Discrete Tangent and Adjoint Sensitivity Analysis for Discontinuous Solutions of Hyperbolic Conservation Laws

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der
RWTH Aachen University zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Jonathan Jakob Hüser, M.Sc.

aus Peine

Berichter: Universitätsprofessor Dr. Uwe Naumann
Universitätsprofessor Dr. Michael Herty

Tag der mündlichen Prüfung: 9. Mai 2022

Abstract

English Version

We consider the discrete tangent and adjoint sensitivities computed via algorithmic differentiation of shock capturing numerical methods for hyperbolic conservation laws which are widely used for models of fluid dynamics such as those based on the Euler equations. For discontinuous solutions the discrete sensitivities do not generally converge to the correct sensitivities of the analytical solution as the discretization grid is refined because the analytical sensitivities are singular at the discontinuities of the solution.

In this thesis we propose a convergent numerical approximation of the correct sensitivities of shock discontinuities in discontinuous solutions of hyperbolic conservations laws with respect to the parameters of the initial data. We compute the shock sensitivities by approximating the Rankine-Hugoniot condition taking into consideration the numerical viscosity of shock capturing numerical methods in a way that can be computed by algorithmic differentiation tools. The resulting discrete sensitivities enable for example the gradient-based parameter optimization of optimization problems constrained by a hyperbolic conservation law.

Deutsche Version

Wir betrachten die diskreten tangenten und adjungierten Sensitivitäten berechnet durch algorithmisches Differenzieren von numerischen Methoden für hyperbolische Erhaltungsgleichungen, die durch numerische Viskosität unstetige Stoßwellen glätten. Solche Verfahren werden für Modelle der Strömungsmechanik genutzt, z.B. die Euler Gleichungen. Für unstetige Lösungen konvergieren die diskreten Sensitivitäten nicht allgemein zu den korrekten Sensitivitäten der analytischen Lösung, wenn das Diskretisierungsgitter verfeinert wird, da die analytischen Sensitivitäten an den Unstetigkeitsstellen singulär sind.

In dieser Arbeit unterbreiten wir eine konvergente numerische Approximation der korrekten Sensitivitäten der Stoßwellen-Unstetigkeiten in unstetigen Lösungen hyperbolischer Erhaltungsgleichungen hinsichtlich der Parameter der Anfangsbedingung. Wir berechnen die Sensitivitäten der Stoßwellen durch Approximation der Rankine-Hugoniot Bedingung unter Rücksichtnahme der numerischen Viskosität des numerischen Verfahrens, sodass die Approximation mit Tools für algorithmisches Differenzieren berechnet werden kann. Die resultierenden diskreten Sensitivitäten ermöglichen z.B. Gradienten-basierte Parameteroptimierung für Optimierungsprobleme mit hyperbolischen Erhaltungsgleichungen als Nebenbedingung.

Acknowledgements

The completion of this thesis would not have been possible without the help of several people.

First of all, I want to thank my advisor Prof. Dr. Uwe Naumann for his guidance and constant support and for introducing me to the field of algorithmic differentiation which today is as relevant as ever. His constructive suggestions and valuable advice contributed significantly to the success of this work.

I also want to thank Prof. Dr. Michael Herty for his in-depth comments and insightful inputs on the technical matters related to hyperbolic conservation laws and their numerical approximation.

I am grateful to all of my co-workers, HiWis and students at STCE for providing such a great working atmosphere. In particular, I would like to thank Jens Deussen, Klaus Leppkes, Johannes Lotz, Markus Towara, Patrick Peltzer, Simon Märtens, Erik Schneidereit, Philippe Suchsland, Viktor Mosenkis, Niloofar Safiran, Arindam Sen, Azar Dastouri and Jan Riehme for their continued support throughout the years.

Special thanks goes to my wife, Anne, for her constant encouragement during the final years of writing this thesis and to my son, Johannes, for all the joy he has brought into our lives.

Finally, I want to thank my family and friends for their patience and never ending support. Thank you so much!

Contents

Chapter 1

Introduction

1.1 Overview

1.1.1 Problem Setting

We consider the parametric scalar conservation law initial value problem

$$u(0,x,p) = u_0(x,p) \tag{1.1.1}$$

$$0 = \partial_t u(t,x,p) + \partial_x f(u(t,x,p)) \tag{1.1.2}$$

for all $t \in [0,T]$, $x \in \Omega \equiv (-\infty,\infty)$ with $p \in D \subseteq \mathbb{R}^d$ and f twice continuously differentiable. The given partial differential equation (PDE) is a conservation law in the following sense. When integrating Equation (1.1.2) over the space interval $[L,R]$ we get an ordinary differential equation (ODE) that describes the dynamics of the amount of the quantity u in that interval

$$\partial_t \int_L^R u(t,x,p)\mathrm{d}x = -\int_L^R \partial_x f(u(t,x,p))\mathrm{d}x = f(u(t,L,p)) - f(u(t,R,p)) \ .$$

The amount of u contained in the interval $[L,R]$ changes according to the inflow and outflow of u at the boundary of the interval. Consequently, the total amount of u over the space domain Ω does not change and so u is a conserved quantity. The inflow and outflow at the boundary are determined by the flux function f.

Because of the parametric initial data the PDE solution u is implicitly a function of p. In this thesis we consider the discrete tangent and adjoint sensitivity analysis of the solution u with respect to the parameter p. We present algorithmic differentiation (AD) methods applicable to numerical integrators of Equations (1.1.1)-(1.1.2). The contributions of this thesis are related to the correct treatment of parameter sensitivities of the solution around shock discontinuities. In the following we briefly introduce the challenges that arise in the presence of shock discontinuities.

If the solution u is smooth and continuously differentiable with respect to p at $p = p_0$ then the derivative gives a linear approximation with quadratic error

$$u(t,x,p_0+\dot{p}) - u(t,x,p_0) - \mathrm{d}_p u(t,x,p_0)\dot{p} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$ according to the Taylor expansion. In a sufficiently smooth setting it is possible to obtain the derivatives with respect to p by implicit differentiation of the PDE. Both classical tangent and adjoint sensitivity analyses are based on implicit differentiation. Tangent and adjoint AD methods also work based on similar regularity assumptions even though with AD we are differentiating a discrete numerical approximation of the solution.

Conservation laws can have discontinuous solutions and, hence, do not necessarily satisfy the regularity conditions required for classical differentiation (see Section 1.3.2). Shock discontinuities in the solution arise for example in models for computational fluid dynamics (e.g. [SS19; Sch+16]). Due to the lack of regularity in the presence of shock discontinuities the classical approach to tangent and adjoint sensitivity and also standard AD methods fail. The analytic tangent sensitivities of discontinuous solutions are singular at the shock locations [BJ99] and the discrete tangent sensitivities obtained by AD methods applied to shock capturing schemes approximate these singularities. Consequently, the tangent sensitivities do not give the desired quadratic error term for the first-order Taylor expansion in the L_1 norm [BM95b] and standard AD methods applied to integral objectives for optimal control do not give the correct gradient in the presence of shock discontinuities [Gil03].

1.1.2 Contributions

Based on the analytic generalized differential calculus for conservation laws with discontinuous solution by Bressan and Marson [BM95b] we propose numerical methods that give a nonlinear generalized Taylor expansion with quadratic error term in the sense of the L_1 norm, i.e. with

$$\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \mathrm{d}_p u(t,\cdot,p_0)\dot{p}\|_{L_1(\Omega)} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$. Our numerical method for the generalized Taylor expansion is a discrete tangent sensitivity analysis.

Similar to Giles [Gil03] and Giles and Ulbrich [GU10a; GU10b] we also solve the problem of finding a numerical method for the gradient of integral objective functions Φ with respect to the parameter p, where Φ has the form

$$\Phi(p) \equiv \int_\Omega \varphi(u(t,x,p))\mathrm{d}x \tag{1.1.3}$$

where φ is twice continuously differentiable. As opposed to the approach by Giles and Ulbrich the methods we present do not require the use of a specific numerical integrator. Our method for integral objective gradients is available as a discrete adjoint sensitivity analysis with the benefit of computing the full gradient in one reverse sweep at a cost that is a constant factor of the forward simulation.

The critical part in both the generalized Taylor expansion and the gradient of the integral objective is the sensitivity of the location of shock discontinuities in the solution of the conservation law. The sensitivity analysis for the location of shock discontinuities is particulary challenging when we only have a discrete numerical approximation of the solution. In this thesis we present methods for the discrete sensitivity analysis of shock locations in conservation laws based on shock capturing discrete numerical solutions that potentially involve numerical viscosity and oscillations around shock discontinuities.

We make the following contributions:

We give a derivation of the analytic sensitivity analysis for piecewise smooth solutions of a scalar conservation law in the sense of distributions based on an assumed solution structure that makes discontinuities explicit (see Section 5.2.2). For the assumed solution structure we also give a derivation of the generalized tangent of Bressan and Marson [BM95b] (see Section 5.3). Furthermore, we give a derivation of the analytic adjoint sensitivity of an integral objective functional of a piecewise smooth solution with the assumed structure (see Section 5.4).

We propose a numerical approximation of the shock locations in the shock capturing solution of a scalar conservation law and prove that the approximation converges to the correct value as the discretization grid is refined (see Section 6.2.1). We also propose a numerical approximation of the shock sensitivities based on the shock capturing solution and its discrete tangent sensitivity and prove that the approximation converges to the correct value as the discretization grid is refined (see Section 6.2.2). For both of the numerical approximations we make some assumptions about the shock capturing solution of the conservation law that we show empirically hold for some examples and provably hold for the traveling wave analytic viscosity solution (see Section 6.1).

Based on the numerical approximation of the shock sensitivities we propose a numerical approximation of the generalized tangent of Bressan and Marson [BM95b] and prove that the approximation converges to the analytic generalized tangent of the assumed solution structure as the discretization grid is refined (see Section 6.3). Based on the numerical approximation of the shock sensitivities we also propose a numerical approximation of the adjoint sensitivity of an integral objective and prove that the approximation converges to the analytic adjoint as the discretization grid is refined (see Section 6.4).

We show that the error of our proposed numerical approximation of the sensitivities of the shock locations in scalar conservation laws is $O(\Delta x^{\alpha})$ for some $0 < \alpha < 1$ as the uniform grid spacing Δx is refined, i.e. as $\Delta x \to 0$. We also show that the approximation error of the resulting generalized tangent and adjoint approximations is $O(\Delta x^{\alpha})$ as $\Delta x \to 0$ for the same α as a consequence of using the shock sensitivity approximation.

We show how the numerical approximation of the shock sensitivity computation can be implemented via AD tools using both the forward and reverse mode of algorithmic differentiation (see Section 6.5). The AD tool based approach produces the discrete adjoint for the numerical approximation of the shock sensitivity without having to explicitly simulate the adjoint PDE of the conservation law.

In computational experiments with the one-dimensional convex scalar Burgers flux and initial data with a known analytic solution we validate the empirical convergence of our proposed numerical approximations (see Section 6.6). We validate the empirical convergence of the numerical approximations for the shock location, shock sensitivity, the components of the generalized tangent and the adjoint sensitivity of an integral objective based on both Lax-Friedrichs and Upwind schemes. Additionally, we also validate that the presented methods empirically converge for a nonconvex scalar flux function (see Section 6.6.3).

Using the example of a system of isentropic gas equations we show how our proposed numerical methods generalize from scalar conservation laws to hyperbolic systems of conservation laws (see Chapter 7). We validate the empirical convergence of our proposed numerical approximations for the isentropic gas equations in computational experiments.

Using the example of the two-dimensional Burgers equation we show how our proposed numerical methods can be generalized to track two-dimensional points on a one-dimensional shock manifold in a two-dimensional conservation law and how the sensitivities of these points can be computed. An analogous approach can be used for two-dimensional shock manifolds in three-dimensional conservation laws. Based on such a point tracking additional geometric reconstructions are needed to approximate the full shock front. We validate the empirical convergence of our proposed numerical approximation for one point on a linear one-dimensional shock manifold in a two-dimensional Burgers inital value problem with known analytic solution.

1.2 Related Work

Bressan and Marson [BM95b] discuss the analytic tangent sensitivity analysis of solutions of hyperbolic systems of n conservation laws. They consider the case of discontinuous solutions with shocks where the tangent sensitivity does not exist in the classical sense. As motivation they introduce the ramp example we also introduce in Example 7 and that we use throughout the thesis. For a parametric family $\bar{u}^\varepsilon$ of initial data depending on a parameter ε the authors give a description of the corresponding solution $u^\varepsilon(t, \cdot)$ for $t > 0$ that is first-order accurate with respect to the L_1 norm as $\varepsilon \to 0$. To that end they define the generalized tangent as a tuple $(v, \xi) \in L_1(\Omega, \mathbb{R}^n) \times \mathbb{R}^n$ consisting of the broad tangent v that leaves out the singular part of the tangent sensitivity (see Section 5.3.1) and ξ the sensitivity of the shock location. In the context of our problem formulation with $p \in D \subseteq \mathbb{R}^d$ we consider perturbations of p in the direction $\dot{p}$, i.e. what corresponds to $\bar{u}^\varepsilon(x) = u_0(x, p_0 + \epsilon\dot{p})$ in Bressan and Marson [BM95b].

The authors describe a tangent variation that is nonlinear in the parameter because it approximates the shock travel as a non-singular perturbation of the shock location. They show that as long as the shocks do not interact the generalized tangent (v, ξ) satisfies a linear hyperbolic system coupled with one ODE per shock derived from the Rankine-Hugoniot condition (see Section 3.1.4). We do not go into the technical details of the generalized tangent approach at this point since the approach is also mentioned in Section 1.3.7 of the introduction and treated in detail for the scalar case with a piecewise smooth solution in Section 5.3.

Bressan and Marson [BM95b] use the method of characteristics to analyze piecewise Lipschitz continuous solutions of systems of hyperbolic conservation laws with a finite number of shocks. Their analysis is more general than ours in Section 5.3 since we formally consider piecewise smooth solutions and we only treat the scalar case. However, we validate via computational experiments that our proposed numerical methods extend to hyperbolic systems of conservation laws in Chapter 7. Bressan and Marson [BM95b] also consider the setting where shocks interact at some time $t > 0$ such that the number of shocks changes. We leave the numerical treatment of that setting to further work.

Bressan and Marson [BM95a] describe the application of the generalized tangent approach to an optimal control problem with an integral objective. In this second paper the authors also derive the analytic adjoint equations to the coupled linearized equations of the generalized tangent (v, ξ) yielding the corresponding generalized adjoint (v^*, ξ^*).

The work in this thesis builds on the analytic idea of the generalized tangent sensitivity formalism and its resulting adjoint sensitivities introduced in [BM95b] and [BM95a]. Our main contribution is the derivation of a convergent numerical approximation of the required broad tangent sensitivities, shock sensitivities and integral objective sensitivities. The corresponding discrete adjoint sensitivities follow directly from the application of the adjoint inner product identity and can be obtained by AD tools.

Bressan and Guerra [BG97] expand on [BM95b] and introduce the notion of shift-differentials for maps to functions of bounded variation. According to Dafermos [Daf77] classical prototypes of weak solutions of scalar conservation laws are constructed in the class of piecewise smooth functions with jump discontinuities across smooth curves. For the homogeneous case given in Equations (1.1.1)-(1.1.2) with smooth initial data solutions are piecewise smooth [Sch73]. But the class of piecewise smooth solutions does not include all solutions of general scalar conservation laws even for smooth initial data. A more general and natural framework in which solutions should be studied is the class of functions with bounded variation [CS66]. Bressan and Guerra [BG97] prove that the flow generated by a scalar conservation law is generically shift-differentiable even though such solutions can generally have complicated shock structure, e.g. dense shock sets.

Cliff, Heinkenschloss, and Shenoy [CHS97] show the differentiability of an optimal control problem for the discontinuous solution of a duct flow problem with a shock. The authors make the point that the shock discontinuity poses a problem when using good shock capturing schemes with low continuity properties since they cannot be combined with efficient optimization methods requiring smooth functions. In that sense the authors address a similar numerical issue to the one we solve in this thesis and by a related method; they also introduce the shock location as an explicit variable with dynamics according to the Rankine-Hugoniot condition. The authors first present a reformulation of the analytic optimal control problem with the shock location as explicit variable and show the existence of optimal solutions, Lagrange multipliers and differentiability. Based on the reformulation the authors present a numerical discretization where the grid depends explicitly on the shock location. In that sense the approach of Cliff, Heinkenschloss, and Shenoy [CHS97] is a shock fitting discretization with an adaptive grid and differs from the approach presented in this thesis in that our approach aims to work for any underlying shock capturing scheme without altering the numerical integrator.

Cliff, Heinkenschloss, and Shenoy [CHS98] continue the work of Cliff, Heinkenschloss, and Shenoy [CHS97] by deriving the corresponding adjoint equations and, hence, the adjoint sensitivity gradient for the optimal control formulation with the shock location as an explicit variable.

Bouchut and James [BJ99] discuss the analytic tangent sensitivity analysis with respect to the initial data of discontinuous solutions of scalar conservation laws with convex flux. They linearize the nonlinear scalar conservation law to obtain a linear conservation law that is the tangent PDE (see Section 4.1.2). In the case of a discontinuous solution of the nonlinear conservation law the linearized conservation law has discontinuous coefficients. The authors solve the linearized conservation law with discontinuous coefficients in the space of measures on $\mathbb{R}$, i.e. in the sense of distributions (see Section 2.4), and show existence and uniqueness of the solution using the notion of duality solutions [BJ98]. The resulting derivative is continuous in the weak sense of measures and can be interpreted as a weak derivative with respect to the initial data for the nonlinear scalar conservation law.

Godlewski and Raviart [GR99] discuss the analytic tangent sensitivity analysis of discontinuous solutions of hyperbolic systems of conservation laws with respect to the initial data from the perspective of linearized stability. They extend on Bouchut and James [BJ99] and show that for nonlinear systems with discontinuous solutions the linearized conservation law with discontinuous coefficients can also be solved in the space of measures. The solution of the linearized system is singular on the discontinuity curve of the solution of the nonlinear system. Since the coefficients in the linearized system are discontinuous where the solution of the nonlinear system is discontinuous defining a solution based on singular Dirac delta distributions requires the product of a Dirac delta distribution with a discontinuous function. The authors define the necessary product of a Dirac delta distribution and a discontinuous function in the sense of the Volpert product [DLM95]. In this thesis we formally resolve the same issue in the sense of the Colombeau algebra (see Section 2.4.5). The authors also derive the singular solution as the limit of solutions to a regularized system where the discontinuity is smoothed out. The limiting definition is perhaps a useful intuition to keep in mind independent of how the product of distributions is formally defined. Furthermore, the authors discuss the linearization of the numerical solution and how the sensitivity of the numerical solution approximates the singular solution which we also demonstrate in Section 1.3.5.

Giles and Pierce [GP01] examine the analytic properties of the adjoint sensitivities for a system of conservation laws with a discontinuous solution. The authors derive the analytic adjoint equations for a design optimization objective for a variant of the Euler equations with the explicit treatment of a shock location via the Rankine-Hugoniot condition. They arrive at a different conclusion about the value of the adjoint solution on the shock curve than previous work, e.g. [IS96] and [CHS98]. Their derivation illustrates the necessity of an internal adjoint boundary condition owing to the disparity of characteristics entering and leaving the shock. We discuss the value of the adjoint solution on the shock curve in the sense of the Colombeau algebra in Section 5.4.2. The discrete adjoint approach we propose based on the coupling of the Rankine-Hugoniot condition dynamics for the shock location and sensitivities implicitly implements the internal adjoint boundary condition of Giles and Pierce [GP01] for the adjoint characteristics flowing out of the shock.

Ulbrich [Ulb02] discusses the analytic adjoint sensitivity analysis of discontinuous solutions of scalar conservation laws in the context of an optimal control formulation with a source term. In the considered setting the initial data and the source term depend on an infinite dimensional control parameter. A challenge arising from the source term is that the characteristics are not straight lines. Even though we do not explicitly treat the case of a source term or an infinite dimensional parameter the ideas presented in this thesis can be extended to the discrete analogue of the optimal control setting discussed by Ulbrich [Ulb03] via a standard AD-based discrete adjoint approach to optimal control based on the discretization of the control variable (see e.g. Walther [Wal07]).

Ulbrich [Ulb02] notes that the result of Bouchut and James [BJ99] stating that a discontinuous solution is differentiable in a weak topology yielding a Dirac delta distribution located at the discontinuity is not strong enough to obtain the derivative of the proposed optimal control integral objective without assuming additional structural information. To that end the author introduces a shift-differentiability calculus based on a shift-variation similar to that of Bressan and Marson [BM95b] that allows for a differentiability in the strong topology of L_1 and that holds for general solutions of the type of conservation laws the author considers. In this thesis we rather assume stronger structural information with a finite number of isolated shocks in a piecewise smooth solution. In our case the weak differentiability result is sufficient to obtain the derivatives of the type of integral objective we consider (see Equation (1.1.3)).

Ulbrich [Ulb02] also shows that with his proposed approach the shock structure does not have to be fixed and that shock generation is possible. In the analytic sensitivity analysis in Chapter 5 we do assume a fixed shock structure but preliminary experiments not presented in this thesis have shown that our methods can be extended to handle shock generation via the detection of shock formation based on the difference between neighboring grid points in relation to the grid spacing. The derivation of the adjoint solution in Ulbrich [Ulb02] is based on the stability of the adjoint equation with respect to its discontinuous coefficients. The author notes that the adjoint derivation also applies to hyperbolic systems and gives the correct shock sensitivities if the necessary stability properties of the shock and adjoint equations hold.

Ulbrich [Ulb03] extends the results further by showing how to obtain a unique reversible solution of the adjoint equation that is a linear transport equation with discontinuous coefficients based on taking a limit in the so-called averaged adjoint equation. The presented averaged adjoint equation formalism also presents an analytical framework for the numerical analysis of adjoint solutions that contain an averaging due to the discretization. Contrary to the approaches introduced in the methods of Ulbrich [Ulb02] and Ulbrich [Ulb03] the adjoint state does not contain an explicit sensitivity of the shock location.

Bardos and Pironneau [BP03] consider the discontinuous solution of the scalar Burgers equation and give a derivation of the tangent sensitivity in the sense of distributions where the tangent has a Dirac delta distribution along the shock curve. The authors consider the optimal control via an integral objective and present a theorem that states that the adjoint sensitivity according to the linearized equations and the gradient of the integral are identical even in the presence of shocks.

Giles [Gil03] discusses the discretization of the adjoint sensitivity equation of nonlinear hyperbolic PDEs when there is a shock in the solution. The author considers the discrete adjoint as obtained from reverse-mode AD that is based on the linearization of the discretization of the nonlinear PDE. The author demonstrates that the discrete sensitivities of a shock capturing scheme do not converge to the correct values in the presence of a shock via computational experiments with the Burgers equation. We give a similar demonstration to illustrate the nonconvergence in Section 1.3.8. The author then proposes to resolve the convergence issue by increasing the numerical smoothing of the shock resulting in an increasing number of grid points in the shock region. The proposed increase of numerical smoothing requires a change to the numerical integrator of the nonlinear PDE and results in a sublinear convergence rate.

Giles and Ulbrich [GU10a] formalize the results of Giles [Gil03]. The authors analyze the convergence of the discrete linearized equations for an unsteady one-dimensional hyperbolic equation with convex flux where the solution has a shock. They consider a modified Lax-Friedrichs discretization on a uniform grid where the numerical smoothing increases the number of points across the discretization of the discontinuity as the grid is refined. The authors prove pointwise convergence almost everywhere for the tangent sensitivities of the discretization and also for the tangent sensitivities of the discrete approximation of a linearized output functional, e.g. an integral objective. With Δx the spacing of grid points they show that for an appropriate choice of numerical viscosity the order of convergence is $O(\Delta x^{\alpha})$ as $\Delta x \to 0$ for $\alpha < 1$. The authors employ discrete stability estimates to show the convergence of the nonlinear and the linearized equations. They analyze the analytic viscous traveling wave that we also analyze in Section 6.1.1 and the corresponding discrete traveling wave (see Jennings [Jen74] for the correspondence of analytic viscous and discrete traveling waves in monotone conservative difference schemes). The authors perform an asymptotic approximation of the discrete solutions obtained with additional numerical smoothing that is more general than what we show for the viscous traveling wave in Section 6.1.1. The convergence results of Giles and Ulbrich [GU10a], therefore, serve as a formal proof of the assumptions we make in Section 6.1.1 for the specific case of the authors' proposed Lax-Friedrichs scheme.

Giles and Ulbrich [GU10b] continue the analysis of Giles and Ulbrich [GU10a] and prove the convergence of discrete approximations of linearized output functionals for Dirac delta initial perturbations and the pointwise convergence almost everywhere for the discrete adjoint sensitivities. The authors also show that their convergence results hold for general nonlinear initial data containing many shocks and also for the case where shocks merge or form at a later time.

Schäfer Aguilar, Schmitt, Ulbrich, and Moos [Sch+19] also analyze the convergence of discretization schemes for the adjoint equations arising from optimal control problems governed by entropy solutions of conservation laws. The authors address a similar problem to the problem addressed in this thesis and the previous two papers, namely, that of the nonconvergence of the adjoint sensitivity as motivated by Giles and Ulbrich [GU10a]. The authors propose a discrete adjoint for monotone difference schemes in conservation form that in contrast to the approach in Giles and Ulbrich [GU10a] does not require numerical viscosity of order $O(\Delta x^\alpha)$ with $\alpha < 1$ but instead gives a modification of the end data of the discrete adjoint sensitivities such that the solution of the standard discrete adjoint scheme with viscosity of order $O(\Delta x)$ converges to the correct values. The authors formally consider a setting of a scalar one-dimensional conservation law with strongly convex flux and an integral objective depending on the solution at some time $\bar{t} > 0$ that is similar to what we consider in out numerical analysis in Chapter 6. Their proposed approach consists of i) the detection of shock discontinuities in the solution at time $\bar{t}$ via the detection of shock areas where there are changes between grid points of order $O(\sqrt{\Delta x})$ as $\Delta x \to 0$ and choosing the shock approximation as the center of those regions and ii) manually setting the adjoint sensitivity data at $\bar{t}$ based on the finite difference approximation of the adjoint at the shock by approximating the limiting points to the shock at location x_s via $x_s \pm \Delta x^{1/3}$ instead of seeding the discrete adjoint sensitivity of the discrete approximation of the integral. For example for a smooth solution u the adjoint sensitivity $\bar{u}$ of an integral objective is

$$\bar{u}(\cdot) = \mathrm{d}_{u(\cdot)} \int_{-\infty}^{\infty} \varphi(u(x)) \mathrm{d}x = \varphi'(u(\cdot)) \ .$$

But if u is discontinuous at x_s the adjoint at $x = x_s$ is not defined in the classical sense. We can consider the definition of the adjoint at $x = x_s$ via the the limit of a central difference around the discontinuity, i.e.

$$\bar{u}(x_s) = \frac{\varphi(u(x_s+)) - \varphi(u(x_s-))}{u(x_s+) - u(x_s-)} \ .$$

Such a definition is motivated by the limit of the integral on a domain around the discontinuity of the adjoint of a smoothed version of the discontinuous solution u (see Section 5.4.2). We consider the idea of a central difference around the shock in the context of the tangent sensitivity in Section 1.4.1 and Section 7.3.2. Schäfer Aguilar, Schmitt, Ulbrich, and Moos [Sch+19] propose to approximate the central difference by a finite difference around the approximated shock location and use the finite difference approximation to replace the discrete adjoint in the shock region where the discrete approximation of the integral adjoint is not convergent according to Giles [Gil03]. As in our approach, the width of the shock region plays a critical role. The authors propose that using $x_s \pm \Delta x^{1/3}$ should be a safe choice to cover the shock region and perform numerical experiments where the empirical order of convergence of the adjoint sensitivity in the sense of the L_1 norm is observed to be $O(\Delta x^{1/2})$ as $\Delta x \to 0$.

Herty, Hüser, Naumann, Schilden, and Schröder [Her+21] describe the AD approach for the numerical approximation of the generalized tangent of Bressan and Marson [BM95b] developed in this thesis. The authors present numerical results for the one-dimensional Burgers and Euler equations, based on a state-of-the-art CFD solver, that are not contained in this thesis and show that using the method for the computation of the shock sensitivities presented in this thesis yields the correct results in contrast to the naive application of black-box AD.

1.3 Motivations for Shock Sensitivities

In this section we aim to motivate the fundamental problem solved in this thesis; the numerical approximation of the sensitivities of shock locations in solutions of conservation laws. The goal is to give an informal introduction to the ideas that are treated in a more formal manner within the thesis. The introduction uses a driving example that is also used for explanations throughout the rest of the thesis.

As a motivational setting we consider the PDE constrained optimization problem

$$\min_{u,p} \Phi(u(p)) \text{ such that Equations (1.1.1)-(1.1.2) are satisfied}$$

where Φ is defined as the integral objective in Equation (1.1.3). Gradient-based optimization methods directly motivate the need for the gradient $\mathrm{d}_p\Phi(u(p)) \in \mathbb{R}^d$.

One approach to PDE constrained optimization for problems where the PDE solution is found by numerical methods is to discretize-then-optimize [Bet10, Chapter 4.12]. In optimal control problems where the optimization variable is infinite dimensional the discretize-then-optimize approach reduces an infinite dimensional optimization problem to a finite dimensional problem through discretization. In our setting we already consider a finite dimension parameter $p \in \mathbb{R}^d$ so here the idea is just to optimize the numerical approximation of the PDE and objective function instead of the analytical problem. For some optimization problems and numerical integrators it is possible to prove the convergence of the optimum of the discrete problem to the optimum of the analytical problem as the discretization is refined (see e.g. Dontchev, Hager, and Veliov [DHV00]).

When following the discretize-then-optimize approach a local numerical optimization method will minimize the norm of the gradient of the discretization instead of the analytical gradient. If the discretization is differentiable then the gradient of the discretization can typically be obtained by discrete adjoint methods made available through AD tools such as e.g. `dco/c++` [LLN11], ADOL-C [GJU96] and Tapenade [HP13].

In this section we show how discrete sensitivity methods fail to produce correct gradients given the presence of shock discontinuities in the solution of the conservation law.

1.3.1 Example

We consider a classical Riemann problem instance [LeV92, Chapter 3.2] given by Equations (1.1.1)-(1.1.2) with the flux function

$$f(u) = \frac{1}{2}u^2$$

resulting in Burgers equation with the following parametric initial data

$$u_0(x,p) = (1+p)H(-x)$$

where H is the Heaviside step function. The PDE solution for the given inital data is

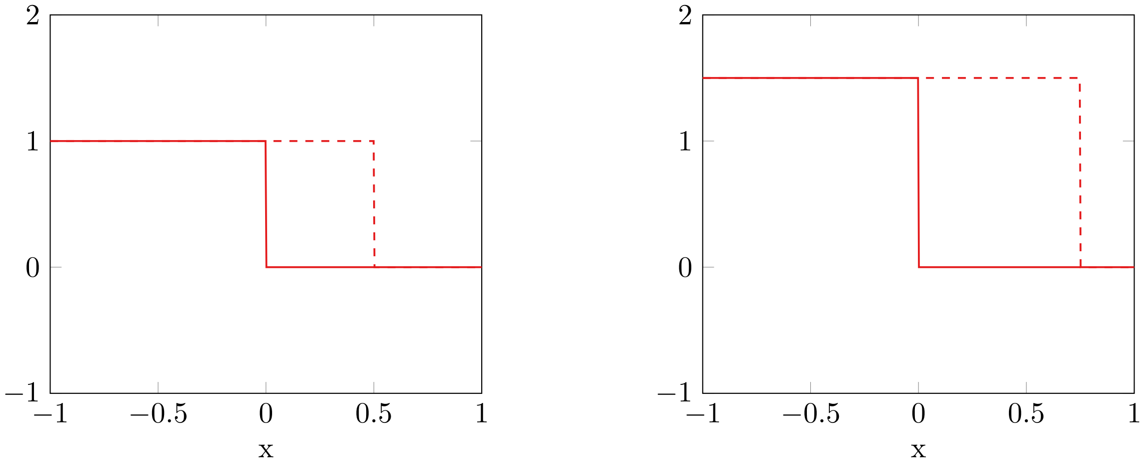
$$u(t,x,p) = (1+p)H(\xi(t,p) - x)$$

where

$$\xi(t,p) = \frac{1}{2}(1+p)t$$

is the location of the discontinuity. Technically, the solution u is a weak solution (see Section 3.1.1 for the definition of weak solution).

The initial data has a discontinuity at $x = 0$ that is a shock discontinuity (see Section 3.1.3 for the definition of shock discontinuity). For $p > -1$ the shock travels to the right over time and $\xi(t,p)$ is the shock location at time t. Figure 1.1 shows the solution on the domain $x \in [-1,1]$ at $t = 0$ (solid) and $t = 1$ (dashed) for two different values of p. For higher values of p the shock travels further to the right in the same amount of time.



(a) $p = 0$

(b) $p = \frac{1}{2}$

Figure 1.1: Weak solution $u(t,x,p)$ at $t = 0$ (solid), $t = 1$ (dashed)

For the given example problem instance we consider the least-squares objective function

$$\Phi(p) \equiv \frac{1}{2}\int_{-1}^{1} \Big(u(t=1,x,p) - 1\Big)^2 \mathrm{d}x \tag{1.3.1}$$

as a concrete example of an integral objective function. The minimum of the corresponding PDE constrained optimization problem with the additional constraint that $p \in (-1,1)$ is at

$$p^\star \equiv \operatorname*{arg\,min}_{p\in(-1,1)} \Phi(p) = \frac{2}{\sqrt{3}} - 1 \, .$$

In the next section we show that the gradient of Φ with respect to p for the example as described above is

$$\mathrm{d}_p\Phi(p) = \frac{3}{4}p^2 + \frac{3}{2}p - \frac{1}{4} \, .$$

For the domain of $p \in (-1, 1)$ the objective function $\Phi(p)$ is strictly convex. The first-order optimality condition $\mathrm{d}_p\Phi(p) = 0$ is necessary and sufficient for defining the unique minimum. By solving the first-order optimality condition for p we get

$$p = \pm\frac{2}{\sqrt{3}} - 1 \, .$$

Since $p > -1$ we get the minimum $p^\star$ as given above. The other critical point of $\mathrm{d}_p\Phi(p)$ is actually a local maximum. Figure 1.2 shows the function Φ on the domain $p \in (-1, 1)$.

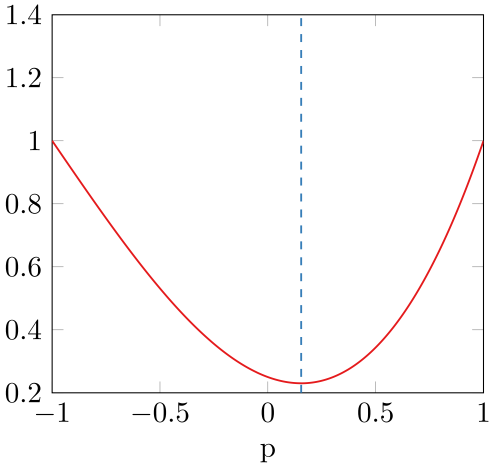


Figure 1.2: Objective function Φ (solid red) with minimum at $p = p^\star$ (dashed blue)

1.3.2 Analytic Derivative of the Weak Solution

We now describe how the sensitivity of the shock discontinuity at $x = \xi(t, p)$ with respect to the parameter p causes the analytic derivative of the weak solution to be singular. The classical definition of a derivative is based on taking the limit of the sensitivity with respect to a small perturbation of the parameter of interest. If we consider such a limit for the weak solution u of the parametric Riemann example described above we get the derivative with respect to p at $p = p_0$ as

$$\mathrm{d}_p u(t, x, p_0) = \lim_{\dot{p}\to 0} \frac{1}{\dot{p}}\big(u(t, x, p_0 + \dot{p}) - u(t, x, p_0)\big) \, .$$

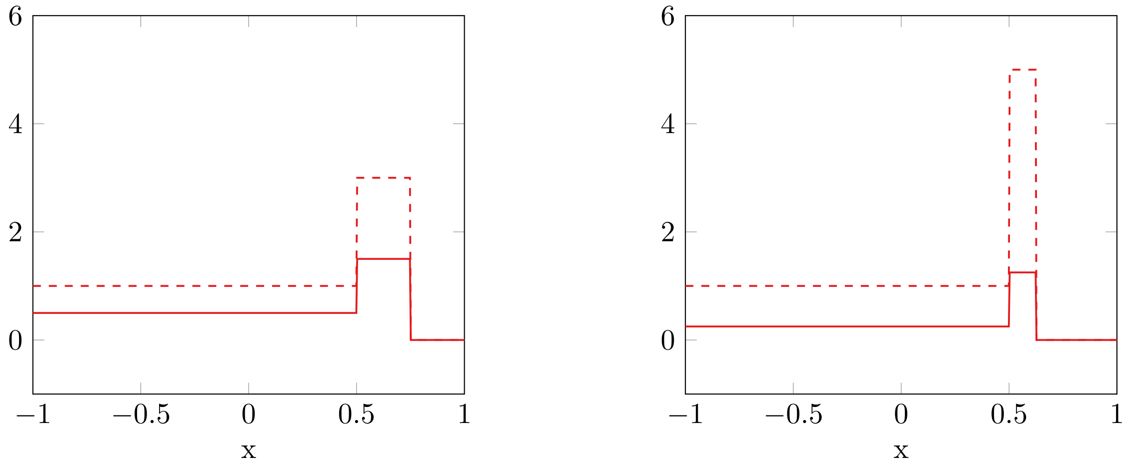
For $\dot{p} > 0$ we have the difference

$$\begin{aligned} u(t, x, p_0 + \dot{p}) - u(t, x, p_0) &= (1 + p_0 + \dot{p})H(\xi(t, p_0 + \dot{p}) - x) - (1 + p_0)H(\xi(t, p_0) - x) \\ &= \begin{cases} \dot{p} & \text{if } x < \xi(t, p_0) \\ 1 + p_0 + \dot{p} & \text{if } \xi(t, p_0) < x < \xi(t, p_0 + \dot{p}) \\ 0 & \text{if } \xi(t, p_0 + \dot{p}) < x \end{cases} . \end{aligned}$$

We are not concerned with the exact definition of the difference at the points $x = \xi(t, p_0)$ and $x = \xi(t, p_0 + \dot{p})$. Figure 1.3 shows the difference and the difference divided by $\dot{p}$ on the domain $x \in [-1, 1]$ for $p = p_0 = 0$ and two different values of $\dot{p}$.

We first note that as $\dot{p} \to 0$ the domain of the middle part of the difference has the width

$$\xi(t, p_0 + \dot{p}) - \xi(t, p_0) = \frac{1}{2}(1 + p_0)t - \frac{1}{2}(1 + p_0 + \dot{p})t \qquad (1.3.2)$$



(a) $\dot{p} = \frac{1}{2}$

(b) $\dot{p} = \frac{1}{4}$

Figure 1.3: Difference $u(t, x, \dot{p}) - u(t, x, 0)$ (solid), divided difference $\frac{1}{\dot{p}}\big(u(t, x, \dot{p}) - u(t, x, 0)\big)$ (dashed) at $t = 1$

$$
\begin{aligned}
&= \frac{1}{2} t \dot{p} && (1.3.3)\\
&= \mathrm{d}_p \xi(t, p_0) \dot{p}\ . && (1.3.4)
\end{aligned}
$$

In the case of a shock location that is a differentiable nonlinear function of p_0 the above equation does not generally hold. Instead, the sensitivity of the shock location $\mathrm{d}_p \xi(t, p_0)\dot{p}$ is a first-order approximation of the difference in shock locations $\xi(t, p_0 + \dot{p}) - \xi(t, p_0)$ according to the Taylor expansion. We can use the sensitivity of the shock location to approximate the additional distance that the shock travels due to a perturbed value of p.

The difference on the domain of the middle part divided by $\dot{p}$ as it occurs in the limit definition of the derivative is singular in the limit

$$
\lim_{\dot{p} \to 0} \frac{1}{\dot{p}} (1 + p_0 + \dot{p}) = \infty\ .
$$

The derivative of the weak solution is, hence, not bounded and not an integrable function, i.e. not in L_1.

The limit of the integral over the middle part of the divided difference that becomes singular is

$$
\begin{aligned}
&\lim_{\dot{p} \to 0} \int_{\xi(t,p_0)}^{\xi(t,p_0+\dot{p})} \frac{1}{\dot{p}} (1 + p_0 + \dot{p}) \mathrm{d}x \\
&= \lim_{\dot{p} \to 0} \left(1 + \frac{1 + p_0}{\dot{p}}\right) \big(\xi(t, p_0 + \dot{p}) - \xi(t, p_0)\big) && \text{(integration of a constant)}\\
&= \lim_{\dot{p} \to 0} \left(1 + \frac{1 + p_0}{\dot{p}}\right) \mathrm{d}_p \xi(t, p_0) \dot{p} && \text{(Equation (1.3.2))}\\
&= \mathrm{d}_p \xi(t, p_0)(1 + p_0)\ . && (\lim_{x \to 0} x = 0, \lim_{x \to 0} 1 = 1)
\end{aligned}
$$

In the formalism of distributions (see Section 2.4 for a definition of distributions) the singularity can be expressed by a Dirac delta distribution multiplied by the the sensitivity of the shock location with respect to the parameter p times $u(\xi-) - u(\xi+) = 1 + p_0$.

The analytic derivative of the weak solution u in the sense of distributions at $p = p_0$ is

$$\mathrm{d}_p u(t,x,p_0) = H(\xi(t,p_0) - x) + \delta(x - \xi(t,p_0))\mathrm{d}_p\xi(t,p_0)(1+p_0) \tag{1.3.5}$$

where δ is the Dirac delta distribution that is informally defined by its integral

$$\int_{-\infty}^{\infty} \delta(x)\mathrm{d}x = 1\ .$$

1.3.3 Analytic Gradient of the Integral Objective

We now describe how the sensitivity of the shock discontinuity at $x = \xi(t,p_0)$ with respect to the parameter p contributes to the gradient of the integral objective function

$$\Phi(p) = \int_{-1}^{1} \varphi(u(t,x,p))\mathrm{d}x$$

at $p = p_0$. We split the domain $(-1,1)$ at the shock location and get

$$\mathrm{d}_p\Phi(p_0) = \mathrm{d}_p \int_{-1}^{\xi(t,p_0)} \varphi(u(t,x,p_0))\mathrm{d}x + \mathrm{d}_p \int_{\xi(t,p_0)}^{1} \varphi(u(t,x,p_0))\mathrm{d}x\ .$$

By differentiation of the integral boundaries and the chain rule we have

$$\begin{aligned}\mathrm{d}_p\Phi(p_0) = &\int_{-1}^{\xi(t,p_0)} \varphi'(u(t,x,p_0))\mathrm{d}_p u(t,x,p_0)\mathrm{d}x + \int_{\xi(t,p_0)}^{1} \varphi'(u(t,x,p_0))\mathrm{d}_p u(t,x,p_0)\mathrm{d}x && (1.3.6)\\ &+ \mathrm{d}_p\xi(t,p_0)\big(\varphi(u(t,\xi(t,p_0)-,p_0)) - \varphi(u(t,\xi(t,p_0)+,p_0))\big)\ . && (1.3.7)\end{aligned}$$

Specifically, for the least-squares example for $t = 1$ we get

$$\begin{aligned}\mathrm{d}_p\Phi(p_0) &= \int_{-1}^{\frac{1}{2}(1+p_0)} \mathrm{d}_p\frac{1}{2}(1+p_0-1)^2\mathrm{d}x + \int_{\frac{1}{2}(1+p_0)}^{1} \mathrm{d}_p\frac{1}{2}(0-1)^2\mathrm{d}x\\ &\quad + \Big(\mathrm{d}_p\frac{1}{2}(1+p_0)\Big)\Big(\frac{1}{2}(1+p_0-1)^2 - \frac{1}{2}(0-1)^2\Big)\\ &= \frac{3}{4}{p_0}^2 + \frac{3}{2}p_0 - \frac{1}{4}\ .\end{aligned}$$

We previously used the gradient $\mathrm{d}_p\Phi(p)$ to find the minimum $p^\star$ of the example optimization problem.

For the example it is also possible to find the gradient by solving the integral definition Φ as a polynomial expression of p and differentiating the polynomial. The more general Equations (1.3.6)-(1.3.7) are instructive because they show how the sensitivity of the shock location ξ with respect to p plays into the value of the gradient. Since u is discontinuous at $x = \xi$ the difference $\varphi(u(\xi-)) - \varphi(u(\xi+))$ is nonzero and the last term of Equations (1.3.6)-(1.3.7) is a significant contribution to the gradient.

In the next sections we show that the contribution of the shock sensitivity to the gradient of the integral objective is not correctly approximated in the standard discrete sensitivity analysis applied to shock capturing numerical methods.

1.3.4 Discrete Solution

We now introduce an example of a shock capturing numerical integrator for the conservation law initial value problem in Equations (1.1.1)-(1.1.2) and show how it applies to the Riemann example problem. Shock capturing methods do not explicitly treat shock discontinuities in a different way than the smooth parts of the solution (as opposed to shock fitting methods). As a result the solution of a shock capturing method is usually not sharp at the shock discontinuity but rather approximates the shock over multiple discrete points in the numerical solution and may even contain oscillations near the shock [LeV92, Chapter 1.4].

The goal of this section is to use the discrete tangent or adjoint sensitivities of the shock capturing method to approximate the analytic gradient of the integral objective derived in the previous section. For the purpose of demonstration we use the Lax-Friedrichs method. The issues that we see with the convergence of the discrete sensitivities of the Lax-Friedrichs method also occur with other shock capturing methods.

The Lax-Friedrichs time step for the numerical approximation $U_i^j \approx u(t_j, x_i)$ on a uniform space grid with $x_{i+1} - x_i = \Delta x$ for $i \in \{0, \ldots, n-1\}$ and a time grid with $t_{j+1} - t_j = \Delta t_j$ for $j \in \{0, \ldots, m-1\}$ is

$$U_i^{j+1} := \frac{1}{2}\big(U_{i-1}^j + U_{i+1}^j\big) - \frac{\Delta t_j}{2\Delta x}\big(f(U_{i+1}^j) + f(U_{i-1}^j)\big)\ .$$

The Lax-Friedrichs method is similar to an explicit Euler discretization. The second term on the right-hand side contains the discretization of the $\partial_x f(u)$ term in the conservation law. But unlike the explicit Euler time step the Lax-Friedrichs method averages two grid points of the previous time step in the first term on the right-hand side. The averaging of the solution at different space points has a smoothing effect, i.e. the averaging introduces so-called numerical viscosity. The amount of viscosity in the discrete solution depends on the size of the time steps Δt_j relative to the distance of the space grid points Δx.

We use the time step

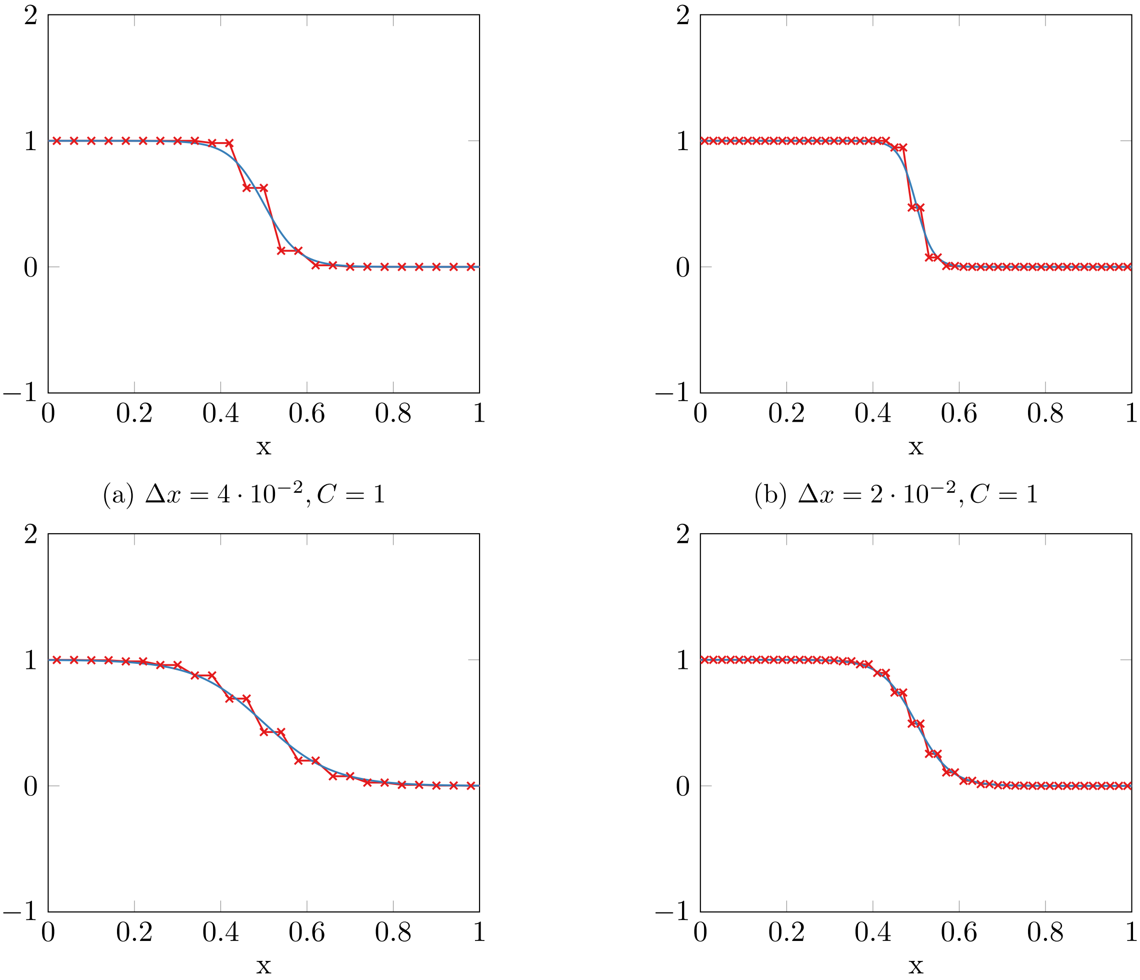
$$\Delta t_j := \frac{C}{\max_i |f'(U_i^j)|}\Delta x$$

based on the CFL condition with the CFL number C such that $0 < C \leq 1$ (see Section 3.2.4 for the definition of the CFL condition). For all C satisfying the CFL condition we have $\Delta t_j = O(\Delta x)$ as $\Delta x \to 0$ for all $j \in \{0, \ldots, m-1\}$ and a choice of the factor between time and space discretization that guarantees that the numerical method is stable.

Figure 1.4 shows the discrete Lax-Friedrichs solution of the Riemann example in red for different values of Δx and C. For the same value of C the smaller Δx the smaller the region in which the shock is smoothed out. It is crucial to observe that the number of discrete grid points on which the shock is smoothed out is constant for different values of Δx. For the same value of Δx the lower C the wider the region in which the shock is smoothed out and also the higher the number of grid points in the shock region, i.e. a lower value of C leads to more numerical viscosity for the same value of Δx.

At this point we should briefly note that a discontinuous solution approximated using a method with numerical viscosity does not necessarily have a smoothing region that has a width that is linear in Δx. For example discontinuous initial data transported by a linear flux function only approximates the discontinuous analytic solution with an error that is $O(\sqrt{\Delta x})$ as $\Delta x \to 0$

because the smoothing region also has that width [LeV92, Chapter 11.2]. Intuitively, the reason for the wider shock region in that case is the absence of a compressive shock at the discontinuity. The diffusion is not counteracted by compression and is relatively stronger. The square root rate represents the best possible convergence rate for general initial data even for nonlinear flux functions [Sab97]. However, under reasonable assumptions the approximation of a solution with shock discontinuities and a nonlinear flux function converges to the analytic solution at a linear rate $O(\Delta x)$ as $\Delta x \to 0$ which implies that the shock region also has linear width [TZ97; TT97].



(a) $\Delta x = 4 \cdot 10^{-2}, C = 1$

(b) $\Delta x = 2 \cdot 10^{-2}, C = 1$

(c) $\Delta x = 4 \cdot 10^{-2}, C = \frac{1}{2}$

(d) $\Delta x = 2 \cdot 10^{-2}, C = \frac{1}{2}$

Figure 1.4: Lax-Friedrichs solution (red), analytic viscosity solution (blue), at $t = 1$ with $p = 0$

In order to better understand the behavior of the numerical viscosity it is helpful to realize that the Lax-Friedrichs discretization can be derived from the finite differencing of the PDE

$$\partial_t u(t,x) + \partial_x f(u(t,x)) = \epsilon \partial^2_{xx} u(t,x) \ . \tag{1.3.8}$$

with $\epsilon = \Delta x^2/(2\Delta t)$ (see Section 3.2.8 for the derivation). The left-hand side is identical to the conservation law introduced in Equation (1.1.2). But the right-hand side is analogous to that of the heat equation

$$\partial_t u(t,x) = \epsilon \partial^2_{xx} u(t,x) \ .$$

The right-hand side of the equation introduces diffusion according to the coefficient $\epsilon > 0$. The diffusion produces smoothing or viscosity in the solution. Since $\epsilon = O(\Delta x)$ as $\Delta x \to 0$ the amount of viscosity depends on the numerical discretization and goes down linearly as the grid is refined.

The viscous Burgers equation, i.e. Equation (1.3.8) with $f(u) = u^2/2$ has an analytic solution for a sideways traveling wave analogous to the solution of the Riemann example we introduced for the inviscid Burgers Equation. The analytic viscosity solution analogous to the parametric Riemann solution is

$$u_{\text{smooth}}(t, x, p) = \frac{1+p}{2} + \frac{1+p}{2} \tanh\left(\frac{\Delta t}{2\Delta x^2}(1+p)(\xi(t,p) - x)\right)$$

with $\xi(t, p)$ the same as in the inviscid Riemann example (see Section 3.3.2 for the proof). To understand the definition of u_{smooth} it is helpful to realize that the function

$$h(x) = \frac{1}{2} + \frac{1}{2}\tanh(ax)$$

is a smooth approximation of the Heaviside step function where a is the scaling factor that determines smoothness. For a comparison of the discrete Lax-Friedrichs solution with u_{smooth} in the context of the Riemann example it is possible to use constant Δt since

$$\Delta t_j := \frac{C}{\max_i |f'(U_i^j)|}\Delta x = \frac{C}{\max_i |U_i^j|}\Delta x = \frac{C}{1+p}\Delta x = \Delta t$$

for all time steps $j \in \{0, \ldots, m-1\}$.

Figure 1.4 shows the analytic viscosity solutions u_{smooth} in blue that correspond to the Lax-Friedrichs discretizations in terms of the numerical viscosity via the Δx, Δt used in the viscosity coefficient ϵ. It is important to note that the Lax-Friedrichs discretization is not as smooth as u_{smooth} in the shock region. The finite difference derivative with respect to x of the Lax-Friedrichs discretization may be much larger or smaller than the derivative of u_{smooth} at the same point.

1.3.5 Discrete Tangent Sensitivity

We now consider the result of a discrete sensitivity analysis of the Lax-Friedrichs integrator for the Riemann example. The discrete sensitivity analysis we introduce here is a tangent sensitivity analysis in the sense that we are considering directional derivatives of the discretization. In the context of the example there is no computational efficiency reason for performing an adjoint sensitivity analysis as the parameter p is one-dimensional. We do not consider adjoint sensitivities in the introduction in order to keep things simple and because the convergence issues that arise in adjoint methods are the same as with the tangent sensitivities.

A possible discrete approximation of the example least-squares objective function in Equation (1.3.1) based on the discrete Lax-Friedrichs solution U is by midpoint rule

$$\Phi_{\text{discrete}} := \frac{1}{2}\sum_{i=1}^{n}(U_i^m - 1)^2\Delta x \; .$$

Directional differentiation of Φ_{discrete} with respect to p yields the discrete tangent sensitivity

$$\dot{\Phi}_{\text{discrete}} \equiv \mathrm{d}_p\Phi_{\text{discrete}} \cdot \dot{p} = \sum_{i=1}^{n}(U_i^m - 1)\dot{U}_i^m\Delta x$$

where

$$\dot{U}_i^m \equiv \mathrm{d}_p U_i^m \cdot \dot{p}$$

is the discrete tangent sensitivity of the Lax-Friedrichs method. In practice the discrete tangent of a numerical integrator can often be obtained by an AD tool without having to differentiate the algorithm by hand.

In the context of the Lax-Friedrichs method the differentiation of the integrator is straight forward since no complex mathematical expressions or multiple assignments per step are involved. The time step of the discrete tangent is obtained by directional differentiation of the Lax-Friedrichs time step

$$\dot{U}_i^{j+1} := \frac{1}{2}\big(\dot{U}_{i-1}^j + \dot{U}_{i+1}^j\big) - \frac{\Delta t}{2\Delta x}\big(f'(U_{i+1}^j)\dot{U}_{i+1}^j + f'(U_{i-1}^j)\dot{U}_{i-1}^j\big) \ .$$

The initial data for the discrete tangent is a discretization of the tangent of the initial data u_0 of the example problem instance, i.e.

$$\dot{U}_i^0 = \mathrm{d}_p u_0(x, p_0) \cdot \dot{p} = H(-x_i) \cdot \dot{p} \ .$$

Figure 1.5 shows the discrete Lax-Friedrichs tangent of the Riemann example in red for different values of Δx and C. We observe that in the shock region the tangent contains a smooth impulse similar to the box impulse in the middle part of the graph in Figure 1.3. The smooth impulse is the numerical approximation of the singular part of the analytical derivative $\mathrm{d}_p u$. We note that the smaller Δx the higher the numerical values in the impulse. Furthermore, we again observe analogous phenomena regarding the grid points and approximation values in the shock region as we observed for the discrete approximation of the weak solution in Figure 1.4.

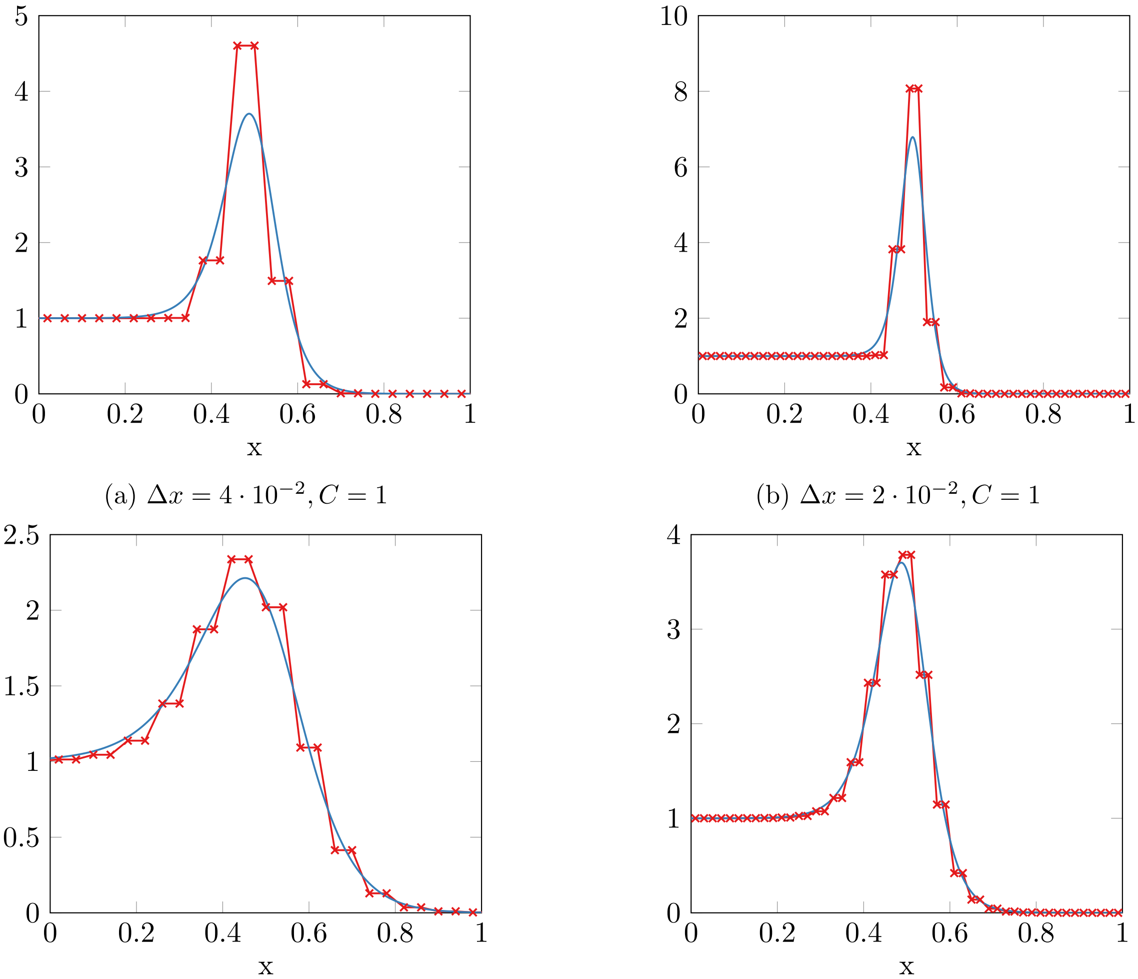
In the previous section we analyzed the numerical viscosity of the Lax-Friedrichs solution for the Burgers equation with the help of the analytical solution of the viscous Burgers equation. Similary, it is helpful to consider the tangent sensitivity $\dot{u}_{\text{smooth}}(t, x, p_0) \equiv \mathrm{d}_p u_{\text{smooth}}(t, x, p_0)\dot{p}$ of the analytical viscosity solution u_{smooth} to analyze the discrete tangent. According to the chain rule we have

$$\begin{aligned}\dot{u}_{\text{smooth}}(t, x, p_0) = {} & \left(\frac{1}{2} + \frac{1}{2}\tanh\left(\frac{\Delta t}{2\Delta x^2}(1 + p_0)(\xi(t, p_0) - x)\right)\right)\dot{p} + \\ & \frac{1 + p_0}{2}\operatorname{sech}\left(\frac{\Delta t}{2\Delta x^2}(1 + p_0)(\xi(t, p_0) - x)\right)^2 \cdot \\ & \frac{\Delta t}{2\Delta x^2}\big((\xi(t, p_0) - x) + (1 + p_0)\mathrm{d}_p\xi(t, p_0)\big)\dot{p} \ .\end{aligned}$$

Figure 1.5 shows the analytic viscosity solution tangents $\dot{u}_{\text{smooth}}$ in blue that correspond to the Lax-Friedrichs discretizations in terms of the numerical viscosity via the Δx, Δt used in the viscosity coefficient ϵ. We observe that like to the relationship between the discrete solution and the analytic viscosity solution the discrete tangent and the corresponding analytic viscosity tangent behave similarly at least empirically.

We now go on to analyze how the analytic viscosity tangent relates to the inviscid analytic tangent of the weak solution. Based on Equation (1.3.5) the analytic tangent $\dot{u}(t, x, p_0) \equiv \mathrm{d}_p u(t, x, p_0)\dot{p}$ in the sense of distributions is

$$\dot{u}(t, x, p_0) = H(\xi(t, p_0) - x)\dot{p} + \delta(x - \xi(t, p_0))\mathrm{d}_p\xi(t, p_0)(1 + p_0)\dot{p} \ . \tag{1.3.9}$$



(a) $\Delta x = 4 \cdot 10^{-2}, C = 1$

(b) $\Delta x = 2 \cdot 10^{-2}, C = 1$

(c) $\Delta x = 4 \cdot 10^{-2}, C = \frac{1}{2}$

(d) $\Delta x = 2 \cdot 10^{-2}, C = \frac{1}{2}$

Figure 1.5: Lax-Friedrichs tangent (red), analytic viscosity tangent (blue), at $t = 1$ with $p = p_0 = 0, \dot{p} = 1$

The term in $\dot{u}_{\text{smooth}}$ involving $\tanh(\cdot)$ is again a smooth approximation of the Heaviside step function in $\dot{u}$ similar to what we saw for u_{smooth}. The $\operatorname{sech}(\cdot)^2$ term is a smooth approximation of the Dirac delta distribution.

The convergence of the function $\dot{u}_{\text{smooth}}$ to the distribution $\dot{u}$ (that is not a function in the classical sense) needs to be understood in a weak sense (see Section 2.4.4 for a definition of weak convergence). We have the weak convergence of $\dot{u}_{\text{smooth}}$ to $\dot{u}$ since

$$\Big| \int_{-\infty}^{\infty} \dot{u}(t, x, p_0)\phi(x)\mathrm{d}x - \int_{-\infty}^{\infty} \dot{u}_{\text{smooth}}(t, x, p_0)\phi(x)\mathrm{d}x \Big| \to 0$$

with $\Delta t = O(\Delta x)$ as $\Delta x \to 0$ for all smooth test functions ϕ with compact support (see Section 2.4.1 for a definition of test functions).

To better understand how $\dot{u}_{\text{smooth}}$ approximates the Dirac delta distribution we consider how the $\operatorname{sech}(\cdot)^2$ term behaves as $\Delta x \to 0$. We have $\Delta t = O(\Delta x)$ as $\Delta x \to 0$ and so the last term of $\dot{u}_{\text{smooth}}$ at $x = \frac{1}{2}$ is

$$O\big(\Delta x^{-1}\big) \operatorname{sech}(0)^2 = O\big(\Delta x^{-1}\big)$$

at $t = 1, p = p_0 = 0, \dot{p} = 1$ as $\Delta x \to 0$. The discrete approximation of an integral of the smoothed impulse that has only a single grid point at $x = \frac{1}{2}$ and grid spacing Δx is thus $O(1)$ as it should be in order to correctly approximate the integral of a Dirac delta distribution.

In the next sections we analyze different questions about the convergence of the discrete sensitivity methods as $\Delta x \to 0$ and also the convergence of the linear approximations obtained based on the tangent sensitivities as $\dot{p} \to 0$.

1.3.6 Weak Convergence of the Discrete Tangent of the Weak Solution

The first question we consider regarding convergence is whether the discrete tangent sensitivity weakly converges to the analytic tangent sensitivity including the singular Dirac delta distribution due to the shock sensitivity as $\Delta x \to 0$. The weak convergence of the discrete Lax-Friedrichs tangent $\dot{U}$ to the analytic tangent $\dot{u}$ implies that

$$\Big| \int_{-1}^{1} \dot{u}(t_m, x, p_0) \mathrm{d}x - \sum_{i=1}^{n} \dot{U}_i^m \Delta x \Big| \to 0$$

as $\Delta x \to 0$. Figure 1.6 shows how the error above behaves for different values of Δx at $t_m = 1$ and with $p = p_0 = 0, \dot{p} = 1$ where we we have

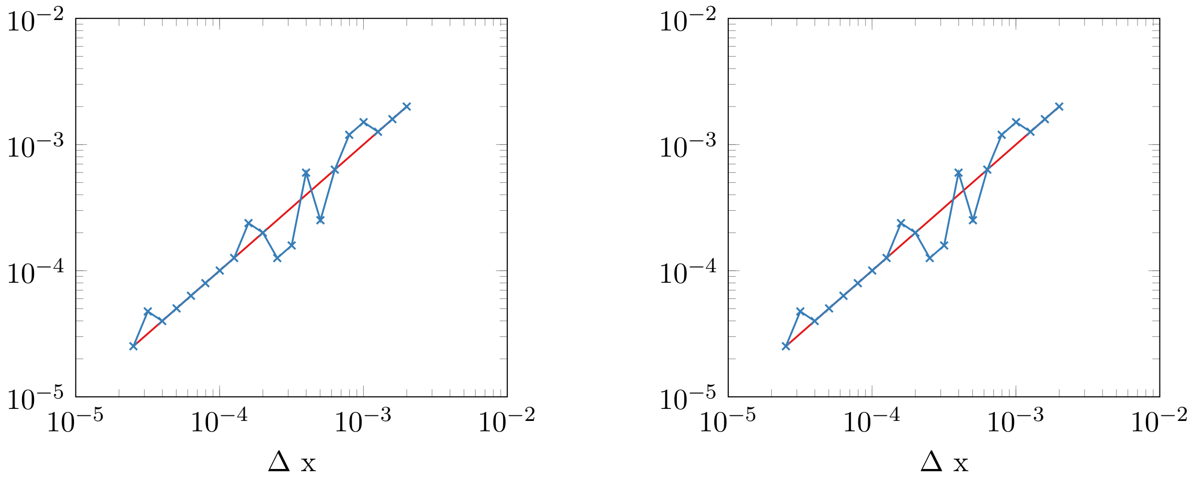
$$\begin{aligned}
&\int_{-1}^{1} \dot{u}(t_m, x, p_0) \mathrm{d}x \\
&= \int_{-1}^{\xi(t,p_0)} \dot{p} \mathrm{d}x + \int_{-1}^{1} \delta(x - \xi(t, p_0)) \mathrm{d}_p \xi(t, p_0)(1 + p_0)\dot{p} \mathrm{d}x && \text{(Equation (1.3.9))} \\
&= \mathrm{d}_p \xi(t, p_0) + \xi(t, p_0) + 1 && \text{(substitution of } p_0, \dot{p}) \\
&= \frac{1}{2} + \frac{1}{2}(1 + p_0) + 1 = 2\ . && \text{(substitution of } \mathrm{d}_p \xi, \xi)
\end{aligned}$$

By checking whether the error between the integral over the domain $x \in [-1, 1]$ and its discrete approximation converges we are checking a necessary but not sufficient condition for weak convergence.

Furthermore, in the following we do not check the true asymptotic convergence as $\Delta x \to 0$ but instead we only observe how the error behaves empirically for geometrically decreasing values of $\Delta x > 0$. Therefore, the analysis done here is by no means a formal proof of convergence but in combination with Figure 1.5 should plausibly convince us that the discrete tangent of the Lax-Friedrichs solution is probably weakly convergent to the analytic tangent in practice.

Figure 1.6 shows the error between the integral of the analytic tangent and the integral of the discrete approximation for two different values of the CFL number C, i.e. two different amounts of numerical viscosity. The graphs for the different CFL numbers are virtually indistinguishable. Both go down at a linear rate except that the discrete solution and, hence, the discrete sum approximation does not smoothly vary as a function of changing the grid spacing Δx. The grid effect occurs because the initial data is discontinuous and changes in a nonsmooth way as grid points move across the discontinuity. Smoothing the initial data according to the numerical viscosity would result in a smooth error graph.

In Figure 1.5 we make the observation that the discrete tangent approximates the singular part of the analytic tangent at the shock. We also observe that the discrete tangent does not match



(a) $C = \frac{1}{2}$

(b) $C = 1$

Figure 1.6: Weak convergence error of the discrete tangent $\left|\int_{-1}^{1} \dot{u}(t_m, x, p)\mathrm{d}x - \sum_{i=1}^{n} \dot{U}_i^m \Delta x\right|$ (blue), $O(\Delta x)$ (red), at $t = 1$ with $p = p_0 = 0, \dot{p} = 1$

the pointwise values of the tangent of the analytic viscosity solution. But in Figure 1.6 we make the observation that the integral correctly approximates the integral of the singular part and the integral does not change for different amounts of numerical viscosity (c.f. discrete conservation in Section 3.2.2).

We empirically observe the weak convergence of the discrete tangent to the analytic tangent and so far we have not yet run into an issue arising from the discontinuity in the solution of the conservation law. After informally understanding the behavior of the discrete tangent in a weak sense we now move on to the first challenge that arises due to the discontinuity of the solution.

1.3.7 Taylor Convergence of the Discrete Tangent of the Weak Solution

The second question we consider regarding convergence is whether the linearization via the discrete tangent converges to the finite difference in the sense of a first-order Taylor expansion in the L_1 norm, i.e. do we have

$$\|u(t_m, \cdot, p_0 + \dot{p}) - u(t_m, \cdot, p_0) - \dot{U}^m(\cdot)\|_{L_1([-1,1])} = O(|\dot{p}|^2)$$

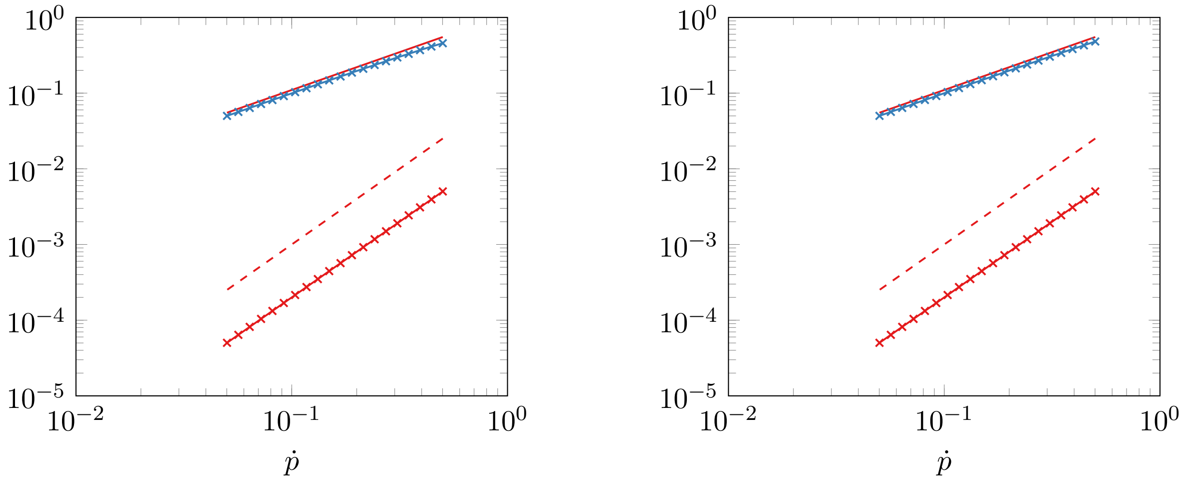
where $\dot{U}^m(\cdot)$ is a piecewise constant interpolation of $\dot{U}^m$ with $\Delta x = O(|\dot{p}|^2)$ as $\dot{p} \to 0$.

We estimate the above L_1 norm error by

$$\sum_{i=1}^{n} |u(t_m, x_i, p_0 + \dot{p}) - u(t_m, x_i, p_0) - \dot{U}_i^m|\Delta x$$

which is a midpoint rule discrete approximation of the integral. The error between this discrete sum and the L_1 norm above is $O(\Delta x)$ as $\Delta x \to 0$, i.e. $O(|\dot{p}|^2)$ as $\dot{p} \to 0$ for the following reason. The error due to locating the discontinuity incorrectly in the discretization of u is the value of the wrong integral between two grid points, where the midpoint is only evaluated on one side of the discontinuity, times the distance between the two grid points. That error is a constant times Δx, i.e. $O(\Delta x)$ as $\Delta x \to 0$.

Figure 1.7 shows the discrete approximation of the L_1 error integral in blue for geometrically decreasing values of $\dot{p} > 0$ for two different CFL numbers. We observe that the error does not decrease at the rate of $O(|\dot{p}|^2)$ but rather at the rate of $O(|\dot{p}|)$. That means there is no convergence in the sense of the Taylor approximation in the L_1 norm. Since the graphs for both values of the CFL number $C = 1/2$ and $C = 1$ look virtually identical the numerical viscosity does not seem to make a difference for this observation.



(a) $C = \frac{1}{2}$

(b) $C = 1$

Figure 1.7: Taylor convergence error of the discrete tangent $\sum_{i=1}^{n} |u(t_m, x_i, p_0 + \dot{p}) - u(t_m, x_i, p_0) - \dot{U}_i^m|\Delta x$ (blue), $O(|\dot{p}|)$ (red solid line), $O(|\dot{p}|^2)$ (red dashed line), Δx (red with markers), at $t = 1$ with $p = p_0 = 0$

Since Figure 1.7 does not convey a visual understanding of why there is no convergence it may be helpful to consider the way the middle part box in the graph of Figure 1.3 converges to the singular impulse as opposed to how the smooth approximation of the Dirac delta distribution converges to the singular impulse in Figure 1.5. The box is only to the right of the shock location while the approximation of the impulse is symmetrically smoothed out around the shock location.

Another perspective helps to understand why the discrete tangent of the Lax-Friedrichs solution does not converge to the finite difference for $\dot{p} \to 0$. We previously saw that $\dot{U}$ weakly converges to $\dot{u}$. We cannot have convergence of the finite difference to $\dot{u}$ in the L_1 norm since $\dot{u}$ is a distribution and not an integrable function. The piecewise constant interpolation of $\dot{U}$ is an integrable function, however, it is not bounded as $\Delta x \to 0$.

Bressan and Marson [BM95b] suggest a generalized variational calculus for solutions of conservation laws with discontinuities. Their generalized Taylor expansion uses the shock sensitivity $\mathrm{d}_p\xi$ to reconstruct the box that is seen in the difference $u(p_0 + \dot{p}) - u(p_0)$ as an effect of the additional shock travel due to the perturbation of p. The height of the box is determined by the jump height of the shock discontinuity.

In the case of the example given $\dot{p}$ such a generalized tangent is

$$v(t, x, p_0, \dot{p}) = \dot{u}_{\text{broad}}(t, x, p_0) +$$

$$H(x-\xi(t,p_0))H(\xi(t,p_0)+\dot{\xi}(t,p_0)-x)\big(u(t,\xi(t,p_0)-,p_0)-u(t,\xi(t,p_0)+,p_0)\big)$$

where $\dot{u}_{\mathrm{broad}}$ is the so-called broad tangent that is defined as the L_1 part of the distributional tangent $\dot{u}$, i.e. it is $\dot{u}$ minus the Dirac delta distribution (see Section 5.3.1 for a definition of the broad tangent). In the generalized tangent we subtract the singular impulses and instead add boxes that have the height of one of the factors of the singular impulse, the jump height at the shock, and the width of the other factor of the impulse, the shock sensitivity.

The analytic generalized tangent satisfies the L_1 Taylor convergence

$$\|u(t,\cdot,p_0+\dot{p})-u(t,\cdot,p_0)-v(t,\cdot,p_0,\dot{p})\|_{L_1([-1,1])}=O(|\dot{p}|^2)$$

as $\dot{p}\to 0$ by using the shock sensitivity to first-order approximate the shock travel in the perturbation. In this thesis we show how to create a numerical approximation of the generalized tangent based on a discrete sensitivity analysis of the shock location.

1.3.8 Convergence of the Discrete Tangent of the Integral Objective Gradient

The third and last question we consider regarding convergence is whether the discrete tangent sensitivity of the integral objective converges to its analytic tangent sensitivity, the analytic gradient derived in Equations (1.3.6)-(1.3.7) times $\dot{p}$, i.e. we consider the whether

$$\left|\dot{\Phi}_{\mathrm{discrete}}-\mathrm{d}_p\Phi(p_0)\dot{p}\right|\to 0$$

as $\Delta x\to 0$. Figure 1.8 shows how the error behaves for different values of Δx at $t_m=1$ and with $p=p_0=0,\dot{p}=1$ where we we have $\mathrm{d}_p\Phi(p_0)\dot{p}=-1/4$.

Again, we evaluate how the error between the discrete approximation and the analytic value behaves empirically for geometrically decreasing values of $\Delta x>0$. The observed rates from this kind of convergence analysis do not necessarily extrapolate to an asymptotic convergence rate as $\Delta x\to 0$.

Figure 1.8 shows the error between the analytic tangent of the integral objective and the discrete approximation for three different values of the CFL number C, i.e. three different amounts of numerical viscosity in the solution. The purple graph for $C=1$ suggests that the gradient is not convergent as it jumps back to a high value even for small values of Δx if we choose the minimum amount of numerical viscosity admissible under the CFL condition. But even with more numerical viscosity with CFL number $C=4/5$ the green graph also jumps back up for small values of Δx. The blue graph for $C=1/2$ seems like for this even higher amount of numerical viscosity the method does in fact converge at first-order. But like we said, the continued behavior of the graph for even smaller values of Δx is not necessarily guaranteed in the asymptote of $\Delta x\to 0$.

In fact, we now show why in general the discrete sensitivity of the integral objective function does not converge to the correct value. As we mentioned earlier, the number of grid points in the shock region over which the shock is smoothed out due to numerical viscosity stays constant as $\Delta x\to 0$ for a constant CFL number C. If a shock capturing method resolves the shock with a constant instead of an increasing number of grid points as $\Delta x\to 0$ then the resulting discrete

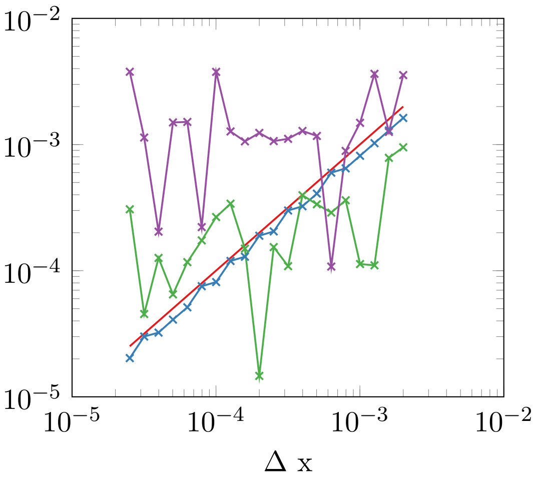


Figure 1.8: Error of the discrete tangent $\left|\dot{\Phi}_{\text{discrete}} - \mathrm{d}_p\Phi(p)\dot{p}\right|$ with $C = \frac{1}{2}$ (blue), $C = \frac{4}{5}$ (green), $C = 1$ (purple), at $t = 1$ with $p = p_0 = 0, \dot{p} = 0$, $O(\Delta x)$ (red)

sensitivities of the integral objective will not converge. Giles [Gil03] shows why that statement holds and we here use a similar example to that of Giles for demonstration.

For the purpose of demonstration we consider the approximation of the more general integral objective $\Phi = \int_{-1}^{1} \varphi(u(t, x, p))\mathrm{d}x$ with two different functions $\varphi = \varphi_1$ and $\varphi = \varphi_2$. The concrete definition of the two functions φ_1, φ_2 is not relevant as long as they satisfy certain properties that we explain later. We use the analytic viscosity solution u_{smooth} to understand the behavior of the discrete solution with numerical viscosity that is proportional to the discretization Δx.

The discrete tangent of the integral objective function behaves similarly to the discrete integral approximation of the integral objective function using the analytic viscosity solution

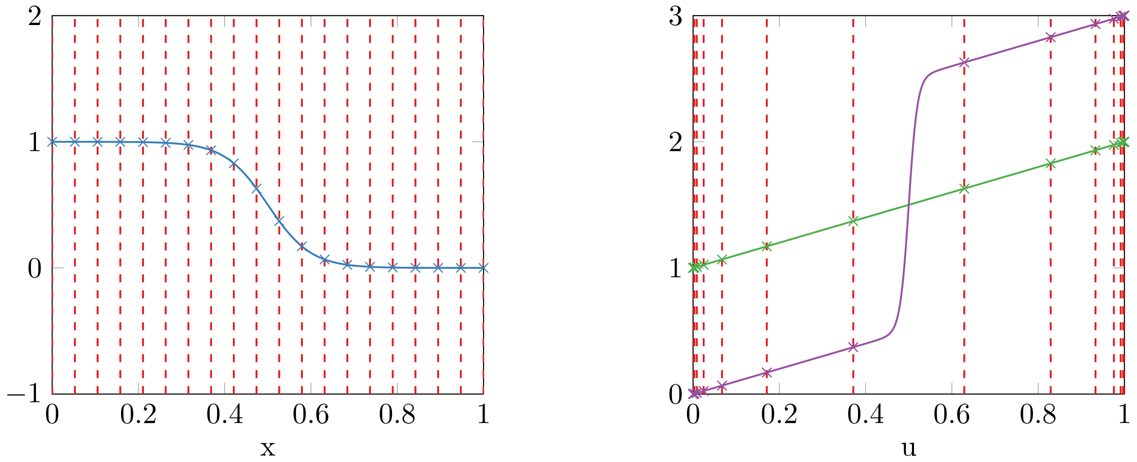
$$\sum_{i=1}^{n} \varphi(u_{\text{smooth}}(t_m, x_i, p_0))\Delta x \approx \int_{-1}^{1} \varphi(u_{\text{smooth}}(t_m, x, p_0))\mathrm{d}x \equiv \Phi_{\text{smooth}} \ .$$

The corresponding tangent sensitivity that we analyze in order to understand the discrete tangent is

$$\mathrm{d}_p\Phi_{\text{smooth}} \cdot \dot{p} = \int_{-1}^{1} \varphi'(u_{\text{smooth}}(t_m, x, p_0))\dot{u}_{\text{smooth}}(t_m, x, p_0)\mathrm{d}x \tag{1.3.10}$$

$$\approx \sum_{i=1}^{n} \varphi'(u_{\text{smooth}}(t_m, x_i, p_0))\dot{u}_{\text{smooth}}(t_m, x_i, p_0)\Delta x \ . \tag{1.3.11}$$

Both Figure 1.9 and Figure 1.10 on the left show u_{smooth} in blue at $t_m = 1$ with $p = p_0 = 0$ on the domain $x \in [0, 1]$ with two different associated discretization grid spacings Δx. The discretization grid is visualized using dashed red lines in each Figure. In the visualization we see that the number of grid points in the region where the shock is smoothed out is roughly eight in both figures when one takes into account those points where the blue graph is clearly not one or zero. So the number of grid points involved in the discretization of the smoothing of the shock region is constant as $\Delta x \to 0$ because even though the distance between grid points becomes smaller the shock region becomes smaller at the same rate. As we see in Figure 1.4 the same is true for the discrete Lax-Friedrichs solution.



(a) u_{smooth} (solid blue), discrete grid (dashed red)

(b) φ_1 (solid green), φ_2 (solid purple), values taken by u_{smooth} on discrete grid (dashed red)

Figure 1.9: $\Delta x = \frac{1}{4} \cdot 10^{-1}$

On the right of Figure 1.9 and Figure 1.10 we see a graph where the red dashed lines mark those values taken by u_{smooth} based on the discretization shown on the corresponding left picture. Due to the narrowing shock region for smaller Δx the number of times that values close to one and zero are taken in the discretization becomes higher. But the number of grid points where u_{smooth} takes values between 4/10 and 6/10 are the same for both discretizations.

The right of both figures also shows two function graphs for φ_1 in green and φ_2 in purple. The functions φ_1 and φ_2 are defined such that the tangent of the continuous integral $\mathrm{d}_p \Phi_{\text{smooth}} \cdot \dot{p}$ is different depending on which of the two functions we substitute for φ. The derivative φ_1' is a constant since φ_1 is linear. The derivative φ_2' is also roughly the same constant for values outside of the interval $[4/10, 6/10]$ because both graphs exhibit the same slope outside of that interval. But the derivative φ_2' has a much higher value than φ_1' for $u = 1/2$ because at that point the slope of φ_2 is much steeper than φ_1. The function $\dot{u}_{\text{smooth}}$ for $t = 1, p = p_0 = 0$ takes a value that is $O(\Delta x^{-1})$ as $\Delta x \to 0$ where $u_{\text{smooth}} = 1/2$. So we have

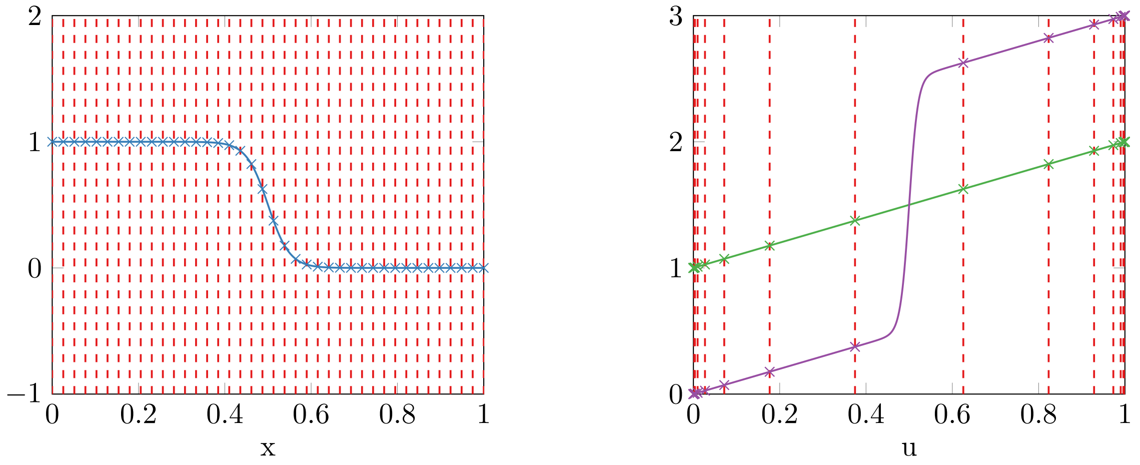
$$\int_{-1}^{1} \varphi_1'(u_{\text{smooth}}(t_m, x, p_0))\dot{u}_{\text{smooth}}(t_m, x, p_0)\mathrm{d}x \neq \int_{-1}^{1} \varphi_2'(u_{\text{smooth}}(t_m, x, p_0))\dot{u}_{\text{smooth}}(t_m, x, p_0)\mathrm{d}x$$

as $\Delta x \to 0$ and the difference between the two integrals may in fact be large.

Using the discretization of u_{smooth} it is impossible to quantify the difference between φ_1' and φ_2' since the values of u_{smooth} where the difference occurs are not taken by the discretization even as $\Delta x \to 0$. We observe this in both figures on the right by the fact that at all points that are passed by red dashed lines the slope of φ_1 and φ_2 are roughly the same and none of them is near $u = 1/2$. From the perspective of the discretization we have

$$\sum_{i=1}^{n} \varphi_1'(u_{\text{smooth}}(t_m, x_i, p_0))\dot{u}_{\text{smooth}}(t_m, x_i, p_0)\Delta x \approx \sum_{i=1}^{n} \varphi_2'(u_{\text{smooth}}(t_m, x_i, p_0))\dot{u}_{\text{smooth}}(t_m, x_i, p_0)\Delta x$$

as $\Delta x \to 0$. The same holds true for the discrete integral approximation based on the Lax-Friedrichs solution. For the Lax-Friedrichs solution we also have the same effect of the discretization not taking on more values between $u(\xi-)$ and $u(\xi+)$.



(a) u_{smooth} (solid blue), discrete grid (dashed red)

(b) φ_1 (solid green), φ_2 (solid purple), values taken by u_{smooth} on discrete grid (dashed red)

Figure 1.10: $\Delta x = \frac{1}{8} \cdot 10^{-1}$

If the continuous integrals are different but the discrete integrals are the same as $\Delta x \to 0$ then we know that the discrete integrals cannot be converging to the continuous integrals as $\Delta x \to 0$. We know that the continuous integral tangent $\mathrm{d}_p\Phi_{\text{smooth}} \cdot \dot{p}$ converges to $\mathrm{d}_p\Phi \cdot \dot{p}$ because u_{smooth} converges to u and $\dot{u}_{\text{smooth}}$ weakly converges to the distribution $\dot{u}$. So the discrete tangent of the integral is not converging to the analytic tangent sensitivity $\dot{\Phi}$ as $\Delta x \to 0$.

By the above argument we have explained the nonconvergence shown in Figure 1.8 for $C = 1$ and $C = 4/5$. The behavior of the $C = 1/2$ graph can be explained by having enough numerical viscosity such that the smoothing is resolved to a degree where the error due to missing values in the shock region only becomes visible for smaller values of Δx than are contained in the figure.

Going back to the derivation of the analytic gradient of the integral objective in Equations (1.3.6)-(1.3.7) we can see that it involves a split of the integral domain at the shock location. After splitting the integral domain in such a way the gradient contains a term with the explicit dependence on the shock sensitivity $\mathrm{d}_p\xi$. No computation involving the Dirac delta distribution or its approximation is needed in that case. As we will see in the next section the domain split used in the analytic derivation can be used in the discrete setting if we have a convergent discrete approximation of both the shock location ξ and its sensitivity $\mathrm{d}_p\xi$.

So far we have described two challenges that arise in the context of the discrete sensitivity analysis of conservation laws where the solution contains shock discontinuities. The first challenge is the approximation of a generalized Taylor expansion according to the ideas in Bressan and Marson [BM95b] but based only on information available through numerical approximation. The second challenge is in dealing with the nonconvergence of the standard discrete sensitivity analysis of the integral objective function. Both challenges can be attacked with a convergent discrete sensitivity analysis of the location of the shock discontinuities in the solution of the conservation law. In the next section we give a brief introduction to the basic ideas behind the discrete shock sensitivity analysis presented in this thesis.

1.4 Shock Sensitivities via the Rankine-Hugoniot Condition

In this section we give an introduction to the main ideas behind the numerical methods for shock sensitivities presented in this thesis. We show how to obtain a convergent discretization of the discontinuous ODE resulting from the generalized characteristics of the conservation law. We demonstrate that the standard discrete sensitivity analysis of that ODE does not result in a convergent discretization of the shock sensitivities and show how to solve that problem by an explicit discretization of the Rankine-Hugoniot condition around the shock. The discretization we suggest takes into account the numerical viscosity of the discrete solution of the conservation law in order to converge to the correct shock sensitivity $\mathrm{d}_p\xi$ as $\Delta x \to 0$.

The ideas for the correct discretization of the shock sensitivities are related to the issues discussed regarding the nonconvergence of the discrete sensitivities of the integral objective. We start off by explaining how a convergent approximation of the integral objective gradient can be obtained by explicitly avoiding the values of the discretization in the shock region and instead splitting the integral around the shock region.

1.4.1 Integrating around the Shock Region

In order to develop a convergent method for the discrete sensitivity analysis of the integral objective it is again helpful to consider the analytic viscosity solution u_{smooth} because it resembles the behavior of the discrete Lax-Friedrichs solution. We consider the limiting behavior of u_{smooth} and its sensitivities as $\Delta x \to 0$ in the context of the integral objective tangent derived via chain rule in Equation (1.3.10). We also compare the resulting sensitivities to the first derivation of the analytic integral objective gradient we gave that was based on splitting the integral at the shock location.

Analogous to Equation (1.3.10) for the analytic weak solution and based on the definition of $\dot{u}$ in Equation (1.3.9) we have the tangent sensitivity

$$\begin{aligned}\mathrm{d}_p\Phi \cdot \dot{p} &= \int_{-1}^{1} \varphi'(u(t,x,p_0))\dot{u}(t,x,p_0)\mathrm{d}x \\ &= \int_{-1}^{1} \varphi'(u(t,x,p_0))\Big(H(\xi(t,p_0)-x)\dot{p} + \delta(x-\xi(t,p_0))\mathrm{d}_p\xi(t,p_0)(1+p_0)\dot{p}\Big)\mathrm{d}x \,.\end{aligned}$$

We further separate the tangent sensitivity integral into two terms

$$\mathrm{d}_p\Phi \cdot \dot{p} = \int_{-1}^{1} \varphi'(u(t,x,p))H(\xi(t,p)-x)\dot{p}\mathrm{d}x \tag{1.4.1}$$

$$+ \int_{-1}^{1} \varphi'(u(t,x,p))\delta(x-\xi(t,p))\mathrm{d}_p\xi(t,p)(1+p)\dot{p}\mathrm{d}x \,. \tag{1.4.2}$$

In the following we consider both terms separately and describe how they relate to the limit of the analogous integral for the smooth analytic viscosity solution and the analytic gradient in Equations (1.3.6)-(1.3.7).

The first term, Equation (1.4.1), corresponds perfectly to the first two terms in Equation (1.3.6) since in the case of the Riemann example we have

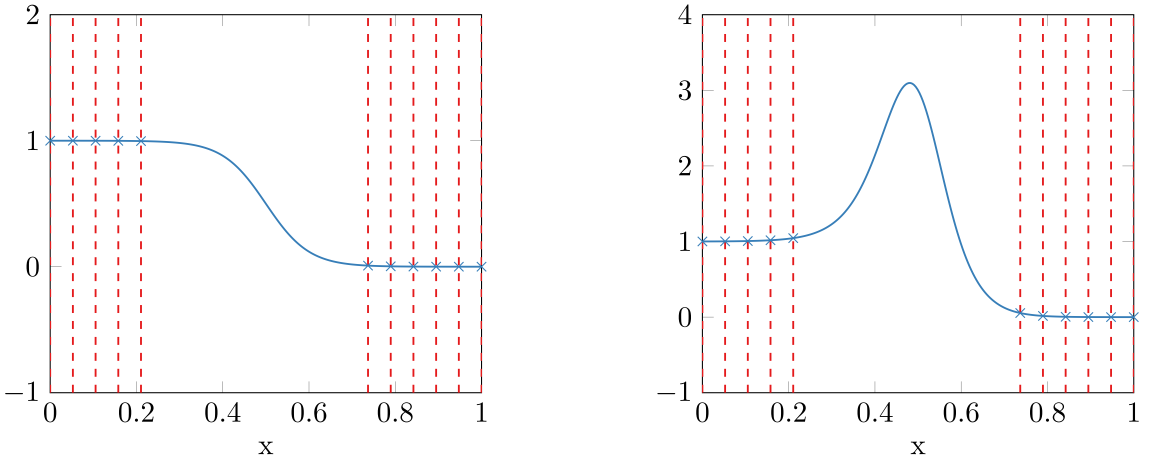
$$\int_{-1}^{1} \varphi'(u(t,x,p))H(\xi(t,p)-x)\dot{p}\mathrm{d}x = \int_{-1}^{\xi(t,p)} \varphi'(u(t,x,p))\dot{p}\mathrm{d}x$$

$$= \int_{-1}^{\xi(t,p)} \varphi'(u(t,x,p))\mathrm{d}_p u(t,x,p)\dot{p}\mathrm{d}x$$
$$+ \int_{\xi(t,p)}^{1} \varphi'(u(t,x,p))\mathrm{d}_p u(t,x,p)\dot{p}\mathrm{d}x \; .$$

The additional factor $\dot{p}$ comes from the fact that in the tangent sensitivity we are dealing with a directional derivative.

The part of the tangent sensitivity integral that corresponds to the first term, Equation (1.4.1), can be approximated by a discrete sum and based on a solution with numerical viscosity in a way that converges at an almost linear rate as $\Delta x \to 0$. The discrete approximation of the weak solution has a shock region that has width $O(\Delta x)$ as $\Delta x \to 0$ for nonlinear flux functions such as the Burgers flux. The idea is to cut out a region of at least that width. In that case the effect of the jump smoothing due to the shock and its singularity approximating tangent does not play into the discrete approximation as $\Delta x \to 0$. But the integral still converges at a rate the is almost linear since the additional error made due to leaving out that region is bounded by the width of the left out shock region times a constant because the weak solution is bounded. We do not give a formal analysis at this point but rather just aim to convey the intuition visually. For the purpose of visualization we use u_{smooth} instead of the discrete Lax-Friedrichs solution while keeping in mind their resemblance.

