in order to have the second discontinuity associated with the rarefaction still in the domain at the final time. The results shown in Figure 3.17 are obtained with a mesh $\Delta x = 10^{-4}$: even in this particular case, with three discontinuities, we are able to approximate well the sensitivity provided that the mesh is fine enough.



(a) $\tau_a(x,T)$ (b) $u_a(x,T)$

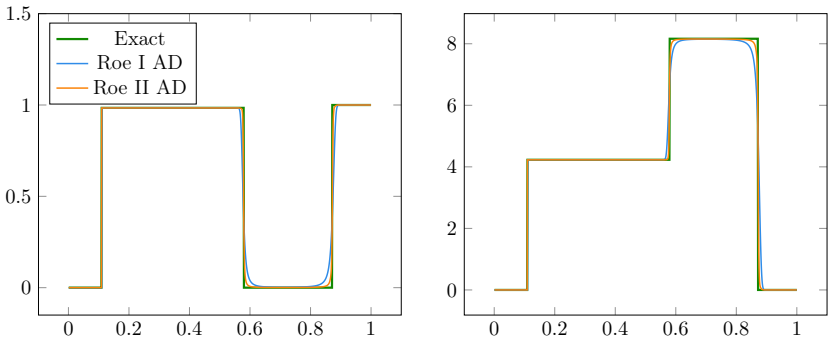
Figure 3.15 – Roe-type schemes without numerical diffusion for the sensitivity - zoom.



(a) $\|\tau(x,T) - \tau_{ex}(x,T)\|_{L^1}$ (b) $\|u(x,T) - u_{ex}(x,T)\|_{L^1}$

(c) $\|\tau_a(x,T) - \tau_{a,ex}(x,T)\|_{L^1}$ (d) $\|u_a(x,T) - u_{a,ex}(x,T)\|_{L^1}$

Figure 3.16 – Convergence of Roe-type schemes, with and without numerical diffusion.



(a) $\tau_a(x, T = 0.07)$ (b) $u_a(x, T = 0.07)$

Figure 3.17 – Test case shock-rarefaction, $a = \tau_R$: sensitivity. $\Delta x = 10^{-4}$, $T = 0.07$.

4 The Euler system

4.1 Introduction

In this chapter we deal with the complete Euler system. In the last part of the chapter an uncertainty quantification problem is defined.

4.1.1 The state system

The Euler system is written:

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) = 0, \\ \partial_t(\rho u) + \partial_x(\rho u^2 + p) = 0, \\ \partial_t(\rho E) + \partial_x(u(\rho E + p)) = 0, \end{cases} \tag{4.1}$$

where ρ is the density, u is the velocity, ρE the total energy per volume unit, and p the pressure. The system is closed by the following algebraic equation:

$$p = (\gamma - 1)\left(\rho E - \frac{1}{2}\rho u^2\right), \tag{4.2}$$

where $\gamma = 1.4$ is the heat capacity ratio. We introduce two other quantities which will be useful in the following: the total enthalpy $H = E + \frac{p}{\rho}$ and the speed of sound $c = \sqrt{(\gamma - 1)(H - \frac{1}{2}u^2)}$. We can rewrite the system (4.1) in the vectorial form:

$$\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = 0, \tag{4.3}$$

where

$$\mathbf{U} = \begin{bmatrix} \rho \\ \rho u \\ \rho E \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ u(\rho E + p) \end{bmatrix} = \begin{bmatrix} w_2 \\ \frac{w_2^2}{w_1} + (\gamma - 1)\left(w_3 - \frac{1}{2}\frac{w_2^2}{w_1}\right) \\ \gamma \frac{w_2 w_3}{w_1} - \frac{(\gamma-1)}{2}\frac{w_2^3}{w_1^2} \end{bmatrix}.$$

One can also write (4.1) in the nonconservative form:

$$\partial_t \mathbf{U} + \mathbf{A}(\mathbf{U})\partial_x \mathbf{U} = 0, \tag{4.4}$$

where the Jacobian matrix $\mathbf{A}$ is:

$$\mathbf{A}(\mathbf{U}) = \frac{\partial \mathbf{F}}{\partial \mathbf{U}} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{\gamma-3}{2}u^2 & (3-\gamma)u & \gamma - 1 \\ \frac{\gamma-2}{2}u^3 - \frac{c^2 u}{\gamma-1} & \frac{3-2\gamma}{2}u^2 + \frac{c^2}{\gamma-1} & \gamma u \end{bmatrix},$$

its eigenvalues are $\lambda_1 = u - c$, $\lambda_2 = u$, and $\lambda_3 = u + c$ and its eigenvectors are:

$$\mathbf{r}_1 = \begin{bmatrix} 1 \\ u - c \\ H - uc \end{bmatrix}, \qquad \mathbf{r}_2 = \begin{bmatrix} 1 \\ u \\ \frac{u^2}{2} \end{bmatrix}, \qquad \mathbf{r}_3 = \begin{bmatrix} 1 \\ u + c \\ H + uc \end{bmatrix}.$$

Therefore $\mathbf{A}$ is $\mathbb{R}$-diagonalisable and the system (4.1) is strictly hyperbolic. At last, (4.3) will be supplemented with a given initial data $\mathbf{U}(x, t = 0) = \mathbf{U}_0(x)$, $\forall x \in \mathbb{R}$.

4.1.2 The sensitivity system

Considering only smooth solutions of (4.1), one can apply the Continuous Sensitivity Equation (CSE) [HEPB04, BB97, DPB06] method which consists in differentiating (4.1) with respect to the parameter of interest a. One can then formally exchange the derivatives in time and space with the ones with respect to a (see [BP02] for the theoretical aspects) and obtain the following sensitivity system:

$$\begin{cases} \partial_t \rho_a + \partial_x (\rho u)_a = 0, \\ \partial_t (\rho u)_a + \partial_x (\rho_a u^2 + 2\rho u u_a + p_a) = 0, \\ \partial_t (\rho E)_a + \partial_x (u_a(\rho E + p) + u((\rho E)_a + p_a)) = 0, \end{cases} \tag{4.5}$$

which can be written in vectorial form as

$$\partial_t \mathbf{U}_a + \partial_x \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = 0, \tag{4.6}$$

where we used the following notation:

$$\mathbf{U}_a = \partial_a \mathbf{U} = \begin{bmatrix} \rho_a \\ (\rho u)_a \\ (\rho E)_a \end{bmatrix}, \qquad \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = \partial_a \mathbf{F}(\mathbf{U}) = \begin{bmatrix} (\rho u)_a \\ \rho_a u^2 + 2\rho u u_a + p_a \\ u_a(\rho E + p) + u((\rho E)_a + p_a) \end{bmatrix}.$$

Note that differentiating (4.2) one has:

$$p_a = (\gamma - 1)((\rho E)_a - \frac{1}{2}\rho_a u^2 - \rho_a u u_a)$$

which acts as a closure relation for (4.5). The initial data for the sensitivity is nothing but $\mathbf{U}_a(x, t = 0) = \partial_a \mathbf{U}_0(x)$.

4.1.3 The global system

In order to write the global system, i.e. the state and sensitivity system, in a more compact way, we introduce the following vectors:

$$\mathbf{V} = \begin{bmatrix} \mathbf{U} \\ \mathbf{U}_a \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{bmatrix},$$

$$\mathbf{G}(\mathbf{V}) = \begin{bmatrix} \mathbf{F}(\mathbf{U}) \\ \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) \end{bmatrix} = \begin{bmatrix} w_2 \\ \frac{w_2^2}{w_1} + (\gamma - 1)\left(w_3 - \frac{1}{2}\frac{w_2^2}{w_1}\right) \\ \gamma\frac{w_2 w_3}{w_1} - \frac{(\gamma-1)}{2}\frac{w_2^3}{w_1^2}) \\ w_5 \\ \frac{\gamma-3}{2}\frac{w_2^2 w_4}{w_1^2} - (\gamma-3)\frac{w_2 w_5}{w_1} + (\gamma-1)w_6 \\ \gamma\frac{w_3 w_5}{w_1} - \gamma\frac{w_2 w_3 w_4}{w_1^2} - \frac{3}{2}(\gamma-1)\frac{w_2^2 w_5}{w_1} + (\gamma-1)\frac{w_2^3 w_4}{w_1^3} + \gamma\frac{w_2 w_6}{w_1} \end{bmatrix}.$$

Therefore, the complete system writes:

$$\begin{cases} \partial_t \mathbf{V} + \partial_x \mathbf{G}(\mathbf{V}) = 0, \\ \mathbf{V}(x, 0) = \mathbf{V}_0(x), \end{cases} \tag{4.7}$$

with $\mathbf{V}_0(x) = (\mathbf{U}_0(x), \partial_a \mathbf{U}_0(x))^t$. The Jacobian matrix of the complete system has the following form:

$$\frac{\partial \mathbf{G}(\mathbf{V})}{\partial \mathbf{V}} = \mathbf{M}(\mathbf{V}) = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}$$

where $\mathbf{A}$ is the Jacobian matrix of the state system and $\mathbf{B}$ writes:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ (\gamma-3)uu_a & (3-\gamma)u_a & 0 \\ (\star) & (\bullet) & \gamma u_a \end{bmatrix}, \tag{4.8}$$

with

$$(\star) = -\frac{c^2}{\gamma - 1}\frac{p_a}{p}u + \frac{3}{2}(\gamma-2)u^2 u_a + \frac{c^2}{\gamma-1}\frac{u\rho_a}{\rho} - \frac{c^2}{\gamma-1}u_a + \gamma\frac{u^3\rho_a}{\rho},$$

and

$$(\bullet) = \frac{\gamma}{2}u^2\rho_a - \frac{c^2}{\gamma-1}\rho_a + \frac{6-5\gamma}{2}\frac{u^2\rho_a}{\rho} + (3-2\gamma)uu_a + 3(\gamma-1)\frac{u\rho_a}{\rho^2} + \frac{c^2}{\gamma-1}\frac{p_a}{p}.$$

The matrix $\mathbf{M}$ has three repeated eigenvalues, which are the eigenvalues of the matrix $\mathbf{A}$. More precisely, one can prove the following result.

Proposition 1. *The global system (4.7) is weakly hyperbolic.*

Proof. A system of the form (4.7) is weakly hyperbolic if its Jacobian matrix has real eigenvalues and it is not $\mathbb{R}$-diagonalisable. We want to investigate whether or not the matrix $\mathbf{M}$ is $\mathbb{R}$-diagonalisable. A matrix is diagonalisable if and only if its minimal polynomial splits in distinct roots. Since the characteristic polynomial of the matrix $\mathbf{M}$ is the following:

$$p_M(x) = (x - \lambda_1)^2(x - \lambda_2)^2(x - \lambda_3)^2, \tag{4.9}$$

the minimal polynomial, in order to have distinct roots, can be at most of degree 3. Therefore, if $\mathbf{M}$ is diagonalisable, it must be:

$$(\mathbf{M} - \lambda_1 I_6)(\mathbf{M} - \lambda_2 I_6)(\mathbf{M} - \lambda_3 I_6) = 0. \tag{4.10}$$

Let us write (4.10) by blocks:

$$\begin{bmatrix} A-\lambda_1 I_3 & 0 \\ B & A-\lambda_1 I_3 \end{bmatrix} \begin{bmatrix} A-\lambda_2 I_3 & 0 \\ B & A-\lambda_2 I_3 \end{bmatrix} \begin{bmatrix} A-\lambda_3 I_3 & 0 \\ B & A-\lambda_3 I_3 \end{bmatrix} = 0 \tag{4.11}$$

Developing the left-hand side products one obtains the following matrix:

$$\begin{bmatrix} (A-\lambda_1 I_3)(A-\lambda_2 I_3)(A-\lambda_3 I_3) & 0 \\ (\blacksquare) & (A-\lambda_1 I_3)(A-\lambda_2 I_3)(A-\lambda_3 I_3) \end{bmatrix}, \tag{4.12}$$

where

$$(\blacksquare) = B(A-\lambda_2 I_3)(A-\lambda_3 I_3) + (A-\lambda_1 I_3)B(A-\lambda_3 I_3) + (A-\lambda_1 I_3)(A-\lambda_2 I_3)B.$$

The top-left and bottom-right coefficients are equal to the characteristic polynomial of $\mathbf{A}$ evaluated in $\mathbf{A}$, thus they are zero. Therefore, the matrix $\mathbf{M}$ is diagonalisable if and only if $(\blacksquare) = 0$. Let us compute the coefficient $(1,1)$ of $(\blacksquare)$:

$$\begin{aligned}(\blacksquare)_{(1,1)} &= 0 + [c-u,\ 1,\ 0] \begin{bmatrix} 0 \\ (3-\gamma)u^2 u_a - \frac{(\gamma-3)^2}{2} u^2 u_a \\ \diamondsuit \end{bmatrix} + \\ &+ [c-u,\ 1,\ 0] \begin{bmatrix} (\gamma-3)uu_a \\ (\gamma-2)(3-\gamma)u^2 u_a + (\gamma-1)(\star) \\ \triangle \end{bmatrix} = \\ &= -\frac{3}{2}(\gamma-1)u^2 u_a + (\gamma-3)cuu_a - c^2 u\frac{p_a}{p} + c^2 u\frac{\rho_a}{\rho} - c^2 u_a + \gamma(\gamma-1)u^3\frac{\rho_a}{\rho},\end{aligned}$$

where there is no need to specify $\diamondsuit$ and $\triangle$. There is no reason why the quantity above should be always be zero. Therefore, the matrix is not diagonalisable and the complete system is not hyperbolic in general. However, as the eigenvalues are real, the system is weakly hyperbolic. □

4.2 Source term

The sensitivity system (4.5) was derived assuming that the state solution $\mathbf{U}$ is regular. However, this is not generally true for hyperbolic systems such as the one considered [BP02]. Therefore, we add a correction term in the Rankine-Hugoniot conditions of (4.5), as done in the previous chapters in order to make this system valid also in the present framework of hyperbolic systems with possibly discontinuous solutions:

$$\mathbf{S} = \sum_{k=1}^{N_s} \delta_k \boldsymbol{\rho}_k, \tag{4.13}$$

where N_s is the number of discontinuities, which can be either shocks or contact discontinuities, $\delta_k = \delta(x - x_{k,s})$ is the Dirac delta function with $x_{k,s}$ position of the $k-$th shock and $\boldsymbol{\rho}_k$ is the amplitude of the $k-$th correction which is

$$\boldsymbol{\rho}_k(t) = \sigma_{k,a}(\mathbf{U}^+ - \mathbf{U}^-). \tag{4.14}$$

The definition of the source term comes from the derivation with respect to a of the Rankine-Hugoniot relations associated with (4.3):

$$-\sigma(\mathbf{U}^+ - \mathbf{U}^-) + \mathbf{F}(\mathbf{U}^+) - \mathbf{F}(\mathbf{U}^-) = 0,$$

which gives:

$$\begin{aligned}(\mathbf{U}_a^- - \mathbf{U}_a^+)\sigma_k + (\mathbf{U}^- - \mathbf{U}^+)\sigma_{k,a} + \sigma_k(\partial_x\mathbf{U}^- - \partial_x\mathbf{U}^+)\partial_a x_{k,s}(t) = \\ = \mathbf{F}_a^- - \mathbf{F}_a^+ + \left(\frac{\partial\mathbf{F}(\mathbf{U}^-)}{\partial\mathbf{U}}\partial_x\mathbf{U}^- - \frac{\partial\mathbf{F}(\mathbf{U}^+)}{\partial\mathbf{U}}\partial_x\mathbf{U}^+\right)\partial_a x_{k,s}(t),\end{aligned} \tag{4.15}$$

where the terms with $\partial_a x_s(t)$ are very difficult to estimate, however they are all zero in a first order finite volume framework, therefore they can be neglected. A special treatment, which will be detailed later, is necessary for a second order discretisation. The new system can thus be written as:

$$\begin{cases}\partial_t\mathbf{U} + \partial_x\mathbf{F}(\mathbf{U}) = 0 \\ \partial_t\mathbf{U}_a + \partial_x\mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = \mathbf{S}.\end{cases} \tag{4.16}$$

Before going to the design of a numerical scheme to approximate the solution of (4.16), we briefly discuss in the next section the typical structure of these solutions and compare it with the ones of (4.3). We specify the solution for a given initial data of Riemann type.

4.3 Riemann problem

In this section, we write the exact solution for the system (4.16) in a specific case (cf. [App97]), which will be used as a test case to check the convergence of the numerical schemes proposed. We consider a Riemann problem, i.e.:

$$\mathbf{V}_0(x) = \begin{cases}\mathbf{V}_L & x < x_c, \\ \mathbf{V}_R & x > x_c.\end{cases}$$

The general solution for this kind of problem is quite complicated, especially for the sensitivity (the last three components of $\mathbf{V}$). First, we study the state (the first three components of $\mathbf{V}$): the pair $(\lambda_2, \mathbf{r}_2)$ is linearly degenerate, i.e. $\nabla\lambda_2 \cdot \mathbf{r}_2 = 0$, therefore the middle wave is always a contact discontinuity; concerning the $1-$wave and the $3-$wave, they are genuinely nonlinear therefore they can either be shocks or rarefaction waves. In Figure 4.1 we show the structure of the state in the case of a rarefaction-contact-shock. Concerning the sensitivity, it has the same structure as the state (cf. Figure 4.2 in the case rarefaction-contact-shock): the middle wave is always a contact wave, and the $1-$ and $2-$wave are of the same type as for the state. The only difference is that the sensitivity presents discontinuities in the two extrema of the rarefaction fan (and this is why in Figure 4.2 the external lines of the rarefaction fan are thicker).

In the following, we illustrate this analysis of the wave structure by giving the detailed solution for the state and for the sensitivity in a specific case. The initial data for the state on the physical variables is the following:

$$\rho_L = 1,\ u_L = 0,\ p_L = 1, \qquad \rho_R = 0.125,\ u_R = 0,\ p_R = 0.1.$$

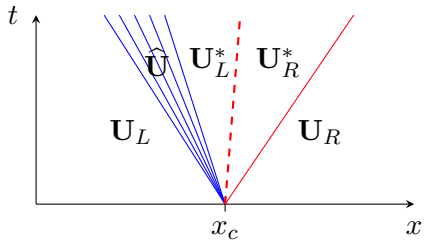


Figure 4.1 – Structure of the solution for the Riemann problem for the state.

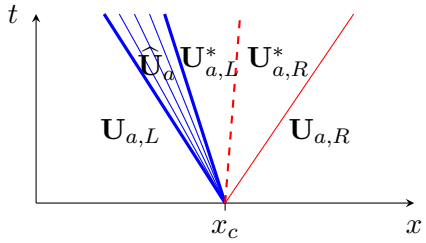


Figure 4.2 – Structure of the solution for the Riemann problem for the sensitivity.

We consider as parameter of interest $a = p_L$, therefore the initial data for the sensitivity is:

$$\rho_{a,L} = \rho_{a,R} = u_{a,L} = u_{a,R} = p_{a,R} = 0, \quad p_{a,L} = 1.$$

This choice of initial data leads to the structure in Figures 4.1-4.2, for the state as well as for the sensitivity: the $1-$wave is a rarefaction and the $3-$wave is a shock. For the notation, please refer to Figure 4.1 for the state and Figure 4.2 for the sensitivity. Let us now give the exact formulas for the state and for the sensitivity.

State solution: the exact solution for the physical variables is given in [App97]. Every variable is given as a function of the pressure in the right-star zone p_R^*, which is computed numerically from the following implicit relation:

$$p_L = p_R^* \left(1 - \frac{(\gamma-1)\frac{c_R}{c_L}(\frac{p_R^*}{p_R} - 1)}{\sqrt{2\gamma\left(2\gamma + (\gamma+1)(\frac{p_R^*}{p_R} - 1)\right)}} \right)^{-\frac{2\gamma}{\gamma-1}}, \tag{4.17}$$

where $c_\ell = \sqrt{\dfrac{\gamma p_\ell}{\rho_\ell}}$, with $\ell = L, R$. In the star regions, we have:

$$p_L^* = p_R^* = p^*,$$

$$u_L^* = u_R^* = u^* = c_R \left(\frac{p^*}{p_R} - 1 \right) \sqrt{\frac{2}{\gamma(\gamma+1)\frac{p^*}{p_R} + \gamma(\gamma-1)}},$$

because the velocity u and the pressure p are Riemann invariants across the $2-$wave; as for the density ρ, we have:

$$\rho_R^* = \rho_R \frac{p^*}{p_R} \left(\frac{1 + \frac{\gamma-1}{\gamma+1}\frac{p_R}{p^*}}{1 + \frac{\gamma-1}{\gamma+1}\frac{p^*}{p_R}} \right),$$

$$\rho_L^* = \rho_L \left(\frac{p^*}{p_L}\right)^{\frac{1}{\gamma}}.$$

In the rarefaction wave, we have:

$$\hat{u}(x,t) = \frac{2(u^* - u_L)}{(\gamma+1)u^*}\left(\frac{x-x_c}{t}\right) + 2\frac{c_L u^* - u_L\left(c_L - \frac{\gamma+1}{2}u^*\right)}{(\gamma+1)u^*},$$

$$\hat{\rho}(x,t) = \rho_L\left(1 - (\gamma-1)\frac{\hat{u}(x,t)}{2c_L}\right)^{\frac{2}{\gamma-1}},$$

$$\hat{p}(x,t) = p_L\left(1 - (\gamma-1)\frac{\hat{u}(x,t)}{2c_L}\right)^{\frac{2\gamma}{\gamma-1}}.$$

Finally, the solution is given as:

$$\mathbf{U}(x,t) = \begin{cases} \mathbf{U}_L & x - x_c < -c_L t, \\ \widehat{\mathbf{U}} & -c_L t < x - x_c < \left(\frac{\gamma+1}{2}u^* - c_L\right)t, \\ \mathbf{U}_L^* & \left(\frac{\gamma+1}{2}u^* - c_L\right)t < x - x_c < u^* t, \\ \mathbf{U}_R^* & u^* t < x - x_c < c_R\sqrt{\frac{\gamma-1}{2\gamma} + \frac{\gamma+1}{2\gamma}\frac{p^*}{p_R}}\,t, \\ \mathbf{U}_R & x - x_c > c_R\sqrt{\frac{\gamma-1}{2\gamma} + \frac{\gamma+1}{2\gamma}\frac{p^*}{p_R}}\,t. \end{cases} \tag{4.18}$$

Sensitivity solution: by differentiating (4.17) with respect to a, one obtains the following explicit formula for $p_{a,R}^*$:

$$p_{a,R}^* = p_{a,L}^* = p_a^* = \frac{1 + \Theta^{\frac{1-3\gamma}{\gamma-1}}\Xi p^*}{\Theta^{-\frac{2\gamma}{\gamma-1}} + \Theta^{\frac{1-3\gamma}{\gamma-1}}(\Lambda - \Psi)p^*},$$

where:

$$\Theta = 1 - \frac{(\gamma-1)c_R\left(\frac{p^*}{p_R} - 1\right)}{c_L\sqrt{4\gamma^2 + 2\gamma(\gamma-1)\left(\frac{p^*}{p_R} - 1\right)}},$$

$$\Xi = \frac{c_R\left(\frac{p^*}{p_R} - 1\right)c_{a,R}\sqrt{2\gamma}}{c_L^2\sqrt{2\gamma + (\gamma+1)\left(\frac{p^*}{p_R} - 1\right)}},$$

$$\Lambda = \frac{\sqrt{2\gamma}\,c_R}{c_L p_R\sqrt{2\gamma + (\gamma+1)\left(\frac{p^*}{p_R} - 1\right)}},$$

$$\Psi = \frac{\gamma(\gamma+1)c_R\left(\frac{p^*}{p_R} - 1\right)}{c_L p_R\sqrt{2\gamma}\left(2\gamma + (\gamma+1)\left(\frac{p^*}{p_R} - 1\right)\right)^{\frac{3}{2}}}.$$

In the star regions, by differentiating the corresponding state, one finds:

$$u_a^* = \frac{2c_{a,L}}{\gamma-1}\left(1-\left(\frac{p^*}{p_L}\right)^{\frac{\gamma-1}{2\gamma}}\right) - \frac{c_L}{\gamma}\left(\frac{p^*}{p_L}\right)^{\frac{-\gamma-1}{2\gamma}}\left(\frac{p_L p_a^* - p^*}{p_L^2}\right),$$

$$\rho_{a,R}^* = \frac{\rho_R p_a^*}{p_R}\frac{\left(1+\frac{\gamma-1}{\gamma+1}\frac{p_R}{p^*}\right)}{\left(1+\frac{\gamma-1}{\gamma+1}\frac{p^*}{p_R}\right)} + \rho_R\frac{p^*}{p_R}\frac{\gamma-1}{\gamma+1}\left(\frac{-\frac{p_R p_a^*}{p^{*2}}\left(1+\frac{\gamma-1}{\gamma+1}\frac{p^*}{p_R}\right) - \frac{p_a^*}{p_R}\left(1+\frac{\gamma-1}{\gamma+1}\frac{p_R}{p^*}\right)}{(1+\frac{\gamma-1}{\gamma+1}\frac{p^*}{p_R})^2}\right),$$

$$\rho_{a,L}^* = \frac{\rho_L}{\gamma}\frac{p_L p_a^* - p^*}{p_L^2}\left(\frac{p^*}{p_L}\right)^{\frac{1-\gamma}{\gamma}}.$$

Finally, in the rarefaction:

$$\hat{u}_a(x,t) = \frac{2u_L u^*}{(\gamma+1)u^{*2}}\frac{x-x_c}{t} + 2\frac{c_{a,L}u^{*2} - c_{a,L}u_L u^* + c_L u_L u_a^*}{(\gamma+1)u^{*2}},$$

$$\hat{\rho}_a(x,t) = -\rho_L\left(\frac{\hat{u}_a(x,t)c_L - \hat{u}(x,t)c_{a,L}}{c_L^2}\right)\left(1-\frac{(\gamma-1)\hat{u}(x,t)}{2c_L}\right)^{\frac{3-\gamma}{\gamma-1}},$$

$$\hat{p}_a(x,t) = \left(1-\frac{(\gamma-1)\hat{u}(x,t)}{2c_L}\right)^{\frac{2\gamma}{\gamma-1}} - p_L\gamma\left(\frac{\hat{u}_a(x,t)c_L - \hat{u}(x,t)c_{a,L}}{c_L^2}\right)\left(1-\frac{(\gamma-1)\hat{u}(x,t)}{2c_L}\right)^{\frac{\gamma+1}{\gamma-1}}.$$

The sensitivity has the same structure as the state, therefore:

$$\mathbf{U}_a(x,t) = \begin{cases} \mathbf{U}_{a,L} & x - x_c < -c_L t, \\ \widehat{\mathbf{U}}_a\left(\frac{x-x_c}{t}\right) & -c_L t < x - x_c < \left(\frac{\gamma+1}{2}u^* - c_L\right)t, \\ \mathbf{U}_{a,L}^* & \left(\frac{\gamma+1}{2}u^* - c_L\right)t < x - x_c < u^* t, \\ \mathbf{U}_{a,R}^* & u^* t < x - x_c < c_R\sqrt{\frac{\gamma-1}{2\gamma} + \frac{\gamma+1}{2\gamma}\frac{p^*}{p_R}}\,t, \\ \mathbf{U}_{a,R} & x - x_c > c_R\sqrt{\frac{\gamma-1}{2\gamma} + \frac{\gamma+1}{2\gamma}\frac{p^*}{p_R}}\,t. \end{cases} \tag{4.19}$$

We remark that if one writes the Rankine-Hugoniot conditions across the shock one finds:

$$-c_R\sqrt{\frac{\gamma-1}{2\gamma} + \frac{\gamma+1}{2\gamma}\frac{p^*}{p_R}}\,(\mathbf{U}_{a,R} - \mathbf{U}_{a,R}^*) + \mathbf{F}_a(\mathbf{U}_R, \mathbf{U}_{a,R}) - \mathbf{F}_a(\mathbf{U}_R^*, \mathbf{U}_{a,R}^*) = \mathbf{S}.$$

4.4 Numerical methods

In this section we consider the numerical approximation of (4.16). We derive first and second order Roe-type numerical schemes and we pay particular attention to the numerical diffusion effects induced by these approaches. Indeed and as we will see it may prevent the numerical solution from converging to the correct solution. We consider a uniform grid in space with a constant step Δx, x_j is the center of the $j-$th cell C_j, whose extrema are $x_{j-1/2}$ and $x_{j+1/2}$ (cf. Figure 4.3). We use an adaptive time step Δt^n, chosen according to a CFL condition, and the intermediate times are $t^{n+1} = t^n + \Delta t^n$. We indicate with $\mathbf{V}_j^n = (\mathbf{U}_j^n, \mathbf{U}_{a,j}^n)^t$ the average value of the state and the sensitivity in the cell C_j at time t^n.

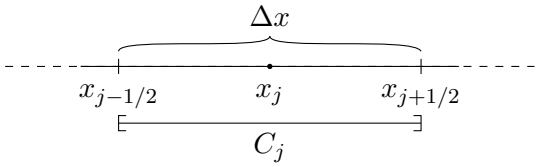


Figure 4.3 – Spatial discretisation.

We use Godunov-type schemes, which consist of two main steps: first, one solves the Riemann problem at each interface $x_{j-1/2}$ at time t^n, obtaining in this way a solution at time t^{n+1}, $\mathbf{v}(x, t^{n+1}) = (\mathbf{u}(x, t^{n+1}), \mathbf{u}_a(x, t^{n+1}))^t$; the second step is to project $\mathbf{v}(x, t^{n+1})$ in order to obtain a piecewise constant solution on the mesh. How to compute $\mathbf{v}(x, t)$ is the topic of the next subsections: different choices for the solution of the Riemann problem lead to different numerical schemes.