Figure 1.11 and Figure 1.12 on the left show u_{smooth} in blue for different values of Δx and the corresponding discretization grid using dashed red lines. On the right both of the figures show $\dot{u}_{\text{smooth}}$ with the same corresponding discretizations. The discretization grids leave out a region of width proportional to Δx around the shock location. By only evaluating an integral approximation on that grid it is visually clear that we do not take into account the part of $\dot{u}_{\text{smooth}}$ approximating the singularity. There might be some grid points close to the shock region where the approximations contain some smoothing that is not in the discontinuous solution. But if the width of the shock region grows even slightly in proportion to the discretization grid spacing this remaining error also vanishes as $\Delta x \to 0$.

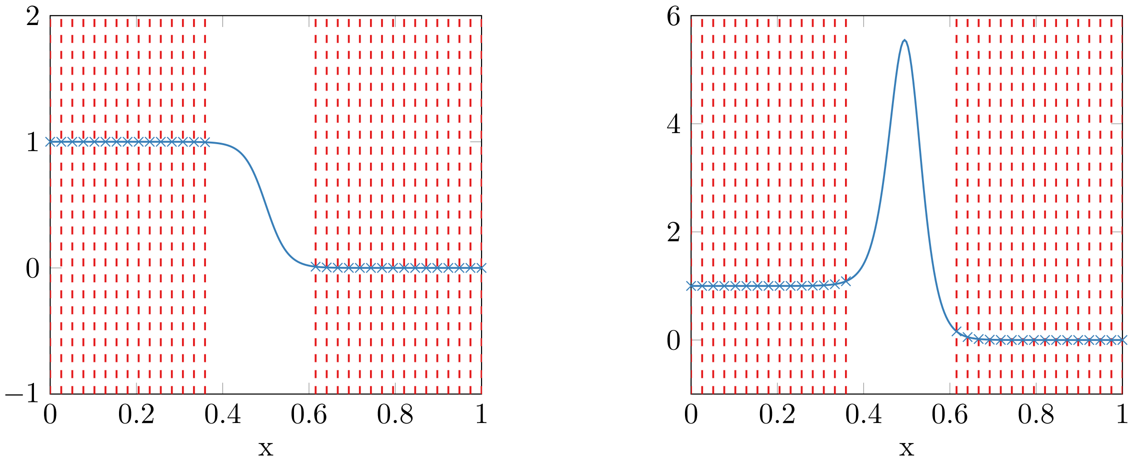


(a) u_{smooth} (solid blue), discrete grid (dashed red)

(b) $\dot{u}_{\text{smooth}}$ (solid blue), discrete grid (dashed red)

Figure 1.11: $\Delta x = \frac{1}{4} \cdot 10^{-1}$

We want to emphasize that in order to implement the idea we just described as a numerical method we need to numerically approximate the shock location $\xi(t,p)$ with sufficiently high



(a) u_{smooth} (solid blue), discrete grid (dashed red)

(b) $\dot{u}_{\text{smooth}}$ (solid blue), discrete grid (dashed red)

Figure 1.12: $\Delta x = \frac{1}{8} \cdot 10^{-1}$

accuracy based on the given discrete solution in order to correctly locate the shock region to cut out of the discretization. How to then get a convergent approximation of the integral according to the above ideas is formalized in Section 6.4.1.

We now move on to the second term, Equation (1.4.2), that arises from the sensitivity integral and involves the Dirac delta distribution. The derivation of that term is an application of the chain rule where it formally cannot be applied because the functions involved are not sufficiently regular to permit the use of classical differential calculus. The term involves a Dirac delta distribution times a discontinuous function and it is not directly obvious how it needs to be consistently interpreted to make sense. In Section 2.4.5 we introduce a formalism of generalized functions called Colombeau algebra in which the definition of such operations arises from the limit of smooth functions such as u_{smooth} and $\dot{u}_{\text{smooth}}$.

Here, we use another way to make sense of that part of the integral sensitivity. We compare the tangent sensitivity with the earlier derivation of the gradient where we did not use the Dirac delta distribution but instead split the integral at the shock location. Since the first term, Equation (1.4.1), corresponds perfectly to the first two terms in Equation (1.3.6) we know that the second term, Equation (1.4.2), has to correspond to the last term in Equation (1.3.7), i.e. we need to have

$$\begin{aligned}&\int_{-1}^{1} \varphi'(u(t,x,p))\delta(x-\xi(t,p))\mathrm{d}_p\xi(t,p)(1+p)\dot{p}\mathrm{d}x\\&= \mathrm{d}_p\xi(t,p)\big(\varphi(u(t,\xi(t,p)-,p)) - \varphi(u(t,\xi(t,p)+,p))\big)\dot{p}\\&= \mathrm{d}_p\xi(t,p)\big(\varphi(1+p) - \varphi(0)\big)\dot{p}\ .\end{aligned}$$

The way the Dirac delta distribution is defined to act on smooth functions would suggest we have

$$\int_{-1}^{1} \varphi'(u(t,x,p))\delta(x-\xi(t,p))\mathrm{d}_p\xi(t,p)(1+p)\dot{p}\mathrm{d}x = \varphi'(u(t,\xi(t,p),p))\mathrm{d}_p\xi(t,p)(1+p)\dot{p}\ .$$

But $u(t,x,p)$ is discontinuous precisely at $x=\xi(t,p)$ so the way the Dirac delta distribution is defined when acting on a smooth functions does not apply. An intuitive way to still make sense of the above is by defining $\varphi'(u)$ via the limit of a central difference, i.e. as

$$\varphi'(u(t,\xi(t,p),p)) = \frac{\varphi(u(t,\xi(t,p)+,p)) - \varphi(u(t,\xi(t,p)-,p))}{u(t,\xi(t,p)+,p) - u(t,\xi(t,p)-,p)} = \frac{\varphi(1+p)-\varphi(0)}{1+p} \ . \tag{1.4.3}$$

The central difference limit can be derived via a change of variable (see Section 5.4.2).

If we substitute that definition into the integral we get from the definition of the Dirac delta distribution when acting on a smooth function we also get

$$\begin{aligned}\varphi'(u(t,\xi(t,p),p))\mathrm{d}_p\xi(t,p)(1+p)\dot{p} &= \frac{\varphi(1+p)-\varphi(0)}{1+p}\mathrm{d}_p\xi(t,p)(1+p)\dot{p} \\ &= \mathrm{d}_p\xi(t,p)\big(\varphi(1+p)-\varphi(0)\big)\dot{p} \ .\end{aligned}$$

The above is the same result that we obtained by comparing to the derivation of the gradient of the integral objective by splitting the integral domain at the shock location.

The part of the sensitivity integral arising from the second term, Equation (1.4.2), can also be approximated based on a numerical solution of the conservation law. We suggest a method that does not require us to use the nonconvergent part of the discrete integral. The method we suggest has two components. First, we approximate both $u(t,\xi(t,p)+,p)$ and $u(t,\xi(t,p)-,p)$ in a manner that is convergent as $\Delta x \to 0$ in order to be able to approximate the discontinuity jump height factor

$$\varphi(u(t,\xi(t,p)-,p)) - \varphi(u(t,\xi(t,p)+,p)) \ .$$

Second, we approximate the shock sensitivity factor $\mathrm{d}_p\xi(t,p)$ with sufficient accuracy based on the discrete solution of the conservation law and the numerical approximation of the shock location $\xi(t,p)$.

To recap, if we have a discrete approximation $\Xi \approx \xi$ and a discrete tangent $\dot{\Xi} \approx \dot{\xi} = \mathrm{d}_p\xi \cdot \dot{p}$ then we can cut out the nonconvergent shock region from the discrete integral based on Ξ and approximate the part involving the shock sensitivity due to the discontinuity in the solution based on $\dot{\Xi}$.

1.4.2 Discrete Generalized Tangent

Based on the same techniques used for the integral objective we can also arrive at a numerical approximation of the generalized Taylor expansion that is convergent in the sense of the L_1 norm. We briefly summarize the essential ideas for the construction of a discrete generalized tangent.

As a first component the generalized tangent involves the broad tangent $\dot{u}_{\mathrm{broad}}$ that is a tangent sensitivity of the weak solution minus the singular parts such that it is in L_1. In the context of the discrete tangent we can approximate the broad tangent by using the same idea we presented in the previous section for cutting out the shock region.

The smoothing of the shock represents an approximation of the Dirac delta distribution singularity in the tangent that is unbounded as $\Delta x \to 0$. We can cut out the shock region from the discrete tangent $\dot{U}$ by cutting out an at least $O(\Delta x)$ wide region around the shock location as $\Delta x \to 0$. The value of the discrete broad tangent in that area can be set to a constant, for example zero. Since the analytic broad tangent is bounded the error in the cut out region is constant as $\Delta x \to 0$ and the shock region has a width that is almost linear in Δx. If we cut out a sufficiently wide region relative to Δx we get rid of the singularity and the resulting discrete broad tangent converges to the analytic broad tangent in the sense of the L_1 norm at an almost linear rate since the width of the cut out region goes to zero as $\Delta x \to 0$.

The second component of the generalized tangent requires the numerical approximation of the width of the box used to represent the additional shock travel due to the perturbation of p. The width of that box is approximated by the tangent of the shock location $\dot{\xi}$ up to first-order. By a numerical approximation $\dot{\Xi} \approx \dot{\xi}$ that is convergent we get a sufficiently accurate box width approximation.

The height of the box used to represent the shock travel is determined by the discontinuity jump height in the weak solution

$$u(t, \xi(t,p)-, p) - u(t, \xi(t,p)+, p) \ .$$

That is the same factor we also need to approximate in the context of the integral objective gradient term that involves the shock sensitivity. In order to approximate the jump height based on the discrete solution of the conservation law in a convergent manner we need to also correctly evaluate around the shock region based on the discrete approximation of the shock location $\Xi \approx \xi$.

From these last two sections it has become clear that we require three components in order to attack the challenges posed in the previous section. That is to obtain both a discrete generalized tangent that is convergent in the sense of the Taylor expansion in the L_1 norm and to obtain a convergent discrete tangent sensitivity for the integral objective we need the following three ingredients. We need a numerical approximation of

(i) the shock location $\Xi \approx \xi$,

(ii) the shock sensitivity $\dot{\Xi} \approx \dot{\xi}$ and

(iii) the discontinuity jump height.

All three numerical approximations need to be convergent as $\Delta x \to 0$. In the next section we explain in more detail how to approximate the discontinuity jump height and discuss the best possible convergence rates we can hope to achieve based on the presence of numerical viscosity in the discrete solution of the conservation law.

1.4.3 Numerical Approximation of the Discontinuity Jump Height

For both the generalized tangent and the integral objective sensitivity we need an approximation of $u(t, \xi(t,p)\pm, p)$ based on the discrete Lax-Friedrichs solution U. We now discuss how this approximation can be obtained assuming that the discrete Lax-Friedrichs solution behaves in a similar manner as the analytic viscosity solution for pointwise evaluations near the shock region. In the context of the Riemann example this assumption is reasonable as we have seen in the figures visualizing the Lax-Friedrichs solution in the previous section.

If we consider the question of how to approximate $u(\xi(t,p)+)$ based on the analytic viscosity solution u_{smooth} then we come to the realization that we cannot do so based on evaluating near the shock location ξ in the smoothed out region where values lie between $u(\xi+)$ and $u(\xi-)$. In the case of some numerical integrators the values of the discrete solution in the shock region might even oscillate, i.e. they may not be stable with respect to small perturbations of ξ.

Instead of evaluating at the shock location we need to evaluate outside of the shock region by moving away from the shock by a distance that is at least proportional to the width of the shock region. For example we could try to evaluate $u_{\text{smooth}}(\xi + C\Delta x)$ for some $C > 0$ to approximate $u(\xi+)$. But choosing $u_{\text{smooth}}(\xi+C\Delta x)$ also does not quite converge to $u(\xi+)$ because the point $\xi+C\Delta x$ is not moving out of the $O(\Delta x)$ wide shock region as $\Delta x \to 0$. If there is a slight error $|u(\xi+) - u_{\text{smooth}}(\xi + C\Delta x)|$ due to some residual shock smoothing at $\xi + C\Delta x$ then that error does not get smaller as $\Delta x \to 0$. The point we are at relative to the smoothing region width stays at a constant distance.

Figure 1.11 and Figure 1.12 on the left illustrate this effect if we consider the value of the blue graph at the red dashed grid line that is closest to the right of the cut out shock region. If the blue graph evaluated at those points has a smoothing error in Figure 1.11 that may be too small to see then it also has the same smoothing error in Figure 1.12 because relative to the width of the shock region the point of evaluation has not moved further away from the shock. Of course it is always possible to choose C large enough such that for relevant values of Δx the residual smoothing error will be so small that it becomes irrelevant for practical purposes. But in the context of an asymptotic convergence analysis we have to choose a different approach.

The other extreme is to evaluate at constant distance from the shock location, i.e. $u_{\text{smooth}}(\xi + C)$ for some $C > 0$ but that of course does not converge to $u(\xi+)$ as $\Delta x \to 0$ either. As the shock region gets smaller $u_{\text{smooth}}(\xi + C)$ converges to $u(\xi + C)$. In the case of our example that may be the same value as $u(\xi+)$ but for a general weak solution that is not piecewise constant the approach of moving out of the shock by a constant does not work.

We have to use the approach that lies in between the two previously considered alternatives and that is to evaluate at $u_{\text{smooth}}(\xi + C\Delta x^{\alpha})$ for some $C > 0$ and $0 < \alpha < 1$. The point $\xi + C\Delta x^{\alpha}$ moves further and further out of the shock region relative to its width of $O(\Delta x)$ but also still converges to ξ from the right as $\Delta x \to 0$.

Since we have to use $u_{\text{smooth}}(\xi + C\Delta x^{\alpha})$ with $\alpha < 1$ to approximate $u(\xi+)$ the convergence is not quite at a linear rate. At best the overall error obtained for the discrete sensitivities based on the approximation is $O(\Delta x^{\alpha})$ as $\Delta x \to 0$. That is the same sublinear rate achieved by the discrete sensitivity methods suggested in Giles and Ulbrich [GU10a; GU10b].

Based on the ideas in this section the discontinuity jump height can be approximated by

$$\varphi(U^m(\Xi^m - C\Delta x^{\alpha})) - \varphi(U^m(\Xi^m + C\Delta x^{\alpha})) \approx \varphi(u(t_m, \xi(t_m,p)-, p)) - \varphi(u(t_m, \xi(t_m,p)+, p))$$

where $U^m(\cdot)$ is a linear interpolation of the discrete Lax-Friedrichs solution and Ξ^m is a numerical approximation of the shock location at time t_m. We have yet to discuss how to compute Ξ which is the topic of the next section.

1.4.4 Numerical Approximation of Shock Location

Characteristics are the lines along which information flows through the solution of a conservation law as time passes (see Section 3.1.2 for a definition of characteristics and their visual illustration for the Riemann example). A feature of shock discontinuities is that shocks appear when the information flow on both sides of the discontinuity, i.e. the characteristics around the shock, point in the direction of the shock discontinuity and there is compression at the shock.

The dynamics of the characteristic lines of a conservation law are determined by the ODE

$$\partial_t \xi(t) = f'(u(t, \xi(t))$$

where f is the flux function. If the weak solution of the conservation law u is discontinuous then the right-hand side of that ODE is discontinuous as well.

The topic of discontinuous ODEs has a vast literature and there are multiple different possible definitions of the solution of a discontinuous ODE depending for example on whether the trajectory potentially stays on the discontinuity. An established way of defining the solution of discontinuous ODEs where the solution can slide along the discontinuity, as would be the case for a characteristic along a shock, is according to Filippov [Cor08; Fil88]. Filippov makes use of the concept of a differential inclusion where the right-hand side of the ODE is turned into a set-valued map, i.e. a map that takes a point in some space to a set of points in another space. The solution of a differential inclusion is a continuous function of time such that the dynamics on the left-hand side belongs to the set on the right-hand side for every point in time.

The Filippov approach defines the right-hand side as the convex closure of points surrounding the discontinuity. In the scalar case we take the convex closure of the left and right limits and end up with a closed interval on the right-hand side to get the differential inclusion

$$\partial_t \xi(t) \in [f'(u(t, \xi(t)+), f'(u(t, \xi(t)-)] \; .$$

We note that if $u(t, x)$ does not have a discontinuity at $x = \xi(t)$ then the interval is a point set and we end up with the original differential equation. The Filippov interpretation of characteristics for discontinuous solutions of conservation laws is also consistent with the theory of generalized characteristics according to Dafermos [Daf77].

Since in the case of a shock that persists over time the characteristics flow into the shock and then stay on the shock we can simulate the shock location by starting out with the initial value being the shock location at $t = 0$. The shock location needs to be known from the initial data of the conservation law. In the case of the Riemann example we use the initial value condition of $\xi(0) = 0$ for the shock location. Simulating the generalized characteristic differential inclusion tracks the shock location.

The numerical simulation of discontinuous ODEs where sliding of the solution along discontinuities occurs is also a topic of a range of papers (see e.g. [DL09]). If we are not necessarily interested in high accuracy then the basic explicit Euler method can be used to obtain convergent numerical approximations of solutions to differential inclusions. Under reasonable assumptions the solution set of the explicit Euler method converges to the solution set of the differential inclusion at a first-order rate as $\Delta t \to 0$ where Δt is the spacing of a uniform time discretization grid [DL92, Theorem 6.1].

Motivated by the success of the explicit Euler method in finding Filippov solutions for discontinuous ODEs in the presence of sliding we consider its use for the simulation of the generalized characteristics by the following time stepping iteration

$$\Xi^{j+1} := \Xi^j + \Delta t f'(U^j(\Xi^j))$$

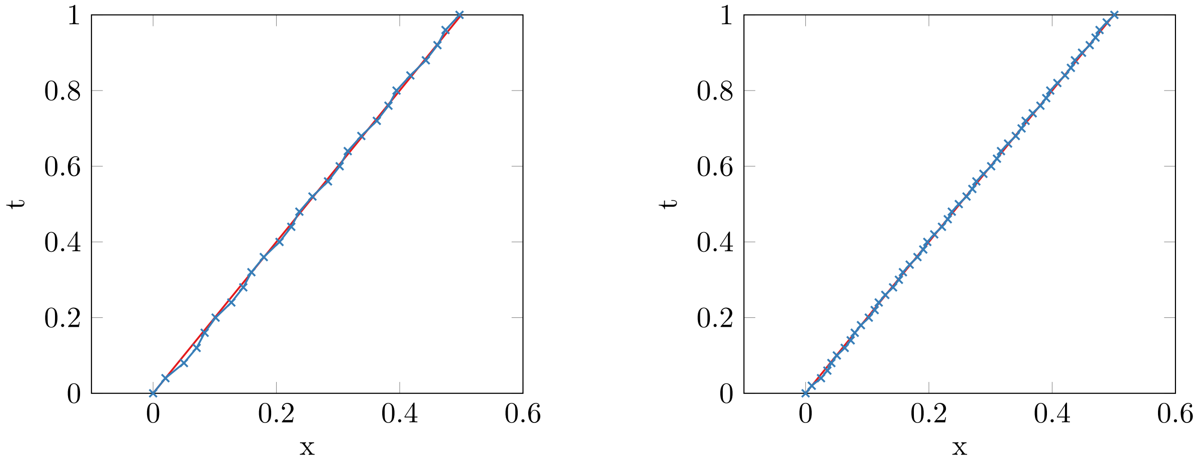
where $U^j(\cdot)$ is a linear interpolation of the discrete Lax-Friedrichs solution of the conservation law.

The explicit Euler method applied to the generalized characteristic differential inclusion is actually

$$\Xi^{j+1} \in \Xi^j + \Delta t f'(u(t_k, \Xi^j))$$

but since we only have access to the numerical approximation of u we cannot implement that method. The linear interpolation of the discrete Lax-Friedrichs solution is not a discontinuous function. For the purpose of the method we describe here we could also utilize a piecewise constant interpolation that is discontinuous. But the relevant observation is that due to the numerical viscosity both of these interpolations do not represent the shock discontinuity in the original solution in a sharp manner relative to Δt. If the shock region does not contain oscillations the simulation of the characteristic ODE based on the discrete solution in practice is more similar to an artificial smoothing of the right-hand side of the ODE. Smoothing also presents a convergent way of simulating discontinuous ODEs [SA12].

Figure 1.13 shows the trajectory for Ξ obtained by the explicit Euler method based on the linear interpolation of U for two different discretization grid spacings Δx in blue compared to the analytic ξ with $p = 0$ in red. Especially on the left we can see that the discrete solution oscillates around the exact shock discontinuity. We do not show a convergence plot here but the discrete shock location from an explicit Euler discretization will converge at a linear rate since the distance at which it oscillates around the shock is linear in $\Delta t = O(\Delta x)$ as $\Delta x \to 0$.



(a) $\Delta x = 4 \cdot 10^{-2}$

(b) $\Delta x = 2 \cdot 10^{-2}$

Figure 1.13: Explicit Euler discretization of shock location $\xi(t_j)$ (red), Ξ^j (blue)

In Section 6.2.1 we present a numerical convergence analysis of the explicit Euler discretization for Ξ. The convergence analysis is not based on a smoothing view of the numerical viscosity but rather just the assumption that the discrete solution U is bounded in the shock region. That assumption also holds for methods that try to sharpen the shock in the discrete solution rather than to make it more smooth.

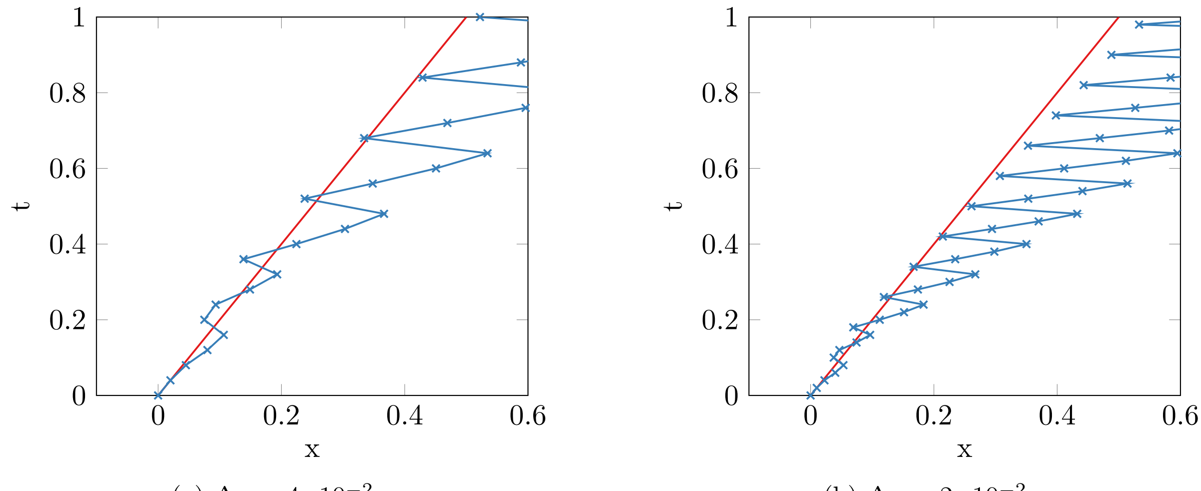
1.4.5 Numerical Approximation of Shock Sensitivity

The naive approach to the discrete tangent sensitivity analysis of the shock location Ξ is by directional differentiation of the explicit Euler step. The resulting discrete tangent step is

$$\dot{\Xi}^{j+1} := \dot{\Xi}^{j} + \Delta t f''(U^j(\Xi^j))\big(\dot{U}^j(\Xi^j) + \partial_x U^j(\Xi^j)\dot{\Xi}^j\big)$$

where $\dot{U}^j(\cdot)$ is the linear interpolation of the discrete tangent of the Lax-Friedrichs method and $\partial_x U^j(\cdot)$ is the derivative of the linear interpolation of the discrete Lax-Friedrichs solution. The initial condition arises from the sensitivity of the initial condition of the shock location. In the case of the Riemann example we have $\mathrm{d}_p\xi(0,p) = 0$ as the shock location in the initial data is always at $\xi(0,p) = 0$ and does not depend on p.

Figure 1.14 shows the trajectory for the naive discrete tangent $\dot{\Xi}$ in blue compared to the analytic tangent sensitivity $\dot{\xi} = \mathrm{d}_p\xi \cdot \dot{p}$ with $p = 0, \dot{p} = 1$ in red. We observe that the discrete sensitivity $\dot{\Xi}$ strongly oscillates and moves away from the analytic solution. The discrete sensitivity gives the wrong results and an empirical convergence analysis will reveal that it does not converge to the analytic solution as $\Delta x \to 0$.



(a) $\Delta x = 4 \cdot 10^{-2}$

(b) $\Delta x = 2 \cdot 10^{-2}$

Figure 1.14: Naive discrete tangent of explicit Euler for shock loation $\dot{\xi}(t_j)$ (red), $\dot{\Xi}^j$ (blue)

The nonconvergence of the discrete tangent of the explicit Euler method results from the fact that the underlying mathematical structure is not differentiable. Stewart and Anitescu [SA10] perform a discrete sensitivity analysis of discontinuous ODEs with a solution that slides along the discontinuity of the right-hand side. They show that the derivatives of the original explicit Euler discretization do not give the correct sensitivities. Even if we do not use the discontinuous u but instead the discrete solution U with numerical viscosity we still observe that Ξ oscillates around the true shock location ξ just as it would for the explicit Euler discretization with a sharp discontinuity. The smoothing of the shock region due to the numerical viscosity is not sufficient to make the discretization differentiable.

The proposed sensitivity analysis of Stewart and Anitescu [SA10] is based on a smoothing of the discontinuity that is sufficiently wide such that the derivatives of an explicit Euler discretization converge. Their smoothing method requires the smoothing region to grow relative to $\Delta t = O(\Delta x)$, i.e. they suggest what would be the analogy to a numerical viscosity that results in a shock region of width $O(\Delta x^\alpha)$ with $0 < \alpha < 1$ as $\Delta x \to 0$. The increased viscosity solution corresponds to the approach by Giles and Ulbrich [GU10a; GU10b] that is not compatible with the requirement for a sharp shock resolution in the discrete solution of the conservation law.

The discrete sensitivity analysis we introduce in this thesis is instead based on a convergent discretization of the Rankine-Hugoniot condition that holds for generalized characteristics along shocks. The Rankine-Hugoniot condition for the conservation law with flux f is

$$\partial_t \xi(t,p) = \frac{f(u(t,\xi(t)+,p)) - f(u(t,\xi(t)-,p))}{u(t,\xi(t)+,p) - u(t,\xi(t)-,p)} \,.$$

We note that the right-hand side is a generalized interpretation of the derivative $f'(u)$ where u is discontinuous that is analogous to the one we introduced for $\varphi'(u)$ in Equation (1.4.3).

In order to obtain a convergent discretization of the Rankine-Hugoniot condition we can use the same idea that we used for the discretization of the discontinuity jump height. Going back to the previous section we recall that in order to obtain a convergent approximation of $u(t_j, \xi(t_j)\pm)$ we have to use the discretization evaluated at $\xi(t_j) \pm C\Delta x^\alpha$ for some $C > 0$ and $0 < \alpha < 1$. The purpose of choosing $\alpha < 1$ is to make sure that we move further and further out of the region with smoothing as $\Delta x \to 0$. We only have access to $\Xi^j \approx \xi(t_j)$ and so we require that Ξ_j is sufficiently accurate to still guarantee that $\Xi^j \pm C\Delta x^\alpha$ stays left and right of the shock region as $\Delta x \to 0$.

If $U^j(\cdot)$ is the linear interpolation of the Lax-Friedrichs discretization we discretize the Rankine-Hugoniot condition using an explicit Euler time step as

$$\Xi^{j+1} := \Xi^j + \Delta t \frac{f(U^j(\Xi^j + C\Delta x^\alpha)) - f(U^j(\Xi^j - C\Delta x^\alpha))}{U^j(\Xi^j + C\Delta x^\alpha) - U^j(\Xi^j - C\Delta x^\alpha)} \,.$$

The above discretization presents an alternative way of computing the shock location Ξ instead of the discrete tangent sensitivity $\dot{\Xi}$. Before answering the question of how to compute the discrete tangent based on the Rankine-Hugoniot discretization we briefly discuss which approach to use to obtain the approximate shock location Ξ.

The above discretization based on the Rankine-Hugoniot condition at best obtains an approximation error that is $O(\Delta x^\alpha)$ as $\Delta x \to 0$ just based on the involved approximation around the shock region. But the correct choice of the parameters C and α is delicate because the further we move out of the shock region the higher the approximation error in Ξ becomes. A higher approximation error in Ξ in turn requires us to choose the parameters such that we go further out of the shock region in order to guarantee that we are still left and right of the smoothing region on both sides. It is not immediately clear whether the discretization even converges to the correct solution.

The approximation based on the direct discretization of the discontinuous ODE that describes the dynamics of the generalized characteristic does not require us to choose any parameters and even though it oscillates around the shock its error to the analytic solution can be upper bounded based on bounds for the values of the disrete solution of the conservation law. The choice for the discretization of the shock location goes to the approach described in the previous section. We only use the Rankine-Hugoniot discretization to obtain the numerical approximation of the shock sensitivities and not the shock location.

We now go on to derive the directional derivative of the explicit Euler discretization of the Rankine-Hugoniot condition in order to obtain the discrete tangent sensitivity time step. We have

$$\dot{\Xi}^{j+1} := \dot{\Xi}^{j} + \Delta t \left(\frac{f'(U^+)\dot{U}^+ - f'(U^-)\dot{U}^-}{U^+ - U^-} - \frac{f(U^+) - f(U^-)}{(U^+ - U^-)^2} (\dot{U}^+ - \dot{U}^-) \right)$$

where

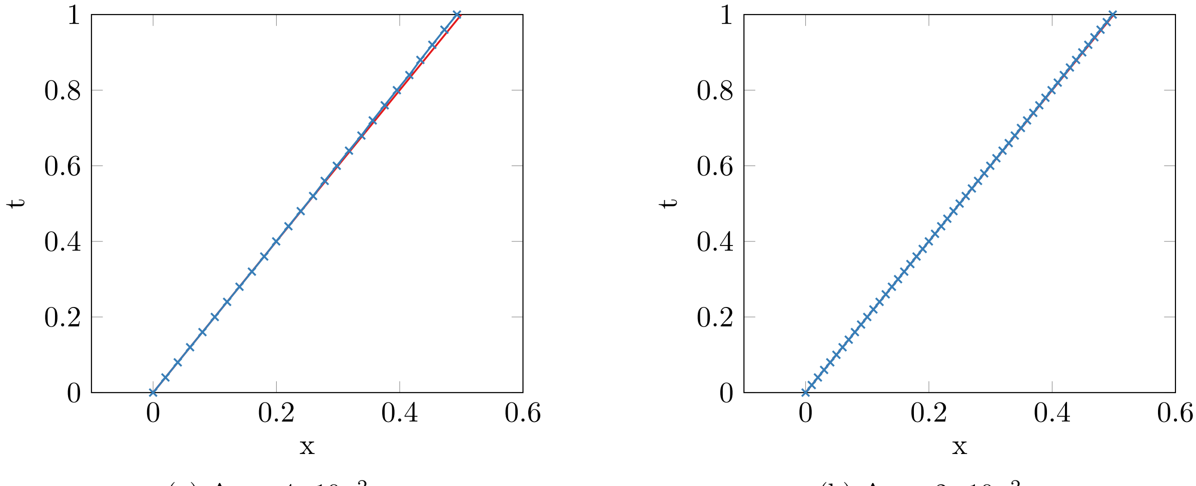
$$U^{\pm} \equiv U^j(\Xi^j \pm C\Delta x^{\alpha})$$

and

$$\dot{U}^{\pm} \equiv \dot{U}^j(\Xi^j \pm C\Delta x^{\alpha}) + \partial_x U^j(\Xi^j \pm C\Delta x^{\alpha})\dot{\Xi}^j$$

with $U^j(\cdot), \dot{U}^j(\cdot), \partial_x U^j(\cdot)$ the corresponding interpolations as defined before.

Figure 1.15 shows the trajectory of the discrete tangent $\dot{\Xi}$ resulting from the discretization based on the Rankine-Hugoniot condition in blue for two different values of Δx or Δt. The shock location Ξ on the right-hand side of the explicit Euler time step is computed as described in the previous section. The shock sensitivity discretization uses the parameters $C = 1$ and $\alpha = 1/2$. For comparison the figure shows the analytic tangent sensitivity $\dot{\xi}$ with $p = 0, \dot{p} = 1$ in red. We observe that the solutions for both discretization accuracies are without oscillations which indicates that the discrete solution and its tangent is not evaluated inside of the shock region. For the more accurate discretization the solution is not visually distinguishable from the analytic shock sensitivity.



(a) $\Delta x = 4 \cdot 10^{-2}$

(b) $\Delta x = 2 \cdot 10^{-2}$

Figure 1.15: Rankine-Hugoniot discrete tangent of explicit Euler for shock loation $\dot{\xi}(t_j)$ (red), $\dot{\Xi}^j$ (blue)

In Section 6.2.2 we present the formal convergence analysis of the discrete shock sensitivity based on the Rankine-Hugoniot method presented here. As expected from the derivation the convergence rate of the error obtained by the discrete sensitivities is $O(\Delta x^{\alpha})$ as $\Delta x \to 0$.

1.5 Structure of the Thesis

In Chapter 2 we give the necessary mathematical background for the analytic tangent and adjoint sensitivity analysis of PDEs with discontinuous solutions in the sense of distributions.

In Chapter 3 we introduce the basic concepts for scalar conservation laws with discontinuous solutions and their numerical approximation via shock capturing schemes.

In Chapter 4 we give a derivation of the analytic and the discrete tangent and adjoint sensitivity analysis of smooth solutions of scalar conservation laws. We also introduce the ideas of forward and reverse mode AD and how AD tools can be used to generate discrete sensitivity analyses from numerical integrators.

In Chapter 5 we introduce an assumed solution structure of piecewise smooth discontinuous solutions of scalar conservation laws and give the derivation of its analytic sensitivity analysis. We give a derivation of the generalized tangent of Bressan and Marson [BM95b] for the assumed solution structure and the adjoint sensitivity of an integral objective.

In Chapter 6 we introduce an algorithm for the numerical approximation of the sensitivities of shock locations in a discontinuous solution of a scalar conservation law. We give a numerical analysis proving its convergence to the analytic shock sensitivities as the discrete grid is refined. We also show how to numerically approximate the generalized tangent and adjoint sensitivities of an integral objective based on the shock sensitivity approximation. We validate the presented methods via computational experiments with a convex and a nonconvex flux function.

In Chapter 7 we extend our approach for the numerical approximation of shock sensitivities to one-dimensional hyperbolic systems of conservation laws and validate the presented methods via computational experiments with a system of isentropic gas equations.

In Chapter 8 we extend our approach for sensitivities of point shocks in a one-dimensional conservation law to points on a one-dimensional shock manifold in a two dimensional conservation law and validate the presented method via computational experiments with multiple parameterization of a two dimensional Burgers equation.

In Chapter 9 we summarize conclusions and propose directions for further work building on the ideas presented in this thesis.

Chapter 2

Mathematical Background

In this chapter we introduce the necessary background in mathematical analysis required for the tangent and adjoint sensitivity analysis of PDEs and to define the analytic sensitivity analysis of conservation laws with discontinuous solutions. We also introduce some basic mathematical tools required in proofs throughout the thesis.

We give the basic definitions and theorems related to normed linear spaces and the operators between them. We show how to (implicitly) differentiate in function spaces assuming sufficient regularity and define the basic notions of Lebesgue integration. Furthermore, we define two formalisms of generalized functions; the theory of distributions and the Colombeau algebra.

Most of the definitions we give are standard and can be found in many text books on mathematical analysis. The references we use are Cheney [Che13], Traub [Tra03], and Stakgold and Holst [SH11]. Additional references will be explicitly stated throughout the text.

2.1 Linear Spaces and Operators

2.1.1 Normed Linear Spaces

We now define the notion of (normed) linear spaces and operators between them.

Definition 1. A *linear space* (or vector space) over $\mathbb{R}$ is a nonempty set X that satisfies

(i) $x + y \in X$ and $x + y = y + x$

(ii) $(x + y) + z = x + (y + z)$

(iii) there exists an element 0 in X such that $0 + x = x$

(iv) $\alpha x \in X$ and $\alpha(\beta x) = (\alpha\beta)x$

(v) $(\alpha + \beta)x = \alpha x + \beta x$

(vi) $\alpha(x + y) = \alpha x + \alpha y$

(vii) $1x = x$

for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{R}$.

Examples of linear spaces are $X = \mathbb{R}^n$ the space of n-dimensional vectors of real numbers, $X = C(\Omega)$ the space of continuous functions over $\Omega \subseteq \mathbb{R}$ and $X = L_1(\Omega)$ the space of functions f for which the Lebesgue integral $\int_\Omega |f(x)| \mathrm{d}x$ exists and is finite.

Definition 2. Let X, Y be linear spaces. An operator (or mapping) $A\colon X \to Y$ is a *linear operator* if and only if

$$A(\alpha x) = \alpha A(x)$$

and

$$A(x + y) = A(x) + A(y)$$

for all $x, y \in X$ and $\alpha \in \mathbb{R}$.

The set of all linear operators from X to Y is a linear space denoted by $\mathcal{L}(X, Y)$.

The following example is based on the Riemann problem initial data from the example used in Section 1.3.1.

Example 1. An example of a linear operator $A\colon \mathbb{R} \to L_1([-1, 1])$ is

$$A \equiv p \mapsto \Big(x \mapsto pH(x)\Big)$$

where H is the Heaviside step function. We can see that the operator A is linear since

$$\begin{aligned} \alpha A(p_1) + \beta A(p_2) &= x \mapsto \alpha p_1 H(x) + \beta p_2 H(x) \\ &= x \mapsto (\alpha p_1 + \beta p_2) H(x) = A(\alpha p_1 + \beta p_2) \ . \end{aligned}$$

A related operator is

$$B \equiv p \mapsto \Big(x \mapsto (1 + p) H(x)\Big) \ .$$

The operator B is nonlinear since

$$\begin{aligned} \alpha B(p_1) + \beta B(p_2) &= x \mapsto \alpha(1 + p_1) H(x) + \beta(1 + p_2) H(x) \\ &= x \mapsto (\alpha(1 + p_1) + \beta(1 + p_2)) H(x) \\ &= x \mapsto (\alpha + \beta + \alpha p_1 + \beta p_2) H(x) \\ &\neq x \mapsto (1 + \alpha p_1 + \beta p_2) H(x) = B(\alpha p_1 + \beta p_2) \end{aligned}$$

if $\alpha + \beta \neq 1$. In the next section we show that the linear operator A is actually the derivative of the nonlinear operator B.

The following example is also inspired by the example used in Section 1.3.1 and is useful for understanding the solution of the Riemann problem where the location of a discontinuity shifts depending on a parameter.

Example 2. The operator

$$C \equiv p \mapsto \Big(x \mapsto H(x - p)\Big)$$

is nonlinear since

$$\begin{aligned} \alpha C(p_1) + \beta C(p_2) &= x \mapsto \alpha H(x - p_1) + \beta H(x - p_2) \\ &\neq x \mapsto H(x - \alpha p_1 + \beta p_2) = C(\alpha p_1 + \beta p_2) \end{aligned}$$

if not $\alpha, \beta = 1$. In the next section we also consider the question of the derivative of C.

We use the terminology of operator, mapping and function interchangeably for both finite and infinite dimensional vector spaces.

Definition 3. A *norm* on a linear space X is any function $\|\cdot\|\colon X \to \mathbb{R}^+$ that satisfies

(i) $\|\alpha x\| = |\alpha|\|x\|$

(ii) $\|x + y\| \leq \|x\| + \|y\|$

(iii) $\|x\| = 0$ implies $x = 0$

for all $x, y \in X$ and $\alpha \in \mathbb{R}$.

If X is equipped with a norm we call it a *normed linear space.*

Examples of norms are the Euclidean norm

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

for $X = \mathbb{R}^n$, the supremum norm $\|f\|_{\text{sup}} = \sup_{x\in\Omega} |f(x)|$ for $X = C(\Omega)$ and the L_1 norm $\|f\|_{L_1(\Omega)} = \int_\Omega |f(x)|\mathrm{d}x$ for $X = L_1(\Omega)$.

Norms allow us to measure the distance between elements of a linear space and to define notions of convergence.

Definition 4. Let X be a normed linear space then X is a *Banach space* if and only if every Cauchy sequence $\{x_i\}$ with respect to the norm of X converges to an element in X. That means X is a Banach space if and only if for every $\delta > 0$ there exist N such that for all $m, n \geq N$ we have $\|x_n - x_m\| \leq \delta$ implies that $\lim_i x_i$ exists and $\lim_i x_i \in X$. We call the property of a Banach space *completeness.*

Examples of Banach spaces are again $X = \mathbb{R}^n$ with any norm, $X = C(\Omega)$ with $\|\cdot\|_{\text{sup}}$ and $X = L_1(\Omega)$ with $\|\cdot\|_{L_1(\Omega)}$.

2.1.2 Continuous Linear Operators

For normed linear spaces the continuity of a linear operator is related to the boundedness of the norm of the elements it maps to.

Definition 5. Let X, Y be linear spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. An operator $A\colon X \to Y$ is a *continuous operator* if and only if for every sequence $\{x_i\}$ converging to $x^\star$ in X the sequence $\{A(x_i)\}$ converges to $y^\star = A(x^\star)$ in Y.

Definition 6. An operator A is a *bounded operator* if and only if

$$\sup_{\|x\|_X \leq 1} \|A(x)\|_Y < \infty \; .$$

A linear operator is continuous if and only if it is bounded.

For example the linear operator between the finite dimensional linear spaces $X = \mathbb{R}^m, Y = \mathbb{R}^n$ is a matrix $A \in \mathbb{R}^{n\times m}$. If the elements of A are bounded then A is a continuous operator.

Definition 7. The space of linear operators $\mathcal{L}(X, \mathbb{R})$ that map to scalars consists of *functionals.* The space of continuous linear functionals is denoted by X^* and is called the *dual space* of X.

An example of a linear functional in the context of sensitivity analysis is the gradient $\mathrm{d}_x f(x) \in \mathbb{R}^m$ of a scalar multivariate function $f\colon \mathbb{R}^m \to \mathbb{R}$ evaluated at some point $x \in \mathbb{R}^m$. In the case that the scalar function takes a linear space X to $\mathbb{R}$, i.e. $f\colon X \to \mathbb{R}$, and is differentiable the gradient evaluated at some $x \in X$ is an element of the dual space $\mathrm{d}_x f(x) \in X^*$.

2.1.3 Adjoint Operators

We now introduce the notion of the adjoint operator by a definition based on inner products.

Definition 8. An *inner product* on a linear space X is any function $\langle\cdot,\cdot\rangle$ that satisfies

(i) $\langle x, y\rangle \geq 0$

(ii) $\langle x, y\rangle = \langle y, x\rangle$

(iii) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$

(iv) $\langle x, x\rangle = 0$ implies $x = 0$

for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{R}$.

Definition 9. For every bounded linear operator $A\colon X \to Y$ there exists a unique bounded linear operator $A^*\colon Y \to X$ such that

$$\langle A(x), y\rangle_Y = \langle x, A^*(y)\rangle_X$$

for all $x \in X, y \in Y$. The operator A^* is called the *adjoint operator* of A.

An example of a bounded linear operator in the context of sensitivity analysis is the Jacobian matrix of a multivariate differentiable function $f\colon \mathbb{R}^m \to \mathbb{R}^n$ evaluated at some $x \in \mathbb{R}^m$. When talking about adjoint sensitivity analysis we refer to a sensitivity analysis that involves the adjoint operator of the derivative as a linear operator.

The adjoint of a finite dimensional real matrix A is the transposed matrix A^T.

Notably, the definition of the adjoint operator requires an inner product to be defined on both domain and codomain. If we consider the differentiable operator f that maps between the two linear spaces X, Y, i.e. $f\colon X \to Y$ then the derivative A of f at some $x \in X$ also maps between X and Y, i.e. $A\colon X \to Y$ (see the next section for the definition of the derivative on linear spaces). In order for the adjoint of the derivative A to be defined both X and Y need an inner product.

Example 3. The adjoint operator of the operator A from Example 1 is an operator $A^*\colon I \to \mathbb{R}$ where I is the image of A, i.e. $I = \{x \mapsto pH(x) \mid p \in \mathbb{R}\} \subseteq L_1([-1,1])$. A possible definition of the adjoint is $A^*(f(\cdot)) \equiv f(1)$ as that operator satisfies the inner product

$$\begin{aligned}\langle A(p_1), p_2 H(\cdot)\rangle_{L_1([-1,1])} = \langle p_1 H(\cdot), p_2 H(\cdot)\rangle_{L_1([-1,1])} &= \int_{-1}^{1} p_1 p_2 H(x)\mathrm{d}x = p_1 p_2 \int_0^1 \mathrm{d}x \\ &= \langle p_1, p_2\rangle_{\mathbb{R}} = \langle p_1, p_2 H(1)\rangle_{\mathbb{R}} = \langle p_1, A^*(x \mapsto p_2 H(x))\rangle_{\mathbb{R}} \,.\end{aligned}$$

2.2 Differentiation in Normed Linear Spaces

2.2.1 Fréchet Derivative

We now introduce the basics of classical differentiation for operators on normed linear spaces.

Definition 10. Let X, Y be normed linear spaces, $f\colon D \subseteq X \to Y$ be an operator from an open set D and $x \in D$. If there is a bounded linear operator $A\colon X \to Y$ such that

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - Ah\|}{\|h\|} = 0$$

then f is said to be *Frechét differentiable* (or differentiable) at x. The linear operator A is called the *Fréchet derivative* (or derivative) of f at x.

Example 4. For the operators A, B defined in Example 1 we have

$$\frac{\|B(p+h) - B(p) - A(h)\|_{L_1([-1,1])}}{|h|} = \frac{\|(1+p+h)H(\cdot) - (1+p)H(\cdot) - hH(\cdot)\|_{L_1([-1,1])}}{|h|} = 0 \;,$$

i.e. A is the derivative of B.

Example 5. The derivative of the operator C defined in Example 2 is not as straight forward. For the linear operator D that is the derivative of C in the limit of $h \to 0$ we need to have

$$\begin{aligned}
0 &= \frac{\|C(p+h) - C(p) - D(h)\|_{L_1([-1,1])}}{|h|} && \text{(definition of derivative)}\\
&= \frac{\|H(\cdot - p - h) - H(\cdot - p) - D(h)\|_{L_1([-1,1])}}{|h|} && \text{(definition of } C\text{)}\\
&= \frac{1}{|h|}\int_{-1}^{1} |H(x-p-h) - H(x-p) - D(h)|\mathrm{d}x && \text{(definition of } L_1 \text{ norm)}\\
&\geq \frac{1}{|h|}\left|\int_{-1}^{1} |H(x-p-h) - H(x-p)|\mathrm{d}x - \int_{-1}^{1} |D(h)|\mathrm{d}x\right| && \text{(triangle inequality)}\\
&= \frac{1}{|h|}\left|\int_{p-h}^{p} |1|\mathrm{d}x - \int_{-1}^{1} |D(h)|\mathrm{d}x\right| && \text{(w.l.o.g. assuming } h>0\text{)}\\
&= \frac{1}{|h|}\left||h| - \int_{-1}^{1} |D(h)|\mathrm{d}x\right| && \text{(definite integral of constant)}\\
&= \left|1 - \frac{1}{|h|}\int_{-1}^{1} |D(h)|\mathrm{d}x\right| && \text{(definition of norm) .}
\end{aligned}$$

In the section on distributions we introduce the Dirac delta distribution as a linear operator that can satisfy the above. Since the Dirac delta distribution is defined to have $\int_{-1}^{1} |\delta(x)|\mathrm{d}x = 1$ by letting

$$D \equiv h \mapsto \Big(x \mapsto h\delta(x)\Big)$$

we have

$$\frac{1}{|h|}\int_{-1}^{1} |D(h)|\mathrm{d}x = \frac{1}{|h|}\int_{-1}^{1} |h\delta(x)|\mathrm{d}x = \int_{-1}^{1} |\delta(x)|\mathrm{d}x = 1 \;.$$

The Dirac delta distribution is not an integrable function so D is not a valid derivative in the sense of the L_1 norm. It is a weak derivative since in order to formally define the Dirac delta distribution we need to apply it to a test function (see the section on distributions).

2.2.2 Differential Calculus

We now give some of the classical theorems from differential calculus in a way that they apply to general normed linear spaces.

Theorem 1 (Chain Rule)**.** If f is differentiable at x with derivative $f'(x)$ and g is differentiable at $f(x)$ with derivative $g'(f(x))$ then $g \circ f$ is differentiable at x and the derivative is

$$(g \circ f)' = g'(f(x)) \cdot f'(x) \; .$$

Proof. See Cheney [Che13, Chapter 3.2]. □

Theorem 2 (Mean Value Theorem)**.** Let X, Y be normed linear spaces, $f \colon D \subseteq X \to Y$ an operator from an open set D, and differentiable on D. If the line segment

$$S = \{ta + (1-t)b \colon 0 \leq t \leq 1\}$$

lies in D then

$$\|f(b) - f(a)\| \leq \|b - a\| \sup_{x \in S} \|f'(x)\| \; .$$

Proof. See Cheney [Che13, Chapter 3.2]. □

Based on the chain rule and the mean value theorem we can for example show the quadratic error bound of the linearization of a twice differentiable nonlinear function f (c.f. the first-order Taylor expansion).

Proposition 3. Let X, Y be normed linear spaces, $f \colon D \subseteq X \to Y$ an operator from an open set D, and twice differentiable on D. If the line segment

$$S = \{ta + (1-t)b \colon 0 \leq t \leq 1\}$$

lies in D then

$$\|f(a + t(b-a)) - f(a) - tf'(a)(b-a)\| \leq \|b-a\|^2 \sup_{\tau \in [0,t]} \|f''(a + \tau(b-a))\| \; .$$

Proof. Let $g(t) \equiv f(a + t(b-a))$ and $h(t) \equiv g(t) - tf'(a)(b-a)$. We have

$$\begin{aligned} \|f(a + t(b-a)) - f(a) - tf'(a)(b-a)\| &= \|g(t) - g(0) - tf'(a)(b-a)\| && \text{(definition of } g\text{)} \\ &= \|h(t) - h(0)\| && \text{(definition of } h\text{)} \\ &\leq \sup_{\tau \in [0,t]} \|h'(\tau)\| && \text{(mean value th.)} \; . \end{aligned}$$

We have the upper bound

$$\begin{aligned} \|h'(\tau)\| &= \|g'(\tau) - f'(a)(b-a)\| && \text{(definition of } h \text{ and chain rule)} \\ &= \|f'(a + \tau(b-a))(b-a) - f'(a)(b-a)\| && \text{(definition of } g \text{ and chain rule)} \\ &= \|(f'(a + \tau(b-a)) - f'(a))(b-a)\| && \text{(linearity of the derivative)} \\ &\leq \|b-a\| \|f'(a + \tau(b-a)) - f'(a)\| && \text{(Cauchy-Schwartz inequality)} \end{aligned}$$

$$\leq \|b-a\|^2 \sup_{\sigma \in [0,\tau]} \|f''(a+\sigma(b-a))\| \qquad\qquad \text{(mean value theorem)} .$$

By substitution of the upper bound in the equation above we get the quadratic bound

$$\begin{aligned} \|f(a+t(b-a)) - f(a) - tf'(a)(b-a)\| &\leq \|b-a\|^2 \sup_{\tau \in [0,t]} \sup_{\sigma \in [0,\tau]} \|f''(a+\sigma(b-a))\| \\ &\leq \|b-a\|^2 \sup_{\tau \in [0,t]} \|f''(a+\tau(b-a))\| \end{aligned}$$

assuming that all of the required derivatives and (operator) norms exist. This concludes the proof. □

Theorem 4 (Implicit Function Theorem)**.** Let X, Y, Z be normed linear spaces and Y a Banach space, Ω be an open set in $X \times Y$, $F \colon \Omega \to Z$, $(x_0, y_0) \in \Omega$. Furthermore, let F be continuous at (x_0, y_0), $F(x_0, y_0) = 0$, $\partial_y F$ exist in Ω, $\partial_y F$ be continuous at (x_0, y_0) and $\partial_y F(x_0, y_0)$ be invertible. Then there is a function f defined on a neighborhood of x_0 such that $F(x, f(x)) = 0$, $f(x_0) = y_0$, f is continuous at x_0 and f is unique.

If F is continuously differentiable in Ω then the function f will be continuously differentiable and

$$f'(x) = -\partial_y F(x, f(x))^{-1} \cdot \partial_x F(x, f(x)) .$$

Furthermore, there will exist a neighborhood of x_0 in which f is unique.

Proof. See Cheney [Che13, Chapter 3.4]. □

We note that $f'(x)$ satisfies

$$\begin{aligned} \mathrm{d}_x F(x, f(x)) &= \partial_x F(x, f(x)) + \partial_y F(x, f(x)) f'(x) \\ &= \partial_x F(x, f(x)) + \partial_y F(x, f(x))(-\partial_y F^{-1}(x, f(x)) \cdot \partial_x F(x, f(x))) = 0 . \end{aligned}$$

Intuitively, for an infinitesimal perturbation of x in a neighborhood of x_0 the equation stays at zero because f changes accordingly.

2.2.3 Lipschitz continuity

We now define the notion of (local) Lipschitz continuity of operators.

Definition 11. Let X, Y be normed linear spaces. An operator $f \colon X \to Y$ is *Lipschitz continuous* on X if there exists $C \geq 0$ such that

$$\|f(x) - f(y)\| \leq C\|x - y\|$$

for all $x, y \in X$. An operator is *locally Lipschitz continuous* on D if it is Lipschitz continuous on a neighborhood of x for all $x \in D$.

Lipschitz continuity is a useful property to prove bounds for approximation errors. A local Lipschitz constant for a continuously differentiable function naturally arises from a bound of its derivative.

Lemma 5. Let U be an open set in $\mathbb{R}^m$ and $f\colon U \to \mathbb{R}^n$ is continuously differentiable on U. Then f is locally Lipschitz continuous on U.

Proof. Since f is differentiable the derivative f' exists with $\|f'(x)\| < \infty$ for all $x \in U$. Since f is continuously differentiable the derivative depends continuously on x. So for all $x \in U$ there exists a neighborhood of x in which the derivative is bounded by some constant $L > 0$, i.e. $\|f'(y)\| \leq L$ for all y in the neighborhood of x. By the mean value theorem we have

$$\|f(x) - f(y)\| \leq \|x - y\| \sup_{t\in[0,1]} \|f'(x + t(x - y))\| \leq L\|x - y\|$$

for all y in the neighborhood of X. So the function is locally Lipschitz continuous with local Lipschitz constant L. This concludes the proof. □

2.3 Lebesgue Integration

We now introduce the basics of measure theoretic integration. Measure theoretic integration as a formalism becomes necessary when integrating over function spaces, for example when evaluating expectations of stochastic processes in probability theory. In this thesis we do not integrate over function spaces but rather introduce the notion of the Lebesgue measure in order to formally define the linear space L_1.

2.3.1 Measurable Sets and Functions

We now define what it means for a set to be measurable and based on that definition what it means for a function to be measurable.

Definition 12. Let X be a normed linear space. The *Borel σ-field* $\mathcal{B}(X)$ is the smallest family of sets from X that contains all open subsets of X and satisfies

(i) $A \in \mathcal{B}(X)$ implies $X \setminus A \in \mathcal{B}(X)$

(ii) $A_i \in \mathcal{B}(X)$ for $i = 1, 2, \dots$ implies that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}(X)$.

A subset A of X is *measurable* if and only if $A \in \mathcal{B}(X)$.

A useful intuition is that complex sets in a Borel σ-field consist of unions of at most countably infinite open neighborhoods and their corresponding complements. In the setting of measure theory we, therefore, do not have to directly deal with the uncountably infinite possible subsets of e.g. the real numbers. That is not to say that the Borel σ-field does not consist of an uncountable number of sets, because it does.

An example of the Borel σ-field $\mathcal{B}(\mathbb{R})$ consists of the open intervals with rational end points, their corresponding complements and countable unions. For example the closed point set $\{\sqrt{2}\} \in \mathcal{B}(\mathbb{R})$ is the complement of the countable union

$$\bigcup_{i=1}^{\infty} \big((-\infty, l_i) \cup (r_i, \infty)\big)$$

for a series of rational approximations $l_i, r_i \in \mathbb{Q}$ with $l_i < \sqrt{2} < r_i$ and $l_i, r_i \to \sqrt{2}$ as $i \to \infty$.

Every closed set is measurable.

Definition 13. Let Y be a normed linear space. A function $f\colon X \to Y$ is measurable if and only if the preimage of a measurable set in Y is measurable in X, i.e. $f^{-1}(A) \in \mathcal{B}(X)$ for all $A \in \mathcal{B}(Y)$.

Every continuous f is measurable. A trivial example of a nonmeasurable function is via some nonmeasurable set $S \subseteq \mathbb{R}$. Let

$$f(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

then the preimage of the measurable set $\{1\} \in \mathcal{B}(\mathbb{R})$ is S which is not measurable in $\mathbb{R}$ by definition. Hence, the function f is not measurable.

2.3.2 Lebesgue Measure

We now define the notion of the measure of a measurable set and specifically the Lebesgue measure for measurable subsets of $\mathbb{R}$.

Definition 14. A function $\mu\colon \mathcal{B}(X) \to \mathbb{R}_+ \cup \{\infty\}$ is a *measure* if and only if it satisfies

(i) $\mu(\emptyset) = 0$

(ii) for every sequence of disjoint and measurable sets A_i we have

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i)\ .$$

Definition 15. For $X = \mathbb{R}^n$ the *Lebegue measure* is uniquely defined by $\mu(A) = \prod_{i=1}^{n}(b_i - a_i)$ for the set $A = I_1 \times \cdots \times I_n$ with $I_i = [a_i, b_i]$, $I_i = [a_i, b_i)$, $I_i = (a_i, b_i]$ or $I_i = (a_i, b_i)$ with $a_i < b_i$.

2.3.3 Lebesgue Integration

We now define the notion of Lebesgue integration and Lebesgue integrable functions.

Definition 16. Let f be a measurable nonnegative function, i.e. $f(x) \geq 0$ for all $x \in X$ and let E be a measurable subset of X. The *Lebesgue integral* of f over E is defined by

$$\int_E f(x)\mu(\mathrm{d}x) = \sup_{\{E_i\}} \sum_{i=1}^{\infty} \mu(E_i) \inf_{x \in E_i} f(x)$$

where the supremum is such that the sets E_i are measurable, disjoint and $\bigcup_{i=1}^{\infty} E_i = E$.

Definition 17. The function f is *Lebesgue integrable* if its Lebesgue integral is finite.

The function f is integrable if and only if $|f|$ is integrable and

$$\Big|\int_E f(x)\mu(\mathrm{d}x)\Big| \leq \int_E |f(x)|\mu(\mathrm{d}x) < \infty \, .$$

When integrating over a subset of the reals we typically use the standard integral notation, i.e. for $E \subseteq \mathbb{R}^n$ we write

$$\int_E f(x)\mathrm{d}^n x \equiv \int_E f(x)\mu(\mathrm{d}x)$$

where μ is the Lebesgue measure.

The linear space $L_p(\Omega)$ consists of measurable functions for which the p-th power of the absolute value is integrable over $\Omega \subseteq \mathbb{R}^n$, i.e. for the function $f\colon \Omega \to \mathbb{R}$ we have $f \in L_p(\Omega)$ if and only if

$$\|f\|_{L_p(\Omega)} = \Big(\int_\Omega |f(x)|^p \mathrm{d}^n x\Big)^{-p} < \infty \, .$$

The linear space $L_1(\Omega)$ is the space of integrable functions.

An example of two integrable function $f, g \in L_1(\mathbb{R})$ are

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise} \end{cases} , \qquad g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{otherwise} \end{cases} .$$

We have $f(0) = 1 \neq 0 = g(0)$ but since they are equal almost everywhere, i.e. equal except on a set of Lebesgue measure zero, the functions f and g are identified in the sense of $L_1(\mathbb{R})$. The notion of equivalence modulo differences on sets of measure zero can be algebraically formalized via the quotient space $L_1(\mathbb{R})/\mathcal{N}$ where $\mathcal{N} \equiv \{f \mid f = 0 \text{ almost everywhere}\}$. The quotient of a linear space by a subset $\mathcal{N}$ is obtained by identifying all elements of $\mathcal{N}$ as zero which results in the equivalence relation of $f = g$ if and only if $f - g \in \mathcal{N}$ [Hal87, Chapter 21].

2.3.4 Notions of Convergence

Related to the notion of equivalence according to the L_1 norm we can also consider different kinds of convergence for functions.

Definition 18. Let $\{f_n\}$ be a sequence of functions $f_n\colon D \subseteq \mathbb{R} \to \mathbb{R}$ with $n = 1, 2, \ldots$ we say $\{f_n\}$ is *pointwise convergent* to f if and only if

$$\lim_{n\to\infty} f_n(x) = f(x)$$

for all $x \in D$.

We say $\{f_n\}$ is *almost everywhere convergent* if the set of points $S \subseteq D$ where for $x \in S$ we have $\lim_{n\to\infty} f_n(x) \neq f(x)$ is a set of measure zero.

We say a sequence of integrable function $\{f_n\}$ is convergent in $L_1(D)$ if

$$\lim_{n\to\infty} \|f_n - f\|_{L_1(D)} = 0 \, .$$

Almost everywhere convergence implies convergence in $L_1(D)$.

2.4 Distributions

We now introduce the formalism of distributions as generalized functions that are used for the differentiation of discontinuous solutions of conservation laws.

2.4.1 Test Functions

We start by defining the space of test functions on which distributions act.

Definition 19. A *test function* $\phi\colon \mathbb{R}^n \to \mathbb{R}$ is a function that is infinitely differentiable on $\mathbb{R}^n$ and vanishes outside some bounded domain, i.e. the test function has compact support. The space of test functions on $\mathbb{R}^n$ is denoted by $C_0^\infty(\mathbb{R}^n)$.

An example of such a smooth test function on $\mathbb{R}$ is the following function

$$\phi(x) = \begin{cases} \exp\left(\frac{1}{x^2-1}\right) & \text{if } |x| < 1 \\ 0 & \text{otherwise .} \end{cases} \tag{2.4.1}$$

For this function we have $\phi(x) \to 0$ as $x \to 1$ with $x < 1$. Similarly, we have $\phi'(x) \to 0$ as $x \to 1$ with $x < 1$ and so on for higher-order derivatives.

2.4.2 Distributions

We recall that by definition of a functional on a linear space X the linear operator (or function) f is a linear functional on $X = C_0^\infty(\mathbb{R}^n)$ if it maps every $\phi \in C_0^\infty(\mathbb{R}^n)$ to a real number.

Definition 20. A *distribution* is a continuous linear functional on $C_0^\infty(\mathbb{R}^n)$. We denote the application of the distribution f to the test function ϕ by an inner product notation $\langle f, \phi \rangle$ instead of $f(\phi)$.

For example a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is a distribution by the definition of

$$\langle f, \phi \rangle \equiv \int_{\mathbb{R}^n} f(x)\phi(x)\mathrm{d}^n x \ .$$

A distribution is regular if it can be written as above with f locally integrable.

We now define the Dirac delta as a distribution.

Definition 21. The *Dirac delta* distribution δ_ξ is defined by

$$\langle \delta_\xi, \phi \rangle \equiv \phi(\xi) \ .$$

The Dirac delta distribution is a singular distribution and not a function in the classical sense.

The Dirac delta is a distribution since it is linear and it maps all test functions to a bounded set. For brevity, we often prefer to use a notation as if the Dirac delta distribution were an integrable function by writing for example

$$\int_{\mathbb{R}^n} \delta(x - \xi)\phi(x)\mathrm{d}^n x \equiv \langle \delta_\xi, \phi \rangle \ .$$

2.4.3 Weak Differentiation

If a function $f\colon \mathbb{R} \to \mathbb{R}$ is differentiable and $f'\colon \mathbb{R} \to \mathbb{R}$ is locally integrable then f' is a distribution with

$$\langle f', \phi \rangle = \int_{-\infty}^{\infty} f'(x)\phi(x)\mathrm{d}x = -\int_{-\infty}^{\infty} f(x)\phi'(x)\mathrm{d}x = -\langle f, \phi' \rangle$$

by integration by parts. Since ϕ is infinitely differentiable and ϕ' also has compact support ϕ' is also a test function and the above is a legitimate application of the distribution f to a test function. It turns out that we can use integration by parts as a general approach to define the derivative of distributions.

Definition 22. Let f be a distribution then its *weak derivative* f' is a distribution defined by

$$\langle f', \phi \rangle \equiv -\langle f, \phi' \rangle \ .$$

Let us for example consider the Heaviside step function with

$$\langle H', \phi \rangle = -\langle H, \phi' \rangle = -\int_{0}^{\infty} \phi'(x)\mathrm{d}x = \phi(0) \ .$$

In a distributional sense H' is identical to the Dirac delta distribution δ_0 since their action on the test function ϕ is the same. This idea leads to a calculus of weak differentiation for functions with jump discontinuities as they can be represented by a linear combination of Heaviside step functions multiplied by continuous differentiable functions.

2.4.4 Weak Convergence

The idea of the application of distributions to test functions allows us to define a notion of weak convergence for distributions as follows

Definition 23. Let $\{f_n\}$ be a sequence of distributions with $n = 1, 2, \dots$ we say $\{f_n\}$ is *weakly convergent* to f if and only if

$$\lim_{n\to\infty} \langle f_n, \phi \rangle = \langle f, \phi \rangle$$

for all smooth test functions ϕ with compact support.

The notion of distributions as a formalism for generalized functions does not answer the question of how to make sense of products of distributions, e.g. the product of a Dirac delta distribution with the Heaviside step function. But we can certainly consider the limit of products of smooth functions approximating the distributions such as the smooth approximation of the Dirac delta distribution via $\mathrm{sech}(\cdot)^2$ in the introduction. The Colombeau algebra we introduce in the next section turns the idea of limits of smooth approximations of distributions into a formalism for generalized functions that allows us to multiply distributions.

2.4.5 Colombeau Algebra

We now introduce the Colombeau algebra [Col85] on $\mathbb{R}$ that defines generalized functions like the Dirac delta distribution based on limits of smooth functions and identifies them modulo the differences that occur outside of the limit. The Colombeau algebra enables us to define the adjoint derivative operator of a discontinuous weak solution of a conservation law through an inner product that involves the multiplication of generalized functions (see Section 5.4.2).

We recall that the set of infinitely differentiable functions on $\mathbb{R}$ is $C^\infty(\mathbb{R})$ and the set of infinitely differentiable functions with compact support (test functions) is $C_0^\infty(\mathbb{R})$.

Definition 24. A *(Colombeau) generalized function* is an operator

$$\boldsymbol{\psi}\colon C_0^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$$

that maps a test function $\phi \in C_0^\infty(\mathbb{R})$ to a smooth function.

Based on the definition of the shifted test function $\phi^y(x) = \phi(x-y)$ the generalized function $\boldsymbol{\psi}$ corresponding to the distribution ψ is defined by

$$\boldsymbol{\psi}[\phi] \equiv \Big(y \mapsto \langle \psi, \phi^y \rangle\Big) = \int_{-\infty}^{\infty} \psi(x)\phi(x-\cdot)\mathrm{d}x \; .$$

We denote the set of generalized functions as

$$\mathcal{H}(\mathbb{R}) = \{\boldsymbol{\psi} \mid \boldsymbol{\psi}\colon C_0^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})\} \; .$$

Let $\boldsymbol{\psi}, \boldsymbol{\omega} \in \mathcal{H}(\mathbb{R})$ be two generalized functions. The sum $\boldsymbol{\psi} + \boldsymbol{\omega} \in \mathcal{H}(\mathbb{R})$ is defined as

$$(\boldsymbol{\psi} + \boldsymbol{\omega})[\phi] \equiv \Big(y \mapsto \boldsymbol{\psi}[\phi](y) + \boldsymbol{\omega}[\phi](y)\Big) \; .$$

The product $\boldsymbol{\psi} \cdot \boldsymbol{\omega} \in \mathcal{H}(\mathbb{R})$ is defined as

$$(\boldsymbol{\psi} \cdot \boldsymbol{\omega})[\phi] \equiv \Big(y \mapsto \boldsymbol{\psi}[\phi](y) \cdot \boldsymbol{\omega}[\phi](y)\Big) \; .$$

For example the Dirac delta distribution as a Colombeau generalized function is defined by

$$\boldsymbol{\delta}[\phi] \equiv \int_{-\infty}^{\infty} \delta(x)\phi(x-\cdot)\mathrm{d}x = \Big(y \mapsto \phi(-y)\Big) \; .$$

The product of two such Dirac deltas $\boldsymbol{\delta}^2$ is defined as

$$\boldsymbol{\delta}^2[\phi] \equiv \Big(y \mapsto \phi(-y)^2\Big) \; .$$

We also use the notation as if the Dirac delta distribution were an integrable function by writing for example

$$\int_{-\infty}^{\infty} \delta(x-\xi)\delta(x-\xi)\phi(x)\mathrm{d}x = \phi(\xi)^2 = \boldsymbol{\delta}^2[\phi](-\xi) \; .$$

A generalized function is an operator that consumes a test function and gives a smooth function as a result. For a smooth function f we can, therefore, also define the generalized function

$$\boldsymbol{F}[\phi] \equiv \Big(y \mapsto f(y)\Big) \; .$$

The definition we gave for the generalized function corresponding to f as a distribution is

$$\boldsymbol{f}[\phi] \equiv \Big(y \mapsto \int_{-\infty}^{\infty} f(x)\phi(x-y)\mathrm{d}x\Big) \; .$$

We have two different generalized functions $\boldsymbol{F}$ and $\boldsymbol{f}$ corresponding to f and they are not equal.

The Colombeau algebra on $\mathbb{R}$ is the quotient space

$$\mathcal{G}(\mathbb{R}) = \mathcal{E}(\mathbb{R})/\mathcal{N}(\mathbb{R})$$

of so-called moderate generalized functions $\mathcal{E}(\mathbb{R}) \subseteq \mathcal{H}(\mathbb{R})$ and with $\mathcal{N}(\mathbb{R})$ defined such that

$$\boldsymbol{F} - \boldsymbol{f} \in \mathcal{N}(\mathbb{R}),$$

i.e. the generalized functions $\boldsymbol{F}$ and $\boldsymbol{f}$ as defined above are equal with respect to the Colombeau algebra. We now give a brief informal introduction to what it means for a test funciton to be moderate and how the equivalence of generalized functions is defined in the Colombeau algebra. For a formal definition of both $\mathcal{E}(\mathbb{R})$ and $\mathcal{N}(\mathbb{R})$ see for example Gratus [Gra13].

In order to get an intuition of how a generalized function acts on certain test function we consider a normalized variation of the test function ϕ defined in Equation (2.4.1) in the previous section

$$\phi_{\text{norm}}(x) \equiv \phi(x) / \int_{-\infty}^{\infty} \phi(x)\mathrm{d}x \ .$$

The effect of the normalization is that the test function integrates to one

$$\int_{-\infty}^{\infty} \phi_{\text{norm}}(x)\mathrm{d}x = 1 \ .$$

Further, we consider another modification of the normalized test function that scales the width of the support of the function by a parameter ϵ

$$\phi_\epsilon(x) = \frac{1}{\epsilon}\phi_{\text{norm}}\left(\frac{x}{\epsilon}\right) \ . \tag{2.4.2}$$

The test function ϕ_ϵ has compact support on $[-\epsilon, \epsilon]$ and integrates to one for all $\epsilon > 0$. Given the distribution ψ the function obtained by mapping the test function ϕ_ϵ via the generalized function $\boldsymbol{\psi}$ is a smooth approximation of ψ with the smoothing width ϵ where ϕ_ϵ is the convolution kernel.

Figure 2.1 shows $\boldsymbol{\delta}[\phi_\epsilon] = \phi_\epsilon(\cdot)$ on the left and $\boldsymbol{H}[\phi_\epsilon]$ on the right for different values of ϵ.

Intuitively, two generalized functions $\boldsymbol{\psi}$ and $\boldsymbol{\omega}$ are defined to be equivalent if $\boldsymbol{\psi}[\phi_\epsilon]$ and $\boldsymbol{\omega}[\phi_\epsilon]$ are equal in the limit of $\epsilon \to 0$. For a smooth function f the generalized functions $\boldsymbol{F}$ and $\boldsymbol{f}$ as defined above are equal in the limit of $\epsilon \to 0$. The formal definition of the set $\mathcal{N}(\mathbb{R})$ requires a more involved construction where all the derivatives are ensured to be equal in the limit of $\epsilon \to 0$ as well.

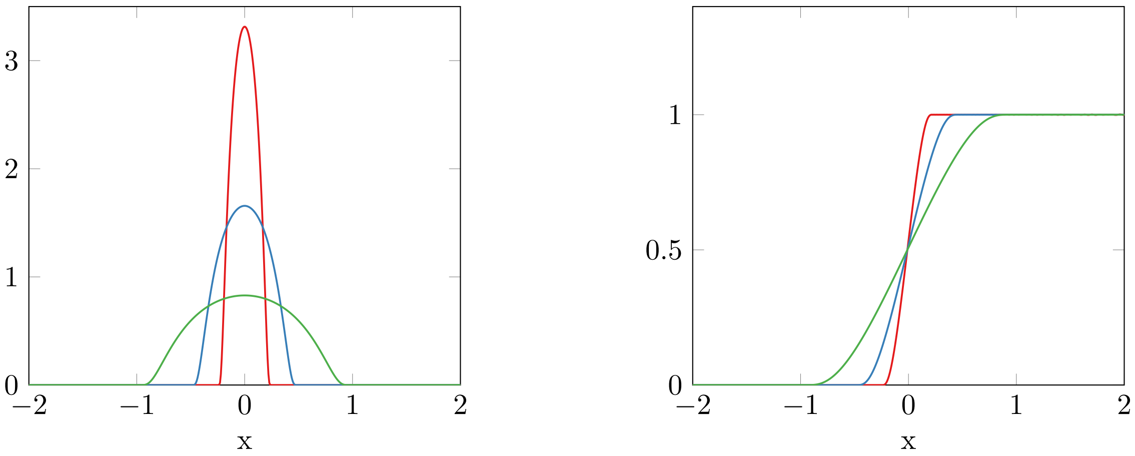
A generalized function $\boldsymbol{\psi}$ is *moderate* if its values do not grow too quickly as $\epsilon \to 0$. More specifically $\boldsymbol{\psi}$ is moderate if there exists a $N \in \mathbb{N}$ such that

$$\max_y |\boldsymbol{\psi}[\phi_\epsilon](y)| = O(\epsilon^{-N}) \ .$$

For example for the Dirac delta distribution we have

$$\max_y |\boldsymbol{\delta}[\phi_\epsilon](y)| = O(\epsilon^{-1}),$$

i.e. the Dirac delta distribution is a moderate generalized function and in the Colombeau algebra. The formal definition of the set of moderate generalized functions $\mathcal{E}(\mathbb{R})$ requires a more involved definition of parametric test functions than we used here.



(a) $\boldsymbol{\delta}[\phi_\epsilon] = \phi_\epsilon(\cdot)$ (Dirac delta applied to symmetric test function)

(b) $\boldsymbol{H}[\phi_\epsilon]$ (Heaviside step function applied to test function)

Figure 2.1: $\epsilon = 1/4$ (red), $\epsilon = 1/2$ (blue), $\epsilon = 1$ (green)

2.5 Auxiliary Results

2.5.1 Divergence Theorem

We now give a theorem that is a useful tool when proving properties by integrating over discontinuous solutions of conservation laws.

Definition 25. Let $f\colon D \subseteq \mathbb{R}^m \to \mathbb{R}$ be a continuous function. The *line integral* of f along a piecewise smooth curve $C \subset D$ is defined as

$$\int_C f(r)\mathrm{d}s \equiv \int_a^b f(r(t))\|r'(t)\|_2 \mathrm{d}t \,.$$

where $r\colon [a,b] \to \mathbb{R}^m$ is a piecewise continuously differentiable parameterization of the curve, i.e. $C = \{r(t) \mid t \in [a,b]\}$.

The line integral gives the area of f under the curve. The notation $\oint_C f(r)\mathrm{d}s$ is used if C is a closed curve. A surface integral is a generalization of the line integral to higher dimensions where the parameterization of the surface requires a multivariate function.

The following is the general divergence theorem that requires a surface integral. Throughout the thesis we only use the case with $m = 2$ where the surface integral in the divergence theorem becomes a line integral.

Theorem 6 (Divergence Theorem)**.** Let $D \subseteq \mathbb{R}^m$ be compact set with piecewise smooth boundary $C \subset D$ and $F\colon R^m \to \mathbb{R}^m$ a continuously differentiable function defined on a neighborhood of D. Then we have

$$\int_D \sum_{i=1}^m \partial_{x_i} F_i(x) \mathrm{d}^m x = \oint_C \langle F(r), n(r)\rangle \mathrm{d}^{m-1}s$$

where the right-hand side denotes the surface integral over C and $n(r)$ denotes the outward pointing unit normal at each point of the boundary.

Proof. The divergence theorem is a special case of Stoke's theorem. For a proof of Stoke's theorem see [Rud64, Chapter 11]. □

2.5.2 Discrete Gronwall Lemma

We now give a lemma that is a useful tool for proving global convergence rates of numerical integrators.

Lemma 7 (Discrete Gronwall Lemma)**.** Let $\{x_n\}$ be a sequence of real numbers with $n = 0, 1, \dots$ and the following inequality holds

$$|x_{n+1}| \leq (1+\delta)|x_n| + \beta$$

for some $\delta > 0$, $\beta \geq 0$. Then we have

$$|x_n| \leq \exp(n\delta)|x_1| + \frac{\exp(n\delta) - 1}{\delta}\beta$$

for all $n = 0, 1, \dots$.

Proof. The proof is by induction.

We have $|x_1| \leq (1+\delta)|x_0| + \beta$.

Due to convexity of exp we have $\exp(\delta) \geq 1 + \delta$, i.e. $(\exp(\delta) - 1)/\delta \geq 1$. Putting these together we get

$$\exp(\delta)|x_1| + \frac{\exp(n\delta) - 1}{\delta}\beta \geq (1+\delta)|x_1| + \beta$$

which concludes the base case.

Now, by induction hypothesis we have

$$|x_n| \leq \exp(n\delta)|x_1| + \frac{\exp(n\delta) - 1}{\delta}\beta$$

and

$$|x_{n+1}| \leq (1+\delta)|x_n| + \beta \ .$$

Then

$$\begin{aligned}
|x_{n+1}| &\leq (1+\delta)\Big(\exp(n\delta)|x_1| + \frac{\exp(n\delta) - 1}{\delta}\beta\Big) + \beta && \text{(substitution of } |x_n|\text{)}\\
&\leq \exp(\delta)\exp(n\delta)|x_1| + \frac{\exp(\delta)\exp(n\delta) - \exp(\delta)}{\delta}\beta + \beta && \text{(convexity of exp)}\\
&= \exp((n+1)\delta)|x_1| + \frac{\exp((n+1)\delta) - \exp(\delta) + (1+\delta) - 1}{\delta}\beta && \text{(adding zero)}\\
&\leq \exp((n+1)\delta)|x_1| + \frac{\exp((n+1)\delta) - \exp(\delta) + \exp(\delta) - 1}{\delta}\beta && \text{(convexity of exp)}\\
&\leq \exp((n+1)\delta)|x_1| + \frac{\exp((n+1)\delta) - 1}{\delta}\beta
\end{aligned}$$

which concludes the induction step. This concludes the proof. □

We gave the necessary background in mathematical analysis to consider the tangent and adjoint sensitivity analysis of PDEs and introduced some of the basic mathematical tools that are used in the thesis. In the next chapter we consider scalar conservation laws with discontinuous solutions and their numerical approximations.

Chapter 3

Scalar Conservation Laws

In this chapter we introduce the necessary concepts from the theory of scalar conservation laws and their numerical approximation. We also introduce some of the driving examples of conservation law problem instances used throughout the thesis.

We give the basic definitions and theorems related to solutions of scalar conservation laws that contain discontinuities. We describe how shock discontinuities that are physically relevant differ from those that are not by satisfying an entropy condition and how the dynamics of a shock discontinuity behaves according to the Rankine-Hugoniot condition. Further, we introduce the concepts of discrete conservation, consistency, stability and monotonicity for some numerical discretization schemes for scalar conservation laws. We also discuss how the numerical viscosity of shock capturing schemes relates to the vanishing viscosity solution of the viscous Burgers equation.

For a formal treatment of the existance and uniqueness of solutions of scalar conservation laws we refer the reader to DiPerna [DiP79], Lax [Lax57, Section5] and the references therein. We refer the reader to Douglis [Dou52] and Oleinik [Ole57] for the existence and uniqueness of solutions of hyperbolic conservation laws and to Hopf [Hop50] for the Burgers equation and Kružkov [Kru70] for general scalar equations.

3.1 Discontinuous Solutions

We now consider discontinuous solutions of scalar conservation laws and the dynamics of shock discontinuities according to the Rankine-Hugoniot condition.

3.1.1 Weak Solutions

We give a definition for the notion of an analytic solution of a conservation law where the solution contains discontinuities.

A solution u of Equations (1.1.1)-(1.1.2) that for some t is discontinuous as a function of x is not differentiable with respect to x in the classical sense. Similarly, a solution u that for some

x is discontinuous as a function of t is not differentiable with respect to t in the classical sense. In both cases a PDE that involves $\partial_t u$ and $\partial_x u$, as is implicitly the case in the conservation law by chain rule $\partial_x f(u) = f'(u)\partial_x u$, does not hold everywhere and, hence, the solution u is also not a solution in the classical sense.

Definition 26. A *weak solution* of Equations (1.1.1)-(1.1.2) satisfies

$$0 = \int_{[0,T]} \int_\Omega \big(\partial_t u(t,x,p) + \partial_x f(u(t,x,p))\big)\phi(t,x)\mathrm{d}x\mathrm{d}t$$

for all smooth test functions ϕ with compact support.

A weak solution can be understood via integration by parts to satisfy

$$\begin{aligned}
&\int_{[0,T]} \int_\Omega \big(\partial_t u(t,x,p) + \partial_x f(u(t,x,p))\big)\phi(t,x)\mathrm{d}x\mathrm{d}t \\
&= \int_{[0,T]} \int_\Omega \partial_t u(t,x,p)\phi(t,x)\mathrm{d}x\mathrm{d}t + \int_{[0,T]} \int_\Omega \partial_x f(u(t,x,p))\phi(t,x)\mathrm{d}x\mathrm{d}t \\
&= \int_\Omega \Big[u(t,x,p)\phi(t,x)\Big]_0^T \mathrm{d}x - \int_{[0,T]} \int_\Omega u(t,x,p)\partial_t\phi(t,x)\mathrm{d}x\mathrm{d}t \\
&\quad + \int_{[0,T]} \Big[u(t,x,p)\phi(t,x)\Big]_{-\infty}^{\infty} \mathrm{d}t - \int_{[0,T]} \int_\Omega f(u(t,x,p))\partial_x\phi(t,x)\mathrm{d}x\mathrm{d}t \\
&= \int_\Omega \Big[u(t,x,p)\phi(t,x)\Big]_0^T \mathrm{d}x - \int_{[0,T]} \int_\Omega u(t,x,p)\partial_t\phi(t,x)\mathrm{d}x\mathrm{d}t \\
&\quad - \int_{[0,T]} \int_\Omega f(u(t,x,p))\partial_x\phi(t,x)\mathrm{d}x\mathrm{d}t \;.
\end{aligned}$$

The last step is due to the compact support of ϕ. We see that the weak solution is defined if both u and $f(u)$ are integrable on the domain of the PDE.

A weak solution with discontinuities is not necessarily unique. Multiple weak solutions may exist for the same initial value problem. In practice we want the solution to additionally satisfy entropy conditions that make it a physically relevant solution [LeV92, Chapter 1.3].

Example 6 (Riemann)**.** We consider the parametric Riemann problem instance first introduced in Section 1.3.1. The weak solution for the initial value problem written in terms of the Heaviside step function is

$$u(t,x,p) = H((1+p)t/2 - x)(1+p)$$

for all p in a neighborhood of $p = 0$. When substituting $t = 0$ we recover the initial data.

In order to apply the calculus of weak differentiation introduced in the previous chapter we also write the flux of the discontinuous weak solution in terms of the Heaviside step function as

$$f(u(t,x,p)) = H((1+p)t/2 - x)f(1+p) \;.$$

The weak derivative of H is the Dirac delta distribution δ. According to the calculus of weak differentiation and the chain rule we have

$$\partial_t u(t,x,p) = \partial_t H((1+p)t/2 - x)(1+p) = \delta((1+p)t/2 - x)(1+p)^2/2 \;.$$

The weak derivative of the flux with respect to x is

$$\begin{aligned}\partial_x f(u(t,x,p)) &= \partial_x H((1+p)t/2 - x) f(1+p) \\ &= -\delta((1+p)t/2 - x) f(1+p) \\ &= -\delta((1+p)t/2 - x)(1+p)^2/2\ .\end{aligned}$$

In the sense of distributions we have $\partial_t u(t,x,p) + \partial_x f(u(t,x,p)) = 0$. If we combine this result together with the definition of the weak solution by means of test functions on which the distributional weak derivatives can act we see that u is in fact the weak solution.

That u is a weak solution can also be shown without Dirac delta distributions only via the integration by parts approach derived previously.

Example 7 (Ramp)**.** Inspired by the example in Bressan and Marson [BM95b] we consider another problem instance for the scalar Burgers equation with the following parametric initial data in the shape of a ramp

$$u_0(x,p) = \begin{cases} (1+p)x & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases} .$$

The weak solution for this initial value problem written in terms of Heaviside step functions is

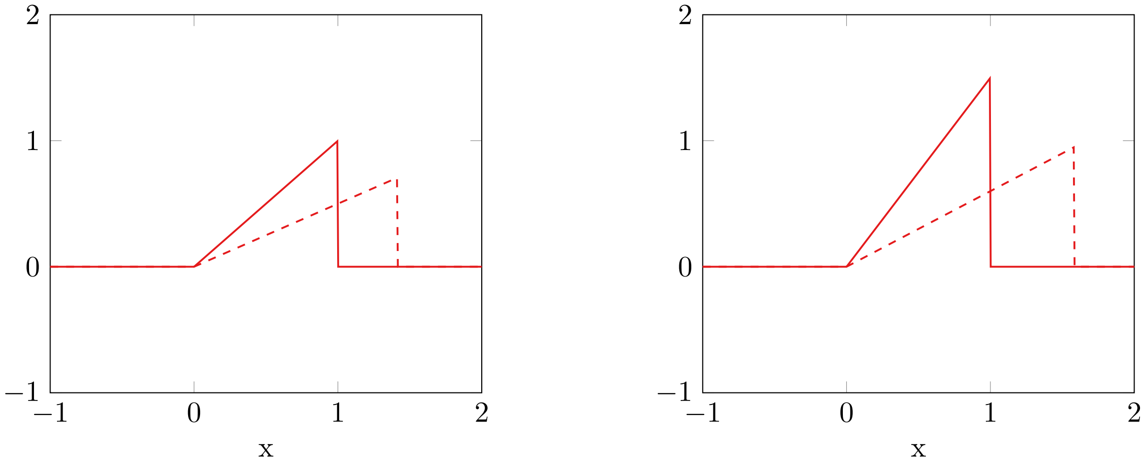
$$u(t,x,p) = H(x) H(\sqrt{1+(1+p)t} - x) \frac{1+p}{1+(1+p)t} x$$

for all p in a neighborhood of $p = 0$. When substituting $t = 0$ we recover the initial data.

Figure 3.1 shows the solution on the domain $x \in [-1, 2]$ at $t = 0$ (solid) and $t = 1$ (dashed) for two different values of p. For higher values of p the shock travels further to the right in the same amount of time. As we see from the analytic solution the shock location for this example is at

$$\xi(t,p) = \sqrt{1+(1+p)t}\ .$$

As opposed to the Riemann problem the shock location is not a linear function of p.



(a) $p = 0$

(b) $p = \frac{1}{2}$

Figure 3.1: Weak solution $u(t,x,p)$ at $t = 0$ (solid), $t = 1$ (dashed)

The flux of the discontinuous weak solution in terms of the Heaviside step function is

$$f(u(t,x,p)) = H(x)H(\sqrt{1+(1+p)t}-x)\frac{1}{2}\frac{(1+p)^2}{(1+(1+p)t)^2}x^2 \ .$$

According to the chain rule the weak derivative of the solution with respect to time t is

$$\begin{aligned}
\partial_t u(t,x,p) &= \partial_t\Big(H(x)H(\sqrt{1+(1+p)t}-x)(1+p)(1+(1+p)t)^{-1}x\Big)\\
&= H(x)\Big(\partial_t H(\sqrt{1+(1+p)t}-x)\Big)(1+p)(1+(1+p)t)^{-1}x\\
&\quad + H(x)H(\sqrt{1+(1+p)t}-x)\Big(\partial_t(1+p)(1+(1+p)t)^{-1}x\Big)\\
&= H(x)\Big(\delta(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+(1+p)t)^{-1/2}(1+p)\Big)(1+p)(1+(1+p)t)^{-1}x\\
&\quad - H(x)H(\sqrt{1+(1+p)t}-x)\Big((1+p)(1+(1+p)t)^{-2}(1+p)x\Big)\\
&= H(x)\delta(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-3/2}x\\
&\quad - H(x)H(\sqrt{1+(1+p)t}-x)(1+p)^2(1+(1+p)t)^{-2}x \ .
\end{aligned}$$

The weak derivative of the flux with respect to space x is

$$\begin{aligned}
\partial_x f(u(t,x,p)) &= \partial_x\Big(H(x)H(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-2}x^2\Big)\\
&= H(x)\Big(\partial_x H(\sqrt{1+(1+p)t}-x)\Big)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-2}x^2\\
&\quad + H(x)H(\sqrt{1+(1+p)t}-x)\Big(\partial_x\frac{1}{2}(1+p)^2(1+(1+p)t)^{-2}x^2\Big)\\
&= -H(x)\delta(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-2}x^2\\
&\quad + H(x)H(\sqrt{1+(1+p)t}-x)(1+p)^2(1+(1+p)t)^{-2}x \ .
\end{aligned}$$

When adding the two derivatives the last terms cancel out respectively and we have

$$\begin{aligned}
\partial_t u(t,x,p) + \partial_x f(u(t,x,p)) &= H(x)\delta(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-3/2}x\\
&\quad - H(x)\delta(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-2}x^2 \ .
\end{aligned}$$

By substitution in Equation (26) that defines the weak solution we have

$$\begin{aligned}
&\int_{[0,T]}\int_\Omega \Big(\partial_t u(t,x,p) + \partial_x f(u(t,x,p))\Big)\phi(t,x)\mathrm{d}x\mathrm{d}t\\
&= \int_{[0,T]}\int_\Omega H(x)\delta(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-3/2}x\phi(t,x)\mathrm{d}x\mathrm{d}t\\
&\quad - \int_{[0,T]}\int_\Omega H(x)\delta(\sqrt{1+(1+p)t}-x)\frac{1}{2}(1+p)^2(1+(1+p)t)^{-2}x^2\phi(t,x)\mathrm{d}x\mathrm{d}t\\
&= \int_{[0,T]} H(\sqrt{1+(1+p)t})\frac{1}{2}(1+p)^2(1+(1+p)t)^{-3/2}\sqrt{1+(1+p)t}\phi(t,\sqrt{1+(1+p)t})\mathrm{d}t\\
&\quad - \int_{[0,T]} H(\sqrt{1+(1+p)t})\frac{1}{2}(1+p)^2(1+(1+p)t)^{-2}\big(\sqrt{1+(1+p)t}\big)^2\phi(t,\sqrt{1+(1+p)t})\mathrm{d}t
\end{aligned}$$

$$
\begin{aligned}
&= \int_{[0,T]} \frac{1}{2}(1+p)^2(1+(1+p)t)^{-1}\phi(t, \sqrt{1+(1+p)t})\mathrm{d}t \\
&\quad - \int_{[0,T]} \frac{1}{2}(1+p)^2(1+(1+p)t)^{-1}\phi(t, \sqrt{1+(1+p)t})\mathrm{d}t = 0\ .
\end{aligned}
$$

We have shown that the solution u is a weak solution of the Burgers equation with the given ramp initial value condition.

3.1.2 Characteristics

We now give the definitions that allow us to distinguish a shock discontinuity from other discontinuities and introduce the entropy condition that holds for physically relevant discontinuities. We also give the Rankine-Hugoniot condition that describes the dynamics of the shock location over time.

Definition 27. A *characteristic* ξ of the solution u is a function of time such that the following ODE holds

$$\partial_t \xi(t,p) = f'(u(t,\xi(t,p),p))$$

for all $t \in [0,T]$.

Assuming differentiability of u its time evolution along the characteristic ξ satisfies the ODE

$$
\begin{aligned}
&\mathrm{d}_t u(t,\xi(t,p),p) \\
&\quad = \partial_t u(t,\xi(t,p),p) + \partial_x u(t,\xi(t,p),p)\partial_t\xi(t,p) && \text{(chain rule)} \\
&\quad = \partial_t u(t,\xi(t,p),p) + \partial_x u(t,\xi(t,p),p) f'(u(t,\xi(t,p),p)) && \text{(definition of characteristic)} \\
&\quad = \partial_t u(t,\xi(t,p),p) + \partial_x f(u(t,\xi(t,p),p)) && \text{(chain rule)} \\
&\quad = 0 && \text{(u solution of conservation law)}
\end{aligned}
$$

for all $t \in [0,T]$. Since $u(t,\xi(t,p),p)$ does not change as t changes that means the solution is constant along the characteristic. In turn $\partial_t\xi(t,p) = f'(u(t,\xi(t,p),p))$ is also constant for all $t \in [0,T]$. So the characteristic is a straight line along which the value of the solution u does not change. The slope of that line, the characteristic speed, is determined by the initial data $f'(u(0,\xi(0,p),p))$.

Intuitively, a characteristic is the line on which information travels starting from each point in the initial data.

3.1.3 Shock Discontinuities

We previously stated that in order for a discontinuous solution to be physically relevant we want to impose an entropy condition. Here, we define a shock discontinuity as the kind of discontinuity that maintains such an entropy condition.

Definition 28. A *shock discontinuity* (or shock) is a discontinuity in u at $x = \xi(t,p)$ where the following entropy condition [LeV92, Chapter 3.8]

$$f'(u(t,\xi(t,p)-,p)) > \partial_t\xi(t,p) > f'(u(t,\xi(t,p)+,p))$$

holds for all $t \in [0,T]$, $p \in D$. A *strong shock discontinuity* is a discontinuity in u at $x = \xi(t,p)$ where the strong entropy condition

$$f'(u(t,\xi(t,p)-,p)) - \delta \geq \partial_t \xi(t,p) \geq f'(u(t,\xi(t,p)+,p)) + \delta$$

with some $\delta > 0$ holds for all $t \in [0,T]$, $p \in D$.

A shock discontinuity is defined by having limiting neighboring characteristics that cross, i.e. information is compressed from left and right of the discontinuity. The speed of the shock itself is greater than the speed of the characteristics to the right of the shock but less than the speed of the characteristics to the left of the shock.

If the entropy condition

$$f'(u(t,\xi(t,p)-,p)) > \partial_t \xi(t,p) > f'(u(t,\xi(t,p)+,p))$$

holds around all discontinuities ξ in u then that is a sufficient condition for the weak solution u to be unique if the flux function f is strongly convex, i.e. $f''(u) \geq \delta$ for some $\delta > 0$ for all $u \in \mathbb{R}$ [Bal70]. A strongly convex function f has a strongly monotonic derivative f', i.e. we have $u^- > u^+$ implies $f'(u^-) > f'(u^+)$. Intuitively, that means that the characteristic speed inside a smooth approximation of the shock lies between the charactistic speeds left and right of the shock.

A more general entropy condition that guarantees a unique solution for nonconvex flux functions is the Oleinik entropy [LeV92, Chapter 3.8]. Even for a nonconvex flux function every physically relevant shock discontinuity will at least satisfy the entropy condition we gave above.

Example 8 (Riemann shock)**.** We consider the parametric Riemann problem example for the Burgers equation with $p = 0$. As we saw in Section 1.3.1 the initial data contains a discontinuity at $x = 0$ and the weak solution $u(t,x) = H(t/2 - x)$ maintains the shock discontinuity at $\xi(t) = t/2$.

Characteristics with $\xi(0) < 0$ have the speed

$$\partial_t \xi(t) = f'(u(0,\xi(0))) = u(0,\xi(0)) = H(-\xi(0)) = 1$$

for all $t \in [0,T]$ and characteristics with $\xi(0) > 0$ have the speed

$$\partial_t \xi(t) = f'(u(0,\xi(0))) = u(0,\xi(0)) = H(-\xi(0)) = 0$$

for all $t \in [0,T]$. The discontinuity at $\xi(t) = t/2$ is a shock discontinuity as it satisfies the entropy condition.

Figure 3.2 shows the characteristics for the Riemann initial data starting from equidistant points along the x axis and running into the shock discontinuity. We observe how the characteristics left and right of the shock point into the shock, i.e. information travels into the shock.

Example 9 (Riemann rarefaction)**.** An example of a discontinuity that does not satisfy the entropy condition of a shock discontinuity is the initial data $u_0(x) = H(x)$. Here, the characteristics point away from the discontinuity and therefore the physically relevant weak solution is a rarefaction wave instead of an entropy violating traveling discontinuity (see LeVeque [LeV92, Chapter 3.5]).

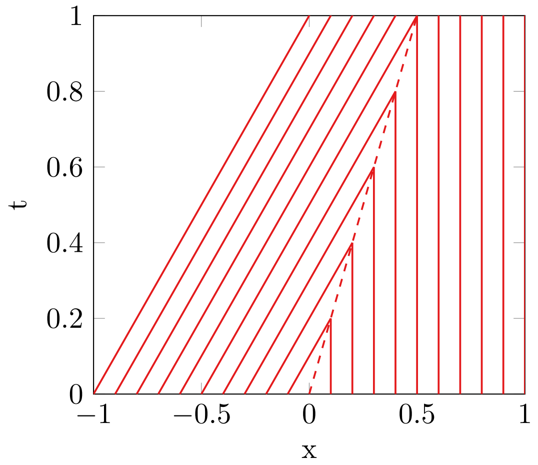


Figure 3.2: Riemann characteristics (solid) and shock location (dashed) for shock wave

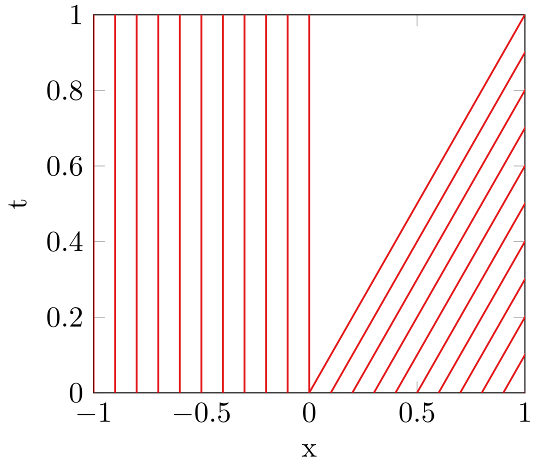


Figure 3.3: Riemann characteristics for rarefaction wave

Figure 3.3 shows the characteristics of the initial data that point away from the shock discontinuity in the initial data at the origin.

We first consider the discontinuous solution $u(t, x) = H(x - t/2)$. Figure 3.4 shows its graph at $t = 1$ on the left.

We have the weak derivative with respect to time

$$\partial_t u(t, x) = -\delta(x - t/2)\frac{1}{2}$$

and the weak derivative with respect to space of the flux

$$\partial_x f(u(t, x)) = \delta(x - t/2)f(1) = \delta(x - t/2)\frac{1}{2} \; .$$

Since in the sense of distributions we have $0 = \partial_t u + \partial_x f(u)$ the given u is a weak solution. But the characteristics right of the shock have the speed one and the characteristics left of the shock have the speed zero, i.e. the entropy condition is violated at the discontinuity. The weak solution u is not a physically relevant solution.

The second solution we consider is the rarefaction wave

$$u(t,x) = \min(1, \max(0, x/t)) \ .$$

Figure 3.4 shows its graph at $t = 1$ on the right.

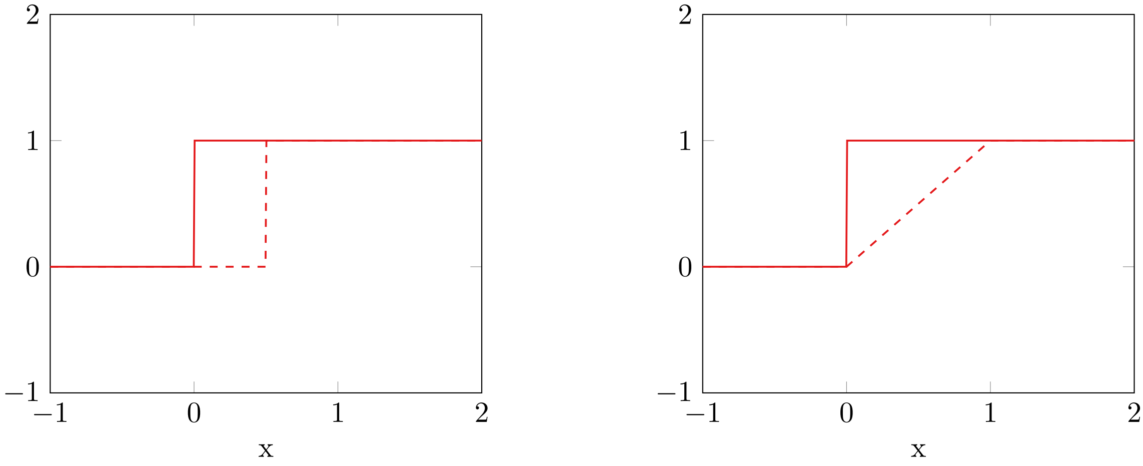
We only show that the rarefaction wave satisfies the conservation law inside of the middle linear part that connects the two constant parts. There, we have the derivative with respect to time

$$\partial_t u(t,x) = -x/t^2$$

and the derivative with respect to space of the flux

$$\partial_x f(u(t,x)) = (x/t)\partial_x(x/t) = x/t^2 \ .$$

Checking the weak derivatives at the kinks of the solution validates the fact that u is also a weak solution of the conservation law and taking the limit of $t \to 0$ converges to the correct initial condition. Since for $t > 0$ the solution does not contain any discontinuities it does not violate the entropy condition and is physically relevant.



(a) Entropy condition violating discontinuity

(b) Entropy condition admissible rarefaction wave

Figure 3.4: Weak solution $u(t,x,p)$ at $t = 0$ (solid), $t = 1$ (dashed)

At each point in time the width of the middle part of the rarefaction wave is as wide as the gap between the two characteristic lines pointing out of the discontinuity at the origin.

Example 10 (Riemann transport)**.** We consider the discontinuous initial data $u_0 = H(-x)$ for the linear conservation law with flux function $f(u) = a \cdot u$. A linear scalar conservation law describes the transport of the solution along the constant characteristics

$$\xi(t) = f'(u(t,\xi(t))) = a \ .$$

The weak solution of a linear scalar conservation law is, therefore, always

$$u(t,x) = u_0(x - a \cdot t) \ .$$

In the case of the discontinuous initial data the solution will be discontinuous as well. But since all characteristics are constant the entropy condition cannot hold for a discontinuity so no discontinuity is a shock discontinuity.

3.1.4 Rankine-Hugoniot Condition

We now give the theorem that describes the dynamics of the shock location providing a generalization of the dynamics of characteristics on the smooth parts of the solution.

Theorem 8 (Rankine-Hugoniot Condition)**.** Let u be a weak solution of Equations (1.1.1)-(1.1.2) such that u has a shock discontinuity along the curve ξ and is smooth on either side of ξ. Then u satisfies the Rankine-Hugoniot condition

$$\xi'(t) = \frac{f(u(t,\xi(t)-)) - f(u(t,\xi(t)+))}{u(t,\xi(t)-) - u(t,\xi(t)+)}$$

across the curve of discontinuity.

Proof. We consider the case of $T = \infty$ and $\Omega = (-\infty, \infty)$ then by integration by parts the weak solution satisfies

$$\int_0^\infty \int_{-\infty}^\infty u(t,x)\partial_t\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}x\mathrm{d}t + \int_{-\infty}^\infty u(0,x)\phi(0,x)\mathrm{d}x = 0$$

for all smooth test functions ϕ with compact support.

We break up the domain of the first integral into two parts; left and right of the discontinuity

$$\begin{aligned}\Omega^- &\equiv \{(t,x)\colon 0 < t < \infty, -\infty < x < \xi(t)\}\\ \Omega^+ &\equiv \{(t,x)\colon 0 < t < \infty, \xi(t) < x < \infty\}.\end{aligned}$$

We make use of the divergence theorem from the previous chapter and consider the case of $C = \partial\Omega^-$ the boundary of the left part of the domain and

$$F(t,x) = \begin{pmatrix} u(t,x)\phi(t,x) \\ f(u(t,x))\phi(t,x) \end{pmatrix} .$$

Since ϕ is a smooth test function with compact support we know that on the left edge of the boundary of Ω^- as $x \to -\infty$ the test function ϕ will take the value zero. By the divergence theorem we have

$$\begin{aligned}&\iint_{\Omega^-} \partial_t(u(t,x)\phi(t,x)) + \partial_x(f(u(t,x)\phi(t,x))\mathrm{d}t\mathrm{d}x\\ &\qquad = \oint_C u(r^-)\phi(r^-)n_1(r^-) + f(u(r^-))\phi(r^-)n_2(r^-)\mathrm{d}s\end{aligned}$$

where $n = (n_1\ n_2)^T$ is the outward pointing normal of C. If we split C into two parts where the line integral does not evaluate to zero we get

$$\begin{aligned}&\oint_C u(r^-)\phi(r^-)n_1(r^-) + f(u(r^-))\phi(r^-)n_2(r^-)\mathrm{d}s\\ &\quad = \oint_{\{(t,\xi(t))\ |\ t\in[0,\infty)\}} u(r^-)\phi(r^-)n_1(r^-) + f(u(r^-))\phi(r^-)n_2(r^-)\mathrm{d}s\\ &\qquad + \oint_{\{(0,x)\ |\ x\in(-\infty,\xi(0))\}} u(r^-)\phi(r^-)n_1(r^-) + f(u(r^-))\phi(r^-)n_2(r^-)\mathrm{d}s\ .\end{aligned}$$

For the second term of the sum we have $n_1 = -1$ and $n_2 = 0$, i.e.

$$\oint_{\{(0,x)\ |\ x\in(-\infty,\xi(0))\}} u(r^-)\phi(r^-)n_1(r^-) + f(u(r^-))\phi(r^-)n_2(r^-)\mathrm{d}s$$
$$= -\oint_{\{(0,x)\ |\ x\in(-\infty,\xi(0))\}} u(r^-)\phi(r^-)\mathrm{d}s = -\int_{-\infty}^{\xi(0)} u(0,x)\phi(0,x)\mathrm{d}x\ .$$

If we differentiate the function in the integral on the left-hand side of the divergence theorem via the chain rule we have

$$\iint_{\Omega^-} \partial_t(u(t,x)\phi(t,x)) + \partial_x(f(u(t,x)\phi(t,x))\mathrm{d}t\mathrm{d}x$$
$$= \iint_{\Omega^-} \partial_t u(t,x)\phi(t,x) + u(t,x)\partial_t\phi(t,x) + \partial_x f(u(t,x)\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}t\mathrm{d}x$$
$$= \iint_{\Omega^-} (\partial_t u(t,x) + \partial_x f(u(t,x))\phi(t,x)\mathrm{d}t\mathrm{d}x + \iint_{\Omega^-} u(t,x)\partial_t\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}t\mathrm{d}x\ .$$

By assumption u is a weak solution and u is smooth on the left side of the shock discontinuity so u is a solution of

$$\partial_t u(t,x) + \partial_x f(u(t,x)) = 0$$

on Ω^-. Consequently, we have the first term above as zero

$$\iint_{\Omega^-} (\partial_t u(t,x) + \partial_x f(u(t,x))\phi(t,x)\mathrm{d}t\mathrm{d}x = 0\ .$$

Substituting the results we obtained so far into the divergence theorem equation we have

$$\iint_{\Omega^-} u(t,x)\partial_t\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}t\mathrm{d}x + \int_{-\infty}^{\xi(0)} u(0,x)\phi(0,x)\mathrm{d}x$$
$$= \oint_{\{(t,\xi(t))\ |\ t\in[0,\infty)\}} u(r^-)\phi(r^-)n_1(r^-) + f(u(r^-))\phi(r^-)n_2(r^-)\mathrm{d}s\ .$$

The analogous steps can be done for Ω^+ to get

$$\iint_{\Omega^+} u(t,x)\partial_t\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}t\mathrm{d}x + \int_{\xi(0)}^{\infty} u(0,x)\phi(0,x)\mathrm{d}x$$
$$= -\oint_{\{(t,\xi(t))\ |\ t\in[0,\infty)\}} u(r^+)\phi(r^+)n_1(r^+) + f(u(r^+))\phi(r^+)n_2(r^+)\mathrm{d}s\ .$$

If we add both left-hand sides together we have

$$\iint_{\Omega^-} u(t,x)\partial_t\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}t\mathrm{d}x + \int_{-\infty}^{\xi(0)} u(0,x)\phi(0,x)\mathrm{d}x$$
$$+ \iint_{\Omega^+} u(t,x)\partial_t\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}t\mathrm{d}x + \int_{\xi(0)}^{\infty} u(0,x)\phi(0,x)\mathrm{d}x$$
$$= \int_0^\infty \int_{-\infty}^{\infty} u(t,x)\partial_t\phi(t,x) + f(u(t,x))\partial_x\phi(t,x)\mathrm{d}t\mathrm{d}x + \int_{-\infty}^{\infty} u(0,x)\phi(0,x)\mathrm{d}x = 0$$

because u is a weak solution. The right-hand sides then necessarily also add to zero and we have

$$\oint_{\{(t,\xi(t))\ |\ t\in[0,\infty)\}} u(r^-)\phi(r^-)n_1(r^-) + f(u(r^-))\phi(r^-)n_2(r^-)\mathrm{d}s$$

$$= \oint_{\{(t,\xi(t))\mid t\in[0,\infty)\}} u(r^+)\phi(r^+)n_1(r^+) + f(u(r^+))\phi(r^+)n_2(r^+)\mathrm{d}s \ .$$

Since the above derivation holds for all smooth test functions ϕ we have

$$u(r^-)n_1(r^-) + f(u(r^-))n_2(r^-) = u(r^+)n_1(r^+) + f(u(r^+))n_2(r^+)$$

almost everywhere. We let $n_1 \equiv n_1(r^-) = n_1(r^+), n_2 \equiv n_2(r^-) = n_2(r^+)$ and get

$$\begin{aligned} n_2(f(u(r^-)) - f(u(r^+))) &= -n_1(u(r^-) - u(r^+)) \\ \frac{f(u(r^-)) - f(u(r^+))}{u(r^-) - u(r^+)} &= -\frac{n_1}{n_2} \ . \end{aligned}$$

Since the normal of the curve $(t, \xi(t))$ is orthogonal to the tangent of ξ its direction at t is

$$\tilde{n}(t) = \begin{pmatrix} \partial_t \xi(t) \\ -1 \end{pmatrix}$$

and $n = \tilde{n}/\|\tilde{n}\|$. We have

$$-\frac{n_1}{n_2} = -\frac{\tilde{n}_1/\|\tilde{n}\|}{\tilde{n}_2/\|\tilde{n}\|} = -\frac{\tilde{n}_1}{\tilde{n}_2} = \partial_t \xi(t) \ .$$

We, therefore, have the Rankine-Hugoniot condition

$$\partial_t \xi(t) = \frac{f(u(t, \xi(t)-)) - f(u(t, \xi(t)+))}{u(t, \xi(t)-) - u(t, \xi(t)+)}$$

for all $t \in [0, \infty)$. This concludes the proof. □

For the Burgers flux $f(u) = u^2/2$ the Rankine-Hugoniot condition can be simplified to

$$\begin{aligned} \partial_t \xi(t) &= \frac{f(u(t, \xi(t)-)) - f(u(t, \xi(t)+))}{u(t, \xi(t)-) - u(t, \xi(t)+)} \\ &= \frac{1}{2} \frac{u(t, \xi(t)-)^2 - u(t, \xi(t)+)^2}{u(t, \xi(t)-) - u(t, \xi(t)+)} \\ &= \frac{1}{2} \big(u(t, \xi(t)-) + u(t, \xi(t)+)\big) \ . \end{aligned}$$

Example 11 (Riemann)**.** We consider the parametric Riemann problem instance for Burgers equation. The weak solution for the initial value problem is

$$u(t, x, p) = H((1 + p)t/2 - x)(1 + p)$$

for all p in a neighborhood of $p = 0$, i.e. its shock location is at

$$\xi(t, p) = \frac{1+p}{2} t \ .$$

By differentiation with respect to t we get

$$\partial_t \xi(t, p) = \frac{1+p}{2} \ .$$

Around the shock discontinuity we have the weak solution values of $u(t, \xi(t)-) = 1 + p$ and $u(t, \xi(t)+) = 0$ so the Rankine-Hugoniot condition is satisfied.

Example 12 (Ramp)**.** We consider the parametric ramp problem instance for Burgers equation defined in Example 7. The weak solution for the initial value problem is

$$u(t,x,p) = H(x)H(\sqrt{1+(1+p)t} - x)\frac{1+p}{1+(1+p)t}x$$

for all p in a neighborhood of $p=0$, i.e. its shock location is at

$$\xi(t,p) = \sqrt{1+(1+p)t}\ .$$

By differentiation with respect to t we get

$$\partial_t \xi(t,p) = -\frac{1}{2}\frac{1+p}{\sqrt{1+(1+p)t}}\ .$$

Around the shock discontinuity we have the weak solution values of

$$u(t,\xi(t)-) = \frac{1+p}{1+(1+p)t}\xi(t)$$

and $u(t,\xi(t)+) = 0$. According to the simplification of the Rankine-Hugoniot condition for the Burgers equation we have

$$\begin{aligned}\partial_t \xi(t,p) &= \frac{1}{2}\Big(\frac{1+p}{1+(1+p)t}\xi(t) + 0\Big)\\ &= \frac{1}{2}\frac{1+p}{1+(1+p)t}\sqrt{1+(1+p)t}\\ &= \frac{1}{2}\frac{1+p}{\sqrt{1+(1+p)t}}\ .\end{aligned}$$

The Rankine-Hugoniot condition is satisfied.

We have introduced the necessary background on discontinuous solutions of scalar conservation laws and the dynamics of shock discontinuities according to the Rankine-Hugoniot condition.

3.2 Numerical Methods

We now consider the numerical simulation of discontinuous solutions of scalar conservation laws via shock capturing numerical methods.

3.2.1 Finite Volume Methods

We define the basic notions related to the numerical approximation of solutions of scalar conservation laws as they are used in this thesis.

We assume the numerical approximation of the weak solution to be defined on a uniform space and time grid such that

$$U_i^j \approx u(t_j, x_i)$$

with $x_{i+1} - x_i = \Delta x$ for all $i \in \{1, \dots, n-1\}$ and $t_{j+1} - t_j = \Delta t = C_t \Delta x$ for some $C_t > 0$ for all $j \in \{1, \dots, m-1\}$. The space and time grid cover the relevant domain, i.e. $t_1 = 0$ and $t_m = T$ and in our examples if u has compact support on $[L, R] \subset \Omega$ for all $t \in [0, T]$ then we use $x_1 < L, x_n > R$. Sometimes the piecewise constant or piecewise linear interpolation of the numerical approximation is also denoted by U. Which method of interpolation is used will be explicitly stated in the context. To differentiate between the numerical approximation and the weak solution we say that u is the analytic solution and U is the numerical solution.

The numerical methods we consider are finite volume methods (see e.g. LeVeque [LeV02, Chapter 12]). For finite volume methods we refer to a cell when talking about a part of the domain that corresponds to a unit of discretization where for each $i \in \{1, \dots, n-1\}$ the corresponding cell has the domain $[x_i - \Delta x/2, x_i + \Delta x/2]$. For finite volume methods the approximated U_i represents the average of the solution u within the corresponding cell i at a specific point in time, i.e.

$$U_i^j \approx \frac{1}{\Delta x} \int_{x_i - \Delta x/2}^{x_i + \Delta x/2} u(t_j, x) \mathrm{d}x \; .$$

The idea then is to discretize the integral form of the conservation law.

We know from the introduction that for the solution of the conservation law we have the in and out flow of the conserved quantity according to the flux at the boundaries of a domain

$$\partial_t \int_{x_i - \Delta x/2}^{x_i + \Delta x/2} u(t_j, x) \mathrm{d}x = -\big(f(u(t_j, x_i + \Delta x/2)) - f(u(t_j, x_i - \Delta x/2))\big) \; .$$

A finite volume method is given by the time discretization of the dynamics of the cell average

$$\partial_t U_i = -\frac{1}{\Delta x}\Big(F_{i+1/2} - F_{i-1/2}\Big)$$

where $F_{i-1/2}, F_{i+1/2}$ are the numerical fluxes at the cell boundaries. For example the time discretization can be by the explicit Euler method

$$U_i^{j+1} := U_i^j - \frac{\Delta t}{\Delta x}\Big(F_{i+1/2}^j - F_{i-1/2}^j\Big) \; .$$

We now go on to further discuss the properties of finite volume methods.

3.2.2 Discrete Conservation

The finite volume interpretation is not crucial to the remainder of the thesis but it motivates the notion of discrete conservation (see e.g. LeVeque [LeV92, Chapter 12.3]). We want to give conditions that guarantee convergence of the numerical approximation to the analytical solution as $\Delta x \to 0$. The following notion of discrete conservation is one of the necessary conditions used to show convergence.

Definition 29. A numerical method is *conservative* if it can be written in the conservation form

$$U_i^{j+1} = U_i^j - \frac{\Delta t}{\Delta x}\Big(F(U_{i-p}^j, \dots, U_{i+q}^j) - F(U_{i-p-1}^j, \dots, U_{i+q-1}^j)\Big)$$

where F is the numerical flux.

We now describe how to derive the discrete conservation property from continuous conservation and how that relates discrete conservation to finite volume methods.

Integrating the analytic solution over $[L, R]$ in the space domain and also integrating over one discrete time step from t_j to t_{j+1} we have

$$\int_{t_j}^{t_{j+1}} \partial_t \int_L^R u(t,x)\mathrm{d}x\mathrm{d}t = -\int_{t_j}^{t_{j+1}} f(u(t,R)) - f(u(t,L))\mathrm{d}t$$

since the continuous conservation law holds. We get the discrete time stepping identity

$$\int_L^R u(t_{j+1},x)\mathrm{d}x = \int_L^R u(t_j,x)\mathrm{d}x - \Big(\int_{t_j}^{t_{j+1}} f(u(t,R)) - \int_{t_j}^{t_{j+1}} f(u(t,L))\Big)\ . \tag{3.2.1}$$

When we sum over the cells of a conservative method by definition we get

$$\Delta x \sum_{i=i_L}^{i_R} U_i^{j+1} = \Delta x \sum_{i=i_L}^{i_R} U_i^j - \Delta t \sum_{i=i_L}^{i_R} \Big(F(U_{i-p}^j, \ldots, U_{i+q}^j) - F(U_{i-p-1}^j, \ldots, U_{i+q-1}^j)\Big)\ .$$

The second sum on the right-hand side is a telescoping sum and we get

$$\Delta x \sum_{i=i_L}^{i_R} U_i^{j+1} = \Delta x \sum_{i=i_L}^{i_R} U_i^j - \Delta t\Big(F(U_{i_R-p}^j, \ldots, U_{i_R+q}^j) - F(U_{i_L-p-1}^j, \ldots, U_{i_L+q-1}^j)\Big)$$

that is a corresponding time stepping identity to that in Equation (3.2.1). If we consider the fact that U_i^j approximates a cell average we can sum over the cells and have

$$\int_L^R u(t_j,x)\mathrm{d}x = \Delta x \sum_{i=1}^n \frac{1}{\Delta x} \int_{x_i-h/2}^{x_i+h/2} u(t_j,x)\mathrm{d}x \approx \Delta x \sum_{i=1}^n U_i^j$$

and analogously for $t = t_{j+1}$. So we see that using the finite volume interpretation of the discretization in the context of a conservative numerical method allows the conserved quantitiy u to change only by the numerical flux at the boundaries and the numerical fluxes correspond to time averaged continuous fluxes.

3.2.3 Consistency

The second necessary condition for convergence that we consider is consistency (see e.g. LeVeque [LeV92, Chapter 10.3 and Chapter 12.2]). A numerical discretization is first-order consistent for a smooth PDE solution if its local truncation error is $O(\Delta x^2)$ as $\Delta x \to 0$. The local truncation error is obtained as the residual of substituting the correct solution in the finite difference scheme [DR06, Chapter 12.3].

For a conservative numerical method to be consistent the numerical flux needs to be equal to the continuous flux f for a constant solution u, i.e. we need

$$F(u, \ldots, u) = f(u)\ .$$

Furthermore, the numerical flux needs to be Lipschitz continuous in its parameters and on the relevant domain of values that occur in the solution, i.e. we need

$$|F(u_1, \ldots, u_n) - F(v_1, \ldots, v_n)| \leq C \max\{|u_1 - v_1|, \ldots, |u_n - v_n|\}$$

for all values of $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ that occur in the solution.

3.2.4 Stability

The third necessary condition for convergence we consider is stability (see e.g. LeVeque [LeV92, Chapter 10.4] on linear stability and LeVeque [LeV92, Chapter 15.7] on nonlinear stability of monotone methods). The von Neumann stability analysis of a numerical method is a local analysis of the eigenvalues of the discrete time stepping operator [Pre+88, Chapter 19.1].

In the context of methods for conservation laws stability can be connected to the speed at which information spreads through the solution. The fastest that information can travel through the solution of a scalar conservation law at any given time t is linked to the absolute speed of the fastest moving characteristic at time t that is $\max_{x\in\Omega} |f'(u(t,x))|$. The Courant-Friedrichs-Lewy (CFL) condition for the time step is

$$\Delta t_j := \frac{C}{\max_i |f'(U_i^j)|}\Delta x$$

with the CFL number $C < 1$ [LeV92, Chapter 10.6]. If the CFL condition is satisfied by the numerical method it will be stable in the sense of von Neumann stability analysis [Pre+88, Chapter 19.1].

We make the assumption that our constant space / time grid ratio $C_t = \Delta x/\Delta t$ is chosen small enough such that the CFL condition is satisfied for every time step. In many of our numerical experiments we also use adaptive time steps depending inversely on the maximum absolute characteristic speed.

3.2.5 Monotonicity and Convergence

In addition to discrete conservation, consistency and stability we consider a fourth condition for convergence and that is the monotonicity property of the numerical method (see e.g. LeVeque [LeV92, Chapter 15.7]).

Definition 30. A numerical method in conservation form

$$U_i^{j+1} = U_i^j - \frac{\Delta t}{\Delta x}\Big(F(U_{i-p}^j,\dots,U_{i+q}^j) - F(U_{i-p-1}^j,\dots,U_{i+q-1}^j)\Big) \equiv G(U_{i-p-1}^j,\dots,U_{i+q}^j)$$

is *monotone* if G is a nondecreasing function of each of its arguments.

Based on monotonicity we have the following convergence result for discrete approximations of solutions of scalar conservation laws.

Theorem 9 (Convergence of Monotone Difference Approximation)**.** Let a numerical method be conservative, consistent, monotone and satisfy the CFL condition. Then as $\Delta t, \Delta x \to 0$ with $\Delta t/\Delta x$ constant the approximation U_i^j based on the initial data $U_i^0 = u_0(x_i)$ converges boundedly almost everywhere to $u(t_j, x_i)$ that is an entropy satisfying weak solution of the conservation law.

Proof. See Crandall and Majda [CM80]. □

We now list some practical examples of numerical methods and discuss how the conditions for convergence apply.

3.2.6 Upwind scheme

We first give an example of a nonconservative discretization to emphasize how critical the conservation property can be especially in the context of discontinuous solutions. We consider Burgers equation rewritten in pseudo-linear form

$$\begin{aligned}\partial_t u(t,x) + \partial_x f(u(t,x)) &= \partial_t u(t,x) + f'(u(t,x))\partial_x u(t,x)\\ &= \partial_t u(t,x) + u(t,x)\partial_x u(t,x) = 0 \ .\end{aligned}$$

Using a foward in time (explicit Euler) and backward in space finite difference discretization for the above equation yields the following integration scheme

$$U_i^{j+1} := U_i^j + \frac{\Delta t}{\Delta x} U_i^j (U_i^j - U_{i-1}^j) \ .$$

This discretization is first-order consistent in the sense that for a smooth solution u we have by Taylor expansion

$$\begin{aligned}u(t_{j+1}, x_i) &= u(t_j, x_i) + \partial_t u(t_j, x_i)\Delta t + O(\Delta x^2)\\ u(t_j, x_{i-1}) &= u(t_j, x_i) - \partial_x u(t_j, x_i)\Delta x + O(\Delta x^2)\end{aligned}$$

since $\Delta t = C_t \Delta x$ as $\Delta x \to 0$. So the local truncation error of the discretization is

$$\begin{aligned}&u(t_j, x_i) + \frac{C_t \Delta x}{\Delta x} u(t_j, x_i)(u(t_j, x_i) - u(t_j, x_{i-1})) - u(t_{j+1}, x_i)\\ &= C_t u(t_j, x_i)(\partial_x u(t_j, x_i)\Delta x + O(\Delta x^2)) + \partial_t u(t_j, x_i) C_t \Delta x + O(\Delta x^2)\\ &= \Big(u(t_j, x_i)\partial_x u(t_j, x_i) + \partial_t u(t_j, x_i)\Big) C_t \Delta x + O(\Delta x^2) = O(\Delta x^2)\end{aligned}$$

as $\Delta x \to 0$.

However, the nonconservative Upwind scheme does not converge to a weak solution u for $\Delta x \to 0$ if the solution is discontinuous. The reason for the nonconvergence can be understood via a Riemann initial condition with $u_0(x) = H(-x)$ and a discretization with $x_{i-1} = -\epsilon$ and $x_i = \epsilon$. We have $U_{i-1}^0 = 1$ and $U_i^0 = 0$ and, hence, the nonconservative Upwind step

$$U_i^1 := U_i^0 + \frac{\Delta t}{\Delta x} U_i^0 (U_i^0 - U_{i-1}^0) = 0 + \frac{\Delta t}{\Delta x} 0(0-1) = 0 \ .$$

The numerical solution to the right of the shock in the initial data stays unchanged. With $U_{i-2}^0 = 1$ the numerical solution to the left of the shock, similarly, stays unchanged. The discrete approximation remains equal to the initial data which is not a convergent approximation of the correct weak solution of the Riemann problem.

If instead of the pseudo-linear form we discretize the standard form of the conservation law by using a forward in time and backward in space finite difference discretization we get

$$U_i^{j+1} := U_i^j + \frac{\Delta t}{\Delta x}\big(f(U_i^j) - f(U_{i-1}^j)\big)$$

that is the conservative Upwind scheme. The conservative Upwind scheme is directly in conservation form with $p = q = 0$ and the numerical flux $F(U) = f(U)$.

In our numerical experiments we use an Upwind scheme where the direction of space discretization depends on the direction of information travel in the solution.

Definition 31. The Upwind scheme is defined by

$$U_i^{j+1} := \begin{cases} U_i^j + \frac{\Delta t}{\Delta x}\big(f(U_i^j) - f(U_{i-1}^j)\big) & \text{if } f'(U_i^j) \geq 0 \\ U_i^j + \frac{\Delta t}{\Delta x}\big(f(U_{i+1}^j) - f(U_i^j)\big) & \text{otherwise} \end{cases} .$$

3.2.7 Lax-Friedrichs Scheme

The other conservative numerical method we use in the computational experiments throughout the thesis is the Lax-Friedrichs scheme.

Definition 32. The Lax-Friedrichs scheme is defined by

$$U_i^{j+1} := \frac{1}{2}(U_{i-1}^j + U_{i+1}^j) - \frac{\Delta t}{2\Delta x}\big(f(U_{i+1}^j) - f(U_{i-1}^j)\big) .$$

The Lax-Friedrichs scheme can be written in conservation form with $p = 0, q = 1$ as

$$U_i^{j+1} := U_i^j - \frac{\Delta t}{\Delta x}\big(F(U_{i+1}^j, U_i^j) - F(U_i^j, U_{i-1}^j)\big)$$

where the numerical flux is

$$F(U_{i+1}, U_i) \equiv \frac{\Delta x}{2\Delta t}(U_i - U_{i+1}) + \frac{1}{2}\big(f(U_i) - f(U_{i+1})\big) .$$

We note that the Lax-Friedrichs scheme is consistent in the sense of conservative numerical methods if f is Lipschitz continuous on any bounded domain since

$$F(u, u) = \frac{\Delta x}{2\Delta t}(u - u) + \frac{1}{2}\big(f(u) - f(u)\big) = 0$$

and F is Lipschitz continuous on the domain taken by values of the solution if f is Lipschitz continuous on any bounded domain, which is for example the case for the Burgers flux $f(u) = u^2/2$. The Lax-Friedrichs scheme is also first-order consistent in the sense that for a smooth solution of the PDE the local truncation error is second-order. By Taylor expansion we have the local truncation error as

$$\begin{aligned} &\frac{1}{2}(u(t_j, x_{i-1}) + u(t_j, x_{i+1})) - \frac{C_t \Delta x}{2\Delta x}\big(f(u(t_j, x_{i+1})) - f(u(t_j, x_{i-1}))\big) - u(t_{j+1}, x_i) \\ &= \frac{1}{2}(u(t_j, x_i) - \partial_x u(t_j, x_i)\Delta x + u(t_j, x_i) + \partial_x u(t_j, x_i)\Delta x + O(\Delta x^2)) \\ &\quad - \frac{C_t}{2}\big(f(u(t_j, x_{i+1})) - f(u(t_j, x_{i-1}))\big) - u(t_j, x_i) - \partial_t u(t_j, x_i) C_t \Delta x + O(\Delta x^2) \\ &= -\frac{C_t}{2}\big(f(u(t_j, x_i) + \partial_x u(t_j, x_i)\Delta x + O(\Delta x^2)) - f(u(t_j, x_i) - \partial_x u(t_j, x_i)\Delta x + O(\Delta x^2))\big) \\ &\quad - \partial_t u(t_j, x_i) C_t \Delta x + O(\Delta x^2) . \end{aligned}$$

Assuming f is continuously differentiable we have

$$\begin{aligned} f(u(t_j, x_i) + \partial_x u(t_j, x_i)\Delta x + O(\Delta x^2)) &= f(u(t_j, x_i)) + f'(u(t_j, x_i))\partial_x u(t_j, x_i)\Delta x + O(\Delta x^2) \\ f(u(t_j, x_i) - \partial_x u(t_j, x_i)\Delta x + O(\Delta x^2)) &= f(u(t_j, x_i)) - f'(u(t_j, x_i))\partial_x u(t_j, x_i)\Delta x + O(\Delta x^2) . \end{aligned}$$

The local truncation error evaluates to

$$
\begin{aligned}
&- C_t\big(f'(u(t_j,x_i))\partial_x u(t_j,x_i)\Delta x\big) - \partial_t u(t_j,x_i)C_t\Delta x + O(\Delta x^2)\\
&= -C_t\Delta x\big(\partial_t u(t_j,x_i) + \partial_x f'(u(t_j,x_i))\big) + O(\Delta x^2) = O(\Delta x^2)
\end{aligned}
$$

as $\Delta x \to 0$.

The Lax-Friedrichs scheme can also be derived based on the discretization of the viscous equation $\partial_t u + \partial_x f(u) = \epsilon \partial^2_{xx} u$. The vanishing viscosity solution of that equation for $\epsilon \to 0$ gives a physically relevant (entropy condition satisfying) weak solution.

3.2.8 Numerical Viscosity

We now show how to obtain the Lax-Friedrichs scheme from a discretization of the viscous equation with numerical viscosity.

We consider the viscous equation

$$
\partial_t u(t,x) + \partial_x f(u(t,x)) = \epsilon \partial^2_{xx} u(t,x)\ .
$$

The right-hand side is like that of the heat equation

$$
\partial_t u(t,x) = \epsilon \partial^2_{xx} u(t,x)
$$

that describes the diffusion of heat according to the heat coefficient ϵ. Due to the diffusion the viscosity solution is smooth and does not contain any discontinuities if $\epsilon > 0$ except perhaps in the initial data. The diffusion also causes any nonphysical discontinuities to smear out into rarefaction waves. As $\epsilon \to 0$ the smoothing of shock discontinuities becomes sharper.

By Taylor expansion and since $\Delta t = O(\Delta x)$ as $\Delta x \to 0$ we can approximate the terms of the viscous equation by the following finite differences

$$
\begin{aligned}
\partial_t u(t,x) &= \frac{U(t+\Delta t,x) - U(t,x)}{\Delta t} + O(\Delta x)\ ,\\
\partial_x f(u(t,x)) &= \frac{f(U(t,x+\Delta x)) - f(U(t,x-\Delta x))}{2\Delta x} + O(\Delta x^2)\ ,\\
\partial^2_{xx} u(t,x) &= \frac{U(t,x+\Delta x) - 2U(t,x) + U(t,x-\Delta x)}{\Delta x^2} + O(\Delta x^2)\ .
\end{aligned}
$$

Based on those finite differences we get the following discretization of the viscous equation

$$
\frac{U_i^{j+1} - U_i^j}{\Delta t} + \frac{f(U_{i+1}^j) - f(U_{i-1}^j)}{2\Delta x} = \epsilon \frac{U_{i+1}^j - 2U_i^j + U_{i-1}^j}{\Delta x^2}
$$

that yields the discrete time step

$$
U_i^{j+1} := U_i^j + \Delta t \left(\epsilon \frac{U_{i+1}^j - 2U_i^j + U_{i-1}^j}{\Delta x^2} - \frac{f(U_{i+1}^j) - f(U_{i-1}^j)}{2\Delta x} \right)\ .
$$

For the viscosity coefficient

$$
\epsilon = \frac{\Delta x^2}{2\Delta t}
$$

we get the Lax-Friedrichs scheme.

The Lax-Friedrichs scheme approximates a solution to the viscous equation with $\epsilon = O(\Delta x)$ as $\Delta x \to 0$. In the next section we consider the analytic solution of the viscous equation for the specific example of a traveling shock discontinuity. The traveling shock wave is helpful for understanding the behavior of the numerical viscosity of shock capturing methods such as the Lax-Friedrichs scheme as the discretization grid is refined.

3.3 Analytic Viscosity Solution

3.3.1 Lax-Friedrichs Convergence Rates

LeVeque [LeV92, Chapter 11.2] argues that the approximation error achieved by the Lax-Friedrichs method can at best be $O(\sqrt{\Delta x})$ as $\Delta x \to 0$ based on the analytic solution of the viscous equation corresponding to the Lax-Friedrichs scheme, i.e. with the corresponding numerical viscosity coefficient. The analytic solution of the viscous equation with linear flux for discontinuous initial data transports the discontinuity like in the equation without viscosity but also applies diffusion to the discontinuity and, thus, smoothes out the discontinuity. The scale of the smoothing is $O(\sqrt{\Delta x})$ as $\Delta x \to 0$ and thus the L_1 error of the numerical approximation converges at best at a square root convergence rate.

However, in Teng and Zhang [TZ97] the L_1 convergence of the numerical approximation by monotone numerical methods such as the Lax-Friedrichs method is analyzed for piecewise constant solutions with shock discontinuities for nonlinear flux functions which gives a better convergence rate. The paper highlights the parallel between the discrete traveling wave in a monotone numerical method [Jen74] and the analytic traveling wave of the corresponding viscous equation. Due to Tang and Teng [TT97] for nonlinear flux functions the L_1 error between the viscous solution and the analytic weak solution with shock discontinuities converges at a linear rate $O(\Delta x)$ as $\Delta x \to 0$. Based on the parallel between analytic viscosity and discrete solution in Teng and Zhang [TZ97] and also from computational experiments we know that the result holds for e.g. the Lax-Friedrichs discrete solution of a scalar conservation law with nonlinear flux.

3.3.2 Analytic Traveling Wave

Inspired by Teng and Zhang [TZ97] we consider the analytic traveling wave solution of a viscous equation. Specifically, we focus on the example of the Burgers flux $f(u) = u^2/2$.

We consider the analogue of the Riemann solution, a traveling wave defined by satisfying

$$u(t, x) = u_0(x - st)$$

where s is the speed at which the wave travels and u_0 is the initial data. In the case of the standard Riemann problem with $u_0(x) = H(-x)$ we have $s = 1/2$ and

$$u(t, x) = H(t/2 - x) = u_0(x - t/2)\ .$$

In order to make the example applicable in the context of sensitivity analysis we consider the analogy to the parametric Riemann example from Chapter 1.3.1.

Proposition 10. For the viscous Burgers equation with $\epsilon > 0$

$$u^\epsilon(t,x) = \frac{1+p}{2} + \frac{1+p}{2}\tanh\left(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - x\Big)\right)$$

is a traveling wave solution with speed $s = (1+p)/2$.

Proof. An alternative form of tanh in terms of exponential functions is

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(-x) + \exp(x)} \ .$$

The derivative of tanh in terms of exponential functions is

$$\tanh'(x) = \frac{4}{\exp(-x) + \exp(x)} \ .$$

We define the short hand notation for the argument of tanh

$$X \equiv \frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - x\Big) \ .$$

By the chain rule the derivative of the solution with respect to time is

$$\begin{aligned}\partial_t u^\epsilon(t,x) &= \frac{1+p}{2}\frac{4}{(\exp(-X)+\exp(X))^2}\frac{(1+p)^2}{8\epsilon}\\ &= \frac{(1+p)^3}{4\epsilon}\frac{1}{(\exp(-X)+\exp(X))^2} \ .\end{aligned}$$

The derivative with respect to space is

$$\begin{aligned}\partial_x u^\epsilon(t,x) &= -\frac{1+p}{2}\frac{4}{(\exp(-X)+\exp(X))^2}\frac{1+p}{4\epsilon}\\ &= -\frac{(1+p)^2}{2\epsilon}\frac{1}{(\exp(-X)+\exp(X))^2} \ .\end{aligned}$$

The second derivative with respect to space is

$$\begin{aligned}\epsilon\partial^2_{xx} u^\epsilon(t,x) &= 2\frac{(1+p)^2}{2}\frac{1}{(\exp(-X)+\exp(X))^3}\Big(\exp(-X)\frac{1+p}{4\epsilon} - \exp(X)\frac{1+p}{4\epsilon}\Big)\\ &= \frac{(1+p)^3}{4\epsilon}\frac{\exp(-X)-\exp(X)}{(\exp(-X)+\exp(X))^3} \ .\end{aligned}$$

For the Burgers flux f we have

$$\partial_x f(u^\epsilon(t,x)) = u^\epsilon(t,x)\partial_x u^\epsilon(t,x) \ .$$

The solution in terms of expontential functions is

$$u^\epsilon(t,x) = \frac{1+p}{2} + \frac{1+p}{2}\frac{\exp(X)-\exp(-X)}{\exp(-X)+\exp(X)} \ .$$

We multiply with the derivative with respect to space to get

$$\begin{aligned} u^\epsilon(t,x)\partial_x u^\epsilon(t,x) &= -\frac{1+p}{2}\frac{(1+p)^2}{2\epsilon}\frac{1}{(\exp(-X)+\exp(X))^2} \\ &\quad -\frac{1+p}{2}\frac{\exp(X)-\exp(-X)}{\exp(X)+\exp(-X)}\frac{(1+p)^2}{2\epsilon}\frac{1}{(\exp(-X)+\exp(X))^2} \\ &= -\frac{(1+p)^3}{4\epsilon}\frac{1}{(\exp(-X)+\exp(X))^2} - \frac{(1+p)^3}{4\epsilon}\frac{\exp(X)-\exp(-X)}{(\exp(-X)+\exp(X))^3}\,. \end{aligned}$$

The first term cancels with $\partial_t u^\epsilon$ and the second term cancels with $\partial^2_{xx} u^\epsilon$.

Substituting all of the above into the viscous Burgers equation we get

$$\partial_t u^\epsilon(t,x) + \partial_x f(u^\epsilon(t,x)) - \epsilon\partial^2_{xx} u^\epsilon(t,x) = 0$$

for all $t, x \in \mathbb{R}$. This concludes the proof □

In Chapter 5 we employ the analytic traveling wave solution to practically motivate our assumptions about the discrete solution and the discrete tangent sensitivities of shock capturing schemes in the vicinity of shocks.

We considered scalar conservation laws with discontinuous solutions and their numerical approximations by shock capturing schemes. In the next chapter we consider the tangent and adjoint sensitivity analysis of PDE solutions and the discrete sensitivities of their numerical approximations via AD methods.

Chapter 4

Tangent and Adjoint Sensitivity Analysis

In this chapter we introduce the tangent and adjoint sensitivity analysis of PDE solutions with respect to a parameter in the initial data in the setting where the solution is smooth and continuously differentiable with respect to the parameter. We also introduce the discrete tangent and adjoint sensitivity analysis of numerical methods and introduce how the discrete sensitivity analysis can be performed by AD methods.

Equations (1.1.1)-(1.1.2) define a scalar conservation law initial value problem where the initial data $u_0(\cdot, p)$ depends on the parameter $p \in \mathbb{R}^d$. Since the solution at times $t > 0$ depends on the initial data the solution is implicitly a function of the parameter p. The sensitivity analysis of the solution u with respect to the parameter p can be accomplished by differentiation.

We give a derivation for the analytic tangent sensitivity

$$\dot{u}(t, x, p) \equiv \mathrm{d}_p u(t, x, p)\dot{p}$$

that is based on the directional differentiation of the solution with the tangent direction $\dot{p} \in \mathbb{R}^d$. We also give a derivation for the analytic adjoint sensitivity

$$\bar{u}(t, \cdot, p) \equiv \mathrm{d}_p \Phi(u(t, \cdot, p))^T \bar{\Phi}$$

that is based on the adjoint direction $\bar{\Phi} \in \mathbb{R}$ for a scalar objective functional $\Phi \colon C^\infty(\Omega) \to \mathbb{R}$ of the solution at time T. We discuss the individual steps involved in these sensitivity analyses according to the chain rule.

We then discuss how to compute the corresponding tangent and adjoint sensitivities for the discrete setting, i.e. when the solution is computed by a numerical method. We also introduce the basics of AD methods used to compute the discrete tangent and adjoint sensitivities algorithmically.

4.1 Analytic Sensitivity Analysis

4.1.1 Asymptotic Analysis

Regarding the parameterization of the solution we focus on an asymptotic analysis in the following sense. We consider the sensitivities of the weak solution with respect to p at $p = p_0$ in a neighborhood

$$D \equiv \{ \ p : \|p - p_0\|_2 < \delta \ \}$$

for some $\delta > 0$. In that setting the weak solution we consider is a function

$$u \colon [0, T] \times \Omega \times D \to \mathbb{R} \ .$$

When performing a sensitivity analysis we are interested in the question: How does u change as p changes? If u is smooth and continuously differentiable with respect to p on D then by first-order Taylor expansion

$$u(t, x, p_0 + \dot{p}) - u(t, x, p_0) - \mathrm{d}_p u(t, x, p_0)\dot{p} = O(\|\dot{p}\|^2)$$

for all $t \in [0, T], x \in \Omega$ as $\dot{p} \to 0$. The derivative of u with respect to p asymptotically describes how u changes as p varies up to first-order.

The sensitivity analysis we present is formally asymptotic in nature in the sense that it only holds in the limit of $\dot{p} \to 0$. However, by some technical effort the sensitivity analysis can be extended to larger sets D. In practice the approximation error also holds for larger sets of perturbations as we see in the numerical experiments in later chapters.

4.1.2 Tangent Sensitivity Analysis

We now give the derivation of the tangent sensitivity as the solution of the tangent PDE obtained from implicit differentiation. Regarding the tangent sensitivity analysis we here consider only the directional differentiation of the solution u with respect to the parameter p as that is a perspective that arises naturally in our problem setting.

In the next sections we see that the tangent sensitivity can be split according to the chain rule into the directional derivative of the initial data with respect to the parameter p and the solution with respect to the initial data. In fact in the section on tangent and adjoint derivative operators we see that the tangent sensitivity analysis of a PDE can be split even further into separate time steps.

Proposition 11. Let $u(\cdot, \cdot, p)$ be a smooth solution of Equations (1.1.1)-(1.1.2) for all $p \in D$ and continuously differentiable with respect to p on D and f twice continuously differentiable. Then the tangent sensitivity

$$\dot{u}(\cdot, \cdot, p_0) \equiv \mathrm{d}_p u(\cdot, \cdot, p_0)\dot{p}$$

satisfies the tangent PDE initial value problem

$$\begin{aligned} \dot{u}(0, x, p_0) &= \mathrm{d}_p u_0(x, p_0)\dot{p} \\ 0 &= \partial_t \dot{u}(t, x, p_0) + f''(u(t, x, p_0))\partial_x u(t, x, p_0)\dot{u}(t, x, p_0) + f'(u(t, x, p_0))\partial_x \dot{u}(t, x, p_0) \end{aligned}$$

for all $t \in [0, T]$, $x \in \Omega$.

Proof. Assuming that $u(\cdot,\cdot,p)$ is a smooth solution for all $p \in D$ then Equations (1.1.1)-(1.1.2) hold for all $t \in [0,T], x \in \Omega$ and not just in a weak sense.

We consider the corresponding operator equation (c.f. the implicit function theorem in Section 2.2.2)

$$F(p,u) \equiv \begin{pmatrix} u(0,\cdot,p) - u_0(\cdot,p) \\ \partial_t u(\cdot,\cdot,p) + \partial_x f(u(\cdot,\cdot,p)) \end{pmatrix} = 0 \; .$$

By chain rule and the linearity of differentiation we have

$$\begin{aligned} \mathrm{d}_p F(p,u) &= \begin{pmatrix} \mathrm{d}_p u(0,\cdot,p) - \mathrm{d}_p u_0(\cdot,p) \\ \mathrm{d}_p \partial_t u(\cdot,\cdot,p) + \mathrm{d}_p \partial_x f(u(\cdot,\cdot,p)) \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{d}_p u(0,\cdot,p) - \mathrm{d}_p u_0(\cdot,p) \\ \partial_t \mathrm{d}_p u(\cdot,\cdot,p) + \mathrm{d}_p \big(f'(u(\cdot,\cdot,p))\partial_x u(\cdot,\cdot,p)\big) \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{d}_p u(0,\cdot,p) - \mathrm{d}_p u_0(\cdot,p) \\ \partial_t \mathrm{d}_p u(\cdot,\cdot,p) + f''(u(\cdot,\cdot,p))\mathrm{d}_p u(\cdot,\cdot,p)\partial_x u(\cdot,\cdot,p) + f'(u(\cdot,\cdot,p))\partial_x \mathrm{d}_p u(\cdot,\cdot,p) \end{pmatrix} \; . \end{aligned}$$

Since $u(\cdot,\cdot,p)$ is a solution for all $p \in D$ we have $0 = \mathrm{d}_p F(p_0,u)$.

By right-multiplying a direction $\dot{p}$ we have

$$\begin{aligned} 0 &= \mathrm{d}_p u(0,x,p_0)\dot{p} - \mathrm{d}_p u_0(x,p_0)\dot{p} \\ 0 &= \partial_t \mathrm{d}_p u(t,x,p_0)\dot{p} + f''(u(t,x,p_0))\mathrm{d}_p u(t,x,p_0)\dot{p}\partial_x u(t,x,p_0) + f'(u(t,x,p_0))\partial_x \mathrm{d}_p u(t,x,p_0)\dot{p} \end{aligned}$$

for all $t \in [0,T], x \in \Omega$.

If we substitute $\mathrm{d}_p u(t,x,p_0)\dot{p}$ by the tangent variable $\dot{u}(t,x,p_0)$ we obtain the tangent PDE

$$\begin{aligned} \dot{u}(0,x,p_0) &= \mathrm{d}_p u_0(x,p_0)\dot{p} \\ 0 &= \partial_t \dot{u}(t,x,p_0) + f''(u(t,x,p_0))\dot{u}(t,x,p_0)\partial_x u(t,x,p_0) + f'(u(t,x,p_0))\partial_x \dot{u}(t,x,p_0) \end{aligned}$$

for all $t \in [0,T]$, $x \in \Omega$. This concludes the proof. □

By the implicit function theorem the uniqueness of the tangent $\dot{u}$ can be shown via the invertibility of the derivative $\partial_u F(p,u)$ of the operator F as defined in the above derivation. For a formal treatment of existence and uniquess of the sensitivity with respect to initial data see e.g. Bouchut and James [BJ99].

Based on the proposition a tangent sensitivity analysis can be performed by solving the tangent PDE for $\dot{u}$.

4.1.3 Adjoint Sensitivity Analysis

We now give the derivation of the adjoint sensitivity as the solution of the adjoint PDE in the context of a PDE constrained optimization problem. Regarding the adjoint sensitivity analysis we here consider the gradient of a scalar objective functional Φ with respect to the parameter p. We also explain how the gradient splits into the sensitivity of the objective function with respect to the solution, the solution with respect to the initial data and the initial data with respect to the parameter.

We consider the PDE constrained optimization problem $\min_{u,p} \Phi(u(T,\cdot,p))$ such that Equations (1.1.1)-(1.1.2) are satisfied.

Proposition 12. Let $u(\cdot,\cdot,p)$ be a smooth solution of Equations (1.1.1)-(1.1.2) for all $p \in D$ and continuously differentiable with respect to p on D. Let $\Phi(\cdot)$ be continuously differentiable with respect to u for all $p \in D$. Then the total derivative of Φ with respect to the parameter p at $p = p_0$ is

$$\mathrm{d}_p\Phi(u(T,\cdot,p_0)) = \mathrm{d}_{u(T,\cdot,p_0)}\Phi(u) \cdot \mathrm{d}_{u(0,\cdot,p_0)}u(T,\cdot,p_0) \cdot \mathrm{d}_p u_0(\cdot,p_0) \ .$$

Proof. Follows directly from the chain rule. □

The derivative of the initial data with respect to the parameter is a linear operator that maps a tangent $\dot{p} \in \mathbb{R}^d$ in the space of the parameter to a continuous tangent of the initial data $\dot{u}(0,\cdot,p_0) \in C(\Omega)$

$$\mathrm{d}_p u_0(\cdot,p) \colon \mathbb{R}^d \to C(\Omega) \ .$$

The tangent of the initial data is continuous by the assumption that u is continuously differentiable in p. The derivative of the solution at time T with respect to the initial data is a linear operator that maps a continuous tangent $\dot{u}(0,\cdot,p_0) \in C(\Omega)$ of the initial data to a continuous tangent of the solution at time T that is $\dot{u}(T,\cdot,p_0) \in C(\Omega)$

$$\mathrm{d}_{u_0(\cdot,p)}u(T,\cdot,p) \colon C(\Omega) \to C(\Omega) \ .$$

The tangent of the solution is continuous again by the assumption that u is continuously differentiable in p. The derivative of the objective functional with respect to the solution at time T is a linear operator that maps the continuous tangent of the solution $\dot{u}(T,\cdot,p) \in C(\Omega)$ to the tangent of the functional $\dot{\Phi}(u) \in \mathbb{R}$

$$\mathrm{d}_{u(T,\cdot,p)}\Phi(u) \colon C(\Omega) \to \mathbb{R} \ .$$

Each of the derivatives listed above is a linear operator with a corresponding adjoint operator. The adjoint of the derivative of the objective functional with respect to the solution at the final time $\mathrm{d}_{u(T,\cdot,p)}\Phi(u)^*$ is often available as a closed form operator directly from the definition of Φ.

Example 13 (Integral objective)**.** For example for an integral objective

$$\Phi(u) \equiv \int_\Omega \varphi(u(T,x,p_0))\mathrm{d}x$$

with continuously differentiable φ we have the tangent

$$\dot{\Phi}(u) = \mathrm{d}_{u(T,\cdot,p)}\Phi(u)\dot{u}(T,\cdot,p_0) = \int_\Omega \varphi'(u(T,x,p_0))\dot{u}(T,x,p_0)\mathrm{d}x \ .$$

From the definition of the adjoint operator we have the inner product identity

$$\begin{aligned}
\langle \dot{\Phi}, \bar{\Phi} \rangle_{\mathbb{R}} &= \langle \mathrm{d}_{u(T,\cdot,p_0)}\Phi(u)\dot{u}(T,\cdot,p_0), \bar{\Phi} \rangle_{\mathbb{R}} && \text{(definition of tangent variable)} \\
&= \left\langle \int_\Omega \dot{u}(T,x,p_0)\varphi'(u(T,x,p_0))\mathrm{d}x, \bar{\Phi} \right\rangle_{\mathbb{R}} && \text{(definition of derivative operator)} \\
&= \int_\Omega \dot{u}(T,x,p_0)\varphi'(u(T,x,p_0))\bar{\Phi}\mathrm{d}x && \text{(inner product in } \mathbb{R})
\end{aligned}$$

$$
\begin{aligned}
&= \langle \dot{u}(T,\cdot,p_0), \varphi'(u(T,\cdot,p_0))\bar{\Phi}\rangle_{C(\Omega)} && \text{(inner product in } C(\Omega))\\
&= \langle \dot{u}(T,\cdot,p_0), \mathrm{d}_{u(T,\cdot,p)}\Phi(u)^*\bar{\Phi}\rangle_{C(\Omega)} && \text{(definition of adjoint operator)}\\
&= \langle \dot{u}(T,\cdot,p_0), \bar{u}(T,\cdot,p_0))\rangle_{C(\Omega)} && \text{(definition of adjoint variable)}
\end{aligned}
$$

We have the adjoint operator

$$\mathrm{d}_{u(T,\cdot,p)}\Phi(u)^* \colon \mathbb{R} \to C(\Omega)$$

given by the map

$$\bar{\Phi} \mapsto \varphi'(u(T,\cdot,p_0))\bar{\Phi} \; .$$

The adjoint of the derivative of the initial data with respect to the parameter also often follows directly in closed form from the definition of the initial data.

Example 14 (Affine initial data)**.** For example if we have affine initial data defined by two parameters as

$$u_0(x,p) \equiv p_1 + p_2 x$$

we have the tangent

$$\dot{u}_0(x,p) = \mathrm{d}_p u_0(x,p)\dot{p} = \dot{p}_1 + \dot{p}_2 x \; .$$

From the definition of the adjoint operator we have the inner product identity

$$
\begin{aligned}
\langle \dot{u}_0(\cdot,p), \bar{u}_0(\cdot,p)\rangle_{C(\Omega)} &= \left\langle \Big(x \mapsto \dot{p}_1 + \dot{p}_2 x\Big), \bar{u}_0(\cdot,p)\right\rangle_{C(\Omega)} && \text{(definition of tangent variable)}\\
&= \int_\Omega \dot{p}_1 \bar{u}_0(x,p) + \dot{p}_2 x \bar{u}_0(x,p)\mathrm{d}x && \text{(inner product in } C(\Omega))\\
&= \dot{p}_1 \int_\Omega \bar{u}_0(x,p)\mathrm{d}x + \dot{p}_2 \int_\Omega x\bar{u}_0(x,p)\mathrm{d}x && \text{(distributivity)}\\
&= \left\langle \dot{p}, \begin{pmatrix} \int_\Omega \bar{u}_0(x,p)\mathrm{d}x \\ \int_\Omega x\bar{u}_0(x,p)\mathrm{d}x \end{pmatrix} \right\rangle_{\mathbb{R}^2} && \text{(inner product in } \mathbb{R}^2)\\
&= \langle \dot{p}, \mathrm{d}_p u_0(\cdot,p)^* \bar{u}_0(\cdot,p)\rangle_{\mathbb{R}^2} && \text{(definition of adjoint operator)}\\
&= \langle \dot{p}, \bar{p}\rangle_{\mathbb{R}^2} && \text{(definition of adjoint variable)}
\end{aligned}
$$

We have the adjoint operator

$$\mathrm{d}_p u_0(\cdot,p)^* \colon C(\Omega) \to \mathbb{R}^2$$

given by the map

$$\bar{u}_0 \mapsto \begin{pmatrix} \int_\Omega \bar{u}_0(x,p)\mathrm{d}x \\ \int_\Omega x\bar{u}_0(x,p)\mathrm{d}x \end{pmatrix} \; .$$

What remains is to consider the adjoint $\mathrm{d}_{u(0,\cdot,p)}u(T,\cdot,p_0)^*$ of the derivative of the solution at time T with respect to the initial data. We give two possible derivations of that adjoint in the following two sections. The first approach is motivated from the first-order optimality condition of the Lagrange multiplier approach for constrained optimization. The derivation via the Lagrangian is the standard in the PDE constrained optimization literature. The second approach is a limiting argument for the continuous time stepping operator of the PDE. The derivation via the time stepping operator is analogous to how the discrete adjoint works and may be more accessible from the perspective of AD. In both approaches the adjoint PDE is derived via integration by parts.

4.1.4 Adjoint PDE from the Lagrangian

We now give the derivation of the adjoint PDE based on the Langrange multiplier approach to constrained optimization. The Lagrange multiplier approach is a standard formulation of the adjoint method in optimal design (see e.g. Jameson [Jam95]).

In the context of PDE constrained optimization we consider the problem

$$\begin{aligned}&\min_{u,p}\ \Phi(u(T,\cdot))\\ &\text{s.t. } 0 = \partial_t u(t,x) + \partial_x f(u(t,x))\\ &\qquad 0 = u(0,x) - u_0(x,p), \qquad \text{for all } t \in [0,T], x \in \Omega\ .\end{aligned}$$

Let us assume that the solution u and the parameter $p = p_0$ give the optimum of such a PDE constrained optimization problem for some continuously differentiable functional Φ.

The Langrangian of the constrained optimization problem is

$$\begin{aligned}\mathcal{L}(u,p,\bar{u},\bar{u}_0) = \Phi(u(T,\cdot)) &- \int_0^T \int_\Omega \bar{u}(t,x)\big(\partial_t u(t,x) + \partial_x f(u(t,x))\big)\mathrm{d}x\mathrm{d}t\\ &- \int_\Omega \bar{u}_0(x)\big(u(0,x) - u_0(x,p)\big)\mathrm{d}x\end{aligned}$$

where $\bar{u}$ is the Lagrange multiplier corresponding to the PDE and $\bar{u}_0$ is the Lagrange multiplier corresponding to the initial value condition. As a necessary optimality condition by the assumption that u, p_0 give a local optimum we have $0 = \mathrm{d}_p\mathcal{L}(u, p_0, \bar{u}, \mu)$ where

$$\mathrm{d}_p\mathcal{L}(u,p,\bar{u},\bar{u}_0) = \mathrm{d}_p\Phi(u(T,\cdot)) - \mathrm{d}_p \int_0^T \int_\Omega \bar{u}(t,x)\big(\partial_t u(t,x) + \partial_x f(u(t,x))\big)\mathrm{d}x\mathrm{d}t \tag{4.1.1}$$

$$- \mathrm{d}_p \int_\Omega \bar{u}_0(x)\big(u(0,x) - u_0(x,p)\big)\mathrm{d}x\ . \tag{4.1.2}$$

By the following proposition we obtain the adjoint of the derivative of the solution at time T with respect to the initial data as it arises in the context of the Lagrangian.

Proposition 13. Let $u(\cdot,\cdot,p)$ be a smooth solution of Equations (1.1.1)-(1.1.2) for all $p \in D$ and continuously differentiable with respect to p on D and with compact support in Ω. Then the Lagrange multiplier $\bar{u}$ that satisfies the backward in time adjoint PDE

$$\begin{aligned}\bar{u}(T,x) &= \mathrm{d}_{u(T,x,p)}\Phi(u(T,\cdot,p_0))\\ 0 &= \partial_t \bar{u}(t,x) + \partial_x \bar{u}(t,x) f'(u(t,x,p_0))\end{aligned}$$

for all $t \in [0,T]$, $x \in \Omega$ and the Lagrange multiplier $\bar{u}_0(x) = \bar{u}(0,x)$ for all $x \in \Omega$ satisfies the stationarity condition $0 = \mathrm{d}_p\mathcal{L}(u, p_0, \bar{u}, \mu)$ if and only if

$$0 = \int_\Omega \bar{u}(0,x)\mathrm{d}_p u_0(x,p)\mathrm{d}x \equiv \bar{p}\ .$$

Proof. The derivation of the adjoint sensitivity for differential equations is obtained via integration by parts of the integral in the second term on the right-hand side of Equation (4.1.1). The result is a differential equation that describes the dynamics of the Lagrange multiplier $\bar{u}$. From here on out we call the Lagrange multiplier $\bar{u}$ the adjoint variable.

For the second term on the right-hand side by linearity of differentiation we have

$$\mathrm{d}_p \int_0^T \int_\Omega \bar{u}(t,x)\big(\partial_t u(t,x) + \partial_x f(u(t,x))\big)\mathrm{d}x\mathrm{d}t \tag{4.1.3}$$

$$= \int_0^T \int_\Omega \bar{u}(t,x)\big(\partial_t \mathrm{d}_p u(t,x) + \partial_x \mathrm{d}_p f(u(t,x))\big)\mathrm{d}x\mathrm{d}t \tag{4.1.4}$$

$$= \int_0^T \int_\Omega \bar{u}(t,x)\partial_t \mathrm{d}_p u(t,x)\mathrm{d}x\mathrm{d}t + \int_0^T \int_\Omega \bar{u}(t,x)\partial_x \mathrm{d}_p f(u(t,x))\mathrm{d}x\mathrm{d}t\ . \tag{4.1.5}$$

For the first term in Equation (4.1.5) by switching the order of integration for a smooth function and then integration by parts we have

$$\int_0^T \int_\Omega \bar{u}(t,x)\partial_t \mathrm{d}_p u(t,x)\mathrm{d}x\mathrm{d}t \tag{4.1.6}$$

$$= \int_\Omega \int_0^T \bar{u}(t,x)\partial_t \mathrm{d}_p u(t,x)\mathrm{d}t\mathrm{d}x \tag{4.1.7}$$

$$= \int_\Omega \Big[\bar{u}(t,x)\mathrm{d}_p u(t,x)\Big]_0^T \mathrm{d}x - \int_\Omega \int_0^T \partial_t \bar{u}(t,x)\mathrm{d}_p u(t,x)\mathrm{d}t\mathrm{d}x\ . \tag{4.1.8}$$

The first term of that equation is relevant for determining the adjoint variable $\bar{u}$ at time T.

For the second term in Equation (4.1.5) by integration by parts and the chain rule we have

$$\int_0^T \int_\Omega \bar{u}(t,x)\partial_x \mathrm{d}_p f(u(t,x))\mathrm{d}x\mathrm{d}t \tag{4.1.9}$$

$$= \int_0^T \Big[\bar{u}(t,x)\mathrm{d}_p f(u(t,x))\Big]_{-\infty}^{\infty} \mathrm{d}t - \int_0^T \int_\Omega \partial_x \bar{u}(t,x)\mathrm{d}_p f(u(t,x))\mathrm{d}x\mathrm{d}t \tag{4.1.10}$$

$$= \int_0^T \Big[\bar{u}(t,x)\mathrm{d}_p f(u(t,x))\Big]_{-\infty}^{\infty} \mathrm{d}t - \int_0^T \int_\Omega \partial_x \bar{u}(t,x) f'(u(t,x))\mathrm{d}_p u(t,x)\mathrm{d}x\mathrm{d}t\ . \tag{4.1.11}$$

The first term of the right-hand side is zero by the assumption that the solution has compact support in Ω.

Adding the second terms of Equation (4.1.8) and Equation (4.1.11), respectively, we get

$$-\int_\Omega \int_0^T \partial_t \bar{u}(t,x)\mathrm{d}_p u(t,x) + \partial_x \bar{u}(t,x) f'(u(t,x))\mathrm{d}_p u(t,x)\mathrm{d}t\mathrm{d}x$$
$$= -\int_\Omega \int_0^T \Big(\partial_t \bar{u}(t,x) + \partial_x \bar{u}(t,x) f'(u(t,x))\Big)\mathrm{d}_p u(t,x)\mathrm{d}t\mathrm{d}x$$

which is zero if the adjoint variable satisfies the PDE

$$0 = \partial_t \bar{u}(t,x) + \partial_x \bar{u}(t,x) f'(u(t,x))$$

for all $t \in [0,T], x \in \Omega$.

The adjoint PDE is solved backwards in time. In order to figure out the initial data for $\bar{u}(T,x)$ we consider the term

$$\int_\Omega \Big[\bar{u}(t,x)\mathrm{d}_p u(t,x)\Big]_0^T \mathrm{d}x = \int_\Omega \bar{u}(T,x)\mathrm{d}_p u(T,x)\mathrm{d}x - \int_\Omega \bar{u}(0,x)\mathrm{d}_p u(0,x)\mathrm{d}x\ . \tag{4.1.12}$$

in Equation (4.1.8).

By chain rule the first term in Equation (4.1.1) is

$$\begin{aligned} \mathrm{d}_p\Phi(u(T,\cdot)) &= \Phi'(u(T,\cdot))\mathrm{d}_p u(T,\cdot) \\ &= \int_\Omega \Phi'(u(T,x))\mathrm{d}_p u(T,x)\mathrm{d}x \,. \end{aligned}$$

Since $\bar{u}(T,x) = \mathrm{d}_{u(T,x)}\Phi'(u(T,\cdot))$ the above cancels out the first term on the right-hand side of Equation (4.1.12).

With $\bar{u}_0(\cdot) = \bar{u}(0,\cdot)$ we have

$$\begin{aligned} \mathrm{d}_p \int_\Omega \bar{u}_0(x)\big(u(0,x) - u_0(x,p)\big)\mathrm{d}x &= \int_\Omega \bar{u}_0(x)\big(\mathrm{d}_p u(0,x) - \mathrm{d}_p u_0(x,p)\big)\mathrm{d}x \\ &= \int_\Omega \bar{u}(0,x)\mathrm{d}_p u(0,x)\mathrm{d}x + \bar{p} \,. \end{aligned}$$

The first term of the above cancels out the second term on the right-hand side of Equation (4.1.12).

Overall, given our choice of Lagrange multipliers we are left with

$$\mathrm{d}_p\mathcal{L}(u,p,\bar{u},\bar{u}_0) = \bar{p} \,.$$

This concludes the proof. □

Since we have $\mathrm{d}_p\mathcal{L} = 0$ if and only if $\mathrm{d}_p\Phi = 0$ and $\mathrm{d}_p\Phi$ factorizes according to Proposition 12 we have the following corollary.

Corollary 14. The adjoint sensitivity

$$\bar{u}(t,\cdot,p_0) \equiv (\mathrm{d}_{u(t,\cdot,p_0)} u_T(\cdot,p_0))^* \bar{u}_T(\cdot,p_0)$$

satisfies the backward in time adjoint PDE

$$\begin{aligned} \bar{u}(T,x,p_0) &= \bar{u}_T(x,p_0) \\ 0 &= \partial_t \bar{u}(t,x,p_0) + \partial_x \bar{u}(t,x,p_0) f'(u(t,x,p_0)) \end{aligned}$$

at the stationary point of some optimization problem with $\Phi'(u) = \bar{u}_T(\cdot,p_0)$.

4.1.5 Adjoint PDE from the Time Stepping Operator

We now show how to derive the adjoint PDE based on the adjoint of the tangent PDE solution operator for the limit of a time step size going to zero.

Proposition 15. Let $\mathcal{S}$ be the solution operator of the linear tangent PDE

$$\begin{aligned} \dot{u}(0,x) &= \dot{u}_0(x) \\ 0 &= \partial_t \dot{u}(t,x) + f''(u(t,x,p_0))\partial_x u(t,x,p_0)\dot{u}(t,x) + f'(u(t,x,p_0))\partial_x \dot{u}(t,x) \end{aligned}$$

for all $t \in [0,T], x \in \Omega$, i.e. $\dot{u}(T,\cdot) = \mathcal{S}(\dot{u}(0,\cdot))$ then the adjoint operator $\mathcal{S}^*$ is the solution operator of the linear adjoint PDE

$$\begin{aligned}\bar{u}(T,x) &= \bar{u}_T(0)\\ 0 &= \partial_t \bar{u}(t,x) + \partial_x \bar{u}(t,x) f'(u(t,x,p_0))\end{aligned}$$

for all $t \in [0,T], x \in \Omega$, i.e. $\bar{u}(0,\cdot) = \mathcal{S}^*(\bar{u}(T,\cdot))$ assuming that both $\dot{u}$ and $\bar{u}$ are bounded, continuously differentiable with respect to t and x for all $t \in [0,T], x \in \Omega$ and with compact support on $\hat{\Omega} \subset \Omega$.

Proof. Since the tangent PDE is a linear PDE the solution operator $\mathcal{S}$ is a linear operator. By the assumption that $\dot{u}$ is bounded $\mathcal{S}$ is continuous.

We consider the solution operator $\mathcal{S}_{t_i,t_{i+1}}$ of the tangent PDE for a smaller time step from $t = t_i$ to $t = t_{i+1}$ with $t_i < t_{i+1}$ such that

$$\dot{u}(t_{i+1},\cdot) = \mathcal{S}_{t_i,t_{i+1}}(\dot{u}(t_i,\cdot)) \ .$$

If we split the time domain into multiple intervals at $t_1 = 0, \ldots, t_n = T$ then we have

$$\dot{u}(T,\cdot) = \mathcal{S}(\dot{u}(0,\cdot)) = \mathcal{S}_{t_{n-1},t_n}(\ldots \mathcal{S}_{t_2,t_3}(\mathcal{S}_{t_1,t_2}(\dot{u}(0,\cdot)))) \ .$$

The adjoint operator $\mathcal{S}^*_{t_i,t_{i+1}}$ is defined by the inner product identity

$$\langle \mathcal{S}_{t_i,t_{i+1}}(\dot{u}(t_i,\cdot)), \bar{u}(t_{i+1},\cdot)\rangle = \langle \dot{u}(t_i,\cdot), \mathcal{S}^*_{t_i,t_{i+1}}(\bar{u}(t_{i+1},\cdot))\rangle \ .$$

By chaining the time stepping operators we get the adjoint operator $\mathcal{S}^*$ as defined by

$$\begin{aligned}\langle \mathcal{S}(\dot{u}(0,\cdot)), \bar{u}(T,\cdot)\rangle &= \langle \mathcal{S}_{t_{n-1},t_n}(\ldots \mathcal{S}_{t_2,t_3}(\mathcal{S}_{t_1,t_2}(\dot{u}(0,\cdot)))), \bar{u}(T,\cdot)\rangle\\ &= \langle \mathcal{S}_{t_{n-2},t_{n-1}}(\ldots \mathcal{S}_{t_2,t_3}(\mathcal{S}_{t_1,t_2}(\dot{u}(0,\cdot)))), \mathcal{S}^*_{t_{n-1},t_n}(\bar{u}(T,\cdot))\rangle\\ &= \langle \dot{u}(0,\cdot), \mathcal{S}^*_{t_1,t_2}(S^*_{t_2,t_3}(\ldots \mathcal{S}^*_{t_{n-1},t_n}(\bar{u}(T,\cdot))))\rangle\\ &= \langle \dot{u}(0,\cdot), \mathcal{S}^*(\bar{u}(T,\cdot))\rangle \ .\end{aligned}$$

Just like $\mathcal{S}$ corresponds to the chaining of forward in time solution operators the adjoint $\mathcal{S}^*$ corresponds to chaining backwards in time adjoint solution operators.

In order to come up with the adjoint PDE we consider a time step t_i to t_{i+1} with $t_{i+1} - t_i$ small and consider the limit of $t_{i+1} - t_i \equiv \Delta t \to 0$.

By Taylor expansion of the smooth tangent $\dot{u}$ we have

$$\begin{aligned}\dot{u}(t_{i+1},\cdot) &= \dot{u}(t_i,\cdot) + \Delta t \partial_t \dot{u}(t_i,\cdot) + O(\Delta t^2)\\ &= \dot{u}(t_i,\cdot) - \Delta t \partial_x \big(f'(u(t_i,\cdot,p_0))\dot{u}(t_i,\cdot)\big) + O(\Delta t^2)\end{aligned}$$

as $\Delta t \to 0$ since the tangent equation is satisfied and by chain rule we have

$$\partial_x\big(f'(u(t_i,\cdot,p_0))\dot{u}(t_i,\cdot,p_0)\big) = f''(u(t_i,\cdot,p_0))\partial_x u(t_i,\cdot,p_0)\dot{u}(t_i,\cdot) + f'(u(t_i,\cdot,p_0))\partial_x \dot{u}(t_i,\cdot) \ .$$

Similary, by Taylor expansion of the smooth adjoint $\bar{u}$ we have

$$\bar{u}(t_{i+1},\cdot) = \bar{u}(t_i,\cdot) + \Delta t \partial_t \bar{u}(t_i,\cdot) + O(\Delta t^2)$$

as $\Delta t \to 0$.

By the definition of the adjoint operator we have

$$\begin{aligned}\langle \dot{u}(t_{i+1},\cdot), \bar{u}(t_{i+1},\cdot)\rangle &= \langle \dot{u}(t_i,\cdot) - \Delta t \partial_x\big(f'(u(t_i,\cdot,p_0)\dot{u}(t_i,\cdot)\big) + O(\Delta t^2),\\ &\qquad \bar{u}(t_i,\cdot) + \Delta t \partial_t \bar{u}(t_i,\cdot) + O(\Delta t^2)\rangle\\ &= \langle \dot{u}(t_i,\cdot), \bar{u}(t_i,\cdot)\rangle + \Delta t \langle \dot{u}(t_i,\cdot), \partial_t \bar{u}(t_i,\cdot)\rangle\\ &\qquad - \Delta t \langle \partial_x\big(f'(u(t_i,\cdot,p_0)\dot{u}(t_i,\cdot)\big), \bar{u}(t_i,\cdot)\rangle + O(\Delta t^2)\\ &= \langle \dot{u}(t_i,\cdot), \bar{u}(t_i,\cdot)\rangle\end{aligned}$$

as $\Delta t \to 0$, i.e. we need

$$\langle \dot{u}(t_i,\cdot), \partial_t \bar{u}(t_i,\cdot)\rangle = \langle \partial_x\big(f'(u(t_i,\cdot,p_0)\dot{u}(t_i,\cdot)\big), \bar{u}(t_i,\cdot)\rangle\ .$$

By integration by parts we have

$$\begin{aligned}&\langle \partial_x\big(f'(u(t_i,\cdot,p_0)\dot{u}(t_i,\cdot)\big), \bar{u}(t_i,\cdot)\rangle\\ &= \int_\Omega \partial_x\big(f'(u(t_i,x,p_0)\dot{u}(t_i,x)\big)\bar{u}(t_i,x)\mathrm{d}x\\ &= \Big[f'(u(t_i,x,p_0)\dot{u}(t_i,x)\bar{u}(t_i,x)\Big]_\Omega - \int_\Omega f'(u(t_i,x,p_0)\dot{u}(t_i,x)\partial_x\bar{u}(t_i,x)\mathrm{d}x\end{aligned}$$

By the assumption that $\dot{u}$ and $\bar{u}$ have compact support on $\hat{\Omega}$ the first term is zero we need to have

$$\begin{aligned}\langle \dot{u}(t_i,\cdot), \partial_t \bar{u}(t_i,\cdot)\rangle &= \int_\Omega \dot{u}(t_i,x)\partial_t\bar{u}(t_i,x)\mathrm{d}x\\ &= -\int_\Omega f'(u(t_i,x,p)\dot{u}(t_i,x)\partial_x\bar{u}(t_i,x)\mathrm{d}x\ .\end{aligned}$$

which since the involved functions are smooth generally holds if and only if

$$\partial_t\bar{u}(t_i,x) = f'(u(t_i,x,p)\partial_x\bar{u}(t_i,x)$$

for all $x \in \Omega, t \in [t_i, t_{i+1}]$, i.e. the adjoint PDE holds. This concludes the proof. □

We have shown an alternative derivation of the continuous adjoint sensitivity analsysis via considering the time stepping operator and now go on to introducing the discrete sensitivity analysis.

4.2 Discrete Sensitivity Analysis

Similar to the continuous time stepping operators considered in the previous section numerical methods provide discrete time stepping operators for the discrete solution. The corresponding discrete sensitivity time stepping operator results from the differentiation of the finite dimensional multivariate function that defines the discrete time stepping operator of the numerical method. The discrete sensitivities of numerical methods will often correspond to the discretization of the corresponding tangent or adjoint PDEs.

4.2.1 Discrete Time Stepping Operator

We now consider how to obtain the discrete tangent and adjoint sensitivities from a general discrete time stepping operator. Let $\mathcal{H}_{t_j,t_{j+1}}$ be the discrete time stepping operator that maps the discretized state U^j at time t_j to the state U^{j+1} at time t_{j+1}, i.e. the numerical method is defined by

$$U^{j+1} := \mathcal{H}_{t_j,t_{j+1}}(U^j) \; .$$

The discrete tangent sensitivity time step is defined by

$$\dot{U}^{j+1} := \mathrm{d}_U\mathcal{H}_{t_j,t_{j+1}}(U^j)\dot{U}^j$$

where $\mathrm{d}_U\mathcal{H}_{t_j,t_{j+1}}$ is the Jacobian matrix that is analogous to the continuous tangent time stepping operator $\mathcal{S}_{t_j,t_{j+1}}$ in Section 4.1.5. The discrete adjoint sensitivity time step is defined backwards in time by

$$\bar{U}^j := \Big(\mathrm{d}_U\mathcal{H}_{t_j,t_{j+1}}(U^j)\Big)^T \bar{U}^{j+1}$$

where $\Big(\mathrm{d}_U\mathcal{H}_{t_j,t_{j+1}}\Big)^T$ is the transposed Jacobian matrix that is analogous to the continuous adjoint time stepping operator $\mathcal{S}^*_{t_j,t_{j+1}}$.

We consider for example the discrete tangent sensitivities of the conservative Upwind scheme as defined in Section 3.2.6. Let the direction of characteristics be such that the Upwind time step is defined by

$$U_i^{j+1} := U_i^j - \frac{\Delta t}{\Delta x}\big(f(U_i^j) - f(U_{i-1}^j)\big) \; .$$

Directional differentiation in the direction of $\dot{U}$ yields the discrete tangent time step

$$\dot{U}_i^{j+1} := \dot{U}_i^j - \frac{\Delta t}{\Delta x}\big(f'(U_i^j)\dot{U}_i^j - f'(U_{i-1}^j)\dot{U}_{i-1}^j\big) \; .$$

The discrete tangent time step directly represents a finite difference discretization of the tangent PDE in the following form

$$0 = \partial_t\dot{u}(t,x,p) + \partial_x\Big(f'(u(t,x,p))\dot{u}(t,x,p)\Big) \; .$$

Proposition 16. Let f be twice continuously differentiable, u be a smooth solution of the conservation law and $\dot{u}$ be a smooth solution of the tangent PDE. Then the discrete tangent of the Upwind time step is first-order consistent.

Proof. The proof is by Taylor expansion of the smooth solution u and smooth tangent solution $\dot{u}$. For $\Delta t = O(\Delta x)$ we have

$$\begin{aligned}
\dot{u}(t_{j+1},x_i) &= \dot{u}(t_j,x_i) + \Delta t\partial_t\dot{u}(t_j,x_i) + O(\Delta x^2)\\
u(t_j,x_{i-1}) &= u(t_j,x_i) - \Delta x\partial_x u(t_j,x_i) + O(\Delta x^2)\\
\dot{u}(t_j,x_{i-1}) &= \dot{u}(t_j,x_i) - \Delta x\partial_x\dot{u}(t_j,x_i) + O(\Delta x^2)
\end{aligned}$$

as $\Delta x, \Delta t \to 0$.

By the assumption that f is twice continuously differentiable f' is continuously differentiable and by Taylor expansion we have

$$f'(u(t_j,x_{i-1})) = f'(u(t_j,x_i) - \Delta x\partial_x u(t_j,x_i) + O(\Delta x^2))$$

$$= f'(u(t_j,x_i)) - f''(u(t_j,x_i))\Delta x \partial_x u(t_j,x_i) + O(\Delta x^2)$$

as $\Delta x \to 0$. Consequently, we also have

$$\begin{aligned} f'(u(t_j,x_{i-1}))\dot{u}(t_j,x_{i-1}) = \Big(f'(u(t_j,x_i)) - f''(u(t_j,x_i))\Delta x \partial_x u(t_j,x_i)\Big)\cdot \\ \Big(\dot{u}(t_j,x_i) - \Delta x \partial_x \dot{u}(t_j,x_i)\Big) + O(\Delta x^2) \end{aligned}$$

as $\Delta x \to 0$. The product on the right-hand side expands to

$$f'(u(t_j,x_i))\dot{u}(t_j,x_i) - f''(u(t_j,x_i))\Delta x \partial_x u(t_j,x_i)\dot{u}(t_j,x_i) - f'(u(t_j,x_i))\Delta x \partial_x \dot{u}(t_j,x_i) + O(\Delta x)^2$$

as $\Delta x \to 0$. By chain rule we have

$$\begin{aligned} & f''(u(t_j,x_i))\Delta x \partial_x u(t_j,x_i)\dot{u}(t_j,x_i) + f'(u(t_j,x_i))\Delta x \partial_x \dot{u}(t_j,x_i) \\ &\quad = \Delta x\Big(\partial_x f'(u(t_j,x_i))\dot{u}(t_j,x_i) + f'(u(t_j,x_i))\partial_x \dot{u}(t_j,x_i)\Big) \\ &\quad = \Delta x \partial_x \Big(f'(u(t_j,x_i))\dot{u}(t_j,x_i)\Big) \end{aligned}$$

and, consequently,

$$f'(u(t_j,x_{i-1}))\dot{u}(t_j,x_{i-1}) = f'(u(t_j,x_i))\dot{u}(t_j,x_i) - \Delta x \partial_x \Big(f'(u(t_j,x_i))\dot{u}(t_j,x_i)\Big) + O(\Delta x^2)$$

as $\Delta x \to 0$.

The local truncation error of the discrete tangent time step is

$$\begin{aligned} & \dot{u}(t_j,x_i) - \frac{\Delta t}{\Delta x}\big(f'(u(t_j,x_i))\dot{u}(t_j,x_i) - f'(u(t_j,x_{i-1}))\dot{u}(t_j,x_{i-1})\big) - \dot{u}(t_{j+1},x_i) \\ &\quad = -\frac{C_t \Delta x}{\Delta x}\big(f'(u(t_j,x_i))\dot{u}(t_j,x_i) - f'(u(t_j,x_{i-1}))\dot{u}(t_j,x_{i-1})\big) - C_t \Delta x \partial_t \dot{u}(t_j,x_i) + O(\Delta x^2) \\ &\quad = -C_t\big(\Delta x \partial_x \big(f'(u(t_j,x_i))\dot{u}(t_j,x_i) + O(\Delta x^2)\big) - C_t \Delta x \partial_t \dot{u}(t_j,x_i) + O(\Delta x^2) \\ &\quad = -C_t \Delta x\big(\partial_x\big(f'(u(t_j,x_i))\dot{u}(t_j,x_i) + \partial_t \dot{u}(t_j,x_i)\big) + O(\Delta x^2) = O(\Delta x^2)\,. \end{aligned}$$

This concludes the proof. □

In order to obtain the discrete adjoint time step we need to consider how to left-multiply the transposed Jacobian of the Upwind time step. In order to produce the adjoint $\bar{U}_i^j$ we need to consider that U_i^j influences both U_i^{j+1} and U_{i+1}^{j+1} and we have

$$\mathrm{d}_{U_i^j} U_i^{j+1} = 1 - \frac{\Delta t}{\Delta x} f'(U_i^j)$$

and

$$\mathrm{d}_{U_i^j} U_{i+1}^{j+1} = \frac{\Delta t}{\Delta x} f'(U_i^j)\,.$$

A systematic method for computing the discrete adjoint sensitivity of any numerical method is by AD as described in Section 4.3.

The transpose Jacobian vector product gives us the discrete adjoint time step

$$\bar{U}_i^j := \mathrm{d}_{U_i^j} U_i^{j+1} \cdot \bar{U}_i^{j+1} + \mathrm{d}_{U_i^j} U_{i+1}^{j+1} \cdot \bar{U}_{i+1}^{j+1}$$

$$
\begin{aligned}
&= \bar{U}_i^{j+1} - \frac{\Delta t}{\Delta x} f'(U_i^j)\bar{U}_i^{j+1} + \frac{\Delta t}{\Delta x} f'(U_i^j)\bar{U}_{i+1}^{j+1} \\
&= \bar{U}_i^{j+1} + \frac{\Delta t}{\Delta x} f'(U_i^j)\left(\bar{U}_{i+1}^{j+1} - \bar{U}_i^{j+1}\right) .
\end{aligned}
$$

The discrete adjoint time step directly represents a backwards in time finite difference discretization of the adjoint PDE in the following form

$$0 = \partial_t \bar{u}(t,x,p) + f'(u(t,x,p))\partial_x \bar{u}(t,x,p) .$$

4.2.2 Consistency of Adjoint and Tangent

We now give a results that is useful for showing the convergence of a discrete adjoint sensitivity based on the definition of a convergent discrete tangent sensitivity.

The adjoint sensitivity of the tangent sensitivity of a function f is the same as the adjoint sensitivity of f. Alternatively, the linearization of a linear function is the same linear function.

Lemma 17. Let $f, g, h \colon X \to Y$ be differentiable functions such that $g(x) \equiv f'(z_0)x$ is the linear operator defined by the derivative of f at z_0 and $h(u) \equiv g'(x_0)u$ is the linear operator defined by the derivative of g at x_0. Then we have

$$h^*(v) = g^*(v)$$

for all $v \in X$, i.e. the adjoint of the derivative of the tangent is the adjoint of the derivative.

Proof. The lemma follows directly from the definitions.

The tangent and adjoint sensitivity of f are

$$g(x) = f'(z_0)x, \qquad g^*(y) = f'(z_0)^* y .$$

The tangent sensitivity of the tangent of f is

$$h(u) = g'(x_0)u = f'(z_0)u .$$

The adjoint sensitivitiy of the tangent of f is

$$h^*(v) = g'(x_0)^* v = f'(z_0)^* v = g^*(v) .$$

This concludes the proof. □

If a tangent approximation converges to the analytic tangent at some rate then the adjoint approximation converges to the analytic adjoint at the same rate.

Lemma 18. Let $f \colon \mathbb{R}^m \to \mathbb{R}^n$ be a differentiable function and

$$\dot{y} \equiv f'(x)\dot{x}, \qquad \bar{x} \equiv f'(x)^T \bar{y}$$

be the tangent and adjoint sensitivities at x.

Let $F^h \colon \mathbb{R}^m \to \mathbb{R}^n$ be a differentiable function and

$$\dot{Y}^h \equiv (F^h)'(x)\dot{X}^h, \qquad \bar{X}^h \equiv \left((F^h)'(x)\right)^T \bar{Y}^h$$

be the tangent and adjoint sensitivity at x for all $h > 0$. Furthermore, for $\dot{X}^h \equiv \dot{x}$ let the tangent $\dot{Y}^h$ be an approximation of the tangent $\dot{y}$ with

$$\dot{y} - \dot{Y}^h = O(g(h))$$

as $h \to 0$.

Then for $\bar{Y}^h \equiv \bar{y}$ the adjoint $\bar{X}^h$ is and approximation of the adjoint $\bar{x}$ with

$$\bar{x} - \bar{X}^h = O(g(h))$$

as $h \to 0$.

Proof. Based on the assumptions and the definition of the adjoint we have

$$\begin{aligned}
\langle \dot{Y}^h, \dot{X}^h \rangle &= \langle \bar{Y}^h, \bar{X}^h \rangle && \text{(definition of adjoint operator)}\\
\langle \dot{y} + O(g(h)), \dot{x} \rangle &= \langle \bar{y}, \bar{X}^h \rangle && \text{(assumptions in lemma)}\\
\langle \dot{y}, \dot{x} \rangle + O(g(h)) &= \langle \bar{y}, \bar{X}^h \rangle && (\dot{x} \text{ constant})\\
\langle \dot{y}, \dot{x} \rangle + O(g(h)) &= \langle \bar{y}, \bar{x} + \bar{X}^h - \bar{x} \rangle && \text{(add zero)}\\
\langle \dot{y}, \dot{x} \rangle + O(g(h)) &= \langle \bar{y}, \bar{x} \rangle + \langle \bar{y}, \bar{X}^h - \bar{x} \rangle && \text{(distributivity)}\\
O(g(h)) &= \langle \bar{y}, \bar{X}^h - \bar{x} \rangle && \text{(definition of adjoint operator)}\\
O(g(h)) &= \bar{X}^h - \bar{x} && (\bar{y} \text{ is constant})
\end{aligned}$$

as $h \to 0$. This concludes the proof. □

We have provided two useful lemmas that illustrate the relationship between the discrete adjoint and tangent sensitivity approximations and now go on to explain how the discrete adjoint and tangent can be computed by algorithmic differentiation.

4.3 Algorithmic Differentiation

In this section we introduce the basic ideas behind algorithmic differentiation (AD) of numerical algorithms. For a complete introduction see Griewank and Walther [GW08] and Naumann [Nau12]. A vast literature exists about AD applied to numerical methods for differential equations, e.g. Hannemann, Marquardt, Naumann, and Gendler [Han+10], Towara, Lotz, and Naumann [TLN19], Sen, Towara, and Naumann [STN14], Cardesa, Hascoët, and Airiau [CHA20], Müller and Cusdin [MC05], Giles, Ghate, and Duta [GGD05], and Jones, Müller, and Christakopoulos [JMC11] just to name a few references.

4.3.1 Numerical Algorithms as Imperative Programs

An imperative programming language uses statements that are executed sequentially in order to change the program's state. Control flow structures are used to control the order of execution of statements.

An imperative programming language can be helpful when defining large and nested mathematical functions in a concise manner.

Example 15 (Simple loop)**.** Let us for example consider the algorithm defined by the following imperative program

Algorithm 1: Simple Loop Example

```
Input: $x$
$y := \exp(x)$ ;
for $i = 1, \dots, n$ do
    $y := \sin(y) \cdot y \cdot x$ ;
end
Output: $y$
```

For $n = 2$ the algorithm corresponds to the closed form mathematical function definition

$$f(x) \equiv \sin(\sin(\exp(x)) \exp(x)x) \sin(\exp(x)) \exp(x)x^2 \ .$$

For $n = 3$ the closed form mathematical function is

$$\begin{aligned} f(x) \equiv \sin(\sin(\sin(\exp(x)) \exp(x)x) \sin(\exp(x)) \exp(x)x^2) \cdot \\ \sin(\sin(\exp(x)) \exp(x)x) \sin(\exp(x)) \exp(x)x^3 \ . \end{aligned}$$

The length of the closed form expression defining the mathematical function grows exponentially in n.

Another example would be a numerical integrator such as the Upwind method or the Lax-Friedrichs method described in Section 3.2. The integrator is a function that maps the discretization of the initial data $\{U_i^0\}_{i \in \{1,\dots,n\}}$ to the approximate solution at the final time step $\{U_i^m\}_{i \in \{1,\dots,n\}}$. The integrator is a numerical algorithm that describes a function

$$f \colon D \subseteq \mathbb{R}^n \to \mathbb{R}^n$$

that is implemented by an imperative program.

In general the goal of algorithmic differentiation is to compute the derivatives of a function

$$f \colon D \subseteq \mathbb{R}^m \to \mathbb{R}^n$$

that is implemented by a computer program. We evaluate the derivatives at a specific argument, or program input, $x = x_0$ and make the following assumptions.

Assumption 1. When evaluated at the input x the program terminates and returns $f(x)$ for all x in the neighborhood of x_0.

Assumption 2. When evaluated at the input x the program executes the same statements as when evaluated at x_0 for all x in the neighborhood of x_0, i.e. the control flow is constant in a neighborhood of the input.

Since we are concerned with the differentiation of the function that is implemented by a program we are primarily concerned with those parts of a program that deals with the floating point representations of real numbers. All variables that are somehow discrete are assumed to be constant in a neighborhood of the given input just like the control flow.

Assumption 3. For all x in the neighborhood of x_0 we assume the output variable y to depend on x only through assignments with differentiable right-hand sides.

By the chain rule and by the assumption that the assignments are unchanged in a neighborhood of x_0 the function f is then also differentiable in the neighborhood of x_0.

Example 16 (Simple loop continued)**.** For example for all $x \in \mathbb{R}$ the statements executed for the example program with $n = 2$ are obtained by unrolling the loop

```
Algorithm 2: Unrolled Simple Loop Example
  Input: $x$
  $y := \exp(x)$ ;
  $y := \sin(y) \cdot y \cdot x$ ;
  $y := \sin(y) \cdot y \cdot x$ ;
  Output: $y$
```

We see that for all assignments the right-hand side is differentiable with respect to all the variables that occur in them.

4.3.2 Chain Rule on the Computational Graph

Since the part of the program we are interested in are the assignments through which the output depends on the input we take a program to be a sequence of assignments to intermediate program variables. We consider the rewriting to single assignment program variables of increasing index. Intermediate program variables are assigned as

$$v_i := f_i(v_1, \ldots, v_{i-1})$$

for $i \in \{1, \ldots, s\}$ where the right-hand side of an assignment only depends on previously assigned program variables. Contrary to what the notation used above might suggest not every right-hand side has to depend on all of the previously assigned program variables.

Example 17 (Simple loop continued)**.** The single assignment program variable version of the unrolled simple loop example is

```
Algorithm 3: Single Assignment Unrolled Simple Loop Example
  Input: $v_1$
  $v_2 := \exp(v_1)$ ;
  $v_3 := \sin(v_2) \cdot v_2 \cdot v_1$ ;
  $v_4 := \sin(v_3) \cdot v_3 \cdot v_1$ ;
  Output: $v_4$
```

with $v_1 = x, y = v_4, s = 4$.

We can define a directed graph based on the data flow dependence of the intermediate program variables. The vertices of the graph represent program variables and the edges represent their direct data flow dependence on previously defined program variables.

Definition 33. The *computational graph* is a directed graph $G = (V, E)$ with $V = \{1, \ldots, s\}$ and $(i, j) \in E$ if and only if v_i occurs on the right-hand side of the assignment of v_j.

For all edges (i, j) we have $i < j$ and the graph is acyclic since we cannot have any backward edges by definition due to the single assignment enumeration.

Example 18 (Simple loop continued)**.** For Example 17 we have $G = (V, E)$ with

$$V = \{1, \ldots, 4\}$$

and

$$E = \{(1, 2), (1, 3), (1, 4), (2, 3), (3, 4)\} .$$

We now augment the computational graph by annotating the edges with the local partial derivatives of the assignments.

Definition 34. The computational graph *annotation* is an edge label function $l\colon E \to \mathbb{R}$ with

$$l((i, j)) \equiv \partial_{v_i} f_j(v_1, \ldots, v_{j-1}) .$$

The numerical value of the label depends on a specific chosen input at which the program is evaluated just like all the intermediate variable values.

Example 19 (Simple loop continued)**.** For the computational graph of Example 17 we have the annotation

$$\begin{aligned} l((1, 2)) &= \mathrm{d}_{v_1} v_2 = \partial_{v_1} f_2(v_1) = \partial_{v_1} \exp(v_1) = \exp(v_1) \\ l((2, 3)) &= \mathrm{d}_{v_2} v_3 = \partial_{v_2} f_3(v_2, v_1) = \partial_{v_2} (\sin(v_2) v_2 v_1) = \cos(v_2) v_2 v_1 + \sin(v_2) v_1 \\ l((1, 3)) &= \mathrm{d}_{v_1} v_3 = \partial_{v_1} f_3(v_2, v_1) = \partial_{v_1} (\sin(v_2) v_2 v_1) = \sin(v_2) v_2 \end{aligned}$$

and so on.

We now consider how to evaluate the derivatives of the function f that is evaluated by the program via the chain rule as it applies to the computational graph.

We give a standard definition of a path that connects two vertices in the computational graph.

Definition 35. An (i, j)*-path* π is a tuple of sequentially connected edges going from i to j, i.e.

$$\pi = \big((i, k_1), (k_1, k_2), (k_2, k_3), \ldots, (k_{Q-1}, k_Q), (k_Q, j)\big)$$

such that $(i, k_1), (k_Q, j) \in E$ and $(k_q, k_{q+1}) \in E$ for all $q \in \{1, \ldots, Q - 1\}$.

We denote the set of all unique (i, j)-paths as

$$\mathrm{paths}(i, j) = \{\pi \mid \pi \text{ is an } (i, j)\text{-path}\} .$$

Proposition 19. The derivative of the intermediate variable v_j with respect to v_i is

$$\begin{aligned} \mathrm{d}_{v_i} v_j &= \sum_{\pi \in \mathrm{paths}(i,j)} \prod_{q=1}^{\dim(\pi)} l((\pi_q)_1, (\pi_q)_2) \\ &= \sum_{\pi \in \mathrm{paths}(i,j)} \prod_{q=1}^{\dim(\pi)} \partial_{v_{(\pi_q)_1}} v_{(\pi_q)_2} \end{aligned}$$

for all $i, j \in \{1, \ldots, s\}$ with $i < j$.