4.4.1 Projection step

The projection step is usually performed by averaging the solution $\mathbf{v}(x, t^{n+1})$ on the cell:

$$\mathbf{V}_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{v}(x, t^{n+1}) dx. \tag{4.20}$$

However, this projection method introduces numerical diffusion. As shown in Chapter 3, numerical diffusion plays a fundamental role in the discretisation of the sensitivity, especially across shocks. For this reason, we propose another projection method, introduced in [CG08] and inspired by Glimm's method [Gli65, Cho76]. First, we define a staggered mesh, whose cells will be denoted $\overline{C}_j^n$, as follows:

$$\overline{C}_j^n = (\bar{x}_{j-1/2}^n, \bar{x}_{j+1/2}^n), \quad \bar{x}_{j-1/2}^n = x_{j-1/2} + \sigma_{j-1/2}^n \Delta t^n,$$

where $\sigma_{j-1/2}^n$ is a proper speed. In this case, we choose it equal to zero if no shock is expected at the interface $j - 1/2$, whilst it is equal to the speed of the shock, if there is one. More details on the choice of $\sigma_{j-1/2}^n$ will be provided in the next subsections. The second step is to perform the average on the staggered mesh, obtaining in this way an intermediate solution $\overline{\mathbf{V}}_j^{n+1}$:

$$\overline{\mathbf{V}}_j^{n+1} = \frac{1}{\Delta x_j^n} \int_{\bar{x}_{j-1/2}}^{\bar{x}_{j+1/2}} \mathbf{v}(x, t^{n+1}) dx, \tag{4.21}$$

where $\Delta x_j^n = \bar{x}_{j+1/2} - \bar{x}_{j-1/2}$. Finally, the last step is a sampling step, in order to go back to the initial uniform grid. Let (α_n) be a random sequence varying in $(0, 1)$, for instance $\alpha_n \sim \mathcal{U}([0, 1])$; then:

$$\mathbf{V}_j^{n+1} = \begin{cases} \overline{\mathbf{V}}_{j-1}^{n+1} & \text{if } \alpha_{n+1} \in \left(0, \frac{\Delta t}{\Delta x} \max(\sigma_{j-1/2}^n, 0)\right), \\ \overline{\mathbf{V}}_j^{n+1} & \text{if } \alpha_{n+1} \in \left[\frac{\Delta t}{\Delta x} \max(\sigma_{j-1/2}^n, 0), 1 + \frac{\Delta t}{\Delta x} \min(\sigma_{j+1/2}^n, 0)\right), \\ \overline{\mathbf{V}}_{j+1}^{n+1} & \text{if } \alpha_{n+1} \in \left[1 + \frac{\Delta t}{\Delta x} \min(\sigma_{j+1/2}^n, 0), 1\right). \end{cases} \tag{4.22}$$

The method is proven to be convergent even if a low discrepancy deterministic sequence is used. In this work, we use the van der Corput sequence for both the state and the sensitivity (cf. [CG08]):

$$\alpha_n = \sum_{k=0}^{m} i_k 2^{-(k+1)}, \quad n = \sum_{k=0}^{m} i_k 2^k,$$

where $i_k = 0, 1$ is the binary expansion of the integers.

In the next subsections we present some Riemann solvers for the state and for the sensitivity, and their two variations: the diffusive standard version with the projection step (4.20), and their anti-diffusive version (4.21)-(4.22) and we compare their results. The strategy is to solve the state system separately from the sensitivity system, since the global system (4.7) is only weakly hyperbolic and some of the standard approaches require strict hyperbolicity. Moreover, notice that the state system (4.3) evolves in an independent way with respect to the sensitivity system (4.6).

4.4.2 Riemann solver for the state

First, we consider the state system, for which the classical numerical schemes can be used: in this work we used the approximate Riemann solver of Roe, because it has the property of being exact for an isolated shock. It turns out that this property plays an important role in the anti-diffusive approach (4.21)-(4.22). Indeed, it makes the method equivalent to the random choice Glimm scheme in the specific case of an isolated shock wave, and therefore is convergent to the correct solution (still in this very specific case). In addition, we remark that it would not be possible to use a solver with only one intermediate star state, such as HLL, because of the definition of the source term (4.14): two intermediate states are necessary in order to be able to compute the correction term across the contact discontinuity.

The main idea of the Roe scheme is to replace the Jacobian matrix $\mathbf{A}(\mathbf{U})$ in (4.4) with a constant matrix $\mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)$, obtaining in this way a linearised system, whose solution to the Riemann problem can be computed exactly. For the Euler system, a proper linearisation is provided by Roe in the original paper [Roe81]. Furthermore, there is no need to assemble the matrix, it is sufficient to know its eigenvalues and eigenvectors, which are the following:

$$\lambda_1^{ROE} = \tilde{u} - \tilde{c}, \quad \lambda_2^{ROE} = \tilde{u}, \quad \lambda_3^{ROE} = \tilde{u} + \tilde{c},$$

$$\tilde{\mathbf{r}}_1 = \begin{pmatrix} 1 \\ \tilde{u} - \tilde{c} \\ \tilde{H} - \tilde{u}\tilde{c} \end{pmatrix}, \quad \tilde{\mathbf{r}}_2 = \begin{pmatrix} 1 \\ \tilde{u} \\ \frac{\tilde{u}^2}{2} \end{pmatrix}, \quad \tilde{\mathbf{r}}_3 = \begin{pmatrix} 1 \\ \tilde{u} + \tilde{c} \\ \tilde{H} + \tilde{u}\tilde{c} \end{pmatrix}.$$

The quantities denoted with a tilde are *Roe averaged* quantities defined as follows:

$$\tilde{u} = \frac{\sqrt{\rho_L}u_L + \sqrt{\rho_R}u_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \quad \tilde{H} = \frac{\sqrt{\rho_L}H_L + \sqrt{\rho_R}H_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \quad \tilde{c} = \sqrt{(\gamma - 1)\left(\tilde{H} - \frac{1}{2}\tilde{u}^2\right)}.$$

Therefore, the Roe solver consists of four constant states ($\mathbf{U}_L$, $\mathbf{U}_L^*$, $\mathbf{U}_R^*$, and $\mathbf{U}_R$, cf. Figure 4.4) connected by three discontinuities travelling at speeds λ_i^{ROE}. To compute

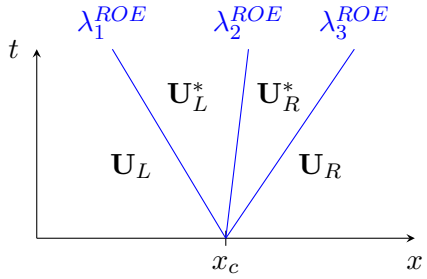


Figure 4.4 – Structure of the Roe solver for the state.

the star states $\mathbf{U}_L^*$ and $\mathbf{U}_R^*$, first we decompose the jump $\mathbf{U}_R - \mathbf{U}_L$ along the eigenvectors of the Jacobian matrix $\mathbf{A}$:

$$\Delta\mathbf{U} = \mathbf{U}_R - \mathbf{U}_L = \sum_{i=1}^{3} \alpha_i \tilde{\mathbf{r}}_i. \tag{4.23}$$

The relation (4.23) is used to compute the coefficients α_i, then one has:

$$\mathbf{U}_L^* = \mathbf{U}_L + \alpha_1\tilde{\mathbf{r}}_1 = \mathbf{U}_R - \alpha_2\tilde{\mathbf{r}}_2 - \alpha_3\tilde{\mathbf{r}}_3, \quad \mathbf{U}_R^* = \mathbf{U}_R - \alpha_3\tilde{\mathbf{r}}_3 = \mathbf{U}_L + \alpha_1\tilde{\mathbf{r}}_1 + \alpha_2\tilde{\mathbf{r}}_2. \tag{4.24}$$

Once all the quantities $\mathbf{U}_L^*$, $\mathbf{U}_R^*$, and λ_ℓ^{ROE} are known at each interface $x_{j-1/2}$, $\mathbf{w}(x, t^{n+1})$ is defined and the integrals (4.20) or (4.21) can easily be computed, since $\mathbf{w}(x, t^{n+1})$ is a piecewise constant.

It is well known that, in case of transonic rarefaction, the Roe solver provides a non-entropic solution. To overcome this problem, we implemented the entropic fix proposed in [HH83].

Definition of $\sigma_{j-1/2}^n$

As already said, $\sigma_{j-1/2}^n$ is defined in order to avoid averaging across a shock. Numerical results show that there is no need to move the mesh for the contact discontinuity (cf. section 4.5). The definition of $\sigma_{j-1/2}^n$ is the following:

$$\sigma_{j-1/2}^n = \begin{cases} \lambda_{1,j-1/2}^{ROE} & \text{if } d_{1,j-1/2} = 1, \\ \lambda_{3,j-1/2}^{ROE} & \text{if } d_{3,j-1/2} = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $d_{\ell,j-1/2}$ are shock detectors, $d_{\ell,j-1/2} = 1$ if there is an $\ell-$shock at the interface $j - 1/2$, it is zero otherwise. They are based on the fact that the velocity u is always decreasing across a shock, whilst the density ρ is increasing across a $1-$shock and it is decreasing across a $3-$shock:

$$d_{1,j-1/2} = \begin{cases} 1 & \text{if } \rho_j > \rho_{j-1} \text{ and } u_j < u_{j-1}, \\ 0 & \text{otherwise,} \end{cases} \qquad d_{3,j-1/2} = \begin{cases} 1 & \text{if } \rho_j < \rho_{j-1} \text{ and } u_j < u_{j-1}, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, we remark that there is no need for a contact detector because it is known that the middle wave is always a contact discontinuity.

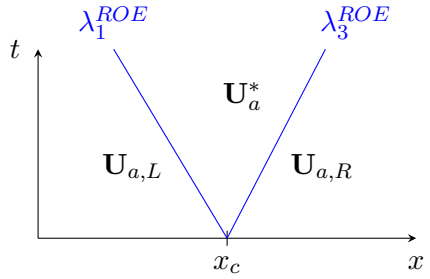


Figure 4.5 – Structure of the HLL-type solver for the sensitivity.

4.4.3 Riemann solvers for the sensitivity

For the sensitivity we propose two different strategies. Indeed and as explained in the previous section, for the state it is necessary to use a Riemann solver with two different star states, in order to be able to compute the source term across the contact discontinuity. However, for the sensitivity an HLL-type approach can be used, which gives a first strategy. Another possible strategy is to keep for the sensitivity the same structure as for the state, and therefore to have an HLLC-type scheme. A third possibility which we will not analyse here, explored in detail in [App97], is to rewrite the sensitivity flux in such a way that the same Roe Riemann solver used for the state can be applied for the sensitivity. Let us now describe the two possibilities considered in detail.

HLL-type scheme

The first Riemann solver proposed for the sensitivity has a simpler structure than the state solver: we neglect the contact discontinuity, therefore the solver consists only of three constant states ($\mathbf{U}_{a,L}$, $\mathbf{U}_a^*$, and $\mathbf{U}_{a,R}$) connected by two discontinuities travelling at speeds λ_1^{ROE} and λ_3^{ROE} (cf. Figure 4.5). The star value of the sensitivity $\mathbf{U}_a^*$ at the interface $j-1/2$ can be computed directly from the Harten, Lax and van Leer conditions [HLL83] applied to a system of conservation laws with source terms. We get:

$$\begin{aligned}\mathbf{U}_{a,j-1/2}^* = \frac{1}{\lambda_3^{ROE}-\lambda_1^{ROE}}\Big(&\lambda_3^{ROE}\mathbf{U}_{a,j}^n - \lambda_1^{ROE}\mathbf{U}_{a,j-1}^n \\ &-\mathbf{F}_a(\mathbf{U}_j,\mathbf{U}_{a,j}) + \mathbf{F}_a(\mathbf{U}_{j-1},\mathbf{U}_{a,j-1}) + \mathbf{S}_{j-1/2}\Big),\end{aligned} \tag{4.25}$$

where the source term is naturally discretised as follows:

$$\begin{aligned}\mathbf{S}_{j-1/2} = \partial_a\lambda_{1,j-1/2}^{ROE}(\mathbf{U}_{L,j-1/2}^* - \mathbf{U}_{j-1})d_{1,j-1/2} + \partial_a\lambda_{2,j-1/2}^{ROE}(\mathbf{U}_{R,j-1/2}^* - \mathbf{U}_{L,j-1/2}^*) \\ +\partial_a\lambda_{3,j-1/2}^{ROE}(\mathbf{U}_j - \mathbf{U}_{R,j-1/2}^*)d_{3,j-1/2}.\end{aligned}$$

HLLC-type scheme

Another possible approach for the sensitivity is to keep the same structure as for the state (cf. Figure 4.4), with the same speeds of propagation for the three discontinuities. We need to compute the two intermediate constant states $\mathbf{U}_{a,L}^*$ and $\mathbf{U}_{a,R}^*$. Again, a

possible strategy to compute $\mathbf{U}^*_{a,L}$ and $\mathbf{U}^*_{a,R}$ is to follow the Harten, Lax and van Leer formalism with source term and to impose the following linear system, created from the Rankine-Hugoniot jump relations:

$$\begin{cases} -\lambda_1(\rho^*_{a,L} - \rho_{a,L}) + (\rho u)^*_{a,L} - (\rho u)_{a,L} = \partial_a\lambda_1(\rho^*_L - \rho_L), \\ -\lambda_2(\rho^*_{a,R} - \rho^*_{a,L}) + (\rho u)^*_{a,R} - (\rho u)^*_{a,L} = \partial_a\lambda_2(\rho^*_R - \rho^*_L), \\ -\lambda_3(\rho_{a,R} - \rho^*_{a,R}) + (\rho u)_{a,R} - (\rho u)^*_{a,R} = \partial_a\lambda_3(\rho_R - \rho^*_R), \\ \frac{(\gamma-3)}{2}\tilde{u}^2(\rho^*_{a,R} - \rho^*_{a,L}) + (2-\gamma)\tilde{u}((\rho u)^*_{a,R} - (\rho u)^*_{a,L}) \\ \qquad +(\gamma-1)((\rho E)^*_{a,R} - (\rho E)^*_{a,L}) = \partial_a\lambda_2((\rho u)^*_R - (\rho u)^*_L), \\ (\lambda_2 - \lambda_1)(\rho u)^*_{a,L} + (\lambda_3 - \lambda_2)(\rho u)^*_{a,R} + \lambda_1(\rho u)_{a,L} - \lambda_3(\rho u)_{a,R} \\ \qquad +\mathbf{F}_{a,R}|_2 - \mathbf{F}_{a,L}|_2 = \Delta x\mathbf{S}|_2, \\ (\lambda_2 - \lambda_1)(\rho E)^*_{a,L} + (\lambda_3 - \lambda_2)(\rho E)^*_{a,R} + \lambda_1(\rho E)_{a,L} - \lambda_3(\rho E)_{a,R} \\ \qquad +\mathbf{F}_{a,R}|_3 - \mathbf{F}_{a,L}|_3 = \Delta x\mathbf{S}|_3, \end{cases} \tag{4.26}$$

where $\lambda_1 = \tilde{u} - \tilde{c}$, $\lambda_2 = \tilde{u}$, and $\lambda_3 = \tilde{u} + \tilde{c}$. The first three equations are the Rankine-Hugoniot condition on ρ across the three waves, differentiated with respect to a. Note that summing up these equations gives the integral condition of the Harten, Lax and van Leer formalism of the density variable. The fourth equation is the Rankine-Hugoniot condition on ρu for the linearised system differentiated with respect to a; the last two equations are the integral conditions on the sensitivities $(\rho u)_a$ and $(\rho E)_a$. If we define the following vectors

$$\mathbf{x} = (\rho^*_{a,L}, \rho^*_{a,R}, (\rho u)^*_{a,L}, (\rho u)^*_{a,R}, (\rho E)^*_{a,L}, (\rho E)^*_{a,R})^t$$

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} = \begin{pmatrix} \partial_a\lambda_1(\rho^*_L - \rho_L) + (\rho u)_{a,L} - \lambda_1\rho_{a,L} \\ \partial_a\lambda_2(\rho^*_R - \rho^*_L) \\ \partial_a\lambda_3(\rho_R - \rho^*_R) - (\rho u)_{a,R} + \lambda_3\rho_{a,R} \\ \partial_a\lambda_2((\rho u)^*_R - (\rho u)^*_L) \\ \Delta x\mathbf{S}|_2 - \lambda_1(\rho u)_{a,L} + \lambda_3(\rho u)_{a,R} - \mathbf{F}_{a,R}|_2 + \mathbf{F}_{a,L}|_2 \\ \Delta x\mathbf{S}|_3 - \lambda_1(\rho E)_{a,L} + \lambda_3(\rho E)_{a,R} - \mathbf{F}_{a,R}|_3 + \mathbf{F}_{a,L}|_3 \end{pmatrix}$$

the system can be rewritten as:

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

where $\mathcal{A}$ is the following matrix:

$$\mathcal{A} = \begin{pmatrix} -\lambda_1 & 0 & 1 & 0 & 0 & 0 \\ \lambda_2 & -\lambda_2 & -1 & 1 & 0 & 0 \\ 0 & \lambda_3 & 0 & -1 & 0 & 0 \\ -\frac{(\gamma-3)}{2}\tilde{u}^2 & \frac{(\gamma-3)}{2}\tilde{u}^2 & -(2-\gamma)\tilde{u} & (2-\gamma)\tilde{u} & -(\gamma-1) & (\gamma-1) \\ 0 & 0 & \tilde{c} & \tilde{c} & 0 & 0 \\ 0 & 0 & 0 & 0 & \tilde{c} & \tilde{c} \end{pmatrix}$$

and we have $\det(\mathcal{A}) = 4\tilde{c}^4(\gamma-1) \neq 0$. The solution of the system has the following form:

$$\mathbf{x} = \begin{pmatrix} \frac{(2\tilde{c}+\tilde{u})b_1+(\tilde{c}+\tilde{u})b_2+\tilde{u}b_3-b_5}{2\tilde{c}^2} \\ \frac{-\tilde{u}b_1+(\tilde{c}-\tilde{u})b_2+(2\tilde{c}-\tilde{u})b_3+b_5}{2\tilde{c}^2} \\ \frac{(\tilde{u}^2+\tilde{c}\tilde{u})b_1+(\tilde{u}^2-\tilde{c}^2)b_2+(\tilde{u}^2-\tilde{c})\tilde{u}b_3+(\tilde{c}-\tilde{u})b_5}{2\tilde{c}^2} \\ \frac{-(\tilde{u}^2+\tilde{c}\tilde{u})b_1+(\tilde{c}^2-\tilde{u}^2)b_2+(\tilde{c}\tilde{u}-\tilde{u}^2)b_3+(\tilde{c}+\tilde{u})b_5}{2\tilde{c}^2} \\ \frac{(\gamma-1)(\tilde{u}^3+\tilde{c}\tilde{u}^2)b_1+((\gamma-1)\tilde{u}^3+2(2-\gamma)\tilde{c}^2\tilde{u})b_2+(\gamma-1)(\tilde{u}^3-\tilde{c}\tilde{u}^2)b_3-2\tilde{c}^2b_4-(\gamma-1)\tilde{u}^2b_5+2(\gamma-1)\tilde{c}b_6}{4(\gamma-1)\tilde{c}^2} \\ \frac{-(\gamma-1)(\tilde{u}^3+\tilde{c}\tilde{u}^2)b_1-((\gamma-1)\tilde{u}^3+2(2-\gamma)\tilde{c}^2\tilde{u})b_2+(\gamma-1)(\tilde{c}\tilde{u}^2-\tilde{u}^3)b_3+2\tilde{c}^2b_4+(\gamma-1)\tilde{u}^2b_5+2(\gamma-1)\tilde{c}b_6}{4(\gamma-1)\tilde{c}^2} \end{pmatrix}.$$

An alternative strategy to compute $\mathbf{U}^*_{a,L}$ and $\mathbf{U}^*_{a,R}$ is to differentiate with respect to a the following relations:

$$\mathbf{U}^*_L = \mathbf{U}_L + \alpha_1\mathbf{r}_1, \qquad \mathbf{U}^*_R = \mathbf{U}_R - \alpha_3\mathbf{r}_3, \tag{4.27}$$

obtaining

$$\mathbf{U}^*_{a,L} = \mathbf{U}_{a,L} + \alpha_{a,1}\mathbf{r}_1 + \alpha_1\mathbf{r}_{a,1}, \qquad \mathbf{U}^*_{a,R} = \mathbf{U}_{a,R} - \alpha_{a,3}\mathbf{r}_3 - \alpha_3\mathbf{r}_{a,3}, \tag{4.28}$$

with

$$\mathbf{r}_1 = \begin{pmatrix} 1 \\ \tilde{u}-\tilde{c} \\ \tilde{H}-\tilde{u}\tilde{c} \end{pmatrix}, \quad \mathbf{r}_{a,1} = \begin{pmatrix} 0 \\ \tilde{u}_a-\tilde{c}_a \\ \tilde{H}_a-\tilde{u}_a\tilde{c}-\tilde{u}\tilde{c}_a \end{pmatrix},$$

$$\mathbf{r}_2 = \begin{pmatrix} 1 \\ \tilde{u} \\ \frac{\tilde{u}^2}{2} \end{pmatrix}, \quad \mathbf{r}_{a,2} = \begin{pmatrix} 0 \\ \tilde{u}_a \\ \tilde{u}\tilde{u}_a \end{pmatrix},$$

$$\mathbf{r}_3 = \begin{pmatrix} 1 \\ \tilde{u}+\tilde{c} \\ \tilde{H}+\tilde{u}\tilde{c} \end{pmatrix}, \quad \mathbf{r}_{a,3} = \begin{pmatrix} 0 \\ \tilde{u}_a+\tilde{c}_a \\ \tilde{H}_a+\tilde{u}_a\tilde{c}+\tilde{u}\tilde{c}_a \end{pmatrix},$$

$$\begin{cases} \alpha_2 = \dfrac{\gamma-1}{\tilde{c}^2}\Big[(\rho_R-\rho_L)(\tilde{H}-\tilde{u}^2) + \tilde{u}\Big((\rho u)_R-(\rho u)_L\Big) - \Big((\rho E)_R-(\rho E)_L\Big)\Big], \\ \alpha_1 = \dfrac{1}{\tilde{c}}\Big[(\rho_R-\rho_L)(\tilde{u}+\tilde{c}) - \Big((\rho u)_R-(\rho u)_L\Big) - \tilde{c}\alpha_2\Big], \\ \alpha_3 = (\rho_R-\rho_L) - (\alpha_1+\alpha_2), \end{cases}$$

$$\begin{cases} \alpha_{a,2} &= -\dfrac{2\tilde{c}_a(\gamma-1)}{\tilde{c}^3}\Big[(\rho_R-\rho_L)(\tilde{H}-\tilde{u}^2) + \tilde{u}\Big((\rho u)_R-(\rho u)_L\Big) - \Big((\rho E)_R-(\rho E)_L\Big)\Big] \\ &\quad + \dfrac{\gamma-1}{\tilde{c}^2}\Big[(\rho_{a,R}-\rho_{a,L})(\tilde{H}-\tilde{u}^2) + (\rho_R-\rho_L)(\tilde{H}_a-2\tilde{u}\tilde{u}_a) \\ &\quad + \tilde{u}_a\Big((\rho u)_R-(\rho u)_L\Big) - \Big((\rho E)_R-(\rho E)_L\Big) + \tilde{u}\Big((\rho u)_{a,R}-(\rho u)_{a,L}\Big) - \Big((\rho E)_{a,R}-(\rho E)_{a,L}\Big)\Big], \\ \alpha_{a,1} &= -\dfrac{\tilde{c}_a}{\tilde{c}^2}\Big[(\rho_R-\rho_L)(\tilde{u}+\tilde{c}) - ((\rho u)_R-(\rho u)_L) - \tilde{c}\alpha_2\Big] \\ &\quad + \dfrac{1}{2\tilde{c}}\Big[(\rho_{a,R}-\rho_{a,L})(\tilde{u}+\tilde{c}) + (\rho_R-\rho_L)(\tilde{u}_a+\tilde{c}_a) - ((\rho u)_{a,R}-(\rho u)_{a,L}) - \tilde{c}_a\alpha_2 - \tilde{c}\alpha_{a,2}\Big], \\ \alpha_{a,3} &= (\rho_{a,R}-\rho_{a,L}) - (\alpha_{a,1}+\alpha_{a,2}). \end{cases}$$

The next proposition states that the two strategies to define $\mathbf{U}^*_{a,L}$ and $\mathbf{U}^*_{a,R}$ are equivalent.

Proposition 2. *The star sensitivities (4.28) solve the system (4.26).*

Proof. We will prove that the star sensitivities defined in (4.28) satisfy the system (4.26).

1. First equation. Writing the first coefficient of (4.27) one easily finds $\rho^*_L - \rho_L = \alpha_1$, and writing the first two coefficients of (4.28) one finds:

$$\rho^*_{a,L} - \rho_{a,L} = \alpha_{a,1}, \qquad (\rho u)^*_{a,L} - (\rho u)_{a,L} = \alpha_{a,1}(\tilde{u} - \tilde{c}) + \alpha_1(\tilde{u}_a - \tilde{c}_a).$$

   We now replace these three expressions in the first equation of (4.26) and we obtain:

$$-\lambda_1 \alpha_{a,1} + \alpha_{a,1}(\tilde{u} - \tilde{c}) + \alpha_1(\tilde{u}_a - \tilde{c}_a) = \partial_a \lambda_1 \alpha_1,$$

   which is always verified, since $\lambda_1 = \tilde{u} - \tilde{c}$.

2. Second equation. We recall that

$$\mathbf{U}_R - \mathbf{U}_L = \sum_{i=1}^{3} \alpha_i \mathbf{r}_i, \qquad \mathbf{U}_{a,R} - \mathbf{U}_{a,L} = \sum_{i=1}^{3} \alpha_{a,i}\mathbf{r}_i + \alpha_i \mathbf{r}_{a,i}.$$

   Therefore, one has:

$$\mathbf{U}^*_R - \mathbf{U}^*_L = \alpha_2 \mathbf{r}_2, \qquad \mathbf{U}^*_{a,R} - \mathbf{U}^*_{a,L} = \alpha_{a,2}\mathbf{r}_2 + \alpha_2 \mathbf{r}_{a,2},$$

   which gives us the following relations:

$$\rho^*_R - \rho^*_L = \alpha_2, \quad \rho^*_{a,R} - \rho^*_{a,L} = \alpha_{a,2}, \quad (\rho u)^*_{a,R} - (\rho u)^*_{a,L} = \alpha_{a,2}\tilde{u} + \alpha_2 \tilde{u}_a.$$

   We now replace them in the second equation of (4.26) and we obtain:

$$-\lambda_2 \alpha_{a,2} + \alpha_{a,2}\tilde{u} + \alpha_2 \tilde{u}_a = \partial_a \lambda_2 \alpha_2,$$

   which is always verified, since $\lambda_2 = \tilde{u}$.

3. Third equation. As we did for the first two equations, one can find the three following expressions:

$$\rho_R - \rho^*_R = \alpha_3, \quad \rho_{a,R} - \rho^*_{a,R} = \alpha_{a,3}, \quad (\rho u)_{a,R} - (\rho u)^*_{a,R} = \alpha_{a,3}(\tilde{u} + \tilde{c}) + \alpha_2(\tilde{u}_a + \tilde{c}_a).$$

   By replacing them in the third equation of (4.26) one can easily check that the equation is always verified, since $\lambda_3 = \tilde{u} + \tilde{c}$.

4. Fourth equation. As we did for the previous equations, one can find the three following expressions:

$$(\rho u)^*_R - (\rho u)^*_L = \alpha_2 \tilde{u}, \quad \rho^*_{a,R} - \rho^*_{a,L} = \alpha_{a,2}, \quad (\rho u)^*_{a,R} - (\rho u)^*_{a,L} = \alpha_{a,2}\tilde{u} + \alpha_2 \tilde{u}_a,$$
$$(\rho E)^*_{a,R} - (\rho E)^*_{a,L} = \alpha_{a,2}\frac{\tilde{u}^2}{2} + \alpha_2 \tilde{u}\tilde{u}_a.$$

   By replacing them in the fourth equation of (4.26) one can easily check that the equation is always verified, since $\lambda_2 = \tilde{u}$.

5. Fifth and sixth equations. The last two equations are the last two components of the following vectorial equation:

$$(\lambda_2 - \lambda_1)\mathbf{U}^*_{a,L} + (\lambda_3 - \lambda_2)\mathbf{U}^*_{a,R} + \lambda_1\mathbf{U}_{a,L} - \lambda_3\mathbf{U}_{a,R} + \mathbf{F}_{a,R} - \mathbf{F}_{a,L} = \Delta x\mathbf{S},$$

which can be rewritten as:

$$\lambda_1(\mathbf{U}_{a,L} - \mathbf{U}^*_{a,L}) + \lambda_2(\mathbf{U}^*_{a,L} - \mathbf{U}^*_{a,R}) + \lambda_3(\mathbf{U}^*_{a,R} - \mathbf{U}_{a,R}) + \mathbf{F}_{a,R} - \mathbf{F}_{a,L} = \Delta x\mathbf{S}.$$

Replacing the definitions (4.28) one finds:

$$-\lambda_1(\alpha_{a,1}\mathbf{r}_1+\alpha_1\mathbf{r}_{a,1})-\lambda_2(\alpha_{a,2}\mathbf{r}_2+\alpha_2\mathbf{r}_{a,2})-\lambda_3(\alpha_{a,3}\mathbf{r}_3+\alpha_3\mathbf{r}_{a,3})+\mathbf{F}_{a,R}-\mathbf{F}_{a,L} = \Delta x\mathbf{S}.$$

We recall that by definition of Roe fluxes, one has:

$$\mathbf{F}_R - \mathbf{F}_L = \sum_{i=1}^{3}\alpha_i\lambda_i\mathbf{r}_i \Rightarrow \mathbf{F}_{a,R} - \mathbf{F}_{a,L} = \sum_{i=1}^{3}\alpha_{a,i}\lambda_i\mathbf{r}_i + \alpha_i\lambda_{a,i}\mathbf{r}_i + \alpha_i\lambda_i\mathbf{r}_{a,i}.$$

Therefore, we obtain:

$$\Delta x\mathbf{S} = \sum_{i=1}^{3}\alpha_i\lambda_{a,i}\mathbf{r}_i = \lambda_{a,1}(\mathbf{U}^*_L - \mathbf{U}_L) + \lambda_{a,2}(\mathbf{U}^*_R - \mathbf{U}^*_L) + \lambda_{a,3}(\mathbf{U}_R - \mathbf{U}^*_R),$$

which is consistent with our discretisation of the source term.

□

4.4.4 Second order MUSCL-type extension

In this section, we extend to the second order the schemes presented above. In time, we use a standard two-step Runge-Kutta method, whilst in space we use a MUSCL-type approach. In a few words the main idea of a MUSCL-type scheme is to consider a replacement of a constant value $\mathbf{V}_j^n$ in each cell by a higher order polynomial $\mathbf{V}_j^n(x)$, $x \in [x_{j-1/2}, x_{j+1/2}]$. The edge values $\mathbf{V}_j^n(x_{j+1/2})$, $\mathbf{V}_{j+1}^n(x_{x+1/2})$ are used as left and right values for the Riemann problem at the interface $j+1/2$; the Riemann problem is then solved as explained in the previous section. However, the definition of the source term (4.13)-(4.14) is valid only if the state is piecewise constant (we refer to the previous Chapter for more details). Therefore, we suggest a piecewise constant state on half of each cell: these two constant values will be denoted $\mathbf{V}_{j\pm1/4}^n$ and correspond to the edge values $\mathbf{V}_j^n(x_{j\pm1/2})$ (see Figure 4.6). In this work, we compute the edge values with a standard approach:

$$\mathbf{V}_{j\pm1/4}^n = \mathbf{V}_j^n \pm \Delta\mathbf{V}_j^n,$$

and the usual choice for $\Delta\mathbf{V}_j^n$ is to use a slope-limiter procedure, for instance:

$$\Delta\mathbf{V}_j^n = \frac{1}{2}\text{minmod}(\mathbf{V}_{j+1}^n - \mathbf{V}_j^n, \mathbf{V}_j^n - \mathbf{V}_{j-1}^n),$$

where

$$\text{minmod}(a,b) = \begin{cases} \text{sgn}(a)\min(|a|,|b|) & \text{if } ab > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We remark that this approach leads to an additional Riemann problem in the middle of the cell: in this way we are able to account for the neglected terms in (4.15).