Proof. By the chain rule, see Bauer [Bau74]. □

Example 20 (Simple loop continued)**.** We consider the derivative $\mathrm{d}_{v_3}v_1$ in Example 17. The two paths from vertex 1 to vertex 3 are

$$\pi_1 = \big((1,2),(2,3)\big), \qquad \pi_2 = \big((1,3)\big) \ .$$

According to Proposition 19 we have

$$\begin{aligned}\mathrm{d}_{v_1}v_3 &= \partial_{v_1}v_2 \cdot \partial_{v_2}v_3 + \partial_{v_1}v_3 \\ &= \exp(v_1)\cdot(\cos(v_2)v_2v_1 + \sin(v_2)v_1) + \sin(v_2)v_2 \\ &= \exp(v_1)^2\cos(v_1)v_1 + \exp(v_1)\sin(\exp(v_1))v_1 + \sin(\exp(v_1))\exp(v_1) \ .\end{aligned}$$

If we write the closed form expression for $v_3 = g(v_1)$ we have

$$g(v_1) = \sin(\exp(v_1))\exp(v_1)v_1$$

and, hence, by chain rule also get the same derivative

$$g'(v_1) = \cos(\exp(v_1))\exp(v_1)^2v_1 + \sin(v_1)\exp(v_1)v_1 + \sin(\exp(v_1)\exp(v_1) \ .$$

That the chain rule applies in such a way motivates the use of AD methods. In the next sections we present the forward and reverse modes of AD that are both methods for multiplying annotated computational graph edge labels according to the chain rule.

4.3.3 Forward Mode AD

The forward mode of AD traverses the vertices of the computational graph in order.

For each vertex j corresponding to the intermediate program variable v_j the algorithm assigns the tangent variable $\dot{v}_j$. At each vertex we sum over all the incoming edges, multiplying their labels with the tangent variable belonging to their source vertex. The tangent variables are assigned by the following rule

$$\dot{v}_j := \sum_{i\in\{i \mid (i,j)\in E\}} l((i,j))\dot{v}_i = \sum_{i\in\{i \mid (i,j)\in E\}} \mathrm{d}_{v_i}v_j \cdot \dot{v}_i \ .$$

The forward mode of AD computes a tangent sensitivity of one node with respect to another.

Proposition 20. Let $\dot{v}_k := 0$ for all $k < i$ then the forward mode of AD started at index $i+1$ computes

$$\dot{v}_j = \mathrm{d}_{v_i}v_j \cdot \dot{v}_i$$

for any two vertices i, j with $i < j$.

Proof. By the chain rule, see e.g. Theorem 2.6 in Naumann [Nau12, Chapter 2.1]. □

We have to assign $\dot{v}_i$ manually because since the algorithm is started at index $i+1$ it does not assign $\dot{v}_k$ for $k < i+1$. We call this manual assignment *seeding* the tangent variable.

If for two vertices i, k we have $\text{paths}(i,k) = \emptyset$ and we initially seed all tangent variables to zero except for $\dot{v}_i$ and $\dot{v}_k$ then after the forward mode computation we have

$$\dot{v}_j = \mathrm{d}_{v_i} v_j \cdot \dot{v}_i + \mathrm{d}_{v_k} v_j \cdot \dot{v}_k \ .$$

The tangent sensitivity for two variables generalizes to more than two path-independent vertices as well. Consequently, if we have m input variables x_i to a function we can seed all of their corresponding tangent variables and obtain the tangent sensitivity

$$\dot{v}_j = \sum_{i=1}^{m} \mathrm{d}_{x_i} v_j \cdot \dot{x}_i = \langle \mathrm{d}_x v_j, \dot{x}_i \rangle$$

where $\mathrm{d}_x v_j \in \mathbb{R}^m$ denotes the gradient of v_j with respect to x.

Example 21 (Simple loop continued)**.** When seeding just $\dot{v}_1$ for Example 17 the forward mode AD computation yields the following tangent variables:

- Visiting node 2:
$$\dot{v}_2 := \mathrm{d}_{v_1} v_2 \cdot \dot{v}_1 = \exp(v_1)\dot{v}_1$$
- Visiting node 3:
$$\dot{v}_3 := \mathrm{d}_{v_2} v_3 \cdot \dot{v}_2 + \mathrm{d}_{v_1} v_3 \cdot \dot{v}_1 = \big(\cos(v_2)v_2v_1 + \sin(v_2)v_1\big)\dot{v}_2 + \sin(v_2)v_2\dot{v}_1$$
- Visiting node 4:
$$\dot{v}_4 := \mathrm{d}_{v_3} v_4 \cdot \dot{v}_3 + \mathrm{d}_{v_1} v_4 \cdot \dot{v}_1 = \big(\cos(v_3)v_3v_1 + \sin(v_3)v_1\big)\dot{v}_3 + \sin(v_3)v_3\dot{v}_1 \ .$$

For $\dot{v}_3$ we see by substitution that we compute the correct tangent sensitivity (c.f. Example 20)

$$\begin{aligned}
\dot{v}_3 &= \big(\cos(v_2)v_2v_1 + \sin(v_2)v_1\big)\dot{v}_2 + \sin(v_2)v_2\dot{v}_1 \\
&= \big(\cos(v_2)v_2v_1 + \sin(v_2)v_1\big)\exp(v_1) \cdot \dot{v}_1 + \sin(v_2)v_2\dot{v}_1 \\
&= \big(\cos(\exp(v_1))\exp(v_1)v_1 + \sin(\exp(v_1))v_1\big)\exp(v_1) \cdot \dot{v}_1 + \sin(\exp(v_1))\exp(v_1)\dot{v}_1 \\
&= \mathrm{d}_{v_1} v_3 \cdot \dot{v}_1 \ .
\end{aligned}$$

Foward mode AD can be implemented via operator overloading.

Without being specific to a programming language we can consider a forward mode AD data type that replaces the standard floating point data type and for which the standard operators and elemental function are defined in such a way that they compute the forward mode AD evaluation interleaved with the standard computation.

The data type of all program variables v_i that are part of the computational graph needs to be replaced with the AD data type. In practice, the AD data type will have two parameters, (i) a value parameter that carries the value v_i, and (ii) a derivative parameter that carries the tangent $\dot{v}_i$.

Example 22 (Operator overloading)**.** To provide examples that sufficiently illustrate how operator overloading is in principle implemented we consider the overloading of a binary operator and two elemental functions for a forward mode AD data type.

The definition of the overloaded exp function is

Algorithm 4: Forward mode AD overloading of $y = \exp(x)$

```
Input: $x$.value, $x$.derivative
$y$.value := $\exp(x$.value$)$ ;
$y$.derivative := $\exp(x$.value$) \cdot x$.derivative ;
Output: $y$.value, $y$.derivative
```

The definition of the overloaded sin function is

Algorithm 5: Forward mode AD overloading of $y = \sin(x)$

```
Input: $x$.value, $x$.derivative
$y$.value := $\sin(x$.value$)$ ;
$y$.derivative := $\cos(x$.value$) \cdot x$.derivative ;
Output: $y$.value, $y$.derivative
```

The definition of the overloaded multiplication operator is

Algorithm 6: Forward mode AD overloading of $y = x_1 \cdot x_2$

```
Input: $x_1$.value, $x_1$.derivative, $x_2$.value, $x_2$.derivative
$y$.value := $x_1$.value $\cdot$ $x_2$.value ;
$y$.derivative := $x_1$.value $\cdot$ $x_2$.derivative $+$ $x_1$.derivative $\cdot$ $x_2$.value ;
Output: $y$.value, $y$.derivative
```

4.3.4 Reverse Mode AD

The reverse mode of AD traverses the vertices of the computational graph in reverse order.

For each vertex i corresponding to the intermediate program variable v_i the algorithm assigns the adjoint variable $\bar{v}_i$. The idea of reverse mode AD is in principle symmetric to the forward mode and the goal is to have

$$\bar{v}_i = \sum_{j \in \{j \mid (i,j) \in E\}} l((i,j))\bar{v}_j = \sum_{j \in \{j \mid (i,j) \in E\}} \mathrm{d}_{v_i} v_j \cdot \bar{v}_j \ . \tag{4.3.1}$$

However, unlike the forward mode, the reverse mode potentially makes multiple assignments per visited vertex. The forward mode sums over incoming edges for each vertex, multiplying the edge label with the tangent variable of the source vertex. The reverse mode sums over outgoing edges for each vertex, multiplying the edge label with the adjoint variable of the target vertex.

In practice reverse mode AD is implemented as writing into the adjoint variables of vertices that are connected through incoming edges as follows. Upon visiting vertex j we perform the following assignment

$$\bar{v}_i := \bar{v}_i + l((i,j))\bar{v}_j$$

for all $(i, j) \in E$. If $\bar{v}_i$ is initialized to zero then after visiting all vertices j with $(i, j) \in E$ we have the state that satisfies Equation (4.3.1).

Proposition 21. Let $\bar{v}_k := 0$ for all $k > j$ then the reverse mode of AD started at index $j - 1$ computes

$$\bar{v}_i = \mathrm{d}_{v_i} v_j \cdot \bar{v}_j$$

for any two vertices i, j with $i < j$.

Proof. By the adjoint of a chained Jacobian product, see e.g. Theorem 2.14 in Naumann [Nau12, Chapter 2.2]. □

We have to assign $\bar{v}_j$ manually because as the algorithm is started at index $j - 1$ it does not assign $\bar{v}_k$ for all $k > j - 1$. We call this manual assignment *seeding* the adjoint variables.

For a scalar intermediate variable v_j and m input variables x_i we have

$$\bar{x}_i = \mathrm{d}_{x_i} v_j \cdot \bar{v}_j$$

for all $i \in \{1, \dots, m\}$, i.e.

$$\bar{x} = \mathrm{d}_x v_j \cdot \bar{v}_j$$

where $\bar{x} \in \mathbb{R}^m$ denotes the adjoints of the inputs as a vector and $\mathrm{d}_x v_j \in \mathbb{R}^m$ denotes the gradient of v_j with respect to x.

Example 23 (Simple loop continued)**.** When seeding $\bar{v}_4$ for Example 17 the reverse mode AD computation yields the following adjoint variables:

- Visiting node 4:

$$\begin{aligned}\bar{v}_3 &:= \bar{v}_3 + \mathrm{d}_{v_3} v_4 \cdot \bar{v}_4 = \bar{v}_3 + \big(\cos(v_3) v_3 v_1 + \sin(v_3) v_1\big) \bar{v}_4 \\ \bar{v}_1 &:= \bar{v}_1 + \mathrm{d}_{v_1} v_4 \cdot \bar{v}_4 = \bar{v}_1 + \sin(v_3) v_3 \bar{v}_4\end{aligned}$$

- Visiting node 3:

$$\begin{aligned}\bar{v}_2 &:= \bar{v}_2 + \mathrm{d}_{v_2} v_3 \cdot \bar{v}_3 = \bar{v}_2 + \big(\cos(v_2) v_2 v_1 + \sin(v_2) v_1\big) \bar{v}_3 \\ \bar{v}_1 &:= \bar{v}_1 + \mathrm{d}_{v_1} v_3 \cdot \bar{v}_3 = \bar{v}_1 + \sin(v_2) v_2 \bar{v}_3\end{aligned}$$

- Visiting node 2:

$$\bar{v}_1 := \bar{v}_1 + \mathrm{d}_{v_1} v_2 \cdot \bar{v}_2 = \bar{v}_1 + \exp(v_1) \bar{v}_2 \ .$$

By substitution we see that when seeding $\bar{v}_i := 0$ for $i \in \{1, 2, 4\}$ we compute the correct adjoint sensitivity (c.f. Example 20)

$$\begin{aligned}\bar{v}_1 &= \exp(v_1) \bar{v}_2 + \sin(v_2) v_2 \bar{v}_3 + \sin(v_3) v_3 \bar{v}_4 \\ &= \exp(v_1)\big(\cos(v_2) v_2 v_1 + \sin(v_2) v_1\big) \bar{v}_3 + \sin(v_2) v_2 \bar{v}_3 \\ &= \exp(v_1)\big(\cos(\exp(v_1)) \exp(v_1) v_1 + \sin(\exp(v_1)) v_1\big) \bar{v}_3 + \sin(\exp(v_1)) \exp(v_1) \bar{v}_3 \\ &= \mathrm{d}_{v_1} v_3 \cdot \bar{v}_3 \ .\end{aligned}$$

Reverse mode AD can be implemented via operator overloading.

A reverse mode AD data type is similar to the forward mode data type in that it contains two parameters, (i) a value parameter that carries the value v_i, and (ii) a derivative parameter that carries the adjoint $\bar{v}_i$ of the corresponding program variable.

As opposed to the forward mode the reverse mode computation requires a representation of the computational graph in memory. The reason that the forward mode does not require a graph data structure is that the vertices are visited in the same order as during the usual forward computation. But in the reverse mode we need to visit the vertices in the reverse order and upon visiting need to access vertices connected through incoming edges in the graph.

The reverse mode AD data type will thus contain some data structural parameters necessary for the graph representation that allows for reversal. The way the computational graph representation is implemented through data structures is not unique and we give only one possible example.

In our example implementation the reverse mode AD data type has a parameter that stores the integer vertex id i corresponding to the intermediate variable v_i. Furthermore, the data type stores a list of pairs of integer source vertex ids and floating point edge labels for all incoming edges.

Maintaining the graph itself additionally requires us to store a list of vertices visited by the program in the forward computation. The reverse sweep of reverse mode AD is implemented via the following algorithm.

Algorithm 7: Reverse sweep

```
for $j = s, s-1, \dots, 1$ do
    $x$ := vertices[$j$] ;
    $l$ := length($x$.incoming_edges) ;
    for $i = 1, \dots, l$ do
        vertices[$x$.incoming_edges[$i$].first].derivative :=
          vertices[$x$.incoming_edges[$i$].first].derivative + $x$.incoming_edges[$i$].second ·
          $x$.derivative ;
    end
end
```

Example 24. To provide examples that sufficiently illustrate how operator overloading is in principle implemented we consider the overloading of an operator and two elemental functions for a reverse mode AD data type.

The definition of the overloaded exp function is

Algorithm 8: Reverse mode AD overloading of $y = \exp(x)$

```
$y$.value := exp($x$.value) ;
$y$.derivative := 0 ;
$y$.vertex_id := length(vertices) ;
$y$.incoming_edges.push(($x$.vertex_id, exp($x$.value))) ;
vertices.push($y$) ;
```

The definition of the overloaded sin function is

Algorithm 9: Reverse mode AD overloading of $y = \sin(x)$

```
$y$.value := sin($x$.value) ;
$y$.derivative := 0 ;
$y$.vertex_id := length(vertices) ;
$y$.incoming_edges.push$\big((x.\text{vertex\_id}, \cos(x.\text{value}))\big)$ ;
vertices.push($y$) ;
```

The definition of the overloaded multiplication operator is

Algorithm 10: Reverse mode AD overloading of $y = x_1 \cdot x_2$

```
$y$.value := $x_1$.value $\cdot$ $x_2$.value ;
$y$.derivative := 0 ;
$y$.vertex_id := length(vertices) ;
$y$.incoming_edges.push$\big((x_1.\text{vertex\_id}, x_2.\text{value})\big)$ ;
$y$.incoming_edges.push$\big((x_2.\text{vertex\_id}, x_1.\text{value})\big)$ ;
vertices.push($y$) ;
```

We introduced the basic aspects of tangent and adjoint sensitivity analysis of smooth PDE solutions. We also discussed the discrete tangent and adjoint time steps obtained via the differentiation of numerical integrators and how these discrete sensitivities can be computed by the forward and reverse mode of AD tools. In the next chapter we consider the tangent and adjoint sensitivity analysis of piecewise smooth discontinuous solutions of scalar conservation laws.

Chapter 5

Sensitivity Analysis for Discontinuous Solutions

In this chapter we extend the analytic sensitivity analysis for smooth solutions of scalar conservations laws from the previous chapter to piecewise smooth but discontinuous weak solutions. We assume a specific solution structure with differentiable shock discontinuity locations. We derive the tangent and generalized tangent sensitivities for the assumed solution structure based on the sensitivities of the shock locations and the weak differentiation in the sense of distributions. We also derive the adjoint sensitivity based on the notion of products of generalized functions in the sense of the Colombeau algebra (see Section 2.4.5).

5.1 Basic Assumptions

5.1.1 Flux Functions

The first technical assumption we make is regarding the flux function f. We already stated this assumption in the introduction and in the context of the tangent sensitivity analysis in the previous chapter.

Assumption 4. Let $f\colon \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable.

For the computational evaluation of the discrete sensitivity methods for scalar conservation laws presented in the next chapter we use two example flux functions. The Burgers flux defined by

$$f(u) \equiv \frac{1}{2}u^2$$

has the first and second derivatives

$$f'(u) = u, \qquad f''(u) = 1 \; .$$

The Buckley-Leverett flux [GZ10; WK13] defined by

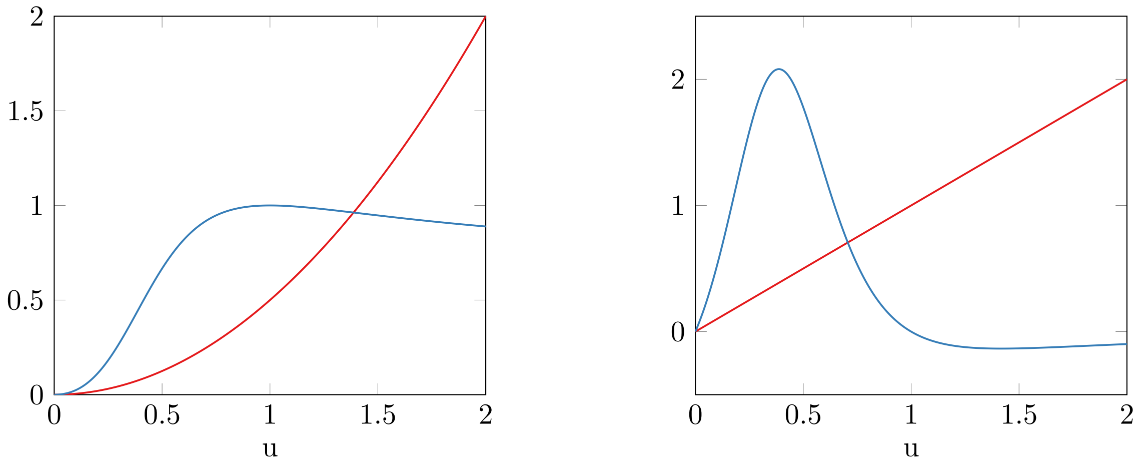
$$f(u) \equiv \frac{u^2}{u^2 + \frac{1}{2}(1-u)^2}$$

has the first and second derivatives

$$f'(u) = \frac{4u(1-u)}{(3u^2-2u+1)^2}, \qquad f''(u) = 4\frac{6u^3-9u^2+1}{(3u^2-2u+1)^3} .$$

Both the Burgers and the Buckley-Leverett flux functions are twice continuously differentiable on the interval $[0,2] \subseteq \mathbb{R}$, the relevant domain of solution values u we are interested in for the examples in this thesis.

Figure 5.1 shows the Burgers and the Buckley-Leverett flux functions and their first derivatives on the domain relevant to the computational results in Section 6.6.



(a) Burgers f (red), Buckley-Leverett f (blue)

(b) Burgers f' (red), Buckley-Leverett f' (blue)

Figure 5.1: Flux functions and their first derivatives

5.1.2 Compact Support of Weak Solution

The second assumption we make is regarding the solution u. Since we want to compute a numerical approximation of u by a finite representation, i.e. a spacial grid with finitely many points, we have to be able to bound the relevant domain. To that end we assume that the solution has compact support.

Assumption 5. For some $\hat{\Omega} \equiv [L,R]$ with $R-L<\infty$ let

$$u(t,x,p_0) = 0$$

for all $x \in \Omega \setminus \hat{\Omega}$, $t \in [0,T]$.

In order to not have to deal with boundary conditions in the PDE we assume that the weak solution u is zero everywhere outside of a domain $\hat{\Omega}$ with bounded width. We can then perform the numerical approximation of u on $\hat{\Omega}$ and use the known exact solution value (zero) on $\Omega \setminus \hat{\Omega}$.

For example, the compact support assumption holds for the ramp initial value problem posed in Example 7 for any $\hat{\Omega} \supseteq \left[0, \sqrt{1+(1+p)T}\right]$.

The compact support assumption does not hold for the Riemann initial value problem posed in Example 6 since for all $x \geq T(1+p)/2$, $t \geq 0$ we have $u(t,x,p) = 1+p \neq 0$ if $p \neq -1$.

In practice the violation of the assumption is not a problem since we are not interested in the approximation of the solution for $x \to \pm\infty$.

5.1.3 Piecewise Smooth Solution Structure

We now state the main assumption regarding the solution structure that explicitly enumerates the shock discontinuities contained in the solution. The structure we assume is not the most general structure but rather one that makes the later derivations not unnecessarily technical by focussing on the essential features of the differentiation around the shock discontinuities. It is, for example, possible to allow for Lipschitz continuous nonsmoothness in the solution where it now is assumed to be continuously differentiable with respect to t and x as we see in the ramp initial value problem example.

The weak solution is piecewise smooth with a finite number of strong shock discontinuities the location curves of which do not intersect over time.

Assumption 6. Let

$$u(t,x,p_0) = \begin{cases} u_1(t,x,p_0) & \text{if } x \in (\xi_0, \xi_1(t,p_0)) \\ u_2(t,x,p_0) & \text{if } x \in (\xi_1(t,p_0), \xi_2(t,p_0)) \\ \quad\vdots & \quad\vdots \\ u_{K+1}(t,x,p_0) & \text{if } x \in (\xi_K(t,p_0), \xi_{K+1}) \end{cases}$$

with $K < \infty$, $\xi_0 = L$, $\xi_{K+1} = R$. The variables $\xi_1(t,p_0) < \cdots < \xi_K(t,p_0)$ are the locations of strong shock discontinuities in u with initial values $\xi_k(0,p_0) = (\xi_k)_0(p_0)$ that are continuously differentiable with respect to p in a neighborhood of p_0 for all $k \in \{1,\ldots,K\}$. The function $u_k(\cdot,\cdot,p_0)$ is twice continuously differentiable with respect to x and continuously differentiable with respect to t on

$$\{(t,x) \mid t \in [0,T] \text{ and } x \in [\xi_{k-1}(t,p_0), \xi_k(t,p_0)]\}$$

for all $k \in \{1,\ldots,K+1\}$. Furthermore, $u_k(t,x,p_0)$ is continuously differentiable with respect to p on in a neighborhood of p_0 for all $t \in [0,T]$, $x \in [\xi_{k-1}(t,p_0), \xi_k(t,p_0)]$ and $\mathrm{d}_p u_k(t,\cdot,p_0)$ is continuously differentiable with respect to x on $[\xi_{k-1}(t,p_0), \xi_k(t,p_0)]$ for all $t \in [0,T]$, both for all $k \in \{1,\ldots,K+1\}$.

The entropy condition (see Section 3.1.3) in the context of the above assumption implies that

$$f'(u_k(t,\xi_k(t,p_0),p_0)) > \partial_t \xi_k(t,p_0) > f'(u_{k+1}(t,\xi_k(t,p_0),p_0))$$

for all $k \in \{1,\ldots,K\}$, $t \in [0,T]$.

Assumption 6 also implies that shocks do not run into each other since the shock locations maintain their order. Furthermore, since there is a finite number of shocks that is strictly ordered the shock locations are isolated.

Corollary 22. With $K < \infty$ and $\xi_1(t,p_0) < \cdots < \xi_K(t,p_0)$ we have

$$\xi_{k+1}(t,p_0) - \xi_k(t,p_0) \geq C_i$$

for some $C_i > 0$ for all $k \in \{1,\ldots,K-1\}$, $t \in [0,T]$.

Example 25 (Riemann)**.** We consider the parametric Riemann initial value problem from Example 6. The weak solution rewritten in the structure of Assumption 6 is

$$u(t,x,p) = \begin{cases} u_1(t,x,p) & \text{if } x \in (-\infty, \xi_1(t,p)) \\ u_2(t,x,p) & \text{if } x \in (\xi_1(t,p), \infty) \end{cases}$$

with

$$u_1(t,x,p) = 1+p, \qquad u_2(t,x,p) = 0$$

and

$$\xi_1(t,p) = \frac{1}{2}(1+p)t$$

for all p in a neighborhood of $p = p_0 = 0$.

The shock location initial value condition $\xi_0(p) = 0$ does not depend on p.

Both $u_1(\cdot,\cdot,p_0)$ and $u_2(\cdot,\cdot,p_0)$ are continuously differentiable with respect to x and t on $[0,T]\times\Omega$ and $u_1(t,x,p)$ and $u_2(t,x,p)$ are continuously differentiable with respect to p on D for all $t \in [0,T]$, $x \in \Omega$.

The Riemann example satisifies Assumption 6 except for the fact that its solution does not have compact support.

Example 26 (Ramp)**.** We consider the ramp initial value problem from Example 7. The weak solution rewritten in the structure of Assumption 6 is

$$u(t,x,p) = \begin{cases} u_1(t,x,p) & \text{if } x \in (-\infty, \xi_1(t,p)) \\ u_2(t,x,p) & \text{if } x \in (\xi_1(t,p), \infty) \end{cases}$$

with

$$u_1(t,x,p) = \max\Big\{0, \frac{1+p}{1+(1+p)t}x\Big\}, \qquad u_2(t,x,p) = 0$$

and

$$\xi_1(t,p) = \sqrt{1+(1+p)t}$$

for all p in a neighborhood of $p = p_0 = 0$.

The shock location initial value condition $\xi_0(p) = 1$ does not depend on p.

The function $u_2(\cdot,\cdot,p_0)$ is continuously differentiable with respect to x and t on $[0,T]\times\Omega$. But the function $u_1(\cdot,\cdot,p_0)$ is Lipschitz continuously nonsmooth and only continuously differentiable with respect to x and t on $[0,T]\times\big(\Omega\setminus\{0\}\big)$, i.e. the function u_1 does not satisfy Assumption 6.

Both $u_1(t,x,p)$ and $u_2(t,x,p)$ are continuously differentiable with respect to p on D for all $t \in [0,T]$ and for $x \in \Omega\setminus\{0\}$.

The function u_1 is not differentiable at $x = 0$ and, thus, the ramp example does not satisfy the solution structure assumption in a strict sense. However, all the derivations regarding sensitivity analysis in this chapter still apply to the ramp example as though u_1 were smooth because in practice we only require the smoothness assumptions in a neighborhood of the shock location.

5.2 Tangent Sensitivity

5.2.1 Shock Location

We now consider the sensitivity analysis of shock locations.

Proposition 23. The shock location $\xi_k(t,p_0)$ is continuously differentiable with respect to p at $p=p_0$ for all $t\in[0,T], k\in\{1,\dots,K\}$.

Proof. By the Rankine-Hugoniot condition (see Theorem 8) the shock location $\xi_k(t,p_0)$ satisfies the ordinary differential equation

$$\partial_t\xi_k(t,p_0)=\frac{f(u_k(t,\xi_k(t,p_0),p_0))-f(u_{k+1}(t,\xi_k(t,p_0),p_0))}{u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0)}$$

for all $t\in[0,T]$ with initial value $\xi_k(0,p_0)=(\xi_k)_0(p_0)$.

By Assumption 6 we have $u_k(t,\xi_k(t,p_0),p_0)$ and $u_{k+1}(t,\xi_k(t,p_0),p_0)$ continuously differentiable with respect to $\xi_k(t,p_0)$ in a neighborhood of $\xi_k(t,p_0)$ for all $t\in[0,T]$. By Assumption 4 the flux $f(u)$ is continuously differentiable with respect to u. Therefore, the right-hand side of the differential equation is continuously differentiable with respect to $\xi_k(t,p_0)$ in a neighborhood of $\xi_k(t,p_0)$ for all $t\in[0,T]$. By Lemma 5 the right-hand side is, hence, also locally Lipschitz continuous with respect to $\xi_k(t,p_0)$. Furthermore, also by Assumption 6 we have $u_k(t,\xi_k(t,p_0),p_0)$ and $u_{k+1}(t,\xi_k(t,p_0),p_0)$ continuous in t and, hence, the right-hand side of the differential equation is continuous in time. By the Picard-Lindelöf theorem [CL55, Theorem 3.1] the solution of the ordinary differential equation is locally unique and continuously differentiable with respect to time, i.e. $\xi_k(\cdot,p_0)\in C^1([0,T])$.

By Assumption 6 we have $u_k(t,\xi_k(t,p_0),p_0)$ and $u_{k+1}(t,\xi_k(t,p_0),p_0)$ continuously differentiable with respect to p in a neighborhood of $\xi_k(t,p_0)$. Furthermore, the initial values $\xi_k(0,p_0)$ are continuously differentiable with respect to p on D.

We consider the differential equation initial value problem as a parametric operator equation solved for $\xi_k(\cdot,p_0)$. By implicit differentiation (see Theorem 4) and the chain rule we have

$$\begin{aligned}
\dot{\xi}_k(0,p_0)&=\mathrm{d}_p(\xi_k)_0(p_0)\dot{p} && (5.2.1)\\
\partial_t\dot{\xi}_k(t,p_0)&=\frac{f'(u_k(t,\xi_k(t,p_0),p_0))\mathrm{d}_pu_k(t,\xi_k(t,p_0),p_0)\dot{p}}{u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0)} && (5.2.2)\\
&\quad-\frac{f'(u_{k+1}(t,\xi_k(t,p_0),p_0))\mathrm{d}_pu_{k+1}(t,\xi_k(t,p_0),p_0)\dot{p}}{u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0)} && (5.2.3)\\
&\quad-\frac{f(u_k(t,\xi_k(t,p_0),p_0))-f(u_{k+1}(t,\xi_k(t,p_0),p_0))}{\big(u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0)\big)^2} && (5.2.4)\\
&\quad\cdot\big(\mathrm{d}_pu_k(t,\xi_k(t,p_0),p_0)\dot{p}-\mathrm{d}_pu_{k+1}(t,\xi_k(t,p_0),p_0)\dot{p}\big) && (5.2.5)\\
&=\frac{f'(u_k(t,\xi_k(t,p_0),p_0))\big(\partial_xu_k(t,\xi_k(t,p_0),p_0)\dot{\xi}_k(t,p_0)+\mathrm{d}_pu_k(t,\xi_k(t,p_0),p_0)\dot{p}\big)}{u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0)} && (5.2.6)\\
&\quad-\frac{f'(u_{k+1}(t,\xi_k(t,p_0),p_0))\big(\partial_xu_{k+1}(t,\xi_k(t,p_0),p_0)\dot{\xi}_k(t,p_0)+\mathrm{d}_pu_{k+1}(t,\xi_k(t,p_0),p_0)\dot{p}\big)}{u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0)} && (5.2.7)
\end{aligned}$$

$$
\begin{aligned}
&-\frac{f(u_k(t,\xi_k(t,p_0),p_0)) - f(u_{k+1}(t,\xi_k(t,p_0),p_0))}{\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big)^2} && (5.2.8)\\
&\cdot \big(\partial_x u_k(t,\xi_k(t,p_0),p_0)\dot{\xi}_k(t,p_0) + \mathrm{d}_p u_k(t,\xi_k(t,p_0),p_0)\dot{p} && (5.2.9)\\
&\qquad - \partial_x u_{k+1}(t,\xi_k(t,p_0),p_0)\dot{\xi}_k(t,p_0) - \mathrm{d}_p u_{k+1}(t,\xi_k(t,p_0),p_0)\dot{p}\big) && (5.2.10)
\end{aligned}
$$

where $\dot{\xi}(t,p_0) \equiv \mathrm{d}_p\xi(t,p_0)\dot{p}$ for some direction $\dot{p} \in \mathbb{R}^d$

By Assumption 6 both $u_k(t,\xi_k(t,p_0),p_0)$ and $u_{k+1}(t,\xi_k(t,p_0),p_0)$ are continuous in time. Since by Assumption 4 the first derivative of the flux $f'(u)$ is continously differentiable and, hence, Lipschitz continuous with respect to u we have $f'(u_k(t,\xi_k(t,p_0),p_0))$ and $f'(u_{k+1}(t,\xi_k(t,p_0),p_0))$ are continuous in time. As above, $u_k(t,\xi_k(t,p_0),p_0)$ and $u_{k+1}(t,\xi_k(t,p_0),p_0)$ are also continuously differentiable with respect to $\xi_k(t,p_0)$ and p respectively. Overall, the right-hand side of the sensitivity differential equation is a Lipschitz continuous function with respect to $\dot{\xi}_k(t,p_0)$ and continuous in time. Again, by Picard-Lindelöf theorem the solution $\dot{\xi}_k$ of the sensitivity differential equation is continously differentiable with respect to time and locally unique.

The elements of the gradient vector of the shock location with respect to p at $p = p_0$ are given by the solution $\dot{\xi}_k$ of the above differential equation with $\dot{p}$ taking the values of the standard basis. This concludes the proof. □

The numerical approximation of the shock location sensitivity is the main goal of this thesis. As already shown in Section 1.4.5 the discrete shock location sensitivity analysis is not straight forward because the piecewise evaluation of the solution required for the Rankine-Hugoniot ODE right-hand side is challenging in the presence of numerical viscosity.

Example 27 (Ramp)**.** We consider the ramp initial value problem from Example 26 since this example has a nonlinear dependence of the shock location on the parameter p. By differentiation of ξ_1 we have

$$
\mathrm{d}_p\xi_1(t,p) = \frac{1}{2}\frac{1}{\sqrt{1+(1+p)t}}t\ .
$$

Figure 5.2a shows the shock location trajectory over time for $p = 0$ and for a perturbuation $p = 0 + \dot{p} = 1/2$ in red. The figure shows the linear approximation of the perturbed shock location based on the derivative with respect to p in blue. We observe that the approximation is good for small values of time t and gets worse for larger values of time in a similar way to how the perturbation differs from the original shock location.

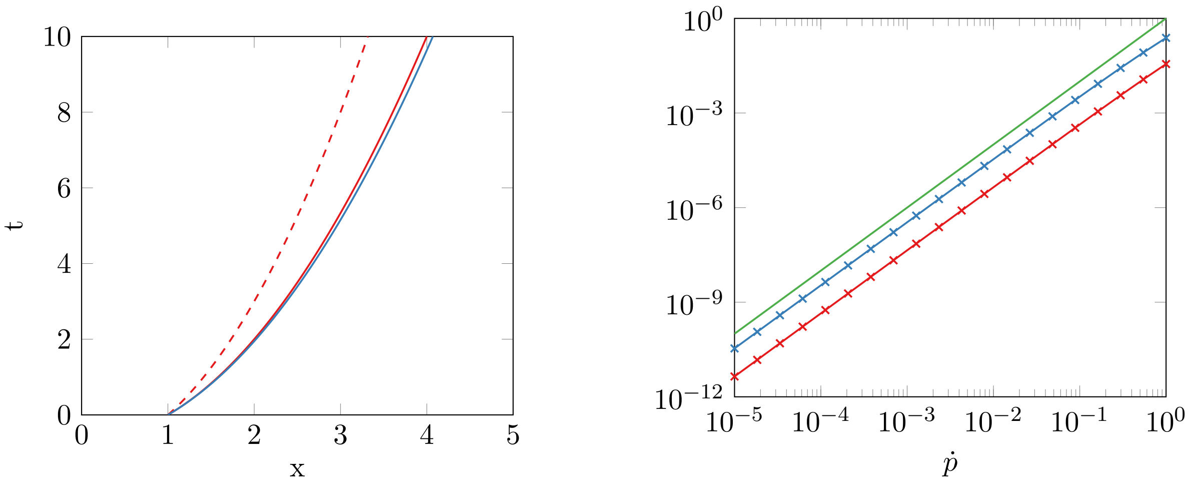
Figure 5.2b shows that the error of the linear approximation for $t = 10$ (blue) is overall larger than for $t = 1$ (red). But the empirical convergence we observe is still quadratic as we would expect from the Taylor expansion based on Proposition 23.

5.2.2 Weak Solution

Based on Assumption 6 and Proposition 23 we now derive the analytic tangent sensitivity of the weak solution in the sense of distributions.

Proposition 24. The distributional tangent sensitivity

$$
\dot{u}(t,x,p_0) \equiv \mathrm{d}_p u(t,x,p_0)\dot{p}
$$



(a) $\xi_1(t,0)$ (dashed, red), $\xi_1(t,\frac{1}{2})$ (solid, red), $\xi_1(t,0)+\frac{1}{2}\mathrm{d}_p\xi_1(t,0)$ (solid, blue)

(b) $\dot{p}^2$ (green), $|\xi_1(t,\dot{p})-\xi_1(t,0)-\mathrm{d}_p\xi_1(t,0)\dot{p}|$ at $t=1$ (red), at $t=10$ (blue)

Figure 5.2: First-order approximation of shock sensitivity over time

of the weak solution u in the direction $\dot{p}\in\mathbb{R}^d$ is

$$\dot{u}(t,x,p_0)=\sum_{k=1}^{K}\delta(x-\xi_k(t,p_0))\dot{\xi}_k(t,p_0)\big(u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0)\big)$$
$$+\begin{cases}\dot{u}_1(t,x,p_0) & \text{if } x\in(\xi_0,\xi_1(t,p_0))\\ \dot{u}_2(t,x,p_0) & \text{if } x\in(\xi_1(t,p_0),\xi_2(t,p_0))\\ \quad\vdots & \quad\vdots\\ \dot{u}_{K+1}(t,x,p_0) & \text{if } x\in(\xi_K(t,p_0),\xi_{K+1})\end{cases}$$

where $\dot{u}_k(t,x,p_0)\equiv\mathrm{d}_pu_k(t,x,p_0)\dot{p}$ and $\dot{\xi}_k(t,p_0)\equiv\mathrm{d}_p\xi_k(t,p_0)\dot{p}$ for all $k\in\{1,\dots,K+1\}$, $t\in[0,T]$, $x\in\Omega$.

Proof. We write the weak solution in terms of Heaviside step functions

$$u(t,x,p_0)=\sum_{k=1}^{K+1}H(x-\xi_{k-1}(t,p_0))H(\xi_k(t,p_0)-x)u_k(t,x,p_0)\ .$$

By weak differentiation and the chain rule we get

$$\begin{aligned}\mathrm{d}_pu(t,x,p_0)&=\mathrm{d}_p\sum_{k=1}^{K+1}H(x-\xi_{k-1}(t,p_0))H(\xi_k(t,p_0)-x)u_k(t,x,p_0)\\ &=\sum_{k=1}^{K+1}\Big(\mathrm{d}_pH(x-\xi_{k-1}(t,p_0))\Big)H(\xi_k(t,p_0)-x)u_k(t,x,p_0)\\ &\quad+\sum_{k=1}^{K+1}H(x-\xi_{k-1}(t,p_0))\Big(\mathrm{d}_pH(\xi_k(t,p_0)-x)\Big)u_k(t,x,p_0)\\ &\quad+\sum_{k=1}^{K+1}H(x-\xi_{k-1}(t,p_0))H(\xi_k(t,p_0)-x)\Big(\mathrm{d}_pu_k(t,x,p_0)\Big)\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K+1} \Big(-\delta(x-\xi_{k-1}(t,p_0))\mathrm{d}_p\xi_{k-1}(t,p_0)\Big) H(\xi_k(t,p_0)-x)u_k(t,x,p_0) \\
&\quad + \sum_{k=1}^{K+1} H(x-\xi_{k-1}(t,p_0))\Big(\delta(\xi_k(t,p_0)-x)\mathrm{d}_p\xi_k(t,p_0)\Big)u_k(t,x,p_0) \\
&\quad + \sum_{k=1}^{K+1} H(x-\xi_{k-1}(t,p_0))H(\xi_k(t,p_0)-x)\Big(\mathrm{d}_p u_k(t,x,p_0)\Big) \ .
\end{aligned}
$$

Since ξ_0 and ξ_{K+1} are not a function of p their derivatives are zero. Furthermore, we have $\delta(x-y) = \delta(y-x)$ for all $y \in \mathbb{R}$. We have

$$
\begin{aligned}
&\sum_{k=1}^{K+1} \Big(-\delta(x-\xi_{k-1}(t,p_0))\mathrm{d}_p\xi_{k-1}(t,p_0)\Big) H(\xi_k(t,p_0)-x)u_k(t,x,p_0) \\
&\quad + \sum_{k=1}^{K+1} H(x-\xi_{k-1}(t,p_0))\Big(\delta(\xi_k(t,p_0)-x)\mathrm{d}_p\xi_k(t,p_0)\Big)u_k(t,x,p_0) \\
&= \sum_{k=2}^{K+1} \Big(-\delta(x-\xi_{k-1}(t,p_0))\mathrm{d}_p\xi_{k-1}(t,p_0)\Big) H(\xi_k(t,p_0)-x)u_k(t,x,p_0) \\
&\quad + \sum_{k=1}^{K} H(x-\xi_{k-1}(t,p_0))\Big(\delta(\xi_k(t,p_0)-x)\mathrm{d}_p\xi_k(t,p_0)\Big)u_k(t,x,p_0) \\
&= \sum_{k=1}^{K} \Big(-\delta(x-\xi_{k}(t,p_0))\mathrm{d}_p\xi_{k}(t,p_0)\Big) H(\xi_{k+1}(t,p_0)-x)u_{k+1}(t,x,p_0) \\
&\quad + \sum_{k=1}^{K} H(x-\xi_{k-1}(t,p_0))\Big(\delta(\xi_k(t,p_0)-x)\mathrm{d}_p\xi_k(t,p_0)\Big)u_k(t,x,p_0) \\
&= \sum_{k=1}^{K} H(x-\xi_{k-1}(t,p_0))\delta(\xi_k(t,p_0)-x)\mathrm{d}_p\xi_k(t,p_0)u_k(t,x,p_0) \\
&\quad - \delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)H(\xi_{k+1}(t,p_0)-x)u_{k+1}(t,x,p_0) \\
&= \sum_{k=1}^{K} H(x-\xi_{k-1}(t,p_0))\delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)u_k(t,x,p_0) \\
&\quad - \delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)H(\xi_{k+1}(t,p_0)-x)u_{k+1}(t,x,p_0) \ .
\end{aligned}
$$

Since $\xi_{k+1} > \xi_k > \xi_{k_1}$ we have

$$H(x-\xi_{k-1}(t,p_0))\delta(x-\xi_k(t,p_0)) = \delta(x-\xi_k(t,p_0))$$

and

$$H(\xi_{k+1}(t,p_0)-x)\delta(x-\xi_k(t,p_0)) = \delta(x-\xi_k(t,p_0) \ .$$

We have

$$
\begin{aligned}
&\sum_{k=1}^{K} H(x-\xi_{k-1}(t,p_0))\delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)u_k(t,x,p_0) \\
&\quad - \delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)H(\xi_{k+1}(t,p_0)-x)u_{k+1}(t,x,p_0)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} \delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)u_k(t,x,p_0) - \delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)u_{k+1}(t,x,p_0)\\
&= \sum_{k=1}^{K} \delta(x-\xi_k(t,p_0))\mathrm{d}_p\xi_k(t,p_0)\big(u_k(t,x,p_0) - u_{k+1}(t,x,p_0)\big)\ .
\end{aligned}
$$

We get the result in the proposition if we consider that

$$
\begin{aligned}
&\sum_{k=1}^{K+1} H(x-\xi_{k-1}(t,p_0))H(\xi_k(t,p_0)-x)\Big(\mathrm{d}_p u_k(t,x,p_0)\Big)\\
&= \begin{cases} \mathrm{d}_p u_1(t,x,p_0) & \text{if } x \in (\xi_0, \xi_1(t,p_0))\\ \mathrm{d}_p u_2(t,x,p_0) & \text{if } x \in (\xi_1(t,p_0), \xi_2(t,p_0))\\ \quad\vdots & \quad\vdots \\ \mathrm{d}_p u_{K+1}(t,x,p_0) & \text{if } x \in (\xi_K(t,p_0), \xi_{K+1}) \end{cases}
\end{aligned}
$$

in the sense of $L_1(\hat{\Omega})$, i.e. ignoring the set of discontinuity locations $\bigcup_{k=1}^{K} \xi_k(t,p_0)$ that has measure zero and multiplying by $\dot{p}$. This concludes the proof. □

The last term in the tangent sensitivity consisting of a $\dot{u}_k$ component represents just the standard tangent sensitivity of each smooth piece between the shocks.

The sum of Dirac delta distribution terms describes the shift of each shock location due to the parameter perturbation. The shock sensitivity multiplied by the jump height of the shock discontinuity gives a singular impulse at the shock. Because of the Dirac delta distribution terms the tangent sensitivity of the weak solution is not a function in the classical sense and not in $L_1(\Omega)$.

Corollary 25. If u has at least one shock discontinuity then $\dot{u}(\cdot,\cdot,p_0) \notin L_1(\Omega)$.

Example 28 (Riemann)**.** We consider the parametric Riemann problem with solution structure as given in Example 25. According to Proposition 24 the tangent sensitivity of the weak solution is

$$
\dot{u}(t,x,p) = \delta(x-\xi_1(t,p))\mathrm{d}_p\xi_1(t,p)\dot{p}(1+p) + \begin{cases} \dot{p} & \text{if } x \in (-\infty, \xi_1(t,p))\\ 0 & \text{if } x \in (\xi_1(t,p), \infty) \end{cases}
$$

with shock sensitivity

$$
\mathrm{d}_p\xi_1(t,p) = \frac{1}{2}t
$$

for p in a neighborhood of $p = p_0 = 0$. The same distributional derivative for this example was already derived in Section 1.3.2.

Example 29 (Ramp)**.** We consider the ramp initial value problem from Example 26. According to Proposition 24 the tangent sensitivity of the weak solution is

$$
\begin{aligned}
\dot{u}(t,x,p) = {}&\delta(x-\xi_1(t,p))\mathrm{d}_p\xi_1(t,p)\dot{p}(u_1(t,\xi_1(t,p),p) - u_2(t,\xi_1(t,p),p))\\
&+ \begin{cases} \dot{u}_1(t,x,p) & \text{if } x \in (-\infty, \xi_1(t,p))\\ \dot{u}_2(t,x,p) & \text{if } x \in (\xi_1(t,p), \infty) \end{cases}
\end{aligned}
$$

with

$$
\dot{u}_1(t,x,p) = \mathrm{d}_p u_1(t,x,p)\dot{p} = H(x)\frac{1}{(1+(1+p)t)^2}x\dot{p}
$$

and

$$\dot{u}_2(t,x,p) = \mathrm{d}_p u_2(t,x,p)\dot{p} = 0$$

for p in a neighborhood of $p = p_0 = 0$. The shock sensitivity $\mathrm{d}_p\xi_1(t,p)\dot{p}$ is as derived in Example 27.

With

$$u_1(t,\xi_1(t,p),p) = \max\left\{0, \frac{1+p}{1+(1+p)t}\sqrt{1+(1+p)t}\right\}$$

and $u_2(t,\xi_1(t,p),p) = 0$ for $t > 0$ and p in a neighborhood of $p = p_0 = 0$ the height of the jump discontinuity is

$$u_1(t,\xi_1(t,p),p) - u_2(t,\xi_1(t,p),p) = \frac{1+p}{\sqrt{1+(1+p)t}}\,.$$

Figure 5.3a shows the weak solution to the ramp problem at time $t = 1$ for both $p = 0$ and $p = 0 + \dot{p} = 1/2$. The shock with higher jump discontinuity travels faster and, hence, is further to the right at the same point in time. The additional shock travel due to the perturbation $p = 0 + \dot{p}$ in this example is

$$\xi_1(1,\frac{1}{2}) - \xi_1(1,0) = \sqrt{\frac{5}{2}} - \sqrt{2}\,.$$

Figure 5.3b shows the difference between the perturbed and unperturbed problem with $\dot{p} = 1/2$ using a dashed red line. We see that the additional shock travel in the perturbed solution is represented by a box that has the width of the difference between shock locations and the height of the perturbed jump discontinuity. The difference $u(1,x,1/2) - u(1,x,0)$ is approximated by the tangent sensitivity $\dot{u}(1,x,0)$ with $\dot{p} = 1/2$. The blue solid line shows the measurable part of $\dot{u}(1,x,0)$.

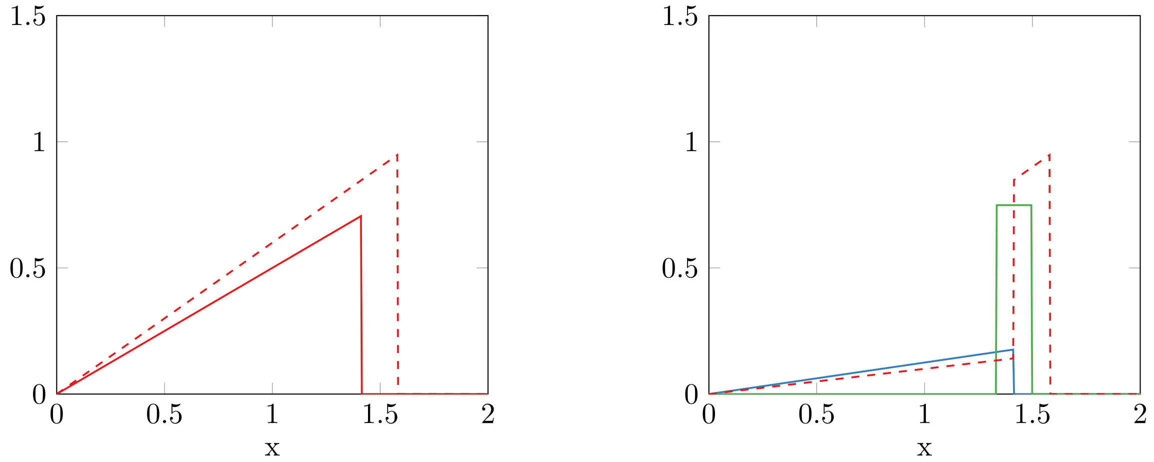
The green solid line shows an approximate visual representation of the singular Dirac delta distribution term of $\dot{u}(1,x,0)$. Since we have the defining property that $\int \delta(x)\mathrm{d}x = 1$ the Dirac delta distribution is approximated by a box with integral 1. We picked the width of the box to be the same as the shock travel due to the perturbation. We observe that when multiplied with the factors in the Dirac delta term as given in Proposition 24 the height of the solid green box approximates the height of the dashed red box for small enough values of $\dot{p}$.

5.3 Generalized Tangent Sensitivity

We now formalize the notion of the generalized tangent of Bressan and Marson [BM95b] as it applies to our assumed piecewise smooth solution structure. We derive the resulting generalized Taylor expansion that we already briefly described in Section 1.4.2.

5.3.1 Broad Tangent

After showing that the tangent sensitivity of a weak solution with at least one shock discontinuity contains a singular part we now introduce the notion of the broad solution of the tangent PDE



(a) $u(t,x,0)$ (solid), $u(t,x,\dot{p})$ (dashed) with $\dot{p} = \frac{1}{2}, t = 1$

(b) $u(t,x,\dot{p}) - u(t,x,0)$ (dashed, red), $\dot{u}_1(t,x,p)$ (solid, blue), box approximation of $\delta(x - \xi_1(t,0))\mathrm{d}_p\xi_1(t,p)\dot{p}(1+p)$ term (solid, green) with $\dot{p} = \frac{1}{2}$

Figure 5.3: Weak solution for the ramp example with pertubation of p

as it is used in Bressan and Marson [BM95b, Section 1.2]. The broad solution of the tangent equation contains only the L_1 part of the tangent sensitivity, i.e. the broad solution does not contain any singular parts due to the shift in shock location.

In the context of our assumed solution structure we use the following definition of the broad tangent.

Definition 36. The *broad tangent* sensitivity of the weak solution in the direction $\dot{p} \in \mathbb{R}^d$ is

$$\dot{u}_{\text{broad}}(t,x,p_0) \equiv \begin{cases} \dot{u}_1(t,x,p_0) & \text{if } x \in (\xi_0, \xi_1(t,p_0)) \\ \dot{u}_2(t,x,p_0) & \text{if } x \in (\xi_1(t,p_0), \xi_2(t,p_0)) \\ \quad \vdots & \quad \vdots \\ \dot{u}_{K+1}(t,x,p_0) & \text{if } x \in (\xi_K(t,p_0), \xi_{K+1}) \end{cases}$$

where $\dot{u}_k(t,x,p_0) \equiv \mathrm{d}_p u_k(t,x,p_0)\dot{p}$ for all $k \in \{1,\ldots,K+1\}$ and for all $t \in [0,T]$, $x \in \Omega$.

The broad tangent does not satisfy a first-order Taylor expansion error bound in the sense of the L_1 norm in the presence of shocks, i.e. we generally do not have

$$\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\text{broad}}(t,\cdot,p_0)\|_{L_1(\Omega)} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$.

The broad tangent satisfies the tangent PDE almost everywhere (everywhere except on the shock location curves) [BM95b, Section 1.2].

Example 30 (Riemann)**.** We consider the parametric Riemann problem from Example 25. According to Definition 36 the broad tangent sensitivity is

$$\dot{u}_{\text{broad}}(t,x,p) = \begin{cases} \dot{p} & \text{if } x \in (-\infty, \xi_1(t,p)) \\ 0 & \text{if } x \in (\xi_1(t,p), \infty) \end{cases}$$

with ξ_1 as before and for p in a neighborhood of $p = p_0 = 0$

Regarding the L_1 error bound for this example assuming $\dot{p} > 0$ we have

$$\begin{aligned}\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\text{broad}}(t,\cdot,p_0)\|_{L_1(\Omega)} &= \|(1+p+\dot{p}) - (1+p) - \dot{p}\|_{L_1((-\infty,\xi_1(t,p_0)))}\\ &\quad + \|(1+p+\dot{p}) - 0 - 0\|_{L_1((\xi_1(t,p),\xi_1(t,p_0+\dot{p})))}\\ &\quad + \|0-0-0\|_{L_1((\xi_1(t,p_0+\dot{p}),\infty))}\\ &= O(\|\dot{p}\|)\end{aligned}$$

as $\dot{p} \to 0$.

Example 31 (Ramp)**.** We consider the ramp initial value problem from Example 26. According to Definition 36 the broad tangent sensitivity is

$$\dot{u}_{\text{broad}}(t,x,p) = \begin{cases} H(x)\frac{1}{(1+(1+p)t)^2}x\dot{p} & \text{if } x \in (-\infty,\xi_1(t,p))\\ 0 & \text{if } x \in (\xi_1(t,p),\infty)\end{cases}$$

with ξ_1 as before and for p in a neighborhood of $p = p_0 = 0$. Figure 5.3b shows the broad tangent in blue.

Regarding the tangent PDE (see Proposition 11) for example left of the shock location curve ξ_1 we have

$$\begin{aligned}&\partial_t\dot{u}_{\text{broad}}(t,x,p) + \partial_x f'(u(t,x,p))\dot{u}_{\text{broad}}(t,x,p) + f'(u(t,x,p))\partial_x\dot{u}_{\text{broad}}(t,x,p)\\ &\quad = \partial_t\frac{1}{(1+(1+p)t)^2}x\dot{p} + \partial_x f'\Big(\frac{1+p}{1+(1+p)t}x\Big)\frac{1}{(1+(1+p)t)^2}x\dot{p}\\ &\qquad + f'\Big(\frac{1+p}{1+(1+p)t}x\Big)\frac{1}{(1+(1+p)t)^2}\dot{p}\\ &\quad = -2\frac{1+p}{(1+(1+p)t)^3}x\dot{p} + \frac{1+p}{1+(1+p)t}\frac{1}{(1+(1+p)t)^2}x\dot{p}\\ &\qquad + \frac{1+p}{1+(1+p)t}x\frac{1}{(1+(1+p)t)^2}\dot{p}\\ &\quad = -2\frac{1+p}{(1+(1+p)t)^3}x\dot{p} + 2\frac{1+p}{(1+(1+p)t)^3}x\dot{p} = 0\ .\end{aligned}$$

5.3.2 Nonlinear Generalized Tangent

We now define the nonlinear generalized tangent variation for the assumed solution structure based on the analytic generalized differential calculus for conservation laws with discontinuous solution from Bressan and Marson [BM95b].

Definition 37. The (nonlinear) *generalized tangent* sensitivity of the weak solution in the direction $\dot{p} \in \mathbb{R}^d$ is

$$\begin{aligned}\dot{u}_{\text{gen}}(t,x,p_0) \equiv &\sum_{\substack{k=1\\ \dot{\xi}_k(t,p_0)>0}}^{K} \chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big)\\ &- \sum_{\substack{k=1\\ \dot{\xi}_k(t,p_0)<0}}^{K} \chi_{[\xi_k(t,p_0)+\dot{\xi}_k(t,p_0),\xi_k(t,p_0)]}(x)\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big)\end{aligned}$$

$$+ \dot{u}_{\mathrm{broad}}(t, x, p_0)$$

where $\dot{\xi}_k(t, x, p_0) \equiv \mathrm{d}_p \xi_k(t, x, p_0)\dot{p}$ for all $k \in \{1, \ldots, K+1\}$ and $\dot{u}_{\mathrm{broad}}$ is the broad tangent sensitivity.

The generalized tangent only uses the tangent sensitivities of the piecewise smooth parts of the solution and the tangent sensitivities of the shock locations. Essentially, the generalized tangent replaces the Dirac delta distributions in the tangent sensitivity by characteristic functions that approximate the additional shock travel in a way that is integrable. The generalized tangent is a tangent sensitivity in the sense that it is a mathematical object obtained by linearization. But as opposed to the standard tangent sensitivity the map that the generalized tangent gives as a function of $\dot{p}$ is not linear (c.f. Example 2).

In contrast to the broad tangent the nonlinear generalized tangent satisfies a first-order Taylor expansion error bound in the sense of the L_1 norm.

Proposition 26.

$$\|u(t, \cdot, p_0 + \dot{p}) - u(t, \cdot, p_0) - \dot{u}_{\mathrm{gen}}(t, \cdot, p_0)\|_{L_1(\hat{\Omega})} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$ for all $t \in [0, T]$.

Proof. We split the domain $\hat{\Omega}$ in such a way that we have regions S_k for all $k \in \{1, \ldots, K\}$ belonging to each shock discontinuity respectively. By Corollary 22 the shocks are isolated and we have

$$\xi_{k+1}(t, p_0) - \xi_k(t, p_0) \geq C_i$$

for some $C_i > 0$. Let

$$S_k = (\xi_k(t, p_0) - \frac{1}{2}C_i, \xi_k(t, p_0) + \frac{1}{2}C_i)$$

for all $k \in \{1, \ldots, K\}$. By Proposition 23 the shock location is continuously differentiable in p and therefore Lipschitz continuous in a neighborhood of p_0 by Lemma 5. We have

$$\|\xi_k(t, p_0 + \dot{p}) - \xi_k(t, p_0)\| = O(\|\dot{p}\|)$$

as $\dot{p} \to 0$ and hence

$$\xi_k(t, p_0 + \dot{p}) \in S_k$$

for all $\dot{p} \leq \delta$ for some $\delta > 0$.

We first consider the difference between the finite difference perturbation and the nonlinear tangent outside of the shock regions. By Assumption 6 we have

$$u(t, x, p_0 + \dot{p}) - u(t, x, p_0) = u_k(t, x, p_0 + \dot{p}) - u_k(t, x, p_0)$$

for some $k \in \{1, \ldots, K+1\}$ for all $x \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} S_k$ because we have $x < \xi_k(t, p_0)$ and $x > \xi_k(t, p_0 + \dot{p})$ for all $k \in \{1, \ldots, K+1\}$. When we are outside of the neighborhood of the shock the shock perturbation does not affect the perturbed solution compared to the unperturbed solution. Similarly, by definition of the nonlinear generalized tangent we have

$$\dot{u}_{\mathrm{gen}}(t, x, p_0) = \dot{u}_{\mathrm{broad}}(t, x, p_0) = \dot{u}_k(t, x, p_0)$$

for some $k \in \{1, \dots, K+1\}$ for all $x \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} S_k$ because we have $x < \xi_k(t, x, p)$ and $x > \xi_k(t, x, p + \dot{p})$ for all $k \in \{1, \dots, K+1\}$. By Assumption 6 we have that $u_k(t, x, p_0)$ is continuously differentiable with respect to p in a neighborhood of p_0 for all $x \in \hat{\Omega}$, $t \in [0, T]$. By Taylor expansion we have

$$\begin{aligned} u_k(t, x, p_0 + \dot{p}) &= u_k(t, x, p_0) + \mathrm{d}_p u_k(t, x, p_0)\dot{p} + O(\|\dot{p}\|^2) \\ &= u_k(t, x, p_0) + \dot{u}_k(t, x, p_0) + O(\|\dot{p}\|^2) \end{aligned}$$

as $\dot{p} \to 0$ for all $x \in \hat{\Omega}$, $t \in [0, T]$. Consequently, by the definition of the L_1 norm we have

$$\begin{aligned} &\|u(t, \cdot, p_0 + \dot{p}) - u(t, \cdot, p_0) - \dot{u}_{\mathrm{gen}}(t, \cdot, p_0)\|_{L_1(\hat{\Omega} \setminus \bigcup_{k=1}^{K} S_k)} \\ &= \int_{\hat{\Omega} \setminus \bigcup_{k=1}^{K} S_k} u(t, x, p_0 + \dot{p}) - u(t, x, p_0) - \dot{u}_{\mathrm{gen}}(t, x, p_0) \mathrm{d}x \\ &= \int_{\hat{\Omega} \setminus \bigcup_{k=1}^{K} S_k} u_k(t, x, p_0 + \dot{p}) - u_k(t, x, p_0) - \dot{u}_k(t, x, p_0) \mathrm{d}x \\ &= \int_{\hat{\Omega} \setminus \bigcup_{k=1}^{K} S_k} O(\|\dot{p}\|^2) \mathrm{d}x \\ &= O(\|\dot{p}\|^2) \int_{\hat{\Omega} \setminus \bigcup_{k=1}^{K} S_k} \mathrm{d}x \\ &\leq O(\|\dot{p}\|^2)(R - L) = O(\|\dot{p}\|^2) \end{aligned}$$

for some $k \in \{1, \dots, K+1\}$ as $\dot{p} \to 0$.

We now consider the difference between the finite difference perturbation and the nonlinear generalized tangent in the shock regions. For each shock region S_k we consider the following cases as $\dot{p} \to 0$:

(i) $\xi_k(t, p_0 + \dot{p}) = \xi_k(t, p_0)$

(ii) $\xi_k(t, p_0 + \dot{p}) > \xi_k(t, p_0)$

(iii) $\xi_k(t, p_0 + \dot{p}) < \xi_k(t, p_0)$

For the first case we consider a split of the set S_k at the shock location. Left of the shock location we have

$$u(t, x, p_0 + \dot{p}) - u(t, x, p_0) = u_k(t, x, p_0 + \dot{p}) - u_k(t, x, p_0)$$

for all $x \in \{y \in S_k : y < \xi_k(t, p_0)\}$. By Taylor expansion we have

$$\xi_k(t, p_0 + \dot{p}) = \xi_k(t, p_0) + \mathrm{d}_p \xi_k(t, p_0)\dot{p} + O(\|\dot{p}\|^2)$$

and since by case assumption

$$\xi_k(t, p_0 + \dot{p}) - \xi_k(t, p_0) = 0$$

we have

$$\dot{\xi}_k(t, p_0) = \mathrm{d}_p \xi_k(t, p_0)\dot{p} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$. If $\dot{\xi}_k(t, p_0) \geq 0$ by definition we have

$$\dot{u}_{\mathrm{gen}}(t, x, p_0) = \dot{u}_k(t, x, p_0)$$

for all $x \in \{y \in S_k : y < \xi_k(t, p_0)\}$ and if $\dot{\xi}_k(t, p_0) < 0$ we have

$$\dot{u}_{\text{gen}}(t, x, p_0) = \dot{u}_k(t, x, p_0) - \chi_{[\xi_k(t,p_0)-\dot{\xi}_k(t,p_0),\xi_k(t,p_0)]}(x)(u_k(t, \xi_k(t, p_0), p_0) - u_{k+1}(t, \xi_k(t, p_0), p_0))$$

for all $x \in \{y \in S_k : y < \xi_k(t, p_0)\}$. Consequently, by definition of the L_1 norm if $\dot{\xi}_k(t, p_0) \geq 0$ we have

$$\begin{aligned}
&\|u(t, \cdot, p_0 + \dot{p}) - u(t, \cdot, p_0) - \dot{u}_{\text{gen}}(t, \cdot, p_0)\|_{L_1(\{y \in S_k : y < \xi_k(t,p_0)\})} \\
&\quad = \int_{\{y \in S_k : y < \xi_k(t,p_0)\}} u(t, x, p_0 + \dot{p}) - u(t, x, p_0) - \dot{u}_{\text{gen}}(t, x, p_0)\mathrm{d}x \\
&\quad = \int_{\{y \in S_k : y < \xi_k(t,p_0)\}} u_k(t, x, p_0 + \dot{p}) - u_k(t, x, p_0) - \dot{u}_k(t, x, p_0)\mathrm{d}x \\
&\quad = \int_{\{y \in S_k : y < \xi_k(t,p_0)\}} O(\|\dot{p}\|^2)\mathrm{d}x = O(\|\dot{p}\|^2)
\end{aligned}$$

as $\dot{p} \to 0$ and if $\dot{\xi}_k(t, p_0) < 0$ we have

$$\begin{aligned}
&\|u(t, \cdot, p_0 + \dot{p}) - u(t, \cdot, p_0) - \dot{u}_{\text{gen}}(t, \cdot, p_0)\|_{L_1(\{y \in S_k : y < \xi_k(t,p_0)\})} \\
&\quad = \int_{\{y \in S_k : y < \xi_k(t,p_0)\}} u(t, x, p_0 + \dot{p}) - u(t, x, p_0) - \dot{u}_{\text{gen}}(t, x, p_0)\mathrm{d}x \\
&\quad = \int_{\{y \in S_k : y < \xi_k(t,p_0)\}} u_k(t, x, p_0 + \dot{p}) - u_k(t, x, p_0)) - \dot{u}_k(t, x, p_0)\mathrm{d}x \\
&\qquad + \int_{\{y \in S_k : y < \xi_k(t,p_0)\}} \chi_{[\xi_k(t,p_0)-\dot{\xi}_k(t,p_0),\xi_k(t,p_0)]}(x)(u_k(t, \xi_k(t, p_0), p_0) - u_{k+1}(t, \xi_k(t, p_0), p_0))\mathrm{d}x \\
&\quad = O(\|\dot{p}\|^2) + \dot{\xi}_k(t, p_0) \cdot (u_k(t, \xi_k(t, p_0), p_0) - u_{k+1}(t, \xi_k(t, p_0), p_0)) \\
&\quad = O(\|\dot{p}\|^2) + O(\|\dot{p}\|^2) = O(\|\dot{p}\|^2)
\end{aligned}$$

as $\dot{p} \to 0$. Analogously, right of the shock location we have

$$u(t, x, p_0 + \dot{p}) - u(t, x, p_0) = u_{k+1}(t, x, p_0 + \dot{p}) - u_{k+1}(t, x, p_0)$$

for all $x \in \{y \in S_k : y > \xi_k(t, p_0)\}$. For the tangent variation we have

$$\dot{u}_{\text{gen}}(t, x, p_0) = \dot{u}_{k+1}(t, x, p_0)$$

if $\dot{\xi}_k(t, p_0) \leq 0$ and

$$\begin{aligned}
\dot{u}_{\text{gen}}(t, x, p_0) &= \dot{u}_{k+1}(t, x, p_0) \\
&\qquad + \chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)(u_k(t, \xi_k(t, p_0), p_0) - u_{k+1}(t, \xi_k(t, p_0), p_0))
\end{aligned}$$

if $\dot{\xi}_k(t, p_0) > 0$ for all $x \in \{y \in S_k : y > \xi_k(t, p_0)\}$. In both cases by analogous steps as above we have

$$\|u(t, \cdot, p_0 + \dot{p}) - u(t, \cdot, p_0) - \dot{u}_{\text{gen}}(t, \cdot, p_0)\|_{L_1(\{y \in S_k : y > \xi_k(t,p_0)\})} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$. Combining the two sets left and right of the shock location we get

$$\begin{aligned}
&\|u(t, \cdot, p_0 + \dot{p}) - u(t, \cdot, p_0) - \dot{u}_{\text{gen}}(t, \cdot, p_0)\|_{L_1(S_k)} \\
&\quad = \|u(t, \cdot, p_0 + \dot{p}) - u(t, \cdot, p_0) - \dot{u}_{\text{gen}}(t, \cdot, p_0)\|_{L_1(\{y \in S_k : y < \xi_k(t,p_0)\})}
\end{aligned}$$

$$+\|u(t,\cdot,p_0+\dot{p})-u(t,\cdot,p_0)-\nu(t,\cdot,p_0)\|_{L_1(\{y\in S_k:y>\xi_k(t,p_0)\})}=O(\|\dot{p}\|^2)$$

as $\dot{p}\to 0$ for the first case.

For the second case by the case assumption we have

$$\dot{\xi}_k(t,p_0)=\mathrm{d}_p\xi_k(t,p_0)\dot{p}>0$$

as $\dot{p}\to 0$ so to the left of the shock we have

$$\|u(t,\cdot,p_0+\dot{p})-u(t,\cdot,p_0)-\dot{u}_{\text{gen}}(t,\cdot,p_0)\|_{L_1(\{y\in S_k:y<\xi_k(t,p_0)\})}=O(\|\dot{p}\|^2)$$

as $\dot{p}\to 0$ by analogous steps as in the first case.

To the right of the shock we differentiate between the subcases

$$\dot{\xi}_k(t,p_0)\geq\xi_k(t,p_0+\dot{p})-\xi_k(t,p_0)$$

and

$$\dot{\xi}_k(t,p_0)<\xi_k(t,p_0+\dot{p})-\xi_k(t,p_0)\ .$$

In the first subcase we have

$$\begin{aligned}
&\|u(t,\cdot,p_0+\dot{p})-u(t,\cdot,p_0)-\dot{u}_{\text{gen}}(t,\cdot,p_0)\|_{L_1(\{y\in S_k:y>\xi_k(t,p_0)\})}\\
&\quad=\int_{S_k^1}u(t,x,p_0+\dot{p})-u(t,x,p_0)-\dot{u}_{\text{gen}}(t,x,p_0)\mathrm{d}x\\
&\qquad+\int_{S_k^2}u(t,x,p_0+\dot{p})-u(t,x,p_0)-\dot{u}_{\text{gen}}(t,x,p_0)\mathrm{d}x\\
&\qquad+\int_{S_k^3}u(t,x,p_0+\dot{p})-u(t,x,p_0)-\dot{u}_{\text{gen}}(t,x,p_0)\mathrm{d}x\\
&\quad=\int_{S_k^1}u_k(t,x,p_0+\dot{p})-u_{k+1}(t,x,p_0)-\dot{u}_{k+1}(t,x,p_0)\mathrm{d}x\\
&\qquad-\int_{S_k^1}\chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)(u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0))\mathrm{d}x\\
&\qquad+\int_{S_k^2}u_{k+1}(t,x,p_0+\dot{p})-u_{k+1}(t,x,p_0)-\dot{u}_{k+1}(t,x,p_0)\mathrm{d}x\\
&\qquad-\int_{S_k^2}\chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)(u_k(t,\xi_k(t,p_0),p_0)-u_{k+1}(t,\xi_k(t,p_0),p_0))\mathrm{d}x\\
&\qquad+\int_{S_k^3}u_{k+1}(t,x,p_0+\dot{p})-u_{k+1}(t,x,p_0)-\dot{u}_{k+1}(t,x,p_0)\mathrm{d}x
\end{aligned}$$

with

$$\begin{aligned}
S_k^1&\equiv\{y\in S_k:y>\xi_k(t,p_0),y<\xi_k(t,p_0+\dot{p})\}\\
S_k^2&\equiv\{y\in S_k:y>\xi_k(t,p_0+\dot{p}),y<\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)\}\\
S_k^3&\equiv\{y\in S_k:y>\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)\}\ .
\end{aligned}$$

We analyze the terms separately. By Taylor expansion we have

$$\dot{u}_{k+1}(t,x,p_0)=u_{k+1}(t,x,p_0+\dot{p})-u_{k+1}(t,x,p_0)+O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$. By Assumption 6 we have u_k and u_{k+1} continously differentiable with respect to x and hence Lipschitz continuous

$$u_k(t,x,p_0+\dot{p}) = u_k(t,\xi_k(t,p_0),p_0+\dot{p}) + O(\|x-\xi_k(t,p_0)\|)$$

and

$$u_{k+1}(t,x,p_0+\dot{p}) = u_{k+1}(t,\xi_k(t,p_0),p_0+\dot{p}) + O(\|x-\xi_k(t,p_0)\|)$$

as $x \to \xi_k(t,p_0)$. On the set $\{y \in S_k : y > \xi_k(t,p_0), y < \xi_k(t,p_0+\dot{p})\}$ we have

$$\|x-\xi_k(t,p_0)\| \leq \|\xi_k(t,p_0+\dot{p}) - \xi_k(t,p_0)\| = O(\|\dot{p}\|)$$

as $\dot{p} \to 0$. The first term, therefore, evaluates to

$$\begin{aligned}
&\int_{S_k^1} u_k(t,x,p_0+\dot{p}) - u_{k+1}(t,x,p_0) - \dot{u}_{k+1}(t,x,p_0)\mathrm{d}x \\
&\quad = \int_{S_k^1} u_k(t,x,p_0+\dot{p}) - u_{k+1}(t,x,p_0) - u_{k+1}(t,x,p_0+\dot{p}) + u_{k+1}(t,x,p_0) + O(\|\dot{p}\|^2)\mathrm{d}x \\
&\quad = \int_{S_k^1} u_k(t,x,p_0+\dot{p}) - u_{k+1}(t,x,p_0+\dot{p}) + O(\|\dot{p}\|^2)\mathrm{d}x \\
&\quad = \int_{S_k^1} u_k(t,\xi_k(t,p_0),p_0+\dot{p}) - u_{k+1}(t,\xi_k(t,p_0),p_0+\dot{p}) + O(\|\dot{p}\|)\mathrm{d}x \\
&\quad = (\xi_k(t,p_0+\dot{p}) - \xi_k(t,p_0))\big(u_k(t,\xi_k(t,p_0),p_0+\dot{p}) - u_{k+1}(t,\xi_k(t,p_0),p_0+\dot{p}) + O(\|\dot{p}\|)\big) \\
&\quad = (\xi_k(t,p_0+\dot{p}) - \xi_k(t,p_0))\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0) + O(\|\dot{p}\|)\big) \\
&\quad = (\xi_k(t,p_0+\dot{p}) - \xi_k(t,p_0))\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big) + O(\|\dot{p}\|^2)
\end{aligned}$$

as $\dot{p} \to 0$.

By the subcase assumption we have

$$\xi_k(t,p_0+\dot{p}) \leq \xi_k(t,p_0) + \dot{\xi}_k(t,p_0)$$

and so by Taylor expansion the second term evaluates to

$$\begin{aligned}
&-\int_{S_k^1} \chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big)\mathrm{d}x \\
&\quad = -(\xi_k(t,p_0+\dot{p}) - \xi_k(t,p_0))\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big)\ .
\end{aligned}$$

So the residual between the first and second term is $O(\|\dot{p}\|^2)$ as $\dot{p} \to 0$.

By Assumption 6 we have u_k, u_{k+1} and $\dot{u}_k$ are bounded and since by Taylor expansion we have

$$\xi_k(t,p_0) + \dot{\xi}_k(t,p_0) - \xi_k(t,p_0+\dot{p}) = O(\|\dot{p}\|^2)$$

the fourth term evaluates to

$$\int_{S_k^2} O(1)\mathrm{d}x = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$. The third term and the fifth term also evaluate to $O(\|\dot{p}\|^2)$ by analogous steps to the first case. By adding all the terms we get that

$$\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\{y\in S_k : y>\xi_k(t,p_0)\})} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$ in the first subcase.

In the second subcase we have

$$\begin{aligned}
&\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\{y\in S_k : y>\xi_k(t,p_0)\})} \\
&\quad = \int_{S_k^1} u(t,x,p_0+\dot{p}) - u(t,x,p_0) - \dot{u}_{\mathrm{gen}}(t,x,p_0)\mathrm{d}x \\
&\qquad + \int_{S_k^2} u(t,x,p_0+\dot{p}) - u(t,x,p_0) - \dot{u}_{\mathrm{gen}}(t,x,p_0)\mathrm{d}x \\
&\qquad + \int_{S_k^3} u(t,x,p_0+\dot{p}) - u(t,x,p_0) - \dot{u}_{\mathrm{gen}}(t,x,p_0)\mathrm{d}x \\
&\quad = \int_{S_k^1} u_k(t,x,p_0+\dot{p}) - u_{k+1}(t,x,p_0) - \dot{u}_{k+1}(t,x,p_0)\mathrm{d}x \\
&\qquad - \int_{S_k^1} \chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0))\mathrm{d}x \\
&\qquad + \int_{S_k^2} u_k(t,x,p_0+\dot{p}) - u_{k+1}(t,x,p_0) - \dot{u}_{k+1}(t,x,p_0)\mathrm{d}x \\
&\qquad - \int_{S_k^2} \chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0))\mathrm{d}x \\
&\qquad + \int_{S_k^3} u_{k+1}(t,x,p_0+\dot{p}) - u_{k+1}(t,x,p_0) - \dot{u}_{k+1}(t,x,p_0)\mathrm{d}x
\end{aligned}$$

with

$$\begin{aligned}
S_k^1 &\equiv \{y \in S_k : y > \xi_k(t,p_0), y < \xi_k(t,p_0) + \dot{\xi}_k(t,p_0)\} \\
S_k^2 &\equiv \{y \in S_k : y > \xi_k(t,p_0) + \dot{\xi}_k(t,p_0), y < \xi_k(t,p_0+\dot{p})\} \\
S_k^3 &\equiv \{y \in S_k : y > \xi_k(t,p_0+\dot{p})\}\ .
\end{aligned}$$

Again, we analyze the terms separately. By analogous steps as in the other subcase the first term evaluates to

$$\begin{aligned}
&\int_{S_k^1} u_k(t,x,p_0+\dot{p}) - u_{k+1}(t,x,p_0) - \dot{u}_{k+1}(t,x,p_0)\mathrm{d}x \\
&\quad = \dot{\xi}_k(t,p_0)\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big) + O(\|\dot{p}\|^2)
\end{aligned}$$

as $\dot{p} \to 0$. The second term evaluates to

$$\begin{aligned}
&-\int_{S_k^1} \chi_{[\xi_k(t,p_0),\xi_k(t,p_0)+\dot{\xi}_k(t,p_0)]}(x)(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0))\mathrm{d}x \\
&\quad = -\dot{\xi}_k(t,p_0)\big(u_k(t,\xi_k(t,p_0),p_0) - u_{k+1}(t,\xi_k(t,p_0),p_0)\big)\ .
\end{aligned}$$

Again, the residual between the first and second term is $O(\|\dot{p}\|^2)$ as $\dot{p} \to 0$. The third and fourth term are both $O(\|\dot{p}\|^2)$ because the integrand is bounded and the domain of those integrals has width

$$\xi_k(t,p_0+\dot{p}) - \xi_k(t,p_0) - \dot{\xi}_k(t,p_0) = O(\|\dot{p}\|)^2$$

by Taylor expansion. The fifth term is $O(\|\dot{p}\|^2)$ by analogous arguments as above based on the continuous differentiability of u_{k+1} with respect to p. By adding the terms we get that

$$\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\{y\in S_k : y>\xi_k(t,p_0)\})} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$ in the second subcase.

Combining the two sets left and right of the shock location we get

$$\begin{aligned}&\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(S_k)}\\&= \|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\{y\in S_k : y<\xi_k(t,p_0)\})}\\&\quad + \|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\{y\in S_k : y>\xi_k(t,p_0)\})} = O(\|\dot{p}\|^2)\end{aligned}$$

as $\dot{p} \to 0$ for both subcases of the second case.

The third case is analogous to the second case by symmetry around the shock location.

We have first shown that

$$\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\hat{\Omega}\setminus\bigcup_{k=1}^K S_k)} = O(\|\dot{p}\|^2)$$

and then shown that

$$\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(S_k)} = O(\|\dot{p}\|^2)$$

for all $k \in \{1,\dots,K\}$ so we have

$$\begin{aligned}&\|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\hat{\Omega})}\\&= \|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(\hat{\Omega}\setminus\bigcup_{k=1}^K S_k)}\\&\quad + \sum_{i=1}^K \|u(t,\cdot,p_0+\dot{p}) - u(t,\cdot,p_0) - \dot{u}_{\mathrm{gen}}(t,\cdot,p_0)\|_{L_1(S_k)}\\&= O(\|\dot{p}\|^2) + K\cdot O(\|\dot{p}\|^2) = O(\|\dot{p}\|^2)\end{aligned}$$

as $\dot{p} \to 0$ since $K < \infty$. The above is the statement of the proposition. This concludes the proof. □

A notable remark that relates the generalized tangent and the distributional tangent sensitivity is the following.

Remark 27. The distributional tangent sensitivity with respect to $\dot{p}$ of the generalized tangent at $\dot{p} = 0$ is the distributional tangent sensitivity of the weak solution.

We do not give a formal proof but the remark follows directly from the weak differentiation of the characteristic function with respect to $\dot{p}$ which will give the Dirac delta distribution terms in the distributional tangent sensitivity of the weak solution.

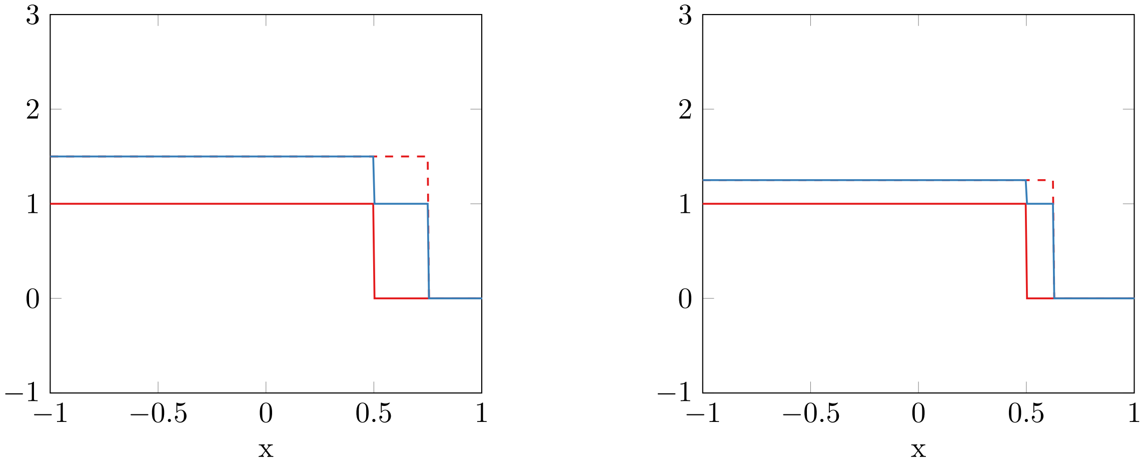
Example 32 (Riemann)**.** We consider the parametric Riemann problem from Example 25. According to Definition 37 the nonlinear generalized tangent sensitivity is

$$\dot{u}_{\mathrm{gen}}(t,x,p) = \chi_{[\xi_1(t,p),\xi_1(t,p)+\dot{\xi}_1(t,p)]}(x)(1+p) + \begin{cases}\dot{p} & \text{if } x \in (-\infty,\xi_1(t,p))\\ 0 & \text{if } x \in (\xi_1(t,p),\infty)\end{cases}$$

with ξ_1 as before, $\dot{\xi}_1 = \mathrm{d}_p\xi_1\dot{p}$ with $\mathrm{d}_p\xi_1$ as before and for p in a neighborhood of $p = p_0 = 0$.

In fact, since the shock location is linear in the parameter $\dot{\xi}_1(t,p_0) = \xi_1(t,p_0+\dot{p}) - \xi_1(t,p_0)$ the shock shift in the generalized tangent is exact. The only difference between the perturbation and the generalized tangent sensitivity in this example is the height of the shock shift. The error in the height of the jump discontinuity is $O(|\dot{p}|)$ as $\dot{p} \to 0$. Since the width of the shock shift due to the perturbation itself is $O(|\dot{p}|)$ the difference in the integral is $O(|\dot{p}|^2)$ as $\dot{p} \to 0$.

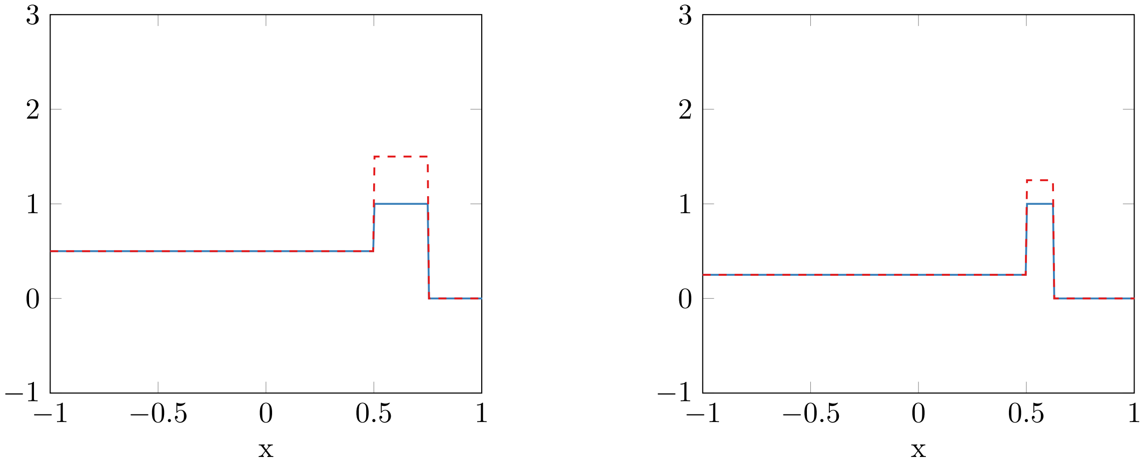
Figure 5.4 shows the solution $u(p_0)$ in solid red, the perturbed solution $u(p_0 + \dot{p})$ in dashed red and the solution plus generalized tangent $u(p_0) + \dot{u}_{\text{gen}}(p_0)$ in blue. Figure 5.5 shows the finite difference $u(p_0 + \dot{p}) - u(p_0)$ in red and the generalized tangent $\dot{u}_{\text{gen}}(p_0)$ in blue. Both figures show the perturbation for $\dot{p} = 1/2$ (left) and $\dot{p} = 1/4$ (right).



(a) $\dot{p} = 1/2$

(b) $\dot{p} = 1/4$

Figure 5.4: $u(t, x, 0)$ (solid red), $u(t, x, \dot{p})$ (dashed red), $u(t, x, 0) + \dot{u}_{\text{gen}}(t, x, 0)$ (solid blue)



(a) $\dot{p} = 1/2$

(b) $\dot{p} = 1/4$

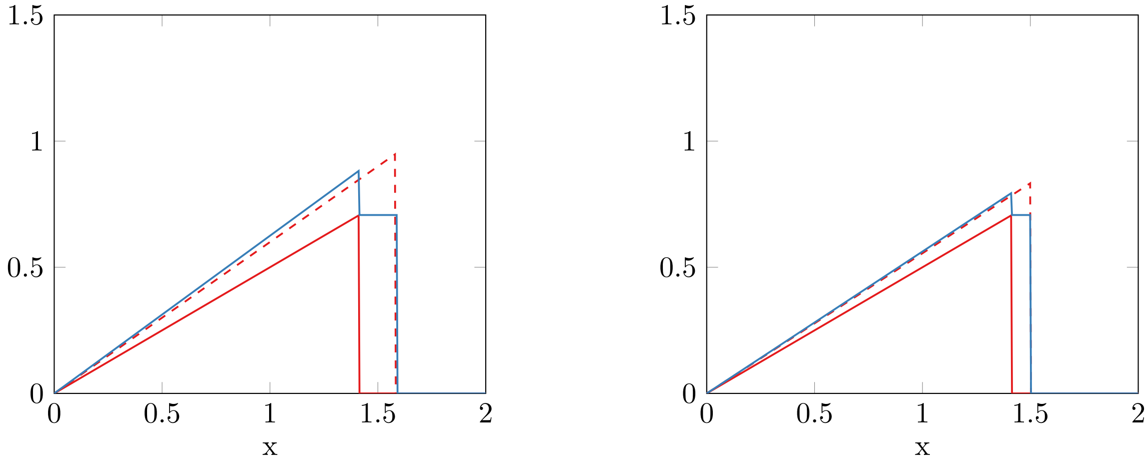
Figure 5.5: $u(t, x, \dot{p}) - u(t, x, 0)$ (dashed red), $\dot{u}_{\text{gen}}(t, x, 0)$ (solid blue)

Example 33 (Ramp)**.** We consider the ramp initial value problem from Example 26. According to Definition 37 the nonlinear generalized tangent sensitivity is

$$\dot{u}_{\text{gen}}(t, x, p) = \chi_{[\xi_1(t,p), \xi_1(t,p) + \dot{\xi}_1(t,p)]}(x) \frac{1+p}{1+(1+p)t} \xi_1(t, p)$$
$$+ \begin{cases} H(x) \frac{1}{(1+(1+p)t)^2} x\dot{p} & \text{if } x \in (-\infty, \xi_1(t, p)) \\ 0 & \text{if } x \in (\xi_1(t, p), \infty) \end{cases}$$

with ξ_1 as before, $\dot{\xi}_1 = \mathrm{d}_p \xi_1 \dot{p}$ with $\mathrm{d}_p \xi_1$ as before and for p in a neighborhood of $p = p_0 = 0$.

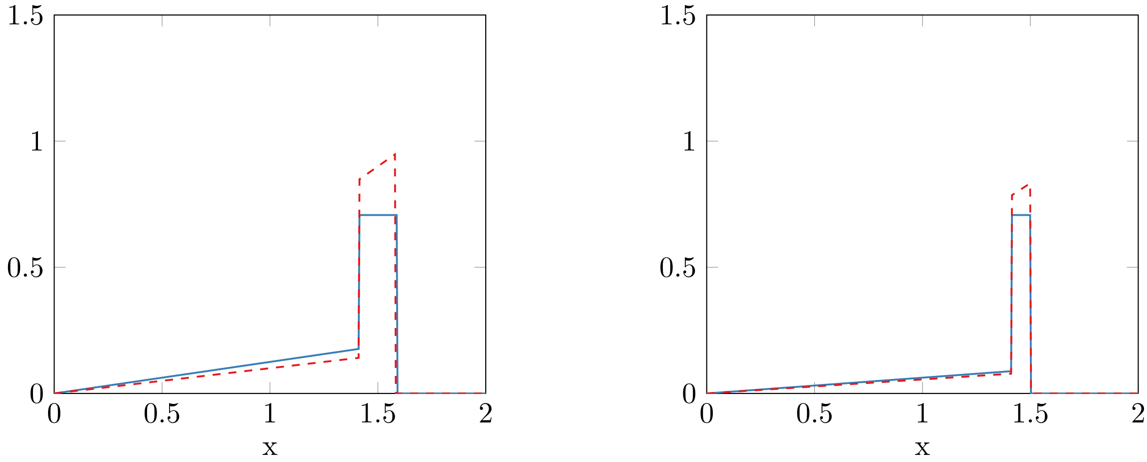
Figure 5.6 and Figure 5.7 are analogous to Figure 5.4 and Figure 5.5. For $\dot{p} = 1/2$ we can see that the approximation of the shock shift by the tangent sensitivity of the shock location is not exact in this case because the shock location is nonlinear in p. Also, the tangent sensitivity of the smooth parts is not an exact approximation of the perturbation because the height of the ramp is nonlinear in p as well.



(a) $\dot{p} = 1/2$

(b) $\dot{p} = 1/4$

Figure 5.6: $u(t, x, 0)$ (solid red), $u(t, x, \dot{p})$ (dashed red), $u(t, x, 0) + \dot{u}_{\text{gen}}(t, x, 0)$ (solid blue)



(a) $\dot{p} = 1/2$

(b) $\dot{p} = 1/4$

Figure 5.7: $u(t, x, \dot{p}) - u(t, x, 0)$ (dashed red), $\dot{u}_{\text{gen}}(t, x, 0)$ (solid blue)

The figures illustrate how the nonlinear generalized tangent is an approximation of the perturbation in the sense of the L_1 norm.

5.4 Adjoint Sensitivity

We now derive the adjoint sensitivity of an integral objective functional of the assumed solution structure. We give a perspective on the weak adjoint sensitivity in the sense of the Colombeau algebra where the intuition is via the limit of smooth approximations to the discontinuous solution. But we focus on the derivation of the adjoint of the generalized tangent decomposition into the broad tangent and the shock location tangent sensitivities as in Bressan and Marson [BM95a]. We note that for the numerical analysis our emphasis will not be on the adjoint sensitivity analysis since for the discrete sensitivities the adjoint is derived as the discrete adjoint of the discretization of the generalized tangent by Lemma 18.

5.4.1 Integral Objective for the Solution Structure

We now consider the adjoint sensitivity of an integral objective of the form

$$\Phi \equiv \int_\Omega \varphi(u(T,x))\mathrm{d}x$$

with respect to the weak solution at the final time $t = T$.

We already stated the following technical assumption regarding the function φ in the introduction but again make it explicit here.

Assumption 7. Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable.

We first consider the sensitivity with respect to the weak solution given in the structure of Assumption 6.

Proposition 28. The adjoint sensitivity

$$\bar{\xi}_k(T) \equiv \left(\mathrm{d}_{\xi_k(T)}\Phi\right)^* \bar{\Phi}$$

of the shock location ξ_k at $t = T$ is

$$\bar{\xi}_k(T) = \left(\varphi(u_k(T,\xi_k(T))) - \varphi(u_{k+1}(T,\xi_k(T)))\right)\bar{\Phi}$$

for all $k \in \{1,\dots,K\}$.

The adjoint sensitivity

$$\bar{u}_k(T,\cdot) \equiv \left(\mathrm{d}_{u_k(T,\cdot)}\Phi\right)^* \bar{\Phi}$$

of the smooth weak solution piece u_k at $t = T$ is defined by

$$\bar{u}_k(T,x) = \varphi'(u_k(T,x))\bar{\Phi}$$

for all $x \in [\xi_{k-1}(T), \xi_k(T)]$ and for all $k \in \{1,\dots,K+1\}$.

Proof. The proof is by splitting of the domain according to the piecewise definition of the weak solution u based on the shock locations. We have

$$\Phi = \int_\Omega \varphi(u(T,x))\mathrm{d}x = \sum_{k=1}^{K+1} \int_{\xi_{k-1}(T)}^{\xi_k(T)} \varphi(u_k(T,x))\mathrm{d}x\ .$$

By chain rule we have the tangent

$$\dot{\Phi} = \sum_{k=1}^{K+1} \int_{\xi_{k-1}(T)}^{\xi_k(T)} \dot{u}_k(T,x)\varphi'(u_k(T,x))\mathrm{d}x + \sum_{k=1}^{K} \dot{\xi}_k(T)\big(\varphi(u_k(T,\xi_k(T))) - \varphi(u_{k+1}(T,\xi_k(T)))\big) \ .$$

By the definition of the adjoint operator we have

$$\begin{aligned}\langle \dot{\Phi}, \bar{\Phi} \rangle &= \sum_{k=1}^{K+1} \int_{\xi_{k-1}(T)}^{\xi_k(T)} \dot{u}_k(T,x)\varphi'(u_k(T,x))\bar{\Phi}\mathrm{d}x \\ &\quad + \sum_{k=1}^{K} \dot{\xi}_k(T)\big(\varphi(u_k(T,\xi_k(T))) - \varphi(u_{k+1}(T,\xi_k(T)))\big)\bar{\Phi} \\ &= \sum_{k=1}^{K+1} \int_{\xi_{k-1}(T)}^{\xi_k(T)} \dot{u}_k(T,x)\bar{u}_k(T,x)\mathrm{d}x + \sum_{k=1}^{K} \dot{\xi}_k(T)\bar{\xi}_k(T) \\ &= \Big\langle \big(\dot{u}_1(T,\cdot), \dot{\xi}_1(T), \dot{u}_2(T,\cdot), \dot{\xi}_2(T), \ldots, \dot{\xi}_K(T), \dot{u}_{K+1}(T,\cdot)\big), \\ &\quad \big(\bar{u}_1(T,\cdot), \bar{\xi}_1(T), \bar{u}_2(T,\cdot), \bar{\xi}_2(T), \ldots, \bar{\xi}_K(T), \bar{u}_{K+1}(T,\cdot)\big)\Big\rangle\end{aligned}$$

that implies the given definitions of the adjoint sensitivities. This concludes the proof. □

Example 34 (Riemann)**.** We consider the parametric Riemann problem from Example 25 as an argument of the integral objective

$$\Phi = \int_0^1 \varphi(u(T,x,p))\mathrm{d}x$$

defined by

$$\varphi(u) \equiv \frac{1}{2}(u-C)^2 \ .$$

This integral objective is similar to the one for which we derived the gradient with respect to a parameter p of the initial data in Section 1.3.3. Here, we consider the adjoint sensitivity of the integral with respect to the weak solution u at the final time step T and not with respect to the parameter p. With $\bar{\Phi} = 1$ the adjoint sensitivity of the shock location is

$$\begin{aligned}\bar{\xi}_1(T,p) &= \varphi(u_1(T,\xi(T),p)) - \varphi(u_2(T,\xi(T),p)) \\ &= \frac{1}{2}(1+p-C)^2 - \frac{1}{2}C^2\end{aligned}$$

for p in a neighborhood of $p = p_0 = 0$. The functional gradient with respect to the smooth pieces of the solution is determined by

$$\varphi'(u) = u - C$$

as, respectively,

$$\begin{aligned}\bar{u}_1(T,x,p) &= \varphi'(u_1(T,x,p)) = 1 + p - C \\ \bar{u}_2(T,x,p) &= \varphi'(u_2(T,x,p)) = -C\end{aligned}$$

for all $x \in [0,1]$, p in a neighborhood of $p = p_0 = 0$.

If we consider the case of e.g. $C = 1$ and $p = p_0 = 0$ as in Section 1.3.3 it seems as though $\bar{u}$ is a discontinuous function in x with a jump discontinuity from $\bar{u}(T,\xi_1(T)-) = 0$

to $\bar{u}(T, \xi_1(T)+) = -1$ at $x = \xi_1(T)$. However, the adjoint $\bar{u}$ is in fact not simply a discontinuous function in L_1.

In the next section we illustrate that the value of the adjoint sensitivity $\bar{u}$ at the point $x = \xi_1(T)$ is crucial when considering $\bar{u}$ as an adjoint of the distributional tangent sensitivity of Proposition 24. In our assumed solution structure the relevant component of the adjoint sensitivity is captured by the adjoint sensitivity $\bar{\xi}_1$ of the shock location and we can safely ignore the value at the point $x = \xi_1(T)$.