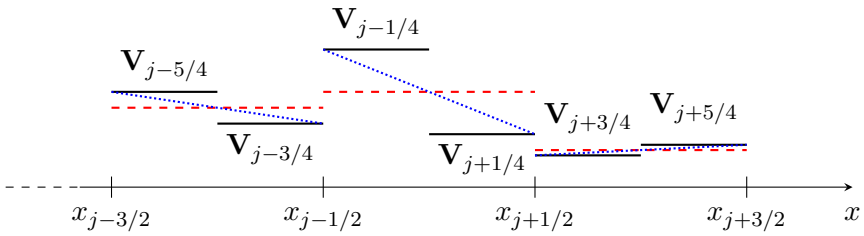
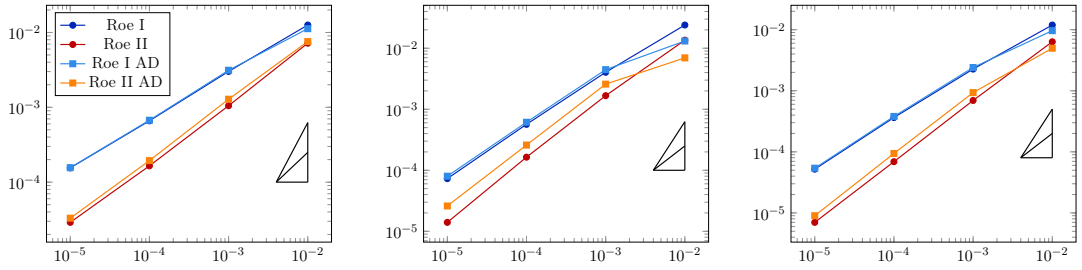


Figure 4.6 – MUSCL discretisation. Dashed red line: first order discretisation. Dotted blue line: classical second order discretisation. Solid black line: second order discretisation used in this work.

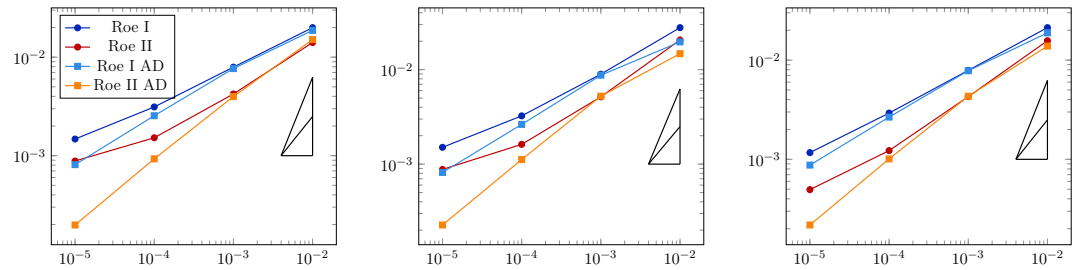


(a) $\|\rho(x,T) - \rho_{ex}(x,T)\|_{L^1}$ (b) $\|u(x,T) - u_{ex}(x,T)\|_{L^1}$ (c) $\|p(x,T) - p_{ex}(x,T)\|_{L^1}$

Figure 4.7 – Convergence test for the state.

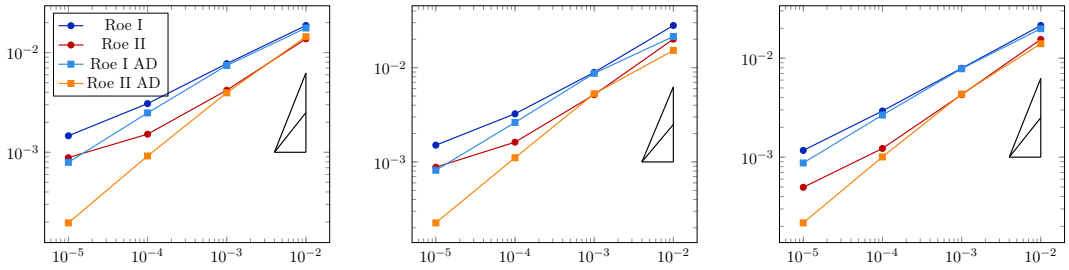
4.5 Convergence tests for the numerical schemes

We consider the Riemann problem of section 4.3. In Figures 4.7-4.8-4.9 we show the convergence of the different numerical schemes presented in Section 4.4. For reference, the L^1 norms of the states ρ_{ex}, u_{ex} and p_{ex} are respectively, 0.5625, 0.2204, and 0.5354. Furthermore, those for the sensitivities $\rho_{a,ex}$, $u_{a,ex}$ and $p_{a,ex}$ are 0.0379, 0.1768, and 0.462. Figure 4.7 shows the convergence for the state: the rate of convergence is the expected one; one can remark that the antidiffusive schemes are slightly less precise than the diffusive ones. In Figures 4.8-4.9 we plot the error for the sensitivity, first with the HLL-type scheme (Figure 4.8) and then with the HLLC-type scheme (Figure 4.9): considering two different star regions for the sensitivity does not seem to make much difference; however one can remark the same effect shown in [CDF17a] for a simpler system: the



(a) $\|\rho_a(x,T) - \rho_{a,ex}(x,T)\|_{L^1}$ (b) $\|u_a(x,T) - u_{a,ex}(x,T)\|_{L^1}$ (c) $\|p_a(x,T) - p_{a,ex}(x,T)\|_{L^1}$

Figure 4.8 – Convergence test for the sensitivity - HLL-type scheme.



(a) $\|\rho_a(x,T) - \rho_{a,ex}(x,T)\|_{L^1}$ (b) $\|u_a(x,T) - u_{a,ex}(x,T)\|_{L^1}$ (c) $\|p_a(x,T) - p_{a,ex}(x,T)\|_{L^1}$

Figure 4.9 – Convergence test for the sensitivity - HLLC-type scheme.

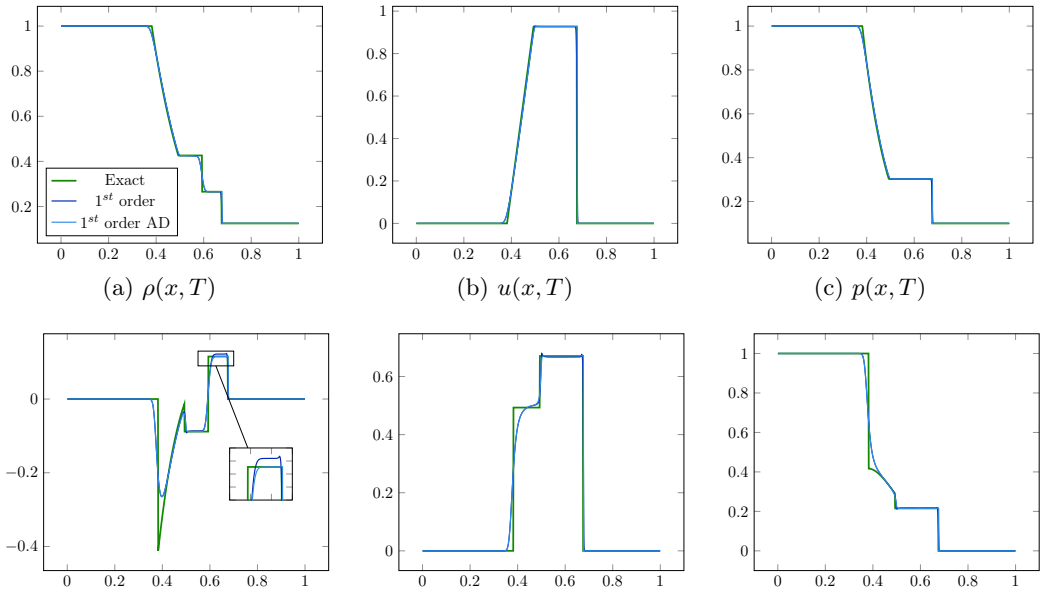
diffusive schemes do not converge for the sensitivity, this is especially evident for the variable ρ_a. In Figure 4.10 we plot the solution at the final time $T = 0.1$, obtained with a mesh $\Delta x = 10^{-3}$ with the first order schemes, both diffusive and antidiffusive (for the sensitivity, the HLL-type scheme has been used): one can notice that the plateau in the right-star zone is not properly captured by the diffusive scheme. This does not change as one refines the mesh, nor with a higher order scheme, as one can see from Figure 4.11. In Figure 4.12 we compare the antidiffusive schemes, first and second order: for the state, the difference is noticeable mainly in the contact discontinuity (therefore only for ρ), whilst for the sensitivity the difference is significant in the neighbourhood of the discontinuities before and after the rarefaction. Finally, in Figure 4.13 we compare the HLL and the HLLC-type schemes for the sensitivity: as anticipated by the error plots, the two schemes are almost equivalent in terms of results. For this reason, the use of HLL-type scheme is preferable, being less expensive from a computational point of view and less complicated to implement.

4.6 Uncertainty Quantification

4.6.1 Problem description

In this section, we show how SA can be used for uncertainty quantification (UQ), cf. [PNTG01, TPB01, Del14]. Many UQ techniques have been developed during the last decades: these methods can be either probabilistic or deterministic. SA falls into the second category while the most well-known of these techniques, the Monte Carlo method, is in the first. Other UQ techniques are polynomial chaos [Wal03, XK03, KM06] and the random space partition [AC12]. The first one is based on a decomposition of the stochastic part of the solution using an orthogonal polynomial basis. Then using a Galerkin method a new system of equations is derived, that provides the coefficients of different statistical quantities, allowing in this way the computation of the statistical moments of the output. Concerning the second approach, the main idea behind it is to consider the random parameters as variable and to solve the system with a finite volume method in a higher dimensional space, the new dimension being the sum of the spatial dimension and the random space dimension. A very good review and comparison of many techniques with applications to fluid dynamics can be found in [WH02].

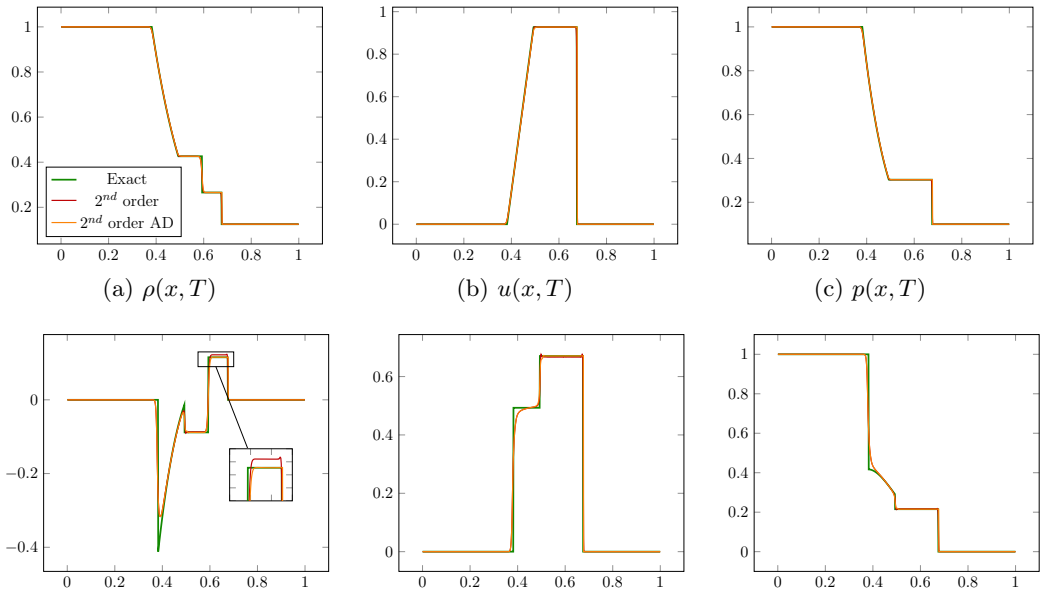
The main aim of UQ is to determine a confidence interval for the output of a model, in our case $\mathbf{U}$, given the uncertainty and the error on the input parameters. According



(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

(d) $\rho_a(x,T)$ (e) $u_a(x,T)$ (f) $p_a(x,T)$

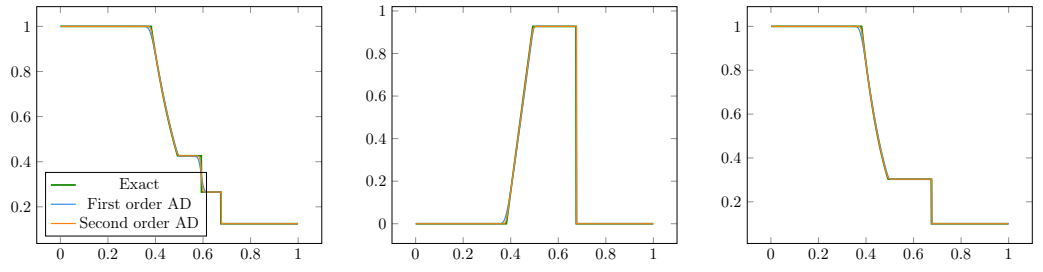
Figure 4.10 – First order schemes, with and without numerical diffusion. HLL-type scheme for the sensitivity.



(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

(d) $\rho_a(x,T)$ (e) $u_a(x,T)$ (f) $p_a(x,T)$

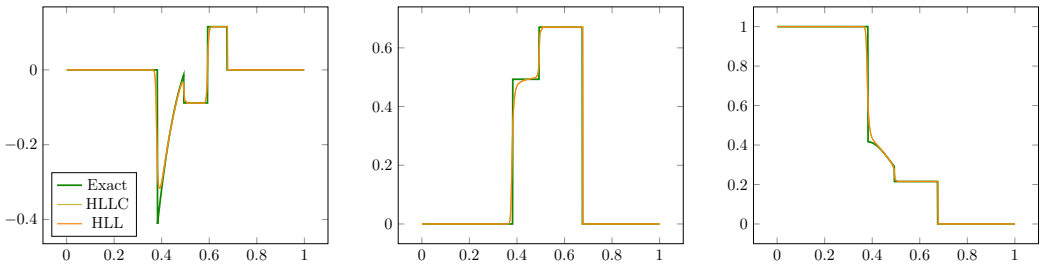
Figure 4.11 – Second order schemes, with and without numerical diffusion. HLL-type scheme for the sensitivity.



(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

(d) $\rho_a(x,T)$ (e) $u_a(x,T)$ (f) $p_a(x,T)$

Figure 4.12 – First and second order schemes, without numerical diffusion. HLL-type scheme for the sensitivity.



(a) $\rho_a(x,T)$ (b) $u_a(x,T)$ (c) $p_a(x,T)$

Figure 4.13 – Second order antidiffusive schemes: HLL and HLLC comparison.

to the AIAA definition, error has a deterministic nature, while uncertainty a stochastic one. Furthermore, uncertainty can be categorised in aleatoric uncertainty and epistemic uncertainty.

In this work, we compare two different UQ methods: Monte Carlo and sensitivity analysis. Both methods aim to provide statistical quantities like moments (mean, variance, ...) of the output of the model. In the following, X will represent one of the variables, considered as random variables, i.e. X can either be ρ, u or p, and X_a the corresponding sensitivity. We use the notation μ_X to indicate the expected value of the variable X and σ_X^2 for its variance. Once these two quantities are known, one can build a confidence interval for the variable X as: $CI_X = [\mu_X - \kappa\sigma_X, \mu_X + \kappa\sigma_X]$. The coefficient κ regulates the amplitude of the interval and it is related to the probability that the variable X will fall within the interval. For instance, the choice $\kappa = 2$ provides an $\sim 95\%$ confidence interval.

Monte Carlo method. Here we briefly introduce the Monte Carlo method, for more details see for instance [CC99]. The Monte Carlo method is a probabilistic technique: to obtain an estimate of the average and of the standard deviation one needs to perform multiple simulations. Let $\mathbf{a}$ be the vector of uncertain parameters, with a known distribution. Then, N random samples $\mathbf{a}_i$ are drawn from this distribution, and for each $\mathbf{a}_i$ the corresponding solution X_i is computed. Then, the unbiased average and variance estimators are used:

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\mu_X - X_i)^2.$$

These estimates are good if N is sufficiently large: the slow convergence, and therefore the high computational cost, is the main limitation of the Monte Carlo method. However, this method is readily parallelisable.

Sensitivity analysis method. SA is a deterministic approach to estimate the average μ_X and the variance σ_X^2 of the output X. Let $\mu_{\mathbf{a}}$ be the average of the uncertain vector $\mathbf{a}$, and $\sigma_{\mathbf{a}}$ the covariance matrix:

$$\mu_{\mathbf{a}} = \begin{bmatrix} \mu_{a_1} \\ \vdots \\ \mu_{a_M} \end{bmatrix}, \qquad \sigma_{\mathbf{a}} = \begin{bmatrix} \sigma_{a_1}^2 & \mathrm{cov}(a_1, a_2) & \dots & \mathrm{cov}(a_1, a_M) \\ \mathrm{cov}(a_1, a_2) & \sigma_{a_2}^2 & \dots & \mathrm{cov}(a_2, a_M) \\ \vdots & & \ddots & \vdots \\ \mathrm{cov}(a_1, a_M) & \dots & & \sigma_{a_M}^2 \end{bmatrix},$$

where M is the number of uncertain parameters, μ_{a_i} the average of the $i-$th parameter, $\sigma_{a_i}^2$ its variance and $\mathrm{cov}(\cdot,\cdot)$ the covariance. Let us consider the first order Taylor expansion for the variable X with respect to the vector of parameters $\mathbf{a}$:

$$X(\mathbf{a}) = X(\mu_{\mathbf{a}}) + \sum_{i=1}^{M}(a_i - \mu_{a_i})X_{a_i}(\mu_{\mathbf{a}}) + o(\|\mathbf{a}\|^2).$$

Then computing the average, since $X(\mu_{\mathbf{a}})$ and $X_{a_i}(\mu_{\mathbf{a}})$ are not random variables, at first order one gets:

$$\mu_X = E[X(\mathbf{a})] = X(\mu_{\mathbf{a}}) + \sum_{i=1}^{M} X_{a_i}(\mu_{\mathbf{a}}) E[a_i - \mu_{a_i}] = X(\mu_{\mathbf{a}}),$$

because $E[(a_i - \mu_{a_i})] = 0$. In the same way, one can compute the variance:

$$\sigma_X^2 = E[(X(\mathbf{a}) - \mu_X)^2] = E\left[\left(\sum_{i=1}^{M} X_{a_i}(\mu_{\mathbf{a}})(a_i - \mu_{a_i})\right)^2\right] =$$

$$= \sum_{i=1}^{M} X_{a_i}^2(\mu_{\mathbf{a}}) E[(a_i - \mu_{a_i})^2] + \sum_{\substack{i,j=1 \\ i \neq j}}^{M} X_{a_i}(\mu_{\mathbf{a}}) X_{a_j}(\mu_{\mathbf{a}}) E[(a_i - \mu_{a_i})(a_j - \mu_{a_j})].$$

Therefore, we obtain the following first order estimates of the average and the variance of the variable X:

$$\mu_X = X(\mu_{\mathbf{a}}), \quad \sigma_X^2 = \sum_{i=1}^{M} X_{a_i}^2 \sigma_{a_i}^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{M} X_{a_i} X_{a_j} \mathrm{cov}(a_i, a_j).$$

Higher order estimates require higher order sensitivities [MD10].

4.6.2 Numerical results

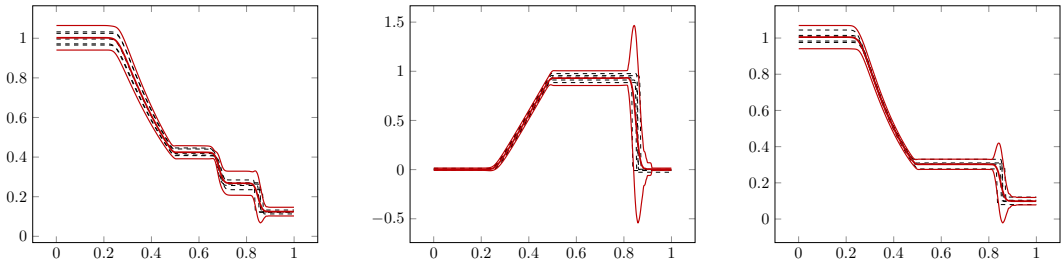
We applied the uncertainty quantification techniques described in the previous subsection to the test case already introduced in section 4.3. The uncertain parameters are the left and right values of the physical variables for the state, i.e.:

$$\mathbf{a} = (\rho_L, \rho_R, u_L, u_R, p_L, p_R)^t,$$

with the following average and covariance matrix:

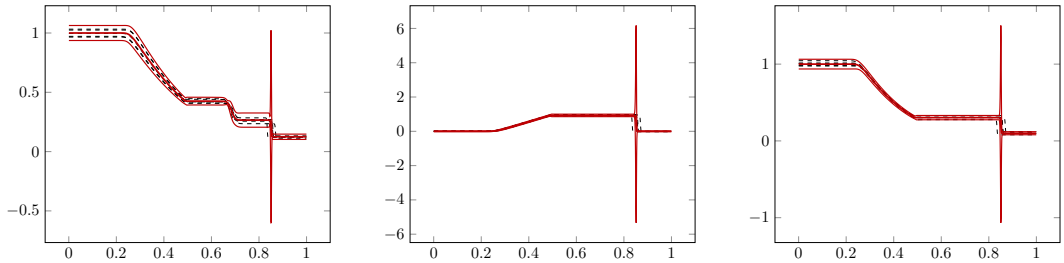
$$\mu_{\mathbf{a}} = (1, 0.125, 0, 0, 1, 0.1)^t, \quad \sigma_{\mathbf{a}} = \mathrm{diag}(0.001, 0.000125, 0.0001, 0.0001, 0.001, 0.0001).$$

This choice means that all the parameters are uncorrelated and we chose as their variance the 0.1% of their average, except for the velocity, whose average is 0. In Figure 4.14 we show the results of the Monte Carlo approach: the average and the average plus and minus twice the standard deviation (i.e. $\kappa = 2$) are plotted in red, five samples are plotted in black. These results are obtained with $N = 1000$ samples, on a mesh with $\Delta x = 10^{-3}$ using a Roe first order diffusive scheme. As one can see, the average process smudges the shock and the standard deviation is larger in that zone. In Figures 4.15-4.16 we show the results of the SA approach, with $\Delta x = 10^{-3}$ and the diffusive first order scheme, when the sensitivity is computed without the correction term (4.13): the spikes in the neighbourhood of the shock are very different with respect to the ones we get with the Monte Carlo approach. On one hand, these peaks lead to non-physical values for the solution (in particular, the confidence intervals contains negative values



(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

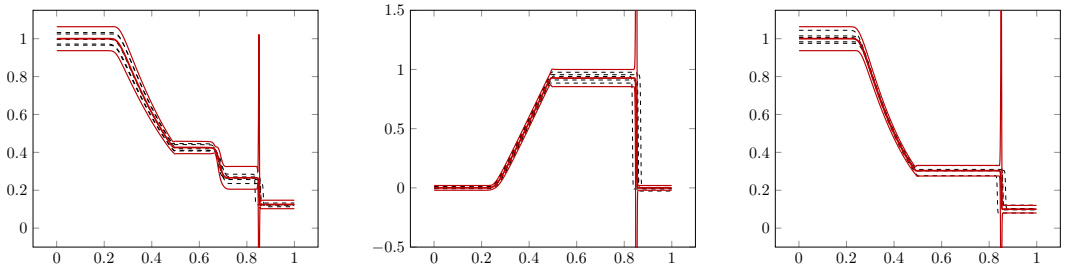
Figure 4.14 – Monte Carlo approach. Average and the average plus and minus twice the standard deviation in red. Five samples in black dashed lines.



(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

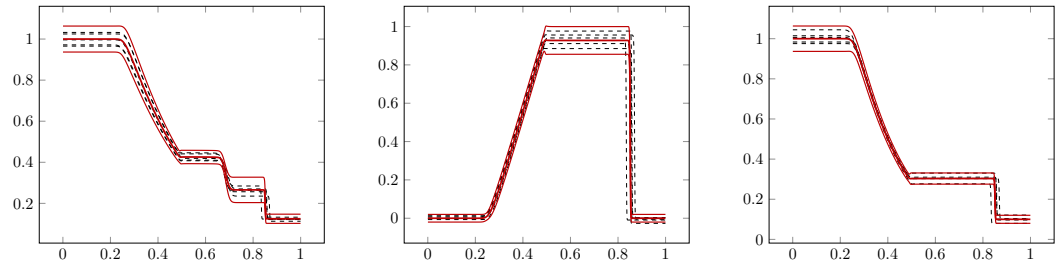
Figure 4.15 – SA approach without correction. Average and the average plus and minus twice the standard deviation in red. Five samples in black dashed lines.

for the pressure and for the density); on the other hand, they do not enlarge the zone enough to contain the majority of the samples: one can observe that four out of five samples fall outside of the predicted interval in the neighbourhood of the shock. For these reasons, we suggest that the corrected sensitivity are more appropriate in this context. The results obtained with the corrected sensitivities are shown in Figure 4.17: the confidence interval obtained correspond to the ones obtained with the Monte Carlo approach, apart for the shock zone. Of course, the SA approach does not capture the uncertainty in the neighbourhood of the shock, because it neglects the dependence of the speed of the shock on the parameters. This is why most of the samples fall out of the zone predicted with the SA approach, and it is the case with and without correction. However, the SA approach is less expensive: the Monte Carlo approach requires 1000 solutions of the state, whilst the SA approach requires only one solution of the state and as many solutions of the sensitivity as the number of uncertain parameters, in this case 6. Furthermore, the solution of the different sensitivities can be done in parallel. Finally, in Figure 4.18 we show the results obtained with the anti-diffusive scheme: the difference with respect to the diffusive scheme is not significant. This is a good news for possible future developments in $2D$: the anti-diffusive scheme is very difficult to adapt in higher dimensional spaces; in fact the Glimm method has been proven not to work in a two-dimensional space. With these results, we underline how the numerical diffusion plays an important role in the convergence of the scheme, but it is not so significant for the final application.



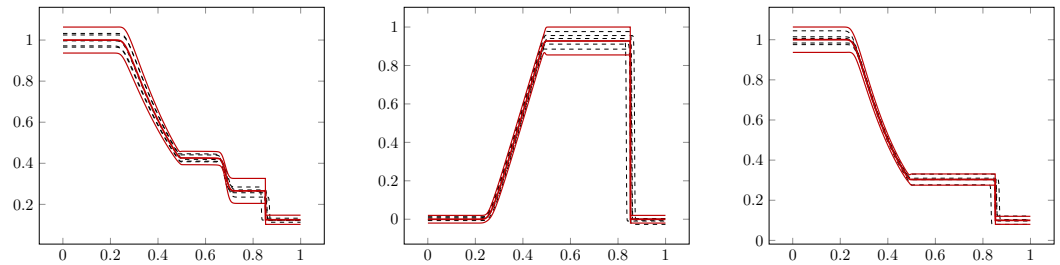
(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

Figure 4.16 – SA approach without correction. Average and the average plus and minus twice the standard deviation in red. Five samples in black dashed lines - zoom.



(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

Figure 4.17 – SA approach with correction. Average and the average plus and minus twice the standard deviation in red. Five samples in black dashed lines.



(a) $\rho(x,T)$ (b) $u(x,T)$ (c) $p(x,T)$

Figure 4.18 – SA approach with correction, anti-diffusive scheme. Average and the average plus and minus twice the standard deviation in red. Five samples in black dashed lines.

5 | Quasi 1D Euler system

5.1 Introduction

5.1.1 State and sensitivity system

The quasi-1D Euler system is written:

$$\begin{cases} \partial_t(h\rho) + \partial_x(h\rho u) = 0, \\ \partial_t(h\rho u) + \partial_x(h\rho u^2 + p) = p\partial_x h, \\ \partial_t(h\rho E) + \partial_x(hu(\rho E + p)) = 0, \end{cases} \tag{5.1}$$

where ρ is the density, u is the velocity, ρE the total energy per mass unit, p the pressure, and $h = h(x) > 0$ is a smooth function of the space x and it is known. This system describes a flow in a nozzle of height h and it allows us to investigate more realistic applications, while remaining in the simpler and computationally less expensive one-dimensional framework. In this work, we considered the following height:

$$h(x) = \begin{cases} 2 - A\sin^2\left(\dfrac{x - x_c}{\ell}\pi - \dfrac{\pi}{2}\right) & x_c - \dfrac{\ell}{2} < x < x_c + \dfrac{\ell}{2}, \\ 2 & \text{otherwise}, \end{cases} \tag{5.2}$$

which is plotted in Figure 5.1. It is described by three parameters: A is the maximal depth, x_c the point of maximal depth and ℓ the length. The system is closed by the following algebraic equation:

$$p = (\gamma - 1)\left(\rho E - \frac{1}{2}\rho u^2\right),$$

where $\gamma = 1.4$ is the heat capacity ratio. We introduce two other quantities which will be useful in the following: the total enthalpy $H = E + \frac{p}{\rho}$ and the speed of sound $c = \sqrt{(\gamma - 1)(H - \frac{1}{2}u^2)}$. We can rewrite the system (5.1) in the vectorial form:

$$\partial_t(h\mathbf{U}) + \partial_x(h\mathbf{F}(\mathbf{U})) = \mathbf{P}\partial_x h, \tag{5.3}$$

where

$$\mathbf{U} = \begin{bmatrix} \rho \\ \rho u \\ \rho E \end{bmatrix} \quad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ u(\rho E + p) \end{bmatrix} \quad \mathbf{P} = \begin{bmatrix} 0 \\ p \\ 0 \end{bmatrix}.$$

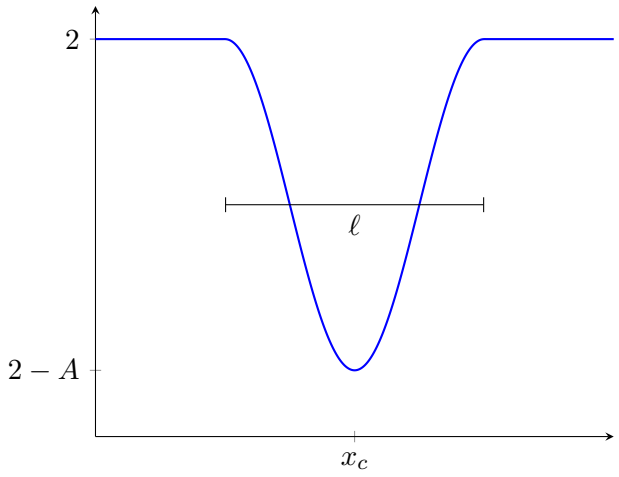


Figure 5.1 – Height of the channel $h(x)$ (5.2).