5.4.2 Colombeau Weak Adjoint Sensitivity of Integral Objective

In Proposition 24 we considered the distributional tangent sensitivity $\dot{u}(T, \cdot)$ instead of the tangents of the shock locations and smooth pieces separately. In order to obtain an intuition of the adjoint sensitivity $\bar{u}(T, \cdot)$ without the structure of Assumption 6 we consider the limiting behavior of a smooth shock approximation.

We consider the example integral objective

$$\Phi = \int_0^1 \varphi(u(x))\mathrm{d}x$$

with $u(x) = H(\xi - x)u_1(x)$, $\xi = 1/2$ and $u_1(x) = 1$ for all $x \in [0, 1]$. As we know from Section 3.3.2 a smooth approximation of u is given by

$$u_{\text{smooth}}(x) = \frac{1}{2} - \frac{1}{2}\tanh\Big(-\frac{1}{\epsilon}(\xi - x)\Big)u_1(x)$$

where $\epsilon > 0$ is the smoothing parameter. For the integral objective

$$\Phi_{\text{smooth}} = \int_0^1 \varphi(u_{\text{smooth}}(x))\mathrm{d}x$$

we have the adjoint sensitivity

$$\bar{u}_{\text{smooth}}(x) = \varphi'(u_{\text{smooth}}(x))$$

for all $x \in [0, 1]$. The analogous adjoint sensitivity is not defined in the classical sense for the discontinuous function because the derivative $\varphi'(u(x))$ is undefined at $x = \xi$.

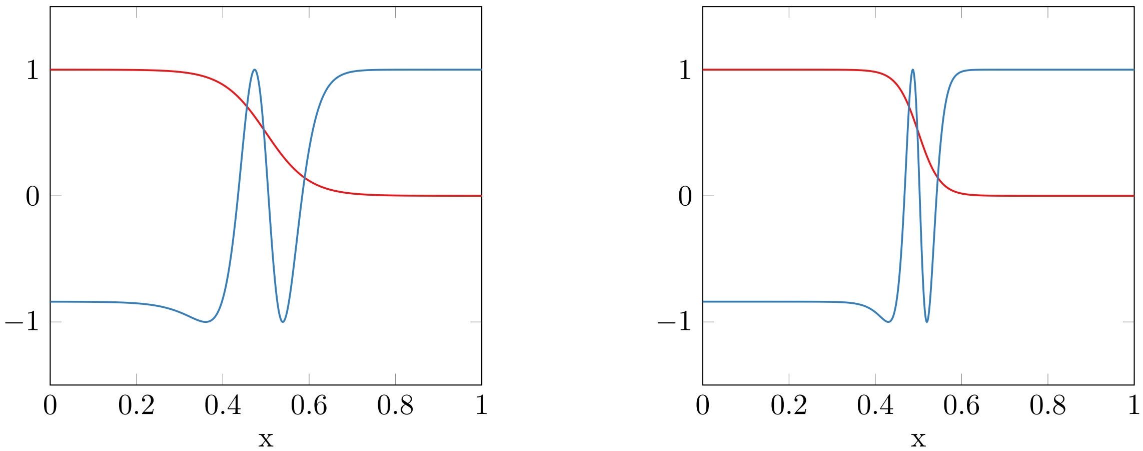
Figure 5.8 shows the graphs of $u_{\text{smooth}}, \bar{u}_{\text{smooth}}$ for different smoothing parameter values $\epsilon > 0$ and

$$\varphi(u) = \sin(10u)/10, \qquad \varphi'(u) = \cos(10u)\ .$$

We observe that outside of the smooth region the adjoint $\bar{u}_{\text{smooth}}$ converges to the piecewise adjoints with values $\varphi'(1)$ left of the discontinuity and $\varphi'(0)$ right of the discontinuity as $\epsilon \to 0$. But inside of the smoothed discontinuity region the adjoint sensitivity does not clearly converge to a specific value rather its graph maintains the same shape compressed into a narrower domain.

The distributional tangent sensitivity of the discontinuous version of u is

$$\dot{u}(x) = \delta(\xi - x)\dot{\xi}u_1(\xi) + H(\xi - x)\dot{u}_1(x)$$



(a) $\epsilon = 1 \cdot 10^{-1}$ (b) $\epsilon = \frac{1}{2} \cdot 10^{-1}$

Figure 5.8: $u_{\text{smooth}}(x)$ (red), $\bar{u}_{\text{smooth}}(x)$ (blue)

for all $x \in [0, 1]$. Based on the naive application of the chain rule and the definition of the adjoint operator we need to have

$$\begin{aligned}\langle \dot{\Phi}, \bar{\Phi} \rangle &= \int_0^1 \varphi'(u(x))\dot{u}(x)\bar{\Phi}\mathrm{d}x \\ &= \int_0^1 \dot{u}(x)\bar{u}(x)\mathrm{d}x = \langle \dot{u}(\cdot), \bar{u}(\cdot) \rangle \ .\end{aligned}$$

The above integral is not defined in the sense of distributions since distributions act on smooth test functions with compact support. We can only hope to find the adjoint in the weak sense where the weak adjoint identity becomes

$$\langle \dot{\Phi}, \bar{\Phi} \rangle = \int_0^1 \dot{u}(x)\bar{u}(x)\phi(x)\mathrm{d}x$$

for all smooth test functions ϕ with compact support.

But even in the sense of distributions it is not clear how we can multiply the Dirac delta distribution $\delta(\xi - x)$ in $\dot{u}$ with the discontinuous $\varphi'(u(x))$ that has the discontinuity at $x = \xi$. In Section 2.4.5 we introduced the notion of Colombeau generalized functions that allow for this sort of multiplication of distributions. We have

$$\begin{aligned}\langle \dot{\Phi}, \bar{\Phi} \rangle &= \int_0^1 \varphi'(u(x))\dot{u}(x)\phi(x)\bar{\Phi}\mathrm{d}x && (5.4.1)\\ &= \int_0^1 \varphi'(u(x))\big(\delta(\xi - x)\dot{\xi}u_1(\xi) + H(\xi - x)\dot{u}_1(x)\big)\phi(x)\bar{\Phi}\mathrm{d}x && (5.4.2)\\ &= \int_0^1 \varphi'(u(x))\delta(\xi - x)\dot{\xi}u_1(\xi)\phi(x)\bar{\Phi}\mathrm{d}x + \int_0^1 \varphi'(u(x))H(\xi - x)\dot{u}_1(x)\phi(x)\bar{\Phi}\mathrm{d}x && (5.4.3)\\ &= \int_0^1 \dot{u}(x)\bar{u}(x)\phi(x)\mathrm{d}x \ . && (5.4.4)\end{aligned}$$

By definition of the Colombeau generalized function from a distribution we have

$$\delta(\xi - \cdot)[\phi](y) = \int_{-\infty}^{\infty} \delta(\xi - x)\phi(x - y)\mathrm{d}x$$

and

$$\varphi'(u(\cdot))[\phi](y) = \int_{-\infty}^{\infty} \varphi'(u(x))\phi(x-y)\mathrm{d}x \, .$$

By definition of the Colombeau algebra we have the product as

$$\begin{aligned}\big(\varphi'(u(\cdot)) \cdot \delta(\xi - \cdot)\big)[\phi](y) &= \int_{-\infty}^{\infty} \varphi'(u(x))\delta(\xi - x)\phi(x-y)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \varphi'(u(x))\phi(x-y)\mathrm{d}x \cdot \int_{-\infty}^{\infty} \delta(\xi - x)\phi(x-y)\mathrm{d}x \\ &= \varphi'(u(\cdot))[\phi](y) \cdot \delta(\xi - \cdot)[\phi](y) \, .\end{aligned}$$

Consequently, we have the first term of Equation (5.4.3) as

$$\begin{aligned}\int_0^1 \varphi'(u(x))\delta(\xi - x)\dot{\xi}u_1(\xi)\phi(x)\bar{\Phi}\mathrm{d}x &= \big(\varphi'(u(\cdot)) \cdot \delta(\xi - \cdot)\big)[\phi](0) \cdot \dot{\xi} \cdot \bar{\Phi} \\ &= \varphi'(u(\cdot))[\phi](0) \cdot \delta(\xi - \cdot)[\phi](0) \cdot \dot{\xi} \cdot \bar{\Phi} \\ &= \int_{-\infty}^{\infty} \varphi'(u(x))\phi(x)\mathrm{d}x \cdot \int_{-\infty}^{\infty} \delta(\xi - x)\phi(x)\mathrm{d}x \cdot \dot{\xi} \cdot \bar{\Phi} \, .\end{aligned}$$

In the sense of the Colombeau algebra we have the weak adjoint $\bar{u}$ as a Colombeau generalized function with

$$\bar{u}[\phi](y) = \int_{-\infty}^{\infty} \varphi'(u(x))\bar{\Phi}\phi(x-y)\mathrm{d}x \, .$$

In order to have an idea about what $\bar{u}$ looks like it makes sense to consider $\bar{u}[\phi_\epsilon]$ with the test function ϕ_ϵ defined in Equation (2.4.2) as $\epsilon \to 0$. The adjoint $\bar{u}$ is, essentially, the limit of a smooth approximation similar to $\bar{u}_{\text{smooth}}$ but not defined in the sense of a classical function.

To gain a better intuition about the "value" of the adjoint $\bar{u}$ at $x = \xi$ we consider the limit of an approximation of the discontinuity by a linear function that interpolates between the function values around the jump discontinuity in a region of width ϵ. We define the linear interpolation by

$$u_{\text{lin}}(x) \equiv u(\xi - \epsilon) + \frac{x - (\xi - \epsilon)}{2\epsilon}(u(\xi + \epsilon) - u(\xi - \epsilon))$$

for $x \in (\xi - \epsilon, \xi + \epsilon)$. We can derive $\bar{u}(\xi)$ as the limit of the average of the adjoint in the domain $(\xi - \epsilon, \xi + \epsilon)$ as $\epsilon \to 0$. In that sense we have

$$\bar{u}(\xi) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_{\xi - \epsilon}^{\xi + \epsilon} \varphi'(u_{\text{lin}}(x))\mathrm{d}x \, .$$

Since on the relevant domain u_{lin} is an invertible linear function we can perform a change of variable by substituting $x = u_{\text{lin}}^{-1}(u)$, i.e. we have

$$\begin{aligned}&\frac{1}{2\epsilon} \int_{\xi - \epsilon}^{\xi + \epsilon} \varphi'(u_{\text{lin}}(x))\mathrm{d}x \\ &= \frac{1}{2\epsilon} \int_{u_{\text{lin}}(\xi - \epsilon)}^{u_{\text{lin}}(\xi + \epsilon)} \varphi'(u_{\text{lin}}(u_{\text{lin}}^{-1}(u))) \cdot (u_{\text{lin}}^{-1})'(u)\mathrm{d}u \\ &= \frac{1}{2\epsilon} \int_{u_{\text{lin}}(\xi - \epsilon)}^{u_{\text{lin}}(\xi + \epsilon)} \varphi'(u_{\text{lin}}(u_{\text{lin}}^{-1}(u))) \cdot u_{\text{lin}}'(u_{\text{lin}}^{-1}(u)) \cdot (u_{\text{lin}}^{-1})'(u) \cdot \frac{1}{u_{\text{lin}}'(u_{\text{lin}}^{-1}(u))}\mathrm{d}u \\ &= \frac{1}{2\epsilon} \int_{u_{\text{lin}}(\xi - \epsilon)}^{u_{\text{lin}}(\xi + \epsilon)} \varphi'(u_{\text{lin}}(u_{\text{lin}}^{-1}(u))) \cdot u_{\text{lin}}'(u_{\text{lin}}^{-1}(u)) \cdot (u_{\text{lin}}^{-1})'(u) \cdot \frac{1}{\frac{1}{2\epsilon}(u(\xi + \epsilon) - u(\xi - \epsilon))}\mathrm{d}u\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{u(\xi+\epsilon)-u(\xi-\epsilon)} \int_{u_{\mathrm{lin}}(\xi-\epsilon)}^{u_{\mathrm{lin}}(\xi+\epsilon)} \varphi'(u_{\mathrm{lin}}(u_{\mathrm{lin}}^{-1}(u))) \cdot u_{\mathrm{lin}}'(u_{\mathrm{lin}}^{-1}(u)) \cdot (u_{\mathrm{lin}}^{-1})'(u)\mathrm{d}u \\
&= \frac{1}{u(\xi+\epsilon)-u(\xi-\epsilon)} \Big(\varphi(u_{\mathrm{lin}}(u_{\mathrm{lin}}^{-1}(u_{\mathrm{lin}}(\xi+\epsilon)))) - \varphi(u_{\mathrm{lin}}(u_{\mathrm{lin}}^{-1}(u_{\mathrm{lin}}(\xi-\epsilon)))) \Big) \\
&= \frac{\varphi(u_{\mathrm{lin}}(\xi+\epsilon)) - \varphi(u_{\mathrm{lin}}(\xi-\epsilon))}{u(\xi+\epsilon)-u(\xi-\epsilon)} \, .
\end{aligned}
$$

We have the value of the adjoint $\bar{u}$ at the shock location as the generalized derivative of a discontinuous function defined by the limit of a central difference around the discontinuity

$$
\bar{u}(\xi) = \lim_{\epsilon\to 0} \frac{\varphi(u_{\mathrm{lin}}(\xi+\epsilon)) - \varphi(u_{\mathrm{lin}}(\xi-\epsilon))}{u(\xi+\epsilon)-u(\xi-\epsilon)} = \frac{\varphi(u(\xi+)) - \varphi(u(\xi-))}{u(\xi+)-u(\xi-)} \, .
$$

For the rest of the adjoint sensitivity analysis we go back to the solution structure where the sensitivitiy analysis is explicitly decomposed into the smooth pieces of the weak solution and the shock locations.

5.4.3 Time Stepping Operator for Solution Structure

We now consider the adjoint sensitivity of the time stepping operator analogous to Section 4.1.5 but for the piecewise smooth solution structure.

Let the PDE solution time stepping operator $S_{t_i,t_{i+1}}$ for a time step from $t = t_i$ to $t = t_{i+1}$ with $t_i < t_{i+1}$ for the solution structure given in Assumption 6 be defined by

$$
\begin{aligned}
&S_{t_i,t_{i+1}}\big(u_1(t_i,\cdot), \xi_1(t_i), u_2(t_i,\cdot), \xi_2(t_i), \dots, \xi_K(t_i), u_{K+1}(t_i,\cdot)\big) \\
&\quad \equiv \big(u_1(t_{i+1},\cdot), \xi_1(t_{i+1}), u_2(t_{i+1},\cdot), \xi_2(t_{i+1}), \dots, \xi_K(t_{i+1}), u_{K+1}(t_{i+1},\cdot)\big) \, .
\end{aligned}
$$

We ignore the parameter p in the notation at this point because the parameter dependence is implicit and not relevant to the derivation of the time stepping operator sensitivites.

The tangent sensitivity time stepping operator $\mathcal{S}_{t_i,t_{i+1}}$ of the solution operator $S_{t_i,t_{i+1}}$ at

$$
u_1(t_i,\cdot), \xi_1(t_i), u_2(t_i,\cdot), \xi_2(t_i), \dots, \xi_K(t_i), u_{K+1}(t_i,\cdot)
$$

is defined by

$$
\begin{aligned}
&\mathcal{S}_{t_i,t_{i+1}}\big(\dot{u}_1(t_i,\cdot), \dot{\xi}_1(t_i), \dot{u}_2(t_i,\cdot), \dot{\xi}_2(t_i), \dots, \dot{\xi}_K(t_i), \dot{u}_{K+1}(t_i,\cdot)\big) \\
&\quad \equiv \big(\dot{u}_1(t_{i+1},\cdot), \dot{\xi}_1(t_{i+1}), \dot{u}_2(t_{i+1},\cdot), \dot{\xi}_2(t_{i+1}), \dots, \dot{\xi}_K(t_{i+1}), \dot{u}_{K+1}(t_{i+1},\cdot)\big) \, .
\end{aligned}
$$

From the entropy condition (see Section 3.1.3) we know that the information flow around shocks in the solution structure is from the smooth pieces of the solution into the shock discontinuities. Therefore, in the forward in time solution no perturbation information flows from the shock location into the smooth pieces of the solution. Since we assume no intersection of shock curves only the two neighboring parts of the solution structure influence a shock location according to the Rankine-Hugoniot condition. Assuming continuous differentiability of the solution with respect to the initial data we have

$$
\dot{u}_k(t_{i+1},\cdot) = \mathrm{d}_{u_k(t_i,\cdot)} u_k(t_{i+1},\cdot)\dot{u}_k(t_i,\cdot)
$$

for all $k \in \{1, \dots, K+1\}$ and

$$\dot{\xi}_k(t_{i+1}) = \mathrm{d}_{\xi_k(t_i)}\xi_k(t_{i+1})\dot{\xi}_k(t_i) + \mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\dot{u}_k(t_i,\cdot) \\ + \mathrm{d}_{u_{k+1}(t_i,\cdot)}\xi_{k+1}(t_{i+1})\dot{u}_{k+1}(t_i,\cdot)$$

for all $k \in \{1, \dots, K\}$.

In principle, the tangent sensitivity $\dot{u}_k$ satisfies the tangent PDE derived in Section 4.1.2 for all $k \in \{1, \dots, K+1\}$. However, the question about the domain on which the PDE is satisfied and on which the initial data is defined is not completely straight forward. Technically, the initial data for $\dot{u}_k$ for a time step from $t = t_i$ to $t = t_{i+1}$ is only given on $(\xi_{k-1}(t_i), \xi_k(t_i))$. Since the information flows into the shocks the PDE solution and its tangent sensitivity can be fully determined on the domain with the shock curves as boundaries, i.e. on $\big(\xi_{k-1}(t), \xi_k(t)\big)$ for all $t \in [t_i, t_{i+1}]$, by the method of characteristics [Eva98, Chapter 3.2], [Bre00, Chapter 3.2]. For the smooth pieces the derivative of the solution with respect to the initial data is continuous due to the continuous differentiability with respect to the initial value of the solution operator of the ODE describing the characteristics (see Section 3.1.2).

The tangent sensitivity $\dot{\xi}_k$ satisfies a tangent ODE analogous to the one derived in the proof of Proposition 23. The only difference is that here we leave out the explicit dependence on the parameter p. We have

$$\partial_t\dot{\xi}_k(t) = \frac{f'(u_k(t,\xi_k(t)))\big(\partial_x u_k(t,\xi_k(t))\dot{\xi}_k(t) + \dot{u}_k(t,\xi_k(t))\big)}{u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))} \tag{5.4.5}$$

$$- \frac{f'(u_{k+1}(t,\xi_k(t)))\big(\partial_x u_{k+1}(t,\xi_k(t))\dot{\xi}_k(t) + \dot{u}_{k+1}(t,\xi_k(t))\big)}{u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))} \tag{5.4.6}$$

$$- \frac{f(u_k(t,\xi_k(t))) - f(u_{k+1}(t,\xi_k(t)))}{\big(u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))\big)^2} \tag{5.4.7}$$

$$\cdot \big(\partial_x u_k(t,\xi_k(t))\dot{\xi}_k(t) + \dot{u}_k(t,\xi_k(t)) - \partial_x u_{k+1}(t,\xi_k(t))\dot{\xi}_k(t) - \dot{u}_{k+1}(t,\xi_k(t))\big) \,. \tag{5.4.8}$$

We can clearly see how the shock location tangent $\dot{\xi}_k$ linearly depends on the tangents of the neighboring smooth pieces of the solution tangents $\dot{u}_k$ and $\dot{u}_{k+1}$.

Naturally, for the adjoint the dependencies will be reversed. The adjoint sensitivities of the neighboring smooth pieces $\bar{u}_k$ and $\bar{u}_{k+1}$ will depend on the adjoint shock location $\bar{\xi}_k$.

Proposition 29. Let the time stepping operator $S_{t_i,t_{i+1}}$ for the solution structure given in Assumption 6 be continuously differentiable at

$$u_1(t_i,\cdot), \xi_1(t_i), u_2(t_i,\cdot), \xi_2(t_i), \dots, \xi_K(t_i), u_{K+1}(t_i,\cdot) \,.$$

Then the adjoint operator $\mathcal{S}^*_{t_i,t_{i+1}}$ of the derivative $\mathcal{S}_{t_i,t_{i+1}}$ of the time stepping operator is defined by

$$\mathcal{S}^*_{t_i,t_{i+1}}\big(\bar{u}_1(t_{i+1},\cdot), \bar{\xi}_1(t_{i+1}), \bar{u}_2(t_{i+1},\cdot), \bar{\xi}_2(t_{i+1}), \dots, \bar{\xi}_K(t_{i+1}), \bar{u}_{K+1}(t_{i+1},\cdot)\big) \\ \equiv \big(\bar{u}_1(t_i,\cdot), \bar{\xi}_1(t_i), \bar{u}_2(t_i,\cdot), \bar{\xi}_2(t_i), \dots, \bar{\xi}_K(t_i), \bar{u}_{K+1}(t_i,\cdot)\big)$$

with

$$\bar{u}_k(t_i,\cdot) = \big(\mathrm{d}_{u_k(t_i,\cdot)}u_k(t_{i+1},\cdot)\big)^*\bar{u}_k(t_{i+1},\cdot) + \big(\mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\big)^*\bar{\xi}_k(t_{i+1})$$

$$+ \big(\mathrm{d}_{u_k(t_i,\cdot)}\xi_{k-1}(t_{i+1})\big)^*\bar{\xi}_{k-1}(t_{i+1})$$

for all $k \in \{2, \dots, K\}$,

$$\begin{aligned}
\bar{u}_1(t_i,\cdot) &= \big(\mathrm{d}_{u_1(t_i,\cdot)}u_1(t_{i+1},\cdot)\big)^*\bar{u}_1(t_{i+1},\cdot) + \big(\mathrm{d}_{u_1(t_i,\cdot)}\xi_1(t_{i+1})\big)^*\bar{\xi}_1(t_{i+1}),\\
\bar{u}_{K+1}(t_i,\cdot) &= \big(\mathrm{d}_{u_{K+1}(t_i,\cdot)}u_{K+1}(t_{i+1},\cdot)\big)^*\bar{u}_{K+1}(t_{i+1},\cdot) + \big(\mathrm{d}_{u_{K+1}(t_i,\cdot)}\xi_K(t_{i+1})\big)^*\bar{\xi}_K(t_{i+1})
\end{aligned}$$

and

$$\bar{\xi}_k(t_i,\cdot) = \big(\mathrm{d}_{\xi_k(t_i)}\xi_k(t_{i+1})\big)^*\bar{\xi}_k(t_{i+1})$$

for all $k \in \{1, \dots, K\}$.

Proof. By the definition of the adjoint operator (see Section 2.1.3) we have

$$\begin{aligned}
&\Big\langle \mathcal{S}_{t_i,t_{i+1}}\big(\dot{u}_1(t_i,\cdot), \dot{\xi}_1(t_i), \dot{u}_2(t_i,\cdot), \dot{\xi}_2(t_i), \dots, \dot{\xi}_K(t_i), \dot{u}_{K+1}(t_i,\cdot)\big),\\
&\qquad \big(\bar{u}_1(t_{i+1},\cdot), \bar{\xi}_1(t_{i+1}), \bar{u}_2(t_{i+1},\cdot), \bar{\xi}_2(t_{i+1}), \dots, \bar{\xi}_K(t_{i+1}), \bar{u}_{K+1}(t_{i+1},\cdot)\big)\Big\rangle\\
&= \Big\langle \big(\dot{u}_1(t_{i+1},\cdot), \dot{\xi}_1(t_{i+1}), \dot{u}_2(t_{i+1},\cdot), \dot{\xi}_2(t_{i+1}), \dots, \dot{\xi}_K(t_{i+1}), \dot{u}_{K+1}(t_{i+1},\cdot)\big),\\
&\qquad \big(\bar{u}_1(t_{i+1},\cdot), \bar{\xi}_1(t_{i+1}), \bar{u}_2(t_{i+1},\cdot), \bar{\xi}_2(t_{i+1}), \dots, \bar{\xi}_K(t_{i+1}), \bar{u}_{K+1}(t_{i+1},\cdot)\big)\Big\rangle\\
&= \sum_{k=1}^{K+1}\langle \dot{u}_k(t_{i+1},\cdot), \bar{u}_k(t_{i+1},\cdot)\rangle + \sum_{k=1}^{K}\langle \dot{\xi}_k(t_{i+1}), \bar{\xi}_k(t_{i+1})\rangle\\
&= \sum_{k=1}^{K+1}\langle \mathrm{d}_{u_k(t_i,\cdot)}u_k(t_{i+1},\cdot)\dot{u}_k(t_i,\cdot), \bar{u}_k(t_{i+1},\cdot)\rangle + \sum_{k=1}^{K}\langle \mathrm{d}_{\xi_k(t_i)}\xi_k(t_{i+1})\dot{\xi}_k(t_i), \bar{\xi}_k(t_{i+1})\rangle\\
&\qquad + \sum_{k=1}^{K}\langle \mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\dot{u}_k(t_i,\cdot), \bar{\xi}_k(t_{i+1})\rangle + \sum_{k=1}^{K}\langle \mathrm{d}_{u_{k+1}(t_i,\cdot)}\xi_{k+1}(t_{i+1})\dot{u}_{k+1}(t_i,\cdot), \bar{\xi}_k(t_{i+1})\rangle\\
&= \sum_{k=1}^{K+1}\langle \dot{u}_k(t_i,\cdot), \big(\mathrm{d}_{u_k(t_i,\cdot)}u_k(t_{i+1},\cdot)\big)^*\bar{u}_k(t_{i+1},\cdot)\rangle + \sum_{k=1}^{K}\langle \dot{\xi}_k(t_i), \big(\mathrm{d}_{\xi_k(t_i)}\xi_k(t_{i+1})\big)^*\bar{\xi}_k(t_{i+1})\rangle\\
&\qquad + \sum_{k=1}^{K}\langle \dot{u}_k(t_i,\cdot), \big(\mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\big)^*\bar{\xi}_k(t_{i+1})\rangle + \sum_{k=1}^{K}\langle \dot{u}_{k+1}(t_i,\cdot), \big(\mathrm{d}_{u_{k+1}(t_i,\cdot)}\xi_{k+1}(t_{i+1})\big)^*\bar{\xi}_k(t_{i+1})\rangle\\
&= \sum_{k=1}^{K}\langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i)\rangle + \sum_{k=1}^{K+1}\langle \dot{u}_k(t_i,\cdot), \big(\mathrm{d}_{u_k(t_i,\cdot)}u_k(t_{i+1},\cdot)\big)^*\bar{u}_k(t_{i+1},\cdot)\rangle\\
&\qquad + \sum_{k=1}^{K}\langle \dot{u}_k(t_i,\cdot), \big(\mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\big)^*\bar{\xi}_k(t_{i+1})\rangle + \sum_{k=2}^{K+1}\langle \dot{u}_k(t_i,\cdot), \big(\mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\big)^*\bar{\xi}_{k-1}(t_{i+1})\rangle\\
&= \sum_{k=1}^{K}\langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i)\rangle + \langle \dot{u}_1(t_i,\cdot), \big(\mathrm{d}_{u_1(t_i,\cdot)}u_1(t_{i+1},\cdot)\big)^*\bar{u}_1(t_{i+1},\cdot) + \big(\mathrm{d}_{u_1(t_i,\cdot)}\xi_1(t_{i+1})\big)^*\bar{\xi}_1(t_{i+1})\rangle\\
&\qquad + \langle \dot{u}_{K+1}(t_i,\cdot), \big(\mathrm{d}_{u_{K+1}(t_i,\cdot)}u_{K+1}(t_{i+1},\cdot)\big)^*\bar{u}_{K+1}(t_{i+1},\cdot) + \big(\mathrm{d}_{u_{K+1}(t_i,\cdot)}\xi_K(t_{i+1})\big)^*\bar{\xi}_K(t_{i+1})\rangle\\
&\qquad + \sum_{k=2}^{K}\langle \dot{u}_k(t_i,\cdot), \big(\mathrm{d}_{u_k(t_i,\cdot)}u_k(t_{i+1},\cdot)\big)^*\bar{u}_k(t_{i+1},\cdot) + \big(\mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\big)^*\bar{\xi}_k(t_{i+1})\\
&\qquad\qquad\qquad + \big(\mathrm{d}_{u_k(t_i,\cdot)}\xi_k(t_{i+1})\big)^*\bar{\xi}_{k-1}(t_{i+1})\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} \langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i) \rangle + \langle \dot{u}_1(t_i,\cdot), \bar{u}_1(t_i,\cdot) \rangle + \langle \dot{u}_{K+1}(t_i,\cdot), \bar{u}_{K+1}(t_i,\cdot) \rangle + \sum_{k=2}^{K} \langle \dot{u}_k(t_i,\cdot), \bar{u}_k(t_i,\cdot) \rangle \\
&= \sum_{k=1}^{K} \langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i) \rangle + \sum_{k=1}^{K+1} \langle \dot{u}_k(t_i,\cdot), \bar{u}_k(t_i,\cdot) \rangle \\
&= \Big\langle \big(\dot{u}_1(t_i,\cdot), \dot{\xi}_1(t_i), \dot{u}_2(t_i,\cdot), \dot{\xi}_2(t_i), \dots, \dot{\xi}_K(t_i), \dot{u}_{K+1}(t_i,\cdot)\big), \\
&\quad \big(\bar{u}_1(t_i,\cdot), \bar{\xi}_1(t_i), \bar{u}_2(t_i,\cdot), \bar{\xi}_2(t_i), \dots, \bar{\xi}_K(t_i), \bar{u}_{K+1}(t_i,\cdot)\big) \Big\rangle \\
&= \Big\langle \big(\dot{u}_1(t_i,\cdot), \dot{\xi}_1(t_i), \dot{u}_2(t_i,\cdot), \dot{\xi}_2(t_i), \dots, \dot{\xi}_K(t_i), \dot{u}_{K+1}(t_i,\cdot)\big), \\
&\quad \mathcal{S}^*_{t_i,t_{i+1}} \big(\bar{u}_1(t_{i+1},\cdot), \bar{\xi}_1(t_{i+1}), \bar{u}_2(t_{i+1},\cdot), \bar{\xi}_2(t_{i+1}), \dots, \bar{\xi}_K(t_{i+1}), \bar{u}_{K+1}(t_{i+1},\cdot)\big) \Big\rangle .
\end{aligned}
$$

This concludes the proof. □

The adjoint sensitivity of the solution structure also satisfies a set of adjoint differential equations that describe its dynamics backwards in time. Just like the tangent ODE for the shock location sensitivity couples the smooth tangent PDE solution pieces the adjoint sensitivity equations couple the smooth adjoint PDE solution pieces with the shock location adjoint.

In the backward in time solution the information flow is reversed and the characteristics flow out of the shock curves back into the adjoint solution for the smooth pieces. As observed by Giles and Pierce [GP01] the reversed flow of characteristics results in internal adjoint boundary conditions at the shock curves. In the following proposition we formulate these internal boundary conditions via Dirac delta distribution source terms in the adjoint PDEs. Contrary to the tangent ODE of the shock location sensitivities via the Rankine-Hugoniot condition that depend on the tangent sensitivities of the neighboring smooth solution pieces due to the reversed information flow the adjoint ODE of the shock location sensitivities do not have any dependence of the adjoint sensitivities on the neighboring smooth pieces.

Proposition 30. Let the assumptions of Proposition 29 be satisfied and let

$$
\begin{aligned}
A_k \equiv\ & \frac{f'(u_k(t,\xi_k(t)))\partial_x u_k(t,\xi_k(t)) - f'(u_{k+1}(t,\xi_k(t)))\partial_x u_{k+1}(t,\xi_k(t))}{u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))} \\
& - \frac{f(u_k(t,\xi_k(t))) - f(u_{k+1}(t,\xi_k(t)))}{\big(u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))\big)^2} \big(\partial_x u_k(t,\xi_k(t)) - \partial_x u_{k+1}(t,\xi_k(t))\big) \\
B_k \equiv\ & \frac{f'(u_k(t,\xi_k(t)))}{u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))} - \frac{f(u_k(t,\xi_k(t))) - f(u_{k+1}(t,\xi_k(t)))}{\big(u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))\big)^2} \\
C_k \equiv & -\frac{f'(u_{k+1}(t,\xi_k(t)))}{u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))} + \frac{f(u_k(t,\xi_k(t))) - f(u_{k+1}(t,\xi_k(t)))}{\big(u_k(t,\xi_k(t)) - u_{k+1}(t,\xi_k(t))\big)^2} .
\end{aligned}
$$

Then the adjoint sensitivity according to Proposition 29

$$\big(\bar{u}_1(t_i,\cdot), \bar{\xi}_1(t_i), \bar{u}_2(t_i,\cdot), \bar{\xi}_2(t_i), \dots, \bar{\xi}_K(t_i), \bar{u}_{K+1}(t_i,\cdot)\big)$$

satisfies the following PDEs in a weak sense

$$\partial_t \bar{u}_k(t,x) = \partial_x \bar{u}_k(t,x) f'(u(t,x)) - B_k \bar{\xi}_k(t)\delta(x - \xi_k(t)) - C_{k-1}\bar{\xi}_{k-1}(t)\delta(x - \xi_{k-1}(t))$$

for all $k \in \{2, \dots, K\}$,

$$\partial_t \bar{u}_1(t,x) = \partial_x \bar{u}_1(t,x) f'(u(t,x)) - B_1 \bar{\xi}_k(t)\delta(x - \xi_k(t))$$

$$\partial_t \bar{u}_{K+1}(t,x) = \partial_x \bar{u}_{K+1}(t,x) f'(u(t,x)) - C_{K+1}\bar{\xi}_K(t)\delta(x-\xi_K(t))$$

and the following ODE

$$\partial_t \bar{\xi}_k(t) = -A_k \bar{\xi}_k(t)$$

for all $k \in \{1,\dots,K\}$, for all $t \in [0,T]$, $x \in \Omega$.

Proof. The proof is an extension of the proof of Proposition 15 to the piecewise smooth solution structure of Assumption 6. We consider a small time step from $t = t_i$ to $t = t_{i+1}$ with $\Delta t = t_{i+1} - t_i$.

By the assumptions of the Proposition we have $\dot{\xi}_k, \bar{\xi}_k$ and $\dot{u}_k, \bar{u}_k$ sufficiently differentiable to get the following approximations from the Taylor expansion. Since $\dot{u}_k$ satisfies the tangent PDE (see Proposition 11) we have

$$\begin{aligned}\dot{u}_k(t_{i+1},\cdot) &= \dot{u}_k(t_i,\cdot) + \Delta t \partial_t u_k(t_i,\cdot) + O(\Delta t^2)\\ &= \dot{u}_k(t_i,\cdot) + \Delta t \partial_x\big(f'(u_k(t_i,\cdot))\dot{u}_k(t_i,\cdot)\big) + O(\Delta t^2)\end{aligned}$$

as $\Delta t \to 0$ for all $k \in \{1,\dots,K+1\}$. Since $\dot{\xi}_k$ satisfies the tangent ODE in Equation (5.4.8) we have

$$\begin{aligned}\dot{\xi}_k(t_{i+1}) &= \dot{\xi}_k(t_i) + \Delta t \partial_t \xi_k(t_i) + O(\Delta t^2)\\ &= \dot{\xi}_k(t_i) + \Delta t\Big(A_k\dot{\xi}_k(t_i) + B_k\dot{u}_k(t_i,\xi_k(t_i)) + C_k\dot{u}_{k+1}(t_i,\xi_k(t_i))\Big) + O(\Delta t^2)\end{aligned}$$

as $\Delta t \to 0$ for all $k \in \{1,\dots,K\}$.

From the definition of the adjoint operator we have

$$\begin{aligned}&\Big\langle \big(\dot{u}_1(t_{i+1},\cdot),\dot{\xi}_1(t_{i+1}),\dot{u}_2(t_{i+1},\cdot),\dot{\xi}_2(t_{i+1}),\dots,\dot{\xi}_K(t_{i+1}),\dot{u}_{K+1}(t_{i+1},\cdot)\big),\\ &\big(\bar{u}_1(t_{i+1},\cdot),\bar{\xi}_1(t_{i+1}),\bar{u}_2(t_{i+1},\cdot),\bar{\xi}_2(t_{i+1}),\dots,\bar{\xi}_K(t_{i+1}),\bar{u}_{K+1}(t_{i+1},\cdot)\big)\Big\rangle\\ &= \sum_{k=1}^{K+1}\langle \dot{u}_k(t_{i+1},\cdot),\bar{u}_k(t_{i+1},\cdot)\rangle + \sum_{k=1}^{K}\langle\dot{\xi}_k(t_{i+1}),\bar{\xi}_k(t_{i+1})\rangle\\ &= \sum_{k=1}^{K+1}\langle \dot{u}_k(t_i,\cdot),\bar{u}_k(t_i,\cdot)\rangle + \sum_{k=1}^{K}\langle\dot{\xi}_k(t_i),\bar{\xi}_k(t_i)\rangle\\ &= \Big\langle \big(\dot{u}_1(t_i,\cdot),\dot{\xi}_1(t_i),\dot{u}_2(t_i,\cdot),\dot{\xi}_2(t_i),\dots,\dot{\xi}_K(t_i),\dot{u}_{K+1}(t_i,\cdot)\big),\\ &\big(\bar{u}_1(t_i,\cdot),\bar{\xi}_1(t_i),\bar{u}_2(t_i,\cdot),\bar{\xi}_2(t_i),\dots,\bar{\xi}_K(t_i),\bar{u}_{K+1}(t_i,\cdot)\big)\Big\rangle\ .\end{aligned}$$

By Taylor expansion and integration by parts and analogous to the proof of Proposition 15 we have

$$\begin{aligned}\langle\dot{u}_k(t_{i+1},\cdot),\bar{u}_k(t_{i+1},\cdot)\rangle &= \langle\dot{u}_k(t_i,\cdot)+\Delta t\partial_t\dot{u}_k(t_i,\cdot),\bar{u}_k(t_i,\cdot)+\Delta t\partial_t\bar{u}_k(t_i,\cdot)\rangle + O(\Delta t^2)\\ &= \langle\dot{u}_k(t_i,\cdot),\bar{u}_k(t_i,\cdot)\rangle + \Delta t\langle\partial_t\dot{u}_k(t_i,\cdot),\bar{u}_k(t_i,\cdot)\rangle\\ &\quad + \Delta t\langle\dot{u}_k(t_i,\cdot),\partial_t\bar{u}_k(t_i,\cdot)\rangle + O(\Delta t^2)\\ &= \langle\dot{u}_k(t_i,\cdot),\bar{u}_k(t_i,\cdot)\rangle + \Delta t\langle\partial_x\big(f'(u_k(t_i,\cdot))\dot{u}_k(t_i,\cdot)\big),\bar{u}_k(t_i,\cdot)\rangle\\ &\quad + \Delta t\langle\dot{u}_k(t_i,\cdot),\partial_t\bar{u}_k(t_i,\cdot)\rangle + O(\Delta t^2)\end{aligned}$$

$$
\begin{aligned}
&= \langle \dot{u}_k(t_i,\cdot), \bar{u}_k(t_i,\cdot)\rangle - \Delta t \langle \dot{u}_k(t_i,\cdot), \partial_x \bar{u}_k(t_i,\cdot) f'(u_k(t_i,\cdot))\rangle \\
&\quad + \Delta t \langle \dot{u}_k(t_i,\cdot), \partial_t \bar{u}_k(t_i,\cdot)\rangle + O(\Delta t^2)
\end{aligned}
$$

as $\Delta t \to 0$ for all $k \in \{1, \dots, K+1\}$.

By Taylor expansion we also have

$$
\begin{aligned}
\langle \dot{\xi}_k(t_{i+1}), \bar{\xi}_k(t_{i+1})\rangle &= \langle \dot{\xi}_k(t_i) + \Delta t \partial_t \dot{\xi}_k(t_i), \bar{\xi}_k(t_i) + \Delta t \partial_t \bar{\xi}_k(t_i)\rangle + O(\Delta t^2) \\
&= \langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i)\rangle + \Delta t \langle \partial_t \dot{\xi}_k(t_i), \bar{\xi}_k(t_i)\rangle \\
&\quad + \Delta t \langle \dot{\xi}_k(t_i), \partial_t \bar{\xi}_k(t_i)\rangle + O(\Delta t^2) \\
&= \langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i)\rangle + \Delta t \langle A_k \dot{\xi}_k(t_i) + B_k \dot{u}_k(t_i, \xi_k(t_i)) + C_k \dot{u}_{k+1}(t_i, \xi_k(t_i)), \bar{\xi}_k(t_i)\rangle \\
&\quad + \Delta t \langle \dot{\xi}_k(t_i), \partial_t \bar{\xi}_k(t_i)\rangle + O(\Delta t^2)
\end{aligned}
$$

as $\Delta t \to 0$ for all $k \in \{1, \dots, K\}$.

By substitution we have

$$
\begin{aligned}
&\sum_{k=1}^{K+1} \langle \dot{u}_k(t_i,\cdot), \bar{u}_k(t_i,\cdot)\rangle + \sum_{k=1}^{K} \langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i)\rangle \\
&\quad = \sum_{k=1}^{K+1} \langle \dot{u}_k(t_{i+1},\cdot), \bar{u}_k(t_{i+1},\cdot)\rangle + \sum_{k=1}^{K} \langle \dot{\xi}_k(t_{i+1}), \bar{\xi}_k(t_{i+1})\rangle \\
&\quad = \sum_{k=1}^{K+1} \langle \dot{u}_k(t_i,\cdot), \bar{u}_k(t_i,\cdot)\rangle - \Delta t \sum_{k=1}^{K+1} \langle \dot{u}_k(t_i,\cdot), \partial_x \bar{u}_k(t_i,\cdot) f'(u_k(t_i,\cdot))\rangle \\
&\quad\quad + \Delta t \sum_{k=1}^{K+1} \langle \dot{u}_k(t_i,\cdot), \partial_t \bar{u}_k(t_i,\cdot)\rangle \\
&\quad\quad + \sum_{k=1}^{K} \langle \dot{\xi}_k(t_i), \bar{\xi}_k(t_i)\rangle + \Delta t \sum_{k=1}^{K} \langle A_k \dot{\xi}_k(t_i) + B_k \dot{u}_k(t_i, \xi_k(t_i)) + C_k \dot{u}_{k+1}(t_i, \xi_k(t_i)), \bar{\xi}_k(t_i)\rangle \\
&\quad\quad + \Delta t \sum_{k=1}^{K} \langle \dot{\xi}_k(t_i), \partial_t \bar{\xi}_k(t_i)\rangle + O(\Delta t^2)
\end{aligned}
$$

as $\Delta t \to 0$, i.e. we need to have

$$
\begin{aligned}
0 = &- \sum_{k=1}^{K+1} \langle \dot{u}_k(t_i,\cdot), \partial_x \bar{u}_k(t_i,\cdot) f'(u_k(t_i,\cdot))\rangle && (5.4.9) \\
&+ \sum_{k=1}^{K+1} \langle \dot{u}_k(t_i,\cdot), \partial_t \bar{u}_k(t_i,\cdot)\rangle && (5.4.10) \\
&+ \sum_{k=1}^{K} \langle A_k \dot{\xi}_k(t_i) + B_k \dot{u}_k(t_i, \xi_k(t_i)) + C_k \dot{u}_{k+1}(t_i, \xi_k(t_i)), \bar{\xi}_k(t_i)\rangle && (5.4.11) \\
&+ \sum_{k=1}^{K} \langle \dot{\xi}_k(t_i), \partial_t \bar{\xi}_k(t_i)\rangle && (5.4.12)
\end{aligned}
$$

which also implies that the same holds for the integral inner products in a weak sense.

By the smoothness of $\dot{u}_k$ for all $k \in \{1, \ldots, K+1\}$ the scalar inner products can be defined in a weak sense via the Dirac delta distribution as follows. Equations (5.4.9)-(5.4.12) hold in the weak sense if

$$\begin{aligned}
0 = &- \sum_{k=1}^{K+1} \int_{-\infty}^{\infty} \dot{u}_k(t_i, x) \partial_x \bar{u}_k(t_i, x) f'(u_k(t_i, x)) \phi(x) \mathrm{d}x \\
&+ \sum_{k=1}^{K+1} \int_{-\infty}^{\infty} \dot{u}_k(t_i, x) \partial_t \bar{u}_k(t_i, x) \phi(x) \mathrm{d}x \\
&+ \sum_{k=1}^{K} \int_{-\infty}^{\infty} \Big(A_k \dot{\xi}_k(t_i) + B_k \dot{u}_k(t_i, x) + C_k \dot{u}_{k+1}(t_i, x) \Big) \bar{\xi}_k(t_i) \delta(x - \xi_k(t_i)) \phi(x) \mathrm{d}x \\
&+ \sum_{k=1}^{K} \int_{-\infty}^{\infty} \dot{\xi}_k(t_i) \partial_t \bar{\xi}_k(t_i) \delta(x - \xi_k(t_i)) \phi(x) \mathrm{d}x + O(\Delta t)
\end{aligned}$$

for all smooth test functions ϕ with compact support as $\Delta t \to 0$.

By the distributional definition of the Dirac delta we have

$$\int_{-\infty}^{\infty} K \delta(x - \xi_k(t_i)) \phi(x) \mathrm{d}x = K \phi(\xi_k(t_i))$$

for any constant K and for all smooth test functions ϕ with compact support. Thus, we have

$$\int_{-\infty}^{\infty} A_k \dot{\xi}_k(t_i) \bar{\xi}_k(t_i) \delta(x - \xi_k(t_i)) \phi(x) \mathrm{d}x = A_k \dot{\xi}_k(t_i) \bar{\xi}_k(t_i) \phi(\xi_k(t_i))$$

and

$$\int_{-\infty}^{\infty} \dot{\xi}_k(t_i) \partial_t \bar{\xi}_k(t_i) \delta(x - \xi_k(t_i)) \phi(x) \mathrm{d}x = \dot{\xi}_k(t_i) \partial_t \bar{\xi}_k(t_i) \phi(\xi_k(t_i)) \ .$$

By rearraging the sums for the weak form equation we get

$$\begin{aligned}
0 = &\sum_{k=2}^{K+1} \int_{-\infty}^{\infty} \dot{u}_k(t_i, x) \Big(\partial_t \bar{u}_k(t_i, x) - \partial_x \bar{u}_k(t_i, x) f'(u_k(t_i, x)) \\
&\qquad + B_k \bar{\xi}_k(t_i) \delta(x - \xi_k(t_i)) + C_k \bar{\xi}_{k-1}(t_i) \delta(x - \xi_{k-1}(t_i)) \Big) \phi(x) \mathrm{d}x \\
&+ \int_{-\infty}^{\infty} \dot{u}_1(t_i, x) \Big(\partial_t \bar{u}_1(t_i, x) - \partial_x \bar{u}_1(t_i, x) f'(u_1(t_i, x)) + B_1 \bar{\xi}_1(t_i) \delta(x - \xi_1(t_i)) \Big) \phi(x) \mathrm{d}x \\
&+ \int_{-\infty}^{\infty} \dot{u}_{K+1}(t_i, x) \Big(\partial_t \bar{u}_{K+1}(t_i, x) - \partial_x \bar{u}_{K+1}(t_i, x) f'(u_{K+1}(t_i, x)) \\
&\qquad + C_{K+1} \bar{\xi}_K(t_i) \delta(x - \xi_K(t_i)) \Big) \phi(x) \mathrm{d}x \\
&+ \sum_{k=1}^{K} \Big(A_k \dot{\xi}_k(t_i) \bar{\xi}_k(t_i) + \dot{\xi}_k(t_i) \partial_t \bar{\xi}_k(t_i) \Big) \phi(\xi_k(t_i))
\end{aligned}$$

for all smooth test functions ϕ with compact support. The above equation generally holds if and only if

$$\partial_t \bar{\xi}_k(t_i) = -A_k \bar{\xi}_k(t_i)$$

for all $k \in \{1, \dots, K\}$,

$$0 = \int_{-\infty}^{\infty} \Big(\partial_t \bar{u}_k(t_i, x) - \partial_x \bar{u}_k(t_i, x) f'(u_k(t_i, x)) \\ + B_k \bar{\xi}_k(t_i) \delta(x - \xi_k(t_i)) + C_k \bar{\xi}_{k-1}(t_i) \delta(x - \xi_{k-1}(t_i)) \Big) \phi(x) \mathrm{d}x$$

for all $k \in \{2, \dots, K\}$ and

$$0 = \int_{-\infty}^{\infty} \Big(\partial_t \bar{u}_1(t_i, x) - \partial_x \bar{u}_1(t_i, x) f'(u_1(t_i, x)) + B_1 \bar{\xi}_1(t_i) \delta(x - \xi_1(t_i)) \Big) \phi(x) \mathrm{d}x,$$
$$0 = \int_{-\infty}^{\infty} \Big(\partial_t \bar{u}_{K+1}(t_i, x) - \partial_x \bar{u}_{K+1}(t_i, x) f'(u_{K+1}(t_i, x)) + C_{K+1} \bar{\xi}_K(t_i) \delta(x - \xi_K(t_i)) \Big) \phi(x) \mathrm{d}x$$

for all smooth test functions ϕ with compact support. That means the adjoint ODE of the proposition is satisfied and the adjoint PDE of the proposition is satisfied in the weak sense. This concludes the proof. □

5.4.4 Initial Data as a Function of the Parameter

We now give the adjoint sensitivity of the parameter p based on the adjoint sensitivity of the weak solution in the structure of Assumption 6.

Proposition 31. The adjoint sensitivity

$$\bar{p} \equiv \Big(\mathrm{d}_p \big(u_1(0, \cdot), \xi_1(0), \dots, \xi_K(0), u_{K+1}(0, \cdot) \big) \Big)^* \big(\bar{u}_1(0, \cdot), \bar{\xi}_1(0), \dots, \bar{\xi}_K(0), \bar{u}_{K+1}(0, \cdot) \big)$$

of the parameter $p \in \mathbb{R}^d$ is

$$\bar{p}_i = \sum_{k=1}^{K+1} \int_{\xi_{k-1}(0)}^{\xi_k(0)} \mathrm{d}_{p_i} u_k(0, x, p_0) \bar{u}_k(0, x) \mathrm{d}x + \sum_{k=1}^{K} \mathrm{d}_{p_i} \xi_k(0, p_0) \bar{\xi}_k(0) \; .$$

for all $i \in \{1, \dots, d\}$.

Proof. By the definition of the adjoint operator we have

$$\begin{aligned} &\Big\langle \big(\dot{u}_1(0, \cdot), \dot{\xi}_1(0), \dots, \dot{\xi}_k(0), \dot{u}_{K+1}(0, \cdot) \big), \big(\bar{u}_1(0, \cdot), \bar{\xi}_1(0), \dots, \bar{\xi}_k(0), \bar{u}_{K+1}(0, \cdot) \big) \Big\rangle \\ &= \sum_{k=1}^{K+1} \int_{\xi_{k-1}(0)}^{\xi_k(0)} \dot{u}_k(0, x) \bar{u}_k(0, x) \mathrm{d}x + \sum_{k=1}^{K} \dot{\xi}_k(0) \bar{\xi}_k(0) \\ &= \sum_{k=1}^{K+1} \int_{\xi_{k-1}(0)}^{\xi_k(0)} \langle \mathrm{d}_p u_k(0, x), \dot{p} \rangle \bar{u}_k(0, x) \mathrm{d}x + \sum_{k=1}^{K} \langle \mathrm{d}_p \xi_k(0), \dot{p} \rangle \bar{\xi}_k(0) \\ &= \sum_{k=1}^{K+1} \int_{\xi_{k-1}(0)}^{\xi_k(0)} \Big(\sum_{i=1}^{d} \mathrm{d}_{p_i} u_k(0, x) \dot{p}_i \Big) \bar{u}_k(0, x) \mathrm{d}x + \sum_{k=1}^{K} \sum_{i=1}^{d} \Big(\mathrm{d}_{p_i} \xi_k(0) \dot{p}_i \Big) \bar{\xi}_k(0) \\ &= \sum_{i=1}^{d} \dot{p}_i \sum_{k=1}^{K+1} \int_{\xi_{k-1}(0)}^{\xi_k(0)} \mathrm{d}_{p_i} u_k(0, x) \bar{u}_k(0, x) \mathrm{d}x + \sum_{i=1}^{d} \dot{p}_i \sum_{k=1}^{K} \mathrm{d}_{p_i} \xi_k(0) \bar{\xi}_k(0) \end{aligned}$$

$$= \left\langle \dot{p}, \Big(\sum_{k=1}^{K+1} \int_{\xi_{k-1}(0)}^{\xi_k(0)} \mathrm{d}_{p_i} u_k(0,x) \bar{u}_k(0,x) \mathrm{d}x + \sum_{k=1}^{K} \mathrm{d}_{p_i} \xi_k(0) \bar{\xi}_k(0) \Big)_{i=1,\ldots,d} \right\rangle$$
$$= \langle \dot{p}, \bar{p} \rangle \ .$$

This concludes the proof. □

We gave the derivation of the generalized tangent and adjoint sensitivity analysis of an assumed structure for the piecewise smooth solution of a scalar conservation law. The analytic sensitivity analysis performed in this chapter forms the basis of the discrete tangent and adjoint sensitivity analysis performed in the next chapter via numerical approximations of the shock location sensitivities, broad tangent sensitivites and integral objectives that explicitly take into consideration the numerical viscosity of shock capturing methods.

Chapter 6

Numerical Methods for Shock Sensitivities in Scalar Conservation Laws

The work presented in this chapter is in parts based on research previously published in:

M. Herty, J. Hüser, U. Naumann, T. Schilden, and W. Schröder. "Algorithmic differentiation of hyperbolic flow problems". In: *Journal of Computational Physics* 430 (2021), p. 110110

In this chapter we introduce a discrete sensitivity analysis for the shock locations in the discontinuous solution of a scalar conservation law. In order to keep the numerical analysis simple we assume the same solution structure as in the previous chapter where the exact analytic weak solution is piecewise smooth with all discontinuities as shock discontinuities and the shock location curves do not intersect over time.

The analysis we present does not require a specific integrator but instead states some assumptions about the numerical approximation and discrete tangents that should be satisfied by most shock capturing integrators. After stating the basic assumptions about the convergence of the numerical approximation of the weak solution we describe a method for simulating the shock locations with a given accuracy bound. After stating assumptions about the convergence of the discrete tangent of the weak solution we describe a method for simulating the shock sensitivities based on the Rankine-Hugoniot condition in a way such that they are convergent as the discretization is refined. We present how the discrete shock location sensitivities can be used to get a numerical version of the generalized tangent sensitivity described in the previous section. We also show how to perform a discrete adjoint sensitivity analysis for an integral objective based on the discrete shock sensitivities. Finally, we show how our approach for the numerical simulation of the discrete sensitivies can be implemented using AD tools and also present various computational results.

6.1 Basic Assumptions

In this section we state the basic assumptions with respect to the convergence of the numerical approximation of the weak solution, i.e. the numerical solution. Based on the Lax-Friedrichs discretizations of an example we demonstrate that the assumptions are empirically valid and reasonable. We also discuss the properties of the analytic viscosity traveling wave of Section 3.3.2 in this context as we know from Section 1.3.4 that it qualitatively captures the properties of the numerical viscosity of the Lax-Friedrichs scheme.

6.1.1 Convergence of the Weak Solution

The first assumption we make is that the numerical solution is pointwise sublinearly convergent outside of the shock regions and bounded inside of the shock regions as the discretization is refined.

Assumption 8. For some $\alpha \in (0,1)$ let

$$U(t,x) - u(t,x,p_0) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all

$$x \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} \big(\xi_k(t,p_0) - C_s\Delta x^\alpha, \xi_k(t,p_0) + C_s\Delta x^\alpha\big)$$

and let

$$U(t,x) - u(t,x,p_0) = O(1)$$

as $\Delta x \to 0$ for all

$$x \in \bigcup_{k=1}^{K} \big(\xi_k(t,p_0) - C_s\Delta x^\alpha, \xi_k(t,p_0) + C_s\Delta x^\alpha\big)$$

for all $t \in [0,T]$ and some $C_s \geq 2C_t$ where $\Delta t = C_t \Delta x$.

We use the notation of Section 3.2 where in this case $U(t,x)$ refers to the piecewise constant interpolation of the numerical solution U.

In order to demonstrate that the assumption is satisfied under reasonable circumstances we analyze the empirical convergence for one of our example problems.

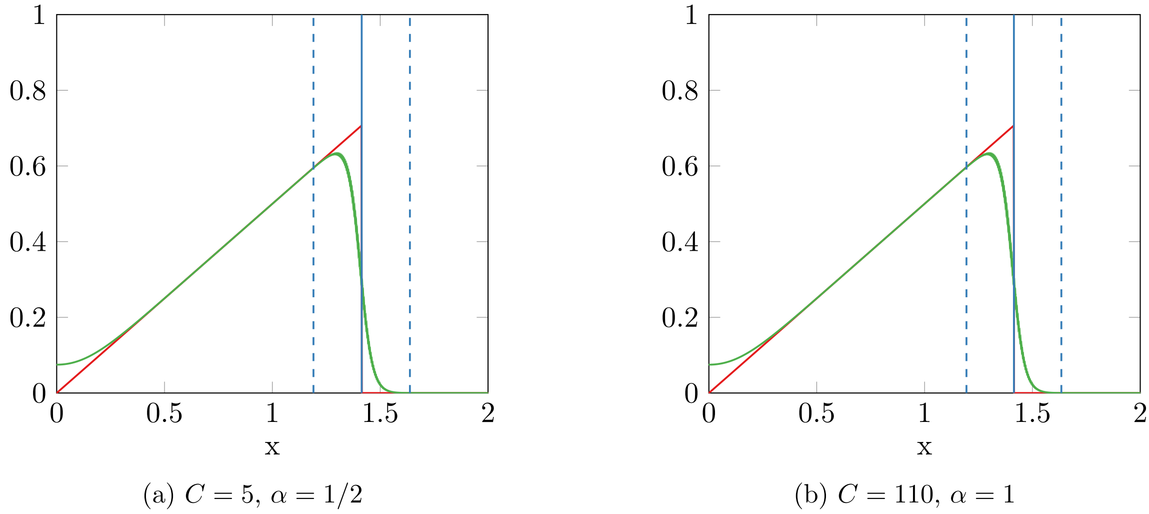
Example 35 (Ramp)**.** We consider the Lax-Friedrichs discretization of the ramp initial value problem from Example 7.

Figure 6.1 shows the analytic solution at $t = 1$ and with $p = 0$ in red and a Lax-Friedrichs discretization in green. One of the first things we see is that the numerical solution seems to be off near the boundary at $x = 0$. That error is a side effect of how the nonexistence of a boundary condition is handled in the simulation and can be ignored because it converges linearly. Specifically, the error is a result of not choosing a large enough domain for the space grid and instead copying wrong values into the boundary cells. As we see in the next figure the error at the boundary visibly vanishes as $\Delta x \to 0$ and in fact it vanishes at a linear rate.

Figure 6.2 shows the same graphs for a finer discretization grid. The Lax-Friedrichs discretization in the figures uses a low choice of C_t for $\Delta t = C_t\Delta x$, i.e. the solutions introduce a high amount of numerical viscosity in order to better visualize a significant width of the shock region where the shock is smoothed out. In Section 6.6.1 we also discuss the empirical convergence for the Upwind discretization and a Lax-Friedrichs discretization with less numerical viscosity.

The error between the analytic solution and the Lax-Friedrichs numerical solution is bounded in the shock region as $\Delta x \to 0$ since the numerical solution stays bounded as $\Delta x \to 0$.

The dashed blue lines show the boundaries of the shock regions for different choices of α and C. In Figure 6.1 we see that for $\alpha = 1$ a larger choice of C can get us far enough out of the shock region for a certain value of Δx. But in Figure 6.2 we see that as $\Delta x \to 0$ for $\alpha = 1/2$ we are clearly moving away from the shock region (as described in Section 1.4.3), whereas, for $\alpha = 1$ the distance from the shock location relative to the smoothing region stays constant.



(a) $C = 5$, $\alpha = 1/2$

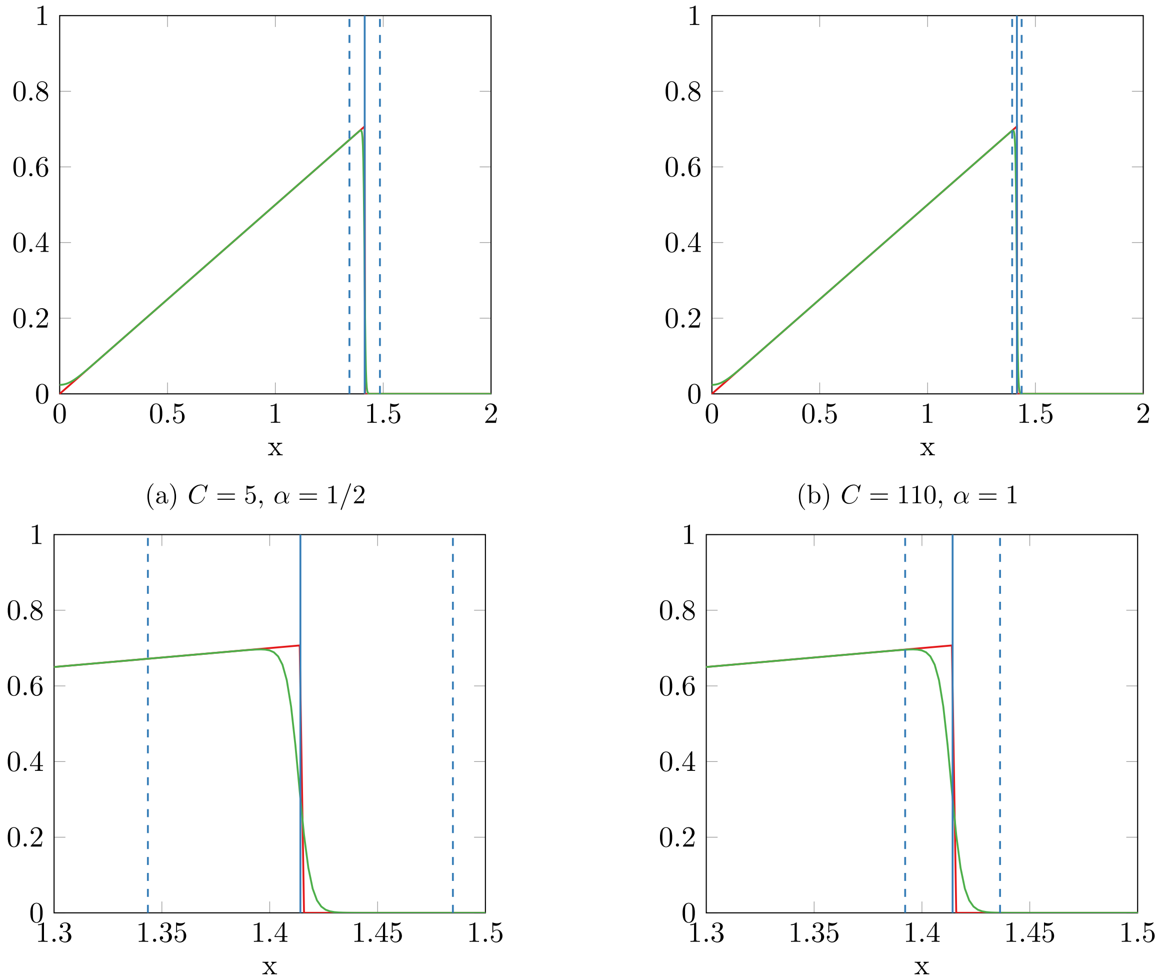
(b) $C = 110$, $\alpha = 1$

Figure 6.1: analytic (red), Lax-Friedrichs U (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed) ($\Delta x = 2 \cdot 10^{-3}$)

In Figure 6.3 we analyze the empirical convergence of the numerical approximation error at the boundary of the shock region, i.e. at the points marked by the dashed blue lines in Figures 6.1 and 6.2, for geometrically decreasing values of Δx and different combinations of values of C and α. The rate Δx^α is shown by the purple graph in each figure. For $\alpha = 1/2$ and $\alpha = 3/4$ we observe the empirical convergence of both $U(\xi_1 + C\Delta x^\alpha)$ and $U(\xi_1 - C\Delta x^\alpha)$ and for $\alpha = 1$ we observe that they both do not converge.

The nonconvergence of $U(\xi_1 \pm C\Delta x^\alpha)$ to $u(\xi_1\pm)$ for $\alpha = 1$ was explained in Section 1.4.3. A similar argument holds for the nonconvergence of $U(\xi_1 + C\Delta x^\alpha)$ to $u(\xi_1 + C\Delta x^\alpha)$ for $\alpha = 1$. Intuitively, we do not move further away from the smoothing as $\Delta x \to 0$, i.e. the error that is due to the numerical viscosity in the shock region stays constant. The size of the error depends on the constant C. Choosing C large enough can make the error small enough so as to be negligible for reasonable values of Δx even for $\alpha = 1$ especially when the numerical viscosity is lower than in these examples where it was purposefully chosen much larger than is required to obtain stability according to the CFL condition.

For $\alpha = 1/2$ and $\alpha = 3/4$ we observe that $U(\xi_1 + C\Delta x^\alpha)$ converges much faster than $U(\xi_1 - C\Delta x^\alpha)$. The reason for the fast convergence to the right of the shock is that the solution is constant $u(x) = 0$ for all $x > \xi_1$ and the numerical method is actually exact for the solution on



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 5$, $\alpha = 1/2$

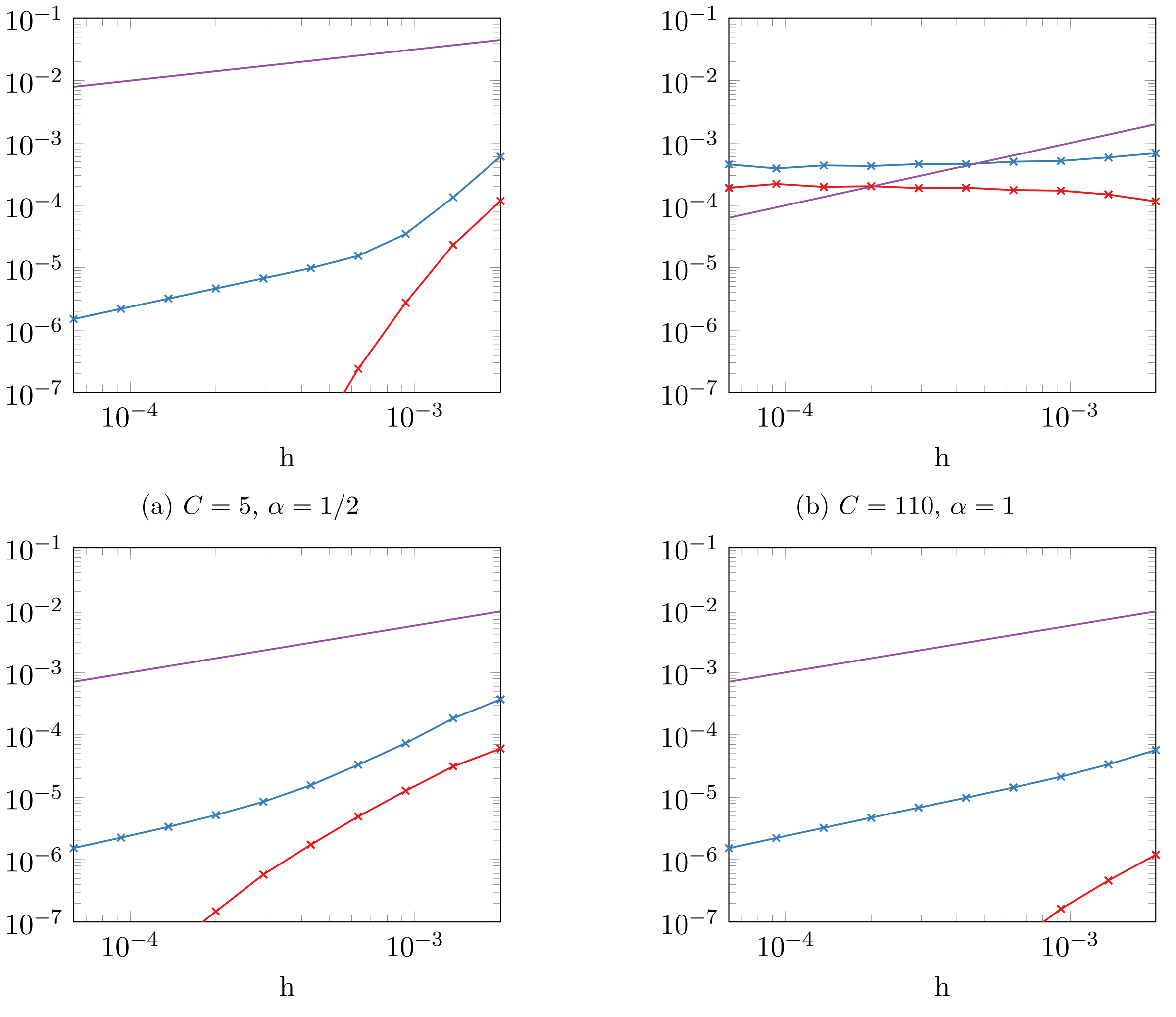
(d) $C = 110$, $\alpha = 1$

Figure 6.2: analytic (red), Lax-Friedrichs U (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed) ($\Delta x = 2 \cdot 10^{-4}$)

that region. But left of the shock the solution is not constant and in the empirical convergence to the left of the shock we can actually observe two phases of convergence behavior especially for $\alpha = 1/2$.

At first the error graph in blue falls quickly for higher values of Δx. The reason for the quick error decay at the beginning is that $\xi_1 - C\Delta x^\alpha$ is moving out of the smoothed shock region and the error due to the smoothing is higher than the approximation error here. Later, where Δx is small enough, $\xi_1 - C\Delta x^\alpha$ is far enough away from ξ_1 relative to to the smoothing region and the approximation error of the numerical method dominates.

For $\alpha = 3/4$ and $C = 25$ we observe a comparable behavior with two phases of convergence in the blue graph. For $\alpha = 3/4$ and $C = 35$ the evaluation starts far enough outside of the shock region due to the higher value of C such that we do not observe the two phases of convergence on the given range of Δx. Instead, for $C = 35$ we observe that for higher values of Δx the approximation error starts out lower. But as Δx becomes smaller the approximation error with $C = 35$ is actually similar to the error with $C = 25$ for $\alpha = 3/4$.



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

(d) $C = 35$, $\alpha = 3/4$

Figure 6.3: $|U(\xi_1 + C\Delta x^\alpha) - u(\xi_1 + C\Delta x^\alpha)|$ (red), $|U(\xi_1 - C\Delta x^\alpha) - u(\xi_1 - C\Delta x^\alpha)|$ (blue), Δx^α (purple)

Based on Assumption 8 we have the L_1 convergence of the weak solution at the same sublinear convergence rate depending on α.

Proposition 32.

$$\|U(t,\cdot) - u(t,\cdot,p_0)\|_{L_1(\hat{\Omega})} = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all $t \in [0,T]$.

Proof. Let

$$S = \bigcup_{k=1}^{K} \left(\xi_k(t,p_0) - C_s\Delta x^\alpha, \xi_k(t,p_0) + C_s\Delta x^\alpha\right) .$$

By Assumption 8 we have

$$\int_{\hat{\Omega}\backslash S} |U(t,x) - u(t,x,p_0)| \mathrm{d}x = \int_{\hat{\Omega}\backslash S} O(\Delta x^\alpha) \mathrm{d}x$$

$$
\begin{aligned}
&= O(\Delta x^\alpha) \int_{\hat{\Omega}\setminus S} \mathrm{d}x \\
&\leq O(\Delta x^\alpha)(R-L) = O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$. We also have

$$
\begin{aligned}
\int_S |U(t,x) - u(t,x,p_0)|\mathrm{d}x &= \int_S O(1)\mathrm{d}x \\
&= O(1) \int_S \mathrm{d}x \\
&= O(1) \sum_{k=1}^{K} 2C_s\Delta x^\alpha \\
&= O(1) 2KC_s\Delta x^\alpha = O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$.

Since $\hat{\Omega} = \hat{\Omega} \cup (\hat{\Omega} \setminus S)$ and by definition of the L_1 norm we have

$$
\begin{aligned}
\|U(t,\cdot) - u(t,\cdot,p_0)\|_{L_1(\hat{\Omega})} &= \int_{\hat{\Omega}} |U(t,x) - u(t,x,p_0)|\mathrm{d}x \\
&= \int_{\hat{\Omega}\setminus S} |U(t,x) - u(t,x,p_0)|\mathrm{d}x + \int_S |U(t,x) - u(t,x,p_0)|\mathrm{d}x \\
&= O(\Delta x^\alpha) + O(\Delta x^\alpha) = O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$. This concludes the proof. □

In practice and according to the theory in Teng and Zhang [TZ97] and Tang and Teng [TT97] we actually observe first-order L_1 convergence for monotone numerical methods for conservation laws with nonlinear flux.

6.1.2 Convergence for the Analytic Viscosity Solution

We now show that for the analytic traveling wave solution u^ϵ of the viscous Burgers equation introduced in Section 3.3.2 the sublinear pointwise convergence outside of the shock region of Assumption 8 holds. A general proof for viscous traveling wave solutions for scalar conservation laws with convex flux functions could potentially form the basis for a proof of Assumption 8 for general monotone discretization schemes (c.f. the approach based on discrete traveling waves in Teng and Zhang [TZ97]).

Proposition 33. Let u^ϵ be the traveling wave solution of Proposition 10 and u the solution of the parametric Riemann problem of Example 6 both with $p > -1$. Then

$$u^\epsilon(t, \xi(t) + C_s\Delta x^\alpha) - u(t, \xi(t) + C_s\Delta x^\alpha) = O(\Delta x^\alpha)$$

and

$$u^\epsilon(t, \xi(t) - C_s\Delta x^\alpha) - u(t, \xi(t) - C_s\Delta x^\alpha) = O(\Delta x^\alpha)$$

with

$$\xi(t) = \frac{1+p}{2}t$$

and

$$\epsilon = \frac{1}{2}\frac{\Delta x}{C_t}$$

for all $t \in [0, \infty)$ as $\Delta x \to 0$ holds for all $0 < \alpha \leq 1/2$. The proposition also holds for higher values of α but the proof is not explicitly given.

Proof. We first consider the error to the right of the shock. The analytic solution to the right of the shock is zero. We have

$$\begin{aligned}
&u^\epsilon(t, \xi(t) + C_s\Delta x^\alpha) - u(t, \xi(t) + C_s\Delta x^\alpha) \\
&\quad = \frac{1+p}{2} + \frac{1+p}{2}\tanh\left(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - \Big(\frac{1+p}{2}t + C_s\Delta x^\alpha\Big)\Big)\right) && \text{(definition of } u^\epsilon,\ u) \\
&\quad = \frac{1+p}{2} + \frac{1+p}{2}\tanh\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big) && \text{(definition of } \epsilon) \\
&\quad = \frac{1+p}{2} - \frac{1+p}{2}\left(1 - 2\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1}\right) && \text{(tanh in terms of exp)} \\
&\quad = (1+p)\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1} && \text{(basic algebra) .}
\end{aligned}$$

In order to show an error convergence rate of $O(\Delta x^\alpha)$ as $\Delta x \to 0$ with $p > -1$ we need to bound the norm of the error as follows

$$\begin{aligned}
&|u^\epsilon(t, \xi(t) + C_s\Delta x^\alpha) - u(t, \xi(t) + C_s\Delta x^\alpha)| \\
&\quad = \Big|(1+p)\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1}\Big| && \text{(norm both sides equation above)} \\
&\quad = (1+p)\Big|\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1}\Big| && (p > -1) \\
&\quad = (1+p)\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1} && (\exp(x) > 0 \text{ for all } x \in \mathbb{R}) \\
&\quad \leq K\Delta x^\alpha
\end{aligned}$$

for some $K > 0$. That error bound holds if and only if

$$1 + p \leq K\Delta x^\alpha\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)$$

as $\Delta x \to 0$ by multiplication of both sides of the last step by the reciprocal of the second factor on the left-hand side above.

Due to the convexity of the exponential function we have the lower bound

$$\exp(x) \geq 1 + x$$

by linearization at $x = 0$. If we use the linear lower bound of exp we can satisfy the error bound with $K\Delta x^\alpha$ if

$$\begin{aligned}
K\Delta x^\alpha\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big) &\geq K\Delta x^\alpha\Big(2 + C_tC_s(1+p)\Delta x^{\alpha-1}\Big) \\
&= 2K\Delta x^\alpha + (1+p)KC_tC_s\Delta x^{2\alpha-1} \\
&\geq 1 + p
\end{aligned}$$

as $\Delta x \to 0$. Since $2K\Delta x^\alpha \to 0$ as $\Delta x \to 0$ we need to have

$$KC_tC_s\Delta x^{2\alpha-1} > 1$$

as $\Delta x \to 0$. That bound is satisfied by some $K > 0$ if

$$\Delta x^{2\alpha-1} > 0$$

as $\Delta x \to 0$, i.e. if $2\alpha - 1 \leq 0$. Thus, with $\alpha \leq 1/2$ we have

$$u^\epsilon(t, \xi(t) + C_s\Delta x^\alpha) - u(t, \xi(t) + C_s\Delta x^\alpha) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$. With a tighter lower bound on exp higher values of α are possible, e.g. a quadratic lower bound gives $3\alpha - 2 \leq 0$, i.e. $\alpha \leq 2/3$ and so on.

The error to the left of the shock is analogous due to the symmetry with respect to the shock location. The analytic solution to the left of the shock has the value $1 + p$.

$$\begin{aligned}
&u^\epsilon(t, \xi(t) - C_s\Delta x^\alpha) - u(t, \xi(t) - C_s\Delta x^\alpha) \\
&\quad = \frac{1+p}{2} + \frac{1+p}{2}\tanh\left(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - \Big(\frac{1+p}{2}t - C_s\Delta x^\alpha\Big)\Big)\right) - (1+p) && \text{(definition of } u^\epsilon,\ u\text{)} \\
&\quad = (1+p)\Big(1 + \exp\Big(-C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1} - (1+p) && \text{(analogous steps)} \\
&\quad = -(1+p)\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1} && \text{(basic algebra) .}
\end{aligned}$$

Again, with $p > -1$ we need to bound the norm of the error as follows

$$\begin{aligned}
&|u^\epsilon(t, \xi(t) - C_s\Delta x^\alpha) - u(t, \xi(t) - C_s\Delta x^\alpha)| \\
&\quad = \Big| - (1+p)\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1}\Big| && \text{(norm both sides equation above)} \\
&\quad = \Big|(1+p)\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1}\Big| && \text{(definition of norm)} \\
&\quad = (1+p)\Big(1 + \exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1} && \text{(analogous steps)} \\
&\quad \leq K\Delta x^\alpha
\end{aligned}$$

for some $K > 0$. The error bound holds if and only if the bound on the right of the shock holds as they are precisely the same error bound. Thus, with $\alpha \leq 1/2$ we also have

$$u^\epsilon(t, \xi(t) - C_s\Delta x^\alpha) - u(t, \xi(t) - C_s\Delta x^\alpha) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$. This concludes the proof. □

6.1.3 Convergence of the Space Derivative

The second assumption we make is that the finite difference with respect to the space grid of the numerical solution is pointwise sublinearly convergent to the derivative with respect to x of the exact weak solution outside of the shock regions as the discretization is refined.

Assumption 9. For some $\alpha \in (0,1)$ let

$$\frac{U(t,x+\Delta x) - U(t,x-\Delta x)}{2\Delta x} - \partial_x u(t,x,p_0) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all

$$x \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} \big(\xi_k(t,p_0) - C_s\Delta x^\alpha, \xi_k(t,p_0) + C_s\Delta x^\alpha\big)$$

for all $t \in [0,T]$ and some $C_s \geq 2C_t$ where $\Delta t = C_t\Delta x$.

Again, we use the notation of Section 3.2 where in this case $U(t,x)$ refers to the piecewise constant interpolation of the numerical solution U.

In order to demonstrate that the assumption is satisfied under reasonable circumstances we analyze the empirical convergence of the finite difference for one of our example problem instances.

Example 36 (Ramp)**.** We consider the Lax-Friedrichs discretization of the ramp initial value problem from Example 7.

Analogous to Figure 6.1 for the weak solution Figure 6.4 shows the derivative with respect to space of the analytic solution at $t = 1$ and with $p = 0$ in red and the Lax-Friedrichs finite difference in green. Again, we note that the numerical solution finite difference has a significant error near the boundary at $x = 0$ that visibly vanishes as $\Delta x \to 0$ as we see in the next figure. The error is a result of how the nonexistant boundary condition is handled on the grid and can be ignored.

Figure 6.4 shows the same graphs for a finer discretization grid. In both figures the analytical space derivative is actually singular at the shock location, i.e. $\partial_x u(\xi_1) = -\infty$. The singularity is not always visible in the graphs due to the resolution of the plotting grid.

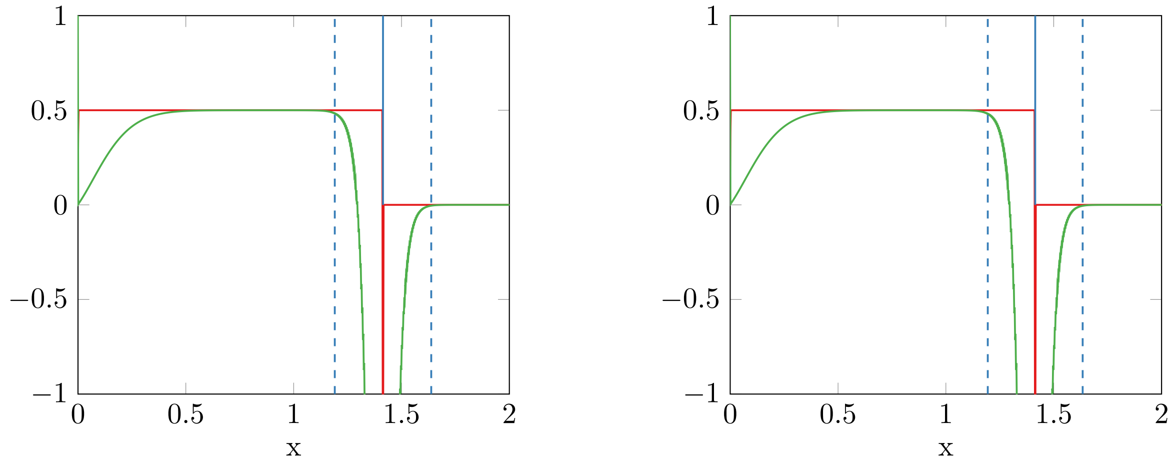
Identical to the previous example for the weak solution value the Lax-Friedrichs discretization uses a higher amount of numerical viscosity than necessary for stability. In Section 6.6.1 we also discuss the Upwind discretization and a Lax-Friedrichs discretization with less numerical viscosity.

The dashed blue lines show the boundaries of the shock regions for different choices of α and C. In Figure 6.5 we see even more clearly than in Figure 6.2 that for $\alpha = 1$ even a large choice of C does not move the point of evaluation $\xi_1 \pm C\Delta x^\alpha$ far enough outside of the shock region as $\Delta x \to 0$. But for $\alpha = 1/2$ we clearly observe that the point of evaluation goes into the region where in the case of the example the derivative with respect to space is constant. In order for $\partial_x u(\xi_1 \pm C\Delta x^\alpha)$ to converge to $\partial_x u(\xi_1\pm)$ smoothly outside of the asymptote it is necessary that the derivative with respect to space is Lipschitz continuous on a sufficiently large subset of the domain to the left and right of the shock location.

Analogous to Figure 6.3 in Figure 6.6 we analyze the empirical convergence of the numerical approximation error at the boundary of the shock region for geometrically decreasing values of Δx. The rate Δx^α is shown by the purple graph in each figure.

We note that since the analytic solution is piecewise linear we have

$$\partial_x u(\xi_1 + C\Delta x^\alpha) = \frac{u(\xi_1 + C\Delta x^\alpha + 2\Delta x) - u(\xi_1 + C\Delta x^\alpha)}{2\Delta x} \qquad \text{and}$$



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

Figure 6.4: analytic (red), Lax-Friedrichs $(U(x + \Delta x) - U(x - \Delta x))/(2\Delta x)$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed) $(\Delta x = 2 \cdot 10^{-3})$

$$\partial_x u(\xi_1 - C\Delta x^\alpha) = \frac{u(\xi_1 - C\Delta x^\alpha) - u(\xi_1 - C\Delta x^\alpha - 2\Delta x)}{2\Delta x}.$$

Even though the graph shows the error evaluated based on a finite difference of the analytic solution it also shows the error between the numerical approximation and the analytic derivative because of the above identities.

For $\alpha = 1/2$ and $\alpha = 3/4$ we observe the empirical convergence of the finite difference both to the left and right of the shock. For $\alpha = 1$ we actually observe that they both diverge, i.e. for $\alpha = 1$ the error actually grows as $\Delta x \to 0$.

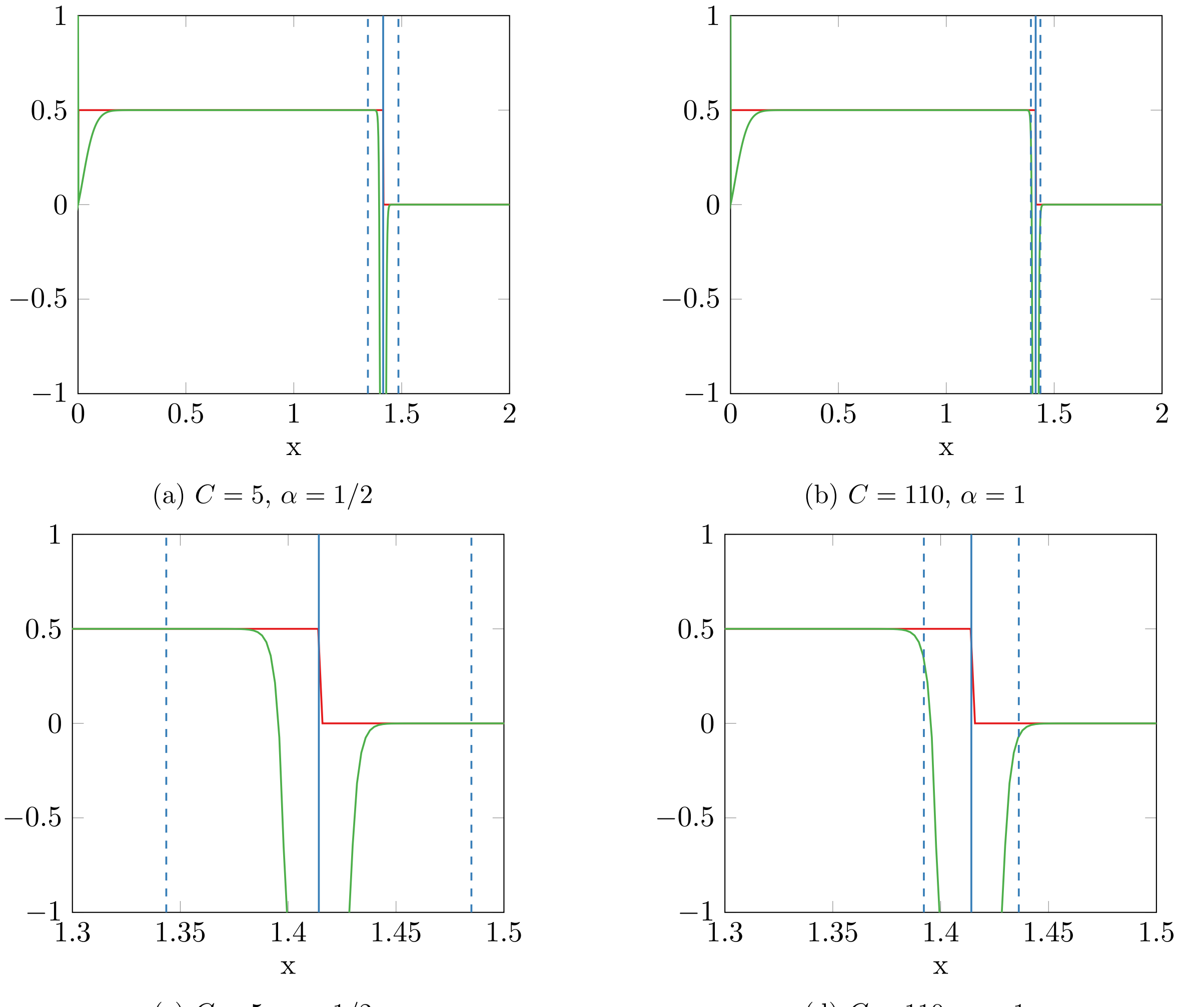
The divergence of the space derivative of the numerical viscosity solution for $\alpha = 1$ is related to the fact that the space derivative inside of the shock region diverges as $\Delta x \to 0$ because it smoothly approximates the singular derivative at $x = \xi$. Maintaining a constant distance to the shock location leads to constant error for the value of the weak solution since it is bounded but to diverging error for the derivative since it is unbounded as $\Delta x \to 0$. Because we utilize the finite difference approximations

$$\partial_x u(\xi_1+) \approx \frac{U(\xi_1 + C\Delta x^\alpha + 2\Delta x) - U(\xi_1 + C\Delta x^\alpha)}{2\Delta x} \qquad \text{and}$$
$$\partial_x u(\xi_1-) \approx \frac{U(\xi_1 - C\Delta x^\alpha) - U(\xi_1 - C\Delta x^\alpha - 2\Delta x)}{2\Delta x}$$

the discrete sensitivity analysis we present in this chapter formally requires a choice of $0 < \alpha < 1$ in order to be asymptotically convergent.

For $\alpha = 1/2$ and $\alpha = 3/4$ we again observe that convergence to the right of the shock is much faster than to the left of the shock. The reason is the same as for the weak solution value. The discrete approximation of the constant part to the right of the shock is exact whereas the discrete approximation to the left of the shock has approximation error.

We also again observe the two phases of convergence for $\alpha = 1/2$ and $\alpha = 3/4, C = 35$ where for high values of Δx convergence is fast because we are still moving out from the shock region. For lower values of Δx the actual numerical approximation error dominates the convergence



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 5$, $\alpha = 1/2$

(d) $C = 110$, $\alpha = 1$

Figure 6.5: analytic (red), Lax-Friedrichs $(U(x + \Delta x) - U(x - \Delta x))/(2\Delta x)$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed) ($\Delta x = 2 \cdot 10^{-4}$)

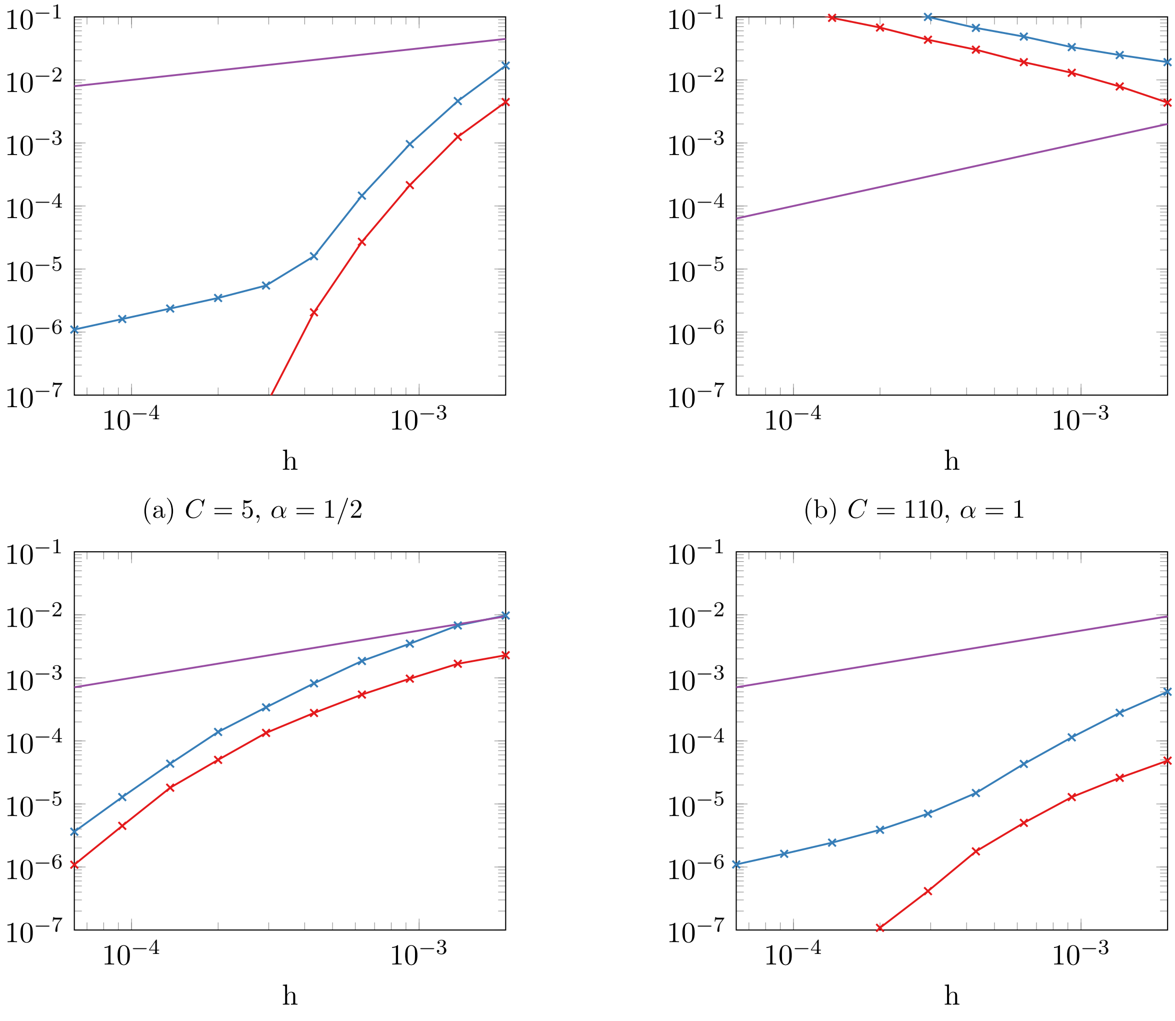
behavior. The lower choice of $C = 25$ for $\alpha = 3/4$ starts out with the points of evaluation being close enough to the shock location such that we do not enter the second phase of convergence behavior for the given range of Δx values. We observe that for $\alpha = 3/4, C = 35$ the difference in slope between the two phases of convergence is less pronounced than for $\alpha = 1/2$.

6.1.4 Convergence of the Space Derivative in the Analytic Viscosity Solution

We now show that for the analytic traveling wave solution u^ϵ of the viscous Burgers equation the sublinear pointwise convergence of the finite difference to the space derivative of Assumption 9 holds.

Proposition 34. Let u^ϵ be the traveling wave solution of Proposition 10 and u the solution of the parametric Riemann problem of Example 6 both with $p > -1$. Then

$$\frac{u^\epsilon(t, \xi(t) + C_s\Delta x^\alpha + 2\Delta x) - u^\epsilon(t, \xi(t) + C_s\Delta x^\alpha)}{2\Delta x} - \partial_x u(t, \xi(t) + C_s\Delta x^\alpha) = O(\Delta x^\alpha)$$



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

(d) $C = 35$, $\alpha = 3/4$

Figure 6.6: $|(U(\xi_1 + C\Delta x^\alpha + 2h) - U(\xi_1 + C\Delta x^\alpha))/(2\Delta x) - (u(\xi_1 + C\Delta x^\alpha + 2\Delta x) - u(\xi_1 + C\Delta x^\alpha))/(2\Delta x)|$ (red), $|(U(\xi_1 - C\Delta x^\alpha) - U(\xi_1 - C\Delta x^\alpha - 2h))/(2h) - (u(\xi_1 - C\Delta x^\alpha) - u(\xi_1 + C\Delta x^\alpha - 2\Delta x))/(2\Delta x)|$ (blue), Δx^α (purple)

and

$$\frac{u^\epsilon(t, \xi(t) - C_s\Delta x^\alpha) - u^\epsilon(t, \xi(t) - C_s\Delta x^\alpha - 2\Delta x)}{2\Delta x} - \partial_x u(t, \xi(t) - C_s\Delta x^\alpha) = O(\Delta x^\alpha)$$

with

$$\xi(t) = \frac{1+p}{2}t$$

and

$$\epsilon = \frac{1}{2}\frac{\Delta x}{C_t}$$

for all $t \in [0, \infty)$ as $\Delta x \to 0$ holds for all $\alpha \leq 1/2$. The proposition also holds for higher values of α but the proof is not explicitly given.

Proof. We first consider the error to the right of the shock. The derivative of the analytic solution with respect to space is zero everywhere except at $x = \xi$. We have

$$
\begin{aligned}
&\frac{u^\epsilon(t, \xi(t) + C_s \Delta x^\alpha + 2\Delta x) - u^\epsilon(t, \xi(t) + C_s \Delta x^\alpha)}{2\Delta x} - \partial_x u(t, \xi(t) + C_s \Delta x^\alpha) \\
&= \frac{1}{2\Delta x}\Bigg(\frac{1+p}{2} + \frac{1+p}{2} \tanh\Bigg(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - \Big(\frac{1+p}{2}t + C_s \Delta x^\alpha + 2\Delta x\Big)\Big)\Bigg) \\
&\qquad - \frac{1+p}{2} - \frac{1+p}{2} \tanh\Bigg(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - \Big(\frac{1+p}{2}t + C_s \Delta x^\alpha\Big)\Big)\Bigg)\Bigg) \qquad \text{(def. of } u^\epsilon,\ u) \\
&= \frac{1+p}{4\Delta x}\Bigg(\tanh\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1} + C_t(1+p)\Big) \\
&\qquad - \tanh\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big)\Bigg) \qquad \text{(def. of } \epsilon) \ .
\end{aligned}
$$

Due to its concave shape as $x \to \infty$ to the right of zero the tanh function can be upper bounded by its linearization, i.e. for all $x \geq x_0 > 0$ we have

$$
\tanh(x) \leq \tanh(x_0) + \tanh'(x_0)(x - x_0) = \tanh(x_0) + \operatorname{sech}^2(x_0)(x - x_0) \ .
$$

Consequently, since $p > -1$ we have

$$
\begin{aligned}
&\tanh\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1} + C_t(1+p)\Big) \\
&\quad \leq \tanh\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) + \operatorname{sech}^2\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) C_t (1+p) \ .
\end{aligned}
$$

By substitution in the equation above and writing sech^2 in terms of exp we get

$$
\begin{aligned}
&\frac{u^\epsilon(t, \xi(t) + C_s \Delta x^\alpha + 2\Delta x) - u^\epsilon(t, \xi(t) + C_s \Delta x^\alpha)}{2\Delta x} - \partial_x u(t, \xi(t) + C_s \Delta x^\alpha) \\
&\leq C_t \frac{(1+p)^2}{4\Delta x} \operatorname{sech}^2\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) \\
&= C_t \frac{(1+p)^2}{\Delta x} \Big(\exp\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) + \exp\Big(-\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big)\Big)^{-2} \ .
\end{aligned}
$$

Due to its strongly convex shape as $x \to \infty$ to the right of zero the exponential function can be lower bounded by a quadratic function based on its second-order Taylor expansion at $x = 0$, i.e. for all $x \geq 0$ we have

$$
\exp(x) \geq 1 + x + \frac{1}{2}x^2 \ .
$$

Consequently, since $p > -1$ we have

$$
\exp\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) \geq 1 + \frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1} + \frac{1}{2}\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big)^2 \ .
$$

Due to the global convexity of the exponential function we also have the lower bound

$$
\exp\Big(-\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) \geq 1 - \frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}
$$

by linearization at $x = 0$.