Since $h(x)$ does not depend on time, one can formally rewrite the system (5.3) as:

$$\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = (\mathbf{P} - \mathbf{F}(\mathbf{U})) \frac{\partial_x h}{h}. \tag{5.4}$$

We remark that the left-hand side corresponds to the classical Euler system.

To obtain the sensitivity system, we differentiate (5.4) with respect to the parameter a. Let h_a be the derivative of the height of the channel with respect to a; then, the sensitivity equations are:

$$\partial_t \mathbf{U}_a + \partial_x \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = (\mathbf{P}_a - \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a)) \frac{\partial_x h}{h} + (\mathbf{P} - \mathbf{F}(\mathbf{U})) \frac{h \partial_x h_a - h_a \partial_x h}{h^2}, \tag{5.5}$$

where we recall that

$$\mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = \partial_a \mathbf{F}(\mathbf{U}) = \begin{bmatrix} (\rho u)_a \\ \rho_a u^2 + 2\rho u u_a + p_a \\ u_a(\rho E + p) + u_a((\rho E)_a + p_a) \end{bmatrix}$$

and $\mathbf{P} = (0, p_a, 0)^t$. Once again, the left-hand side is the sensitivity of the Euler system, and as done previously, a correction term needs to be added to the right-hand side in order to consider the possible discontinuities:

$$\begin{aligned} \partial_t \mathbf{U}_a + \partial_x \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = (\mathbf{P}_a - \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a)) \frac{\partial_x h}{h} \\ + (\mathbf{P} - \mathbf{F}(\mathbf{U}, \mathbf{U}_a)) \frac{h \partial_x h_a - h_a \partial_x h}{h^2} + \mathbf{S}(\mathbf{U}). \end{aligned} \tag{5.6}$$

The source $\mathbf{S}$ is defined as follows:

$$\mathbf{S}(\mathbf{U}) = \sum_{k=1}^{N_s} \boldsymbol{\rho}_k \delta(x - x_{s,k}(t)), \tag{5.7}$$

where $\boldsymbol{\rho}_k$ is the amplitude of the correction, the computation of which is the object of the next subsection.

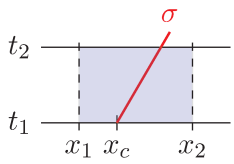


Figure 5.2 – Control volume $\mathcal{C}$ in shaded blue.

5.1.2 Definition of the source term

In this subsection we want to define the amplitudes of the correction terms $\boldsymbol{\rho}_k$. For this purpose, we rewrite equations (5.4) and (5.6) as follows:

$$\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = \mathbf{G}(\mathbf{U}), \tag{5.8}$$

$$\partial_t \mathbf{U}_a + \partial_x \mathbf{F}_a(\mathbf{U}, \mathbf{U}_a) = \mathbf{G}_a(\mathbf{U}, \mathbf{U}_a) + \mathbf{S}(\mathbf{U}). \tag{5.9}$$

We integrate (5.9) over the control volume $\mathcal{C}$ in Figure 5.2, in which it is plotted a single discontinuity propagating at velocity σ, obtaining:

$$\int_{x_1}^{x_2} \mathbf{U}_a(x,t_2) - \mathbf{U}_a(x,t_1)\mathrm{d}x + \int_{t_1}^{t_2} \mathbf{F}_a(x_2,t) - \mathbf{F}_a(x_1,t)\mathrm{d}t = \iint_{\mathcal{C}} \mathbf{G}_a \mathrm{d}x\mathrm{d}t + \int_{t_1}^{t_2} \boldsymbol{\rho}_k(t)\mathrm{d}t.$$

One can formally exchange the double integral over $\mathcal{C}$ with the derivative with respect to a:

$$\sigma\Delta t\mathbf{U}_a^- - \sigma\Delta t\mathbf{U}_a^+ + \Delta t\mathbf{F}_a^+ - \Delta t\mathbf{F}_a^- = \partial_a \iint_{\mathcal{C}} \mathbf{G}\mathrm{d}x\mathrm{d}t + \int_{t_1}^{t_2} \boldsymbol{\rho}_k(t)\mathrm{d}t.$$

Now using equation (5.8), one obtains:

$$\sigma\Delta t\mathbf{U}_a^- - \sigma\Delta t\mathbf{U}_a^+ + \Delta t\mathbf{F}_a^+ - \Delta t\mathbf{F}_a^- = \partial_a \iint_{\mathcal{C}} \partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U})\mathrm{d}x\mathrm{d}t + \int_{t_1}^{t_2} \boldsymbol{\rho}_k(t)\mathrm{d}t.$$

The double integral over $\mathcal{C}$ can now easily be computed:

$$\sigma\Delta t\mathbf{U}_a^- - \sigma\Delta t\mathbf{U}_a^+ + \Delta t\mathbf{F}_a^+ - \Delta t\mathbf{F}_a^- = \partial_a \left(\sigma\Delta t\mathbf{U}^- - \sigma\Delta t\mathbf{U}^+ + \Delta t\mathbf{F}^+ - \Delta t\mathbf{F}^-\right) + \int_{t_1}^{t_2} \boldsymbol{\rho}_k(t)\mathrm{d}t.$$

Finally, dividing by Δt and taking the limit as the control volume goes to zero, one obtains the following definition for the $k-$th amplitude:

$$\boldsymbol{\rho}_k(t) = \sigma_{k,a}(\mathbf{U}^+ - \mathbf{U}^-). \tag{5.10}$$

We remark that we obtain the same definition of the amplitude of the correction term as we did in the previous Chapters for the conservative systems considered. However, here we do not use Rankine-Hugoniot conditions to compute $\boldsymbol{\rho}_k(t)$. This means that we can use all the numerical techniques already developed not only for the conservative part of the system, but also for the discretisation of the correction term. This is the subject of the next section.

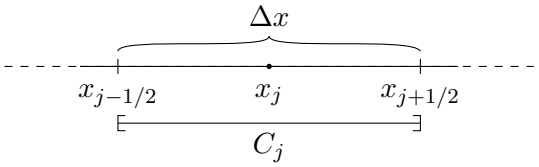


Figure 5.3 – Spatial discretisation.

5.2 Numerical schemes

In this section we briefly describe the numerical schemes implemented for the state and for the sensitivity.

We consider a uniform grid in space with a constant step Δx, x_j is the center of the $j-$th cell C_j, whose extrema are $x_{j-1/2}$ and $x_{j+1/2}$ (cf. Figure 5.3). We use an adaptive time step Δt^n, chosen according to a CFL condition, and the intermediate times are $t^{n+1} = t^n + \Delta t^n$. We indicate with $\mathbf{V}_j^n = (\mathbf{U}_j^n, \mathbf{U}_{a,j}^n)^t$ the average value of the state and the sensitivity in the cell C_j at time t^n.

We remark that the left-hand side of the state equation (5.4) is the classical Euler system (4.3) and the left-hand side of the sensitivity equation (5.6) is identical to the left-hand side of the sensitivity of the Euler system (4.6). Therefore, the same numerical schemes used to solve the Euler system and its sensitivity can be used here to discretise the conservative part. For more details, cf. Chapter 4.

Concerning the right-hand side of the state equation (5.4) and its corresponding part in the sensitivity equation (5.6), we propose a very simple discretisation: the term $(\mathbf{P} - \mathbf{F}(\mathbf{U}))$ is evaluated in each cell, whilst for the derivative of $h(x)$ we can either use the analytical derivative evaluated in the center of each cell, or a centred finite-difference discretisation.

As we did for all the systems considered in this thesis, for the conservative part of the system we use Godunov-type schemes, which consist of two main steps: first, the solution of the Riemann problem at each interface $x_{j-1/2}$ at time t^n; then, the projection step in order to obtain a piecewise constant solution on the mesh:

- *Riemann solvers.* For the first step, we use two different approximate Riemann solvers, one for the state and one for the sensitivity. For the state, we choose the approximate Riemann solver of Roe. This choice is motivated by the fact that to compute a correction term of the form (5.7) we need to be very accurate for the state in the shock and this solver has the property of being exact for an isolated shock. However, different numerical schemes can be used, provided they have two intermediate states, as already discussed in the previous chapter. This means that the HLL solver cannot be used for the state. Concerning the sensitivity, we analysed two different approximate Riemann solvers in the previous chapter: an HLL-type scheme and an HLLC-type scheme. Numerical results show that these two schemes provide almost the same solution (the difference is negligible), the HLL-type scheme being less expensive from a computational point of view. For this reason, in this framework we adopted the HLL-type scheme for the sensitivity.

- *Projection step.* In a classical Godunov method, the projection step is usually per-

formed just by averaging over each cell the solution of the Riemann problem computed in the previous step. At the moment, we have not considered anti-diffusive schemes for this kind of problem. This is mainly motivated by two reasons: first, numerical results in the previous chapter show how the role of numerical diffusion is very important to have the expected order of convergence, but is less significant when it comes to applications; secondly, for this system we are interested in stationary solutions and an anti-diffusive scheme such as the one used in the previous chapters can significantly slow down the convergence in time if, for instance, a shock keeps jumping back and forth between two adjacent cells. For similar reasons we do not consider higher order schemes for the moment.

5.3 Stationary solutions

The system (5.4) admits stationary solutions. As a first step in order to validate the code, we want to be able to reproduce the results obtained in [TAI11, GP01] for two test cases with and without shock and described hereafter in details (subsections 5.3.2-5.3.3). Let us first address the boundary conditions definition.

5.3.1 Boundary conditions

Inlet and outlet boundary conditions have to be considered to compute the fluxes at domain extremities. In [TAI11, GP01], they consider the triplet (H, p, p_{tot}) (p_{tot} being the total pressure, defined as $p_{tot} = p + \frac{1}{2}\rho u^2$), and depending on the test case they impose the value of these physical quantities at the inlet and/or at the outlet in a fictitious external cell, before computing the boundary flux. The quantities not imposed are extrapolated from the interior adjacent cell. The fluxes are based on the conservative variables $(\rho, \rho u, \rho E)$, therefore the first thing to do is to obtain the triplet $(\rho, \rho u, \rho E)$ from (H, p, p_{tot}):

$$\begin{cases} H &= E + \dfrac{p}{\rho}, \\ p &= (\gamma - 1)\left(\rho E - \dfrac{1}{2}\rho u^2\right), \\ p_{tot} &= p + \dfrac{1}{2}\rho u^2, \end{cases} \Rightarrow \begin{cases} \rho &= \left(p_{tot} + \dfrac{p}{\gamma - 1}\right)\dfrac{1}{H} \\ \rho u &= \sqrt{2\rho(p_{tot} - p)} \\ \rho E &= \dfrac{p}{\gamma - 1} + \dfrac{1}{2}\rho u^2. \end{cases} \tag{5.11}$$

We remark two things about this:

(i) this result is valid only if $u > 0$, otherwise we should take $\rho u = -\sqrt{2\rho(p_{tot} - p)}$. In our test cases we will always consider a positive velocity;

(ii) the square root $\sqrt{2\rho(p_{tot} - p)}$ should not present any problem because $\rho > 0$ and $p_{tot} > p$. However, this second inequality is valid cell by cell: if neither p nor p_{tot} are imposed at the boundary, the inequality is naturally verified; if one wants to impose both of them, the constraint needs to be verified; finally, if we are in a case in which we impose only one of them (and the other one's value is taken from the first cell inside the domain) the inequality is not guaranteed to be verified. To avoid this problem, if $p_{tot} < p$ for some time step, we force $\rho u = 0$.

Concerning the sensitivity boundary conditions, for each variable we impose the same type of condition as for the state: for instance, if ρ is imposed as inlet external state, the corresponding sensitivity ρ_a is imposed as inlet external value as well. The values for $(\rho_a, (\rho u)_a, (\rho E)_a)$ are obtained by differentiating with respect to a the second system of (5.11):

$$\begin{cases} \rho_a & = \left(p_{tot,a} + \dfrac{p_a}{\gamma - 1} - \rho H_a \right) \dfrac{1}{H}, \\ (\rho u)_a & = \dfrac{\rho_a (p_{tot} - p) + \rho (p_{tot,a} - p_a)}{\rho u}, \\ (\rho E)_a & = \dfrac{p_a}{\gamma - 1} + u(\rho u)_a - \dfrac{1}{2} \rho_a u^2. \end{cases}$$

Once again, this is valid only if $u > 0$.

5.3.2 Isentropic transonic case

The first test case, taken from [TAI11, GP01], that we want to reproduce is an isentropic transonic case, i.e. with the following boundary conditions for the state:

$$\begin{cases} H_L = 4, \\ p_{tot,L} = 2, \end{cases}$$

and all the other variables are extrapolated from the interior solution. The value of the parameters describing h are $A = 1$, $\ell = 0.5$, and $x_c = 0.5$. We do not have an analytical solution for the state, however the analytical Mach number ($Ma = \frac{u\sqrt{\rho}}{\sqrt{\gamma p}}$) is known and it is plotted in Figure 5.4 for the case considered, compared to the numerical one obtained with our numerical scheme: the two perfectly match. The Mach number is continuous in this case, as are all the physical quantities. The passage from the subsonic regime ($Ma < 1$) to the supersonic one ($Ma > 1$) happens in the neck of the nozzle, i.e. $x = x_c$, as expected. To validate the numerical results obtained for the sensitivity, we compute the empirical sensitivity, defined as follows:

$$\mathbf{U}_a^{emp} = \frac{\mathbf{U}(a + \delta a) - \mathbf{U}(a)}{\delta a},$$

for a sufficiently small δa. In Figure 5.5 we show in colour the sensitivity obtained with our numerical schemes, and in black the empirical sensitivity computed with $\delta a = 10^{-3}$, on a mesh of constant spatial step $\Delta x = 10^{-3}$ with a first order Roe scheme. Furthermore, we consider that at the continuous level:

$$\mathrm{ERR}(\mathbf{U}) := \|\mathbf{U}(a + \delta a) - \mathbf{U}(a) - \delta a \mathbf{U}_a(a)\|_{L^1}, \tag{5.12}$$

which is supposed to be $O(\delta a^2)$. Figure 5.7 shows (5.12) for all the variables for different values of δa computed with the discrete solutions. We remark that for ρ and u we have the expected rate of convergence for the bigger δa: the slight change in the slope for smaller δa corresponds to the point in which $O(\Delta x^r)$, r being the order of the scheme used, gets comparable with $O(\delta a^2)$ and consequently the error due to the discretisation gets comparable with the one of the Taylor expansion. We recall that the spatial step is $\Delta x = 10^{-3}$ and the scheme is a first order Roe scheme. Concerning $\mathrm{ERR}(p)$, we do

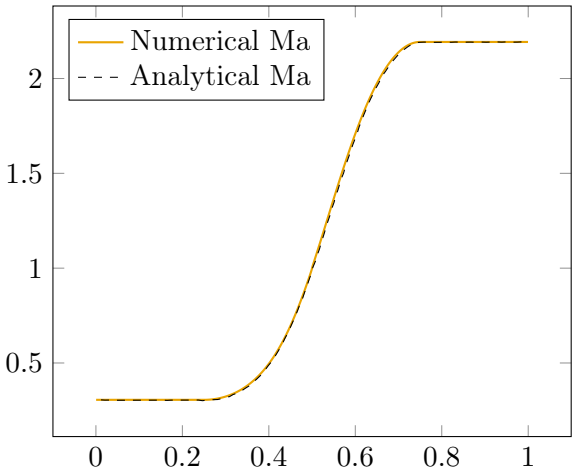


Figure 5.4 – Numerical results Mach number comparision - isentropic transonic case.

not have the expected order of convergence even for larger δa. We remark however, that the case here considered presents a transonic rarefaction and it is well known that a Roe scheme is not able to capture that. To overcome this problem, an entropy fix has been implemented: however, a small jump still remains in the state at $x = 0.5$ (cf. left side of Figure 5.6) and this generates the peak in the sensitivity noticeable in the right side of Figure 5.6, which deteriorates the convergence.

5.3.3 Transonic case with shock

The second test case we consider is a transonic case with shock, i.e. with the following boundary conditions for the state:

$$\begin{cases} H_L = 4, \\ p_{tot,L} = 2, \\ p_R = 1.6, \end{cases}$$

and all the other variables are extrapolated from the interior solution. The value of the parameters describing h are the same as the previous test case, i.e. $A = 1$, $\ell = 0.5$, and $x_c = 0.5$. We show in Figure 5.8 the comparison of the analytical and numerical Mach number for the case considered. As expected, the results are identical. The Mach number and all the physical variables present a discontinuity at $x \simeq 0.6$. As in the previous case, the transition from subsonic to supersonic happens at $x = x_c$. In Figure 5.9 we plotted the sensitivity of the three variables with respect to $a = H_L$, and we compare the scheme with and without correction term $\mathbf{S}(\mathbf{U})$: as one cas see, the two solutions are identical in the whole domain except for the spikes at $x \simeq 0.6$ which are present if the correction term is not taken into account. In this case, we do not compute ERR($\mathbf{U}$) for different δa, because it comes from a first order Taylor expansion, which of course is not valid in case of discontinuity.

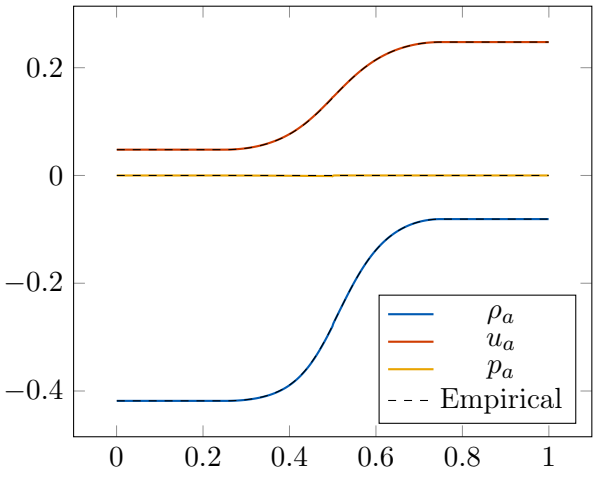


Figure 5.5 – Numerical vs empirical sensitivity for the isentropic transonic case.

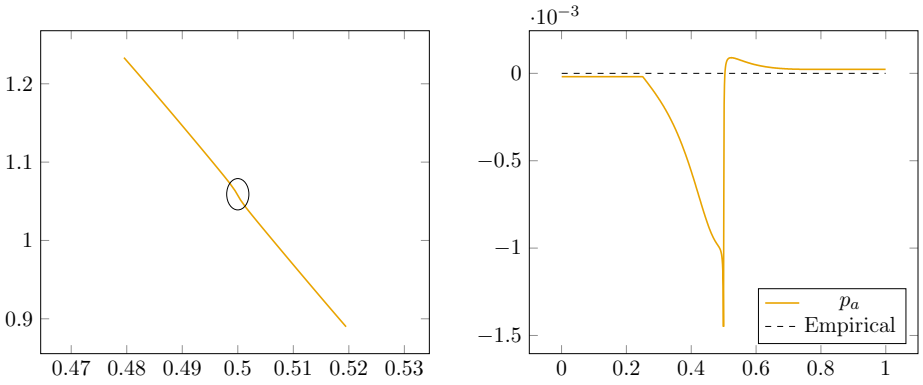


Figure 5.6 – Small pressure shock due to transonic rarefaction on the left and its sensitivity on the right.

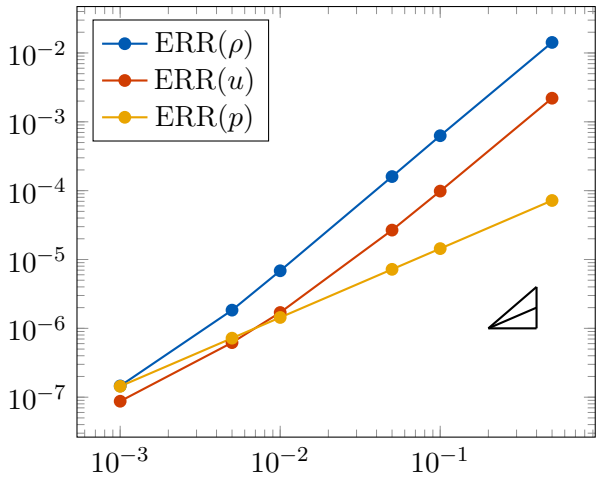


Figure 5.7 – L^1 norm of the error for Roe first order scheme - isentropic transonic case.

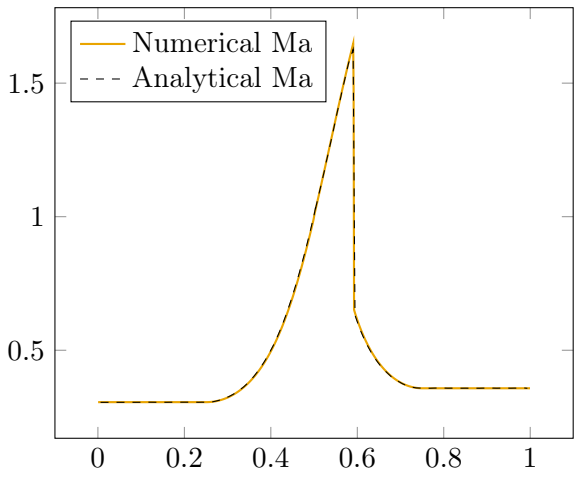


Figure 5.8 – Numerical vs Analytical Mach number - shocked flow.

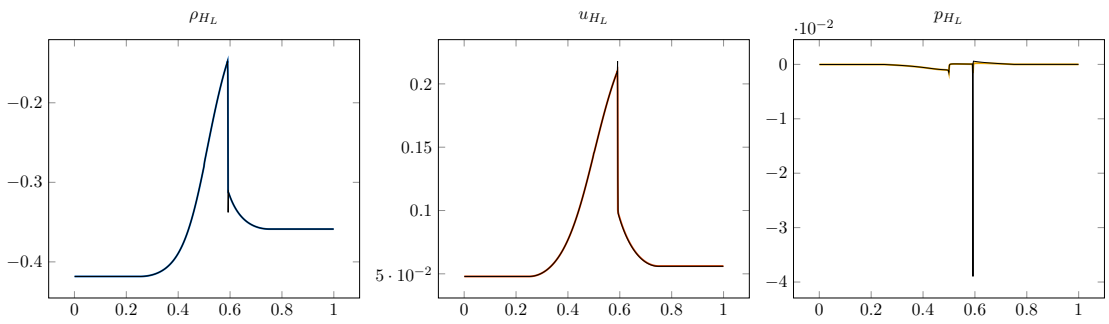


Figure 5.9 – Scheme with correction (in color) and without correction (in black) for the sensitivity.

5.4 Optimization

5.4.1 Problem description

Sensitivity analysis can be used to solve PDE constrained optimization problems such as:

$$\min_{\mathbf{a}\in\mathcal{A}} J(\mathbf{U}),$$

subject to a system of PDEs, for instance (5.3). J is the cost functional and it depends on the vector of parameters $\mathbf{a}$ through the state $\mathbf{U}$. The set $\mathcal{A}$ is the set of admissible parameters. The existence of the minimum is guaranteed under some hypotheses: continuity of the cost functional J and compactness of the set $\mathcal{A}$. Moreover, if the cost functional is convex with respect to the parameters, the minimum is unique.

Let us consider a problem where the cost functional can be written as a bilinear form:

$$J(\mathbf{U}) = \frac{1}{2} b(\mathbf{U}, \mathbf{U}).$$

Classical optimization algorithms require the gradient of cost functional, whose $i-$th component can be written as:

$$[\nabla_{\mathbf{a}} J(\mathbf{U})]_i = \partial_{a_i} J(\mathbf{U}) = b(\mathbf{U}, \mathbf{U}_{a_i}). \tag{5.13}$$

One of the most used techniques to compute such gradients is the *adjoint equation method* [Jam88, MP01, Pir74], which introduces additional adjoint variables to compute the derivative of any functional output with respect to all input parameters. The adjoint equation is independent of the input parameters, thus this approach is very efficient for optimization problems involving a large number of design parameters. However, the adjoint equation should be solved backwards in time, which could lead to practical difficulties for unsteady problems. Here, we use the continuous sensitivity equation method: one sensitivity system has to be solved for each parameter a_i. However, all the sensitivity systems are independent of each other, therefore they can be solved in parallel.

5.4.2 Optimization algorithm

In this section we briefly describe the optimization algorithm used to obtain the results presented in the next section. It is a projected gradient descent method and it is detailed in Algorithm 1. First, the state $\mathbf{U}$ is solved and the cost functional J is evaluated in the initial parameters $\mathbf{a}$. Secondly, all the sensitivity equations are solved (line 7) in order to compute the gradient of the cost functional $\nabla_{\mathbf{a}} J$ (5.13): we remark that this step is performed in parallel, all the sensitivities being independent from each other. Finally, a line search is performed (lines 15-20): notice that for this step, only the solution of the state equation is required. The parameters value is updated and the loop is repeated until convergence.

5.4.3 Test cases

In all the test cases presented in this section, we deal with *pressure matching* problems, i.e. the cost functional is the following:

$$J(\mathbf{U}) = \frac{1}{2} \|p - p^*\|^2_{L^2}, \tag{5.14}$$

Algorithm 1 Projected gradient descent method using sensitivities

Choice of initial parameters $\rightarrow \mathbf{a}$
State solution $\rightarrow \mathbf{U}$
Evaluation of $J(\mathbf{U}) \rightarrow J$
Initialisation of $\mathbf{a}_{new}$
while $\|\mathbf{a}_{new} - \mathbf{a}\| >$toll **do**
    $r \leftarrow r_0$
    Parallel sensitivities solution $\rightarrow \mathbf{U}_{a_i}$
    Computation of $\nabla_{\mathbf{a}} J(\mathbf{U}, \mathbf{U}_{\mathbf{a}})$
    $\mathbf{a}_{new} = \mathbf{a} - r\nabla_{\mathbf{a}} J$
    if $\mathbf{a}_{new} \notin \mathcal{A}$ **then**
        Projection of $\mathbf{a}_{new}$ onto $\mathcal{A}$
    end if
    State solution $\rightarrow \mathbf{U}$
    Evaluation of $J(\mathbf{U}) \rightarrow J_{new}$
    while $J_{new} > J$ **do**
        $r \leftarrow \frac{1}{2} r$
        $\mathbf{a}_{new} = \mathbf{a} - r\nabla_{\mathbf{a}} J$
        State solution $\rightarrow \mathbf{U}$
        Evaluation of $J(\mathbf{U}) \rightarrow J_{new}$
    end while
    $J = J_{new}$ $\mathbf{a} = \mathbf{a}_{new}$
end while

where p^* is the target pressure.

Test case 1

For the first test case, we consider two parameters of optimization: A and ℓ (cf. definition of $h(x)$ (5.2) and Figure 5.1). The set of admissible parameters $\mathcal{A}$ is the rectangle $\mathcal{A} = [0, 2) \times [0, 1]$. We remark that $\mathcal{A}$ is open ($A = 2$ is not admitted because we ask for $h(x) > 0$). This is not a problem because we will consider a test case with minimum inside the domain; on the other hand, $\ell = 0$ is to be intended as $h(x) = \text{const.}\ \forall x \in (0, 1)$. The target pressure is the pressure obtained with $A = 1$ and $\ell = 0.5$: in this way, the optimum is known and reachable and we aim at recovering this value with our algorithm. The gradient of the cost functional is the following:

$$\nabla_{\mathbf{a}} J(\mathbf{U}) = \begin{bmatrix} (p - p^*, p_A)_{L^2} \\ (p - p^*, p_\ell)_{L^2} \end{bmatrix},$$

where we used the notation $(\cdot, \cdot)_{L^2}$ to indicate the L^2 dot product. The two sensitivities $\mathbf{U}_A$ and $\mathbf{U}_\ell$ can be computed by solving the system (5.6), with

$$h_A(x) = \begin{cases} \sin^2\left(\dfrac{x - x_c}{\ell}\pi - \dfrac{\pi}{2}\right) & x_c - \frac{\ell}{2} < x < x_c + \frac{\ell}{2}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$h_\ell(x) = \begin{cases} -\frac{\pi(x - x_c)}{\ell^2} \sin\left(2\pi \frac{x - x_c}{\ell}\right) & x_c - \frac{\ell}{2} < x < x_c + \frac{\ell}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Isentropic transonic flow. First we deal with the isentropic transonic case, already introduced in the previous section. In Figure 5.10 we show the cost functional: as one can see, it is quite flat in one direction, which can significantly slow down the convergence of the gradient descent method. Furthermore, one can remark a discontinuity: this is due to the fact that, when the nozzle is too deep (i.e. bigger A), a shock occurs in the state solution. In Figure 5.11 we show the optimization algorithm steps obtained with a first order Roe scheme with a constant spatial step $\Delta x = 10^{-2}$. We chose as starting parameters $(A, \ell) = (0.5, 0.2)$. The algorithm stops after 154 iterations at $(A, \ell) = (1.0003, 0.5002)$, with a value of the cost functional $J = 2.2013 \times 10^{-9}$. We remark that there is no crossing of the cost functional discontinuity. In Figure 5.12 we show the value of the cost functional J and of the euclidean norm of its gradient $\nabla_{\mathbf{a}} J$ with respect to the number of iterations in a semi-logarithmic scale. One can notice that the decrease of the cost functional is faster at the beginning and then slows down significantly, as we expected, given the shape of the cost functional.

Shocked flow. The second optimization test case we present is the shocked flow case. Here, because of the presence of a shock in the solution, we compare the results of two different numerical schemes: with and without the correction term (5.7). In Figure 5.13 we show the cost functional: it has a quite different shape with respect to the previous test caseIn this test case, first we computed the sensitivity without the correction term,

i.e. by solving the equations (5.5), and then with the correction term, i.e. by solving (5.6). This leads to different gradients: in fact, if we think of the analytical solution for the sensitivities p_A and p_ℓ, with and without the Dirac delta function, there is a missing term in the second case, when computing the integrals to obtain $\nabla_{\mathbf{a}} J(\mathbf{U})$. On the other hand, considering the discrete solution, there is an approximation error due to the Dirac delta function, which cannot be seized numerically. In Figure 5.14 we show the steps of the gradient descent method: in blue the results obtained without the correction and in red the results obtained with the correction. As one can see from the zoom (right side of the Figure), the convergence is smoother if the gradient is computed with the corrected sensitivities. This can be seen also from Figure 5.15, in which the cost functional and its gradient are plotted in a semi-logarithmic scale with respect to the iterations: we remark that the difference in the value of the cost functional is unimportant, while the gap between the gradients is more significant. The algorithm with the corrected sensitivities converges in 169 iterations to the point $(A, \ell) = (1.0003, 0.5002)$ with a value of the cost functional $J = 1.2185 \times 10^{-8}$, as opposed to the one with the uncorrected sensitivities, which stops in 196 iteration at the point $(A, \ell) = (1.0003, 0.5001)$, with $J = 1.6821 \times 10^{-9}$.

Test case 2

Here, we consider two other parameters of optimization: x_c and ℓ (cf. definition of $h(x)$ (5.2) and Figure 5.1). The set of admissible parameters $\mathcal{A}$ is a closed triangle:

$$\mathcal{A} = \{(x_c, \ell) \in \mathbb{R}^2 : 0 \leq x_c \leq 1,\ 0 \leq \ell \leq \min(2x_c, 2 - 2x_c)\},$$

in order for the nozzle not to exit the domain. We remark that $\ell = 0$ means $h(x) =$ const. $\forall x \in (0, 1)$. The target pressure is the pressure obtained with $x_c = 0.5$ and $\ell = 0.5$: in this way, the optimum is again known and reachable. The gradient of the cost functional is the following:

$$\nabla_{\mathbf{a}} J(\mathbf{U}) = \begin{bmatrix} (p - p^*, p_{x_c})_{L^2} \\ (p - p^*, p_\ell)_{L^2} \end{bmatrix}.$$

The two sensitivities $\mathbf{U}_{x_c}$ and $\mathbf{U}_\ell$ can be computed by solving the system (5.6), with

$$h_{x_c}(x) = \begin{cases} -\frac{\pi}{\ell} \sin\left(2\pi \frac{x - x_c}{\ell}\right) & x_c - \frac{\ell}{2} < x < x_c + \frac{\ell}{2}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$h_\ell(x) = \begin{cases} -\frac{\pi(x - x_c)}{\ell^2} \sin\left(2\pi \frac{x - x_c}{\ell}\right) & x_c - \frac{\ell}{2} < x < x_c + \frac{\ell}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Isentropic transonic flow. The first optimization test case we deal with is the isentropic transonic case, already introduced in the previous section. In Figure 5.16 we show the cost functional: as one can see from the left plot, changes in the parameter x_c affect the value of J way more than changes in the parameter ℓ. This leads to a cost functional which is quite flat in one direction and that can significantly slow down the convergence of the gradient descent method. On the right part of Figure 5.16, it

is plotted the cost functional with respect to ℓ for $x_c = 0.5$: as one can see, the minimum is in $(x_c, \ell) = (0.5, 0.5)$. In Figure 5.17 we show the optimization algorithm steps obtained with a first order Roe scheme with a constant spatial step $\Delta x = 10^{-2}$. We chose as starting parameters $(x_c, \ell) = (0.4, 0.4)$. The algorithm stops after 88 iterations at $(x_c, \ell) = (0.499998, 0.49974)$, with a value of the cost functional $J = 4.36936 \times 10^{-9}$. In Figure 5.18 we show the value of the cost functional J and of the euclidean norm of its gradient $\nabla_{\mathbf{a}} J$ with respect to the number of iterations in a semi-logarithmic scale. One can notice that the decrease of the cost functional is faster at the beginning and then slows down significantly, as we expected, given the shape of the cost functional. Concerning the oscillations of the gradient, they are due to the naïf choice of the step and a relaxation step would probably guarantee a faster convergence.

Shocked flow. The second optimization test case we present is the shocked flow case. Here, because of the presence of a shock in the solution, we compare the results obtained with and without the correction term (5.7), with the diffusive Roe first order scheme. In Figure 5.19 we show the cost functional: againt it has a quite different shape with respect to the previous isentropic transonic test case. In this test case, first we computed the sensitivity without the correction term, i.e. by solving the equations (5.5), and then with the correction term, i.e. by solving (5.6). This leads to different gradients. In Figure 5.20 we show the steps of the gradient descent method: in blue the results obtained without the correction and in red the results obtained with the correction. As one can see from the zoom (right side of the Figure), the convergence is faster if the gradient is computed with the corrected sensitivities. This can be seen also from Figure 5.21, in which the cost functional and its gradient are plotted in a semi-logarithmic scale with respect to the iteration: the algorithm with the corrected sensitivities converges in 9 iterations to the point $(x_c, \ell) = (0.5002, 0.498539)$ with a value of the cost functional $J = 4.50852 \times 10^{-7}$, as opposed to the one with the uncorrected sensitivities, which stops in 36 iteration at the point $(x_c, \ell) = (0.502818, 0.477847)$, with $J = 1.12238 \times 10^{-4}$.

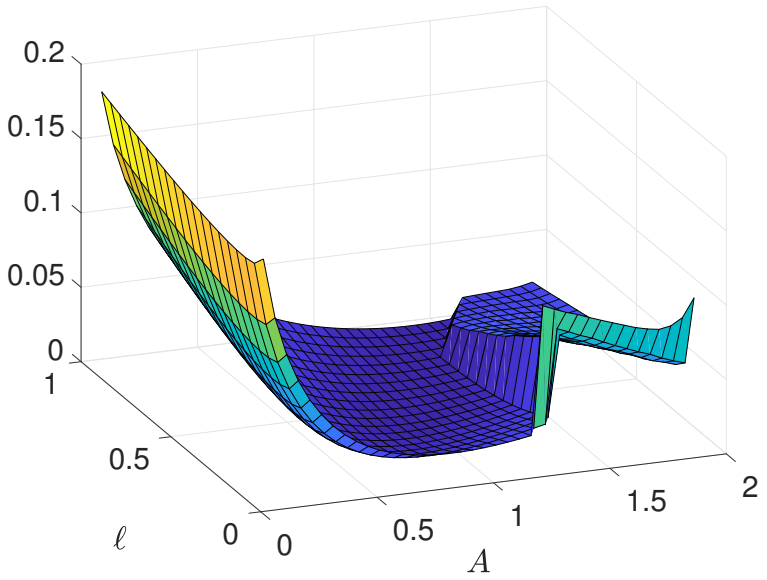


Figure 5.10 – Cost functional (5.14) for the isentropic transonic case, $\mathbf{a} = (A, \ell)$.

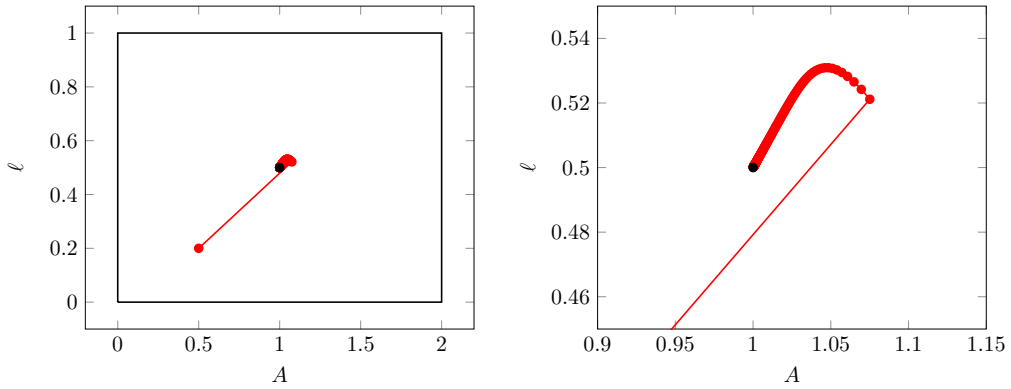


Figure 5.11 – On the left: optimization algorithm steps for the isentropic transonic case, $\mathbf{a} = (A, \ell)$. On the right: emphasising the behaviour near optimality.

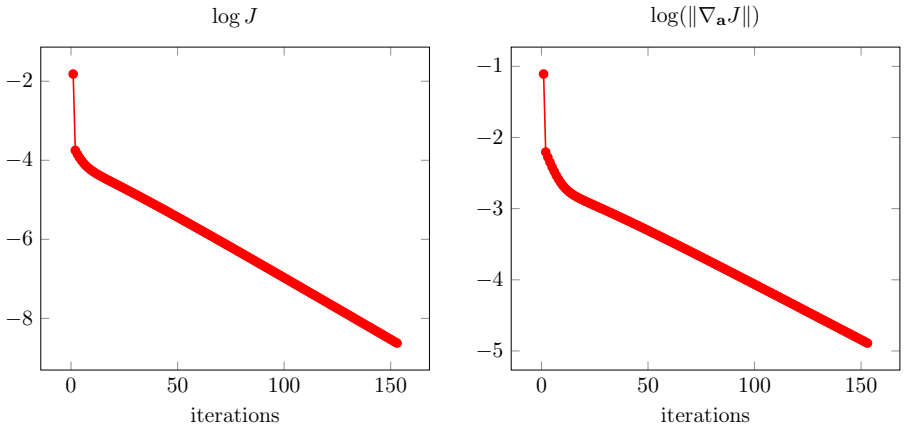


Figure 5.12 – Value of the cost functional and norm of its gradient with respect to iterations in a semilogarithmic scale - isentropic transonic case, $\mathbf{a} = (A, \ell)$.

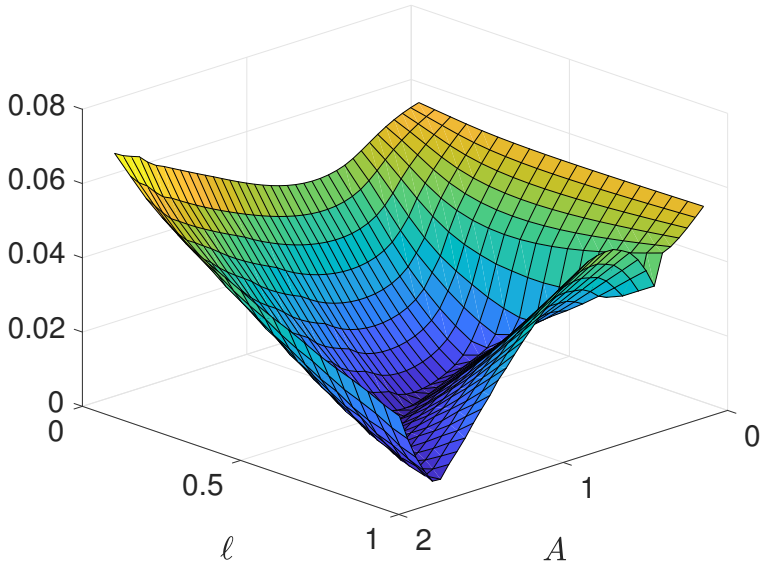


Figure 5.13 – Cost functional (5.14) for the shocked case, $\mathbf{a} = (A, \ell)$.

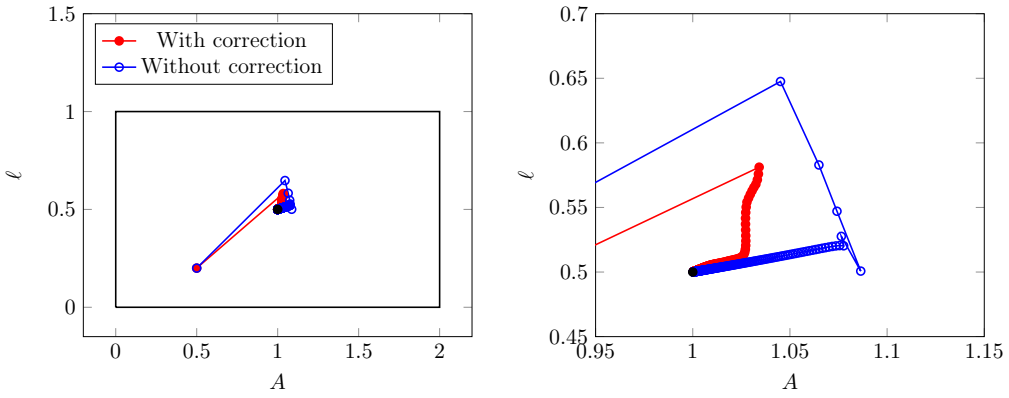


Figure 5.14 – On the left: optimization algorithm steps for the shocked flow case, $\mathbf{a} = (A, \ell)$. On the right: emphasising the behaviour near optimality.

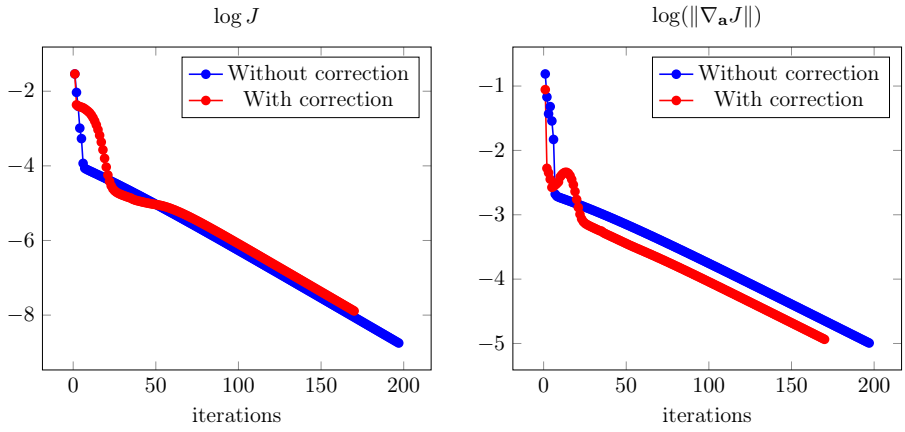


Figure 5.15 – Value of the cost functional and norm of its gradient with respect to iterations in a semilogarithmic scale - shocked flow case, $\mathbf{a} = (A, \ell)$.

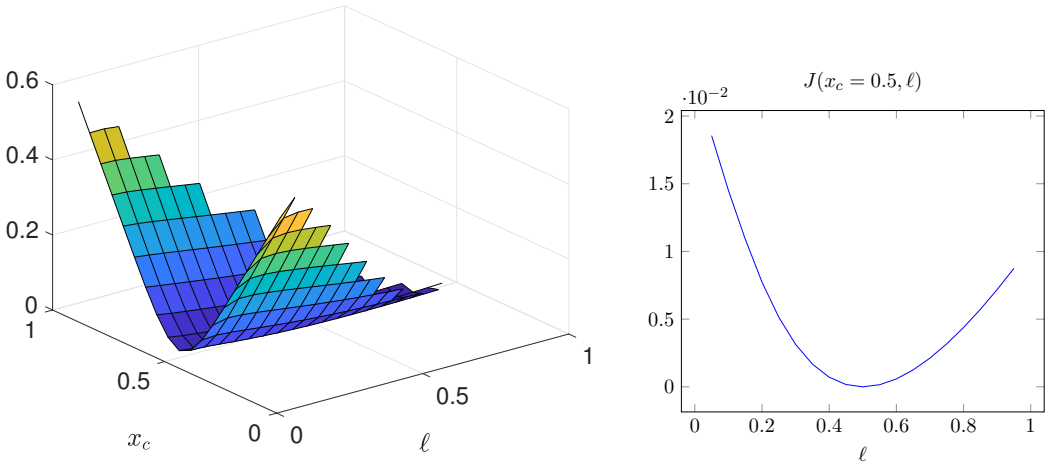


Figure 5.16 – Cost functional (5.14) for the isentropic transonic case, $\mathbf{a} = (x_c, \ell)$.

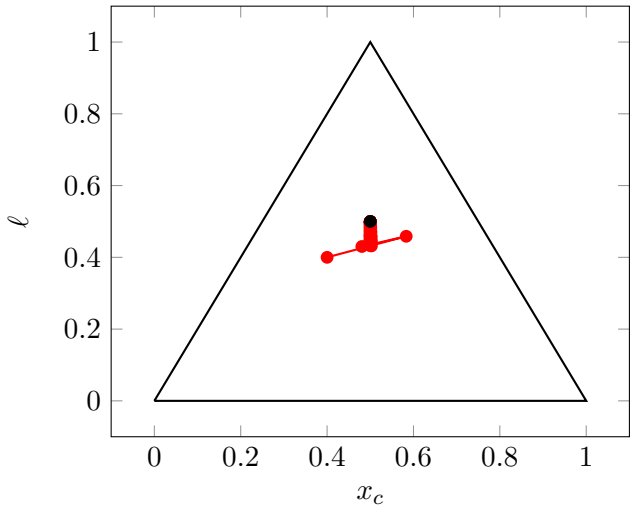


Figure 5.17 – Optimization algorithm steps for the isentropic transonic case, $\mathbf{a} = (x_c, \ell)$.

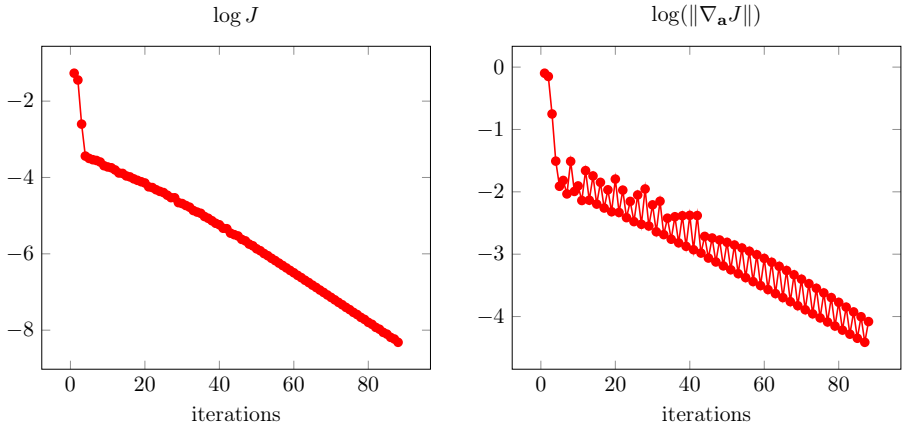


Figure 5.18 – Value of the cost functional and norm of its gradient with respect to iterations in a semilogarithmic scale - isentropic transonic case, $\mathbf{a} = (x_c, \ell)$.

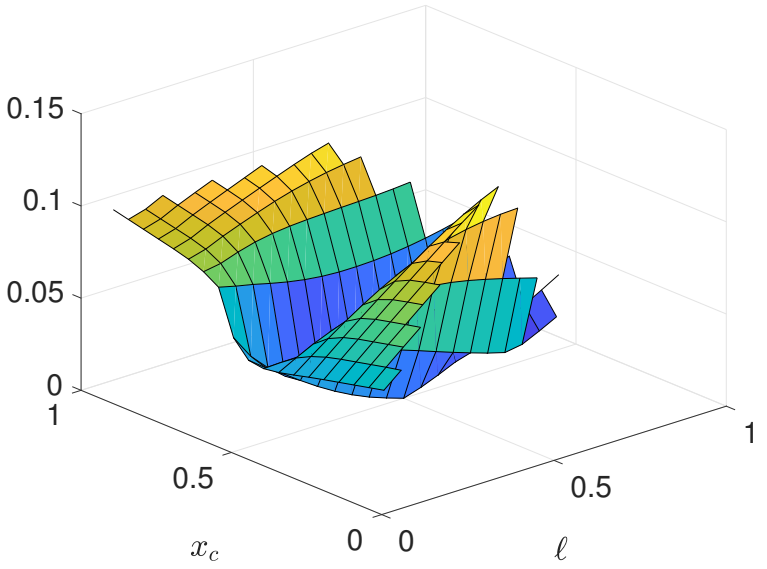


Figure 5.19 – Cost functional (5.14) for the shocked flow case, $\mathbf{a} = (x_c, \ell)$.

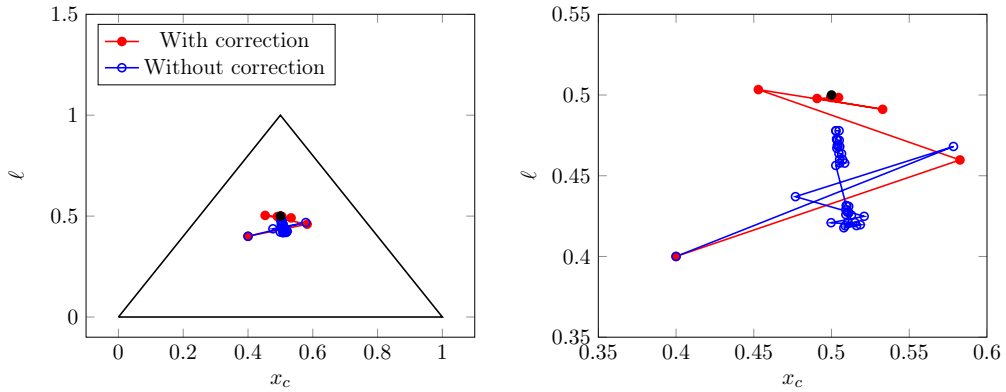


Figure 5.20 – On the left: optimization algorithm steps for the shocked flow case. On the right: emphasising the behaviour near optimality. $\mathbf{a} = (x_c, \ell)$.

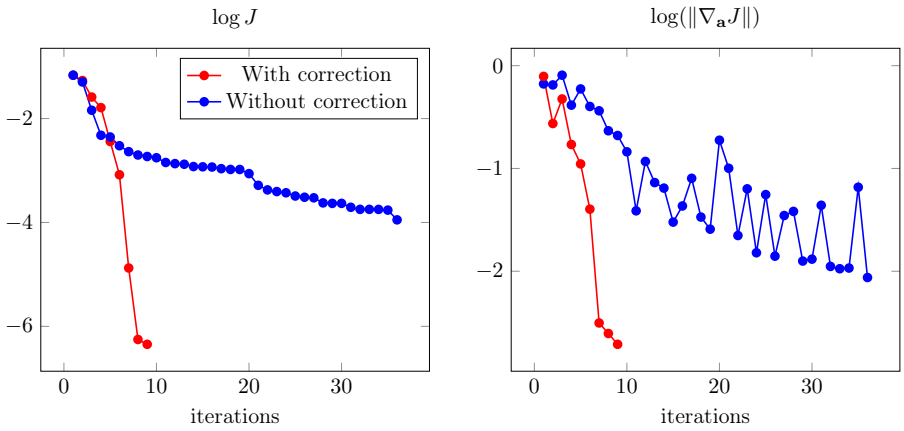


Figure 5.21 – Value of the cost functional and norm of its gradient with respect to iterations in a semilogarithmic scale - shocked flow case, $\mathbf{a} = (A, \ell)$.

6 | Conclusion and perspectives

The first goal of this thesis was to adapt some sensitivity techniques, and in particular the continuous sensitivity equation method, to the hyperbolic framework in case of discontinuous solutions. The first step was to define a sensitivity system valid also in the case of discontinuous state: this was achieved by adding a correction term to the sensitivity equations. The definition of this correction term was given for a general hyperbolic system and, in Chapter 5, we showed how this definition is valid also in the case of non-conservative systems. The scalar case of Chapter 2 allowed us to understand more in depth the Dirac delta functions exhibited by the sensitivities in the absence of the correction term, since we were able to compute analytically their coefficient.

The sensitivity system obtained with the correction term provided us with solutions without spikes approximating the Dirac delta function. This was an objective of this thesis, because Dirac delta function cannot be seized numerically and this corrupts the solution in the neighbourhood of the shocks and makes convergence impossible. Moreover, the spikes change with the numerical discretisation.

The second aim of this thesis was to design proper numerical schemes in order to discretise and approximate the solution of the sensitivity system obtained. This did not present any particular issue in the scalar case of Chapter 2, however the numerical results presented in Chapters 3-4-5 showed that, in the case of systems, the numerical diffusion plays a very important role in the discretisation of the sensitivity, to such an extent that classical finite volume schemes do not converge to the analytical solution: in particular, the value of the plateau in the intermediate zones is not correct.

To overcome this problem, we proposed some anti-diffusive numerical schemes, based on sampling techniques and inspired by Glimm random choice method: with these schemes we were able to discretise more precisely the source term and to obtain a correct solution for the sensitivity.

Two applications were tackled: an uncertainty quantification and an optimization problem. Both of them underlined the importance of the correction term: on one hand, for the uncertainty quantification the results obtained without the correction term provide non-physical estimates for some of the variables, without adding any useful information on the shock displacement; on the other hand, the optimization algorithm converges more quickly and with a smoother trajectory if the corrected sensitivity are employed to compute the gradient of the cost functional. Concerning the anti-diffusive schemes, they were only tested on the uncertainty quantification problem and they did not provide better results, in spite of their higher computational cost. This is a very interesting point because anti-diffusive schemes such as Glimm method are not extendible in 2D: comparing the results obtained with diffusive and anti-diffusive schemes, we can say that

the firsts are suitable for more complex applications, even though they do not provide a solution for the sensitivity as precise as one would expect with respect to the mesh.

This work leads to some new interesting problems. Firstly, it would be interesting to investigate higher order sensitivities: the first order sensitivity exhibits more discontinuities than the state, therefore the correction term for a second order sensitivity would be even more complex. Moreover, the definition of such a correction term would add constraints of the numerical schemes used to solve the first order sensitivity. Secondly, to tackle more realistic fluid dynamics application, one should extend the methods here developed in higher spatial dimension. From a theoretical point of view, it should be straightforward. The main practical difficulties that we can foresee are the definition of a good shock detector and, if needed, the 2D extension of the anti-diffusive numerical schemes. To do this, one would need a parametric description of the shock and then some shock fitting techniques [PB09] could be used in this framework. Finally, one could perform SA on conservation laws presenting non-classical shocks [LeF02] and study the sensitivity of the solutions with respect to the underlying kinetic function, which allows to select the admissible solutions.

It follows a more detailed conclusion, chapter by chapter.

Chapter 2. In this chapter we considered a scalar nonlinear hyperbolic PDE. Starting from the scalar case allowed us to introduce some of the main problems in a simpler framework. In the first part of the chapter, the analytical solution for the state equation was computed using the method of characteristics, which provided us with an implicit solution, unless the initial data was particularly simple, like in the case of a Riemann problem. Secondly, the sensitivity system was defined and its analytical solution was obtained starting from the state solution: interestingly, for the sensitivity we were able to give an explicit solution, although depending on the state, for all initial data. Then, the correction term was defined and the hyperbolicity of the global system was briefly analysed. We remarked that the solution to the sensitivity equation without the correction term presented a Dirac delta function where the state is discontinuous: in this simple scalar case, we were able to compute analytically the coefficient associated to the Dirac and one could notice that the spikes were larger if the dependence of the position of the shock on the parameter of interest was stronger. Finally, a numerical scheme was presented, along with the results obtained for the inviscid Burgers' equation with different initial data: a Riemann problem and a continuous initial conditions. We observed that, for a scalar equation, numerical diffusion does not play a fundamental role, because the problem is too simple: for instance, for the Riemann problem there is no intermediate state to be computed. On the other hand, the test case of a continuous initial data was complicated enough to tackle the problem of the definition of a shock detector: in this case it is way more difficult to numerically detect a discontinuity than in the case of a piecewise constant function (which is the solution to a scalar Riemann problem).

Chapter 3. In this chapter, which is an adaptation of [CDF17a], we dealt with the Euler equations in barotropic conditions and in Lagrangian coordinates. It is a system of two equations with two unknowns. It describes the dynamics of a compressible material in barotropic conditions (i.e. the density is a function only of the pressure), using two physical variables: the co-volume τ, which is the reciprocal of the density, and the Lagrangian speed. The choice of the Lagrangian coordinates was motivated by the fact that in this way the sign of the eigenvalues of the system is known, which simplified the

design of numerical schemes. In the first sections, we presented the state equations and derived the sensitivity equations in the regular case. Next, a section was devoted to the definition of the correction term in order to have a sensitivity system which was valid also in the case of discontinuous state. The correction term was defined by integrating the sensitivity equation over a control volume and then using the Rankine-Hugoniot conditions, which govern the state across a shock. We then detailed the exact resolution of the Riemann problem for the state and for the corrected sensitivity system: the state presents a classical structure, composed by two waves which can be either shocks or rarefaction waves; contrarily, the sensitivity has a quite particular structure, with two discontinuities associated with the beginning and the end of the rarefaction wave and a constant plateau inside the rarefaction itself. Then, some classical numerical methods for the state and their adaptation for the sensitivity were examined. In particular an exact Godunov scheme was implemented for both state and sensitivity, and a Roe approximate Riemann solver was implemented for the state and an adaptation of it for the sensitivity. We extended the Roe-type schemes to the second order, with a two-step Runge-Kutta method in time and a MUSCL-type approach in space. However, the second order extension in space is not classical: a precaution was necessary in order to be able to discretise the source term in the second order framework. Some numerical tests were conducted, which exhibited mesh-convergence issues, regardless of the order of the scheme. In particular, the test case of an isolated shock led us to conclude that the convergence problem was due to numerical diffusion. For this reason, an anti-diffusive version of the Roe-type schemes mentioned above was introduced, inspired to Glimm's random choice method, both for the first and second order. This allowed to overcome the mesh-convergence problems, as shown in some numerical results in the last section of the chapter.

Chapter 4. In this chapter we studied the complete Euler system. It is a system of three PDEs with four unknowns, closed by an algebraic equation. One can choose different sets of variables to describe the flow. A common choice is, for instance, to work with the conserved variables: the density of the fluid ρ, the momentum ρu and the total energy per unit mass ρE. In this work, we presented all the results with the physical, or primitive, variables: the density ρ, the speed u and the pressure p. In the first part of the chapter the state system was described and the sensitivity system was derived in the regular case. It followed a detailed analysis on the hyperbolicity of the global system of state and sensitivity as a whole and we proved that, in the general case, the system is only weakly hyperbolic. Then, the case of discontinuous state was considered and the correction term for the sensitivity equations was computed, in a similar way to the previous chapter. The solution of a specific Riemann problem, known as the Sod shock tube, was detailed for the state. The choice of this case was motivated by the presence of three different waves: a rarefaction, a contact discontinuity and a shock. Starting from the analytical solution for the state, the analytical sensitivity was derived: as we observed in the previous chapter, the sensitivity exhibits two discontinuities, one at the beginning and the other at the end of the rarefaction wave; however in this case there is a proper rarefaction wave for the sensitivity, too. In the following part of the Chapter, some numerical schemes were presented: we illustrated some constraints regarding the schemes that could be used for the state. In particular, what we categorised as HLL-type Riemann solvers (i.e. with only one intermediate star state) are not suitable for

our problem: if used, they would make impossible the definition and the computation of the correction term across the contact discontinuity. For this reason, the approximate Riemann solver of Roe was used. Concerning the schemes for the sensitivity, first we introduced a simpler HLL scheme, we then illustrated two approaches to obtain an HLLC-type scheme (i.e. with two different intermediate zones) and proved that they both lead to the same scheme. Diffusive and anti-diffusive versions of all of the schemes mentioned were implemented at first and second order, as done in the previous chapter, and some mesh-convergence tests were performed. The results obtained were consistent with the ones obtained in the previous chapter: the numerical diffusion affects the plateau values and this corrupts the convergence. Concerning the sensitivity, the computationally most expensive HLLC-type scheme did not provide a more accurate solution.