By adding the two lower bounds we have

$$\begin{aligned}
&\exp\Big(\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)+\exp\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\\
&\geq 2+\frac{1}{2}\Big(\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)^2\\
&=2+\frac{(1+p)^2}{8}C_t^2C_s^2\Delta x^{2\alpha-2}\,.
\end{aligned}$$

Due to the monotonicity of $x\mapsto x^2$ for $x>0$ we have

$$\begin{aligned}
&\Big(\exp\Big(\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)+\exp\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^2\\
&\geq\Big(2+\frac{(1+p)^2}{8}C_t^2C_s^2\Delta x^{2\alpha-2}\Big)^2\\
&=4+\frac{(1+p)^2}{4}C_t^2C_s^2\Delta x^{2\alpha-2}+\frac{(1+p)^4}{64}C_t^4C_s^4\Delta x^{4\alpha-4}\,.
\end{aligned}$$

In order to show an error convergence rate of $O(\Delta x^\alpha)$ as $\Delta x\to 0$ with $p>-1$ we need to bound the norm of the error as follows

$$\begin{aligned}
&\left|\frac{u^\epsilon(t,\xi(t)+C_s\Delta x^\alpha+2\Delta x)-u^\epsilon(t,\xi(t)+C_s\Delta x^\alpha)}{2\Delta x}-\partial_x u(t,\xi(t)+C_s\Delta x^\alpha)\right|\\
&=\left|C_t\frac{(1+p)^2}{\Delta x}\Big(\exp\Big(\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)+\exp\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-2}\right|\\
&=C_t\frac{(1+p)^2}{\Delta x}\Big(\exp\Big(\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)+\exp\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-2}\\
&\leq C_t\frac{(1+p)^2}{\Delta x}\Big(4+\frac{(1+p)^2}{4}C_t^2C_s^2\Delta x^{2\alpha-2}+\frac{(1+p)^4}{64}C_t^4C_s^4\Delta x^{4\alpha-4}\Big)^{-1}\\
&\leq C_t(1+p)^2\Big(4\Delta x+\frac{(1+p)^2}{4}C_t^2C_s^2\Delta x^{2\alpha-1}+\frac{(1+p)^4}{64}C_t^4C_s^4\Delta x^{4\alpha-3}\Big)^{-1}\\
&\leq K\Delta x^\alpha
\end{aligned}$$

for some $K>0$. The first step is taking the norm of the error derivation we gave above, the second step is due to $p>-1$ and exp positive, the third step is by $y\geq x>0$ implies $y^{-1}\leq x^{-1}$ in combination with the bound we derived above and the fourth step is by distributivity. The bound in the last step holds if and only if

$$\begin{aligned}
C_t(1+p)^2&\leq K\Delta x^\alpha\Big(4\Delta x+\frac{(1+p)^2}{4}C_t^2C_s^2\Delta x^{2\alpha-1}+\frac{(1+p)^4}{64}C_t^4C_s^4\Delta x^{4\alpha-3}\Big)\\
&\leq K4\Delta x^{\alpha+1}+K\frac{(1+p)^2}{4}C_t^2C_s^2\Delta x^{3\alpha-1}+K\frac{(1+p)^4}{64}C_t^4C_s^4\Delta x^{5\alpha-3}
\end{aligned}$$

by multiplication of both sides of the last step by the reciprocal of the second factor on the left-hand side above. The bound is satisfied by some $K>0$ if

$$\Delta x^{5\alpha-3}>0$$

as $\Delta x\to 0$, i.e. if $5\alpha-3\leq 0$. Thus, with $\alpha\leq 3/5$ we have

$$\frac{u^\epsilon(t,\xi(t)+C_s\Delta x^\alpha+2\Delta x)-u^\epsilon(t,\xi(t)+C_s\Delta x^\alpha)}{2\Delta x}-\partial_x u(t,\xi(t)+C_s\Delta x^\alpha)=O(\Delta x^\alpha)$$

as $\Delta x \to 0$. With tighter bounds on the exponential function higher values of α that still satisfy the bound are possible.

The proof for the derivative with respect to space to the left of the shock is analogous due to symmetry with respect to the shock location. This concludes the proof. □

6.1.5 Convergence of the Tangent Sensitivity

The third assumption we make is that the discrete tangent sensitivity is pointwise sublinearly convergent to the exact (broad) tangent sensitivity of the weak solution outside of shock regions as the discretization is refined.

Assumption 10. For some $\alpha \in (0, 1)$ let

$$\dot{U}(t, x) - \dot{u}_{\text{broad}}(t, x, p_0) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all

$$x \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} \left(\xi_k(t, p_0) - C_s \Delta x^\alpha, \xi_k(t, p_0) + C_s \Delta x^\alpha\right)$$

for all $t \in [0, T]$ and some $C_s \geq 2C_t$ where $\Delta t = C_t \Delta x$.

Here, the notation $\dot{U}(t, x)$ refers to the piecewise constant interpolation of the discrete tangent solution $\dot{U}$.

In order to demonstrate that the assumption is satisfied under reasonable circumstances we analyze the empirical convergence of the discrete tangent sensitivity for one of our example problem instances.

Example 37 (Ramp)**.** We consider the Lax-Friedrichs discretization of the ramp initial value problem from Example 7.

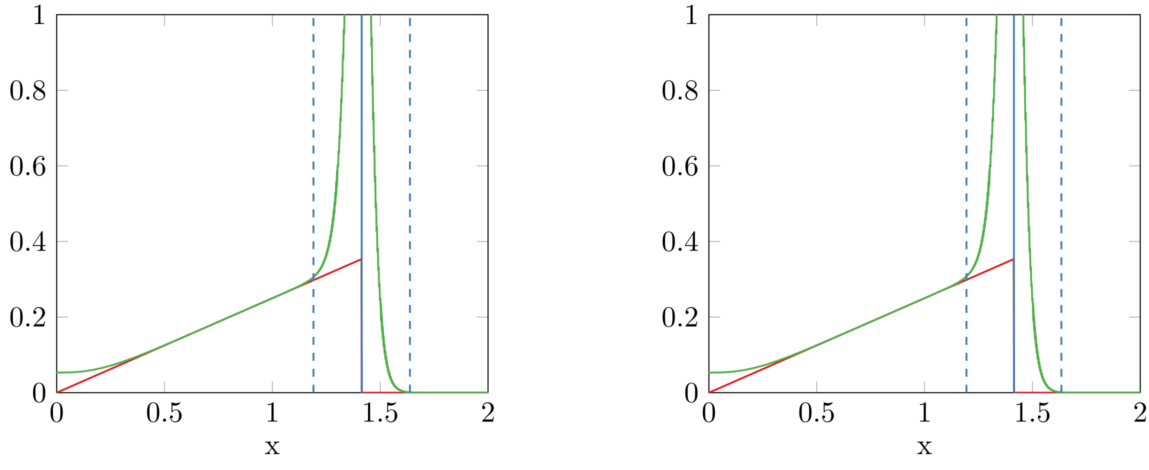
Analogous to Figure 6.1 for the weak solution Figure 6.7 shows the broad tangent sensitivity of the analytic solution at $t = 1$ and with $p = 0, \dot{p} = 1$ in red and the Lax-Friedrichs discrete tangent in green. Again, we note that the discrete tangent has a significant error near the boundary at $x = 0$ that visibly vanishes as $\Delta x \to 0$ as we see in the next figure. The error is a result of how the nonexistant boundary condition is handled on the grid and can be ignored because it converges linearly.

Figure 6.4 shows the same graphs for a finer discretization grid.

Identical to the previous example for the weak solution value the Lax-Friedrichs discretization uses a higher amount of numerical viscosity than necessary for stability. In Section 6.6.1 we also discuss the Upwind discretization and a Lax-Friedrichs discretization with less numerical viscosity.

Similar to the space derivative the distributional analytic tangent, but not the broad tangent, is singular at the shock location, i.e. $\dot{u}(\xi_1) = \infty$. As a result the discrete tangent approximates the singularity and is unbounded as $\Delta x \to 0$.

Like in the previous examples, the dashed blue lines show the boundaries of the shock regions for different choices of α and C. In Figure 6.8 we again clearly see that for $\alpha = 1$ even a large choice of C does not move the point of evaluation $\xi_1 \pm C\Delta x^\alpha$ far enough outside of the shock region as $\Delta x \to 0$. But for $\alpha = 1/2$ we observe that the point of evaluation goes into the region where we are actually evaluating an approximation of the broad tangent without the part that approximates the singularity.



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

Figure 6.7: analytic (red), Lax-Friedrichs $\dot{U}$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed) ($\Delta x = 2 \cdot 10^{-3}$)

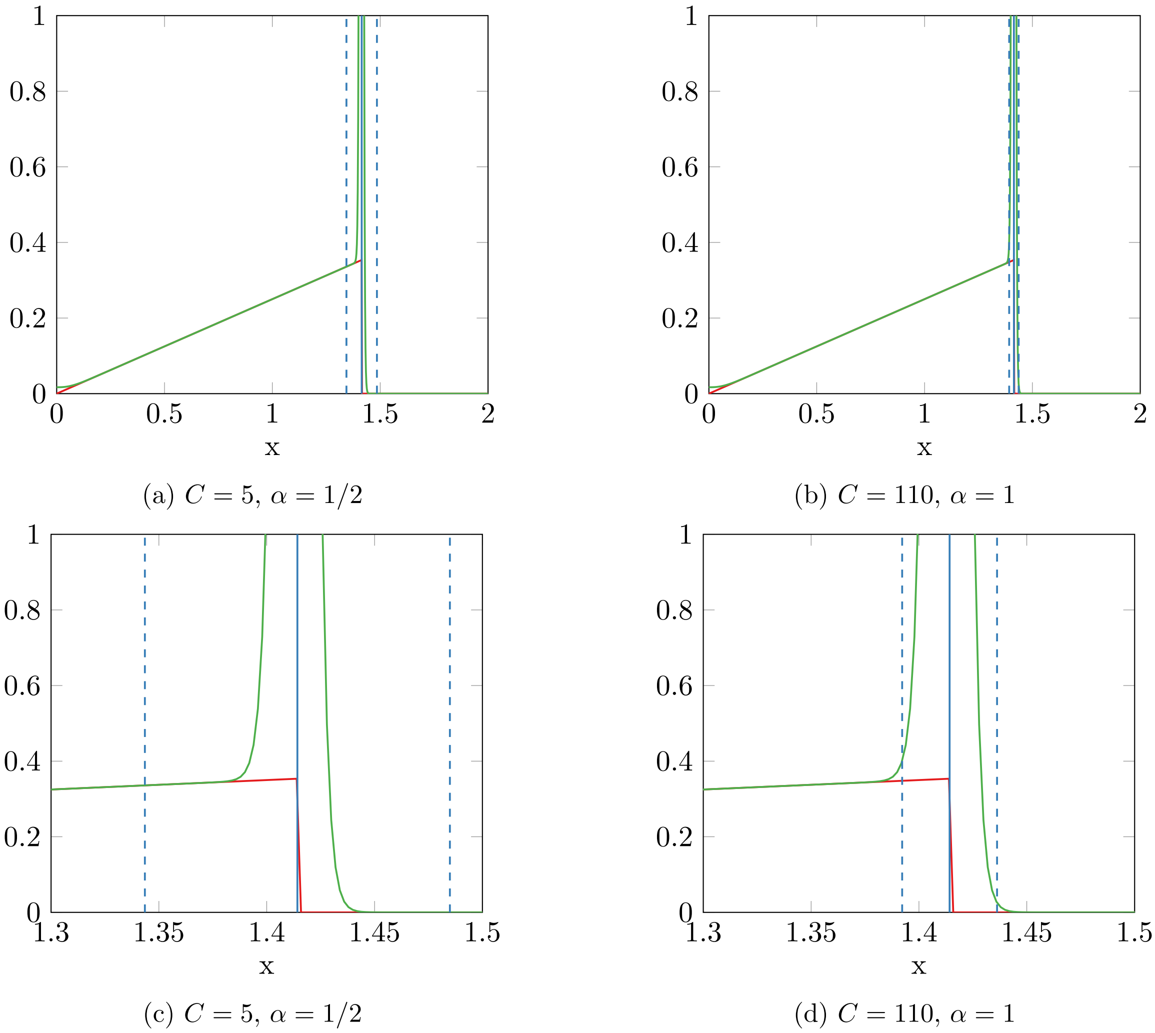
Analogous to Figure 6.3 in Figure 6.9 we analyze the empirical convergence of the numerical approximation error at the boundary of the shock region for geometrically decreasing values of Δx. The rate Δx^α is shown by the purple graph in each figure.

For $\alpha = 1/2$ and $\alpha = 3/4$ we observe the empirical convergence of the discrete tangent both to the left and right of the shock and for $\alpha = 1$ we observe that they both diverge, i.e. for $\alpha = 1$ the error grows as $\Delta x \to 0$, just like we observed for the space derivative previously. The divergence of the discrete tangent for $\alpha = 1$ has analogous reasons to the divergence of the finite difference space derivative approximation in the shock region. Because we utilize the discrete tangent approximations

$$\begin{aligned} \dot{u}(\xi_1+) &\approx \dot{U}(\xi_1 + C\Delta x^\alpha) \qquad \text{and} \\ \dot{u}(\xi_1-) &\approx \dot{U}(\xi_1 - C\Delta x^\alpha) \end{aligned}$$

the discrete sensitivity analysis we present in this chapter formally requires a choice of $0 < \alpha < 1$ in order to be asymptotically convergent.

For $\alpha = 1/2$ and $\alpha = 3/4$ we again observe the same types of convergence patterns and for the same reasons as before. Convergence to the right of the shock is faster because there the numerical solution and hence its discrete tangent is exact. Convergence is in two phases; first, a faster convergence from moving away from the shock relative to the width of the shock region and, second, a slower convergence due to the remaining approximation error. In order to flatten the fast convergence in the first phase and steepen the slower convergence in the second phase it is possible to choose a higher value of α. From the choice of larger C for $\alpha = 3/4$ we see that because we start out further outside of the shock region we enter the second phase of convergence faster (for higher values of Δx).



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 5$, $\alpha = 1/2$

(d) $C = 110$, $\alpha = 1$

Figure 6.8: analytic (red), Lax-Friedrichs $\dot{U}$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed) ($\Delta x = 2 \cdot 10^{-4}$)

6.1.6 Convergence of the Tangent in the Analytic Viscosity Solution

We now show that for the analytic traveling wave solution u^ϵ of the viscous Burgers equation the sublinear pointwise convergence of the tangent sensitivity to the broad tangent of Assumption 10 holds.

Proposition 35. Let u^ϵ be the traveling wave solution of Proposition 10 and u the solution of the parametric Riemann problem of Example 6 both with $p > -1$. Let the tangent sensitivity of u^ϵ be defined by

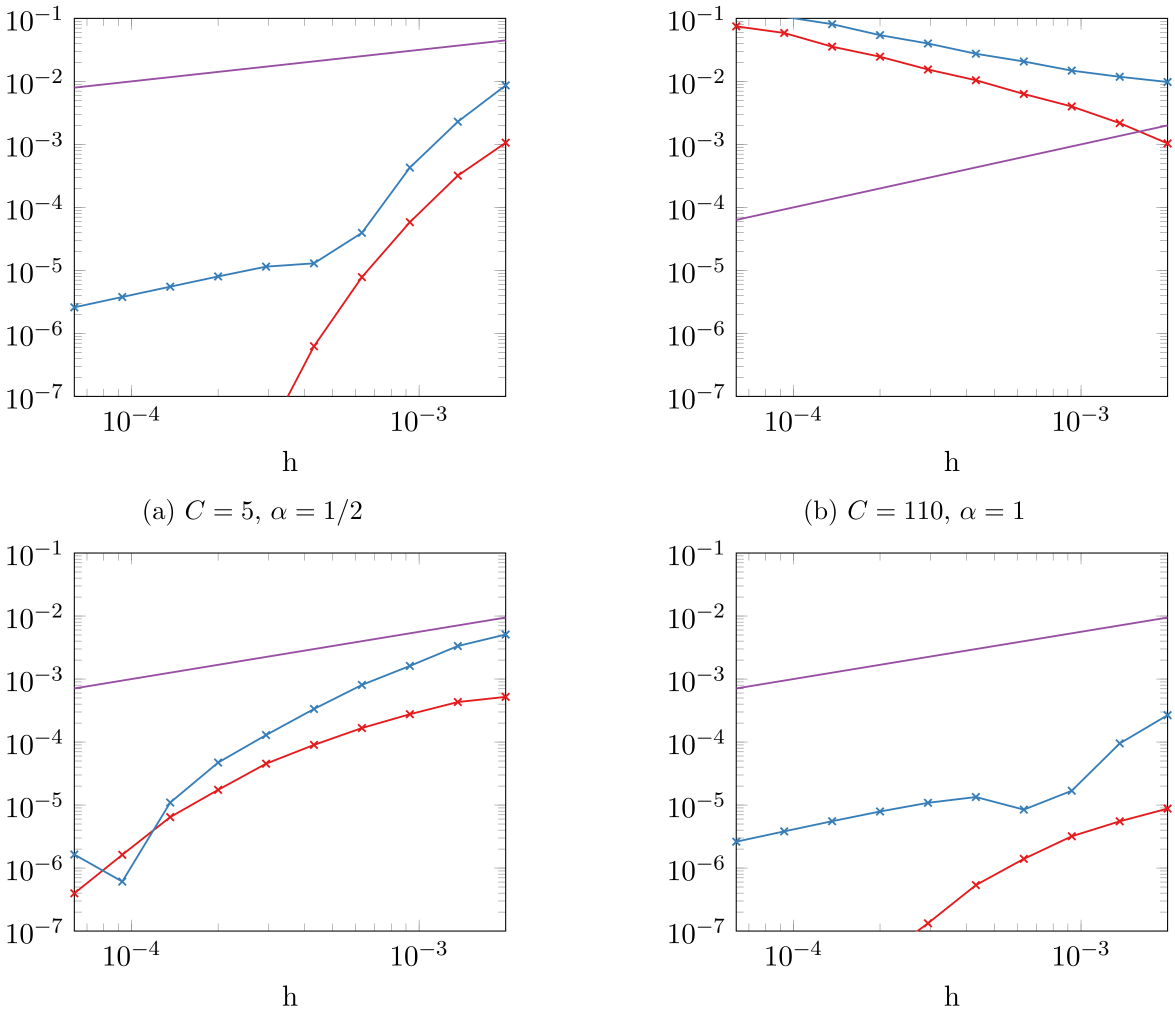
$$\dot{u}^\epsilon(t, x, p) = \mathrm{d}_p u^\epsilon(t, x, p)\dot{p}$$

for all $x \in \Omega$. Then

$$\dot{u}^\epsilon(t, \xi(t) + C_s\Delta x^\alpha, p) - \dot{u}(t, \xi(t) + C_s\Delta x^\alpha, p) = O(\Delta x^\alpha)$$

and

$$\dot{u}^\epsilon(t, \xi(t) - C_s\Delta x^\alpha, p) - \dot{u}(t, \xi(t) - C_s\Delta x^\alpha, p) = O(\Delta x^\alpha)$$



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

(d) $C = 35$, $\alpha = 3/4$

Figure 6.9: $|\dot{U}(\xi_1 + C\Delta x^\alpha) - \dot{u}(\xi_1 + C\Delta x^\alpha)|$ (red), $|\dot{U}(\xi_1 - C\Delta x^\alpha) - \dot{u}(\xi_1 - C\Delta x^\alpha)|$ (blue), Δx^α (purple)

with

$$\xi(t) = \frac{1+p}{2}t$$

and

$$\epsilon = \frac{1}{2}\frac{\Delta x}{C_t}$$

for all $t \in [0, \infty)$ as $\Delta x \to 0$ holds for all $\alpha \leq 1/2$. The proposition also holds for higher values of α but the proof is not explicitly given.

Proof. According to the chain rule we have

$$\begin{aligned} \mathrm{d}_p u^\epsilon(t,x,p) = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - x\Big)\right) \\ + \frac{1+p}{2}\operatorname{sech}^2\left(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - x\Big)\right)\left(\frac{1}{4\epsilon}\Big(\frac{1+p}{2}t - x\Big) + \frac{1+p}{8\epsilon}t\right). \end{aligned}$$

We first consider the error to the right of the shock location ξ. The analytic tangent sensitivity is zero for $x > \xi$. We have

$$\begin{aligned}
&\dot{u}^{\epsilon}(t,\xi(t)+C_s\Delta x^{\alpha}) - \dot{u}(t,\xi(t)+C_s\Delta x^{\alpha})\\
&\quad= \frac{1}{2}\dot{p} + \frac{1}{2}\tanh\left(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - \Big(\frac{1+p}{2}t + C_s\Delta x^{\alpha}\Big)\Big)\right)\dot{p}\\
&\qquad+ \frac{1+p}{2}\operatorname{sech}^2\left(\frac{1+p}{4\epsilon}\Big(\frac{1+p}{2}t - \Big(\frac{1+p}{2}t + C_s\Delta x^{\alpha}\Big)\Big)\right)\\
&\qquad\cdot\left(\frac{1}{4\epsilon}\Big(\frac{1+p}{2}t - \Big(\frac{1+p}{2}t + C_s\Delta x^{\alpha}\Big)\Big) + \frac{1+p}{8\epsilon}t\right)\dot{p} \qquad \text{(def. of } \dot{u}^{\epsilon},\ \dot{u})\\
&\quad= \frac{1}{2}\dot{p} + \frac{1}{2}\tanh\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\dot{p}\\
&\qquad+ \frac{1+p}{2}\operatorname{sech}^2\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big(-\frac{1}{2}C_tC_s\Delta x^{\alpha-1} + \frac{1+p}{4\Delta x}C_tt\Big)\dot{p}\,. \quad \text{(def. of } \epsilon)
\end{aligned}$$

In order to show an error convergence rate of $O(\Delta x^{\alpha})$ as $\Delta x \to 0$ with $p > -1$ we need to bound the norm of the error as follows

$$\begin{aligned}
&\left|\dot{u}^{\epsilon}(t,\xi(t)+C_s\Delta x^{\alpha}) - \dot{u}(t,\xi(t)+C_s\Delta x^{\alpha})\right|\\
&\quad= \left|\frac{1}{2}\dot{p} + \frac{1}{2}\tanh\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\dot{p}\right.\\
&\qquad\left.+ \frac{1+p}{2}\operatorname{sech}^2\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big(-\frac{1}{2}C_tC_s\Delta x^{\alpha-1} + \frac{1+p}{4\Delta x}C_tt\Big)\dot{p}\right|\\
&\quad\le |\dot{p}|\left|\frac{1}{2} + \frac{1}{2}\tanh\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\right|\\
&\qquad+ |\dot{p}|\left|\frac{1+p}{2}\operatorname{sech}^2\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big(-\frac{1}{2}C_tC_s\Delta x^{\alpha-1} + \frac{1+p}{4\Delta x}C_tt\Big)\right|\\
&\quad\le K\Delta x^{\alpha}
\end{aligned}$$

for some $K > 0$. The first step is taking the norm of the error derivation we gave above and the second step is via the triangle inequality.

We bound the two normed terms separately and ignore the factor $|\dot{p}|$ because it is constant and can be summarized into the constant K. For the first term we write tanh in terms of exp to get the following bound

$$\begin{aligned}
\left|\frac{1}{2} + \frac{1}{2}\tanh\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\right| &= \left|\Big(1+\exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1}\right|\\
&= \Big(1+\exp\Big(C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big)^{-1}\\
&\le \Big(2 + C_tC_s(1+p)\Delta x^{\alpha-1}\Big)^{-1}\,.
\end{aligned}$$

The second step is due to the positivity of exp and the third step is due to the convexity of exp. We have

$$\begin{aligned}\Big(2 + C_t C_s (1+p)\Delta x^{\alpha-1}\Big)^{-1} &\le K_1 \Delta x^{\alpha} \\ 1 &\le 2K_1 \Delta x^{\alpha} + K_1 C_t C_s (1+p) \Delta x^{2\alpha-1}\end{aligned}$$

for some $K_1 > 0$ if and only if $2\alpha - 1 \le 0$, or $\alpha \le 0$ which is not admissible. Thus, with $\alpha \le 1/2$ we have

$$\left| \frac{1}{2} + \frac{1}{2} \tanh\Big(-\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big)\right| \le K_1 \Delta x^{\alpha}$$

for some $K_1 > 0$ as $\Delta x \to 0$. Again, tighter bounds on the exponential function exist and lead to higher values of α for which the bound holds.

For the second term we write sech^2 in terms of exp to get

$$\begin{aligned}&\frac{1+p}{2} \operatorname{sech}^2 \Big(-\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) \\ &\quad = 2(1+p) \Big(\exp\Big(\frac{1}{2} C_t C_s (1+p)\Delta x^{\alpha-1}\Big) + \exp\Big(-\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big)\Big)^{-2} .\end{aligned}$$

With the same bounds we used in the previous proof for the space derivative for $p > -1$ we have

$$\exp\Big(\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) \ge 1 + \frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1} + \frac{1}{2}\Big(\frac{1}{2} C_t C_s (1+p)\Delta x^{\alpha-1}\Big)^2$$

and

$$\exp\Big(-\frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1}\Big) \ge 1 - \frac{1}{2} C_t C_s (1+p) \Delta x^{\alpha-1} .$$

Hence, by adding the two lower bounds and due to the monotonicity of $x \mapsto x^2$ for $x > 0$ we have

$$\begin{aligned}&\Big(\exp\Big(\frac{1}{2} C_t C_s (1+p)\Delta x^{\alpha-1}\Big) + \exp\Big(-\frac{1}{2} C_t C_s (1+p)\Delta x^{\alpha-1}\Big)\Big)^{-2} \\ &\le \Big(2 + \frac{(1+p)^2}{8} C_t^2 C_s^2 \Delta x^{2\alpha-2}\Big)^{-2} .\end{aligned}$$

We want to show the bound

$$\begin{aligned}&\left|\frac{1+p}{2} \operatorname{sech}^2 \Big(-\frac{1}{2} C_t C_s (1+p)\Delta x^{\alpha-1}\Big)\Big(-\frac{1}{2} C_t C_s \Delta x^{\alpha-1} + \frac{1+p}{4} C_t t \Delta x^{-1}\Big)\right| \\ &\le 2(1+p)\Big(2 + \frac{(1+p)^2}{8} C_t^2 C_s^2 \Delta x^{2\alpha-2}\Big)^{-2} \left| -\frac{1}{2} C_t C_s \Delta x^{\alpha-1} + \frac{1+p}{4} C_t \Delta x^{-1}\right| \\ &\le K_2 \Delta x^{\alpha}\end{aligned}$$

for some $K_2 > 0$. As $\Delta x \to 0$ we have $\Delta x^{-1} > \Delta x^{\alpha-1}$ for $\alpha \in (0,1)$ and hence we have

$$\left| -\frac{1}{2} C_t C_s \Delta x^{\alpha-1} + \frac{1+p}{4} C_t \Delta x^{-1}\right| = -\frac{1}{2} C_t C_s \Delta x^{\alpha-1} + \frac{1+p}{4} C_t \Delta x^{-1}$$

as $\Delta x \to 0$. We need to show that

$$2(1+p)\Big(2+\frac{(1+p)^2}{8}C_t^2C_s^2\Delta x^{2\alpha-2}\Big)^{-2}\Big(-\frac{1}{2}C_tC_s\Delta x^{\alpha-1}+\frac{1+p}{4}C_tt\Delta x^{-1}\Big)\le K_2\Delta x^\alpha\,.$$

By multiplication of both sides with the quadratic factor we get

$$2(1+p)\Big(-\frac{1}{2}C_tC_s\Delta x^{\alpha-1}+\frac{1+p}{4}C_tt\Delta x^{-1}\Big)\le K_2\Delta x^\alpha\Big(2+\frac{(1+p)^2}{8}C_t^2C_s^2\Delta x^{2\alpha-2}\Big)^2\,.$$

Adding the first term on the left-hand side we get

$$\begin{aligned}\frac{(1+p)^2}{2}C_tt\Delta x^{-1}&\le K_2\Delta x^\alpha\Big(2+\frac{(1+p)^2}{8}C_t^2C_s^2\Delta x^{2\alpha-2}\Big)^2+(1+p)C_tC_s\Delta x^{\alpha-1}\\
\frac{(1+p)^2}{2}C_tt&\le K_2\Delta x^{\alpha+1}\Big(2+\frac{(1+p)^2}{8}C_t^2C_s^2\Delta x^{2\alpha-2}\Big)^2+(1+p)C_tC_s\Delta x^{\alpha}\\
&\le K_2\Delta x^{\alpha+1}\Big(4+\frac{(1+p)^2}{4}C_t^2C_s^2\Delta x^{2\alpha-2}+\frac{(1+p)^4}{64}C_t^4C_s^4\Delta x^{4\alpha-4}\Big)\\
&\quad+(1+p)C_tC_s\Delta x^\alpha\\
&\le 4K_2\Delta x^{\alpha+1}+K_2\frac{(1+p)^2}{4}C_t^2C_s^2\Delta x^{3\alpha-1}\\
&\quad+K_2\frac{(1+p)^4}{64}C_t^4C_s^4\Delta x^{5\alpha-3}+(1+p)C_tC_s\Delta x^\alpha\end{aligned}$$

if and only if $3\alpha-1\le 0$ or $5\alpha-3\le 0$ as $\Delta x\to 0$. For $\alpha\le 3/5$ and, thus, also for $\alpha\le 1/2$ we have

$$\left|\frac{1+p}{2}\operatorname{sech}^2\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big(-\frac{1}{2}C_tC_s\Delta x^{\alpha-1}+\frac{1+p}{4}C_tt\Delta x^{-1}\Big)\right|\le K_2\Delta x^\alpha$$

for some $K_2>0$ as $\Delta x\to 0$.

By adding the two bounds derived separately above for $\alpha\le 1/2$ we have

$$\begin{aligned}&\Big|\dot u^\epsilon(t,\xi(t)+C_s\Delta x^\alpha)-\dot u(t,\xi(t)+C_s\Delta x^\alpha)\Big|\\
&\quad\le|\dot p|\left|\frac{1}{2}+\frac{1}{2}\tanh\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\right|\\
&\qquad+|\dot p|\left|\frac{1+p}{2}\operatorname{sech}^2\Big(-\frac{1}{2}C_tC_s(1+p)\Delta x^{\alpha-1}\Big)\Big(-\frac{1}{2}C_tC_s\Delta x^{\alpha-1}+\frac{1+p}{4\Delta x}C_tt\Big)\right|\\
&\quad\le|\dot p|K_1\Delta x^\alpha+|\dot p|K_2\Delta x^\alpha=K\Delta x^\alpha\end{aligned}$$

for $K=|\dot p|(K_1+K_2)>0$.

The proof for the tangent sensitivity to the left of the shock is analogous due to symmetry with respect to the shock location. This concludes the proof. □

6.2 Numerical Analysis

In this section we present the numerical analysis of our method for the numerical simulation of the tangent shock sensitivities in scalar conservation laws based on the assumptions made in the previous section. The proposed method itself was already introduced in Section 1.4.4 and Section 1.4.5. We first present a convergence analysis for the numerical approximation of the shock locations. Based on the shock locations we then present a convergence analysis for the numerical approximation of the shock sensitivitities.

6.2.1 Shock Location

We now give the numerical analysis of an algorithm for the simulation of the shock locations in a discontinuous solution of a scalar conservation law based on a shock capturing numerical method that satisfies Assumption 8.

As explained in Section 1.4.4 the shock location algorithm is motivated by the explicit Euler discretization of the ODE that gives the dynamics of the characteristics.

Definition 38. The *shock location algorithm* is defined by the following iteration

$$\Xi_k^{j+1} := \Xi_k^j + \Delta t f'\big(U(t_j, \Xi_k^j)\big)$$

for all $j \in \{0, \dots, m-1\}$ with $\Xi_k^0 := \xi_k(0, p_0)$ and where $U(t_j, \cdot)$ is the piecewise constant interpolation of U^j.

The shock location is sublinearly convergent with the convergence rate $O(\Delta x^\alpha)$ with the same $\alpha \in (0,1)$ for which Assumption 8 holds. Furthermore, we give an explicit constant factor A for the upper bound of its numerical approximation error. The explicit constant factor is relevant to the definition of the shock sensitivity algorithm in the next section.

Proposition 36. Let

$$\begin{aligned} A_1 &\equiv \max\{\ |f'(u(t,x,p_0))| : t \in [0,T], x \in \hat{\Omega}\ \} \\ A_2 &\equiv 2 + \max\{\ |f'(U(t,x))| : t \in [0,T], x \in \hat{\Omega}\ \} \\ A &\equiv \max\{A_1, A_2\} \end{aligned}$$

and let Ξ_k^j be the result of the shock location algorithm with $\Delta t = C_t \Delta x$. Then

$$\|\Xi_k^j - \xi_k(t_j, p_0)\| \leq A C_s \Delta x^\alpha$$

for all $\Delta x \leq \delta$ for some $\delta > 0$ and for all $j \in \{0, \dots, m\}$ where C_s and α are such that Assumption 8 holds.

Proof. The proof is by induction over j. For the base case $j = 1$ we have

$$\|\Xi_k^j - \xi_k(t_j, p_0)\| = 0$$

by definition of the algorithm.

By Assumption 6 and Assumption 8 both $u(t,x,p_0)$ and $U(t,x)$ are bounded for all $t \in [0,T], x \in \hat{\Omega}$ and by Assumption 4 the flux function is twice continuously differentiable, i.e. f' is continuous and we have $A_1, A_2 < \infty$ and, consequently, also $A < \infty$.

For the induction step we consider three cases:

(i) $\xi_k(t_j, p_0) - AC_s\Delta x^\alpha \le \Xi_k^j \le \xi_k(t_j, p_0) - C_s\Delta x^\alpha$

(ii) $\xi_k(t_j, p_0) + C_s\Delta x^\alpha \le \Xi_k^j \le \xi_k(t_j, p_0) + AC_s\Delta x^\alpha$

(iii) $\xi_k(t_j, p_0) - C_s\Delta x^\alpha < \Xi_k^j < \xi_k(t_j, p_0) + C_s\Delta x^\alpha$

We show that if one of the cases holds for j then one of the cases also holds for $j+1$ as $\Delta x \to 0$.

The shock speed is given by the Rankine-Hugoniot condition

$$\partial_t \xi_k(t, p_0) = \frac{f(u_k(t, \xi_k(t, p_0), p_0)) - f(u_{k+1}(t, \xi_k(t, p_0), p_0))}{u_k(t, \xi_k(t, p_0), p_0) - u_{k+1}(t, \xi_k(t, p_0), p_0)}$$

which according to Assumption 6 and Assumption 4 has a right-hand side that for its unique solution is continuously differentiable with respect to t on $[0, T]$. According to Lemma 5 we have $\partial_t \xi_k(t, p_0)$ is locally Lipschitz continuous and hence by Taylor expansion we have

$$\xi_k(t_{j+1}, p_0) = \xi_k(t_j, p_0) + C_t\Delta x \partial_t \xi_k(t_j, p_0) + O(\Delta x^2)$$

as $\Delta x \to 0$.

We first consider case (i).

With $\Xi_k^j < \xi_k(t_j, p_0) - C_s\Delta x^\alpha$ as the case assumption we have

$$\Xi_k^j \in \hat{\Omega} \setminus \left(\xi_k(t_j, p_0) - C_s\Delta x^\alpha, \xi_k(t_j, p_0) + C_s\Delta x^\alpha\right) .$$

The numerical shock location approximation lies to the left of the shock region belonging to the exact shock location we are trying to simulate.

By Corollary 22 we have

$$\xi_k(t, p_0) - \xi_{k-1}(t, p_0) \ge C_i .$$

The shock location to the left of the shock location we are considering is at least at a constant distance. With $\alpha > 0$ we have

$$(A+1)C_s\Delta x^\alpha \le C_i$$

for all $\Delta x \le \delta$ for some $\delta > 0$ and consequently also as $\Delta x \to 0$. So combining the two inequalities we have

$$\begin{gathered}\xi_k(t, p_0) - \xi_{k-1}(t, p_0) \ge (A+1)C_s\Delta x^\alpha \\ \xi_k(t, p_0) - AC_s\Delta x^\alpha \ge \xi_{k-1}(t, p_0) + C_s\Delta x^\alpha\end{gathered}$$

as $\Delta x \to 0$. With $\xi_k(t_j, p_0) - AC_s\Delta x^\alpha \le \Xi_k^j$ as the case assumption we have

$$\Xi_k^j \ge \xi_{k-1}(t, p_0) - C_s\Delta x^\alpha$$

and so

$$\Xi_k^j \in \hat{\Omega} \setminus \left(\xi_{k-1}(t_j, p_0) - C_s\Delta x^\alpha, \xi_{k-1}(t_j, p_0) + C_s\Delta x^\alpha\right) .$$

The left side bound on the numerical shock location plus some additional buffer guarantees that eventually the numerical shock location is to the right of the shock region belonging to the shock to the left of the one we are currently considering as $\Delta x \to 0$.

Since $\Xi_k^j \in (\xi_{k-1}(t_j,p_0), \xi_k(t_j,p_0))$ we have

$$\Xi_k^j \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} \left(\xi_k(t_j,p_0) - C_s\Delta x^\alpha, \xi_k(t_j,p_0) + C_s\Delta x^\alpha\right) .$$

The numerical shock location lies between the corresponding exact shock and the one to its left and, hence, it lies outside of all shock regions in the numerical solution U according to Assumption 8. By Assumption 8 we then have

$$U(t_j,\Xi_k^j) - u(t_j,\Xi_k^j,p_0) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$.

By definition of the shock location algorithm we have

$$\begin{aligned}\Xi_k^{j+1} &= \Xi_k^j + C_t\Delta x f'\big(U(t_j,\Xi_k^j)\big)\\ &= \Xi_k^j + C_t\Delta x f'\big(u(t_j,\Xi_k^j,p_0) + O(\Delta x^\alpha)\big)\end{aligned}$$

as $\Delta x \to 0$. By Assumption 4 the function f is twice continuously differentiable. So by Lemma 5 f' is locally Lipschitz continuous around $u(t_j,\Xi_k^j,p_0)$ and we have

$$\begin{aligned}\Xi_k^{j+1} &= \Xi_k^j + C_t\Delta x\Big(f'\big(u(t_j,\Xi_k^j,p_0)\big) + O(\Delta x^\alpha)\Big)\\ &= \Xi_k^j + C_t\Delta x f'\big(u(t,\Xi_k^j,p_0)\big) + O(\Delta x^{1+\alpha})\end{aligned}$$

as $\Delta x \to 0$. The numerical shock dynamics can be interpreted to be based on the characteristic of the exact solution of the conservation law evaluated at the numerical shock location, i.e. to the left of the exact shock, plus an approximation error.

We now show how the characteristic evaluated at the numerical shock location relates to the characteristic of u_k evaluated at the exact shock location. By Assumption 6 the solution piece $u_k(t_j,x,p_0)$ is continuously differentiable and thus according to Lemma 5 locally Lipschitz continuous with respect to x on $[\xi_{k-1}(t_j,p_0), \xi_k(t_j,p_0)]$. With the induction hypothesis that

$$\|\Xi_k^j - \xi_k(t_j,p_0)\| \leq AC_s\Delta x^\alpha$$

we have

$$\begin{aligned}u_k(t_j,\Xi_k^j,p_0) &= u_k(t_j,\xi_k(t_j,p_0) + (\Xi_k^j - \xi_k(t_j,p_0)),p_0)\\ &= u_k(t_j,\xi_k(t_j,p_0)),p_0) + O(\|\Xi_k^j - \xi_k(t_j,p_0)\|)\\ &= u_k(t_j,\xi_k(t_j,p_0)),p_0) + O(\Delta x^\alpha)\end{aligned}$$

as $\Delta x \to 0$. The exact weak solution evaluated at the numerical shock location is the exact weak solution evaluated at the left limit of the exact shock location plus an approximation error.

Since $\Xi_k^j \in (\xi_{k-1}(t_j,p_0), \xi_k(t_j,p_0))$ we have $u(t_j,\Xi_k^j,p_0) = u_k(t_j,\Xi_k^j,p_0)$ and so using the previous derivation we have

$$\begin{aligned}\Xi_k^{j+1} &= \Xi_k^j + C_t\Delta x f'(u_k(t_j,\Xi_k^j,p_0)) + O(\Delta x^{1+\alpha})\\ &= \Xi_k^j + C_t\Delta x f'(u_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)) + O(\Delta x^{1+\alpha})\end{aligned}$$

as $\Delta x \to 0$. And, finally, by the same argument regarding the local Lipschitz continuity of f' as above we have

$$\Xi_k^{j+1} = \Xi_k^j + C_t \Delta x f'(u_k(t_j, \xi_k(t_j, p_0), p_0)) + O(\Delta x^{1+\alpha})$$

as $\Delta x \to 0$. The numerical shock dynamics can be interpreted to be based on the characteristic of the exact solution of the conservation law evaluated at the left limit of the exact shock location plus an approximation error.

We now bring in the crucial element of the proof that causes the dynamics of the approximate numerical shock location to be such that the numerical shock location always goes in the direction of the exact shock location if it is too far away. This crucial element is the strong entropy condition that holds around strong shock discontinuities (see Definition 28). By the strong entropy condition we have

$$f'(u_k(t_j, \xi_k(t_j, p_0), p_0)) \geq \partial_t \xi_k(t_j, p_0) + \delta$$

for some $\delta > 0$. The characteristic at the left limit of the exact shock location points into the curve of the shock discontinuity.

If we substitute the entropy condition in the above approximation of the numerical shock dynamics based on the left limit of the exact shock location we have

$$\Xi_k^{j+1} \geq \Xi_k^j + C_t \Delta x \partial_t \xi_k(t_j, p_0) + \delta C_t \Delta x + O(\Delta x^{1+\alpha})$$

as $\Delta x \to 0$. We subtract the exact shock location of the next time step from both sides and then approximate it by first-order Taylor expansion to get

$$\begin{aligned}
\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) &\geq \Xi_k^j + C_t \Delta x \partial_t \xi_k(t_j, p_0) - \xi_k(t_{j+1}, p_0) + \delta C_t \Delta x + O(\Delta x^{1+\alpha}) \\
&= \Xi_k^j + C_t \Delta x \partial_t \xi_k(t_j, p_0) - \xi_k(t_j, p_0) - C_t \Delta x \partial_t \xi_k(t_j, p_0) \\
&\quad + \delta C_t \Delta x + O(\Delta x^{1+\alpha}) + O(\Delta x^2) \\
&= \Xi_k^j - \xi_k(t_j, p_0) + \delta C_t \Delta x + O(\Delta x^{1+\alpha}) + O(\Delta x^2)
\end{aligned}$$

as $\Delta x \to 0$. By multiplying both sides by negative one we have

$$\xi_k(t_{j+1}, p_0) - \Xi_k^{j+1} \leq \xi_k(t_j, p_0) - \Xi_k^j - \delta C_t \Delta x + O(\Delta x^{1+\alpha}) + O(\Delta x^2)$$

as $\Delta x \to 0$. We have upper bounded the error of the numerical shock location of the next time step based on the error of the numerical shock location of the current time step.

For any $K_1, K_2 > 0$ and $\alpha > 0$ there exists a $\delta_2 > 0$ such that for all $\Delta x \leq \delta_2$ we have

$$-\delta C_t \Delta x + K_1 \Delta x^{1+\alpha} + K_2 \Delta x^2 < 0 \ .$$

Substituting the case assumption $\xi_k(t_j, p_0) - \Xi_k^j \leq A C_s \Delta x^\alpha$ in the above bound we have

$$\xi_k(t_{j+1}, p_0) - \Xi_k^{j+1} \leq A C_s \Delta x^\alpha - \delta C_t \Delta x + O(\Delta x^{1+\alpha}) + O(\Delta x^2) < A C_s \Delta x^\alpha$$

as $\Delta x \to 0$. We have shown one side of the bound for the induction step of case (i).

It remains to show that also

$$\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) \leq A C_s \Delta x^\alpha$$

as $\Delta x \to 0$. We show this second bound by showing that we cannot move too far to the right via a step that is only $O(\Delta x)$ long when the region we are allowed to reach is $O(\Delta x^\alpha)$ wide.

By definition of A_1 we have $f'(u_k(t_j, \xi_k(t_j, p_0), p_0)) \leq A_1$ and so

$$\begin{aligned}\Xi_k^{j+1} &= \Xi_k^j + C_t \Delta x f'(u_k(t_j, \xi_k(t_j, p_0), p_0)) + O(\Delta x^{1+\alpha})\\ &\leq \Xi_k^j + C_t \Delta x A_1 + O(\Delta x^{1+\alpha})\end{aligned}$$

as $\Delta x \to 0$. We can upper bound how far the numerical shock location moves to the right in the next time step based on an upper bound of the characteristic speeds in the exact solution. By the case assumption we have $\Xi_k^j \leq \xi_k(t_j, p_0) - C_s \Delta x^\alpha$, i.e. the numerical shock location is to the left of the shock region and so by substitution above we have

$$\Xi_k^{j+1} \leq \xi_k(t_j, p_0) - C_s \Delta x^\alpha + C_t \Delta x A_1 + O(\Delta x^{1+\alpha})$$

as $\Delta x \to 0$. By subtracting the exact shock location in the next time step from both sides and approximating the shock location in the next time step by its Taylor expansion we get

$$\begin{aligned}\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) &\leq \xi_k(t_j, p_0) - \xi_k(t_{j+1}, p_0) - C_s \Delta x^\alpha + C_t \Delta x A_1 + O(\Delta x^{1+\alpha})\\ &= -C_t \Delta x \partial_t \xi_k(t_j, p_0) - C_s \Delta x^\alpha + C_t \Delta x A_1 + O(\Delta x^{1+\alpha})\\ &= -C_s \Delta x^\alpha + \big(A_1 - \partial_t \xi_k(t_j, p_0)\big) C_t \Delta x + O(\Delta x^{1+\alpha})\end{aligned}$$

as $\Delta x \to 0$. By Assumption 8 we have $C_s \geq 2C_t$ so

$$\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) \leq -C_s \Delta x^\alpha + \big(A_1 - \partial_t \xi_k(t_j, p_0)\big) \frac{1}{2} C_s \Delta x + O(\Delta x^{1+\alpha})$$

as $\Delta x \to 0$. We have a second bound for the numerical shock location in the next time step in terms of how far away it is to the right of the exact shock location. Now we just need to interpret that bound to get the correct one for the induction step.

By the entropy condition and by the definition of A_1 we have both

$$\partial_t \xi_k(t_j, p_0) < f'(u_k(t_j, \xi_k(t_j, p_0), p_0)) \leq A_1$$

and

$$\partial_t \xi_k(t_j, p_0) > f'(u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)) \geq -A_1 \ .$$

We thus also have the bound $|\partial_t \xi_k(t_j, p_0)| < A_1$. Based on that bound we can bound one of the factors in the equation above

$$A_1 - \partial_t \xi_k(t_j, p_0) \leq A_1 + |\partial_t \xi_k(t_j, p_0)| < 2A_1 \leq 2A \ .$$

Since $\alpha < 1$ we have $\Delta x < \Delta x^\alpha$ for all $\Delta x < 1$. We also have $K \Delta x^{1+\alpha} < \Delta x^\alpha$ for any constant $K > 0$ as $\Delta x \to 0$, i.e. for all $\Delta x \leq \delta$ there exists some $\delta > 0$ such that the inequality holds. By the previous two inequalities we have the second bound on the distance between the numerical and exact shock location above as

$$\begin{aligned}\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) &\leq -C_s \Delta x^\alpha + 2A \frac{1}{2} C_s \Delta x + O(\Delta x^{1+\alpha})\\ &< -C_s \Delta x^\alpha + A C_s \Delta x^\alpha + O(\Delta x^{1+\alpha})\\ &< (A-1) C_s \Delta x^\alpha + O(\Delta x^{1+\alpha})\\ &< A C_s \Delta x^\alpha\end{aligned}$$

as $\Delta x \to 0$. We have shown the second side of the bound for the induction step for case (i) and have in total shown that

$$\|\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0)\| \leq AC_s \Delta x^\alpha \ .$$

Case (ii) is analgous to case (i) by symmetry about the exact shock discontinuity. Every step applies with a switched sign when considering the corresponding shock location ξ_{k+1} to the right of ξ_k instead of ξ_{k-1} and the weak solution piece u_{k+1} to the right of ξ_k instead of u_k.

Case (iii) is similar to the second bound of case (i). In case (iii) the numerical shock location we consider is inside of the shock region. The discrete step we take is $O(\Delta x)$ long because by Assumption 8 the numerical solution is bounded even inside of the shock region. We show case (iii) also by showing that we cannot move too far either to the left or the right via a step that is only $O(\Delta x)$ long when the region we are allowed to reach is $O(\Delta x^\alpha)$ wide.

By definition of A_2 we have $f'(U(t_j, \Xi_k^j)) \leq A_2 - 2$ and so by definition of the shock location algorithm we have

$$\begin{aligned} \Xi_k^{j+1} &= \Xi_k^j + C_t \Delta x f'(U_k(t_j, \Xi_k^j)) \\ &\leq \Xi_k^j + C_t \Delta x (A_2 - 2) \ . \end{aligned}$$

By subtraction of the exact shock location from both sides and then approximating the exact shock location via its Taylor expansion we have

$$\begin{aligned} \Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) &\leq \Xi_k^j + C_t \Delta x (A_2 - 2) - \xi_k(t_{j+1}, p_0) \\ &\leq \Xi_k^j + C_t \Delta x (A_2 - 2) - \xi_k(t_j, p_0) - C_t \Delta x \partial_t \xi_k(t_j, p_0) + O(\Delta x^2) \end{aligned}$$

as $\Delta x \to 0$. We have an upper bound for how far right of the exact shock location the numerical shock location is and need to interpret it such that we get one side of the induction step.

By the case assumption we have $\Xi_k^j - \xi_k(t_j, p_0) \leq C_s \Delta x^\alpha$ and so

$$\begin{aligned} \Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) &\leq C_s \Delta x^\alpha + C_t \Delta x (A_2 - 2) - C_t \Delta x \partial_t \xi_k(t_j, p_0) + O(\Delta x^2) \\ &= C_s \Delta x^\alpha + (A_2 - \partial_t \xi_k(t_j, p_0) - 2) C_t \Delta x + O(\Delta x^2) \end{aligned}$$

as $\Delta x \to 0$. By Assumption 8 we have $C_s \geq 2C_t$ and as shown before due to the entropy condition and the definition of A_1 we have $|\partial_t \xi_k(t_j, p_0)| < A_1$. We upper bound one of the factors above by

$$\begin{aligned} (A_2 - \partial_t \xi_k(t_j, p_0) - 2) C_t \Delta x &\leq (A_2 - \partial_t \xi_k(t_j, p_0) - 2) \frac{1}{2} C_s \Delta x \\ &< (A_2 + A_1 - 2) \frac{1}{2} C_s \Delta x \\ &\leq (2 \max\{A_1, A_2\} - 2) \frac{1}{2} C_s \Delta x \\ &\leq (A - 1) C_s \Delta x \ . \end{aligned}$$

Since $\alpha < 1$ we have the distance between the numerical shock and the exact shock to the right side in the next time step upper bounded by

$$\begin{aligned} \Xi_k^{j+1} - \xi_k(t_{j+1}, p_) &\leq C_s \Delta x^\alpha + (A - 1) C_s \Delta x + O(\Delta x^2) \\ &< AC_s \Delta x^\alpha \end{aligned}$$

as $\Delta x \to 0$. We have shown that the numerical shock location in the next time step does not move too far to the right.

The bound in the other direction is analogous by switching signs. By definition of A_2 we also have $-f'(U(t_j, \Xi_k^j)) \leq A_2 - 2$, i.e. $f'(U(t_j, \Xi_k^j)) \geq 2 - A_2$ and, hence, we have

$$\begin{aligned}\Xi_k^{j+1} &= \Xi_k^j + C_t \Delta x f'(U_k(t_j, \Xi_k^j)) \\ &\geq \Xi_k^j + C_t \Delta x (2 - A_2)\ .\end{aligned}$$

By subtraction of the shock location and using the Taylor expansion we get

$$\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0) \geq \Xi_k^j + C_t \Delta x (2 - A_2) - \xi_k(t_j, p_0) - C_t \Delta x \partial_t \xi_k(t_j, p_0) + O(\Delta x^2)$$

as $\Delta x \to 0$. Multiplying by negative one, using the case assumption and assumed bounds with analogous steps to above we get

$$\begin{aligned}\xi_k(t_{j+1}, p_0) - \Xi_k^{j+1} &\leq \xi_k(t_j, p_0) - \Xi_k^j - C_t \Delta x (2 - A_2) + C_t \Delta x \partial_t \xi_k(t_j, p_0) + O(\Delta x^2) \\ &\leq C_s \Delta x^\alpha - C_t \Delta x (2 - A_2) + C_t \Delta x \partial_t \xi_k(t_j, p_0) + O(\Delta x^2) \\ &= C_s \Delta x^\alpha + C_t \Delta x (A_2 - 2 + \partial_t \xi_k(t_j, p_0)) + O(\Delta x^2) \\ &< C_s \Delta x^\alpha + C_t \Delta x (A_2 - 2 + A_1) + O(\Delta x^2) \\ &\leq C_s \Delta x^\alpha + C_t \Delta x (2A - 2) + O(\Delta x^2) \\ &\leq C_s \Delta x^\alpha + C_s \Delta x (A - 1) + O(\Delta x^2) \\ &< A C_s \Delta x^\alpha\end{aligned}$$

as $\Delta x \to 0$. We have shown that the numerical shock location in the next time step does not move too far to the left. Combining the two bounds we have shown the induction step

$$\|\Xi_k^{j+1} - \xi_k(t_{j+1}, p_0)\| \leq A C_s \Delta x^\alpha\ .$$

This concludes the proof. □

6.2.2 Shock Sensitivity

We now give the numerical analysis of an algorithm for the simulation of the shock sensitivities in a discontinuous solution of a scalar conservation law based on a shock capturing numerical method that satisfies Assumption 8 and Assumption 9, its discrete tangent sensitivities that satisfy Assumption 10 and based on Proposition 36.

As explained in Section 1.4.5 the shock sensitivity algorithm is motivated by the explicit Euler discretization of the tangent ODE resulting from the implicit differentiation of the Rankine-Hugoniot condition (see Equations (5.2.1)-(5.2.10)).

Definition 39. The *shock sensitivity algorithm* is defined by the following iteration

$$\begin{aligned}\dot{\Xi}_k^{j+1} &:= \dot{\Xi}_k^j \\ &+ \Delta t \left(\frac{f'(U(t_j, \Xi_k^j - C_r \Delta x^\alpha)) \left(\frac{U(t_j, \Xi_k^j - C_r \Delta x^\alpha) - U(t_j, \Xi_k^j - C_r \Delta x^\alpha - 2\Delta x)}{2\Delta x} \dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j - C_r \Delta x^\alpha) \right)}{U(t_j, \Xi_k^j - C_r \Delta x^\alpha) - U(t_j, \Xi_k^j + C_r \Delta x^\alpha)} \right.\end{aligned}$$

$$
\begin{aligned}
&- \frac{f'(U(t_j, \Xi_k^j + C_r\Delta x^\alpha))\Big(\frac{U(t_j,\Xi_k^j+C_r\Delta x^\alpha+2\Delta x)-U(t_j,\Xi_k^j+C_r\Delta x^\alpha)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j + C_r\Delta x^\alpha)\Big)}{U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)} \\
&- \frac{f(U(t_j, \Xi_k^j - C_r\Delta x^\alpha)) - f(U(t_j, \Xi_k^j + C_r\Delta x^\alpha))}{\big(U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big)^2} \\
&\cdot \Big(\frac{U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j - C_r\Delta x^\alpha) \\
&- \frac{U(t_j, \Xi_k^j + C_r\Delta x^\alpha + 2\Delta x) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)}{2\Delta x}\dot{\Xi}_k^j - \dot{U}(t_j, \Xi_k^j + C_r\Delta x^\alpha)\Big)\Bigg)
\end{aligned}
$$

for all $j \in \{0, \dots, m-1\}$ with $\dot{\Xi}_k^0 := \dot{\xi}_k(0, p_0)$ and with

$$C_r \equiv (A+1)C_s$$

where A is defined as in Proposition 36, where $U(t_j, \cdot), \dot{U}(t_j, \cdot)$ are the piecewise constant interpolations of $U^j, \dot{U}^j$ respectively and where C_s and α are such that Assumption 8, Assumption 9 and Assumption 10 all hold.

For the weak solution and its tangent sensitivity evaluated at the left and right limits of the shock location ξ_k we use the approximations

$$
\begin{aligned}
u(t_j, \xi_k(t_j, p_0)\pm, p_0) &\approx U(t_j, \Xi_k^j \pm C_r\Delta x^\alpha) \\
\dot{u}(t_j, \xi_k(t_j, p_0)\pm, p_0) &\approx \dot{U}(t_j, \Xi_k^j \pm C_r\Delta x^\alpha)
\end{aligned}
$$

with a constant C_r that based on Proposition 36 guarantees that the limit approximation is evaluated outside of the shock region and where the numerical solution and the discrete tangent sensitivity pointwise converge. Similarly, for the derivative of the weak solution with respect to x we also use the following finite difference approximations

$$
\begin{aligned}
\partial_x u(t_j, \xi_k(t_j, p_0)+, p_0) &\approx \frac{U(t_j, \Xi_k^j + C_r\Delta x^\alpha + 2\Delta x) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)}{2\Delta x} \\
\partial_x u(t_j, \xi_k(t_j, p_0)-, p_0) &\approx \frac{U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}
\end{aligned}
$$

to make sure that the approximations are evaluated outside of the shock region and instead at points where they pointwise converge.

Based on Assumption 8, Assumption 9 and Assumption 10 the approximation errors of the weak solution, tangent sensitivity and space derivative evaluated at the left and right limits of the shock location ξ_k have a convergence rate of $O(\Delta x^\alpha)$.

Lemma 37. Let Ξ_k^j be the result of the shock location algorithm with $\Delta t = C_t\Delta x$. Then the weak solution left and right limits to the shock location have approximation errors

$$
\begin{aligned}
U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - u_k(t_j, \xi_k(t_j, p_0), p_0) &= O(\Delta x^\alpha) \\
U(t_j, \Xi_k^j + C_r\Delta x^\alpha) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0) &= O(\Delta x^\alpha),
\end{aligned}
$$

the tangent sensitivity limits to the shock location have approximation errors

$$\dot{U}(t_j, \Xi_k^j - C_r\Delta x^\alpha) - \dot{u}_k(t_j, \xi_k(t_j, p_0), p_0) = O(\Delta x^\alpha)$$

$$\dot{U}(t_j, \Xi_k^j + C_r \Delta x^\alpha) - \dot{u}_{k+1}(t_j, \xi_k(t_j, p_0), p_0) = O(\Delta x^\alpha)$$

and the space derivative limits to the shock location have approximation errors

$$\frac{U(t_j, \Xi_k^j - C_r \Delta x^\alpha) - U(t_j, \Xi_k^j - C_r \Delta x^\alpha - 2\Delta x)}{2\Delta x} - \partial_x u_k(t_j, \xi_k(t_j, p_0), p_0) = O(\Delta x^\alpha)$$
$$\frac{U(t_j, \Xi_k^j + C_r \Delta x^\alpha + 2\Delta x) - U(t_j, \Xi_k^j + C_r \Delta x^\alpha)}{2\Delta x} - \partial_x u_{k+1}(t_j, \xi_k(t_j, p_0), p_0) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all $j \in \{0, \dots, m\}$ where C_r is as defined in the shock sensitivity algorithm and where α is such that Assumption 8, Assumption 9 and Assumption 10 hold, correspondingly.

Proof. We start by showing that the points $\Xi_k^j \pm C_r \Delta x^\alpha$ lie outside of the shock regions where the weak solution, tangent sensitivity and space derivative are potentially smoothed out, oscillating or not numerically stable due to the numerical viscosity.

By Proposition 36 the approximation error of the numerical shock location is bounded by

$$\|\Xi_k^j - \xi_k(t_j, p_0)\| \leq A C_s \Delta x^\alpha$$

for all $\Delta x \leq \delta$ for some $\delta > 0$. The bound on the error norm implies that the numerical shock location is not too far to the left of the exact shock location, i.e. we have

$$\Xi_k^j \geq \xi_k(t_j, p_0) - A C_s \Delta x^\alpha$$

as $\Delta x \to 0$. By adding $C_r \Delta x^\alpha$ to both sides of the above inequality and using the definition of C_r we have

$$\begin{aligned}\Xi_k^j + C_r \Delta x^\alpha &\geq \xi_k(t_j, p_0) - A C_s \Delta x^\alpha + C_r \Delta x^\alpha \\ &= \xi_k(t_j, p_0) - A C_s \Delta x^\alpha + (A+1) C_s \Delta x^\alpha \\ &= \xi_k(t_j, p_0) + C_s \Delta x^\alpha\end{aligned}$$

for all $\Delta x \leq \delta$ for some $\delta > 0$. By the definition of C_r and the bound on the numerical approximation error adding $C_r \Delta x^\alpha$ to the numerical shock location, thus, guarantees that we get a point with a distance at least $C_s \Delta x^\alpha$ to the right of the exact shock location. That point is to the right of the shock region and at a point where we have pointwise convergence of the weak solution, its derivative with respect to space and its tangent sensitivity. At least if we are not in the shock region of the next shock to the right. We have

$$\Xi_k^j + C_r \Delta x^\alpha \in \hat{\Omega} \setminus (\xi_k(t_j, p_0) - C_s \Delta x^\alpha, \xi_k(t_j, p_0) + C_s \Delta x^\alpha) \,.$$

By Corollary 22 we have at least a constant distance to the shock location to the right of the one we are considering

$$\xi_k(t_j, p_0) \leq \xi_{k+1}(t_j, p_0) - C_i \,.$$

The bound on the approximation error norm of the numerical shock location above also implies that the numerical shock location is not too far to the right of the exact shock location, i.e. we have

$$\Xi_k^j \leq \xi_k(t_j, p_0) + A C_s \Delta x^\alpha$$

as $\Delta x \to 0$. By adding $C_r \Delta x^\alpha$ to both sides of the above inequality and using the definition of C_r and the constant distance to the shock location to the right we have

$$\Xi_k^j + C_r \Delta x^\alpha \leq \xi_k(t_j, p_0) + A C_s \Delta x^\alpha + C_r \Delta x^\alpha$$

$$
\begin{aligned}
&\leq \xi_{k+1}(t_j,p_0) - C_i + (2A+1)C_s\Delta x^\alpha \\
&= \xi_{k+1}(t_j,p_0) - C_s\Delta x^\alpha - C_i + (2A+2)C_s\Delta x^\alpha \; .
\end{aligned}
$$

For all $\Delta x \leq \delta$ with

$$\delta = \Big(\frac{C_i}{(2A+2)C_s}\Big)^{1/\alpha}$$

we have

$$(2A+2)C_s\Delta x^\alpha \leq C_i$$

and so

$$\Xi_k^j + C_r\Delta x^\alpha \leq \xi_{k+1}(t_j,p_0) - C_s\Delta x^\alpha \; .$$

For some small enough Δx adding the $C_r\Delta x^\alpha$ required to move out of the shock region belonging to the shock we are considering does not move us too far to the right such that we would move into the shock region of the next shock. We have

$$\Xi_k^j + C_r\Delta x^\alpha \in \hat{\Omega} \setminus (\xi_{k+1}(t_j,p_0) - C_s\Delta x^\alpha, \xi_{k+1}(t_j,p_0) + C_s\Delta x^\alpha) \; .$$

Combining the results that we are to the right of the shock region belonging to the shock we are considering and to the left of the next shock we have

$$\Xi_k^j + C_r\Delta x^\alpha \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} (\xi_k(t_j,p_0) - C_s\Delta x^\alpha, \xi_k(t_j,p_0) + C_s\Delta x^\alpha)$$

as $\Delta x \to 0$.

The same result can analogously be shown for taking a $C_r\Delta x^\alpha$ step to the left of the numerical shock location. The numerical shock location approximation error bound implies

$$\Xi_k^j - C_r\Delta x^\alpha \leq \xi_k(t_j,p_0) - C_s\Delta x^\alpha$$

as $\Delta x \to 0$. When moving to the left by $C_r\Delta x^\alpha$ from the numerical shock location we are not in the shock region of that shock. The constant distance to the next shock to the left additionally implies

$$\Xi_k^j - C_r\Delta x^\alpha \geq \xi_{k-1}(t_j,p_0) + C_s\Delta x^\alpha$$

as $\Delta x \to 0$. When moving to the left by $C_r\Delta x^\alpha$ from the numerical shock location we are also not in the shock region of the next shock to the left. Again, combining the results we have

$$\Xi_k^j - C_r\Delta x^\alpha \in \hat{\Omega} \setminus \bigcup_{k=1}^{K} (\xi_k(t_j,p_0) - C_s\Delta x^\alpha, \xi_k(t_j,p_0) + C_s\Delta x^\alpha)$$

as $\Delta x \to 0$.

The fundamental idea is to evaluate at a point far enough away from the shock location such that we are not evaluating the numerical solution of the conservation law and its sensitivities in the area where numerical viscosity dominates. But we still evaluate at a point that converges to the exact shock location from the correct side, i.e. we have

$$\Xi_k \pm C_r\Delta x^\alpha \to \Xi_k\pm \to \xi_k\pm$$

as $\Delta x \to 0$. We can now utilize the pointwise convergence of Assumption 8, Assumption 9 and Assumption 10.

We now show the approximation error convergence rate for the weak solution left limit to the shock. The approximation error for the right limit to the shock is analogous. By Assumption 8 we have

$$U(t_j, \Xi_k^j - C_r\Delta x^\alpha) = u(t_j, \Xi_k^j - C_r\Delta x^\alpha, p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$. When showing that the numerical shock location plus or minus the step we take to move outside of the shock region satisfies the above conditions we implicitly showed the weaker bounds

$$\xi_{k-1}(t_j, p_0) < \Xi_k^j - C_r\Delta x^\alpha < \xi_k(t_j, p_0)\ .$$

Based on the definition of the solution structure in Assumption 6 we then have

$$\begin{aligned} U(t_j, \Xi_k^j - C_r\Delta x^\alpha) &= u(t_j, \Xi_k^j - C_r\Delta x^\alpha, p_0) + O(\Delta x^\alpha) && (6.2.1)\\ &= u_k(t_j, \Xi_k^j - C_r\Delta x^\alpha, p_0) + O(\Delta x^\alpha) && (6.2.2) \end{aligned}$$

as $\Delta x \to 0$.

By Assumption 6 the smooth solution piece u_k is continuously differentiable with respect to x on $[\xi_{k-1}(t_j, p_0), \xi_k(t_j, p_0)]$ so by Lemma 5 we have

$$\begin{aligned} u_k(t_j, \Xi_k^j - C_r\Delta x^\alpha, p_0) &= u_k(t_j, \xi_k(t_j, p_0) + (\Xi_k^j - \xi_k(t_j, p_0) - C_r\Delta x^\alpha), p_0) && (6.2.3)\\ &= u_k(t_j, \xi_k(t_j, p_0), p_0) + O(\|\Xi_k^j - \xi_k(t_j, p_0) - C_r\Delta x^\alpha\|) && (6.2.4)\\ &= u_k(t_j, \xi_k(t_j, p_0), p_0) + O(\Delta x^\alpha) && (6.2.5) \end{aligned}$$

as $\Delta x \to 0$. In the last step we used the approximation error bound of the numerical shock location according to Proposition 36 and the triangle inequality.

Combining Equation (6.2.2) and Equation (6.2.5) we get

$$U(t_j, \Xi_k^j - C_r\Delta x^\alpha) = u_k(t_j, \xi_k(t_j, p_0), p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$. Since we earlier showed that

$$\xi_k(t_j, p_0) < \Xi_k^j + C_r\Delta x^\alpha < \xi_{k+1}(t_j, p_0)$$

we also have

$$U(t_j, \Xi_k^j + C_r\Delta x^\alpha) = u_{k+1}(t_j, \xi_k(t_j, p_0), p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ by analogous steps. We have shown the weak solution approximation error convergence rate of the lemma.

We now show the approximation error convergence rate for the tangent sensitivity left limit to the shock. As before, the approximation error for the right limit to the shock is analogous. This step is, overall, analogous to the previous step where we showed the approximation error convergence rate for the weak solution left limit of the shock.

By Assumption 10 and based on the definition of the solution structure in Assumption 6 we have

$$\dot{U}(t_j, \Xi_k^j - C_r\Delta x^\alpha) = \dot{u}_k(t_j, \Xi_k^j - C_r\Delta x^\alpha, p_0) + O(\Delta x^\alpha)\ . \qquad (6.2.6)$$

as $\Delta x \to 0$.

By Assumption 6 the weak solution derivative $\mathrm{d}_p u_k$ with respect to the parameter is continuously differentiable with respect to x on $[\xi_{k-1}(t_j,p_0),\xi_k(t_j,p_0)]$ so by Lemma 5 we have

$$\begin{aligned}\dot{u}_k(t_j,\Xi_k^j - C_r\Delta x^\alpha, p_0) &= \mathrm{d}_p u_k(t_j,\Xi_k^j - C_r\Delta x^\alpha, p_0)\dot{p} && (6.2.7)\\ &= \mathrm{d}_p u_k(t_j,\xi_k(t_j,p_0) + (\Xi_k^j - \xi_k(t_j,p_0) - C_r\Delta x^\alpha), p_0)\dot{p} && (6.2.8)\\ &= \mathrm{d}_p u_k(t_j,\xi_k(t_j,p_0),p_0)\dot{p} + O(\|\Xi_k^j - \xi_k(t_j,p_0) - C_r\Delta x^\alpha\|) && (6.2.9)\\ &= \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha) && (6.2.10)\end{aligned}$$

as $\Delta x \to 0$.

Combining Equation (6.2.6) and Equation (6.2.10) we get

$$\dot{U}(t_j,\Xi_k^j - C_r\Delta x^\alpha) = \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$. We also have

$$\dot{U}(t_j,\Xi_k^j + C_r\Delta x^\alpha) = \dot{u}_{k+1}(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ by analogous steps. We have shown the tangent sensitivity approximation error convergence rate of the lemma.

Finally, we show the approximation error convergence rate for the space derivative left limit to the shock. As before, the approximation error for the right limit to the shock is analogous. This step is, overall, analogous to the previous two steps where we showed the approximation error convergence rate for the weak solution and the tangent sensitivity left limit of the shock.

By Assumption 9 and based on the definition of the solution structure in Assumption 6 we have

$$\frac{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x} = \partial_x u_k(t_j,\Xi_k^j - C_r\Delta x^\alpha, p_0) + O(\Delta x^\alpha)\;. \tag{6.2.11}$$

as $\Delta x \to 0$.

By Assumption 6 the smooth solution piece u_k is twice continuously differentiable with respect to x and, hence, the space derivative $\partial_x u_k$ is continuously differentiable with respect to x on $[\xi_{k-1}(t_j,p_0),\xi_k(t_j,p_0)]$ so by Lemma 5 we have

$$\begin{aligned}\partial_x u_k(t_j,\Xi_k^j - C_r\Delta x^\alpha, p_0) &= \partial_x u_k(t_j,\xi_k(t_j,p_0) + (\Xi_k^j - \xi_k(t_j,p_0) - C_r\Delta x^\alpha), p_0) && (6.2.12)\\ &= \partial_x u_k(t_j,\xi_k(t_j,p_0),p_0) + O(\|\Xi_k^j - \xi_k(t_j,p_0) - C_r\Delta x^\alpha\|) && (6.2.13)\\ &= \partial_x u_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha) && (6.2.14)\end{aligned}$$

as $\Delta x \to 0$. Combining Equation (6.2.6) and Equation (6.2.10) we get

$$\frac{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x} = \partial_x u_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$. We also have

$$\frac{U(t_j,\Xi_k^j + C_r\Delta x^\alpha + 2\Delta x) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)}{2\Delta x} = \partial_x u_{k+1}(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ by analogous steps. We have shown the space derivative approximation error convergence rate of the lemma. This concludes the proof. □

Based on the above lemma we can now go on to the numerical analysis of the shock sensitivity algorithm. The shock sensitivity is sublinearly convergent with the convergence rate $O(\Delta x^\alpha)$ with the same $\alpha \in (0,1)$ for which Assumption 8, Assumption 9 and Assumption 10 hold.

Proposition 38. Let $\dot{\Xi}_k^j$ be the result of the shock sensitivity algortihm with $\Delta t = C_t \Delta x$. Then

$$\|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all $j \in \{0, \dots, m\}$ where α is defined as in Definition 39.

Proof. The proof strategy is similar to that of the standard convergence proof for the explicit Euler discretization of an ODE. The extension we make to the standard proof is to show that the approximations used for the right-hand side are sufficiently accurate in order to obtain the $O(\Delta x^\alpha)$ global convergence rate.

According to Proposition 23 the dynamics of the shock sensitivity are given by the implicit differentiation of the Rankine-Hugoniot condition

$$\begin{align}
\partial_t \dot{\xi}_k(t, p_0) = {}& \frac{f'(u_k(t, \xi_k(t,p_0), p_0))\big(\partial_x u_k(t, \xi_k(t,p_0), p_0)\dot{\xi}_k(t,p_0) + \dot{u}_k(t, \xi_k(t,p_0), p_0)\big)}{u_k(t, \xi_k(t,p_0), p_0) - u_{k+1}(t, \xi_k(t,p_0), p_0)} \tag{6.2.15}\\
& - \frac{f'(u_{k+1}(t, \xi_k(t,p_0), p_0))\big(\partial_x u_{k+1}(t, \xi_k(t,p_0), p_0)\dot{\xi}_k(t,p_0) + \dot{u}_{k+1}(t, \xi_k(t,p_0), p_0)\big)}{u_k(t, \xi_k(t,p_0), p_0) - u_{k+1}(t, \xi_k(t,p_0), p_0)} \tag{6.2.16}\\
& - \frac{f(u_k(t, \xi_k(t,p_0), p_0)) - f(u_{k+1}(t, \xi_k(t,p_0), p_0))}{\big(u_k(t, \xi_k(t,p_0), p_0) - u_{k+1}(t, \xi_k(t,p_0), p_0)\big)^2} \tag{6.2.17}\\
& \cdot \big(\partial_x u_k(t, \xi_k(t,p_0), p_0)\dot{\xi}_k(t,p_0) + \dot{u}_k(t, \xi_k(t,p_0), p_0) \tag{6.2.18}\\
& \quad - \partial_x u_{k+1}(t, \xi_k(t,p_0), p_0)\dot{\xi}_k(t,p_0) - \dot{u}_{k+1}(t, \xi_k(t,p_0), p_0)\big) \tag{6.2.19}
\end{align}$$

which according to Assumption 6 and Assumption 4 has a right-hand side that for its unique solution is continuously differentiable with respect to t on $[0,T]$. According to Lemma 5 the derivative $\partial_t \dot{\xi}_k(t, p_0)$ is locally Lipschitz continuous and hence by Taylor expansion we have

$$\dot{\xi}_k(t_{j+1}, p_0) = \dot{\xi}_k(t_j, p_0) + C_t \Delta x \partial_t \dot{\xi}_k(t_j, p_0) + O(\Delta x^2)$$

as $\Delta x \to 0$.

Just like in the standard convergence proof for the explicit Euler method we decompose the approximation error at time $t = t_{j+1}$ into the error of the previous time step at time $t = t_j$ and the error made due to the step from t_j to t_{j+1}. By the Taylor expansion of the smooth exact tangent sensitivity we have the approximation error

$$\begin{aligned}
&\|\dot{\Xi}_k^{j+1} - \dot{\xi}_k(t_{j+1}, p_0)\| \\
&\quad = \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0) + \dot{\Xi}_k^{j+1} - \dot{\Xi}_k^j - \partial_t \dot{\xi}_k(t_j, p_0) C_t \Delta x + O(\Delta x^2)\| && \text{(Taylor expansion)}\\
&\quad \leq \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + \|\dot{\Xi}_k^{j+1} - \dot{\Xi}_k^j - C_t \Delta x \partial_t \dot{\xi}_k(t_j, p_0)\| + O(\Delta x^2) && \text{(triangle inequality)}\\
&\quad = \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + C_t \Delta x \left\| \frac{\dot{\Xi}_k^{j+1} - \dot{\Xi}_k^j}{C_t \Delta x} - \partial_t \dot{\xi}_k(t_j, p_0) \right\| + O(\Delta x^2) && \text{(definition of norm)}
\end{aligned}$$

as $\Delta x \to 0$.

We first analyze the second term. By Equations (6.2.15)-(6.2.19) and the definition of the algorithm we have

$$E^{j+1} \equiv \frac{\dot{\Xi}_k^{j+1} - \dot{\Xi}_k^j}{C_t h} - \partial_t \dot{\xi}_k(t_j, p_0) = E_1^{j+1} + E_2^{j+1} + E_3^{j+1} + E_4^{j+1}$$

with

$$\begin{aligned}
E_1^{j+1} \equiv{}& \frac{f'(U(t_j, \Xi_k^j - C_r\Delta x^\alpha))\Big(\frac{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j - C_r\Delta x^\alpha)\Big)}{U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)} \\
&- \frac{f'(u_k(t_j, \xi_k(t_j,p_0),p_0))\big(\partial_x u_k(t_j, \xi_k(t_j,p_0),p_0)\dot{\xi}_k(t_j,p_0) + \dot{u}_k(t_j, \xi_k(t_j,p_0),p_0)\big)}{u_k(t_j, \xi_k(t_j,p_0),p_0) - u_{k+1}(t_j, \xi_k(t_j,p_0),p_0)} \\
E_2^{j+1} \equiv{}& -\frac{f'(U(t_j, \Xi_k^j + C_r\Delta x^\alpha))\Big(\frac{U(t_j,\Xi_k^j + C_r\Delta x^\alpha + 2\Delta x) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j + C_r\Delta x^\alpha)\Big)}{U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)} \\
&+ \frac{f'(u_{k+1}(t_j, \xi_k(t_j,p_0),p_0))\big(\partial_x u_{k+1}(t_j, \xi_k(t_j,p_0),p_0)\dot{\xi}_k(t_j,p_0) + \dot{u}_{k+1}(t_j, \xi_k(t_j,p_0),p_0)\big)}{u_k(t_j, \xi_k(t_j,p_0),p_0) - u_{k+1}(t_j, \xi_k(t_j,p_0),p_0)} \\
E_3^{j+1} \equiv{}& -\frac{f(U(t_j, \Xi_k^j - C_r\Delta x^\alpha)) - f(U(t_j, \Xi_k^j + C_r\Delta x^\alpha))}{\big(U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big)^2} \\
&\cdot \Big(\frac{U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j - C_r\Delta x^\alpha)\Big) \\
&+ \frac{f(u_k(t_j, \xi_k(t_j,p_0),p_0)) - f(u_{k+1}(t_j, \xi_k(t_j,p_0),p_0))}{\big(u_k(t_j, \xi_k(t_j,p_0),p_0) - u_{k+1}(t_j, \xi_k(t_j,p_0),p_0)\big)^2} \\
&\cdot \big(\partial_x u_k(t_j, \xi_k(t_j,p_0),p_0)\dot{\xi}_k(t_j,p_0) + \dot{u}_k(t_j, \xi_k(t_j,p_0),p_0)\big) \\
E_4^{j+1} \equiv{}& -\frac{f(U(t_j, \Xi_k^j - C_r\Delta x^\alpha)) - f(U(t_j, \Xi_k^j + C_r\Delta x^\alpha))}{\big(U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big)^2} \\
&\cdot \Big(\frac{U(t_j, \Xi_k^j + C_r\Delta x^\alpha + 2\Delta x) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j + C_r\Delta x^\alpha)\Big) \\
&+ \frac{f(u_{k+1}(t_j, \xi_k(t_j,p_0),p_0)) - f(u_{k+1}(t_j, \xi_k(t_j,p_0),p_0))}{\big(u_k(t_j, \xi_k(t_j,p_0),p_0) - u_{k+1}(t_j, \xi_k(t_j,p_0),p_0)\big)^2} \\
&\cdot \big(\partial_x u_{k+1}(t_j, \xi_k(t_j,p_0),p_0)\dot{\xi}_k(t_j,p_0) + \dot{u}_{k+1}(t_j, \xi_k(t_j,p_0),p_0)\big)\ .
\end{aligned}$$

The four error terms represent differences between the corresponding parts on the right-hand sides of the shock sensitivity algorithm and the analytic tangent ODE.

We start by bounding the error of the term E_1^{j+1} in the total error E^{j+1} that we defined above. Since by Assumption 4 the flux function f is twice continuously differentiable the derivative f' is Lipschitz continuous in a neighborhood of $u_k(t_j, \xi_k(t_j,p_0),p_0)$. By Lemma 37 and the local Lipschitz continuity of f' we have

$$\begin{aligned}
f'(U(t_j, \Xi_k^j - C_r\Delta x^\alpha)) &= f'(u_k(t_j, \xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)) \\
&= f'(u_k(t_j, \xi_k(t_j,p_0),p_0)) + O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$. Also by Lemma 37 and the fact that all relevant variables are bounded we have the overall approximation in the E_1^{j+1} term as

$$\frac{f'(U(t_j, \Xi_k^j - C_r\Delta x^\alpha))\Big(\frac{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j, \Xi_k^j - C_r\Delta x^\alpha)\Big)}{U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)}$$

$$
\begin{aligned}
&= \frac{f'(u_k(t_j,\xi_k(t_j,p_0),p_0)) + O(\Delta x^\alpha)}{u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)} \\
&\quad \cdot \Big(\big(\partial_x u_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)\big)\dot{\Xi}_k^j + \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)\Big) \\
&= \frac{f'(u_k(t_j,\xi_k(t_j,p_0),p_0))\Big(\partial_x u_k(t_j,\xi_k(t_j,p_0),p_0)\dot{\Xi}_k^j + \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0)\Big) + O(\Delta x^\alpha)}{u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)}
\end{aligned}
$$

as $\Delta x \to 0$.

The right-hand side of the above equation has the structure

$$g(\epsilon) = \frac{a+\epsilon}{b+\epsilon}$$

where $\epsilon = O(\Delta x^\alpha)$. The first derivative of g with respect to ϵ is

$$g'(\epsilon) = \frac{1}{b+\epsilon} - \frac{a+\epsilon}{(b+\epsilon)^2}\ .$$

By Taylor expansion at $\epsilon = 0$ we have

$$
\begin{aligned}
g(\epsilon) &= g(0) + g'(0)\epsilon + O(\epsilon^2) \\
&= g(0) + \epsilon\Big(\frac{1}{b+\epsilon} - \frac{a+\epsilon}{(b+\epsilon)^2}\Big)\Big|_{\epsilon=0} + O(\epsilon^2) \\
&= g(0) + \epsilon\Big(\frac{1}{b} - \frac{a}{b^2}\Big) + O(\epsilon^2) \\
&= g(0) + O(\epsilon)
\end{aligned}
$$

as $\epsilon \to 0$. Consequently, we have the overall approximation in the E_1^{j+1} term as

$$
\begin{aligned}
&\frac{f'(U(t_j,\Xi_k^j - C_r\Delta x^\alpha))\Big(\frac{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j,\Xi_k^j - C_r\Delta x^\alpha)\Big)}{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)} \\
&= \frac{f'(u_k(t_j,\xi_k(t_j,p_0),p_0))\Big(\partial_x u_k(t_j,\xi_k(t_j,p_0),p_0)\dot{\Xi}_k^j + \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0)\Big)}{u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)} + O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$.

By Assumption 6 and Assumption 4 both $f'(u_k(t_j,\xi_k(t_j,p_0),p_0))$ and $\partial_x u_k(t_j,\xi_k(t_j,p_0),p_0)$ are bounded and

$$u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)$$

is bounded away from zero since u is assumed to have a strong shock discontinuity at ξ_k. Hence, by substitution of

$$\dot{\Xi}_k^j = \dot{\xi}_k(t_j,p_0) + \dot{\Xi}_k^j - \dot{\xi}_k(t_j,p_0)$$

in the right-hand side of the above equation we get

$$
\begin{aligned}
&\frac{f'(u_k(t_j,\xi_k(t_j,p_0),p_0))\Big(\partial_x u_k(t_j,\xi_k(t_j,p_0),p_0)\dot{\xi}(t_j,p_0) + \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0)\Big)}{u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)} \\
&\quad + L_1^{j+1}(\dot{\Xi}_k^j - \dot{\xi}_k(t_j,p_0)) + O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$ with

$$\|L_1^{j+1}\| = \left\| \frac{f'(u_k(t_j,\xi_k(t_j,p_0),p_0))\partial_x u_k(t_j,\xi_k(t_j,p_0),p_0)}{u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)} \right\| < \infty\ .$$

By substitution of the overall approximation into the full E_1^{j+1} term we get

$$\|E_1^{j+1}\| = \|L_1^{j+1}\| \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j,p_0)\| + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ since the first term cancels out against the exact dynamics. This concludes bounding the error in the term E_1^{j+1}.

Bounding the error of the term E_2^{j+1} in the total error E^{j+1} is analogous to the bound for E_1^{j+1} by symmetry about the shock location. We, thus, also have

$$\|E_2^{j+1}\| = \|L_2^{j+1}\| \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j,p_0)\| + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ with $\|L_2^{j+1}\| < \infty$.

We now go on to bound the error of the term E_3^{j+1} in the total error E^{j+1}. By Lemma 37, utilizing the Lipschitz continuity of f and a similar asymptotic argument as above we have

$$\begin{aligned}
&\frac{f(U(t_j,\Xi_k^j - C_r\Delta x^\alpha)) - f(U(t_j,\Xi_k^j + C_r\Delta x^\alpha))}{\big(U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)\big)^2} \\
&= \frac{f(u_k(t_j,\xi_k(t_j,p_0),p_0)) - f(u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)) + O(\Delta x^\alpha)}{\big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)\big)^2} \\
&= \frac{f(u_k(t_j,\xi_k(t_j,p_0),p_0)) - f(u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)) + O(\Delta x^\alpha)}{\big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)^2 + O(\Delta x^\alpha)} \\
&= \frac{f(u_k(t_j,\xi_k(t_j,p_0),p_0)) - f(u_{k+1}(t_j,\xi_k(t_j,p_0),p_0))}{\big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)^2} + O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$. As before also by Lemma 37 we have

$$\begin{aligned}
&\frac{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j,\Xi_k^j - C_r\Delta x^\alpha) \\
&= \partial_x u_k(t_j,\xi_k(t_j,p_0),p_0)\dot{\Xi}_k^j + \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0) + O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$. Taking the product of the two factors we have the approximation in E_3^{j+1} as

$$\begin{aligned}
&-\frac{f(U(t_j,\Xi_k^j - C_r\Delta x^\alpha)) - f(U(t_j,\Xi_k^j + C_r\Delta x^\alpha))}{\big(U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)\big)^2} \\
&\quad \cdot \Big(\frac{U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j - C_r\Delta x^\alpha - 2\Delta x)}{2\Delta x}\dot{\Xi}_k^j + \dot{U}(t_j,\Xi_k^j - C_r\Delta x^\alpha)\Big) \\
&= -\frac{f(u_k(t_j,\xi_k(t_j,p_0),p_0)) - f(u_{k+1}(t_j,\xi_k(t_j,p_0),p_0))}{\big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)^2} \\
&\quad \cdot \big(\partial_x u_k(t_j,\xi_k(t_j,p_0),p_0)\dot{\Xi}_k^j + \dot{u}_k(t_j,\xi_k(t_j,p_0),p_0)\big) + O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$.

Again, since

$$u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)$$

is bounded away from zero and all relevant factors are also bounded according to Assumption 6 by substitution of

$$\dot{\Xi}_k^j = \dot{\xi}_k(t_j, p_0) + \dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)$$

in the right-hand side of the above equation we get

$$\begin{aligned}
&- \frac{f(u_k(t_j, \xi_k(t_j, p_0), p_0)) - f(u_{k+1}(t_j, \xi_k(t_j, p_0), p_0))}{\big(u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)\big)^2} \\
&\quad \cdot \big(\partial_x u_k(t_j, \xi_k(t_j, p_0), p_0)\dot{\xi}_k(t_j, p_0) + \dot{u}_k(t_j, \xi_k(t_j, p_0), p_0)\big) + L_3^{j+1}(\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)) + O(\Delta x^\alpha)
\end{aligned}$$

with $\|L_3^{j+1}\| < \infty$.

By substitution of the overall approximation into the full E_3^{j+1} term we get

$$\|E_3^{j+1}\| = \|L_3^{j+1}\| \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ since the first term cancels out against the exact dynamics. This concludes bounding the error in the term E_3^{j+1}.

Bounding the error of the term E_4^{j+1} in the total error E^{j+1} is analogous to the bound for E_3^{j+1} by symmetry about the shock location. We, thus, also have

$$\|E_4^{j+1}\| = \|L_4^{j+1}\| \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + O(\Delta x^\alpha)$$

as $\Delta x \to \alpha$ with $\|L_4^{j+1}\| < \infty$.

Going back to bounding the total error by the triangle inquality we have

$$\begin{aligned}
\|E^{j+1}\| &= \|E_1^{j+1} + E_2^{j+1} + E_3^{j+1} + E_4^{j+1}\| \\
&\leq \|E_1^{j+1}\| + \|E_2^{j+1}\| + \|E_3^{j+1}\| + \|E_4^{j+1}\| \\
&= \Big(\|L_1^{j+1}\| + \|L_2^{j+1}\| + \|L_3^{j+1}\| + \|L_4^{j+1}\|\Big) \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + O(\Delta x^\alpha) \\
&\leq L\|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$ with $L = \max_{i,j} \|L_i^{j+1}\| < \infty$ such that $i \in \{1, \ldots, 4\}, j \in \{0, \ldots, m-1\}$.

Now, going even further back to the approximation error at time $t = t_{j+1}$ we have

$$\begin{aligned}
&\|\dot{\Xi}_k^{j+1} - \dot{\xi}_k(t_{j+1}, p_0)\| \\
&\quad = \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + C_t \Delta x \|E^{j+1}\| + O(\Delta x^2) \\
&\quad \leq \|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + C_t \Delta x \Big(L\|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + O(\Delta x^\alpha)\Big) + O(\Delta x^2) \\
&\quad = (1 + LC_t \Delta x)\|\dot{\Xi}_k^j - \dot{\xi}_k(t_j, p_0)\| + O(\Delta x^{1+\alpha}) \qquad\qquad (\alpha < 1)
\end{aligned}$$

as $\Delta x \to 0$.

By definition of the shock sensitivity algorithm we have

$$\|\dot{\Xi}_k^0 - \dot{\xi}_k(0, p_0)\| = 0\ .$$

By the Discrete Gronwall Lemma (see Lemma 7) we then have

$$\begin{aligned}\|\dot{\Xi}_k^j - \dot{\xi}_k(t_j,p_0)\| &\leq \exp(jLC_t\Delta x)\|\dot{\Xi}_k^0 - \xi_k(0,p_0)\| + \frac{\exp(jLC_t\Delta x)-1}{LC_t\Delta x}O(\Delta x^{1+\alpha})\\ &= \frac{\exp(jLC_t\Delta x)-1}{LC_t\Delta x}O(\Delta x^{1+\alpha})\\ &= \frac{\exp(jLC_t\Delta x)-1}{LC_t}O(\Delta x^{\alpha})\end{aligned}$$

as $\Delta x \to 0$ for all $j \in \{1,\dots,m\}$. By definition of the time grid we have $m \cdot C_t \cdot \Delta x = T$ and since $j \leq m$ we have $jC_t\Delta x \leq T$. By substituting that inequality in the above bound we get

$$\|\dot{\Xi}_k^j - \dot{\xi}_k(t_j,p_0)\| \leq \frac{\exp(L\cdot T)-1}{C_tL}O(\Delta x^{\alpha}) = O(\Delta x^{\alpha})$$

as $\Delta x \to 0$ for all $j \in \{1,\dots,m\}$. This concludes the proof. □

6.3 Discrete Generalized Tangent Sensitivity

In this section we show how the generalized tangent introduced in Section 5.3 can be numerically approximated based on the assumptions made in the first section of this chapter and the error convergence results of the previous section. We first present a numerical approximation of the broad tangent that converges to the exact broad tangent of Section 5.3.1 in the sense of the L_1 norm as $\Delta x \to 0$. Based on the numerical approximation of the broad tangent and the numerical approximation of the shock locations and sensitivities we then present the numerical approximation of the generalized tangent first-order Taylor expansion that has a quadratic error bound in terms of $\dot{p}$ in the sense of the L_1 norm when choosing the grid distance Δx dependent on the value of the parameter perturbation $\dot{p}$.

6.3.1 Broad Tangent

Based on the assumed convergence of the discrete tangent sensitivity in Assumption 10 we now define the numerical broad tangent sensitivity.

Definition 40. The *numerical broad tangent* sensitivity is

$$\dot{U}_{\text{broad}}(t,x) \equiv \begin{cases} 0 & \text{if } x \in \left(\Xi_k^j - C_r\Delta x^{\alpha}, \Xi_k^j + C_r\Delta x^{\alpha}\right) \text{ for some } k \in \{1,\dots,K\} \\ \dot{U}(t,x) & \text{otherwise}\end{cases}$$

with C_r and α as in Definition 39.

We use the notation of Section 3.2 where in this case $\dot{U}(t,x)$ refers to a piecewise constant interpolation of the discrete tangent sensitivity $\dot{U}$.

The construction of the numerical broad tangent sensitivity is the discrete analogy to the analytic broad tangent sensitivity in the sense that it is the mathematical object we obtain by removing the divergent part of the discrete tangent sensitivity that is due to the singularity at the shock location in the analytic tangent sensitivity. The reason that the approximation converges to the analytic broad tangent in the sense of the L_1 norm is that the width of the shock regions that are cut out converges to zero.

The numerical broad tangent converges to the analytic broad tangent in the sense of the L_1 norm.

Proposition 39.

$$\|\dot{U}_{\text{broad}}(t,\cdot) - \dot{u}_{\text{broad}}(t,\cdot,p_0)\|_{L_1(\hat{\Omega})} = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all $t \in [0,T]$ with α as in Definition 39.

Proof. We show separate error bounds for inside and outside of the shock regions and then combine them to get an L_1 error bound for the whole domain.

Let

$$S = \bigcup_{k=1}^{K} \left(\xi_k(t,p_0) - C_r\Delta x^\alpha, \xi_k(t,p_0) + C_r\Delta x^\alpha\right) .$$

By definition of the numerical broad tangent we have

$$\dot{U}_{\text{broad}}(t,x) = \dot{U}(t,x)$$

for all $x \in \hat{\Omega} \setminus S$ and so by Assumption 10 we have

$$\begin{aligned}
\int_{\hat{\Omega}\setminus S} |\dot{U}_{\text{broad}}(t,x) - \dot{u}_{\text{broad}}(t,x,p_0)|\mathrm{d}x &= \int_{\hat{\Omega}\setminus S} O(\Delta x^\alpha)\mathrm{d}x \\
&= O(\Delta x^\alpha) \int_{\hat{\Omega}\setminus S} \mathrm{d}x \\
&\leq O(\Delta x^\alpha)(R-L) = O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$.

By definition of the numerical broad tangent we also have

$$\dot{U}_{\text{broad}}(t,x) = 0$$

for all $x \in S$ and so

$$\int_S |\dot{U}_{\text{broad}}(t,x) - \dot{u}_{\text{broad}}(t,x,p_0)|\mathrm{d}x = \int_S |\dot{u}_{\text{broad}}(t,x,p_0)|\mathrm{d}x .$$

By Assumption 6 and the definition of the broad tangent $\dot{u}_{\text{broad}}$ is bounded on $\hat{\Omega}$ and, thus, also on $S \subseteq \hat{\Omega}$. We have

$$\begin{aligned}
\int_S |\dot{u}_{\text{broad}}(t,x,p_0)|\mathrm{d}x &\leq \int_S O(1)\mathrm{d}x \\
&= O(1)\int_S \mathrm{d}x \\
&= O(1)\sum_{k=1}^{K} 2C_r\Delta x^\alpha \\
&= O(1)2KC_r\Delta x^\alpha = O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$.

Since $\hat{\Omega} = (\hat{\Omega} \setminus S) \cup S$ by definition of the L_1 norm we have

$$\begin{aligned}
&\|\dot{U}_{\text{broad}}(t,\cdot) - \dot{u}_{\text{broad}}(t,\cdot,p_0)\|_{L_1(\hat{\Omega})} \\
&\quad = \int_{\hat{\Omega}} |\dot{U}_{\text{broad}}(t,\cdot) - \dot{u}_{\text{broad}}(t,\cdot,p_0)| \mathrm{d}x \\
&\quad = \int_{\hat{\Omega}\setminus S} |\dot{U}_{\text{broad}}(t,\cdot) - \dot{u}_{\text{broad}}(t,\cdot,p_0)| \mathrm{d}x + \int_S |\dot{U}_{\text{broad}}(t,\cdot) - \dot{u}_{\text{broad}}(t,\cdot,p_0)| \mathrm{d}x \\
&\quad = O(\Delta x^\alpha) + O(\Delta x^\alpha) = O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x \to 0$. This concludes the proof. □

6.3.2 Nonlinear Generalized Tangent

Based on the numerical broad tangent and the numerical shock locations and sensitivities we now define the numerical generalized tangent sensitivity.

Definition 41. The (nonlinear) *numerical generalized tangent* sensitivity of the weak solution in the direction $\dot{p} \in \mathbb{R}^d$ is

$$\begin{aligned}
\dot{U}_{\text{gen}}(t_j, x) \equiv & \sum_{\substack{k=1 \\ \dot{\Xi}_k^j > 0}}^{K} \chi_{[\Xi_k^j, \Xi_k^j + \dot{\Xi}_k^j]}(x) \big(U(t_j, \Xi_k^j - C_r \Delta x^\alpha) - U(t_j, \Xi_k^j + C_r \Delta x^\alpha) \big) \\
& - \sum_{\substack{k=1 \\ \dot{\Xi}_k^j < 0}}^{K} \chi_{[\Xi_k^j + \dot{\Xi}_k^j, \Xi_k^j]}(x) \big(U(t_j, \Xi_k^j - C_r \Delta x^\alpha) - U(t_j, \Xi_k^j + C_r \Delta x^\alpha) \big) \\
& + \dot{U}_{\text{broad}}(t_j, x)
\end{aligned}$$

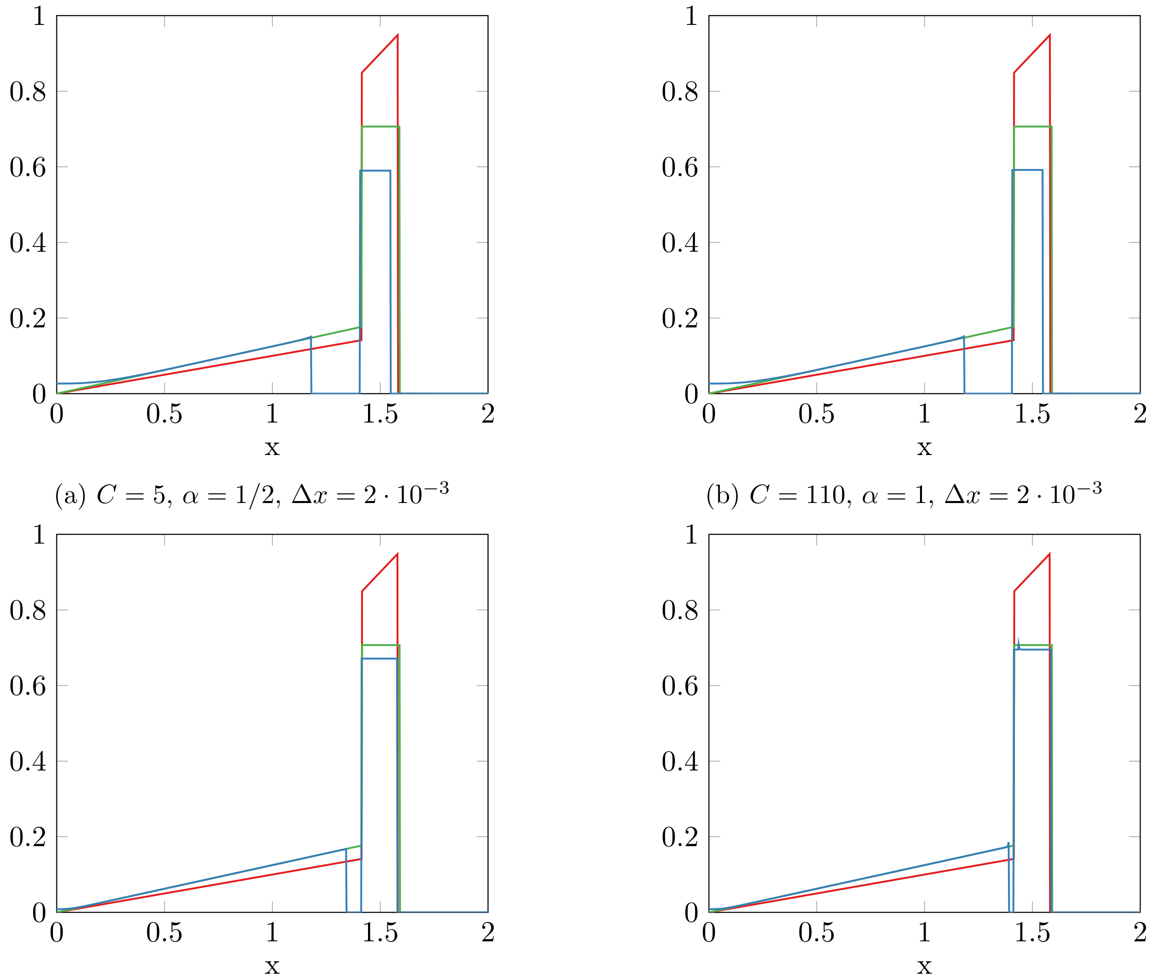
with C_r and α as in Definition 39.

We use the notation of Section 3.2 where in this case $\dot{U}_{\text{broad}}(t,x)$ refers to a piecewise constant interpolation of the numerical broad tangent $\dot{U}_{\text{broad}}$.

The construction of the numerical generalized tangent is analogous to the analytic generalized tangent. Instead of the exact shock locations, shock sensitivities, weak solution and broad tangent we use their numerical approximations. Notably, the jump height of the shock discontinuities in the weak solution is approximated according to the idea of Lemma 37.

Example 38 (Ramp)**.** We consider the Lax-Friedrichs discretization of the ramp initial value problem from Example 7. The shock location is obtained using the shock location algorithm and the shock sensitivity is obtained with the shock sensitivity algorithm for different values of $C_r \equiv C$ and α.

Figure 6.10 shows the divided difference of the analytic weak solution evaluated at $p = p_0$ and $p = p_0 + \dot{p}$ with $p_0 = 0$ and $\dot{p} = 1$. The analytic generalized tangent sensitivity is shown in green and the numerical generalized tangent sensitivity is shown in blue. The two figures on the top



(a) $C = 5$, $\alpha = 1/2$, $\Delta x = 2 \cdot 10^{-3}$

(b) $C = 110$, $\alpha = 1$, $\Delta x = 2 \cdot 10^{-3}$

(c) $C = 5$, $\alpha = 1/2$, $\Delta x = 2 \cdot 10^{-4}$

(d) $C = 110$, $\alpha = 1$, $\Delta x = 2 \cdot 10^{-4}$

Figure 6.10: $u(p_0 + \dot{p}) - u(p_0)$ (red), analytic $\dot{u}_{\text{gen}}$ (green), Lax-Friedrichs $\dot{U}_{\text{gen}}$ (blue)

show a less accurate discretization and the two figures on the bottom show a more accurate discretization.

The first observation we make is that for the less accurate discretization the height of the box that approximates the shock travel in the numerical generalized tangent is less than that of the analytic generalized tangent. The reason for that is that the weak solution further to the left of the shock decreases linearly so since we approximate the jump height based on evaluating at $\Xi_1 - C\Delta x^\alpha$ instead of ξ_1- we underestimate the analytic jump height. But the difference and, hence, the approximation error becomes smaller as $\Delta x \to 0$ and evidently faster so with the choice of $\alpha = 1$.

The second observation we make is that the numerical generalized tangent is zero in a region left to the shock where the analytical generalized tangent is not zero. This approximation error results from the way the numerical broad tangent is defined. We cut out the divergent part of the discrete tangent sensitivity approximation in a $\Xi_1 \pm C\Delta x^\alpha$ region. In that region the numerical broad tangent and, hence, the corresponding term in the numerical generalized tangent is zero. That zero region and as a result the approximation error becomes smaller as $\Delta x \to 0$ and, again, evidently faster so with the choice of $\alpha = 1$.

The third and final observation we make is that there is a small spike at the top of the shock travel approximating box in the plot with $C = 110, \alpha = 1, \Delta x = 2 \cdot 10^{-4}$. The small spike relates to the previous point as it is a part of the discrete tangent sensitivity where the sensitivity starts to diverge approximating the singularity. That means the shock region is possibly not wide enough as $\Delta x \to 0$ for $\alpha = 1$. In order to better understand this phenomenon we can again consider the shape of the discrete tangent in Figure 6.7 and Figure 6.8.

Based on the convergence of all involved numerical approximation errors at the rate of $O(\Delta x^\alpha)$ the numerical generalized tangent converges to the analytic generalized tangent in the sense of the L_1 norm at the same rate.

Proposition 40.

$$\|\dot{U}_{\text{gen}}(t_j, \cdot) - \dot{u}_{\text{gen}}(t_j, \cdot, p_0)\|_{L_1(\hat{\Omega})} = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all $j \in \{0, \dots, m\}$ with α as in Definition 39.

Proof. We bound the errors corresponding to the three terms in the definition of the analytic and generalized tangents separately.

By triangle inequality we have

$$\begin{aligned}\|\dot{U}_{\text{gen}}(t_j, \cdot) - \dot{u}_{\text{gen}}(t_j, \cdot, p_0)\|_{L_1(\hat{\Omega})} &= \|E_1 + E_2 + E_3\|_{L_1(\hat{\Omega})} \\ &\leq \|E_1\|_{L_1(\hat{\Omega})} + \|E_2\|_{L_1(\hat{\Omega})} + \|E_3\|_{L_1(\hat{\Omega})}\end{aligned}$$

with

$$\begin{aligned}\|E_1\|_{L_1(\hat{\Omega})} = \Bigg\| & \sum_{\substack{k=1 \\ \dot{\Xi}_k^j > 0}}^{K} \chi_{[\Xi_k^j, \Xi_k^j + \dot{\Xi}_k^j]}(\cdot) \big(U(t_j, \Xi_k^j - C_r \Delta x^\alpha) - U(t_j, \Xi_k^j + C_r \Delta x^\alpha)\big) \\ & - \sum_{\substack{k=1 \\ \dot{\xi}_k(t_j, p_0) > 0}}^{K} \chi_{[\xi_k(t_j,p_0), \xi_k(t_j,p_0) + \dot{\xi}_k(t_j,p_0)]}(\cdot) \\ & \qquad \cdot \big(u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)\big) \Bigg\|_{L_1(\hat{\Omega})} \\ \|E_2\|_{L_1(\hat{\Omega})} = \Bigg\| & \sum_{\substack{k=1 \\ \dot{\Xi}_k^j < 0}}^{K} \chi_{[\Xi_k^j + \dot{\Xi}_k^j, \Xi_k^j]}(\cdot) \big(U(t_j, \Xi_k^j - C_r \Delta x^\alpha) - U(t_j, \Xi_k^j + C_r \Delta x^\alpha)\big) \\ & - \sum_{\substack{k=1 \\ \dot{\xi}_k(t_j, p_0) < 0}}^{K} \chi_{[\xi_k(t_j,p_0) + \dot{\xi}_k(t_j,p_0), \xi_k(t_j,p_0)]}(\cdot) \\ & \qquad \cdot \big(u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)\big) \Bigg\|_{L_1(\hat{\Omega})} \\ \|E_3\|_{L_1(\hat{\Omega})} = \| & \dot{U}_{\text{broad}}(t_j, \cdot) - \dot{u}_{\text{broad}}(t_j, \cdot, p_0)\|_{L_1(\hat{\Omega})} .\end{aligned}$$

We bound the three errors separately. For the third error we directly have

$$\|E_3\|_{L_1(\hat{\Omega})} = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ by Proposition 39.

The first and second error are analogous by symmetry and we only show the details for the E_1 bound. First we need to look at what terms actually occur in these errors based on the sign of the shock sensitivities.

By Proposition 38 we have

$$\dot{\Xi}_k^j = \dot{\xi}_k(t_j, p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all $k \in \{1, \dots, K\}$. For all $\dot{\xi}_k(t_j, p_0) > 0$ we differentiate two cases.

In the first case there is no $\delta > 0$ such that $\dot{\xi}_k(t_j, p_0) \geq \delta$, i.e. the shock sensitivity is "infinitely small". For the corresponding numerical shock sensitivity $\dot{\Xi}_k^j$ it is not clear whether it is greater or less than zero for $\Delta x > 0$.

In this case the term corresponding to ξ_k belongs to the sum in E_1 but the term corresponding to Ξ_k might belong to the sum in E_2. But the width of the shock travel approximation due to the characteristic function is only $O(\Delta x^\alpha)$ as $\Delta x \to 0$. And the characteristic function is multiplied by a constant. So the overall contribution of the L_1 error for such shocks can at most be $O(\Delta x^\alpha)$ as well no matter whether as part of E_1 or E_2. In that sense shock sensitivities that are infinitely small can be treated the same as shock sensitivities that are zero valued in the context of our error analysis.

In the second case the shock sensitivity is bounded away from zero and we have $\dot{\xi}_k(t_j, p_0) \geq \delta$ with some $\delta > 0$. Then for some $C > 0$ and $(\delta/C)^{1/\alpha} > \delta_2 > 0$ for all $\Delta x \leq \delta_2$ we have

$$\begin{aligned} \dot{\Xi}_k^j &\geq \dot{\xi}_k(t_j, p_0) - C\Delta x^\alpha \\ &\geq \delta - C\Delta x^\alpha \\ &\geq \delta - C\delta_2^\alpha \\ &> \delta - \delta = 0\,. \end{aligned}$$

So in this case we have $\dot{\xi}(t_j, p_0) > 0$ and $\dot{\Xi}_k^j > 0$ as $\Delta x \to 0$.

Differentiating the two cases where we use the notation $\delta > \dot{\xi}_k(t_j, p_0) > 0$ to denote the first case and $\dot{\xi}_k(t_j, p_0) \geq \delta > 0$ to denote the second case we have

$$\begin{aligned} \|E_1\|_{L_1(\hat{\Omega})} = \Big\| &\sum_{\substack{k=1 \\ \dot{\Xi}_k^j > 0}}^{K} \chi_{[\Xi_k^j, \Xi_k^j + \dot{\Xi}_k^j]}(\cdot)\big(U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big) \\ &- \sum_{\substack{k=1 \\ \dot{\xi}_k(t_j,p_0) \geq \delta > 0}}^{K} \chi_{[\xi_k(t_j,p_0), \xi_k(t_j,p_0) + \dot{\xi}_k(t_j,p_0)]}(\cdot) \\ &\qquad\qquad \cdot \big(u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)\big) \\ &- \sum_{\substack{k=1 \\ \delta > \dot{\xi}_k(t_j,p_0) > 0}}^{K} \chi_{[\xi_k(t_j,p_0), \xi_k(t_j,p_0) + \dot{\xi}_k(t_j,p_0)]}(\cdot) \\ &\qquad\qquad \cdot \big(u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)\big)\Big\|_{L_1(\hat{\Omega})} \end{aligned}$$

$$
\begin{aligned}
&= \Big\| \sum_{\substack{k=1\\ \dot{\xi}_k(t_j,p_0)\geq\delta>0}}^{K} \Big(\chi_{[\Xi_k^j,\Xi_k^j+\dot{\Xi}_k^j]}(\cdot)\big(U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)\big) \\
&\qquad\qquad - \chi_{[\xi_k(t_j,p_0),\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)]}(\cdot) \\
&\qquad\qquad \cdot \big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)\Big) \\
&\quad + \sum_{\substack{k=1\\ \dot{\Xi}_k^j>0\\ \delta<\dot{\xi}_k(t_j,p_0)<0}}^{K} \chi_{[\Xi_k^j,\Xi_k^j+\dot{\Xi}_k^j]}(\cdot)\big(U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)\big) \\
&\quad - \sum_{\substack{k=1\\ \delta>\dot{\xi}_k(t_j,p_0)>0}}^{K} \chi_{[\xi_k(t_j,p_0),\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)]}(\cdot) \\
&\qquad\qquad \cdot \big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)\Big\|_{L_1(\hat{\Omega})} \\
&\leq \sum_{\substack{k=1\\ \dot{\xi}_k(t_j,p_0)\geq\delta>0}}^{K} \Big\| \chi_{[\Xi_k^j,\Xi_k^j+\dot{\Xi}_k^j]}(\cdot)\big(U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)\big) \\
&\qquad\qquad - \chi_{[\xi_k(t_j,p_0),\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)]}(\cdot) \\
&\qquad\qquad \cdot \big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)\Big\|_{L_1(\hat{\Omega})} \\
&\quad + \sum_{\substack{k=1\\ \dot{\Xi}_k^j>0\\ \delta<\dot{\xi}_k(t_j,p_0)<0}}^{K} \Big\| \chi_{[\Xi_k^j,\Xi_k^j+\dot{\Xi}_k^j]}(\cdot)\big(U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)\big)\Big\|_{L_1(\hat{\Omega})} \\
&\quad + \sum_{\substack{k=1\\ \delta>\dot{\xi}_k(t_j,p_0)>0}}^{K} \Big\| \chi_{[\xi_k(t_j,p_0),\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)]}(\cdot) \\
&\qquad\qquad \cdot \big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)\Big\|_{L_1(\hat{\Omega})}
\end{aligned}
$$

as $\Delta x \to 0$ where the last step is due to the triangle inequality.

The last two terms are both $O(\Delta x^\alpha)$ since they are finite sums of L_1 norms that are $O(\Delta x^\alpha)$ because the corresponding $\dot{\xi}_k(t_j,p_0)$ are infinitely small and the corresponding $\dot{\Xi}_k^j$ are $O(\Delta x^\alpha)$. By definition of the L_1 norm we have

$$
\Big\| \chi_{[\Xi_k^j,\Xi_k^j+\dot{\Xi}_k^j]}(\cdot)\big(U(t_j,\Xi_k^j - C_r\Delta x^\alpha) - U(t_j,\Xi_k^j + C_r\Delta x^\alpha)\big)\Big\|_{L_1(\hat{\Omega})} = \dot{\Xi}_k^j O(1) = O(\Delta x^\alpha)
$$

as $\Delta x \to 0$ for all $k \in \{1,\ldots,K\}$ with $\dot{\xi}_k < 0$ infinitely small. We also have

$$
\begin{aligned}
&\Big\| \chi_{[\xi_k(t_j,p_0),\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)]}(\cdot)\big(u_k(t_j,\xi_k(t_j,p_0),p_0) - u_{k+1}(t_j,\xi_k(t_j,p_0),p_0)\big)\Big\|_{L_1(\hat{\Omega})} \\
&= \dot{\xi}_k(t_j,p_0)O(1) \\
&= O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$ for all $k \in \{1, \dots, K\}$ with $\dot{\xi}_k > 0$ infinitely small.

We now further analyze the first term.

By Proposition 36 we have

$$\Xi_k^j = \xi_k(t_j, p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all $k \in \{1, \dots, K\}$. With $\dot{\xi}_k(t_j, p_0) \geq \delta > 0$ there are some $(\delta/C)^{1/\alpha} > \delta_2 > 0$, $C > 0$ such that for all $\Delta x \leq \delta_2$ we have

$$\begin{aligned}
\Xi_k^j &\leq \xi_k(t_j, p_0) + C\Delta x^\alpha \\
&\leq \xi_k(t_j, p_0) + C\delta_2^\alpha \\
&< \xi_k(t_j, p_0) + \delta \\
&\leq \xi_k(t_j, p_0) + \dot{\xi}_k(t_j, p_0)\ .
\end{aligned}$$

Similary, we have with $\dot{\xi}_k(t_j, p_0) \geq \delta > 0$ there are some $(\delta/C)^{1/\alpha} > \delta_2 > 0$, $C > 0$ such that for all $\Delta x \leq \delta_2$ we have

$$\begin{aligned}
\Xi_k^j + \dot{\Xi}_k^j &\geq \xi_k(t_j, p_0) + \dot{\xi}_k(t_j, p_0) - C\Delta x^\alpha \\
&\geq \xi_k(t_j, p_0) + \delta - C\delta_2^\alpha \\
&> \xi_k(t_j, p_0)\ .
\end{aligned}$$

Given that we have both $\Xi_k^j < \xi_k(t_j, p_0) + \dot{\xi}_k(t_j, p_0)$ and $\xi_k(t_j, p_0) < \Xi_k^j + \dot{\Xi}_k^j$ we have to differentiate four cases for each term corresponding to a shock location. We treat the following four cases:

(i) $\xi_k(t_j, p_0) \leq \Xi_k^j$, $\xi_k(t_j, p_0) + \dot{\xi}_k(t_j, p_0) \leq \Xi_k^j + \dot{\Xi}_k^j$

(ii) $\xi_k(t_j, p_0) > \Xi_k^j$, $\xi_k(t_j, p_0) + \dot{\xi}_k(t_j, p_0) \leq \Xi_k^j + \dot{\Xi}_k^j$

(iii) $\xi_k(t_j, p_0) > \Xi_k^j$, $\xi_k(t_j, p_0) + \dot{\xi}_k(t_j, p_0) > \Xi_k^j + \dot{\Xi}_k^j$

(iv) $\xi_k(t_j, p_0) \leq \Xi_k^j$, $\xi_k(t_j, p_0) + \dot{\xi}_k(t_j, p_0) > \Xi_k^j + \dot{\Xi}_k^j$

We recall that by Lemma 37 we have

$$\begin{aligned}
U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - u_k(t_j, \xi_k(t_j, p_0), p_0) = O(\Delta x^\alpha) \\
U(t_j, \Xi_k^j + C_r\Delta x^\alpha) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0) = O(\Delta x^\alpha)\ .
\end{aligned}$$

For case (i) by definition of the L_1 norm we have

$$\begin{aligned}
&\Big\| \chi_{[\Xi_k^j, \Xi_k^j + \dot{\Xi}_k^j]}(\cdot)\big(U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big) \\
&\qquad - \chi_{[\xi_k(t_j,p_0), \xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)]}(\cdot)\big(u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)\big)\Big\|_{L_1(\hat{\Omega})} \\
&= \int_{\xi_k(t_j,p_0)}^{\Xi_k^j} \big|u_k(t_j, \xi_k(t_j, p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j, p_0), p_0)\big| \mathrm{d}x \\
&\qquad + \int_{\Xi_k^j}^{\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)} \big|\big(U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big)
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad - \big(u_k(t_j, \xi_k(t_j,p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j,p_0), p_0)\big)\big|\mathrm{d}x \\
&\quad + \int_{\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)}^{\Xi_k^j + \dot{\Xi}_k^j} \big|U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big|\mathrm{d}x \\
&= \int_{\xi_k(t_j,p_0)}^{\Xi_k^j} O(1)\mathrm{d}x + \int_{\Xi_k^j}^{\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)} O(\Delta x^\alpha)\mathrm{d}x + \int_{\xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)}^{\Xi_k^j+\dot{\Xi}_k^j} O(1)\mathrm{d}x \\
&= \big(\Xi_k^j - \xi_k(t_j,p_0)\big)O(1) + \big(\xi_k(t_j,p_0) + \dot{\xi}_k(t_j,p_0) - \Xi_k^j\big)O(\Delta x^\alpha) \\
&\quad + \big(\Xi_k^j + \dot{\Xi}_k^j - (\xi_k(t_j,p_0) + \dot{\xi}_k(t_j,p_0))\big)O(1) \\
&= O(\Delta x^\alpha) + O(\Delta x^\alpha) + O(\Delta x^\alpha) = O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$.

The other three cases are analogous to case (i) in that the L_1 norm decomposes into three parts of the domain where the integral defining the norm is nonzero. The outer parts integrate a bounded constant over a domain of width $O(\Delta x^\alpha)$ and the inner part integrates an $O(\Delta x^\alpha)$ approximation error over a domain of bounded constant width. We have

$$
\begin{aligned}
&\Big\|\chi_{[\Xi_k^j, \Xi_k^j + \dot{\Xi}_k^j]}(\cdot)\big(U(t_j, \Xi_k^j - C_r\Delta x^\alpha) - U(t_j, \Xi_k^j + C_r\Delta x^\alpha)\big) \\
&\qquad - \chi_{[\xi_k(t_j,p_0), \xi_k(t_j,p_0)+\dot{\xi}_k(t_j,p_0)]}(\cdot)\big(u_k(t_j, \xi_k(t_j,p_0), p_0) - u_{k+1}(t_j, \xi_k(t_j,p_0), p_0)\big)\Big\|_{L_1(\hat{\Omega})} \\
&= O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$ for all $k \in \{1, \dots, K\}$ with $\dot{\xi}_k \geq \delta > 0$.

In combination with the previous bounds for the case where the shock sensitivity is infinitely small and by $K < \infty$ we have

$$
\|E_1\|_{L_1(\hat{\Omega})} = \sum_{\substack{k=1 \\ \dot{\xi}_k(t_j,p_0) \geq \delta > 0}}^{K} O(\Delta x^\alpha) + \sum_{\substack{k=1 \\ \dot{\Xi}_k^j > 0 \\ \delta < \dot{\xi}_k(t_j,p_0) < 0}}^{K} O(\Delta x^\alpha) + \sum_{\substack{k=1 \\ \delta > \dot{\xi}_k(t_j,p_0) > 0}}^{K} O(\Delta x^\alpha) = O(\Delta x^\alpha)
$$

as $\Delta x \to 0$. This concludes the bound for E_1.

The bound for E_2 is completely analogous to that of E_1 by symmetry of the shock sensitivity about zero and can be shown by flipping signs correspondingly. We, thus, also have

$$
\|E_2\|_{L_1(\hat{\Omega})} = O(\Delta x^\alpha)
$$

as $\Delta x \to 0$. The bound for E_3 follows directly from Proposition 39 and so by substitution of all three bounds we have

$$
\|\dot{U}_{\mathrm{gen}}(t_j, \cdot) - \dot{u}_{\mathrm{gen}}(t_j, \cdot, p_0)\|_{L_1(\hat{\Omega})} \leq \|E_1\|_{L_1(\hat{\Omega})} + \|E_2\|_{L_1(\hat{\Omega})} + \|E_3\|_{L_1(\hat{\Omega})} = O(\Delta x^\alpha)
$$

as $\Delta x \to 0$. This concludes the proof. □

We now show that by picking a sufficiently quickly descreasing discretization grid distance Δx the numerical generalized tangent sensitivity defined in the previous section satisfies the same first-order Taylor expansion error bound in terms of $\dot{p}$ as we showed for the analytic generalized tangent sensitivity in Proposition 26.

Proposition 41. Let $\Delta x = O(\|\dot{p}\|^{2/\alpha})$ with α as in Definition 39 then

$$\|u(t_j, \cdot, p_0 + \dot{p}) - u(t_j, \cdot, p_0) - \dot{U}_{\text{gen}}(t_j, \cdot)\|_{L^1(\hat{\Omega})} = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$.

Proof. We combine the bound of the error between the analytic generalized tangent and the perturbation with the bound of the error between the numerical generalized tangent and the analytic generalized tangent for the assumed choice of Δx.

By adding zero and by the triangle inequality we have

$$\begin{aligned}
&\|u(t_j, \cdot, p_0 + \dot{p}) - u(t_j, \cdot, p_0) - \dot{U}_{\text{gen}}(t_j, \cdot)\|_{L^1(\hat{\Omega})} \\
&\quad = \|u(t_j, \cdot, p_0 + \dot{p}) - u(t_j, \cdot, p_0) - \dot{u}_{\text{gen}}(t_j, \cdot, p_0) + \dot{u}_{\text{gen}}(t_j, \cdot, p_0) - \dot{U}_{\text{gen}}(t_j, \cdot)\|_{L^1(\hat{\Omega})} \\
&\quad \leq \|u(t_j, \cdot, p_0 + \dot{p}) - u(t_j, \cdot, p_0) - \dot{u}_{\text{gen}}(t_j, \cdot, p_0)\|_{L^1(\hat{\Omega})} + \|\dot{u}_{\text{gen}}(t_j, \cdot, p_0) - \dot{U}_{\text{gen}}(t_j, \cdot)\|_{L^1(\hat{\Omega})} .
\end{aligned}$$

By Proposition 26 and Proposition 40 we have

$$\begin{aligned}
&\|u(t_j, \cdot, p_0 + \dot{p}) - u(t_j, \cdot, p_0) - \dot{u}_{\text{gen}}(t_j, \cdot, p_0)\|_{L^1(\hat{\Omega})} + \|\dot{u}_{\text{gen}}(t_j, \cdot, p_0) - \dot{U}_{\text{gen}}(t_j, \cdot)\|_{L^1(\hat{\Omega})} \\
&\quad = O(\|\dot{p}\|^2) + O(\Delta x^\alpha)
\end{aligned}$$

as $\Delta x, \dot{p} \to 0$.

By assumption of the proposition we have $\Delta x = O(\|\dot{p}\|^{2/\alpha})$, i.e. $\Delta x^\alpha = O(\|\dot{p}\|^2)$ as $\dot{p} \to 0$. Hence, we have the total approximation error norm bounded by

$$\|u(t_j, \cdot, p_0 + \dot{p}) - u(t_j, \cdot, p_0) - \dot{U}_{\text{gen}}(t_j, \cdot)\|_{L^1(\hat{\Omega})} = O(\|\dot{p}\|^2) + O(\Delta x^\alpha) = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$. This concludes the proof. □

6.4 Discrete Adjoint Sensitivity

In this section we show how the adjoint sensitivity analysis of an integral objective of the form

$$\Phi \equiv \int_\Omega \varphi(u(T, x, p_0))\mathrm{d}x$$

as introduced in Section 1.3.1 and analytically derived for the piecewise smooth solution structure in Section 5.4 can be numerically approximated based on the assumptions and the error convergence results of the first two sections of this chapter. We first present a numerical approximation of the tangent sensitivity of the integral objective that contrary to the naive approach discussed in Section 1.3.8 actually converges to the analytic tangent sensitivities as $\Delta x \to 0$. We then show the convergence of the discrete adjoint sensitivities of the integral objective based on Proposition 38, i.e. by utilizing the discrete adjoint of the convergent discrete tangent sensitivity.

6.4.1 Sensitivities of the Integral Objective

We now define a numerical tangent sensitivity of the integral objective that explicitly makes use of the numerical approximations of shock locations and shock sensitivities.

Definition 42. The *explicit shock tangent* sensitivity of the integral objective in the direction $\dot{p} \in \mathbb{R}^d$ is

$$\begin{aligned}\dot{\Phi}_{\text{shock}} \equiv \sum_{k=1}^{K+1} \sum_{i=\mathrm{r}(\Xi_{k-1}^m + C_r \Delta x^\alpha)}^{\mathrm{l}(\Xi_k^m - C_r \Delta x^\alpha)} & \varphi'(U(t_m, x_i))\dot{U}(t_m, x_i)\Delta x \\ + \sum_{k=1}^{K} & \dot{\Xi}_k^m \big(\varphi(U(t_m, \Xi_k^m - C_r\Delta x^\alpha)) - \varphi(U(t_m, \Xi_k^m + C_r \Delta x^\alpha))\big)\end{aligned}$$

with C_r and α as in Definition 39 and where

$$\begin{aligned}\mathrm{r}(x) &= x + \min\{x_i - x \mid x_i \geq x, i \in \mathbb{N}\} \\ \mathrm{l}(x) &= x - \min\{x - x_i \mid x_i \leq x, i \in \mathbb{N}\}\ .\end{aligned}$$

The explicit shock tangent is an explicit approximation of the analytical tangent sensitivity

$$\begin{aligned}\dot{\Phi} = \sum_{k=1}^{K+1} \int_{\xi_{k-1}(T,p_0)}^{\xi_k(T,p_0)} & \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x \\ + \sum_{k=1}^{K} & \dot{\xi}_k(T, p_0)\big(\varphi(u_k(T, \xi_k(T, p_0), p_0)) - \varphi(u_{k+1}(T, \xi_k(T, p_0), p_0))\big)\end{aligned}$$

that is obtained by chain rule after splitting the integral at the shock locations (see Section 1.3.3). Based on Assumption 8 and Assumption 10 the approximation of the integrals in the first sum requires us to cut out sufficiently large approximations of the shock regions (also see the explanation in Section 1.4.1). The $\mathrm{l}(\cdot)$ and $\mathrm{r}(\cdot)$ maps are used to pick grid points that are guaranteed to lie left of or right of the corresponding shock regions respectively.

Proposition 42.

$$\|\dot{\Phi}_{\text{shock}} - \dot{\Phi}\| = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ with α as in Definition 39.

Proof. By triangle inequality we can decompose the error bound into the following sum

$$\begin{aligned}\|\dot{\Phi}_{\text{shock}} - \dot{\Phi}\| = \Bigg\| \sum_{k=1}^{K+1} \sum_{i=\mathrm{r}(\Xi_{k-1}^m + C_r \Delta x^\alpha)}^{\mathrm{l}(\Xi_k^m - C_r \Delta x^\alpha)} & \varphi'(U(t_m, x_i))\dot{U}(t_m, x_i)\Delta x && (6.4.1) \\ - \int_{\xi_{k-1}(T,p_0)}^{\xi_k(T,p_0)} & \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x && (6.4.2) \\ + \sum_{k=1}^{K} \dot{\Xi}_k^m & \big(\varphi(U(t_m, \Xi_k^m - C_r\Delta x^\alpha)) - \varphi(U(t_m, \Xi_k^m + C_r \Delta x^\alpha)) && (6.4.3) \\ - \dot{\xi}_k(T, p_0) & \big(\varphi(u_k(T, \xi_k(T, p_0), p_0)) - \varphi(u_{k+1}(T, \xi_k(T, p_0), p_0))\big)\big)\Bigg\| && (6.4.4) \\ \leq \sum_{k=1}^{K+1} \Bigg\| \sum_{i=\mathrm{r}(\Xi_{k-1}^m + C_r \Delta x^\alpha)}^{\mathrm{l}(\Xi_k^m - C_r \Delta x^\alpha)} & \varphi'(U(t_m, x_i))\dot{U}(t_m, x_i)\Delta x && (6.4.5)\end{aligned}$$

$$- \int_{\xi_{k-1}(T,p_0)}^{\xi_k(T,p_0)} \varphi'(u_k(T,x,p_0))\dot{u}_k(T,x,p_0)\mathrm{d}x \Big\| \tag{6.4.6}$$

$$+ \sum_{k=1}^{K} \Big\| \dot{\Xi}_k^m \big(\varphi(U(t_m, \Xi_k^m - C_r\Delta x^\alpha)) - \varphi(U(t_m, \Xi_k^m + C_r\Delta x^\alpha)) \tag{6.4.7}$$

$$- \dot{\xi}_k(T,p_0)\big(\varphi(u_k(T,\xi_k(T,p_0),p_0)) - \varphi(u_{k+1}(T,\xi_k(T,p_0),p_0))\big)\big)\Big\| . \tag{6.4.8}$$

We first consider the individual terms of the second sum over the shock index.

Because according to Assumption 7 the function φ is continuously differentiable by Lemma 7 it is also locally Lipschitz continuous. By Lemma 37 and the local Lipschitz continuity of φ we have

$$\begin{aligned}\varphi(U(t_m, \Xi_k^m - C_r\Delta x^\alpha)) &= \varphi(u_k(t_m, \xi_k(t_m,p_0),p_0) + O(\Delta x^\alpha)) \\ &= \varphi(u_k(t_m, \xi_k(t_m,p_0),p_0)) + O(\Delta x^\alpha)\end{aligned}$$

as $\Delta x \to 0$. Analogously, we also have

$$\varphi(U(t_m, \Xi_k^m + C_r\Delta x^\alpha)) = \varphi(u_{k+1}(t_m, \xi_k(t_m,p_0),p_0)) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$. Furthermore, by Proposition 38 we have

$$\dot{\Xi}_k^m = \dot{\xi}_k(t_m,p_0) + O(\Delta x^\alpha)$$

as $\Delta x \to 0$. By combining these error bounds and since the tangent sensitivity and weak solution are bounded, φ is locally Lipschitz continuous and $t_m = T$ we have the approximation error of the terms of the second sum over shock locations in the explicit shock tangent sensitivity as

$$\begin{aligned}&\dot{\Xi}_k^m \big(\varphi(U(t_m, \Xi_k^m - C_r\Delta x^\alpha)) - \varphi(U(t_m, \Xi_k^m + C_r\Delta x^\alpha)) \\ &\quad= \big(\dot{\xi}_k(T,p_0) + O(\Delta x^\alpha)\big)\big(\varphi(u_k(T,\xi_k(T,p_0),p_0)) - \varphi(u_{k+1}(T,\xi_k(T,p_0),p_0)) + O(\Delta x^\alpha)\big) \\ &\quad= \dot{\xi}_k(T,p_0)\big(\varphi(u_k(T,\xi_k(T,p_0),p_0)) - \varphi(u_{k+1}(T,\xi_k(T,p_0),p_0))\big) + O(\Delta x^\alpha)\end{aligned}$$

as $\Delta x \to 0$. By substitution the normed approximation error we have

$$\begin{aligned}&\Big\|\dot{\Xi}_k^m \big(\varphi(U(t_m, \Xi_k^m - C_r\Delta x^\alpha)) - \varphi(U(t_m, \Xi_k^m + C_r\Delta x^\alpha)) \\ &\quad - \dot{\xi}_k(T,p_0)\big(\varphi(u_k(T,\xi_k(T,p_0),p_0)) - \varphi(u_{k+1}(T,\xi_k(T,p_0),p_0))\big)\big)\Big\| = O(\Delta x^\alpha)\end{aligned}$$

as $\Delta x \to 0$ for all $k \in \{1,\dots,K\}$. Since K is finite the second sum of Equations (6.4.5)-(6.4.8) in total is $O(\Delta x^\alpha)$ as $\Delta x \to 0$.

We now consider the individual terms of the first sum over the shock index that contains the integral approximations.

By Proposition 36 we have

$$\Xi_k^m \leq \xi_k(t_m,p_0) + AC_s\Delta x^\alpha$$

as $\Delta x \to \alpha$. By subtracting $C_r\Delta x^\alpha$ from both sides and using its definition we have

$$\Xi_k^m - C_r\Delta x^\alpha \leq \xi_k(t_m,p_0) + AC_s\Delta x^\alpha - C_r\Delta x^\alpha$$

$$
\begin{aligned}
&= \xi_k(t_m, p_0) + AC_s\Delta x^\alpha - (A+1)C_s\Delta x^\alpha \\
&= \xi_k(t_m, p_0) - C_s\Delta x^\alpha
\end{aligned}
$$

as $\Delta x \to 0$. By the definition of $\mathrm{l}(\cdot)$ we have

$$
\begin{aligned}
x_i &\leq \Xi_k^m - C_r\Delta x^\alpha \\
&\leq \xi_k(t_m, p_0) - C_s\Delta x^\alpha
\end{aligned}
$$

for all $i \leq \mathrm{l} \equiv \mathrm{l}(\Xi_k^m - C_r\Delta x^\alpha)$. Analogously, by Proposition 36 and the definition of C_r and of $\mathrm{r}(\cdot)$ we have

$$
\begin{aligned}
x_i &\geq \Xi_{k-1}^m + C_r\Delta x^\alpha \\
&\geq \xi_{k-1}(t_m, p_0) + C_s\Delta x^\alpha
\end{aligned}
$$

for all $i \geq \mathrm{r} \equiv \mathrm{r}(\Xi_{k-1}^m + C_r\Delta x^\alpha)$. The terms of the integral approximating sum over the index i are between the two shocks ξ_{k-1} and ξ_k as $\Delta x \to 0$ since the shock locations are isolated and the x_i are far enough outside of the respective shock regions where the shock is smeared out. To show that for Δx sufficiently small all x_i are actually between the shock locations a similar argument as in the proof of Lemma 37 can be used.

By Assumption 7 the function φ is twice continuously differentiable and hence the derivative φ' is continuously differentiable and locally Lipschitz continuous. By Assumption 8 and the local Lipschitz continuity of φ' we have

$$
\begin{aligned}
\varphi'(U(t_m, x_i)) &= \varphi'(u_k(t_m, x_i, p_0) + O(\Delta x^\alpha)) \\
&= \varphi'(u_k(t_m, x_i, p_0)) + O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$ for all $i \in \{\mathrm{r}, \ldots, \mathrm{l}\}$. By Assumption 10 we also have

$$
\dot{U}(t_m, x_i) = \dot{u}_k(t_m, x_i, p_0) + O(\Delta x^\alpha)
$$

as $\Delta x \to 0$ for all $i \in \{\mathrm{r}, \ldots, \mathrm{l}\}$. Since both the weak solution and tangent sensitivity are bounded and φ' is Lipschitz continuous we have

$$
\begin{aligned}
\varphi'(U(t_m, x_i))\dot{U}(t_m, x_i) &= \big(\varphi'(u_k(t_m, x_i, p_0)) + O(\Delta x^\alpha)\big)\big(\dot{u}_k(t_m, x_i, p_0) + O(\Delta x^\alpha)\big) \\
&= \varphi'(u_k(t_m, x_i, p_0))\dot{u}_k(t_m, x_i, p_0) + O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$ for all $i \in \{\mathrm{r}, \ldots, \mathrm{l}\}$.

Since $\alpha < 1$ we have $\Delta x < C_s\Delta x^\alpha$ for all $\Delta x \leq \delta$ for some $\delta > 0$. We thus have

$$
\begin{aligned}
x_{\mathrm{r}} - \frac{1}{2}\Delta x &\geq \xi_{k-1}(t_m, p_0) + C_s\Delta x^\alpha - \frac{1}{2}\Delta x \\
&> \xi_{k-1}(t_m, p_0)
\end{aligned}
$$

and, analogously

$$
x_{\mathrm{l}} + \frac{1}{2}\Delta x < \xi_k(t_m, p_0)
$$

as $\Delta x \to 0$. We can, therefore, define the following error between an integral and its approximation by the mid point rule

$$
E_k \equiv \int_{x_i - \frac{1}{2}\Delta x}^{x_i + \frac{1}{2}\Delta x} \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x - \varphi'(U(t_m, x_i))\dot{U}(t_m, x_i)\Delta x
$$

for all $i \in \{\mathrm{r}, \ldots, \mathrm{l}\}$. By Assumption 6 both u_k and $\mathrm{d}_p u_k$ and hence $\dot{u}_k$ are continuously differentiable with respect to x on $[\xi_{k-1}(t_m, p_0), \xi_k(t_m, p_0)]$. By the Lipschitz continuity of φ and the necessary boundedness of factors we have

$$\begin{aligned}
&\int_{x_i - \frac{1}{2}\Delta x}^{x_i + \frac{1}{2}\Delta x} \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x \\
&= \int_{x_i - \frac{1}{2}\Delta x}^{x_i + \frac{1}{2}\Delta x} \varphi'(u_k(T, x_i, p_0) + O(|x - x_i|)(\dot{u}_k(T, x_i, p_0) + O(|x - x_i|))\mathrm{d}x \\
&= \int_{x_i - \frac{1}{2}\Delta x}^{x_i + \frac{1}{2}\Delta x} \varphi'(u_k(T, x_i, p_0))\dot{u}_k(T, x_i, p_0) + O(|x - x_i|)\mathrm{d}x \\
&= \varphi'(u_k(T, x_i, p_0))\dot{u}_k(T, x_i, p_0)\Delta x + \int_{x_i - \frac{1}{2}\Delta x}^{x_i + \frac{1}{2}\Delta x} O(\Delta x)\mathrm{d}x \\
&= \varphi'(u_k(T, x_i, p_0))\dot{u}_k(T, x_i, p_0)\Delta x + O(\Delta x^2)
\end{aligned}$$

as $\Delta x \to 0$. By the above approximation error, $T = t_m$ and since $\alpha < 1$ we have

$$\begin{aligned}
E_k &= \varphi'(u_k(T, x_i, p_0))\dot{u}_k(T, x_i, p_0)\Delta x + O(\Delta x^2) - \varphi'(U(t_m, x_i))\dot{U}(t_m, x_i)\Delta x \\
&= \varphi'(u_k(T, x_i, p_0))\dot{u}_k(T, x_i, p_0)\Delta x - \varphi'(u_k(t_m, x_i, p_0))\dot{u}_k(t_m, x_i, p_0)\Delta x + O(\Delta x^{1+\alpha}) \\
&= O(\Delta x^{1+\alpha})
\end{aligned}$$

as $\Delta x \to 0$.

We now show a bound for the error between the integral on the domain between the shock regions as they are approximated in the discrete approximation and the exact integral between the exact shocks. We later come back to use the E_k error bounds in combination with the result that follows to bound the total error of the sum of integral approximations.

We have $x_\mathrm{r} - \Delta x/2 > \xi_{k-1}$ and $x_\mathrm{l} + \Delta x/2 < \xi_k$ as $\Delta x \to 0$. By the definition of $\mathrm{l}(\cdot)$ and $\mathrm{r}(\cdot)$ we have

$$x_\mathrm{l} = x_{\mathrm{l}(\Xi_k^m - C_r \Delta x^\alpha)} \geq \Xi_k^m - C_r \Delta x^\alpha - \Delta x$$

and

$$x_\mathrm{r} = x_{\mathrm{r}(\Xi_{k-1}^m + C_r \Delta x^\alpha)} \leq \Xi_{k-1}^m + C_r \Delta x^\alpha + \Delta x\ .$$

In combination with the previous bounds we, therefore, have

$$x_\mathrm{l} - \xi_k(T, p_0) = O(\Delta x^\alpha)$$

and

$$x_\mathrm{r} - \xi_{k-1}(T, p_0) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$. Due to the boundedness of $u_k, \dot{u}_k$ and the Lipschitz continuity of φ' we have the error between integrals as

$$\begin{aligned}
&\sum_{i=1}^{\mathrm{r}} \int_{x_i - \frac{1}{2}\Delta x}^{x_i + \frac{1}{2}\Delta x} \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x - \int_{\xi_{k-1}(T, p_0)}^{\xi_k(T, p_0)} \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x \\
&= \int_{x_\mathrm{l} - \frac{1}{2}\Delta x}^{x_\mathrm{r} + \frac{1}{2}\Delta x} \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x - \int_{\xi_{k-1}(T, p_0)}^{\xi_k(T, p_0)} \varphi'(u_k(T, x, p_0))\dot{u}_k(T, x, p_0)\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= -\int_{\xi_{k-1}(T,p_0)}^{x_\mathrm{l}-\frac{1}{2}\Delta x} \varphi'(u_k(T,x,p_0))\dot{u}_k(T,x,p_0)\mathrm{d}x - \int_{x_\mathrm{r}+\frac{1}{2}\Delta x}^{\xi_k(T,p_0)} \varphi'(u_k(T,x,p_0))\dot{u}_k(T,x,p_0)\mathrm{d}x \\
&= \Big(x_\mathrm{l} - \frac{1}{2}\Delta x - \xi_{k-1}(T,p_0)\Big)O(1) + \Big(\xi_k(T,p_0) - x_\mathrm{r} + \frac{1}{2}\Delta x\Big)O(1) \\
&= O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$.

Combining the integral error approximation with that of the individual terms of the sum over i by adding zero and by the triangle inquality we have

$$
\begin{aligned}
&\Big\| \sum_{i=\mathrm{r}}^{\mathrm{l}} \varphi'(U(t_m,x_i))\dot{U}(t_m,x_i)\Delta x - \int_{\xi_{k-1}(T,p_0)}^{\xi_k(T,p_0)} \varphi'(u_k(T,x,p_0))\dot{u}_k(T,x,p_0)\mathrm{d}x \Big\| \\
&\leq \Big\| \sum_{i=\mathrm{r}}^{\mathrm{l}} \varphi'(U(t_m,x_i))\dot{U}(t_m,x_i)\Delta x - \sum_{i=\mathrm{l}}^{\mathrm{r}} \int_{x_i-\frac{1}{2}\Delta x}^{x_i+\frac{1}{2}\Delta x} \varphi'(u_k(T,x,p_0))\dot{u}_k(T,x,p_0)\mathrm{d}x \Big\| \\
&\quad + \Big\| \sum_{i=\mathrm{l}}^{\mathrm{r}} \int_{x_i-\frac{1}{2}\Delta x}^{x_i+\frac{1}{2}\Delta x} \varphi'(u_k(T,x,p_0))\dot{u}_k(T,x,p_0)\mathrm{d}x - \int_{\xi_{k-1}(T,p_0)}^{\xi_k(T,p_0)} \varphi'(u_k(T,x,p_0))\dot{u}_k(T,x,p_0)\mathrm{d}x \Big\| \\
&\leq \sum_{i=\mathrm{r}}^{\mathrm{l}} \|E_k\| + O(\Delta x^\alpha) = O(\Delta x^{-1} - \Delta x^{\alpha-1})O(\Delta x^{1+\alpha}) + O(\Delta x^\alpha) \\
&= O(\Delta x^\alpha) + O(\Delta x^{2\alpha}) + O(\Delta x^\alpha) = O(\Delta x^\alpha)
\end{aligned}
$$

as $\Delta x \to 0$.

The integral approximation error is given by a sum over $i = \mathrm{r}, \ldots, \mathrm{l}$ that has $O(\Delta x^{-1} - \Delta x^{\alpha-1})$ terms $\|E_k\| = O(\Delta x^\alpha)$. With $\xi_k - \xi_{k-1} \geq C_d$ for some constant $C_d > 0$ the number of terms in a uniform discretization with Δx grid spacing of that interval is $C_d \Delta x^{-1}$. Since we subtract a shock region of width $O(\Delta x^\alpha)$ the number of terms between r and l is instead

$$
(C_d - O(\Delta x^\alpha))\Delta x^{-1} = O(\Delta x^{-1} - \Delta x^{\alpha-1}) \,.
$$

Since C_d is finite the first sum of Equations (6.4.5)-(6.4.8) in total is $O(\Delta x^\alpha)$ as $\Delta x \to 0$. This concludes the proof. □

6.4.2 Discrete Adjoint Sensitivities

We now derive the discrete adjoint sensitivity that corresponds to the explicit shock tangent of Definition 42.

Proposition 43. Let $\mathrm{d}_p\Phi_\mathrm{shock}$ be the Jacobian implicitly defined by the explicit shock tangent sensitivity. The discrete adjoint sensitivity

$$
\bar{p}_\mathrm{shock} \equiv (\mathrm{d}_p\Phi_\mathrm{shock})^T\bar{\Phi}
$$

with respect to the parameter $p \in \mathbb{R}^d$ is

$$
\bar{p}_\mathrm{shock} = \sum_{k=1}^{K+1} \sum_{i=\mathrm{r}(\Xi_{k-1}^m + C_r\Delta x^\alpha)}^{\mathrm{l}(\Xi_k^m - C_r\Delta x^\alpha)} (\mathrm{d}_p U(t_m,x_i))^T \varphi'(U(t_m,x_i))\bar{\Phi}\Delta x
$$

$$+\sum_{k=1}^{K}(\mathrm{d}_p\Xi_k^m)^T\big(\varphi(U(t_m,\Xi_k^m-C_r\Delta x^\alpha))-\varphi(U(t_m,\Xi_k^m+C_r\Delta x^\alpha))\big)\bar{\Phi}\ .$$

with C_r, α, $\mathrm{r}(x)$, $\mathrm{l}(x)$ as in Definition 42.

Proof. By the definition of the adjoint operator we have

$$\begin{aligned}
&\langle\dot{\Phi}_{\text{shock}},\bar{\Phi}_{\text{shock}}\rangle\\
&\quad=\Big\langle\sum_{k=1}^{K+1}\sum_{i=\mathrm{r}(\Xi_{k-1}^m+C_r\Delta x^\alpha)}^{\mathrm{l}(\Xi_k^m-C_r\Delta x^\alpha)}\varphi'(U(t_m,x_i))\dot{U}(t_m,x_i)\Delta x\\
&\qquad\quad+\sum_{k=1}^{K}\dot{\Xi}_k^m\big(\varphi(U(t_m,\Xi_k^m-C_r\Delta x^\alpha))-\varphi(U(t_m,\Xi_k^m+C_r\Delta x^\alpha))\big),\bar{\Phi}_{\text{shock}}\Big\rangle\\
&\quad=\Big\langle\sum_{k=1}^{K+1}\sum_{i=\mathrm{r}(\Xi_{k-1}^m+C_r\Delta x^\alpha)}^{\mathrm{l}(\Xi_k^m-C_r\Delta x^\alpha)}\varphi'(U(t_m,x_i))\mathrm{d}_pU(t_m,x_i)\dot{p}\Delta x\\
&\qquad\quad+\sum_{k=1}^{K}\mathrm{d}_p\Xi_k^m\dot{p}\big(\varphi(U(t_m,\Xi_k^m-C_r\Delta x^\alpha))-\varphi(U(t_m,\Xi_k^m+C_r\Delta x^\alpha))\big),\bar{\Phi}_{\text{shock}}\Big\rangle\\
&\quad=\Big\langle\dot{p},\sum_{k=1}^{K+1}\sum_{i=\mathrm{r}(\Xi_{k-1}^m+C_r\Delta x^\alpha)}^{\mathrm{l}(\Xi_k^m-C_r\Delta x^\alpha)}(\mathrm{d}_pU(t_m,x_i))^T\varphi'(U(t_m,x_i))\bar{\Phi}_{\text{shock}}\Delta x\\
&\qquad\quad+\sum_{k=1}^{K}(\mathrm{d}_p\Xi_k^m)^T\big(\varphi(U(t_m,\Xi_k^m-C_r\Delta x^\alpha))-\varphi(U(t_m,\Xi_k^m+C_r\Delta x^\alpha))\big)\bar{\Phi}_{\text{shock}}\Big\rangle\\
&=\langle\dot{p},\bar{p}_{\text{shock}}\rangle\ .
\end{aligned}$$

This concludes the proof. □

The above adjoint sensitivity result requires the evaluation of the adjoint shock sensitivities

$$(\mathrm{d}_p\Xi_k^m)^T\bar{\Xi}_k^m$$

with

$$\bar{\Xi}_k^m\equiv\big(\varphi(U(t_m,\Xi_k^m-C_r\Delta x^\alpha))-\varphi(U(t_m,\Xi_k^m+C_r\Delta x^\alpha))\big)\bar{\Phi}_{\text{shock}}$$

and the adjoint weak solution sensitivities

$$(\mathrm{d}_pU(t_m,x_i))^T\bar{U}_i^m$$

with

$$\bar{U}_i^m\equiv\varphi'(U(t_m,x_i))\bar{\Phi}_{\text{shock}}\ .$$

In order to obtain those adjoint sensitivities we can use the idea of Lemma 17. Due to linearity the adjoint sensitivity of the tangent sensitivity of a function is the same as the adjoint sensitivity of that function.

The shock sensitivity algorithm defined in Definition 39 gives a recipe for computing the tangent sensitivities $\dot{\Xi}_k$. Taking the adjoint of that tangent sensitivity will provide us with the adjoint sensitivity of the shock location. The time stepping scheme is a linear function of the tangent sensitivities of both the shock location and the weak solution, i.e. it has the form

$$\dot{\Xi}_k^{j+1} := \mathrm{d}_{\Xi_k^j} \Xi_k^{j+1} \cdot \dot{\Xi}_k^j + \mathrm{d}_{U^j} \Xi_k^{j+1} \cdot \dot{U}^j \ .$$

The corresponding adjoint sensitivity time stepping schemes need to have the form

$$\bar{\Xi}_k^j := \bar{\Xi}_k^j + (\mathrm{d}_{\Xi_k^j} \Xi_k^{j+1})^T \cdot \bar{\Xi}_k^{j+1}$$

and

$$\bar{U}^j := \bar{U}^j + (\mathrm{d}_{U^j} \Xi_k^{j+1})^T \cdot \bar{\Xi}_k^{j+1}$$

where $\bar{U}^j$ is required to additionally accumulate the discrete adjoint of the numerical integrator used for the solution of the scalar conservation law.

We say more about the practical implementation of the adjoint sensitivities via AD tools in the next section. According to Lemma 17 if we have a correct implementation for the tangent $\dot{\Phi}_{\text{shock}}$ then we can get the adjoint sensitivity of that tangent sensitivity via a nested application of an AD tool without having to explicitly write the above adjoint implementation by hand.

We now show that the convergence result that holds for the discrete tangent sensitivity also holds for the discrete adjoint sensitivity. The discrete adjoint sensitivity resulting from the explicit shock tangent sensitivity converges to the analytic adjoint sensitivity.

Proposition 44.

$$\|\bar{p}_{\text{shock}} - \bar{p}\| = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ with α as in Definition 39.

Proof. The proof follows directly from Lemma 18.

We have

$$\dot{\Phi}_{\text{shock}} = (\mathrm{d}_p \Phi_{\text{shock}})\dot{p}, \qquad \dot{\Phi} = (\mathrm{d}_p \Phi)\dot{p}$$

and

$$\bar{p}_{\text{shock}} = (\mathrm{d}_p \Phi_{\text{shock}})^T \bar{\Phi}_{\text{shock}}, \bar{p} = (\mathrm{d}_p \Phi)^* \bar{\Phi}_{\text{shock}} \ .$$

By Proposition 42 we have

$$\dot{\Phi}_{\text{shock}} - \dot{\Phi} = O(\Delta x^\alpha) \ .$$

Hence, by Lemma 18 we have

$$\bar{p}_{\text{shock}} - \bar{p} = O(\Delta x^\alpha) \ .$$

This concludes the proof. □

6.5 Implementation via Algorithmic Differentiation

In this section we explain how the shock sensitivity algorithm can be generated by an AD tool without having to explicitly implement the formula for the linearization of the Rankine-Hugoniot condition in Definition 39 that requires the derivative of the flux function. We also describe the resulting discrete adjoint sensitivities and show how both discrete tangent and adjoint sensitivities can be generated automatically by an operator overloading AD tool.

6.5.1 Shock Sensitivities from Rankine-Hugoniot Condition

The shock sensitivity algorithm of Definition 39 is motivated by the tangent sensitivity of the ODE resulting from the Rankine-Hugoniot condition. We now show how we can utilize the discrete tangents of the explicit Euler discretization of the Rankine-Hugoniot condition ODE instead of explictly finding a discretization of the analytic tangent ODE. With the correct choice of shock region width the resulting discrete tangent is convergent in the same way as the shock sensitivity algorithm of Definition 39.

In the following we use the notation

$$\Xi_k^{j\,\pm} \equiv \Xi_k^j \pm C\Delta x^\alpha$$

where $C > 0$ is the constant that depends on the width of the shock region in which the shock discontinuity is smeared out due to numerical viscosity which is, e.g., determined by the ratio between the space and time step sizes in the case of the Lax-Friedrichs scheme. We also use the notation $\mathcal{U}$ for different kinds of interpolations of the weak solution and $\dot{\mathcal{U}}$ its discrete tangent sensitivity to explicitly distinguish e.g. the linear interpolation from the discrete vector that contains the solution on the grid.

The explicit Euler discretization of the Rankine-Hugoniot condition ODE given by

$$\Xi_k^{j+1} := \Xi_k^j + \Delta t \frac{f(\mathcal{U}(t_j, \Xi_k^{j\,-})) - f(\mathcal{U}(t_j, \Xi_k^{j\,+}))}{\mathcal{U}(t_j, \Xi_k^{j\,-}) - \mathcal{U}(t_j, \Xi_k^{j\,+})}$$

uses the same ideas to obtain convergent approximations of the weak solution at the left and right limits of the shock that were already used in the previous sections.

In Section 1.4.5 we argued that we do not want to use that discretization to simulate the shock location because its error cannot be bounded as tightly as that of the shock location algorithm of Definition 38. But we can use its discrete tangent and adjoint sensitivities

The discrete tangent sensitivity of the Rankine-Hugoniot discretization is

$$\begin{aligned}\dot{\Xi}_k^{j+1} := \dot{\Xi}_k^j + \Delta t \Bigg(& \frac{f'(\mathcal{U}(t_j, \Xi_k^{j\,-}))\dot{\mathcal{U}}(t_j, \Xi_k^{j\,-}) - f'(\mathcal{U}(t_j, \Xi_k^{j\,+}))\dot{\mathcal{U}}(t_j, \Xi_k^{j\,+})}{\mathcal{U}(t_j, \Xi_k^{j\,-}) - \mathcal{U}(t_j, \Xi_k^{j\,+})} \\ & - \frac{f(\mathcal{U}(t_j, \Xi_k^{j\,-})) - f(\mathcal{U}(t_j, \Xi_k^{j\,+}))}{\big(\mathcal{U}(t_j, \Xi_k^{j\,-}) - \mathcal{U}(t_j, \Xi_k^{j\,+})\big)^2} \big(\dot{\mathcal{U}}(t_j, \Xi_k^{j\,-}) - \dot{\mathcal{U}}(t_j, \Xi_k^{j\,+})\big)\Bigg)\end{aligned}$$

where the structure of the discrete tangent $\dot{\mathcal{U}}$ depends on the kind of interpolation used for the weak solution.

In the case that $\mathcal{U}$ represents a piecewise constant interpolation we have

$$\begin{aligned}\dot{\mathcal{U}}(t_j, \Xi_k^{j\,\pm}) &= \partial_U \mathcal{U}(t_j, \Xi_k^{j\,\pm}) \cdot \dot{U}^j + \underbrace{\partial_x \mathcal{U}(t_j, \Xi_k^{j\,\pm})}_{0} \cdot \dot{\Xi}_k^j \\ &= \dot{U}(t_j, \Xi_k^{j\,\pm})\end{aligned}$$

where $\dot{U}$ is the piecewise constant interpolation of the discrete tangent sensitivity. In that case the term consisting of the derivative with respect to space as a factor is not correctly approximated because a piecewise constant interpolation is not differentiable. The discrete tangent based on the piecewise constant interpolation is not generally convergent to the analytic tangent sensitivity as $\Delta x \to 0$ as a consequence of the missing term.

In the case that $\mathcal{U}$ represents a linear interpolation we have

$$\begin{aligned}\dot{\mathcal{U}}(t_j, \Xi_k^{j\,\pm}) &= \partial_U \mathcal{U}(t_j, \Xi_k^{j\,\pm}) \cdot \dot{U}^j + \partial_x \mathcal{U}(t_j, \Xi_k^{j\,\pm}) \cdot \dot{\Xi}_k^j \\ &= \dot{U}_i^j + (\Xi_k^{j\,\pm} - x_i)\frac{\dot{U}_{i+1}^j - \dot{U}_i^j}{\Delta x} + \frac{U_{i+1}^j - U_i^j}{\Delta x} \cdot \dot{\Xi}_k^j\end{aligned}$$

with i such that $x_i < \Xi_k^{j\,\pm} \leq x_i + \Delta x$ correspondingly. The space derivative is approximated by a finite difference similar to that in Definition 39. Therefore, we expect the discrete adjoint resulting from a linear interpolation to converge as $\Delta x \to 0$.

In practice with $x_i < \Xi_k^{j\,+} \leq x_i + \Delta x$ we have

$$\frac{U(t_j, x_i + \Delta x) - U(t_j, x_i)}{\Delta x} - \frac{U(t_j, \Xi_k^j + C\Delta x^\alpha + 2\Delta x) - U(t_j, \Xi_k^j + C\Delta x^\alpha)}{2\Delta x} = O(\Delta x)$$

and with $x_i < \Xi_k^{j\,-} \leq x_i + \Delta x$ we have

$$\frac{U(t_j, x_i + \Delta x) - U(t_j, x_i)}{\Delta x} - \frac{U(t_j, \Xi_k^j - C\Delta x^\alpha) - U(t_j, \Xi_k^j - C\Delta x^\alpha - 2\Delta x)}{2\Delta x} = O(\Delta x)$$

and so the convergence result of Proposition 38 holds for $\dot{\Xi}$ obtained by the above time stepping with the linear interpolation if C is large enough such that the finite difference is evaluated outside of the shock region.

For $C \geq C_r + 1$ and $\alpha < 1$ we have

$$x_i + \Delta x \geq \Xi_k^j + C\Delta x^\alpha$$

and, hence, by subtraction of Δx from both sides we have

$$\begin{aligned}x_i &\geq \Xi_k^j + C\Delta x^\alpha - \Delta x \\ &\geq \Xi_k^j + C_r\Delta x^\alpha + \Delta x^\alpha - \Delta x \\ &\geq \Xi_k^j + C_r\Delta x^\alpha\ .\end{aligned}$$

The left side of the linear interpolation finite difference to the right of the shock is outside of the shock region according to Assumption 8 and Assumption 9. Similary, we have

$$x_i < \Xi_k^j - C\Delta x^\alpha$$

and, hence, by addition of Δx to both sides we have

$$\begin{aligned}x_{i+1} &< \Xi_k^j - C\Delta x^\alpha + \Delta x \\ &\leq \Xi_k^j - C_r\Delta x^\alpha - \Delta x^\alpha + \Delta x \\ &\leq \Xi_k^j - C_r\Delta x^\alpha\ .\end{aligned}$$

The right side of the linear interpolation finite difference to the left of the shock is also outside of the shock region. For those reasons we observe the same type of convergence behavior for the discrete tangents obtained by the above method as the one we proved for the shock sensitivity algorithm of Definition 39.

The discrete adjoint sensitivities of the explicit Euler Rankine-Hugoniot discretization are computed by the following adjoint time stepping schemes

$$\bar{\Xi}_k^j := \bar{\Xi}_k^j + (\mathrm{d}_{\Xi_k^j}\Xi_k^{j+1})^T \cdot \bar{\Xi}_k^{j+1}$$

and

$$\bar{U^j} := \bar{U^j} + (\mathrm{d}_{U^j}\Xi_k^{j+1})^T \cdot \bar{\Xi}_k^{j+1}$$

where $\bar{U^j}$ also accumulates the discrete adjoint time step of the numerical integrator used for the weak solution of the scalar conservation law.

The first critical factor is the sensitivity of the time step with respect to the shock location given by

$$\begin{aligned}(\mathrm{d}_{\Xi_k^j}\Xi_k^{j+1})^T = 1 + \Delta t\Bigg(&\frac{f'(\mathcal{U}(t_j,\Xi_k^{j^-}))\mathrm{d}_{\Xi_k^j}\mathcal{U}(t_j,\Xi_k^{j^-}) - f'(\mathcal{U}(t_j,\Xi_k^{j^+}))\mathrm{d}_{\Xi_k^j}\mathcal{U}(t_j,\Xi_k^{j^+})}{\mathcal{U}(t_j,\Xi_k^{j^-}) - \mathcal{U}(t_j,\Xi_k^{j^+})} \\ &- \frac{f(\mathcal{U}(t_j,\Xi_k^{j^-})) - f(\mathcal{U}(t_j,\Xi_k^{j^+}))}{\big(\mathcal{U}(t_j,\Xi_k^{j^-}) - \mathcal{U}(t_j,\Xi_k^{j^+})\big)^2}\big(\mathrm{d}_{\Xi_k^j}\mathcal{U}(t_j,\Xi_k^{j^-}) - \mathrm{d}_{\Xi_k^j}\mathcal{U}(t_j,\Xi_k^{j^+})\big)\Bigg)\end{aligned}$$

where in the case that $\mathcal{U}$ is a linear interpolation we have

$$\mathrm{d}_{\Xi_k^j}\mathcal{U}(t_j,\Xi_k^{j^\pm}) = \frac{U(t_j,x_i+\Delta x) - U(t_j,x_i)}{\Delta x}$$

with i such that $x_i < \Xi_k^{j^\pm} \leq x_i + \Delta x$.

The second critical factor is the sensitivity of the time step with respect to the weak solution given by

$$\begin{aligned}(\mathrm{d}_{U^j}\Xi_k^{j+1})^T = \Delta t\Bigg(&\frac{f'(\mathcal{U}(t_j,\Xi_k^{j^-}))\mathrm{d}_{U^j}\mathcal{U}(t_j,\Xi_k^{j^-}) - f'(\mathcal{U}(t_j,\Xi_k^{j^+}))\mathrm{d}_{U^j}\mathcal{U}(t_j,\Xi_k^{j^+})}{\mathcal{U}(t_j,\Xi_k^{j^-}) - \mathcal{U}(t_j,\Xi_k^{j^+})} \\ &- \frac{f(\mathcal{U}(t_j,\Xi_k^{j^-})) - f(\mathcal{U}(t_j,\Xi_k^{j^+}))}{\big(\mathcal{U}(t_j,\Xi_k^{j^-}) - \mathcal{U}(t_j,\Xi_k^{j^+})\big)^2}\big(\mathrm{d}_{U^j}\mathcal{U}(t_j,\Xi_k^{j^-}) - \mathrm{d}_{U^j}\mathcal{U}(t_j,\Xi_k^{j^+})\big)\Bigg).\end{aligned}$$

In the case that $\mathcal{U}$ is a linear interpolation we have

$$\mathcal{U}(t_j,\Xi_k^{j^\pm}) = U_i^j + \frac{\Xi_k^{j^\pm} - x_i}{\Delta x}(U_{i+1}^j - U_i^j)$$

and, hence,

$$\mathrm{d}_{U_i^j}\mathcal{U}(t_j,\Xi_k^{j^\pm}) = 1 + \frac{\Xi_k^{j^\pm} - x_i}{\Delta x}, \qquad \mathrm{d}_{U_{i+1}^j}\mathcal{U}(t_j,\Xi_k^{j^\pm}) = -\frac{\Xi_k^{j^\pm} - x_i}{\Delta x}$$

with i such that $x_i < \Xi_k^{j^\pm} \leq x_i + \Delta x$.

The discrete tangent is convergent with $C \geq C_r + 1$ given Assumption 8, Assumption 10 and the slight modification of Assumption 9 for the one step finite difference hold. According to Lemma 18 the corresponding discrete adjoint that we just described is also convergent if the discrete tangent is convergent.

6.5.2 Implementation via Operator Overloading AD

We now present an algorithm that for both the forward mode and the reverse mode of AD computes the correct sensitivities for the shock locations. To simplify the notation we consider the case of only one shock location. The numerical integrator used for the scalar conservation law is the Lax-Friedrichs scheme for which we empirically showed that Assumption 8, Assumption 9 and Assumption 10 are satisfied for one of our examples in Section 6.1.

In this section we use a similar but shorter notation to that of Section 4.3 where for a program variable x the value component is denoted as x.v, instead of x.value, and the derivative component is denoted as x.d, instead of x.derivative. The following algorithm only contains the computation performed in the context of the simulation time stepping but does not explicitly list initializations and assignments of constants.

Algorithm 11: Lax-Frierichs Simulation with Differentiable Shock Location

```
Result: $U$.v, $U$.d, $\Xi$.v, $\Xi$.d
for $j = 0, \dots, m-1$ do
    // Lax-Friedrichs Scheme (copy boundary from interior)
    $U_1^{j+1} := U_2^j$ ;
    $U_n^{j+1} := U_{n-1}^j$ ;
    for $i = 2, \dots, n-1$ do
        $U_i^{j+1} := \frac{1}{2}(U_{i-1}^j + U_{i+1}^j) - \frac{\Delta t}{2\Delta x}\big(f(U_{i+1}^j) - f(U_{i-1}^j)\big)$ ;
    end
    // Differentiable Shock Location
    $\Xi^j.\mathrm{v} := \Lambda^j.\mathrm{v}$ ;
    $\Lambda^{j+1} := \Lambda^j + \Delta t f'(\mathcal{U}^j(\Lambda^j))$ ;
    $\Xi^{j+1} := \Xi^j + \Delta t \frac{f(\mathcal{U}^j(\Xi^j - C\Delta x^\alpha)) - f(\mathcal{U}^j(\Xi^j + C\Delta x^\alpha))}{\mathcal{U}^j(\Xi^j - C\Delta x^\alpha) - \mathcal{U}^j(\Xi^j + C\Delta x^\alpha)}$ ;
end
$\Xi^m.\mathrm{v} := \Lambda^m.\mathrm{v}$ ;
```

We first discuss the Lax-Friedrichs scheme section of the algorithm.

The first two lines represent a way of dealing with the boundary cells of the grid without having defined a boundary condition. By copying from the interior of the solution of the previous time step the error introduced at the boundary is $O(\Delta x)$ as $\Delta x \to 0$. The result of assigning the solution values of the boundary cells in such a way was discussed in Example 35.

The loop going over the nonboundary cells is the standard Lax-Friedrichs implementation for which both the forward mode and the reverse mode of an AD tool will correctly compute the corresponding tangent and adjoint sensitivities. The time stepping for the numerical solution U can be replaced by e.g. the Upwind scheme and other shock capturing integrators that satisfy the necessary assumptions about convergence outside of the shock region.

We now discuss the differentiable shock location section of the algorithm and show that both forward mode and reverse mode AD applied to Algorithm 11 compute the correct sensitivities as described in the previous sections.

The value component $\Lambda^j.\mathrm{v}$ of Λ^j represents the correct computation of the shock location

$$\Lambda^{j+1}.\mathrm{v} := \Lambda^j.\mathrm{v} + \Delta t f'(\mathcal{U}^j(\Lambda^j.\mathrm{v}))$$

according to the explicit Euler discretization of the characteristic speed (and not based on the Rankine-Hugoniot condition). The derivative component $\Lambda^j.\mathrm{d}$ is irrelevant because it is not used.

The shock location discretization here uses a linear interpolation of the numerical solution instead of the piecewise constant interpolation used in Definition 38. Just like we showed for the shock sensitivities in the previous sections using a linear interpolation for the shock location does not fundamentally change any of the steps in the convergence proof, i.e. a comparable convergence result to that of Proposition 36 holds for $\Lambda^j.\mathrm{v}$ as well.

The value component $\Xi^j.\mathrm{v}$ of Ξ^j is explicitly overridden at the beginning of every time step according to the line

$$\Xi^j.\mathrm{v} := \Lambda^j.\mathrm{v}\ .$$

The value component time step is, hence, computed as

$$\Xi^{j+1}.\mathrm{v} := \Lambda^j.\mathrm{v} + \Delta t \frac{f(\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha)) - f(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha))}{\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)}$$

according to the assignment to Ξ^{j+1}. But the value computed due to that time step is lost in the next time step when $\Xi^{j+1}.\mathrm{v}$ is overriden by $\Lambda^{j+1}.\mathrm{v}$ and in the end when $\Xi^m.\mathrm{v}$ is overriden by $\Lambda^m.\mathrm{v}$.

The crucial part is the derivative component $\Xi^j.\mathrm{d}$ of Ξ^j.

We first consider what happens in the forward mode computation. Due to the overriding of $\Xi^m.\mathrm{v}$ by $\Lambda^m.\mathrm{v}$ the forward mode tangent derivative component $\Xi^j.\mathrm{d}$ is computed as

$$\begin{aligned}
\Xi^{j+1}.\mathrm{d} := \Xi^j.\mathrm{d} + \Delta t \Bigg(& \frac{f'(\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha))}{\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)} \\
& \cdot \big(\partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) \cdot U^j.\mathrm{d} + \partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) \cdot \Xi^j.\mathrm{d}\big) \\
& - \frac{f'(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha))}{\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)} \\
& \cdot \big(\partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) \cdot U^j.\mathrm{d} + \partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) \cdot \Xi^j.\mathrm{d}\big) \\
& - \frac{f(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)) - f(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha))}{\big(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)\big)^2} \\
& \cdot \big(\partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) \cdot U^j.\mathrm{d} + \partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) \cdot \Xi^j.\mathrm{d} \\
& \quad - \partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) \cdot U^j.\mathrm{d} - \partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) \cdot \Xi^j.\mathrm{d}\big) \Bigg)\ .
\end{aligned}$$

Since $\Lambda^j.\mathrm{v}$ contains the correct approximation of the shock location and $U.\mathrm{d}$ contains the discrete tangent sensitivity of the numerical solution the tangent derivative component time step gives the computation for the discrete tangent sensitivity that we described in the previous sections.

We now consider what happens in the reverse mode computation. The reverse mode assigns

$$\Xi^j.\mathrm{d} := \Xi^j.\mathrm{d} + \mathrm{d}_{\Xi^j}\Xi^{j+1} \cdot \Xi^{j+1}.\mathrm{d}$$

and

$$U^j.\mathrm{d} := U^j.\mathrm{d} + \mathrm{d}_{U^j}\Xi^{j+1} \cdot \Xi^{j+1}.\mathrm{d}$$

where the local derivatives $\mathrm{d}_{\Xi^j}\Xi^{j+1}$ and $\mathrm{d}_{U^j}\Xi^{j+1}$ are evaluated using $\Xi^j.\mathrm{v} = \Lambda^j.\mathrm{v}$ due to the overriding, i.e. with

$$\begin{aligned}
\mathrm{d}_{\Xi^j}\Xi^{j+1} = 1 + \Delta t \Bigg(& \frac{f'(\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha))}{\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)} \cdot \partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) \\
& - \frac{f'(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha))}{\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)} \cdot \partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) \\
& - \frac{f(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)) - f(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)}{\big(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)\big)^2} \\
& \cdot \big(\partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \partial_x \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)\big) \Bigg)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{d}_{U^j}\Xi^{j+1} = \Delta t \Bigg(& \frac{f'(\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha))}{\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)} \cdot \partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) \\
& - \frac{f'(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha))}{\mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)} \cdot \partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) \\
& - \frac{f(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)) - f(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha))}{\big(\mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha) - \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)\big)^2} \\
& \cdot \big(\partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} - C\Delta x^\alpha) - \partial_U \mathcal{U}^j(\Lambda^j.\mathrm{v} + C\Delta x^\alpha)\big) \Bigg) .
\end{aligned}$$

Since $\Lambda^j.\mathrm{v}$ contains the correct approximation of the shock location and in the reverse mode computation $U.\mathrm{d}$ contains the discrete adjoint sensitivity of the numerical solution the adjoint derivative component time step gives the computation for the discrete adjoint sensitivity that we described in the previous sections.

6.6 Computational Results

We now validate the presented numerical methods via computational experiments. We first demonstrate that the basic assumptions made in Section 6.1 are also satisfied for the Upwind scheme and the Lax-Friedrichs scheme with a time step determined by the CFL condition with the CFL number set to one, i.e. the scheme with minimal numerical viscosity still guaranteed to remain stable. We then analyze the empirical convergence of the numerical approximation error of the weak solution, the shock location, the broad tangent sensitivity of the weak solution, the shock location sensitivity, the generalized tangent and the discrete adjoint sensitivity of an integral objective for an example with the convex Burgers flux function. Finally, we also validate the numerical approximation of the generalized tangent for a nonconvex flux function.

6.6.1 Basic Assumptions

We analyze whether Assumption 8, Assumption 9 and Assumption 10 also apply more generally for different settings of numerical integrators. We consider the numerical approximations of the weak solution, derivative with respect to space and tangent sensitivities for the ramp example Burgers equation initial value problem just like we did in Example 35, Example 36 and Example 37.

We analyze the result of using the minimum amount of numerical viscosity with the Lax-Friedrichs scheme by using an adaptive time step based on the CFL number set to one. We also analyze the result of using an Upwind scheme for numerical integration.

CFL Time Step

We now consider the Lax-Friedrichs discretization of the ramp initial value problem from Example 7 that uses the CFL time step with

$$\Delta t_j := \frac{1}{\max_i |f'(U_i^j)|} \Delta x$$

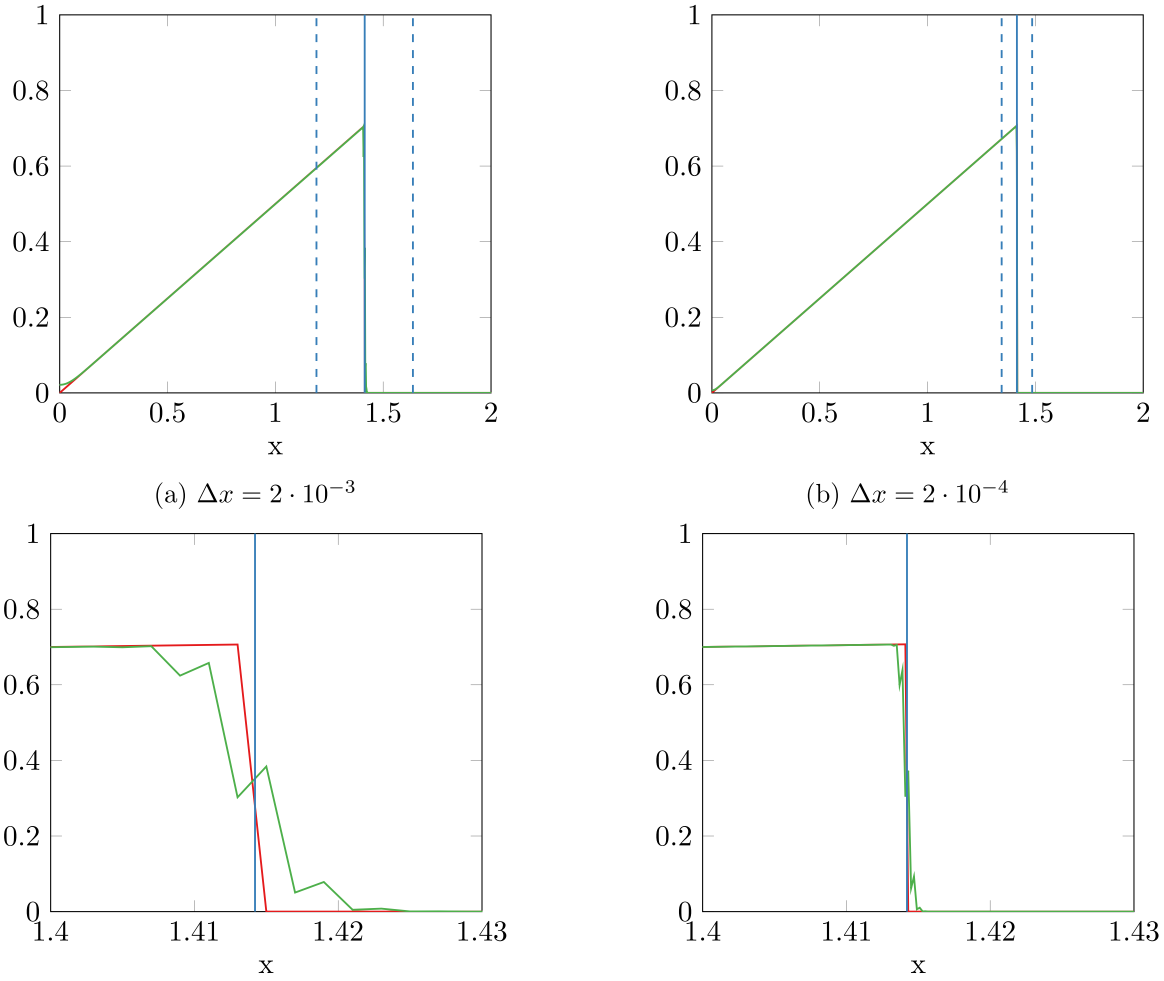
for $j \in \{0, \ldots, m\}$. According to Section 3.2.4 using the CFL time step guarantees numerical stability of the method for a CFL number less than or equal to one. Here, we use the CFL number set to one so the resulting method is at the edge of where it would become unstable. As we see in the following figures due to oscillations the resulting numerical solution is not necessarily stable to evaluate in the shock regions around shock discontinuities.

Figure 6.11 shows the analytic solution at $t = 1$ and with $p = 0$ in red and the Lax-Friedrichs discretization in green. The figure is similar to Figure 6.1 and Figure 6.2 of Example 35. The two figures on the left show the numerical solution with $\Delta x = 2 \cdot 10^{-3}$ and the two figures on the right show the numerical solution with $\Delta x = 2 \cdot 10^{-4}$. The two figures on the top show the numerical solution on the full domain and the two figures on the bottom show the numerical solution zoomed in to the shock region.

In the zoomed in figures we observe that the numerical solution in the shock region actually oscillates, i.e. it is nonsmooth as a function of x and in that sense it is not stable to evaluate within the shock region. This is in contrast to Example 35 where we used a time step Δt much smaller relative to Δx, i.e. with higher numerical viscosity. The higher numerical viscosity does not result in oscillation but instead in a wider and more smoothed out shock region. Nonetheless, the width of the shock region is linear in the discretization grid distance Δx as $\Delta x \to 0$ for the CFL time step as well.

In the full domain figures we can see that the graph of the numerical solution has a higher approximation error or difference to the analytic solution near the boundary because it uses the same method for dealing with the boundary conditions that we discussed in the previous section and in Example 35. We also observe that the shock region boundaries at $\xi_1 \pm C\Delta x^\alpha$ with $\alpha = 1/2$ and $C = 5$ marked by the dashed blue lines are outside of the area where the oscillations occur.

We observe that for the Lax-Friedrichs scheme with a CFL time step evaluating the weak solution outside of a $\xi_1 \pm C\Delta x^\alpha$ shock region will result in a pointwise convergent approximation error satisfying Assumption 8 for $\alpha = 1/2$ and $C = 5$ but also for other combinations just like we saw for the higher numerical viscosity earlier.



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

(c) $\Delta x = 2 \cdot 10^{-3}$

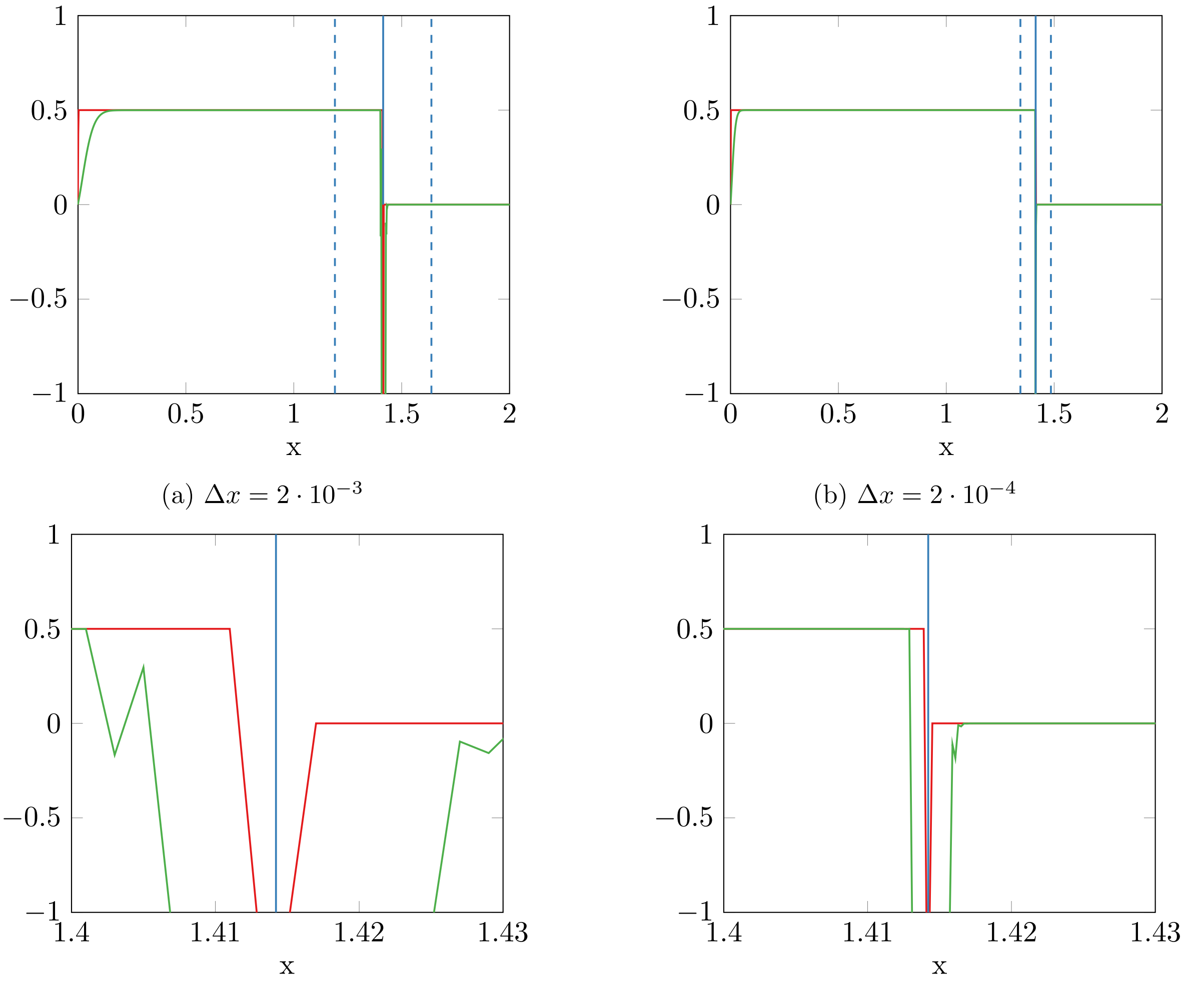
(d) $\Delta x = 2 \cdot 10^{-4}$

Figure 6.11: CFL time step, $C = 5$, $\alpha = 1/2$, analytic (red), Lax-Friedrichs U (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed)

Analogous to Figure 6.11 for the weak solution Figure 6.12 shows the derivative with respect to space of the analytic solution at $t = 1$ and with $p = 0$ in red and the Lax-Friedrichs finite difference in green. The figure is similar to Figure 6.4 and Figure 6.5 of Example 36. The only notable difference between the CFL time step Lax-Friedrichs discretization and the discretization that uses a smaller time step is that the oscillation of the weak solution in the shock region naturally also leads to oscillations in the derivative with respect to space.

We observe that for the Lax-Friedrichs discretization with a CFL time step evaluating the weak solution space derivative outside of a $\xi_1 \pm C\Delta x^\alpha$ shock region will result in a pointwise convergent approximation error satisfying Assumption 9 for $\alpha = 1/2$ and $C = 5$ but also for other combinations just like we saw for the higher numerical viscosity earlier.

Also analogous to Figure 6.11 for the weak solution Figure 6.13 shows the broad tangent sensitivity of the analytic solution at $t = 1$ and with $p = 0, \dot{p} = 1$ in red and the Lax-Friedrichs discrete tangent in green. The figure is similar to Figure 6.7 and Figure 6.8 of Example 37.



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

(c) $\Delta x = 2 \cdot 10^{-3}$

(d) $\Delta x = 2 \cdot 10^{-4}$

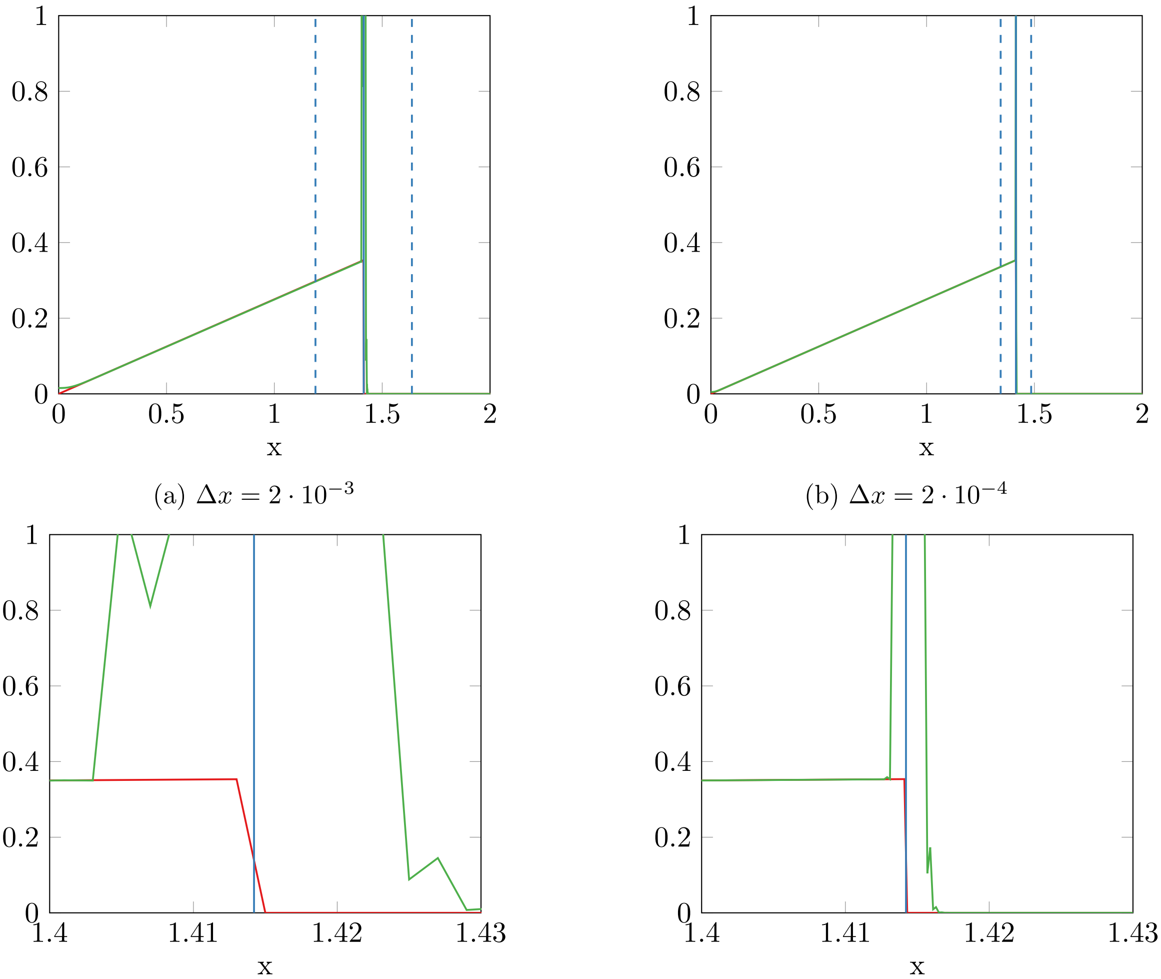
Figure 6.12: CFL time step, $C = 5$, $\alpha = 1/2$, analytic (red), Lax-Friedrichs $(U(x+\Delta x) - U(x-\Delta x))/(2\Delta x)$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed)

Again, the only notable difference between the CFL time step Lax-Friedrichs discretization and the discretization that uses a smaller time step is that the oscillation of the weak solution in the shock region naturally also leads to oscillations in the discrete tangent.

We observe that for the Lax-Friedrichs discretization with a CFL time step evaluating the weak solution tangent sensitivity outside of a $\xi_1 \pm C\Delta x^\alpha$ shock region will result in a pointwise convergent approximation error satisfying Assumption 10 for $\alpha = 1/2$ and $C = 5$ but also for other combinations just like we saw for the higher numerical viscosity earlier.

We conclude that the choice of time step does not change the empirical convergence observation we made for the Lax-Friedrichs scheme in Section 6.1.

The figures for the CFL time step illustrate why the approximation of the Rankine-Hugoniot condition tangent is not numerically stable inside of the shock region. The oscillations of the numerical solution, finite differences and discrete tangent lead an evaluation of the Rankine-Hugoniot condition inside of the shock region to be unstable in the sense that small numerical errors in the shock location approximation can result in completely different tangent approxi-



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

(c) $\Delta x = 2 \cdot 10^{-3}$

(d) $\Delta x = 2 \cdot 10^{-4}$

Figure 6.13: CFL time step, $C = 5$, $\alpha = 1/2$, analytic (red), Lax-Friedrichs $\dot{U}$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^{\alpha}$ (blue, dashed)

mations.

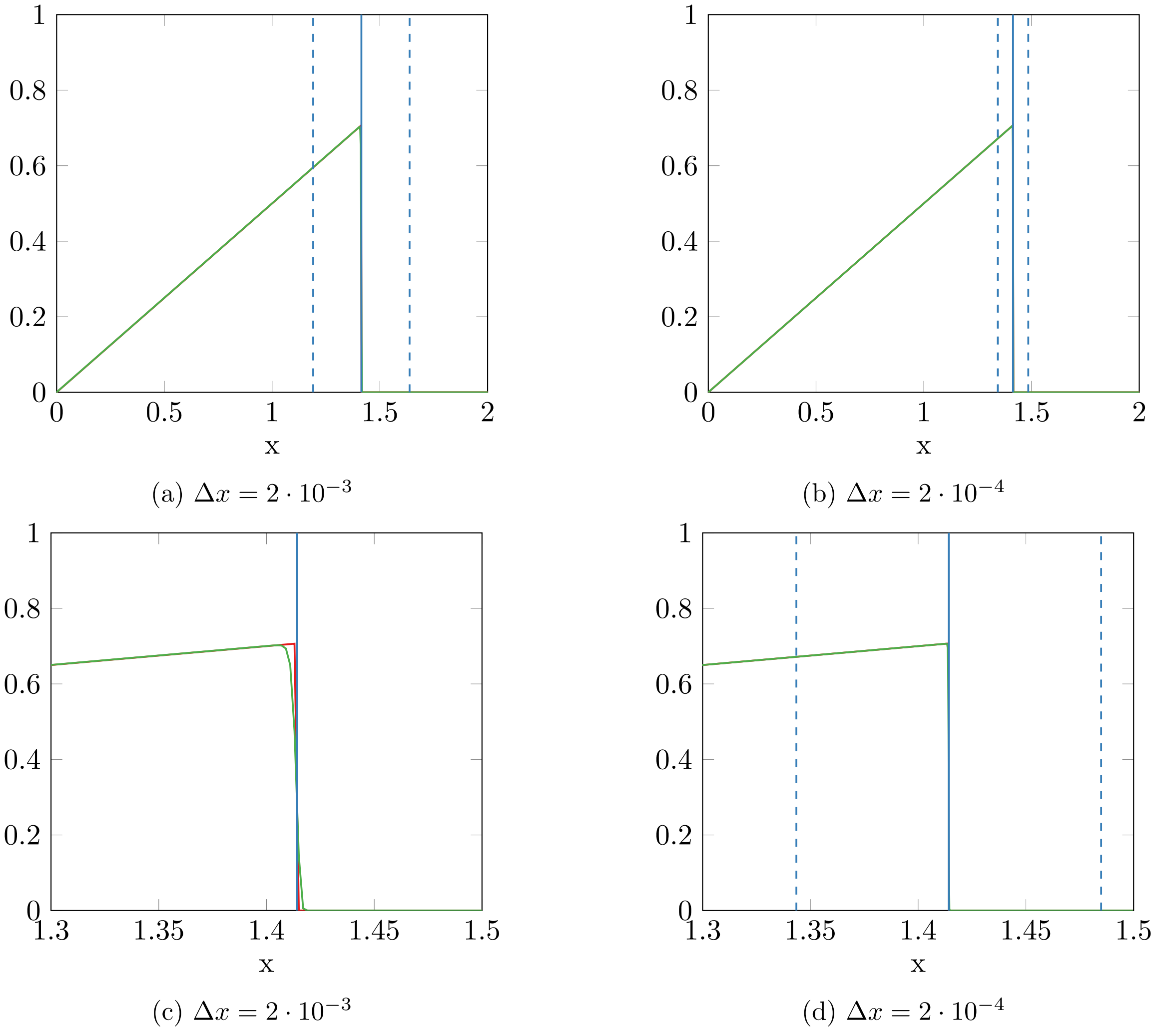
Upwind Scheme

We now consider the Upwind discretization of the ramp initial value problem from Example 7 that uses the Upwind scheme as defined in Section 3.2.6 with the direction of the finite difference discretization depending on the direction of characteristic lines. As we see in the following figures the resulting numerical solution has a shock resolution that is sharper than that of the Lax-Friedrichs scheme and there are no oscillations in the shock region.

Figure 6.14 shows the analytic solution at $t = 1$ and with $p = 0$ in red and the Upwind discretization in green. The figure is analogous to Figure 6.11 in that the two figures on the left show the numerical solution with $\Delta x = 2 \cdot 10^{-3}$ and the two figures on the right show the numerical solution with $\Delta x = 2 \cdot 10^{-4}$. The two figures on the top show the numerical solution on the full domain and the two figures on the bottom show the numerical solution zoomed in to the shock region.

In all figures we observe that the shock resolution is better than with the Lax-Friedrichs scheme both for the small time steps and the time steps with CFL number set to one. In the zoomed in figures we see that the numerical approximation around the shock does not contain any oscillation but still spreads over multiple cells. The shock location in the numerical solution accurately approximates the exact shock location and seems to converge as $\Delta x \to 0$.

We observe that for the Upwind scheme evaluating the weak solution outside of a $\xi_1 \pm C\Delta x^\alpha$ shock region will result in a pointwise convergent approximation error satisfying Assumption 8 for $\alpha = 1/2$ and $C = 5$ but also for other higher values of $\alpha < 1$ and smaller values of $C > 0$.