In the final part of the Chapter, the first application was tackled. A UQ analysis was performed on the Riemann problem: confidence intervals were calculated for all the physical variables, using the sensitivities. We computed the sensitivities in three different ways: with and without the correction term, and with the diffusive and anti-diffusive numerical schemes. We showed the confidence intervals obtained with the well-known Monte Carlo method and with the SA methods for each physical variable. In the regular zones the two approaches gave the same results, with a significantly lower computational cost of the SA method. In the neighbourhood of the discontinuity there was a loss of precision of the SA methods, due to the fact that the influence of the uncertain parameters on the position of the shock was neglected. However, the corrected sensitivities gave more reasonable results: the sensitivities obtained without the correction term, due to the presence of the spikes, provided non-physical confidence intervals for the density and the pressure. An important observation is that the anti-diffusive schemes did not provide a significantly different solution: this is interesting in the perspective of more realistic, two-dimensional applications because the random choice methods, on which our anti-diffusive methods are based, are not easily extendible in 2D.

Chapter 5. In this chapter, we investigated the quasi 1D Euler system: this is an extension of the Euler system considered in the previous chapter, but it models a flow happening in a channel whose section is not constant. To describe that, a smooth function of the space, $h(x;\mathbf{a})$, depending on some parameters $\mathbf{a}$, was introduced and a term depending on $h(x;\mathbf{a})$ and on its derivative in space was added to the right-hand side of the Euler system. Once again, the sensitivity system was defined at first in the regular case and then a correction term was added to account for possible discontinuities in the state. Since the state system considered in this chapter was not conservative, the computation of the correction term was slightly different, because it could not rely on the Rankine-Hugoniot conditions. Nonetheless the same conclusion as in the previous cases was reached. As the conservative part of the system coincided with the Euler system of the previous chapter, the same numerical schemes were used. However, in light of the results obtained, in particular the UQ ones, we chose to focus our attention on the first order diffusive Roe scheme. A brief description of the discretisation of the non-conservative part was done. This system admits stationary solutions. We considered in particular two test cases: an isentropic transonic case and a case presenting a shock. The boundary conditions for these cases were discussed in details. We compared the numerical results obtained for the state to the analytical solution given in [GP01] for the Mach number. Concerning the sensitivity, we do not have an analytical solution for this

case, therefore we compared the numerical sensitivities obtained with our schemes with the empirical sensitivities. The empirical sensitivities were computed by finite differences, starting from two different realisations of the state for two close enough values of the parameter of interest.

In the final part of the Chapter, a second application was tackled: we defined an optimization problem. The optimization method used was a projected gradient algorithm which used sensitivities to compute the gradient of the cost functional. The problem considered is an inverse problem, and in particular a pressure matching one: we set a target pressure p^*, which is a pressure obtained with a certain vector of parameters $\mathbf{a}^*$, and we aimed at recovering this value with the optimization algorithm. We performed the same test in the isentropic transonic case and in the case with shock, for two different choices of optimization parameters. In the case with shock, we compared the results obtained if the sensitivities, and therefore the gradient, were computed with or without the correction term. The results showed that the correction term speeds up the convergence of the optimization algorithm, and therefore confirmed the importance of the correction.

Conclusion et perspectives

Le premier objectif de cette thèse était d'adapter certaines techniques d'AS, et en particulier la méthode de l'équation de sensibilité continue, au cadre hyperbolique en cas de solutions discontinues. La première étape consistait à définir un système de sensibilité valable également en cas d'état discontinu : ceci a été réalisé en ajoutant un terme de correction aux équations de sensibilité. La définition de ce terme de correction a été donnée pour un système hyperbolique général et, dans le chapitre 5, nous avons montré que cette définition est valable aussi dans le cas de systèmes non conservatifs. Le cas scalaire du chapitre 2 nous a permis de comprendre plus en détail les Dirac présents dans les sensibilités en l'absence du terme de correction, puisque nous avons pu calculer analytiquement leur amplitude.

Le système de sensibilité obtenu avec le terme de correction nous a fourni des solutions sans pics approchant les Dirac. Ceci était un objectif de cette thèse, car une distribution de Dirac ne peut pas être évaluée numériquement et cela dégrade la solution au voisinage des chocs et rend la convergence vers la solution analytique impossible. De plus, les pics changent avec la discrétisation numérique.

Le deuxième objectif de cette thèse était de définir des schémas numériques appropriés afin de discrétiser et d'approximer la solution du système de sensibilité obtenu. Cela n'a pas posé de problèmes particuliers dans le cas scalaire traité dans le chapitre 2, cependant les résultats numériques présentés dans les chapitres 3-4-5 ont montré que, dans le cas des systèmes, la diffusion numérique joue un rôle très important dans la discrétisation de la sensibilité, à tel point que les schémas volumes finis classiques ne convergent pas vers la solution analytique : en particulier, la valeur du plateau dans les zones intermédiaires n'est pas correcte.

Pour surmonter ce problème, nous avons proposé des schémas numériques anti-diffusifs, basés sur des techniques d'échantillonnage et inspirés de la méthode de choix aléatoire de Glimm : avec ces schémas, nous avons pu discrétiser plus précisément le terme source et obtenir la bonne solution pour la sensibilité.

Nous nous sommes attaqués à deux applications : un problème de quantification d'incertitude et un d'optimisation. Les deux ont souligné l'importance du terme de correction : d'une part, pour la quantification d'incertitude, les résultats obtenus sans le terme de correction fournissent des estimations non physiques pour certaines variables, sans cependant ajouter d'informations utiles sur le déplacement du choc ; d'autre part, l'algorithme d'optimisation converge plus rapidement et avec une trajectoire plus lisse si la sensibilité corrigée est utilisée pour calculer le gradient de la fonctionnel coût. Concernant les schémas anti-diffusifs, nous ne les avons testés que sur le problème de quantification d'incertitude et ils n'ont pas donné de meilleurs résultats, malgré un coût de calcul plus

élevé. C'est un point très intéressant car les schémas anti-diffusifs comme la méthode de Glimm ne sont pas facilement extensibles en deux dimensions : en comparant les résultats obtenus avec des schémas diffusifs et anti-diffusifs, on peut dire que les premiers sont adaptés à des applications plus complexes, même si ils ne fournissent pas une solution de sensibilité au niveau de précision que l'on s'attend par rapport au maillage.

Ce travail mène à des nouveaux problèmes intéressants. D'abord, il serait intéressant d'étudier des sensibilités d'ordre supérieur : la sensibilité de premier ordre présente plus de discontinuités que l'état, donc le terme de correction pour une sensibilité de second ordre serait encore plus complexe. De plus, la définition d'un tel terme de correction ajouterait des contraintes aux schémas numériques utilisés pour résoudre la sensibilité de premier ordre. Deuxièmement, pour aborder une application plus réaliste de la dynamique des fluides, on devrait étendre les méthodes développées ici en dimension spatiale supérieure. D'un point de vue théorique, cela devrait être simple. Les principales difficultés pratiques que nous pouvons prévoir sont la définition d'un bon détecteur de choc et, si nécessaire, l'extension 2D des schémas numériques anti-diffusifs. Pour cela, il faudrait avoir une description paramétrique du choc et les techniques de *shock fitting* [PB09] pourraient être adaptées à ce contexte. Enfin, on pourrait appliquer l'AS à des lois de conservation présentant des chocs non classiques [LeF02] et étudier ainsi la sensibilité des solutions par rapport à la fonction cinétique sous-jacente qui permet de sélectionner les solutions admissibles.

Ci-dessous une conclusion plus détaillée, chapitre par chapitre.

Chapitre 2. Dans ce chapitre, nous avons considéré une EDP hyperbolique non linéaire scalaire. Partir du cas scalaire nous a permis d'affronter certains des problèmes principaux dans un cadre plus simple. Dans la première partie du chapitre, la solution analytique pour l'équation d'état a été calculée en utilisant la méthode des caractéristiques, qui nous a fourni une solution implicite, sauf dans des cas de donnée initiale particulièrement simple, comme un problème de Riemann. Deuxièmement, le système de sensibilité a été défini et sa solution analytique a été obtenue à partir de la solution d'état : il est intéressant de noter que, pour la sensibilité, nous avons pu donner une solution explicite, en fonction de l'état, pour toute donnée initiale. Ensuite, le terme de correction a été défini et l'hyperbolicité du système global a été brièvement analysée. Nous avons observé que la solution de l'équation de sensibilité sans le terme de correction présentait un Dirac là où l'état était discontinu. Dans ce cas scalaire simple, nous avons pu calculer analytiquement le coefficient associé au Dirac et nous avons constaté que les pics étaient plus petits si l'influence du paramètre d'intérêt sur la position du choc était faible. Enfin, un schéma numérique a été présenté, ainsi que les résultats obtenus pour l'équation de Burgers non visqueux avec différentes données initiales : un problème de Riemann et un problème avec conditions initiales continues. Nous avons observé que, pour une équation scalaire, la diffusion numérique ne joue pas un rôle fondamental, car le problème est trop simple : par exemple, pour le problème de Riemann, il n'y a pas d'état intermédiaire à calculer. D'autre part, le cas d'une donnée initiale continue était suffisamment compliqué pour aborder le problème de la définition d'un détecteur de choc : dans ce cas, il est beaucoup plus difficile de détecter numériquement une discontinuité que dans le cas d'une fonction constante par morceaux, qui est la solution du problème scalaire de Riemann.

Chapitre 3. Dans ce chapitre, qui est une adaptation de [CDF17a], nous avons traité

les équations d'Euler dans des conditions barotropes et en coordonnées lagrangiennes. Il s'agit d'un système de deux équations à deux inconnues. Il décrit la dynamique d'un matériau compressible dans des conditions barotropes (c'est-à-dire la densité est une fonction seulement de la pression), utilisant deux variables physiques : le co-volume τ, qui est l'inverse de la densité, et la vitesse lagrangienne. Le choix des coordonnées lagrangiennes a été motivé par le fait que l'on connaît le signe des valeurs propres du système, ce qui a simplifié la conception des schémas numériques. Dans les premières sections, nous avons présenté les équations d'état et nous avons dérivé les équations de sensibilité dans le cas régulier. Ensuite, une section a été consacrée à la définition du terme de correction, pour avoir un système de sensibilité valable également dans le cas d'état discontinu. Le terme de correction a été défini en intégrant l'équation de sensibilité sur un volume de contrôle, puis en utilisant les conditions de Rankine-Hugoniot, qui gouvernent l'état à travers un choc. Nous avons ensuite détaillé la résolution analytique du problème de Riemann pour l'état et pour le système de sensibilité corrigé : l'état présente une structure classique, composée par deux ondes qui peuvent être soit des chocs, soit des ondes de détente ; au contraire, la sensibilité a une structure assez particulière, avec deux discontinuités associées au début et à la fin de l'onde de détente et un plateau constant à l'intérieur de la détente elle-même. Ensuite, quelques méthodes numériques classiques pour l'état et leur adaptation pour la sensibilité ont été examinées. En particulier un schéma de Godunov exact a été implémenté pour l'état et la sensibilité, et un solveur de Riemann approché, le solveur de Roe, a été implémenté pour l'état et une adaptation de celui-ci pour la sensibilité. Nous avons étendu les schémas de type Roe au second ordre, avec une méthode Runge-Kutta en temps et une approche de type MUSCL en espace. Cependant, l'extension au second ordre en espace n'est pas classique : une précaution était nécessaire pour pouvoir discrétiser le terme source au second ordre. Des tests numériques ont été faits, qui présentaient des problèmes de convergence en maillage, indépendamment de l'ordre du schéma. En particulier, le cas test d'un choc isolé nous a mené à conclure que le problème de convergence était dû à la diffusion numérique. Pour cette raison, une version anti-diffusive des schémas de type Roe mentionnés ci-dessus a été introduite, inspirée de la méthode de choix aléatoire de Glimm, à la fois pour le premier et le second ordre. Cela a permis de surmonter les problèmes de convergence en maillage, comme le montrent certains résultats numériques dans la dernière section du chapitre.

Chapitre 4. Dans ce chapitre, nous avons étudié le système d'Euler complet. C'est un système de trois EDP à quatre inconnues, fermé par une équation algébrique. On peut choisir différentes variables pour décrire le flux. Un choix commun est, par exemple, de travailler avec les variables conservatives : la densité du fluide ρ, la quantité de mouvement ρu et l'énergie totale par unité de volume ρE. Dans ce travail, nous avons présenté tous les résultats avec les variables physiques, ou primitives : la densité ρ, la vitesse u et la pression p. Dans la première partie du chapitre, nous avons présenté le système d'état et nous avons dérivé le système de sensibilité dans le cas régulier. Il a suivi une analyse détaillée sur l'hyperbolicité du système global formé par l'état et la sensibilité et nous avons prouvé que, dans le cas général, le système n'est que faiblement hyperbolique. Ensuite, nous avons considéré le cas d'état discontinu et nous avons calculé le terme de correction pour les équations de sensibilité, de la même façon que dans le chapitre précédent. La solution d'un problème de Riemann spécifique, connu sous le nom de tube à choc de Sod, a été détaillée pour l'état. Le choix de ce cas est motivé par la présence

de trois ondes différentes : une détente, une discontinuité de contact et un choc. À partir de la solution analytique de l'état, nous avons dérivé celle de la sensibilité : comme nous l'avons observé dans le chapitre précédent, la sensibilité présente deux discontinuités, une au début et une à la fin de l'onde de détente ; cependant, dans ce cas, il y a une vraie onde de détente aussi pour la sensibilité. Dans la partie suivante du chapitre, des schémas numériques ont été présentés : nous avons illustré certaines contraintes concernant les schémas qui peuvent être utilisés pour l'état. En particulier, ce que nous avons classé comme solveurs de Riemann de type HLL (c'est-à-dire avec un seul état étoile intermédiaire) ne conviennent pas à notre problème : si utilisés, ils rendraient impossible la définition et le calcul du terme de correction à travers la discontinuité de contact. Pour cette raison, le solveur de Riemann approché de Roe a été utilisé. Concernant les schémas de sensibilité, nous avons d'abord introduit un schéma HLL classique, puis nous avons illustré deux approches pour obtenir un schéma de type HLLC (c'est-à-dire avec deux zones intermédiaires différentes) et montré qu'elles conduisaient au même schéma. Des versions diffusives et anti-diffusives de tous les schémas mentionnés ont été implémentées au premier et au second ordre, comme dans le chapitre précédent, et des tests de convergence en maillage ont été également réalisés. Les résultats obtenus sont en accord avec ceux obtenus dans le chapitre précédent : la diffusion numérique affecte les valeurs du plateau et cela dégrade la convergence. En ce qui concerne la sensibilité, le schéma de type HLLC, plus coûteux en termes de calcul, n'a pas fourni une solution plus précise.

Dans la dernière partie du chapitre, nous nous sommes attaqués à la première application. Une analyse de quantification d'incertitude a été effectuée sur le problème de Riemann : nous avons calculé les intervalles de confiance pour toutes les variables physiques, en utilisant les sensibilités. Nous avons calculé les sensibilités de trois manières différentes : avec et sans le terme de correction, et avec les schémas numériques diffusifs et anti-diffusifs. Nous avons montré les intervalles de confiance obtenus avec la méthode Monte Carlo et avec les méthodes d'AS pour chaque variable physique. Dans les zones régulières, les deux approches ont donné les mêmes résultats, avec un coût de calcul significativement plus bas pour la méthode d'AS. Au voisinage de la discontinuité, il y avait une perte de précision des méthodes d'AS, due au fait que l'influence des paramètres incertains sur la position du choc a été négligée. Cependant, les sensibilités corrigées ont donné des résultats plus raisonnables : les sensibilités obtenues sans le terme de correction, en raison des pics, ont fourni des intervalles de confiance non physiques pour la densité et la pression. Une observation importante est que les schémas anti-diffusifs ne donnent pas une solution significativement différente : ceci est intéressant dans la perspective d'applications bidimensionnelles plus réalistes car les méthodes de choix aléatoires, sur lesquelles sont basées nos méthodes anti-diffusives, ne sont pas facilement extensibles en 2D.

Chapitre 5. Dans ce chapitre, nous avons étudié le système d'Euler quasi-1D : il s'agit d'une extension du système d'Euler considéré dans le chapitre précédent, mais il modélise un écoulement dans un canal dont la section n'est pas constante. Pour décrire cela, une fonction lisse de l'espace, $h(x; \mathbf{a})$, dépendante d'un vecteur de paramètres $\mathbf{a}$, a été introduite et un terme dépendant de $h(x; \mathbf{a})$ et de sa dérivée en espace ont été ajoutés à la partie droite du système d'Euler. Encore une fois, le système de sensibilité a été défini d'abord dans le cas régulier, puis un terme de correction a été ajouté pour tenir compte

des éventuelles discontinuités dans l'état. Comme le système d'état considéré dans ce chapitre n'était pas conservatif, le calcul du terme de correction était légèrement différent, car il ne pouvait pas s'appuyer sur les conditions de Rankine-Hugoniot. Néanmoins, la même conclusion que dans les cas précédents a été obtenue. Comme la partie conservative du système coïncidait avec le système d'Euler du chapitre précédent, les mêmes schémas numériques ont été utilisés. Cependant, vus les résultats obtenus, en particulier ceux de la quantification d'incertitude, nous avons choisi de concentrer notre attention sur le schéma de Roe diffusif au premier ordre. Une brève description de la discrétisation de la partie non conservative a été faite. Ce système admet des solutions stationnaires. Nous avons considéré en particulier deux cas tests : un cas isentropique transsonique et un cas transsonique avec un choc. Les conditions aux limites de ces deux cas ont été discutées en détail. Nous avons comparé les résultats numériques obtenus pour l'état à la solution analytique donnée dans [GP01] pour le nombre de Mach. Concernant la sensibilité, nous n'avons pas de solution analytique pour ce cas, donc nous avons comparé les sensibilités numériques obtenues avec nos schémas avec les sensibilités empiriques. Les sensibilités empiriques ont été calculées par des différences finies, à partir de deux réalisations différentes de l'état pour deux valeurs proches du paramètre d'intérêt.

Dans la dernière partie du chapitre, une deuxième application a été abordée : nous avons défini un problème d'optimisation. La méthode d'optimisation utilisée est un algorithme de gradient projeté qui utilise des sensibilités pour calculer le gradient de la fonctionnelle coût. Le problème considéré est un problème inverse pour la pression : nous établissons une pression cible p^*, qui est une pression obtenue avec un certain vecteur de paramètres $\mathbf{a}^*$, et nous visons à récupérer cette valeur avec l'algorithme d'optimisation. Nous avons effectué le même test dans le cas isentropique transsonique et dans le cas transsonique avec choc, pour deux choix différents de paramètres d'optimisation. Dans le cas du choc, nous avons comparé les résultats obtenus si les sensibilités, et donc le gradient, étaient calculés avec ou sans le terme de correction. Les résultats ont montré que le terme de correction accélère la convergence de l'algorithme d'optimisation, confirmant ainsi l'importance de la correction.

Bibliography

[AC12] R. Abgrall and P. M. Congedo. A semi-intrusive deterministic approach to uncertainty quantification in non-linear fluid flow problems. *J. Comput. Physics*, 2012.

[App97] J. R. Appel. *Sensitivity calculations for conservation laws with application to discontinuous fluid flows*. PhD thesis, PhD thesis, Virginia Polytechnic Institute and State University, 1997.

[BB97] J. Borggaard and J. Burns. A PDE sensitivity equation method for optimal aerodynamic design. *Journal of Computational Physics*, 136(2):366 – 384, 1997.

[BJL03] F. Bouchut, S. Jin, and X. Li. Numerical approximations of pressureless and isothermal gas dynamics. *SIAM Journal on Numerical Analysis*, 41(1):135–158, 2003.

[Bou04] F. Bouchut. *Nonlinear stability of finite Volume Methods for hyperbolic conservation laws: And Well-Balanced schemes for sources*. Springer Science & Business Media, 2004.

[BP02] C. Bardos and O. Pironneau. A formalism for the differentiation of conservation laws. *Compte rendu de l'Académie des Sciences*, 335(10):839–845, 2002.

[CC99] P. R. Christian and G. Casella. Monte Carlo statistical methods, 1999.

[CDF17a] C. Chalons, R. Duvigneau, and C. Fiorini. Sensitivity analysis and numerical diffusion effects for hyperbolic pde systems with discontinuous solutions. the case of barotropic euler equations in lagrangian coordinates. *SIAM Journal on Scientific Computing*, 2017. In review.

[CDF17b] C. Chalons, R. Duvigneau, and C. Fiorini. Sensitivity analysis for the Euler equations in Lagrangian coordinates. In *International Conference on Finite Volumes for Complex Applications*, pages 71–79. Springer, 2017.

[CG08] C. Chalons and P. Goatin. Godunov scheme and sampling technique for computing phase transitions in traffic flow modeling. *Interfaces and Free Boundaries*, 10(2):197–221, 2008.

[Cho76] A. J. Chorin. Random choice solution of hyperbolic systems. *Journal of Computational Physics*, 22(4):517–533, 1976.

[CKM12] C. Chalons, D. Kah, and M. Massot. Beyond pressureless gas dynamics: quadrature-based moment models. *Communications in Mathematical Sciences*, 10(4):1241–1272, 2012.

[Del14] C. Delenne. Propagation de la sensibilité dans les modèles hydrodynamiques., 2014.

[DP06] R. Duvigneau and D. Pelletier. A sensitivity equation method for fast evaluation of nearby flows and uncertainty analysis for shape parameters. *Int. J. of CFD*, 20(7):497–512, August 2006.

[DPB06] R. Duvigneau, D. Pelletier, and J. Borggaard. An improved continuous sensitivity equation method for optimal shape design in mixed convection. *Numerical Heat Transfer part B : Fundamentals*, 50(1):1–24, July 2006.

[Fio17] C. Fiorini. Optimization of running strategies according to the physiological parameters for a two-runners model. *Bulletin of mathematical biology*, 79(1):143–162, 2017.

[FLF92] A. Forestier and P. Le Floch. Multivalued solutions to some non-linear and non-strictly hyperbolic systems. *Japan journal of industrial and applied mathematics*, 9(1):1, 1992.

[GDC09] V. Guinot, C. Delenne, and B. Cappelaere. An approximate riemann solver for sensitivity equations with discontinuous solutions. *Advances in Water Resources*, 32(1):61–77, 2009.

[GLC07] V. Guinot, M. Leménager, and B. Cappelaere. Sensitivity equations for hyperbolic conservation law-based flow models. *Advances in water resources*, 30(9):1943–1961, 2007.

[Gli65] J. Glimm. Solutions in the large for nonlinear hyperbolic systems of equations. *Communications on pure and applied mathematics*, 18(4):697–715, 1965.

[GP01] M. B. Giles and N. A. Pierce. Analytic adjoint solutions for the quasi-one-dimensional Euler equations. *Journal of Fluid Mechanics*, 426:327–345, 2001.

[GR96] E. Godlewski and P. A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*. Springer Science & Business Media, 1996.

[Gui09] V. Guinot. Upwind finite volume solution of sensitivity equations for hyperbolic systems of conservation laws with discontinuous solutions. *Computers & Fluids*, 38(9):1697–1709, 2009.

[Gun03] M. D. Gunzburger. *Perspectives in flow control and optimization*, volume 5. Siam, 2003.

[HEPB04] H. Hristova, S. Etienne, D. Pelletier, and J. Borggaard. A continuous sensitivity equation method for time-dependent incompressible laminar flows. *Int. J. for Numerical Methods in Fluids*, 50:817–844, 2004.

[HH83] A. Harten and J. M. Hyman. Self adjusting grid methods for one-dimensional hyperbolic conservation laws. *Journal of computational Physics*, 50(2):235–269, 1983.

[HLL83] A. Harten, P. D. Lax, and B. van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM review*, 25(1):35–61, 1983.

[HLVL97] A. Harten, P. D. Lax, and B. Van Leer. On upstream differencing and godunov-type schemes for hyperbolic conservation laws. In *Upwind and High-Resolution Schemes*, pages 53–79. Springer, 1997.

[HMB98] P. Hovland, B. Mohammadi, and C. Bischof. Automatic differentiation and Navier-Stokes computations. In *Computational Methods for Optimal Design and Control*, pages 265–284. Springer, 1998.

[Jam88] A. Jameson. Aerodynamic design via control theory. *Journal of Scientific Computing*, 3(97-0101):233–260, 1988.

[Jos93] K. T. Joseph. A riemann problem whose viscosity solutions contain δ-measures. *Asymptotic Analysis*, 7(2):105–120, 1993.

[KM06] O.M. Knio and O.P. Le Maitre. Uncertainty propagation in CFD using polynomial chaos decomposition. *Fluid Dynamics Research*, 38(9):616–640, September 2006.

[LeF90] P. G. LeFloch. An existence and uniqueness result for two nonstrictly hyperbolic systems, ima volumes in math. and its appl. 27. *Nonlinear evolution equations that change type*, pages 126–138, 1990.

[LeF02] P. G. LeFloch. *Hyperbolic Systems of Conservation Laws: The theory of classical and nonclassical shock waves*. Springer Science & Business Media, 2002.

[MD10] M. Martinelli and R. Duvigneau. On the use of second-order derivative and metamodel-based monte-carlo for uncertainty estimation in aerodynamics. *Computers and Fluids*, 37(6), 2010.

[MP01] B. Mohammadi and O. Pironneau. *Applied Optimal Shape Design for Fluids*. Oxford University Press, 2001.

[PB09] R. Paciorri and A. Bonfiglioli. A shock-fitting technique for 2D unstructured grids. *Computers & Fluids*, 38(3):715–726, 2009.

[Pir74] O. Pironneau. On optimum design in fluid mechanics. *J. Fluid Mechanics*, (64), 1974.

[PNTG01] M.M. Putko, P.A. Newman, A.C. Taylor, and L.L. Green. Approach for uncertainty propagation and robust design in cfd using sensitivity derivatives. In *15th AIAA Computational Fluid Dynamics Conference*, Anaheim, CA, June 2001.

[Roe81] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of computational physics*, 43(2):357–372, 1981.

[TAI11] H. Telib, E. Arian, and A. Iollo. The effect of shocks on second order sensitivities for the quasi-one-dimensional euler equations. *Journal of Computational Physics*, 230(23):8603–8618, 2011.

[Tor09] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer Science & Business Media, 2009.

[Tor13] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer Science & Business Media, 2013.

[TPB01] É. Turgeon, D. Pelletier, and J. Borggaard. Sensitivity and uncertainty analysis for variable property flows. In *39th AIAA Aerospace Sciences Meeting and Exhibit*, Reno, NV, Jan. 2001.

[Wal03] R. Walters. Towards stochastic fluid mechanics via polynomial chaos. In *41st AIAA Aerospace Sciences Meeting and Exhibit, Reno, USA*, 2003.

[WH02] R. W. Walters and L. Huyse. Uncertainty analysis for fluid mechanics with applications. Technical report, NATIONAL AERONAUTICS AND SPACE ADMINISTRATION HAMPTON VA LANGLEY RESEARCH CENTER, 2002.

[XK03] D.B. Xiu and G.E. Karniadakis. Modeling uncertainty in flow simulations via generalized polynomial chaos. *Journal of Computational Physics*, (187):137–167, 2003.

[YZ12] H. Yang and Y. Zhang. New developments of delta shock waves and its applications in systems of conservation laws. *Journal of Differential Equations*, 252(11):5951–5993, 2012.

Appendices

A | Modelling of running strategies

A.1 Abstract

In order to describe the velocity and the anaerobic energy of two runners competing against each other for middle-distance races, we present a mathematical model relying on an optimal control problem for a system of ordinary differential equations. The model is based on energy conservation and on Newton's second law: resistive forces, propulsive forces and variations in the maximal oxygen uptake are taken into account. The interaction between the runners provides a minimum for staying one meter behind one's competitor. We perform numerical simulations and show how a runner can win a race against someone stronger by taking advantage of staying behind, or how they can improve their personal record by running behind someone else. Our simulations show when it is the best time to overtake, depending on the difference between the athletes. Finally, we compare our numerical results with real data from the men's $1500m$ finals of different competitions.

A.2 Introduction

The running strategy to win an Olympic medal is quite complex. It relies on outstanding physiology, good preparation, psychological factors and the optimal way to compete with the others to beat them. Quite a few mathematical works starting with Keller's [10], have analysed running strategies for a single runner [1, 2, 12, 13, 22], but very few take into account the competition situation where the point is to beat the others [11, 16].

The point of view of Keller [10] is to write the equations governing the energy and the velocity of a single runner, starting from Newton's second law and energy conservation. He considers a simple problem, in which the athlete runs alone on a straight path of length D, and the aim is to minimise the time T when the runner reaches the final distance D. This model matches the final times of world records. However some hypotheses are not physiologically reasonable, and this leads to a non-realistic velocity profile. Some authors have tried to improve Keller's model: Woodside in [22] and Mathis in [12] introduce a correction by adding a fatigue term for long races. Ward-Smith [21] and Morton [13, 14, 15] follow a different approach. Morton has introduced a three component model to take into account the variations in the oxygen uptake ($\dot{V}O2$) but the full optimal control problem is not solved. Behncke [2] incorporates the hydraulic model of Morton to a biomechanical model that extends the ones of Keller and Ward-Smith: it is more detailed in terms of resistive forces and takes into account the reaction time of the athlete. Aftalion and Bonnans [1] improve the models of Keller [10], Behncke [2] and

Morton [13, 14, 15] and assume that maximal oxygen uptake is a function of the anaerobic energy of the athlete. The aim is to match experimental measurements, in particular, in [8, 9, 20], where Hanon et al. show how the oxygen uptake varies during races of 400, 800 and $1500m$. They solve the full numerical control problem using an optimal control solver Bocop.

As pointed out by Pitcher [16], in middle distance running, it is common practice to try to position oneself behind but within striking distance of the leader for most of the race and then overtake them near the finish line. Pitcher explains that the runner behind can take advantage of the slipstream of the runner in front and relies on analyses of Pugh [17] and Kyle [11]. We believe that it is a combination of slipstream and psychological factors which explain why it is better to stay behind, and the equations can incorporate all this. The weakness of Pitcher's paper is that she imposes a strategy for one of the runners and allows only the second runner to have a free strategy. Therefore, in this paper, based on the recent work of Aftalion-Bonnans [1], we extend the model of Pitcher [16] to include slipstream and psychological factors in a two runners race and we set a realistic optimal control problem with each runner having a free strategy.

A.2.1 Mathematical model for a single runner

The system of Keller couples together the velocity of the runner at instant t, $v(t)$, the energy of the runner at instant t, $e(t)$, and the propulsive force of the runner at instant t, $f(t)$. The first equation is Newton's second law: it involves the propulsive force and the friction. Here, τ is a constant coefficient which gathers together all the friction effects, supposed to be linear in v. The friction term can be modified to include air resistance [11] which adds a term in $-cv^2$ to the first equation.

The second equation is an energy balance incorporating the oxygen uptake, σ, considered constant in Keller's paper, while the second term is the work of the propulsive force f. Both equations are normalised with respect to the runner's mass:

$$\begin{cases} \dot{v}(t) = f(t) - \frac{v(t)}{\tau} & v(0) = 0,\ x(0) = 0,\ x(T) = D \\ \dot{e}(t) = \sigma - f(t)v(t) & e(0) = e^0. \end{cases} \tag{A.1}$$

We use the dot notation to indicate the derivative with respect to time, i.e. $\dot{v} = \frac{dv}{dt}$ and $\dot{e} = \frac{de}{dt}$. We have set $x(t)$ to be the position so that $\dot{x} = v$. The final time T is defined as the time to reach the distance D. Moreover, e^0 is the initial energy. It is necessary to add some physiological constraints to the system (A.1):

- the energy must be positive:

$$e(t) \geq 0 \quad \forall t \geq 0; \tag{A.2}$$

- the propulsive force has an upper bound which depends on the runner's physiology, and a positive lower bound due to the fact that they are moving forwards:

$$0 \leq f(t) \leq f_M \quad \forall t \geq 0. \tag{A.3}$$

Therefore, in this model, the athlete is identified by four parameters: e^0, the initial energy, τ, the friction coefficient, σ, the oxygen uptake, and f_M the maximal propulsive

force. The aim is to solve (A.1)-(A.2)-(A.3) in such a way that, given a distance D, the final time T is minimal. From a mathematical point of view, it is a problem of optimal control: the propulsive force f is the control variable and the time T is the cost functional to be minimised, which depends on f through the states variables v and e. Therefore, the problem can be written as follows:

$$\min_{f\in\mathcal{F}} T(f) \quad \text{s.t (A.1)-(A.2)}, \tag{A.4}$$

where $\mathcal{F}$ is the set of admissible controls:

$$\mathcal{F} = \{f :\ 0 \le f(t) \le f_M \quad \forall t \ge 0\}.$$

In his work, Keller [10] claims that his energy balance takes into account only the aerobic energy (i.e. energy provided by oxygen consumption). However, Aftalion and Bonnans [1] remark that what he encompasses in the balance is the accumulated oxygen deficit: $e^0 - e(t)$. Therefore, $e(t)$ in equation (A.1) is in fact the anaerobic energy (i.e. energy provided by glycogen and lactate). In order to reproduce the results of [7, 8, 9], the oxygen uptake σ introduced in [1] is piecewise defined: in the most part of the race, σ is constant equal to its maximal value $\sigma_{\max}$, but it is increasing at the beginning of the race, and decreasing at the end (see Figure A.1). In fact, σ depends on five parameters: the initial value at rest σ_r, the maximal value $\sigma_{\max}$, the final value σ_f and two parameters φ and e_{cr} which denote the transition point from one zone to another, $\varphi,\ e_{cr} \in (0,1)$:

$$\sigma(e;\sigma_{\max},\sigma_f,\sigma_r,\varphi,e_{cr}) = \begin{cases} \sigma_{\max}\frac{e}{e^0 e_{cr}} + \sigma_f\left(1-\frac{e}{e^0 e_{cr}}\right) & \text{if } e < e^0 e_{cr} \\ \sigma_{\max} & \text{if } e^0 e_{cr} \le e \le e^0\varphi \\ \sigma_r + \frac{(\sigma_{\max}-\sigma_r)(e^0-e)}{e^0(1-\varphi)} & \text{if } e \ge e^0\varphi, \end{cases} \tag{A.5}$$

The oxygen uptake σ as defined in (A.5) is continuous but not C^1. In the numerical simulations, it has been smoothed, since from the physiology, it is clear that the passage from one zone to the other occurs smoothly. The sigma used is shown in Figure A.1. Let us observe that, consistently with Hanon experimental results [7], the functional form of σ does not change with the athlete, whilst the values of $\sigma_{\max}$ and σ_f do.

Moreover, in [1], a second modification is introduced to the energy equation: for sufficiently long races (longer than $1000m$), it has been observed that slowing down recreates some of the anaerobic energy. Therefore, the energy equation results in:

$$\dot{e} = \sigma(e) + \eta(\dot{v}) - fv,$$

where η depends on the acceleration $\dot{v}$ and has the following form:

$$\eta(\dot{v}) = \begin{cases} 0 & \text{if } \dot{v} > 0 \\ c_\eta|\dot{v}|^2 & \text{if } \dot{v} \le 0, \end{cases} \tag{A.6}$$

where c_η is a constant to be tuned. This leads to oscillations in the velocity profile (cf. [1], Figure 2.4). In this paper, we will not focus on the term η and on the causes of the oscillations: we will briefly present some numerical results obtained with the term η in one simple case. For further details, see [1].

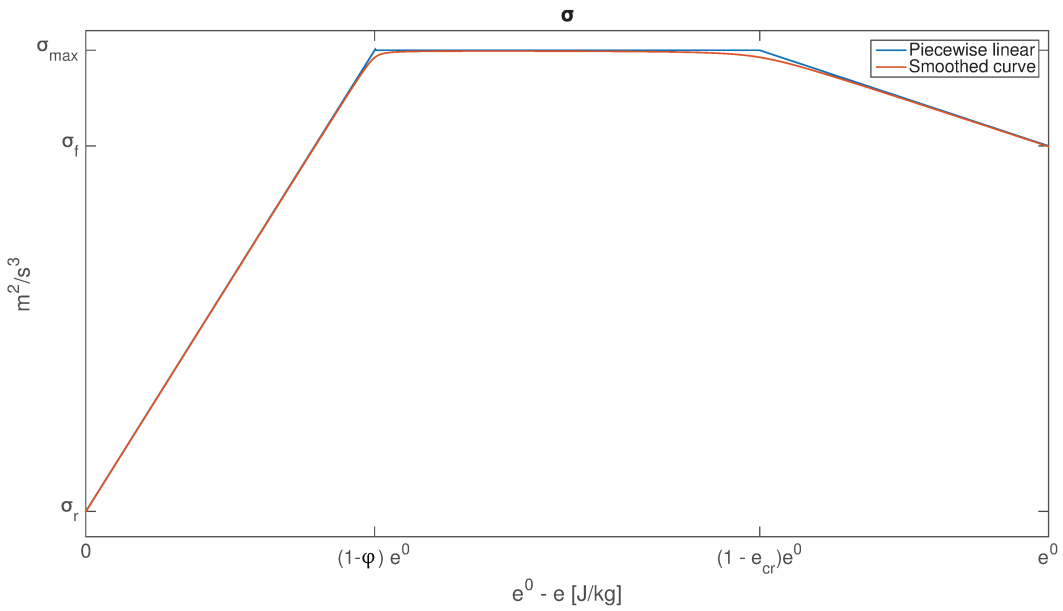


Figure A.1 – Typical curve σ vs $(e^0 - e)$ for a $1500m$ race.

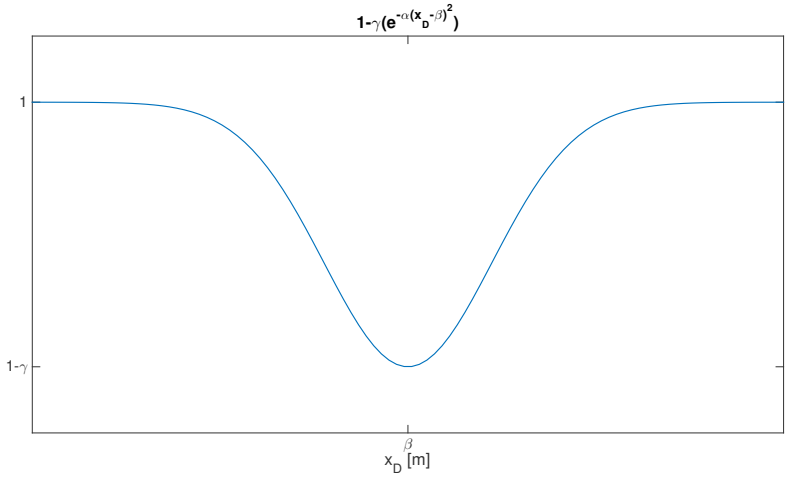


Figure A.2 – Term that modulates the friction with the air in the two-runners model.

A.2.2 Equations for two runners

In real races, the competition between runners has a fundamental impact on the strategy. Starting from the works of Keller [10], Quinn [18] and Kyle [11], Pitcher introduces a two-runners model in [16] based on the slipstream. The observation is that running behind someone can save 1 or 2 seconds per lap in middle distance races. Therefore, in the equations, the friction gets reduced when runners are just behind their competitor.

Some of the quantities in this model have a subscript i, which refers to the runner i: therefore x_i, v_i and e_i are, respectively, the position, the velocity and the energy of runner-i. All the physiological parameters which depend on the runner have the subscript i, too. Finally, the state variable x_D represents the distance between the runners, defined as $x_D := x_2 - x_1$. The energy balance for each runner is the same as in (A.1), with a constant value of σ, while the dynamics equation incorporates an aerodynamical term. Because it is the relative position which is important, instead of using x_i, the position of each runner as parameters, we use x_1 and the relative position x_D. The resulting model is the following: for $i = 1, 2$

$$\begin{cases} \dot{x}_1 = v_1 & x_1(0) = 0 \\ \dot{x}_D = v_2 - v_1 & x_D(0) = 0 \\ \dot{v}_1 = f_1 - \frac{v_1}{\tau_1} - c_1 v_1^2 (1 - \gamma(e^{-\alpha(x_D-\beta)^2})) & v_1(0) = 0 \\ \dot{v}_2 = f_2 - \frac{v_2}{\tau_2} - c_2 v_2^2 (1 - \gamma(e^{-\alpha(x_D+\beta)^2})) & v_2(0) = 0 \\ \dot{e}_i = \sigma_i - f_i v_i & e_i(0) = e_i^0. \end{cases} \tag{A.7}$$

As in Keller's model, the velocities and energies equations are normalized with respect to the mass of the runner. The term $-c_i v_i^2$ is the friction with the air. It is necessary to highlight and separate the effect of the friction with the air from the other ones, because it is the only one that is reduced while running in the slipstream of someone else. This frictional term is modulated by $1 - \gamma(e^{-\alpha(x_D \pm \beta)^2})$, which is shown in Figure A.2.

The parameter β represents the optimal distance a runner should keep from the other in order to obtain the maximal reduction of the air friction, while $\gamma \in (0, 1)$ is the

percentage reduction at the optimal distance. The parameter α is an index of the variance of the phenomenon and is chosen large enough so that the exponential term becomes negligible as we deviate by half a meter from β. We observe that the choice of a non symmetric term would probably be more accurate and realistic, but more complicated and with more parameter to estimate: for these reasons, in this work we use the term suggested by Pitcher in [16], which gives reasonable results when compared to real races. The parameters α, β and γ do not depend on the runner, however the parameter c is related to the drag coefficient and depends on the shape and surface properties of the athlete's body (see [18] for further details). For simplicity, in this work, we consider $c_1 = c_2 = c$. If ever the athletes have different masses, then we would have $m_1 c_1 = m_2 c_2$.

The problem has physiological constraints:

$$e_i(t) \geq 0 \quad \forall t \geq 0,\ i = 1, 2. \tag{A.8}$$

In order to have the state equations for both runners defined on the same time interval, Pitcher chooses a fixed final time T. Moreover, she fixes the running strategy of runner 1 (therefore $f_1(\mathbf{t})$ is given), as the optimal strategy they would adopt if running alone, therefore the only control is $f_2(t)$. Formally, the resulting optimization problem is:

$$\max_{f_2 \in \mathcal{F}} x_D(T) \quad \text{s.t. (A.7)-(A.8).} \tag{A.9}$$

It is clear that this strategy is not realistic, because all the runners adapt their strategy according to the performances of their opponents. In fact, one of the crucial point in a two-runners problem is how to model competition between them: some possible ways are game theory or multi-objective optimization. However, in this work, as explained below, we model the competition by leaving both strategies free, therefore having two controls, f_1 and f_2, and by encompassing in the cost functional the distance between the runners at the final time.

A.3 Mathematical model

In this work, we use Pitcher's two-runners model [16] and the single runner model of Aftalion-Bonnans [1] to build a new model for two runners, which incorporate psychological factors.

In order to simplify the notation, we substitute the expression $\sigma(e_i;\, \sigma_{\max,i},\, \sigma_{f,i},\, \sigma_{r,i},\, \varphi_i,\, e_{cr,i})$ from (A.5) with $\sigma_i(e_i)$. We recall that the friction in Pitcher's paper is $-cv_i^2(1-\gamma e^{-\alpha(x_D \pm \beta)^2})$ and is supposed to model the slipstream. Shielding behind someone has a strong impact when there is wind or when the velocity is high, however in real races this position does not come entirely from slipstream but from strategic factors, too, for which the position one meter behind is the best. Therefore, the potential $1-\gamma e^{-\alpha(x_D \pm \beta)^2}$ can also be considered to model a psychological factor which consists in trying to follow one's competitor, in order to be able to overtake. Indeed, it is a potential which has a minimum at distance β behind and decreases global friction. On the other hand, when the other runner is too far, there is no benefit. Let us observe that this model does not take into account the lateral displacement, and therefore the additional propulsive force, that is necessary to overtake. One can model the fact that overtaking requires some additional energy, by possibly using a non symmetric potential well $1 - \gamma e^{-\alpha(x_D \pm \beta)^2}$,

with a varying α, which is not what we have done in the simulations presented to reduce the number of parameters involved. We obtain the following equations: for $i = 1, 2$

$$\begin{cases} \dot{x}_1 = v_1 & x_1(0) = 0 \\ \dot{x}_D = v_2 - v_1 & x_D(0) = 0 \\ \dot{v_1} = f_1 - \frac{v_1}{\tau_1} - cv_1^2(1 - \gamma(e^{-\alpha(x_D-\beta)^2})) & v_1(0) = 0 \\ \dot{v_2} = f_2 - \frac{v_2}{\tau_2} - cv_2^2(1 - \gamma(e^{-\alpha(x_D+\beta)^2})) & v_2(0) = 0 \\ \dot{e_i} = \sigma_i(e_i) + \eta_i(\dot{v_i}) - f_i v_i & e_i(0) = e_i^0. \end{cases} \tag{A.10}$$

The equations in (A.10) are defined for $t \in (0, T)$, where T is the time at which the first of the two runners reaches the final distance D; to model this, it is necessary to add the following boundary condition to the system:

$$(x_1(T) - D)(x_2(T) - D) = 0. \tag{A.11}$$

As in the previous models, the energy has a lower bound:

$$e_i(t) \geq 0 \quad \forall t \in (0, T),\ i = 1, 2. \tag{A.12}$$

The choice of the cost functional, i.e. the quantity to be minimised, is a key point. As said before, in contrast with Pitcher's choice, in this case none of the strategies is fixed, therefore there are two controls: $f_1(t)$ and $f_2(t)$. Here, we propose to minimise the following quantity, given a proper constant weight $c_w > 0$:

$$J(f_1, f_2) = T + c_w|x_D(T)|. \tag{A.13}$$

The aim of this choice is to minimise the final time of the winner, and the term $c_w|x_D(T)|$ models the fact that the loser has tried to win as well. Different values of c_w can lead to different results, as in real races when two runners compete against each other multiple times the outcome of the race can change.

The resulting problem is:

$$\min_{f_i \in \mathfrak{F}_i} J \quad \text{s.t. (A.10)-(A.11)-(A.12)}, \tag{A.14}$$

where $\mathfrak{F}_i$ is the set of the admissible controls, and depends on the athlete. For physiological reasons, it is necessary to impose a bound to the variations of $\dot{f}$, in addition to the bounds on f already introduced, related to the fact that athletes cannot vary their propulsive force too quickly (see more details in [1]). This leads to the following definition of $\mathfrak{F}_i$:

$$\mathfrak{F}_i := \{f : 0 \leq f(t) \leq f_{M,i},\ |\dot{f}(t)| \leq K_i\ \forall t \in (0, T)\}, \tag{A.15}$$

where K_i and $f_{M,i}$ are constants depending on the athlete, which model the fact that every runner has a limited maximal force ($f_{M,i}$) and cannot vary it too quickly (K_i). The rest of the paper consists in providing numerical simulations of (A.14).

A.4 Numerical results

All the results presented in this section are obtained with the free software BOCOP [3]. The equations are solved with a finite difference scheme (implicit Euler), while the optimization problem is solved with an iterative method, using as stopping criterion the difference between successive iterates, with a tolerance of 10^{-10}.

The aim of these simulations is to find out if a runner can win a race against someone stronger, by running behind the first part of the race, and to quantify in term of variation of some parameters how much weaker they can be and when the best time to overtake is.

The single runner strategy has been mathematically proven and numerically computed in [1], and found experimentally in [8], and it consists in three parts:

- a first part of maximal force with a strong acceleration during which the peak velocity is achieved,
- a second part in which the propulsive force first decreases smoothly and then increases again, with the corresponding decrease and increase of the velocity,
- a final part at maximal force and maximal velocity again, until zero energy level is reached, where the velocity drops.

From the two-runners model, we expect an overall similar strategy: however, the additional term in the velocities equations encourages one of the runners to start slightly slower and to position themselves at distance β from the other. This allows them to keep the same velocity as the other runner while using a smaller propulsive force, which, in turn, leads to a lower energy consumption. We expect that, at a certain point during the race, the runner who is behind, will overtake the other by using the energy they have saved throughout the race and will be able to perform a longer final sprint. Moreover, it is reasonable to think that this moment occurs sooner if the runner who runs the first part of the race behind is stronger. Let us observe that this running strategy could also lead, in some favourable situations, to an improvement of the personal record. Nonetheless, it is important to underline that the decrease in the final time is not the main goal for a runner in some occasions, such as the Olympic finals, in which the final position is definitely more important than the final time: in [19], Thiel et al. study, starting from the Beijing 2008 Olympic Games data and the world records data, how the difference in the goal affects the pacing strategies: win the race versus minimising the final time.

How to estimate the parameters values starting from a race is beyond the scope of this paper. For this reason, the reference values for the parameters used in the simulations are taken from the literature and reported in Table A.1. The initial, maximal and final value of sigma (respectively σ_r, σ_{max} and σ_f) are taken from the $\dot{V}O_2$ values reported in [8]: let us observe that these values are given in $ml\ kg^{-1}min^{-1}$. In order to convert them in the unity of measurement needed (i.e. $m^2\ s^{-3} = J\ kg^{-1}s^{-1}$), we consider that the uptake of $1ml$ of oxygen is often converted into an energy expenditure estimate of $21J$. It is then necessary to convert the minutes in seconds, obtaining in this way the conversion factor 21/60. The values chosen for φ and e_{cr} aim at fitting the $_2$ profile reported in Figure 2 of [8]. The values for f_M and τ are strongly related: in fact, a first order approximation of the maximal velocity v_{peak} a runner can reach is τf_M. Therefore,

starting from the velocity profile plotted in Figure 1 of [8], one can chose a couple of reasonable values. The value of the constant c is taken from [18], c_η from [1], α and β from [16] and, finally, γ from [17].

Table A.1 – Parameters values for $1500m$

Parameter	Unit of Measurement	Value
τ	s	1.33
c	m^{-1}	0.0028
e^0	J/kg	1400
σ_r	m^2/s^3	6
$\sigma_{\max}$	m^2/s^3	24.22
σ_f	m^2/s^3	20.44
φ	-	0.5
e_{cr}	-	0.3
c_η	s	4
α	m^{-2}	10
β	m	1
γ	-	0.8
f_M	N/kg	5
c_w	-	0.1

First of all, let us consider the perfectly symmetric situation, in which the two runners have the same parameters. In this case it is very influential which runner is running the first part of the race behind: this can be mathematically modelled by choosing, in a proper way, the initial guess for the variable x_D given to the iterative method that solves the optimization problem. In fact, giving as initial guess β (or $-\beta$) forces runner-1 (or runner-2) to start the race more slowly and to position themselves behind for the first part. If one gives as initial guess $x_D = 0$, one finds a solution in which the runners overtake each other multiple times, which is not very realistic. The non uniqueness of the solution is not surprising, especially in a perfectly symmetric situation, such as the one considered: if a certain couple of strategies $\{f_1(t), f_2(t)\}$ provides a minimum for the cost functional, the couple $\{f_2(t), f_1(t)\}$ provides a minimum, too. This can be considered to model the fact that if the same two runners run against each other multiple times, the outcome of the race can change. The results of this simulation are shown in Figures A.3-A.4: one can observe that the runner who stays behind can keep the same velocity as the other runner, while using a significantly lower propulsive force and therefore having a much lower energy consumption. All the graphs in Figure A.3 have the position on the x-axis, which means that the velocity of runner-i is plotted with respect to the position of runner-i (and it is the same for the propulsive force, the energy and sigma). The choice of plotting with respect to position, and not time, has been made because it is the most common in the sports literature. Figure A.4 shows the distance between the runners: the overtaking occurs at about 94% of the race, corresponding to $1416m$. This variable is plotted with respect to a normalized time (i.e. t/T), because it does not concern only one runner, but both of them, therefore it does not make sense to plot it with respect to the position of any of them.

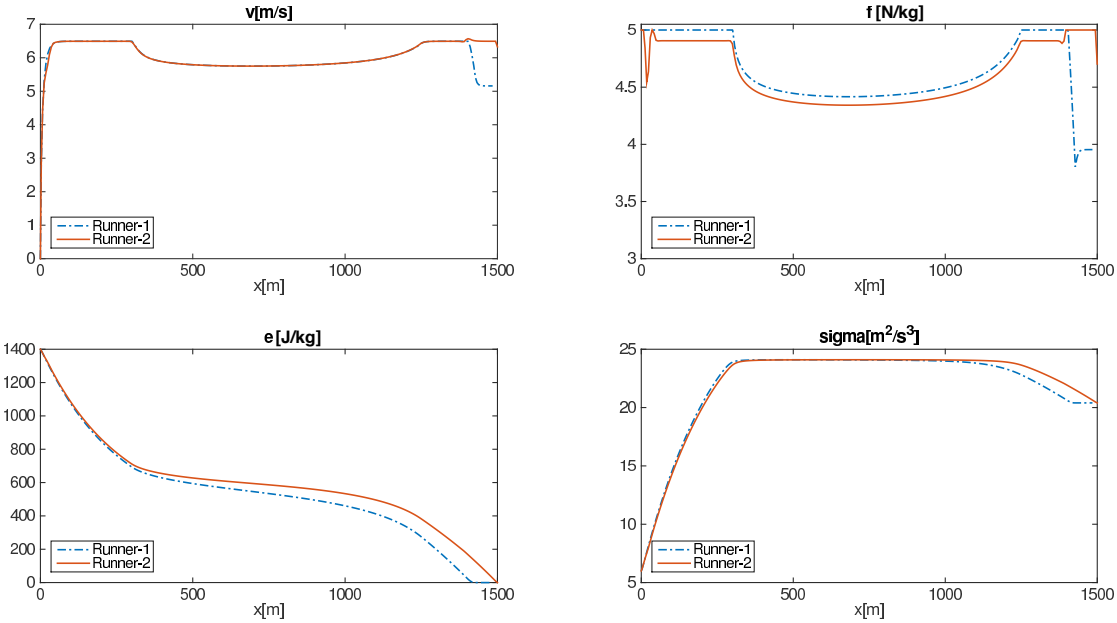


Figure A.3 – Competition between two runners with the same parameters.

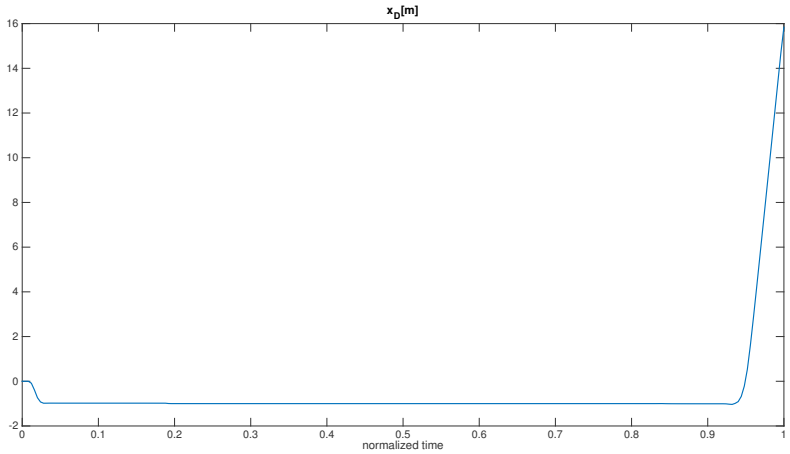


Figure A.4 – Competition between two runners with the same parameters; distance between the runners.

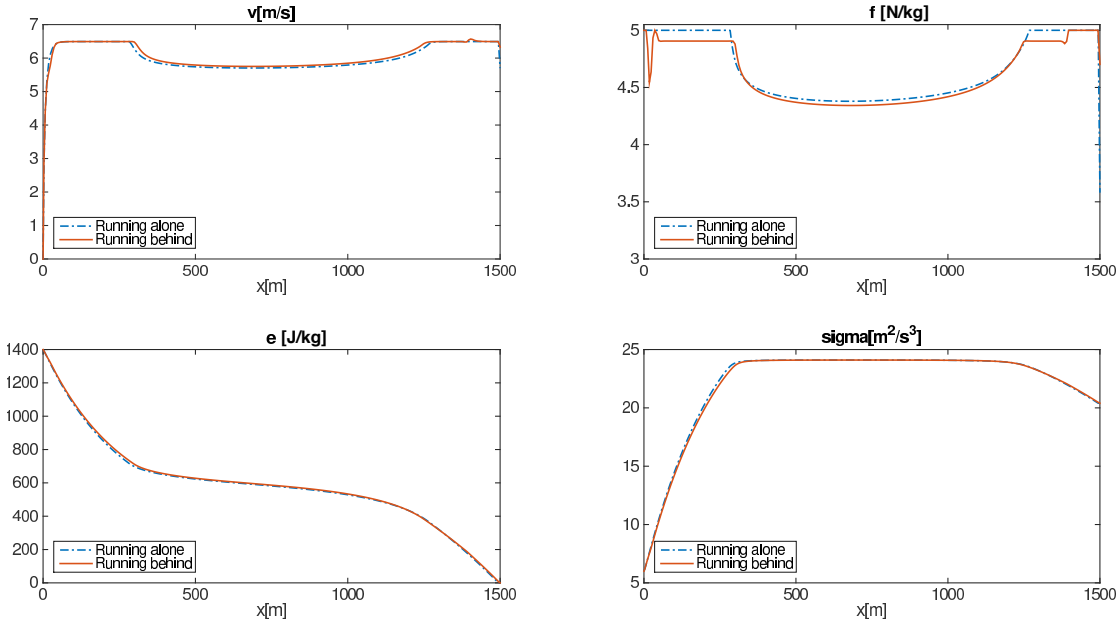


Figure A.5 – Running alone vs running behind.

From this first result, it is clear that running behind someone for the most part of the race allows runners to win against athletes as strong as themselves. We now want to investigate how the strategy of runners changes if they are running alone or behind someone else. In Figure A.5, two performances of the same runner are compared: the blue line represents the optimal strategy of the runner running alone (they complete the race in $249.681s$), the red line represents the strategy adopted when running behind someone as strong as themselves (they completes the race in $247.822s$). This difference in the final time (almost $2s$ of improvement) is equivalent to a difference of about $0.05m/s$ in the mean velocity. However, what we have simulated is not the most favourable situation to reduce the final time, and therefore to improve the mean velocity. If the aim is exclusively to improve the personal best performance and not to win the race, the best scenario possible for a runner is to run behind someone slightly stronger for the whole race. However, let us observe that the opponent must not be too much stronger: in this case, the runner would use too much energy to stay behind and would soon reach the zero energy level, which would cause a drop in the propulsive force too soon in the race. Nevertheless, from Figure A.5 we can still observe how running behind someone else allows an athlete to have a higher velocity in spite of keeping a smaller force throughout the race.

The results presented in Figures A.3-A.4-A.5 are obtained with the model (A.10) without the recreational term η introduced in (A.6). In Figures A.6-A.7 we present the same scenario, but with recreation: the two runners have the same parameters. We recall that, as explained in section A.2.1, the recreational term η leads to oscillations in the velocity profile. One can observe, comparing Figure A.3 with Figure A.6 and Figure A.4 with Figure A.7, that the strategy does not change: runner-2 slows down at the beginning, in order to be behind runner-1; in the middle part of the race the propulsive forces are oscillating (this behaviour is caused by the additional term η and

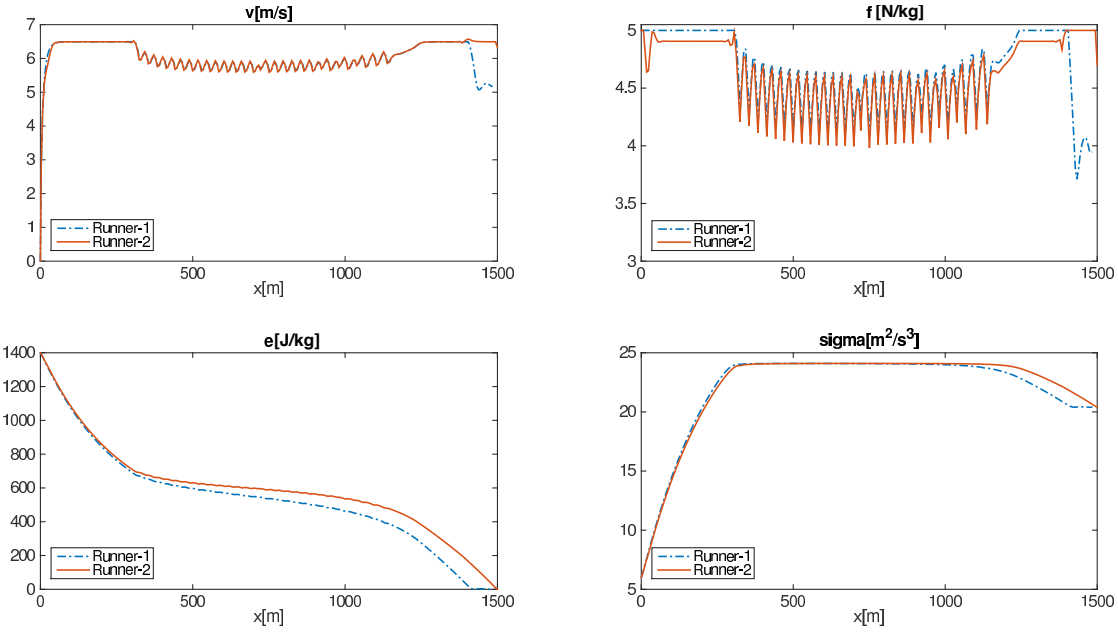


Figure A.6 – Competition between two runners with the same parameters, with the recreation term η.

can be found also in the single-runner problem, see [1] for further details), and the mean value around which the propulsive force of runner-1 is oscillating is slightly bigger than the one around which the force of runner-2 is oscillating; the energy curve does not change significantly, compared to the previous results. From Figure A.7, one can notice that runner-2 overtakes at about 94.85% of the race, which corresponds to $1417m$: one meter later, if compared with the case without oscillations. Let us observe that in this case the final time is smaller: in fact runner-2 completes the race in $247.75s$. This decrease in the final time, when adding the recreation η, is consistent with the results for the single-runner problem presented in [1].

Finally, we can say that the recreation term does not change the strategy of the race, however the results are more difficult to read, due to the oscillations. For this reason, from now on we will present only results obtained without η.

We now want to find the threshold values, i.e. how much runners can be weaker than their opponent and still win the race by running behind. Therefore, we vary one parameter at a time, making runner-2 weaker. Figures A.8-A.9 show the results for two runners who have a different initial energy. Given the reference value for $e_1^0 = 1400$ (as in Table A.1), the lowest initial energy runner-2 can have, while still being able to win the race, is:

$$e_2^0 = 1275 J/kg.$$

At the beginning of the race, runner-2 slows down in order to stay behind: in this way runner-2 manages to keep the same velocity as runner-1 using a smaller force; this leads to a smaller energy consumption, therefore at the end of the race runner-2 has enough energy to speed up and overtake runner-1. For the boundary condition (A.11), the time stops as soon as the first runner finishes the race. Therefore, when runner-2 reaches the

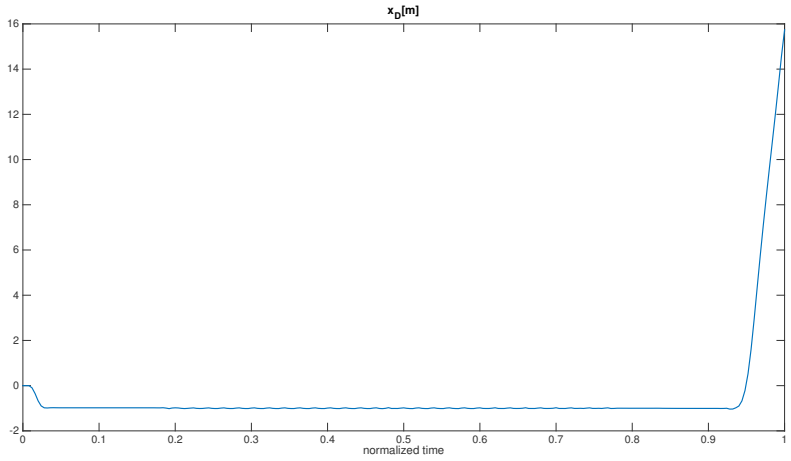


Figure A.7 – Competition between two runners with the same parameters, with the recreation term η; distance between the runners.

finish line (i.e. $1500m$), runner-1 has covered only $1498.13m$. The final time of runner-2 is $249.43s$, while their best performance running alone is $251.403s$, again an improvement of almost $2s$. As shown in Figure A.9, the overtaking occurs later in the race, if compared with the case in which the runners were equally strong: here occurs at 99% of the race (i.e. about $1487m$). This is reasonable: in fact, being weaker, runner-2 has to exploit the advantage of staying behind as long as possible.

In Figures A.10-A.11, the two runners have different oxygen uptake. The threshold values are the following:

$$\sigma_{max,1} = 24.22m^2/s^3,\ \sigma_{f,1} = 20.44m^2/s^3$$
$$\sigma_{max,2} = 23.75m^2/s^3,\ \sigma_{f,2} = 20.18m^2/s^3.$$

The initial strategy is the same as the previous case: runner-2 slows down in order to stay behind; this allows them to keep the same velocity as runner-1 using a smaller force and therefore compensating the smaller σ. The speed up in the final part is less evident than it was in the previous case, because the difference between the energies is smaller. The final distance covered by runner-1 in this case is $1499.72m$. The final time of runner-2 is $249.66s$, compared to $251.665s$ if running alone. Figure A.11 shows the distance between the runners during the race: as in the previous case, the overtaking occurs late in the race.

Figures A.12-A.13 show the results for two runners who have a different value for τ. The threshold values are:

$$\tau_1 = 1.33s \qquad \tau_2 = 1.31s.$$

A smaller τ indicates a bigger drop in velocity due to frictional effects, therefore, a greater force is necessary to keep the same velocity. Being behind another runner is a way to compensate this weakness: as shown in Figure A.12, runner-2 manages to keep the same velocity as runner-1, using a slightly smaller force, in spite of having a smaller τ. In the final part, when the difference in the energies is sufficiently big, runner-2 overtakes runner-1. In this case, at the end of the race, runner-1 has covered a distance

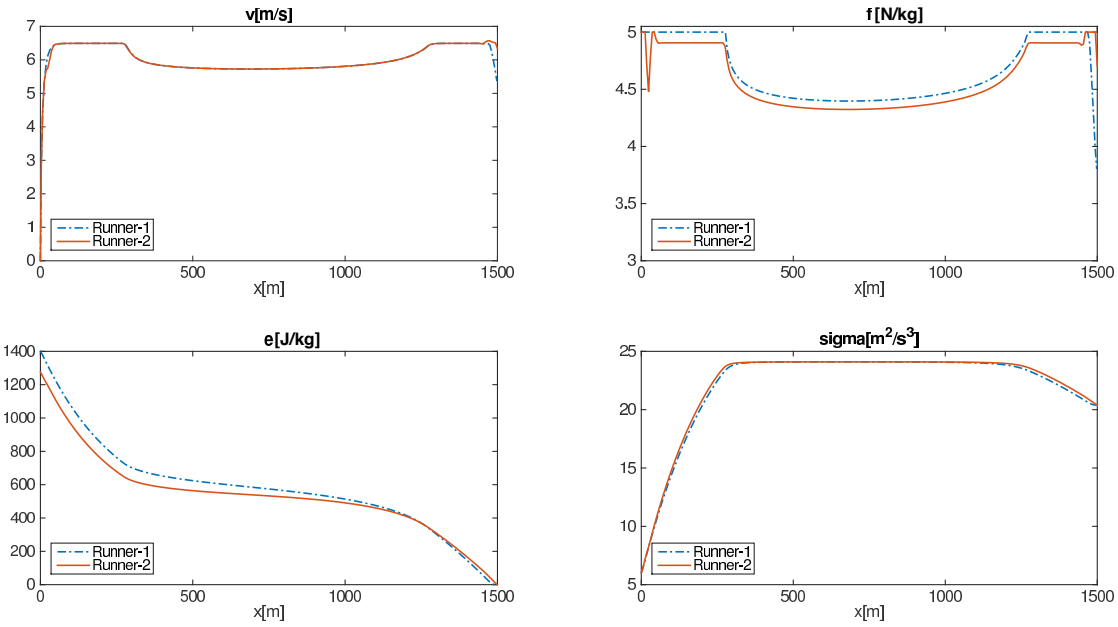


Figure A.8 – Competition between two runners with different e^0.

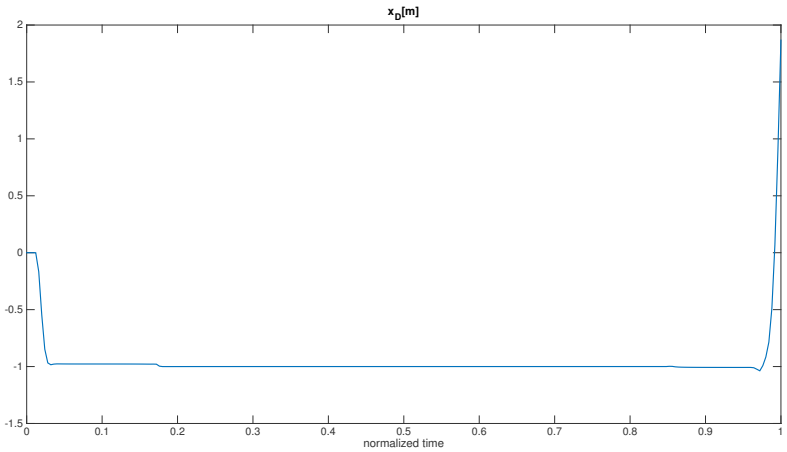


Figure A.9 – Competition between two runners with different e^0; distance between the runners.

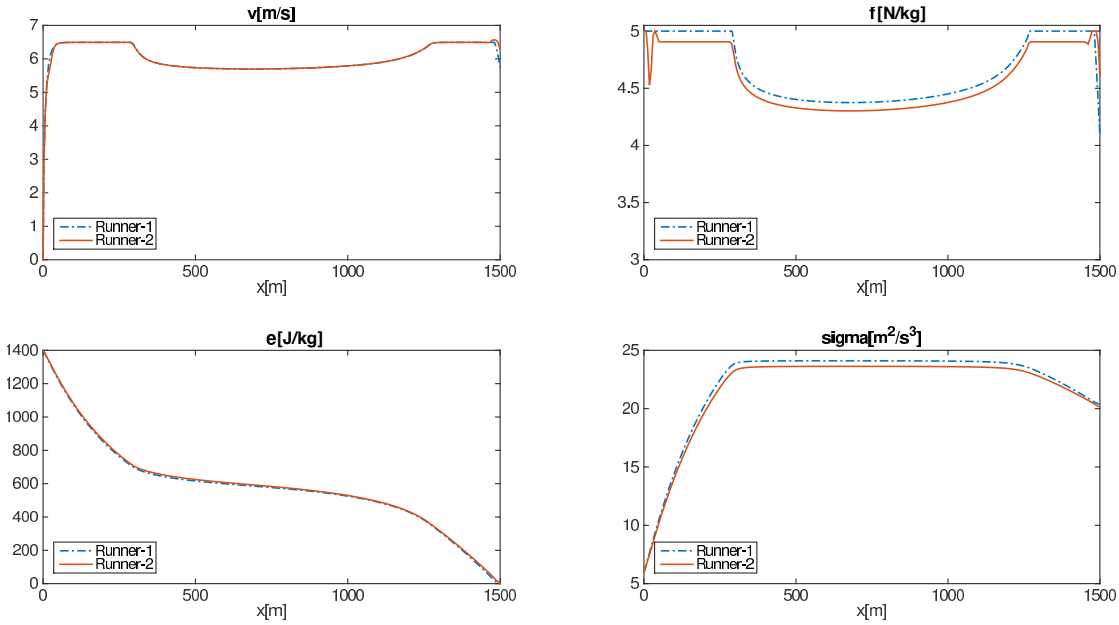


Figure A.10 – Competition between two runners with different σ.

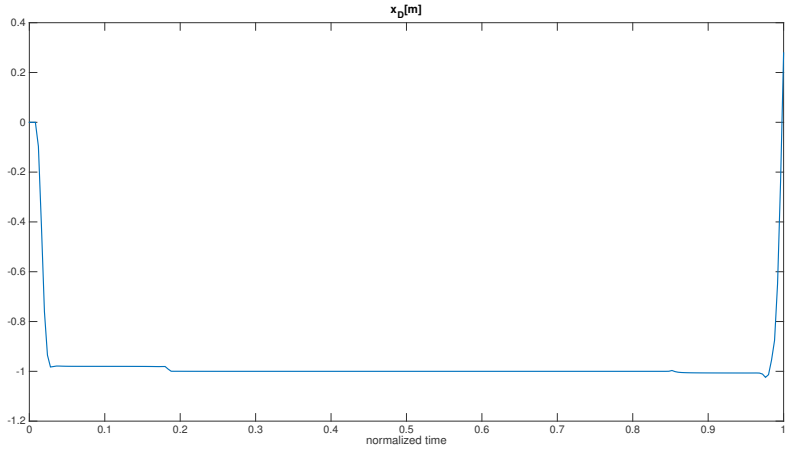


Figure A.11 – Competition between two runners with different σ; distance between the runners.

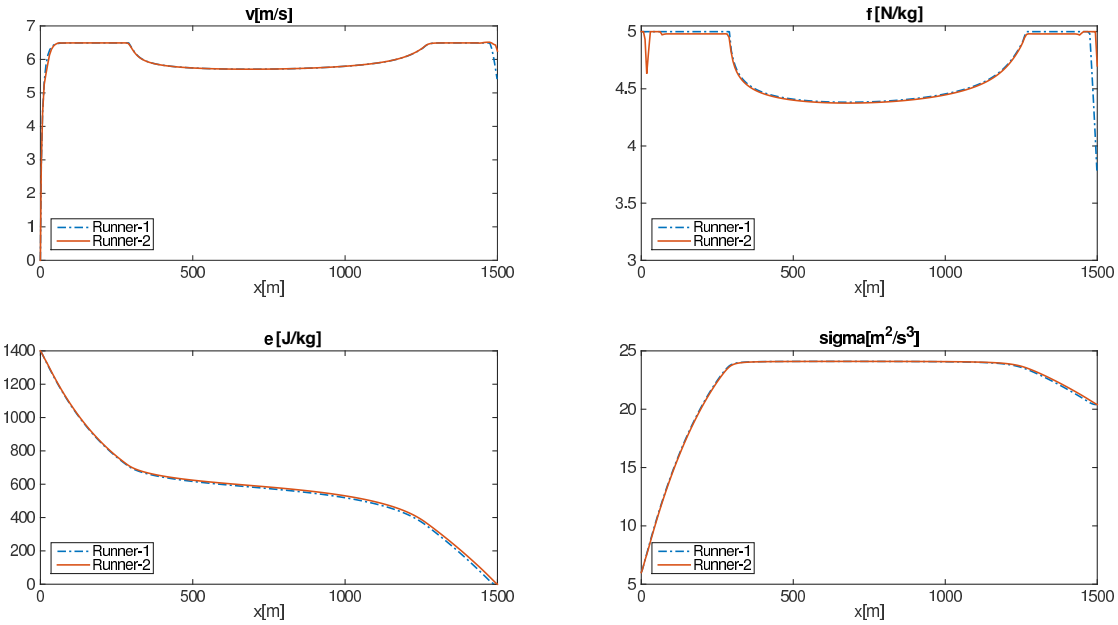


Figure A.12 – Competition between two runners with different τ.

of $1498.82m$. The final time of runner-2 is $249.536s$, while their best performance running alone is $251.83s$, i.e. they have an improvement of more than $2s$.

Finally, let us consider a case in which the runner who starts behind is stronger. For this purpose, we use the following parameters:

$$\tau_1 = 1.29s \qquad \tau_2 = 1.33s.$$

All the other parameters remain unvaried (reference values from Table A.1). Figures A.14-A.15 show the results obtained: in this case, the difference between the athletes is much bigger than in the previous ones, and this is evident from all the curves in Figure A.14. From Figure A.15, one can notice that the overtaking occurs very early in the race: at about 87.1%, i.e. $1290m$. What is interesting in this case is that runner-2 completes the race in $248.726s$, which is almost $1s$ less than their best performance running alone: therefore, in order to improve a personal record, it is not necessary to run behind someone stronger.

Let us now compare the results obtained here with Pitcher's ones. In [16], Figure 5.2, when the weaker runner is the one with the fixed strategy, the stronger runner remains only slightly ahead of their opponent until nearly the end of the race. This is in order that the weaker runner does not gain the advantage of running in the slipstream of the stronger for a long part of the race. However, it is the runner who stays behind who should adjust their position with respect to the other one, and not the opposite. The advantage of having two strategies free allow us to avoid this unrealistic result, and in this case, we get that the weaker runner stays one meter behind, and either wins the race if the difference in energy is not too big or drops following if they do not have enough energy.

Finally, we want to analyse the strategy and to see when the overtaking occurs in real races and compare them with our results. For this purpose, we have considered

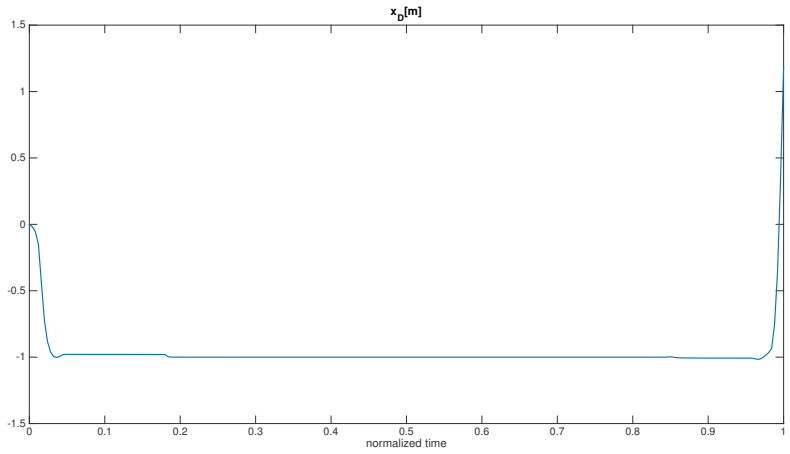


Figure A.13 – Competition between two runners with different τ; distance between the runners.

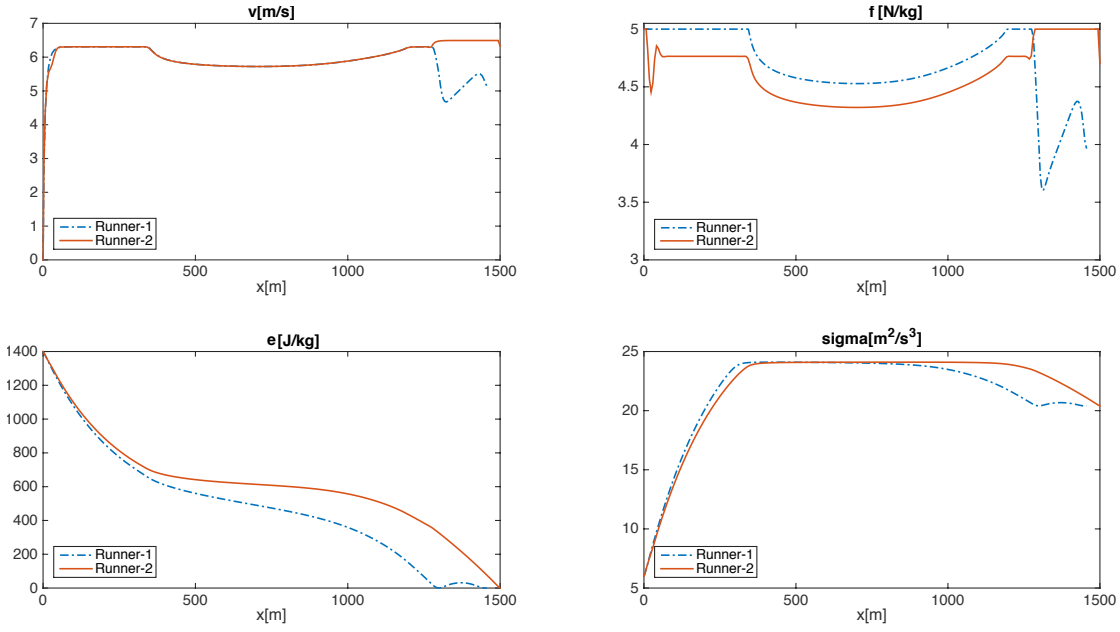


Figure A.14 – Competition between two runners with different τ; stronger runs the first part of the race behind.

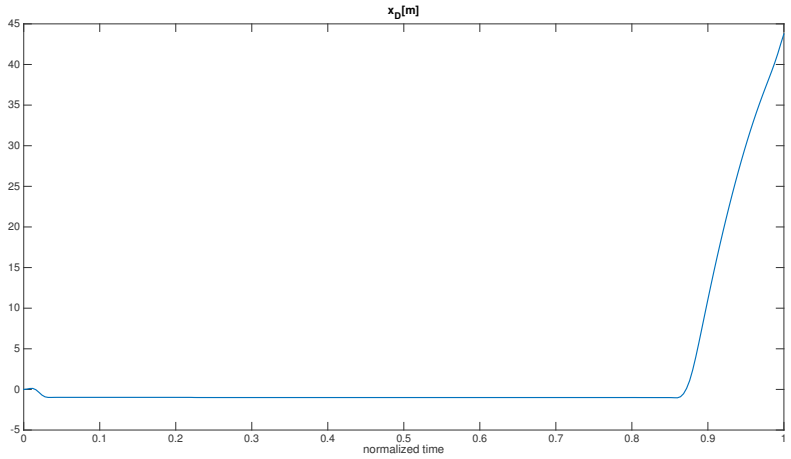


Figure A.15 – Competition between two runners with different τ; stronger runs the first part of the race behind; distance between the runners.

the men's $1500m$ finals of three different competitions: Beijing 2008 Olympic Games, Rome 2014 IAAF Diamond League and Singapore 2015 SEA Games. Videos of the races can be found on the internet [4, 5, 6]. We want to point out that the athlete who won the Beijing 2008 Olympics was disqualified one year later for doping and his gold medal was reassigned. Nonetheless, it is still interesting to analyse the race, knowing that doping increases the maximal value of σ but delays the time at which the peak velocity is reached: this should lead to a slower start, but it provides a capacity to keep a higher velocity for a longer part of the race. From the three videos [4, 5, 6], one can observe that the winner always runs the first part of the race behind, and this is consistent with our numerical results. The overtaking occurs at 84.6% of the race in Beijing 2008, at 96.9% in Rome 2014 and at 91.8% in Singapore 2015: these values are close to our numerical results, that vary between 87% and 99% of the race depending on the difference between the athletes.

A.5 Conclusion

In this work, we have presented a new model for a two-runners problem, starting from the single runner model of Aftalion and Bonnans [1] and from the two-runners model of Pitcher [16], changing the optimal control problem. The key of our simulations is that they quantify very precisely in terms of physiological parameters, optimal control problems and numerical simulations, phenomena which are only qualitatively understood. In this paper, we do not take into account the curvature of the track, which is the aim of an upcoming paper, since it requires more effort in the modelling.

As expected, going from one runner to two runners does not change the main characteristics of the velocity profile individuated already in [1]. We can still clearly distinguish the different phases of the race: the fast start, with maximal propulsive force and strong acceleration until the peak velocity is reached; an intermediate phase in which the propulsive force and the velocity first smoothly decrease and then increase; a final part at

maximal force again, where the runner speeds up (final sprint), followed by a very short zero energy arc, in which there is a drop in force and velocity. Running behind someone allows to keep a velocity with a smaller propulsive force than the one needed when running alone; this leads to a smaller energy consumption, therefore the zero energy level occurs later and the speed up at the end of the race is more pronounced.

Our numerical results show that if a runner has a fast start and leads the race for the most part, even if they are slightly stronger than their opponent, at the end they are overtaken: in order to lead the race and win, the physiological difference between the athletes has to be significant. Furthermore, we have shown how runners can improve their personal best performance by exploiting the advantage of running behind someone else, who can be stronger or weaker. The most significant improvements are obtained by running behind someone stronger.

An interesting development, in order to have more realistic results, would be to include a delayed reaction term which takes into account the fact that runners cannot adapt instantaneously their strategy to changes in their competitor's strategy. This could be compared to a stochastic model. Finally, one could increase the number of runners, in order to be able to model real races more accurately. These considerations are outside the scope of this paper, but they can be important for future research.

This model suggests to use special runners to set the pace for others and help improve their racing times in training. The other major application for Olympic training could be for athletes to estimate whether they should stay behind or lead, and when, given their physiology, and that of their opponents, is the best time to overtake.

Acknowledgements: I would like to thank Frédéric Bonnans for his very helpful comments and A. Aftalion for suggesting this interesting topic of research and for her many remarks. A first version of this work, of which she is co-author, can be found on arxiv (http://arxiv.org/abs/1508.00523v1).

Bibliography

[1] A. Aftalion and J.-F. Bonnans. Optimization of running strategies based on anaerobic energy and variations of velocity. *SIAM Journal on Applied Mathematics*, 74(5):1615–1636, 2014.

[2] H. Behncke. A mathematical model for the force and energetics in competitive running. *Journal of mathematical biology*, 31(8):853–878, 1993.

[3] J.-F. Bonnans, D. Giorgi, V. Grelard, S. Maindrault, and P. Martinon. BOCOP - A toolbox for optimal control problems. http://bocop.org.

[4] Y. channel: Athletics. Men's 1500m Final IAAF Diamond League Rome 2014 [video file], 2014. https://www.youtube.com/watch?v=CTVUGLapmPY.

[5] Y. channel: Beijing 2008 Athletics Gymnastics Aquatics. Athletics - Men's 1500M - Beijing 2008 Summer Olympic Games [video file], 2008. https://www.youtube.com/watch?v=0HcGVbDLhI8.

[6] Y. channel: Sport Singapore. Athletics Men's 1500m Final (Day 6) | 28th SEA Games Singapore 2015 [video file], 2015. https://www.youtube.com/watch?v=pfeVSzDnv-I.

[7] C. Hanon, P.-M. Lepretre, D. Bishop, and C. Thomas. Oxygen uptake and blood metabolic responses to a 400-m run. *European journal of applied physiology*, 109(2):233–240, 2010.

[8] C. Hanon, J. M. Leveque, C. Thomas, and L. Vivier. Pacing strategy and VO2 kinetics during a 1500-m race. *International journal of sports medicine*, 29(3):206–211, 2008.

[9] C. Hanon and C. Thomas. Effects of optimal pacing strategies for 400-, 800-, and 1500-m races on the VO2 response. *Journal of sports sciences*, 29(9):905–912, 2011.

[10] J. B. Keller. Optimal velocity in a race. *American Mathematical Monthly*, pages 474–480, 1974.

[11] C. R. Kyle. Reduction of wind resistance and power output of racing cyclists and runners travelling in groups. *Ergonomics*, 22(4):387–397, 1979.

[12] F. Mathis. The effect of fatigue on running strategies. *SIAM review*, 31(2):306–309, 1989.

[13] R. H. Morton. A three component model of human bioenergetics. *Journal of mathematical biology*, 24(4):451–466, 1986.

[14] R. H. Morton. A 3-parameter critical power model. *Ergonomics*, 39(4):611–619, 1996.

[15] R. H. Morton. The critical power and related whole-body bioenergetic models. *European journal of applied physiology*, 96(4):339–354, 2006.

[16] A. B. Pitcher. Optimal strategies for a two-runner model of middle-distance running. *SIAM Journal on Applied Mathematics*, 70(4):1032–1046, 2009.

[17] L. G. C. E. Pugh. The influence of wind resistance in running and walking and the mechanical efficiency of work against horizontal or vertical forces. *The Journal of Physiology*, 213(2):255–276, 1971.

[18] M. Quinn. The effects of wind and altitude in the 400-m sprint. *Journal of sports sciences*, 22(11-12):1073–1081, 2004.

[19] C. Thiel, C. Foster, W. Banzer, and J. De Koning. Pacing in olympic track races: competitive tactics versus best performance strategy. *Journal of sports sciences*, 30(11):1107–1115, 2012.

[20] C. Thomas, C. Hanon, S. Perrey, J. M. Le Chevalier, A. Couturier, H. Vandewalle, et al. Oxygen uptake response to an 800-m running race. *International journal of sports medicine*, 26(4):268–273, 2005.

[21] A. J. Ward-Smith. A mathematical theory of running, based on the first law of thermodynamics, and its application to the performance of world-class athletes. *Journal of biomechanics*, 18(5):337–349, 1985.

[22] W. Woodside. The optimal strategy for running a race (a mathematical model for world records from 50 m to 275 km). *Mathematical and computer modelling*, 15(10):1–12, 1991.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Analyse de sensibilité pour systèmes hyperboliques non linéaires

Mots Clefs : Analyse de sensibilité, EDP hyperboliques, lois de conservation, quantification d'incertitude, optimisation

Résumé : L'analyse de sensibilité (AS) concerne la quantification des changements dans la solution d'un système d'équations aux dérivées partielles (EDP) dus aux variations des paramètres d'entrée du modèle. Les techniques standard d'AS pour les EDP, comme la méthode d'équation de sensibilité continue, requirent de dériver la variable d'état. Cependant, dans le cas d'équations hyperboliques l'état peut présenter des discontinuités, qui donc génèrent des Dirac dans la sensibilité. Le but de ce travail est de modifier les équations de sensibilité pour obtenir un syst'eme valable même dans le cas discontinu et obtenir des sensibilités qui ne présentent pas de Dirac. Ceci est motivé par plusieurs raisons : d'abord, un Dirac ne peut pas être saisi numériquement, ce qui pourvoit une solution incorrecte de la sensibilité au voisinage de la discontinuité ; deuxièmement, les pics dans la solution numérique des équations de sensibilité non corrigées rendent ces sensibilités inutilisables pour certaines applications. Par conséquent, nous ajoutons un terme de correction aux équations de sensibilité. Nous faisons cela pour une hiérarchie de modèles de complexité croissante : de l'équation de Burgers non visqueuse au système d'Euler quasi-1D. Nous montrons l'influence de ce terme de correction sur un problème d'optimisation et sur un de quantification d'incertitude.

Title: Sensitivity analysis for nonlinear hyperbolic equations of conservation laws

Keys words: Sensitivity analysis, hyperbolic PDEs, conservation laws, uncertainty quantification, optimization

Abstract: Sensitivity analysis (SA) concerns the quantification of changes in Partial Differential Equations (PDEs) solution due to perturbations in the model input. Standard SA techniques for PDEs, such as the continuous sensitivity equation method, rely on the differentiation of the state variable. However, if the governing equations are hyperbolic PDEs, the state can exhibit discontinuities yielding Dirac delta functions in the sensitivity. We aim at modifying the sensitivity equations to obtain a solution without delta functions. This is motivated by several reasons: firstly, a Dirac delta function cannot be seized numerically, leading to an incorrect solution for the sensitivity in the neighbourhood of the state discontinuity; secondly, the spikes appearing in the numerical solution of the original sensitivity equations make such sensitivities unusable for some applications. Therefore, we add a correction term to the sensitivity equations. We do this for a hierarchy of models of increasing complexity: starting from the inviscid Burgers' equation, to the quasi 1D Euler system. We show the influence of such correction term on an optimization algorithm and on an uncertainty quantification problem.