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

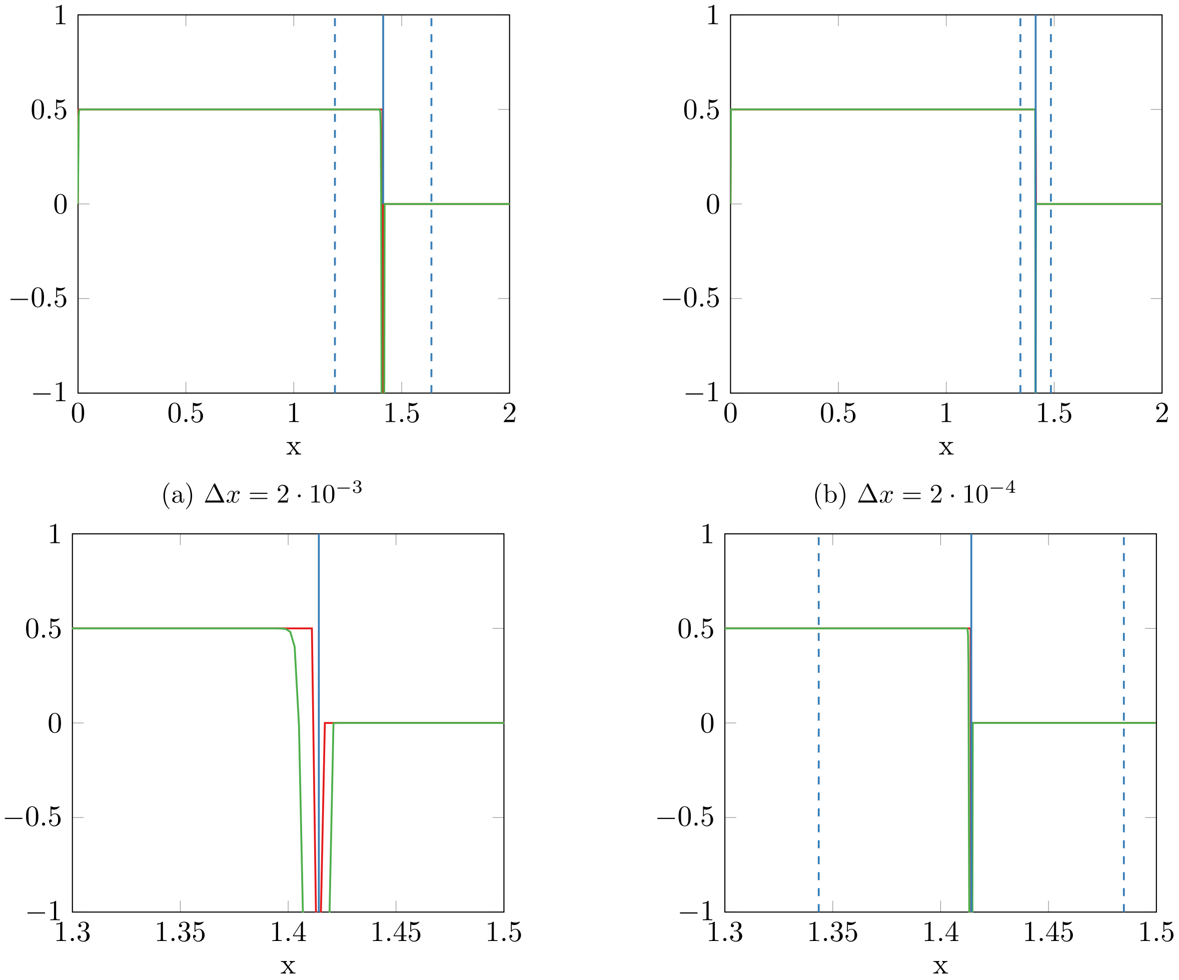
(c) $\Delta x = 2 \cdot 10^{-3}$

(d) $\Delta x = 2 \cdot 10^{-4}$

Figure 6.14: Upwind scheme, $C = 5$, $\alpha = 1/2$, analytic (red), Upwind U (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed)

Figure 6.15 shows the derivative with respect to space of the analytic solution at $t = 1$ and with $p = 0$ in red and the Upwind finite difference in green. The graph of the analytic solution shows a negative signed singularity due to the plotting discretization and the fact that the derivative with respect to space of the analytic weak solution is evaluated via a finite difference, that is still exact everywhere except at the singularity because the analytic solution is piecewise linear. In the shock region the finite difference of the numerical solution diverges because it approximates the derivative with respect to space that is singular at the shock location.

We observe that for the Upwind scheme evaluating the weak solution space derivative outside of a $\xi_1 \pm C\Delta x^\alpha$ shock region will result in a pointwise convergent approximation error satisfying to Assumption 9 for $\alpha = 1/2$ and $C = 5$ but also for other higher values of $\alpha < 1$ and smaller values of $C > 0$.



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

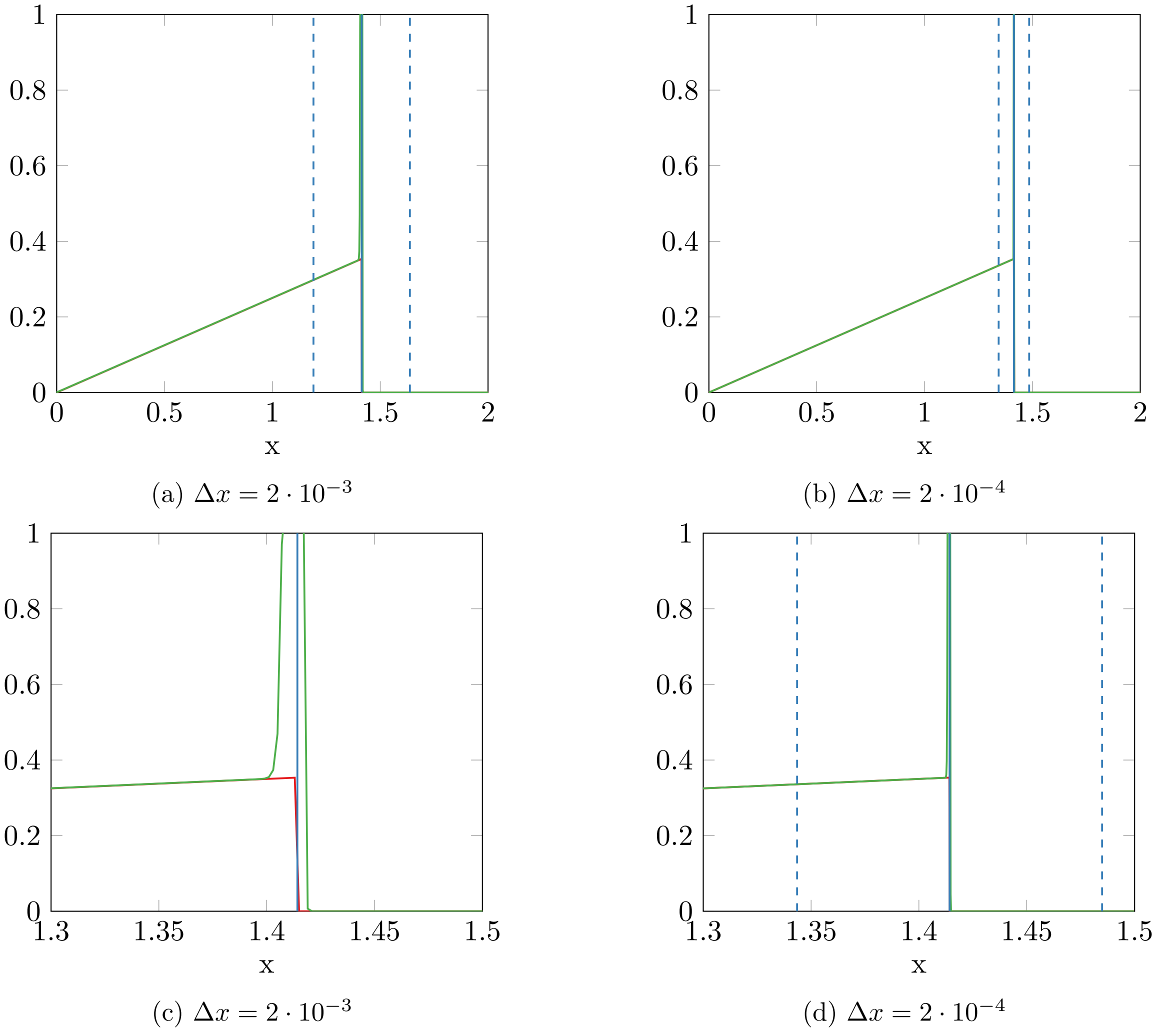
(c) $\Delta x = 2 \cdot 10^{-3}$

(d) $\Delta x = 2 \cdot 10^{-4}$

Figure 6.15: Upwind scheme, $C = 5$, $\alpha = 1/2$, analytic (red), Lax-Friedrichs $(U(x + \Delta x) - U(x - \Delta x))/(2\Delta x)$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^\alpha$ (blue, dashed)

Figure 6.15 shows the broad tangent sensitivity of the analytic solution at $t = 1$ and with $p = 0, \dot{p} = 1$ in red and the Upwind discrete tangent in green. In the shock region the discrete tangent of the numerical solution diverges because it approximates the tangent sensitivity that is singular at the shock location. The region around the shock in which the discrete tangent sensitivity is not pointwise convergent to the broad tangent is wider than the area where we observe a discretization inaccuracy in the weak solution. That is consistent with what we saw for the Lax-Friedrichs scheme both with CFL time step and higher numerical viscosity.

We observe that for the Upwind scheme evaluating the discrete tangent sensitivity outside of a $\xi_1 \pm C\Delta x^\alpha$ shock region will result in a pointwise convergent approximation error satisfying to Assumption 10 for $\alpha = 1/2$ and $C = 5$ but also for other higher values of $\alpha < 1$ and smaller values of $C > 0$.



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

(c) $\Delta x = 2 \cdot 10^{-3}$

(d) $\Delta x = 2 \cdot 10^{-4}$

Figure 6.16: CFL time step, $C = 5$, $\alpha = 1/2$, analytic (red), Lax-Friedrichs $\dot{U}$ (green), ξ_1 (blue, solid), $\xi_1 \pm C\Delta x^{\alpha}$ (blue, dashed)

We conclude that a numerical integrator that has a sharper shock resolution with less numerical viscosity and oscillations allows us to use larger values of $\alpha < 1$ with smaller values of $C > 0$ to bound the shock region. In principle, the empirical convergence observations we make for the Lax-Friedrichs scheme in Section 6.1 apply also to the Upwind scheme and the shock region can be narrower for the Upwind scheme.

6.6.2 Empirical Convergence

We perform computational experiments in order to numerically validate the formal convergence results proved earlier in this chapter by plotting the empirical convergence behavior of different numerical approximation errors. We consider the numerical approximations of the weak solution, shock location, discrete tangent sensitivity, shock sensitivity and generalized discrete tangent sensitivity for the ramp example Burgers equation initial value problem and the adjoint sensitivities of the least-squares integral objective function for the Riemann example initial value problem. Throughout this section the time step Δt is a constant factor of Δx as we assume in the convergence proofs, i.e. the numerical approximation is comparable to the experiments in Section 6.1.

Weak Solution

We now consider the approximation error of the numerical solution U. Figure 6.17 shows the approximation error of the weak solution obtained using a Lax-Friedrichs numerical integrator for geometrically decaying values of Δx.

The approximation error is evaluated as an approximation of the L_1 error. For a piecewise Lipschitz continuous function f with finitely many discontinuities we have

$$\sum_{i=1}^{n} |f(x_i)|\Delta x - \|f(\cdot)\|_{L_1(\Omega)} = O(\Delta x)$$

as $\Delta x \to 0$. If the function f is discontinuous then the L_1 error approximation is not necessarily stable due to grid points x_i potentially moving "across" the discontinuity as $\Delta x \to 0$.

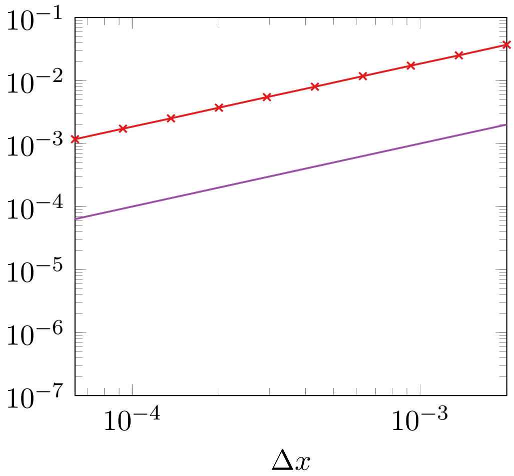


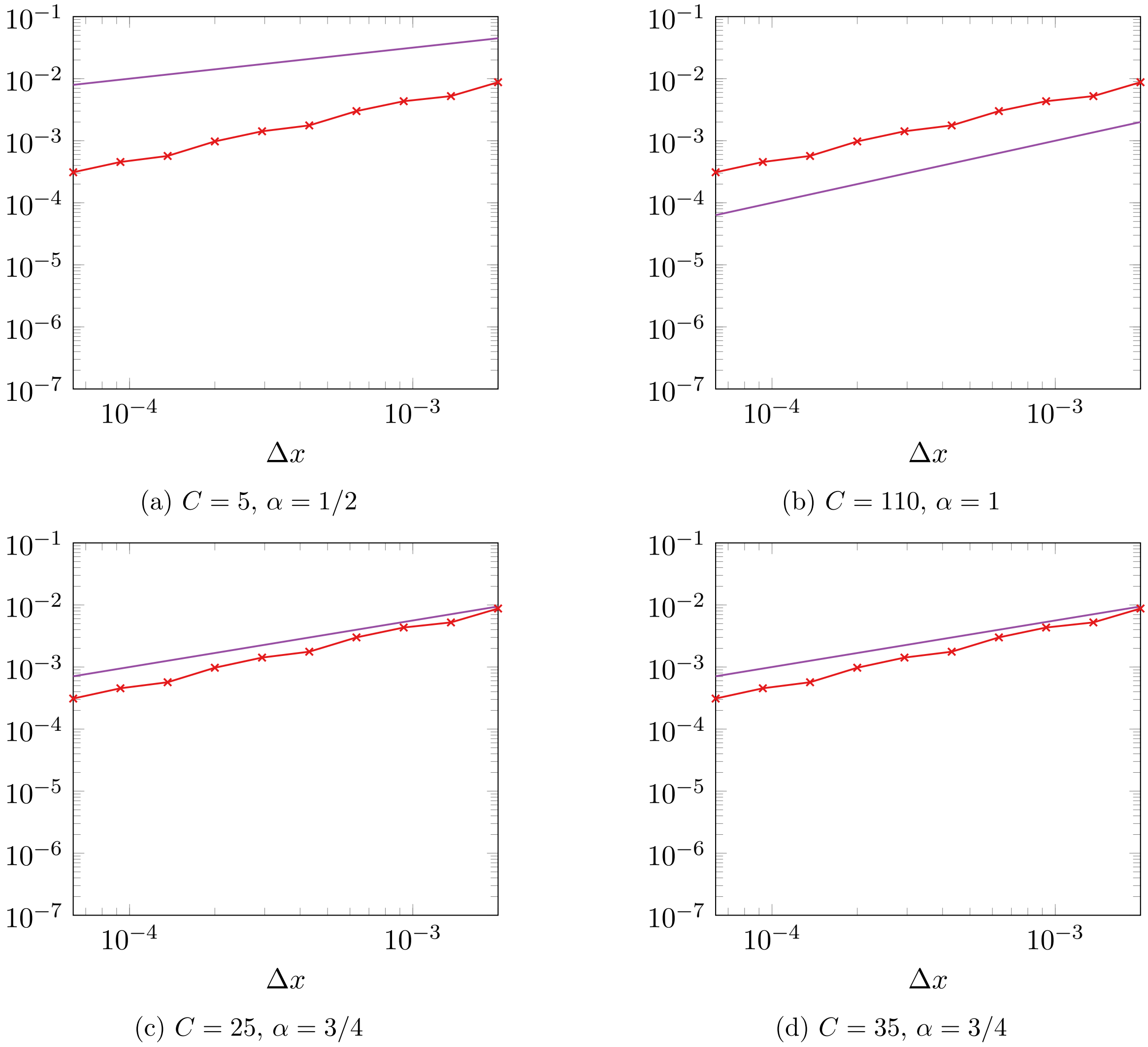
Figure 6.17: $\sum_i |U_i - u(x_i)|\Delta x$ (red), Δx (purple)

We observe a linear convergence behavior for the approximation error as predicted by Teng and Zhang [TZ97] for solutions with shock discontinuities. We earlier showed that the Lax-Friedrichs scheme satisfies Assumption 8. The convergence behavior Figure 6.17 also empirically validates Proposition 32.

Shock Location

We now consider the approximation error of the numerical shock location Ξ. Figure 6.18 shows the approximation error of the shock location obtained using the shock location algorithm of Definition 38 in combination with a Lax-Friedrichs numerical integrator.

The approximation error of the shock location is actually less than the L_1 approximation error of the weak solution. We observe linear convergence behavior for the approximation error,



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

(d) $C = 35$, $\alpha = 3/4$

Figure 6.18: $|\Xi_1 - \xi_1|$ (red), Δx (purple)

i.e. Proposition 36 is actually empirically validated with $\alpha = 1$. That means in the practical application the convergence is actually better than predicted by the theory if Assumption 8 is not made with $\alpha = 1$.

Intuitively, the reason for the practical linear convergence of the shock location can be explained as follows. The Burgers equation has a convex flux function that has a monotone first derivative such that we have $f'(u_L) > f'(u) > f'(u_R)$ for all $u_L > u > u_R$. Due to the entropy condition the charactistics, i.e. the dynamics of the shock location, always point into the direction of the shock location. This leads to the simulated shock location actually always being within a $O(\Delta x)$ distance to where the shock is approximated for the discrete approximation of the weak solution. Where the shock is approximated, i.e. how far the shock region is from the exact shock location, is implicitly determined to be at most $O(\Delta x)$ distant from the exact shock location by the fact that the L_1 norm of the weak solution approximation is also $O(\Delta x)$ as $\Delta x \to 0$. In Assumption 8 we only assume boundedness of the solution within the shock region. But if we consider e.g. the graphs in Figure 6.11 we can see that even with oscillations in the shock region the weak solution actually has values that lie between the left and right limits to the shock region in practice. That is why the simulation of the generalized characteristic is relatively well behaved and does not jump too far out of the shock region. In the convergence proof we did not assume this to be necessarily true so for this example the theoretical bound is not as tight as it could be.

Broad Tangent Sensitivity

We now consider the approximation error of the disrete tangent sensitivity $\dot{U}$ and the numerical broad tangent $\dot{U}_{\text{broad}}$. Figure 6.19 shows the approximation error of the discrete tangent sensitivity of the Lax-Friedrichs numerical integrator.

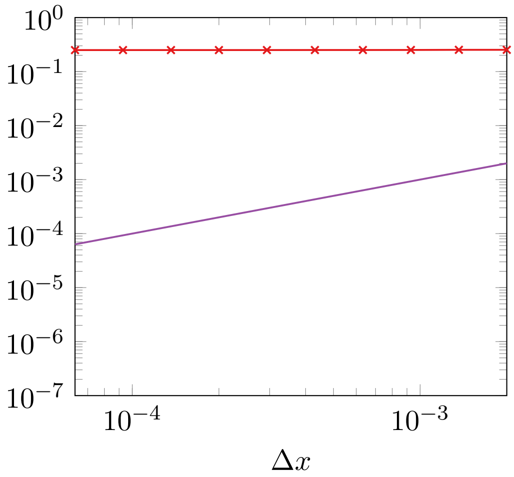
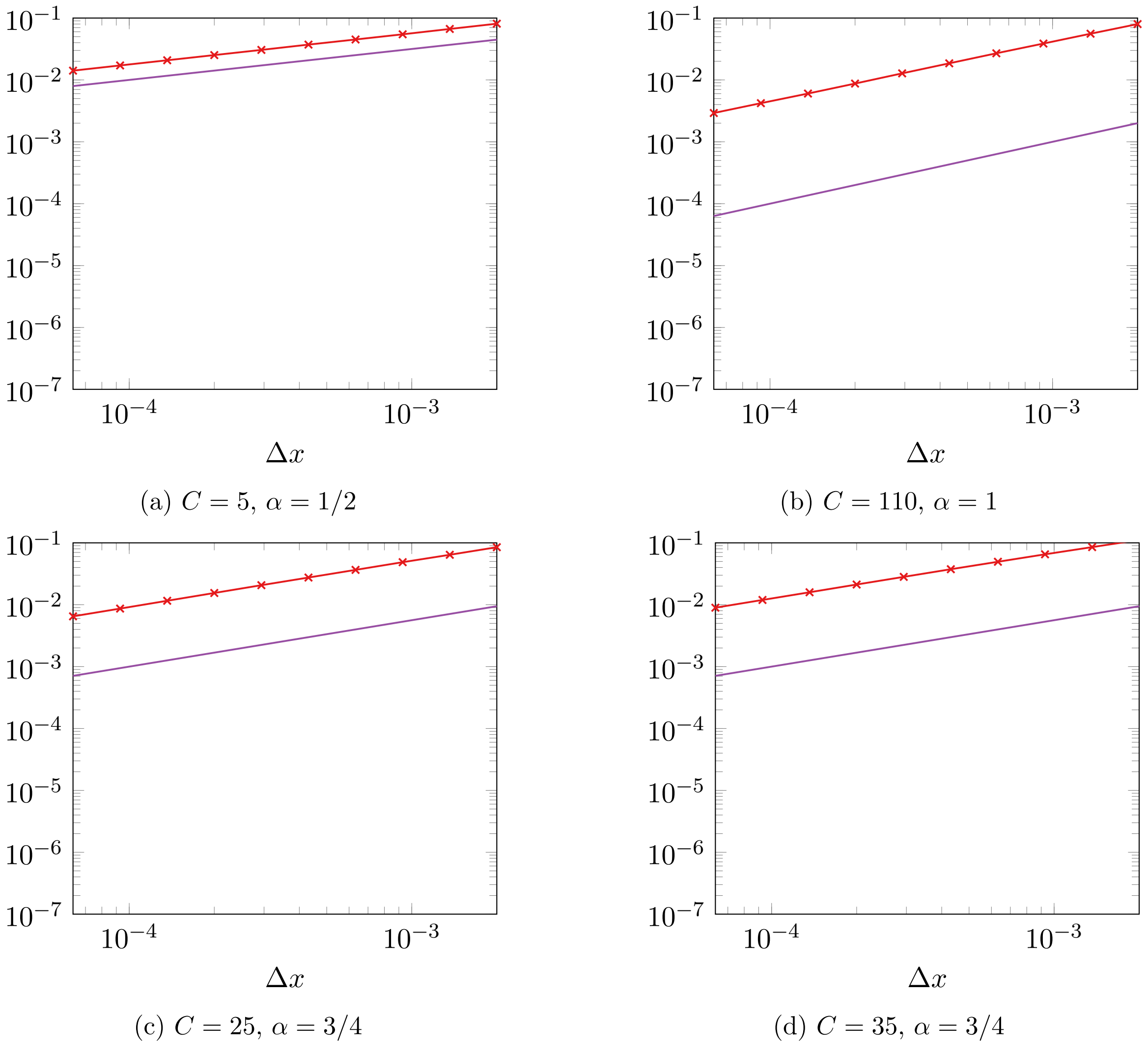


Figure 6.19: $\sum_i |\dot{U}_i - \dot{u}_{\text{broad}}(x_i)|\Delta x$ (red), Δx (purple)

Since the broad tangent $\dot{u}_{\text{broad}}$ does not contain the singularity of the analytic tangent sensitivity at the shock location and we have seen that the discrete tangent $\dot{U}$ empirically weakly converges to the analytic tangent sensitivity in Section 1.3.6 the discrete tangent does not converge to the analytic broad tangent as $\Delta x \to 0$. This nonconvergence is the motivation for the definition of the numerical broad tangent.

Figure 6.20 shows the approximation error of the numerical broad tangent sensitivity of Definition 40 obtained from the discrete tangent sensitivity of the Lax-Friedrichs numerical integrator.

The approximation error of the numerical broad tangent is directly dependent on the choice of α and C. We observe empirical convergence that is visually consistent with an $O(\Delta x^{\alpha})$ rate for all four combinations of α and C, i.e. Proposition 39 is empirically validated. By observation of the discrete tangent in, e.g., Figure 6.13 it is clear that by choosing $\alpha = 1$ and C too small the numerical broad tangent does not cut out all parts of the shock region as the shock region is also $O(\Delta x)$ wide. This phenomenon is not observed empirically here for the given range of Δx in combination with $C = 110$ being a large enough constant number of cells to consider as the shock region such that the potential approximation error from the tangent divergence is negligible compared to the regular numerical approximation error.



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

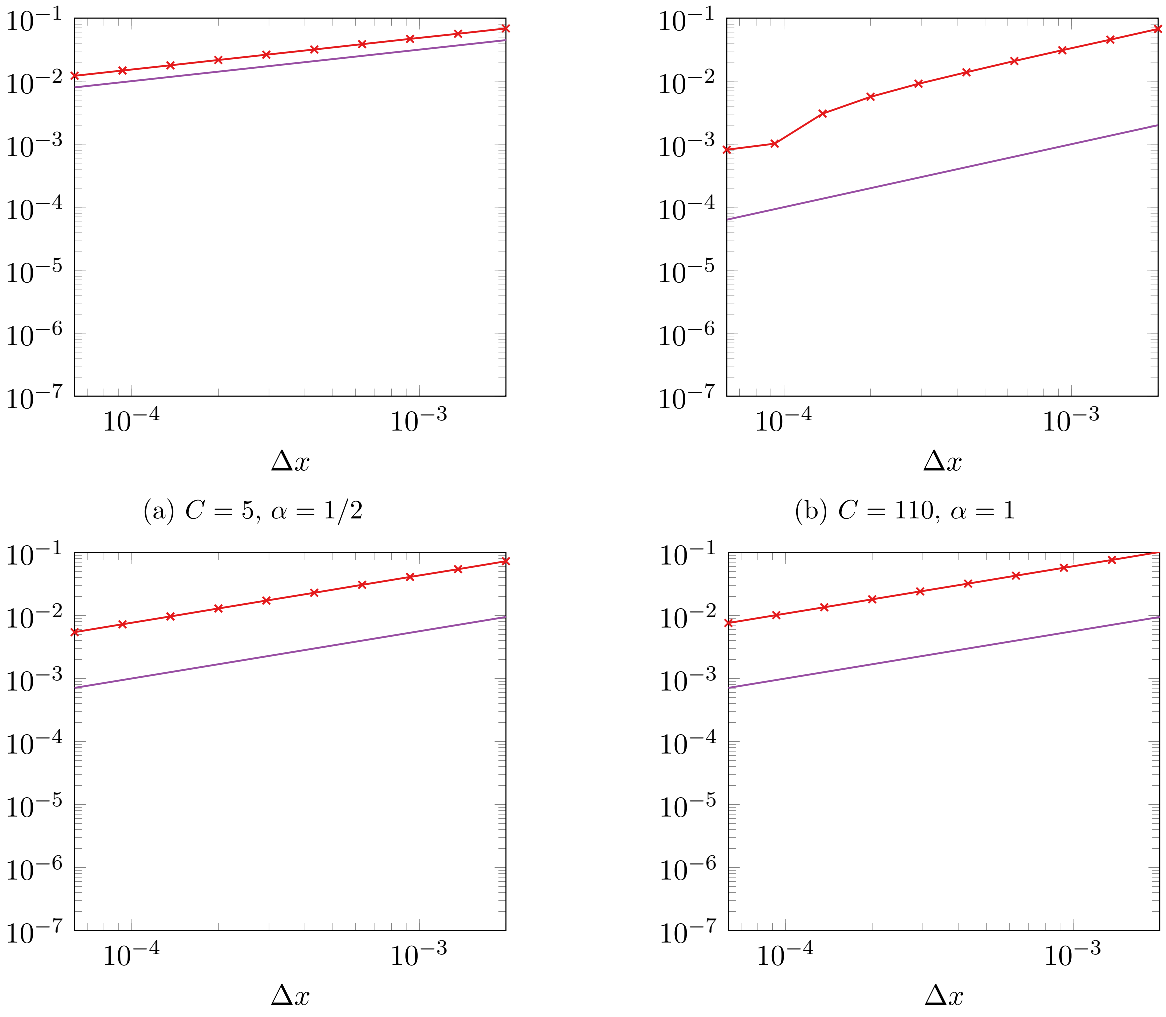
(d) $C = 35$, $\alpha = 3/4$

Figure 6.20: $\sum_i |(\dot{U}_{\text{broad}})_i - \dot{u}_{\text{broad}}(x_i)|\Delta x$ (red), Δx^α (purple)

Shock Sensitivity

We now consider the approximation error of the numerical shock sensitivity $\dot{\Xi}$. Figure 6.21 shows the approximation error of the shock sensitivity obtained using the shock sensitivity algorithm of Definition 39 with $C_r \equiv C$ in combination with the shock location obtained by the shock location algorithm of Definition 38 and with a Lax-Friedrichs numerical integrator.

Just like for the broad tangent sensitivity the approximation error of the shock sensitivity is directly dependent on the choice of α and C. We observe empirical convergence that is visually consistent with a $O(\Delta x^\alpha)$ rate for all four combinations of α and C, i.e. Proposition 38 is empirically validated. The best result as $\Delta x \to 0$ is, hence, obtained by choosing $\alpha = 1$ in this case as the constant C is large enough to stay far enough from the numerical instabilities of the shock region. Although the difference is almost not notable we observe that for the smaller choice of $\alpha = 3/4$ a smaller choice of C leads to a smaller error which makes sense because as long as we are far enough away from the shock region going out further will only increase the approximation error.



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

(d) $C = 35$, $\alpha = 3/4$

Figure 6.21: $|\dot{\Xi}_1 - \dot{\xi}_1|$ (red), Δx^{α} (purple)

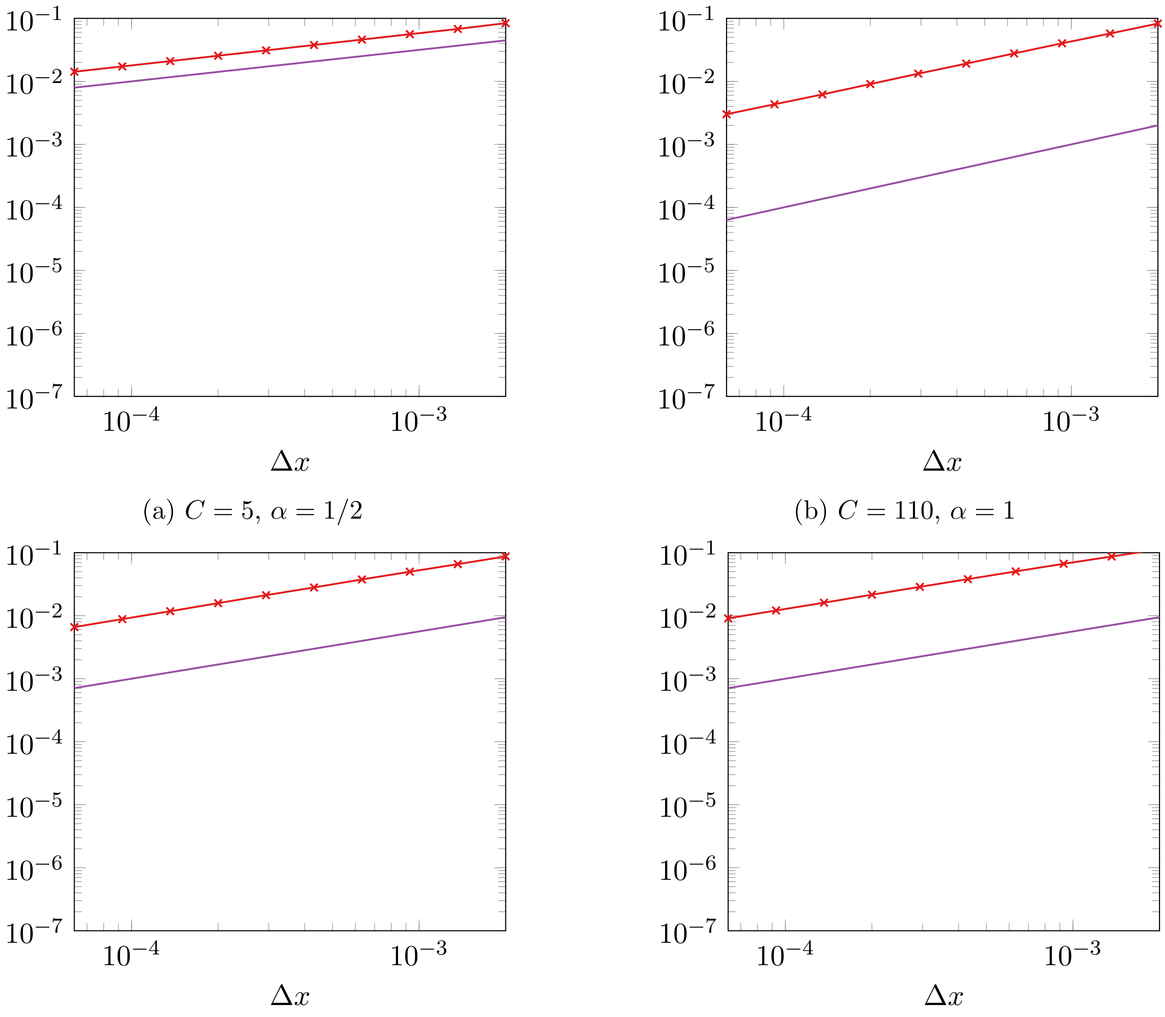
Generalized Tangent

We now consider the approximation error of the numerical generalized tangent sensitivity $\dot{U}_{\mathrm{gen}}$. We also consider the L_1 convergence in the sense of a first-order Taylor expansion as $\dot{p} \to 0$ when letting $\Delta x = O(\|\dot{p}\|^{2/\alpha})$.

Figure 6.22 shows the approximation error of the numerical generalized tangent sensitivity of Definition 41 obtained from the discrete tangent sensitivity of the Lax-Friedrichs numerical integrator.

In order to gain a visual understanding of the numerical generalized tangent compared to the analytic generalized tangent and to the divided difference of the analytic solution we consider Figure 6.10 of Example 38.

The approximation error of the numerical generalized tangent is directly dependent on the choice of α and C. We observe empirical convergence that is visually consistent with a $O(\Delta x^{\alpha})$ rate for all four combinations of α and C, i.e. Proposition 40 is empirically validated. Choosing



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

(d) $C = 35$, $\alpha = 3/4$

Figure 6.22: $\sum_i |(\dot{U}_{\text{gen}})_i - \dot{u}_{\text{gen}}(x_i)|\Delta x$ (red), Δx^{α} (purple)

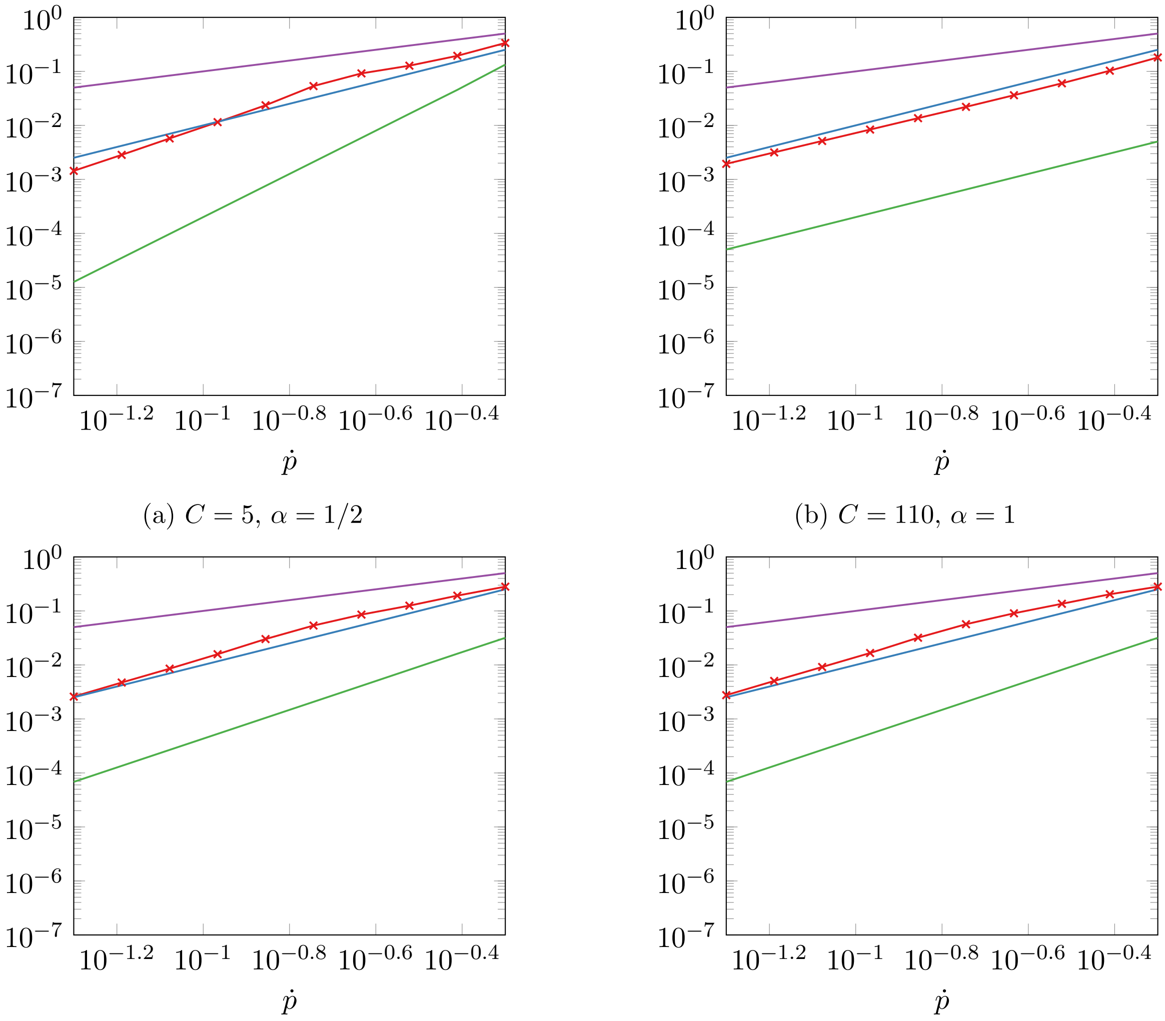
a higher value of α does not improve the approximation error for higher values of Δx initially since we need to pick a higher value of C as well but the difference becomes significant as $\Delta x \to 0$.

Figure 6.23 shows an approximation of the L_1 Taylor expansion error

$$\|u(\cdot, p_0 + \dot{p}) - u(\cdot, p_0) - \dot{u}_{\text{gen}}(\cdot, p_0)\|_{L_1(\hat{\Omega})} \approx \sum_{i=1}^{n} |U(p_0 + \dot{p})_i - U(p_0)_i - (\dot{U}_{\text{gen}})_i|\Delta x$$

with $p_0 = 0$ in red where the difference of the weak solution is approximated by the numerical solution and the generalized tangent sensitivity is approximated by the numerical generalized tangent sensitivity. The blue line represents a rate of $O(\|\dot{p}\|^2)$. The green line is a plot of the discretization grid distance $\Delta x = O(\|\dot{p}\|^{2/\alpha})$ where the constant is slightly different between the different figures.

The empirical convergence, as $\dot{p} \to 0$, shows a quadratic rate as predicted by Proposition 41 for the analytic finite difference. Considering that the numerical solution is also convergent to the weak solution we have quadratic approximation error by the nonlinear generalized tangent.



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

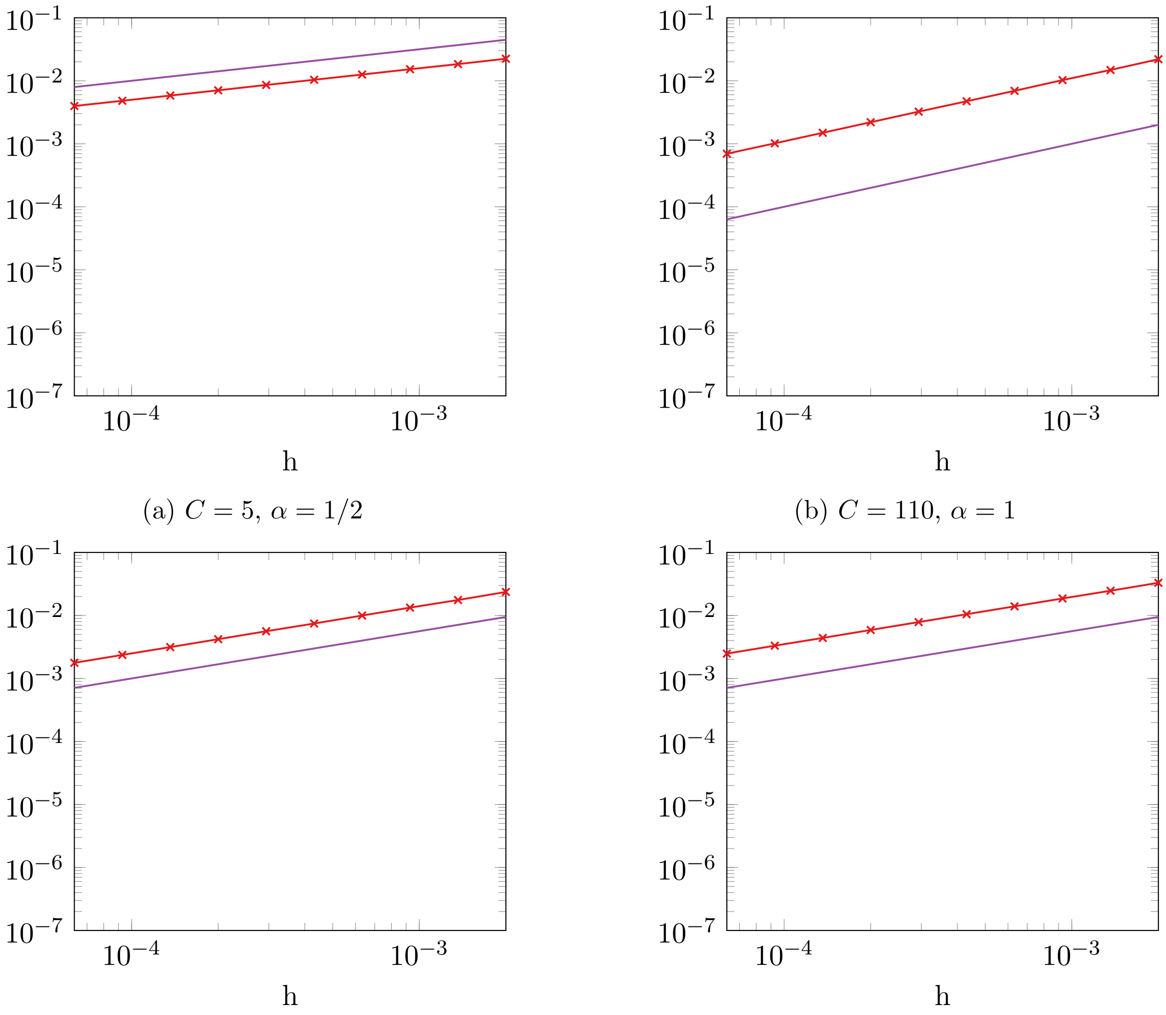
(d) $C = 35$, $\alpha = 3/4$

Figure 6.23: $\sum_i |U(p_0 + \dot{p})_i - U(p_0)_i - (\dot{U}_{\text{gen}})_i|\Delta x$ (red), $\dot{p}$ (purple), $\dot{p}^2$ (blue), Δx (green)

Adjoint Sensitivity

We now consider the approximation error of the discrete adjoint sensitivity $\bar{p}_{\text{shock}}$ with explicit shock handling for the numerical approximation of the least-squares integral objective in Equation (1.3.1) of the introduction where in Section 1.3.3 we derived the exact analytic adjoint sensitivity. We note that this example uses the Riemann initial value problem instead of the Ramp initial value problem used in the previous sections. According to the method described in Section 6.4 the explicit shock handling tangent sensitivity of the integral objective is computed by cutting out the divergent part of the tangent from the finite sum approximation of the integral and adding the jump discontinuity height approximation times the shock sensitivity. The discrete adjoint is obtained by utilizing the AD tool `dco/c++` according to the description in Section 6.5, i.e. using the nested adjoint of the tangent implementation.

Figure 6.24 shows the approximation error of the numerical tangent sensitivity of Definition 42 and the numerical adjoint sensitivity of Proposition 43 obtained from the discrete tangent sensitivitiy of the Lax-Friedrichs numerical integrator with a CFL time step with CFL number set to one and for $p = 1/10$, $\dot{p} = 1$ and $\bar{\Phi} = 1$ respecitvely. The approximation errors of the corresponding tangent and adjoints are equal according to the theory and this is also observed in practice. Each figure, therefore, only contains one approximation error value plotted in red.



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

(c) $C = 25$, $\alpha = 3/4$

(d) $C = 35$, $\alpha = 3/4$

Figure 6.24: $|\bar{p}_{\text{shock}} - \bar{p}|(= |\dot{\Phi}_{\text{shock}} - \dot{\Phi}|)$ (red), Δx^{α} (purple)

The approximation error of the explicit shock handling numerical tangent and adjoint sensitivity of the integral objective is directly dependent on the choice of α and C. We observe empirical convergence that is visually consistent with a $O(\Delta x^{\alpha})$ rate for all four combinations of α and C, i.e. Proposition 42 and Proposition 44 are both empirically validated.

6.6.3 Nonconvex Flux

We perform computational experiments to evaluate the generalized tangent sensitivity based on a shock sensitivity for a scalar conservation law initial value problem resulting from the nonconvex Buckley-Leverett flux function

$$f(u) \equiv \frac{u^2}{u^2 + \frac{1}{2}(1-u)^2}$$

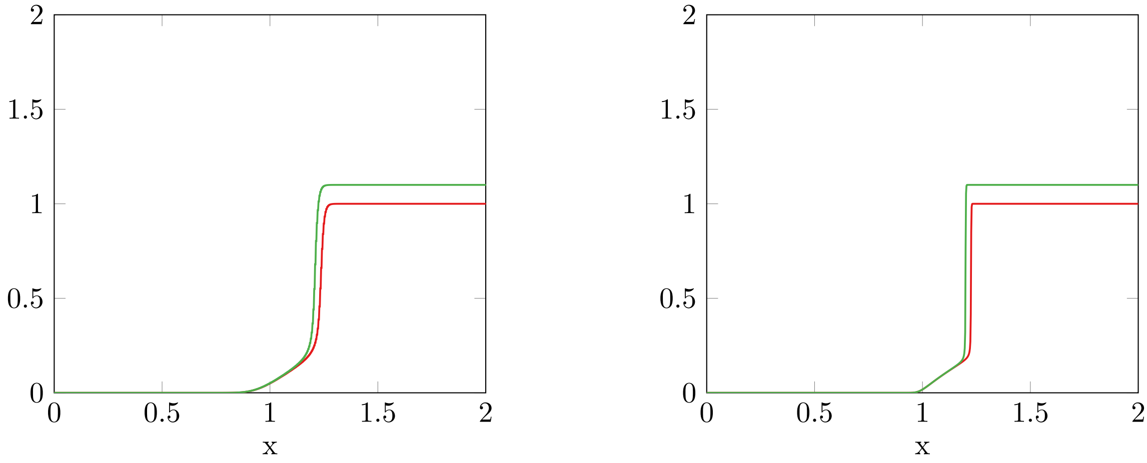
that we already introduced in Section 5.1.1. We analyze a parametric initial value problem and show that the numerical generalized tangent sensitivity exhibits the desired quadratic rate for the error term.

The parametric initial value problem we consider is given by the initial data

$$u_0(x,p) = \begin{cases} 1+p & \text{if } x \geq 1 \\ 0 & \text{otherwise} \end{cases} .$$

The initial data represents a shifted Riemann problem for the Buckley-Leverett flux with a shock discontinuity at $x = 1$. The shift away from zero is not relevant since the flux does not depend on the location x.

Figure 6.25 shows the Lax-Friedrichs numerical solution of the initial value problem at $t = 1/5$ for $p = p_0 = 0$ in red and for $p = p_0 + \dot{p} = 1/10$ in green.



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

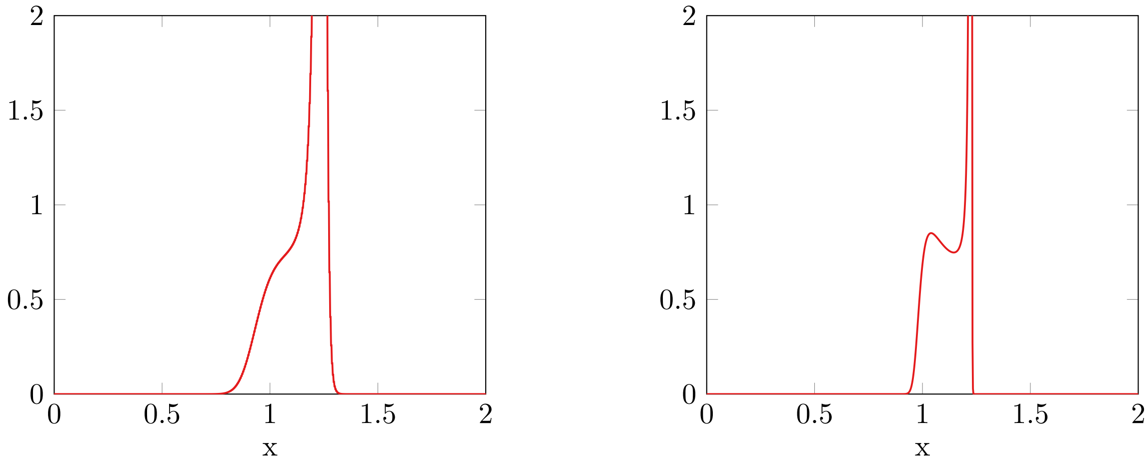
Figure 6.25: $U(p_0)$ (red), $U(p_0 + \dot{p})$ (green)

For both numerical solutions we observe a discontinuity that is traveling to the right compared to the initial data. But due to the nonconvexity of the flux function the shock does not form for solution values u below the value of $1/4$ where the flux function turns concave (see Figure 5.1 in Section 5.1.1). In a sense, the discontinuity in the initial data decomposes into a shock discontinuity and a rarefaction wave. Still, there is a shock discontinuity component that will lead to a divergent discrete tangent the location of which we can simulate and approximate sensitivities for.

Figure 6.26 shows the derivative of the weak solution with respect to space as approximated by a finite difference.

We observe that the derivative with respect to space is singular at the shock location as necessary by the fact that the weak solution is discontinuous at that point. But unlike the previous examples for the Burgers equation the derivative with respect to space is also nonlinear to the left of the shock. Previously, the derivative with respect to space was constant and, hence, its approximation error was not affected by evaluating too far to the left of the shock location.

Figure 6.27 and Figure 6.28 show with $p_0 = 0, \dot{p} = 1/10$ the discrete tangent sensitivity $\dot{U} = \mathrm{d}_p U(p_0)\dot{p}$ in red, the difference of the numerical solutions $U(p_0 + \dot{p}) - U(p_0)$ in blue and the



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

Figure 6.26: $(U(x + \Delta x) - U(x - \Delta x))/(2\Delta x)$ (red)

numerical generalized tangent sensitivity computed based on the shock sensitivity according to Definition 39 and the numerical broad tangent according to Definition 40 in green. Figure 6.27 uses the parameters $C = 5, \alpha = 1/2$ and Figure 6.28 uses the parameters $C = 110, \alpha = 1$. The top two figures represent the whole domain respectively and the bottom two are zoomed into the shock region. The left pictures are of a less accurate discretization and the right pictures are of a more accurate discretization.

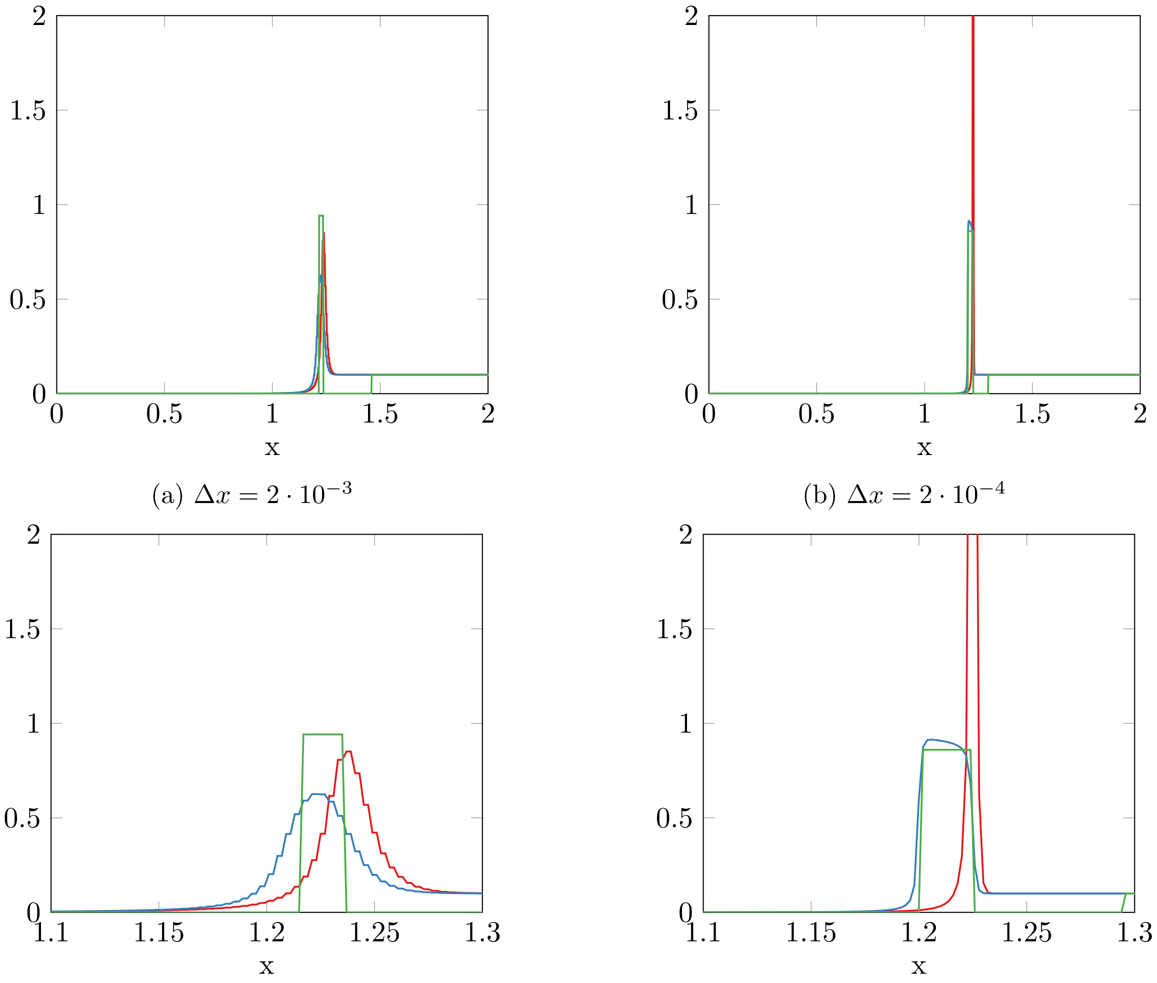
We observe that, naturally, the discrete tangent diverges at the shock location as $\Delta x \to 0$. To the right of the shock region the green, blue and red line all approximate the exact tangent sensitivity of the weak solution with $\dot{p} = 1/10$. Going from right to left up to the shock location the discrete tangent in red coincides with the finite difference in blue, but the green generalized tangent becomes zero intermittently because of the cut out shock region according to the definition of the numerical broad tangent. For the more accurate discretization we observe that the green box clearly approximates the blue finite difference box that describes the additional shock travel due to the perturbation of p by $\dot{p}$. Both the height of the box, i.e. the jump height of the shock discontinuity, and the width, i.e. shock sensitivity, are approximated accurately for the example problem and for $\alpha = 1/2$.

For $\alpha = 1$ we observe that the cut out shock region is much smaller for the more accurate discretization. The result is that the numerical generalized tangent sensitivity has a visibly nonzero component of the discrete tangent sensitivity added to the green box approximating the shock travel. The numerical generalized tangent sensitivity is also a visibly accurate approximation of the finite difference at least for the more accurate discretization.

We now consider the empirical convergence in the sense of the first-order Taylor expansion via the generalized nonlinear tangent sensitivity to the numerical solution difference just like we did in the previous section.

Figure 6.29 shows the approximation error of the numerical generalized tangent sensitivity of Definition 41 obtained from the discrete tangent sensitivity of the Lax-Friedrichs numerical integrator.

In the figure with $\alpha = 1/2$ we initially observe an unusually high approximation error because of



(a) $\Delta x = 2 \cdot 10^{-3}$

(b) $\Delta x = 2 \cdot 10^{-4}$

(c) $\Delta x = 2 \cdot 10^{-3}$

(d) $\Delta x = 2 \cdot 10^{-4}$

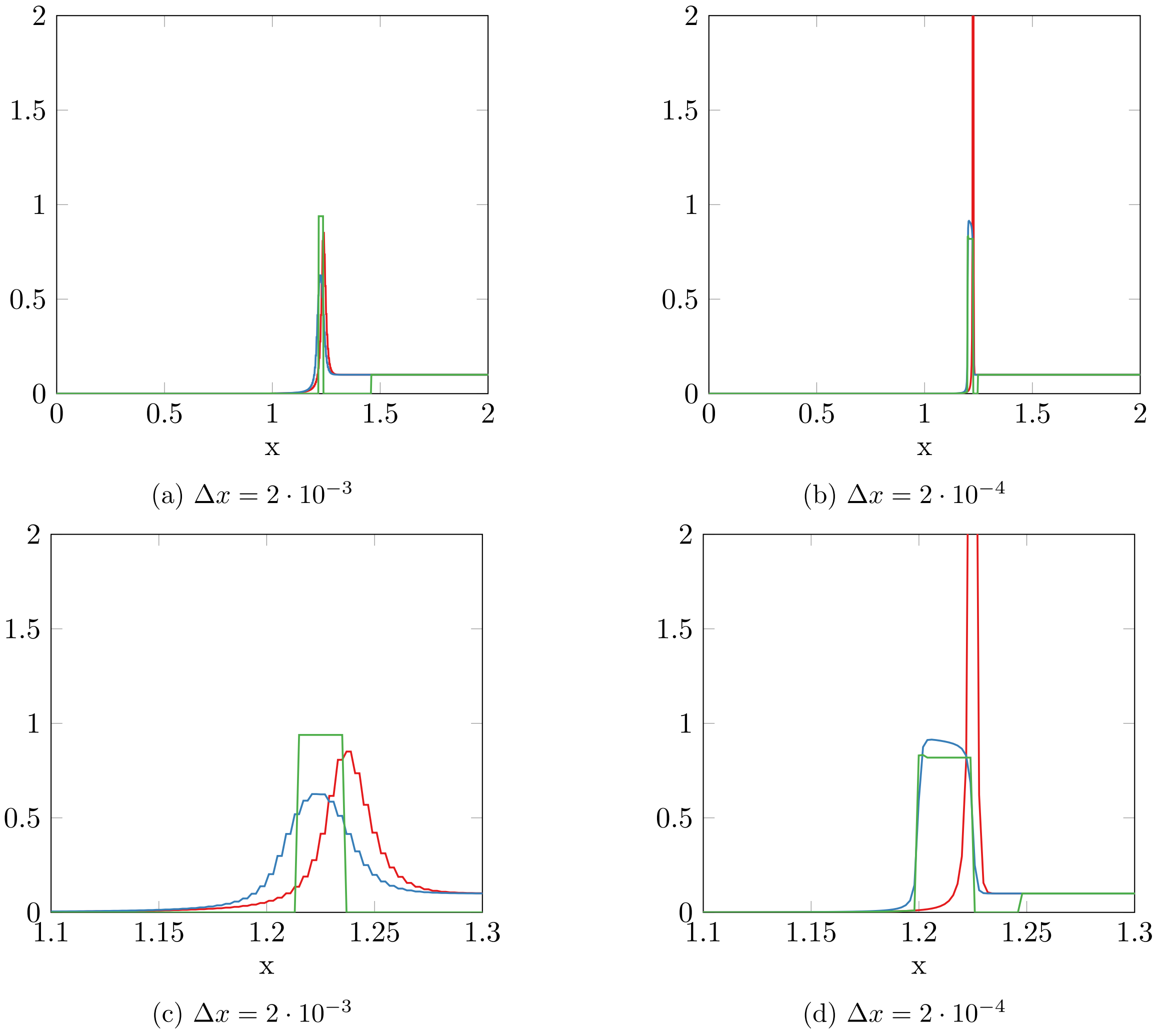
Figure 6.27: $\dot{U}$ (red), $U(p_0 + \dot{p}) - U(p_0)$ (blue), $\dot{U}_{\text{gen}}$ with $C = 5, \alpha = 1/2$ (green)

the low accuracy of the discretization in general; see the green line that represents discretization grid width Δx. But as $\Delta x \to 0$ we observe an empirical convergence that is bounded by the $O(|\dot{p}|^2)$ rate. For $\alpha = 1$ we observe the same bound for the values of Δx shown in the figure.

Hence, for the nonconvex Buckley-Leverett flux function example we make the same observation as we did for the Burgers equation example. The empirical convergence as $\dot{p} \to 0$ has a quadratic rate as predicted by Proposition 41 for the analytic finite difference. Considering that the numerical solution is also convergent to the weak solution we have a quadratic approximation error by the nonlinear generalized tangent. The observation of the quadratic rate here is not possible without a convergent approximation of the shock sensitivity and height of the jump discontinuity.

We conclude that all components of the suggested method for the approximation of shock sensitivities and the resulting generalized tangent also work for nonconvex flux functions in practical applications. Consequently, also the proposed method for the discrete adjoint sensitivity of integral objectives can be expected to work for nonconvex flux functions.

We gave the numerical analysis of methods for the approximation of shock locations, sensi-



(a) $\Delta x = 2 \cdot 10^{-3}$

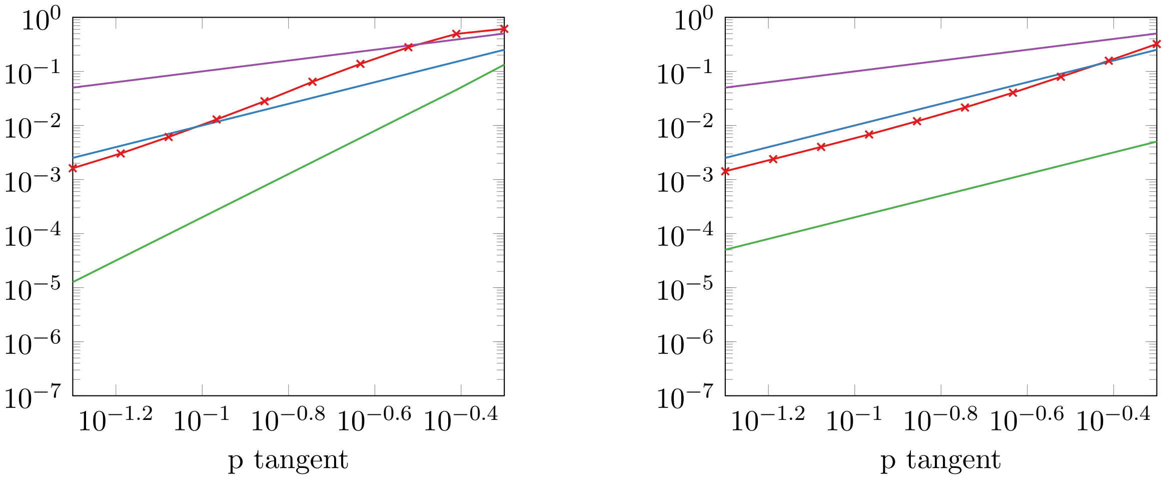
(b) $\Delta x = 2 \cdot 10^{-4}$

(c) $\Delta x = 2 \cdot 10^{-3}$

(d) $\Delta x = 2 \cdot 10^{-4}$

Figure 6.28: $\dot{U}$ (red), $U(p_0 + \dot{p}) - U(p_0)$ (blue), $\dot{U}_{\text{gen}}$ with $C = 110, \alpha = 1$ (green)

tivities and generalized tangent and adjoint sensitivities for discontinuous solutions of scalar conservation laws and validated the numerical methods via computational experiments with a convex and a nonconvex flux function. In the next chapter we extend the presented methods to one-dimensional hyperbolic systems of conservation laws and validate those extensions via computational experiments with a system of isentropic gas equations.



(a) $C = 5$, $\alpha = 1/2$

(b) $C = 110$, $\alpha = 1$

Figure 6.29: $\sum_i |U(p_0 + \dot{p}_0)_i - U(p)_i - (\dot{U}_{\text{gen}})_i|\Delta x$ (red), $\dot{p}$ (purple), $\dot{p}^2$ (blue), Δx (green)

Chapter 7

Discrete Shock Sensitivities in Hyperbolic Systems of Conservation Laws

In this chapter we show how the numerical methods for approximating shock sensitivities in scalar conservation laws introduced in the previous chapter extend to hyperbolic systems of conservation laws.

Example 39 (Isentropic gas equations)**.** As a driving example throughout this chapter we consider the model for isentropic gas dynamics in Lagrangian coordinates introduced in Example 5 of Bressan [Bre99]. We denote the system of isentropic gas equations as

$$\begin{aligned}\partial_t u + \partial_x P(v) &= 0\\ \partial_t v - \partial_x u &= 0\end{aligned}$$

where u is the velocity, $v > 0$ is the volume, with $v = \rho^{-1}$ where ρ is the density, and P is a pressure function.

We introduce the background on hyperbolic systems of conservation laws that is necessary to understand how characteristics and shock discontinuities behave in such systems. Based on the example of the isentropic gas equations we show how to discretize the discontinuous ODE that gives the characteristic speed. We also show how to differentiate the Rankine-Hugoniot condition for a hyperbolic system and specifically for the isentropic gas equations. We then validate the proposed method for hyperbolic systems based on the empirical first-oder Taylor expansion convergence of the discrete shock sensitivities in computational experiments with a parametric Riemann initial value problem for the isentropic gas equations.

7.1 Hyperbolic Systems of Conservation Laws

We briefly give the necessary background to understand the discretization and shock dynamics in a hyperbolic system of conservation laws. For a deeper understanding of methods for hyperbolic systems of conservation laws we refer the reader to Lax [Lax57], Godlewski and Raviart [GR13], LeVeque [LeV02], and Bressan [Bre13; Bre00; Bre99].

7.1.1 Discretization

We now consider the discretization of systems of conservation laws based on the methods we introduced for scalar conservation laws.

In order to keep things simple we consider the example of a 2 × 2 system of conservation laws

$$\begin{aligned}\partial_t u + \partial_x F_1(u,v) &= 0\\ \partial_t v + \partial_x F_2(u,v) &= 0\end{aligned}$$

where u, v are the two variable that are determined by two equations with the flux functions F_1, F_2. For both equations integration over a subset of the space domain will yield dynamics of the integral of both u and v. The integral changes over time only according to the corresponding flux function evaluated at the boundaries of the integral domain just like for a scalar conservation law. Discrete conservation concepts should, therefore, apply in an analogous manner.

We can use the same discretization methods introduced in Section 3.2 for scalar conservation laws by considering the two equations separately. For example the Lax-Friedrichs scheme applied to the above 2 × 2 system yields

$$U_i^{j+1} = \frac{1}{2}(U_{i-1}^j + U_{i+1}^j) - \frac{\Delta t_j}{2\Delta x}(F_1(U_{i+1}^j, V_{i+1}^j) - F_1(U_{i-1}^j, V_{i-1}^j)) \tag{7.1.1}$$

$$V_i^{j+1} = \frac{1}{2}(V_{i-1}^j + V_{i+1}^j) - \frac{\Delta t_j}{2\Delta x}(F_2(U_{i+1}^j, V_{i+1}^j) - F_2(U_{i-1}^j, V_{i-1}^j)) \; . \tag{7.1.2}$$

The finite difference approximation of the derivative of the flux with respect to space requires us to difference in both variables u, v as they both depend on x. Regarding the time step we assume that $\Delta t_j = O(\Delta x)$ as $\Delta x \to 0$ and we can just pick a constant time step $\Delta t_j = \Delta t \equiv C_t \Delta x$ for all $j \in \{1, \dots, m\}$ with $C_t > 0$ small enough to guarantee stability. We later also introduce an adaptive time step based on the CFL condition for the system case.

When considering the viscous system of equations that corresponds to the Lax-Friedrichs discretization with a constant time step (c.f. Section 3.2.8) we have

$$\begin{aligned}\partial_t u + \partial_x F_1(u,v) &= \frac{\Delta x^2}{2\Delta t}\partial_{xx}^2 u\\ \partial_t v + \partial_x F_2(u,v) &= \frac{\Delta x^2}{2\Delta t}\partial_{xx}^2 v \; .\end{aligned}$$

We see that the numerical viscosity applies a diffusion in both u and v on an equal scale.

7.1.2 Hyperbolicity

We now define what it means for a system of conservation laws to be hyperbolic based on the eigenvalues of the flux Jacobian.

We can write a general $N \times N$ system of conservation laws in standard form as

$$\partial_t \boldsymbol{u} + \partial_x F(\boldsymbol{u}) = 0$$

with $\boldsymbol{u}(t,x) \in \mathbb{R}^N$ and $F(\boldsymbol{u}) \in \mathbb{R}^N$. By the chain rule we have

$$\partial_x F(\boldsymbol{u}) = A(\boldsymbol{u})\partial_x \boldsymbol{u}$$

where $A(\boldsymbol{u}) \in \mathbb{R}^{N\times N}$ is the Jacobian of F with respect to $\boldsymbol{u}$. Hence, the same $N \times N$ system of conservation laws as above written in quasilinear form is

$$\partial_t \boldsymbol{u} + A(\boldsymbol{u})\partial_x \boldsymbol{u} = 0 \tag{7.1.3}$$

where $A(\boldsymbol{u})$ is the flux Jacobian.

We consider the quasilinear form because it is useful for intuitively understanding the dynamics of a nonlinear system based on properties of the flux Jacobian by analogy to linear systems over infinitesimal time steps. In the next section we focus on understanding the solution dynamics of linear systems of conservation laws.

Example 40 (Isentropic gas equations)**.** The Jacobian of the isentropic gas equations flux function is

$$A(u,v) = \begin{pmatrix} 0 & P'(v) \\ -1 & 0 \end{pmatrix} .$$

We recall the definition of the eigendecomposition of a matrix (see e.g. [Fra12, Chapter 4], [GV13, Chapter 2.1.7]).

Let A be an $N \times N$ matrix with N linearly independent eigenvectors q_i. Then A can be factorized as

$$A = Q\Lambda Q^{-1}$$

where $Q = [q_1 q_2 \dots q_n]$ is the matrix with the eigenvectors as columns and $\Lambda_{i,i} = \lambda_i$ is the diagonal matrix with eigenvalues as diagonal entries.

Definition 43. The system of conservation laws in Equation (7.1.3) is *hyperbolic* if $A(u)$ has N linearly independent eigenvectors and *strictly hyperbolic* if $A(u)$ has N real, distinct eigenvalues

$$\lambda_1(\boldsymbol{u}) < \lambda_2(\boldsymbol{u}) < \cdots < \lambda_N(\boldsymbol{u})$$

for all values of $\boldsymbol{u}$ taken by the solution.

7.1.3 Linear Systems

We now consider the case of a linear $N \times N$ system of conservation laws in order to gain an understanding about what hyperbolicity implies for the solution of such systems.

A linear $N \times N$ system of conservation laws is given by

$$\partial_t \boldsymbol{u} + A\partial_x \boldsymbol{u} = 0 \tag{7.1.4}$$

where $A \in \mathbb{R}^{N\times N}$ is the operator that defines a linear flux function.

If the system is hyperbolic then A has N linearly independent eigenvectors and, hence, A has the eigendecomposition $A = Q\Lambda Q^{-1}$ as defined above. We then have

$$\begin{aligned} \partial_t \boldsymbol{u} + A\partial_x \boldsymbol{u} &= 0 \\ \partial_t \boldsymbol{u} + Q\Lambda Q^{-1}\partial_x \boldsymbol{u} &= 0 && \text{(eigendecomposition)} \\ Q^{-1}\Big(\partial_t \boldsymbol{u} + Q\Lambda Q^{-1}\partial_x \boldsymbol{u}\Big) &= Q^{-1}0 && \text{(left multiply } Q^{-1}\text{)} \end{aligned}$$

$$\begin{aligned} Q^{-1}\partial_t \boldsymbol{u} + \Lambda Q^{-1}\partial_x \boldsymbol{u} &= 0 && (Q^{-1}Q = I) \\ \partial_t Q^{-1}\boldsymbol{u} + \Lambda \partial_x Q^{-1}\boldsymbol{u} &= 0 && \text{(linearity of differentiation) .} \end{aligned}$$

Since Λ is a diagonal matrix the last equation corresponds to N independent scalar conservation laws for the variable $w \in \mathbb{R}^N$ via the reparameterization to $\boldsymbol{w} = Q^{-1}\boldsymbol{u}$, i.e. we have

$$\partial_t w_i + \lambda_i \partial_x w_i = 0$$

for all $i \in \{1, \dots, N\}$.

From Example 10 we recall that the solution to a linear scalar conservation law is a transport of the initial data by a distance determined by the product of time and the characteristic speed given by the scalar linear factor. For the above set of independent linear scalar conservation laws we have the solution

$$\begin{aligned} w_i(t,x) &= (w_0)_i(x - t\lambda_i) && \text{(scalar solution)} \\ &= \langle Q^{-1}{}_{i,:}, \boldsymbol{u}_0(x - t\lambda_i) \rangle && \text{(definition of } w) \end{aligned}$$

where $\boldsymbol{u}_0, \boldsymbol{w}_0$ are the corresponding initial data of the original system and the reparameterization.

Due to the linearity of the reparameterization the solution of the original system is a superposition of linear transports of the initial data.

Proposition 45. Let Equation (7.1.4) be a hyperbolic system of conservation laws and $A = Q\Lambda Q^{-1}$ the eigendecomposition of A. Then the weak solution of Equation (7.1.4) is

$$\boldsymbol{u}(t,x) = \sum_{i=1}^{N} Q_{:,i} \cdot (u_0)_i(x - t\lambda_i)$$

where u_0 is the initial data.

Proof. To get the first term of Equation (7.1.4) we differentiate u according to the proposition with respect to t and have

$$\begin{aligned} \partial_t \boldsymbol{u}(t,x) &= \partial_t \sum_{i=1}^{N} Q_{:,i} \cdot (u_0)_i(x - t\lambda_i) && (u \text{ according to proposition}) \\ &= \sum_{i=1}^{N} Q_{:,i} \cdot \partial_t (u_0)_i(x - t\lambda_i) && \text{(linearity of differentiation)} \\ &= -\sum_{i=1}^{N} Q_{:,i} \cdot \partial_x (u_0)_i(x - t\lambda_i) \cdot \lambda_i && \text{(chain rule)} \\ &= -\partial_x \sum_{i=1}^{N} \lambda_i \cdot Q_{:,i} \cdot (u_0)_i(x - t\lambda_i) && \text{(linearity of differentiation) .} \end{aligned}$$

To get the second term of Equation (7.1.4) we differentiate u as given in the proposition with respect to x and have

$$\partial_x A\boldsymbol{u}(t,x) = \partial_x A \sum_{i=1}^{N} Q_{:,i} \cdot (u_0)_i(x - t\lambda_i) \qquad (u \text{ according to proposition})$$

$$
\begin{aligned}
&= \partial_x Q \Lambda Q^{-1} \sum_{i=1}^{N} Q_{:,i} \cdot (u_0)_i(x - t\lambda_i) && \text{(eigendecomposition)} \\
&= \partial_x Q \Lambda \sum_{j=1}^{N} Q^{-1}{}_{:,j} \sum_{i=1}^{N} Q_{j,i} \cdot (u_0)_i(x - t\lambda_i) && \text{(matrix vector product)} \\
&= \partial_x Q \Lambda \sum_{i=1}^{N} \sum_{j=1}^{N} Q^{-1}_{:,j} Q_{j,i} \cdot (u_0)_i(x - t\lambda_i) && \text{(distributivity)} \\
&= \partial_x Q \Lambda \sum_{i=1}^{N} I_{:,i} \cdot (u_0)_i(x - t\lambda_i) && (Q^{-1}Q = I) \\
&= \partial_x \sum_{i=1}^{N} \lambda_i \cdot Q_{:,i} \cdot (u_0)_i(x - t\lambda_i) && \text{(matrix vector product)}
\end{aligned}
$$

For weakly differentiable $\boldsymbol{u}$ we have shown that in a weak sense we have

$$\partial_t \boldsymbol{u}(t,x) + \partial_x A \boldsymbol{u}(t,x) = 0$$

for all t, x in the solution domain and, hence, u is a weak solution of Equation (7.1.4). This concludes the proof. □

The solution of a scalar linear conservation law is a traveling wave where the characteristic speed is given by the linear factor. The solution of a hyperbolic linear system of conservation laws is a linear combination of traveling waves where the characteristic speeds of the individual wave components are given by the eigenvalues of the flux matrix A.

Example 41 (Linear system)**.** We consider the example linear system given by Equation (7.1.4) with the flux matrix

$$A = \begin{pmatrix} 1 & 4 \\ \frac{1}{2} & 1 \end{pmatrix}$$

that has the eigenvalues

$$\lambda_1 = 2, \quad \lambda_2 = -2 \ .$$

According to the opposing signs of the eigenvalues we already see that the individual wave components of the solution travel in opposing directions.

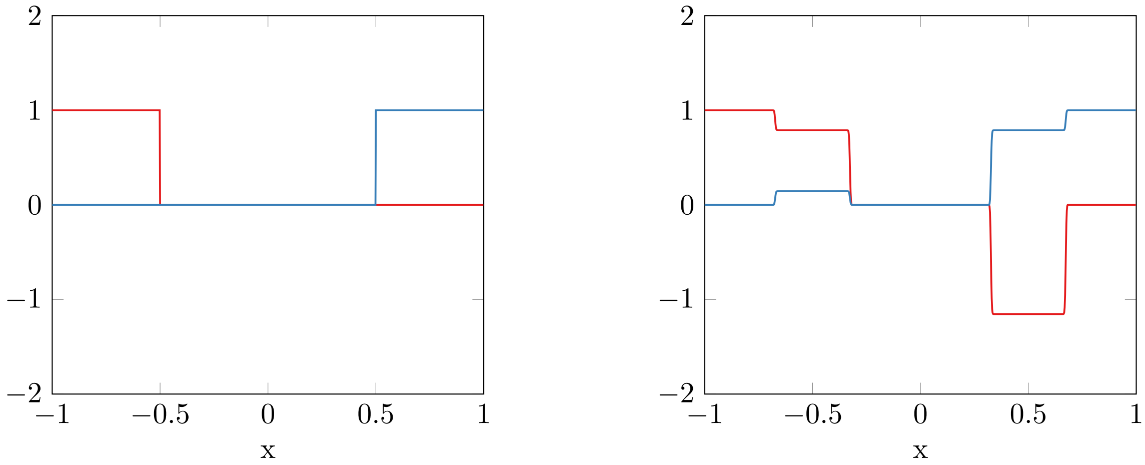
The initial data we consider is given by two discontinuous functions

$$(u_0)_1(x) = H\Big(-\frac{1}{2} - x\Big), \quad (u_0)_2(x) = H\Big(x - \frac{1}{2}\Big)$$

where $(u_0)_1$ jumps down from one to zero at $x = -1/2$ and $(u_0)_2$ jumps up from zero to one at $x = 1/2$.

Figure 7.1 shows a numerical approximation of the weak solution for u_1 in red and u_2 in blue at $t = 0$ on the left and $t = 1/10$ on the right. The numerical approximation is via a Lax-Friedrichs scheme as described in Equations (7.1.1)-(7.1.2).

In the left figure we see a visualization of the initial data where the step function wave components do not mix yet. In the right figure we see a linear combination of two step function



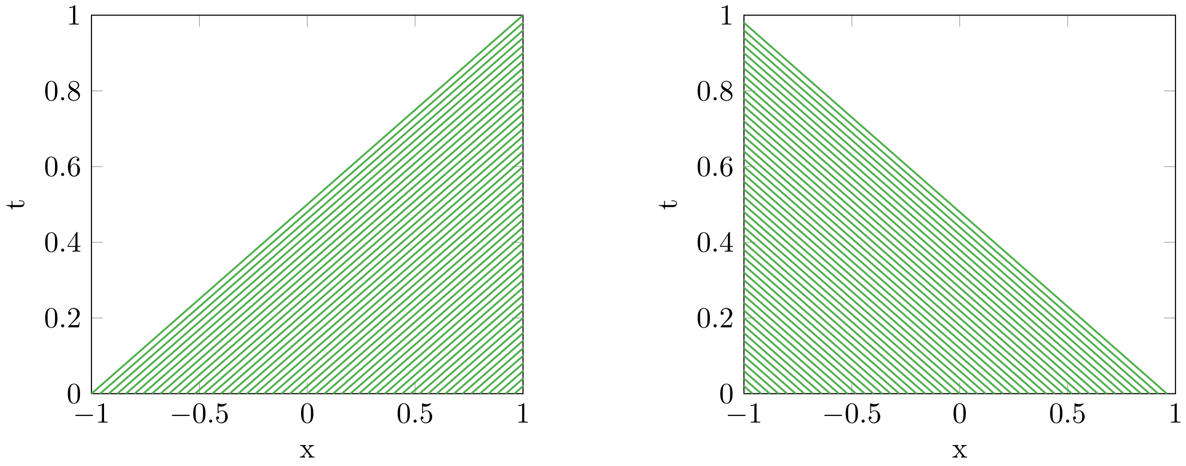
(a) $t = 0$

(b) $t = 1/10$

Figure 7.1: Weak solution of example strictly hyperbolic linear system, u_1 (red), u_2 (blue)

waves in both the red and the blue graphs. The wave component resulting from the u_1 initial data has traveled to the right and the wave component resulting from the u_2 initial data has traveled to the left.

Figure 7.2 shows a visualization of the characteristic lines starting on an equidistant grid along the visible domain with the dynamics according to the characteristic speed determined by the eigenvalues of A. Because the system is strictly hyperbolic the eigenvalues are real and the dynamics of the characteristics have real solutions.



(a) $\partial_t \xi(t) = \lambda_1$

(b) $\partial_t \xi(t) = \lambda_2$

Figure 7.2: Characteristics of the traveling wave components in a hyperbolic linear system

Crucially, even though we have a discontinuous weak solution for the above example the discontinuities that exist in the initial data and those that form due to the linear combination are not shock discontinuities. The characteristics in a linear conservation law do not cross and, hence, they do not produce compression in the solution. In the next section we consider a strictly hyperbolic nonlinear system of conservation laws where shocks form in the solution.

7.2 Shock Location for the Isentropic Gas Equations

We now consider the dynamics of the characteristics of the isentropic gas equations in Example 39 and show their evolution for an initial value problem that develops a shock discontinuity. We then discuss how the shock location of such a shock discontinuity can be numerically approximated based on the method developed for scalar conservation laws in the previous chapter.

Just like for a linear system the characteristic speeds for a nonlinear system at a specific time and space position (t, x) are given by the eigenvalues of the flux Jacobian evaluated at (t, x). According to Example 40 the Jacobian of the isentropic gas equation flux at (t, x) is given by

$$A(u(t,x), v(t,x)) = \begin{pmatrix} 0 & P'(v(t,x)) \\ -1 & 0 \end{pmatrix} .$$

We obtain the eigenvalues as the roots of the characteristic polynomial

$$0 = \det \begin{bmatrix} 0-\lambda & P'(v(t,x)) \\ -1 & 0-\lambda \end{bmatrix} = \lambda^2 + P'(v(t,x)) .$$

We have the two eigenvalues

$$\lambda_1(u,v) = \sqrt{-P'(v(t,x))}, \qquad \lambda_2(u,v) = -\sqrt{-P'(v(t,x))} \tag{7.2.1}$$

as the solution of the quadratic equation.

The function P describes a pressure that in a physically feasible context satisfies the constraints $P > 0$, $P' < 0$, $P'' > 0$. For example in the following computational results we use the pressure function

$$P(v) \equiv \frac{k}{v^\gamma}$$

with some constant $k > 0$ and $\gamma \in (1, 3)$ the adiabatic gas constant. For $v > 0$ that pressure function P satisfies the physical constraints as given above since

$$P'(v) = -\gamma \frac{k}{v^{\gamma+1}}, \qquad P''(v) = (\gamma^2 + \gamma) \frac{k}{v^{\gamma+2}} .$$

In the case of $P' < 0$ we have $-P' > 0$ and $\sqrt{-P'}$ is real and nonzero. So for a physically feasible pressure function the above two eigenvalues λ_1, λ_2 are distinct and real, i.e. the system of isentropic gas equations is strictly hyperbolic.

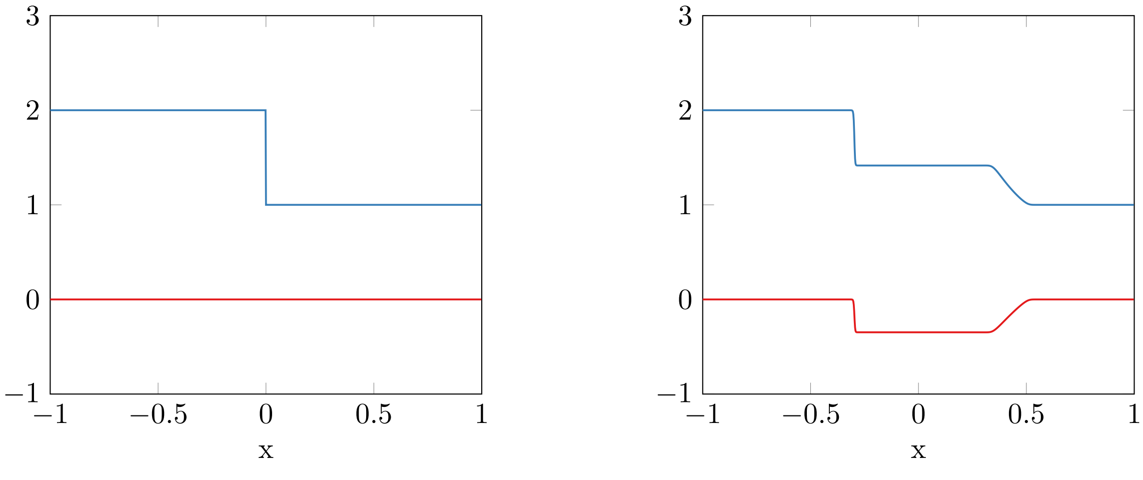
Example 42 (Riemann problem)**.** We consider a Riemann initial value problem for the isentropic gas equations with initial data given by

$$u_0(x) = 0, \qquad v_0(x) = 1 + H(-x)$$

and the pressure function $P(v) = 1/v$.

Figure 7.3 shows a numerical approximation of the weak solution for u in red and v in blue at $t = 0$ on the left and $t = 1/2$ on the right. The numerical approximation is via the Lax-Friedrichs scheme as described in Equations (7.1.1)-(7.1.2).

In the left figure we just see a visualization of the initial data. The velocity u is zero everywhere initially but the volume, or inverse pressure, v is higher to the left and lower to the right of



(a) $t = 0$

(b) $t = 1/2$

Figure 7.3: Weak solution of example strictly hyperbolic nonlinear system, u (red), v (blue)

$x = 0$. As a result of the volume imbalance a shock wave develops at the origin and travels to the left over time. At the same time a rarefaction wave develops at the origin and travels to the right over time. In the right figure we see the state at $t = 1/2$ where we observe a discontinuity in both variables u and v that upon inspection of the characteristics over time turns out to be a shock discontinuity.

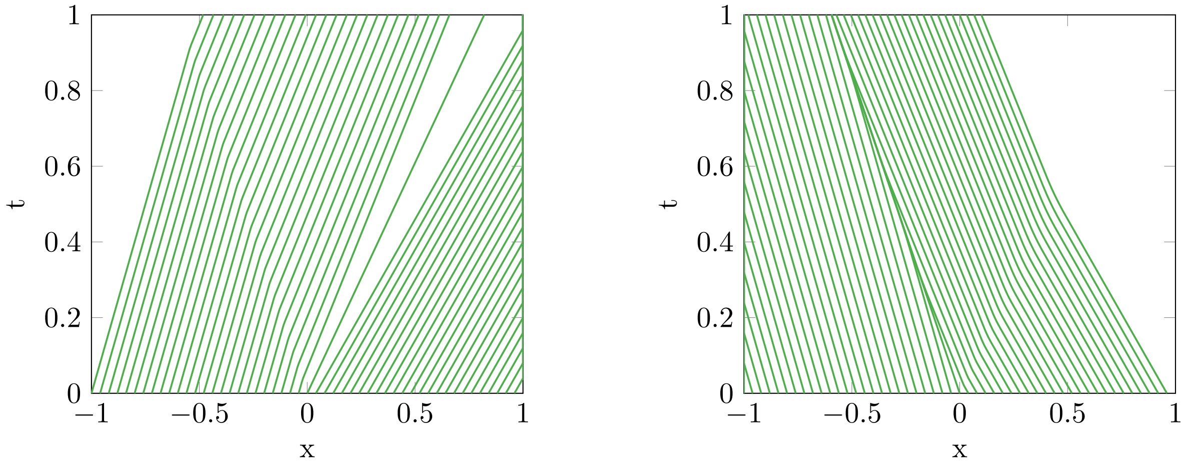
Figure 7.4 shows a visualization of the generalized characteristics obtained from solving the discontinuous ODEs

$$\partial_t \xi_1(t) = \lambda_1(u(t, \xi_1(t)), v(t, \xi_1(t)))$$

and

$$\partial_t \xi_2(t) = \lambda_2(u(t, \xi_2(t)), v(t, \xi_2(t)))$$

in the sense of Filippov (see Section 1.4.4) with initial value conditions $\xi_1(0)$ and $\xi_2(0)$ on an equidistant grid along the visible domain. The plots are generated by an explicit Euler discretization of the discontinuous ODEs.



(a) $\partial_t \xi_1(t) = \lambda_1(u(t, \xi_1(t)), v(t, \xi_1(t)))$

(b) $\partial_t \xi_1(t) = \lambda_2(u(t, \xi_2(t)), v(t, \xi_2(t)))$

Figure 7.4: Characteristics of the strictly hyperbolic nonlinear system

The generalized characteristics belonging to the eigenvalue λ_2 visually show a compression that satisfies the entropy condition along a shock location curve going to the left over time starting from the origin. The shock discontinuity in both u and v at $t = 1/2$ is at the identical location as the curve where the characteristics cross, i.e. the shock location.

The generalized characteristics belonging to λ_1 visually show a rarefaction fan going to the right over time starting from the origin. The corresponding rarefaction wave is also visible in both u and v at $t = 1/2$ in the same location and with the same width.

According to the characteristics for the example we observe a shock traveling to the left and a rarefaction wave traveling to the right. Notably, the characteristics are associated with the eigenvalues of the flux Jacobian and not with an individual variable. So the shock discontinuity due to the compression in the characteristic corresponding to one eigenvalue forms in both variables u and v.

For nonlinear flux functions a shock discontinuity forms where the characteristic lines corresponding to an eigenvalue cross. The associated strict entropy condition for a shock discontinuity along the curve ξ (c.f. the entropy condition in Section 3.1.3) is

$$\lambda_i(\boldsymbol{u}(t,\xi(t)-)) > \partial_t\xi(t) > \lambda_i(\boldsymbol{u}(t,\xi(t)+))$$

where λ_i is the eigenvalue associated with the shock formation and $\boldsymbol{u}(t,\xi(t)) \in \mathbb{R}^N$ is the solution vector containing all variables. The nonstrict version of that entropy condition for a shock to be admissable is called the Lax condition in Bressan [Bre99, Section 1.5] due to its initial introduction in Lax [Lax57, Section 7].

The analogous shock location algorithm to the one introduced for scalar conservation laws in Section 1.4.4 and analyzed in Section 6.2.1 is the explicit Euler discretization of the Filippov differential inclusion for the discontinuous characteristic speed

$$\partial_t\xi(t) \in \big[\lambda_i(\boldsymbol{u}(t,\xi(t)-)), \lambda_i(\boldsymbol{u}(t,\xi(t)+))\big]$$

where again λ_i is the eigenvalue associated with the shock formation and $\boldsymbol{u}(t,\xi(t)) \in \mathbb{R}^N$ is the solution vector.

In the general case where we do not have the symbolic eigendecomposition available it is possible to perform a shock location approximation based on the numerical eigendecomposition of the flux Jacobian performed at every time step. In this chapter we do not consider the numerical eigendecompostion approach but instead make use of the availability of the symbolic eigenvalues as a function of the weak solution.

In the case of the isentropic gas equations we use the following discrete time step for the numerical shock location

$$\begin{aligned}\Xi^{j+1} &:= \Xi^j + \Delta t_j \lambda_2(U^j(\Xi^j), V^j(\Xi^j)) && (7.2.2)\\ &= \Xi^j + \Delta t_j\Big(-\sqrt{-P'(V^j(\Xi^j))}\Big) && (7.2.3)\end{aligned}$$

where $U^j(\cdot), V^j(\cdot)$ are linear interpolations of the approximation of the weak solution obtained by the Lax-Friedrichs scheme. Because we know that the shock curve starts at the origin we have $\Xi^0 = 0$.

For the time step Δt_j we pick the same time step as for the numerical integrator. In Section 3.2.4 we saw that the CFL condition for a scalar conservation law depends on the maximum characteristic speed of the weak solution at a given time t_j. The same type of condition can be formulated based on the characteristic speeds for a system given by the eigenvalues. In the case of a 2×2 system like the isentropic gas equations we use the discrete CFL time step

$$\Delta t_j = \frac{1}{\max\Big\{ \max_i |\lambda_1(U_i^j, V_i^j)|, \max_i |\lambda_2(U_i^j, V_i^j)| \Big\}} \Delta x$$

with CFL number set to one.

7.3 Shock Sensitivity for the Isentropic Gas Equations

We now show how the shock sensitivity with respect to a parameter of the initial data can be numerically approximated in a hyperbolic system of conservation laws based on the method developed for scalar conservation laws in the previous chapter and the shock location algorithm of the previous section.

For a hyperbolic $N \times N$ system we consider the parametric initial data

$$u_i(0,x) = (u_i)_0(x,p)$$

for all $i \in \{1, \dots, n\}$ and with the parameter $p \in \mathbb{R}^d$. We assume that the weak solution has a shock discontinuity along a curve ξ like the example in the previous section. Then the analytic tangent sensitivity of ξ is

$$\dot{\xi}(t) = \mathrm{d}_p \xi(t,p)\dot{p}$$

for all $t \in [0,T]$.

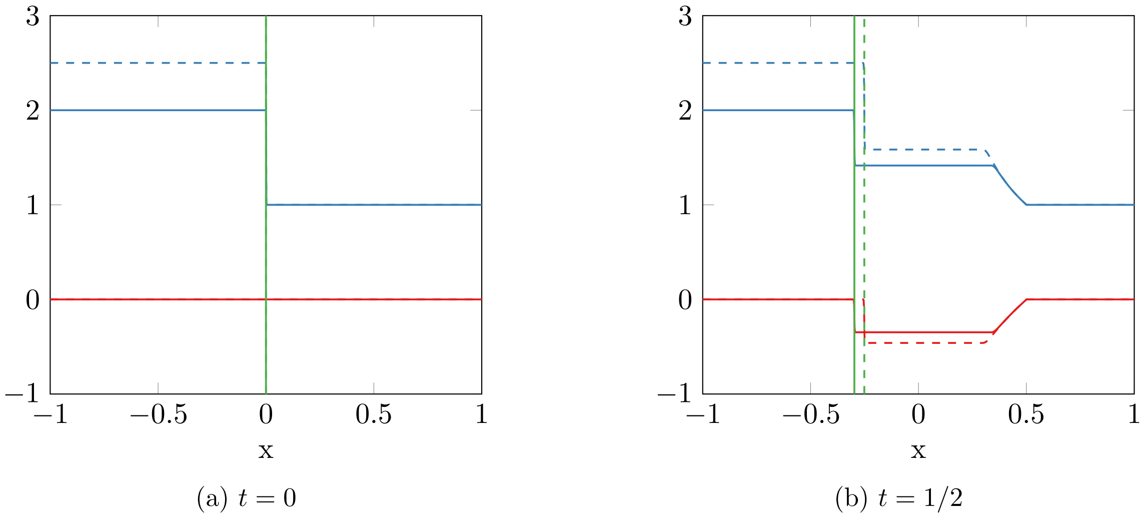
In this section we show how to obtain a numerical approximation $\dot{\Xi}$ of $\dot{\xi}$ such that it converges as the discretization accuracy increases, i.e. such that $\|\dot{\Xi} - \dot{\xi}\| \to 0$ as $\Delta x \to 0$.

Example 43 (Parametric Riemann problem)**.** We consider a parametric version of the Riemann initial value problem for the isentropic gas equations from Example 42. The parametric initial data is given by

$$u_0(x) = 0, \quad v_0(x,p) = 1 + (1+p)H(-x)\ .$$

The parameter p determines the height of the volume jump at the origin. We perform a sensitivity analysis in the neighborhood of $p = p_0 = 0$, i.e. a perturbation of the original problem simulated in Example 42.

Figure 7.5 shows a numerical approximation of the weak solution for u in red, v in blue and of the shock location ξ in green at $t = 0$ on the left and $t = 1/2$ on the right. Each numerical approximation is plotted twice, once with a solid line for the original initial data with $p = p_0 = 0$, and once with a dashed line for the perturbed initial data with $p = p_0 + \dot{p} = 1/2$. The numerical approximation of the weak solution is via the Lax-Friedrichs scheme as described in Equations (7.1.1)-(7.1.2) using the CFL time step as described at the end of the previous section. The numerical approximation of the shock location is via the explicit Euler discretization described in the previous section.



(a) $t = 0$

(b) $t = 1/2$

Figure 7.5: Weak solution and shock location perturbation, $U(p_0)$ (solid red), $U(p_0 + \dot{p})$ (dashed red), $V(p_0)$ (solid blue), $V(p_0 + \dot{p})$ (dashed blue), $\Xi(p_0)$ (solid green), $\Xi(p_0 + \dot{p})$ (dashed green)

In the left figure we see that the shock location initial value is zero in both the perturbed and unperturbed case and, hence, the dashed and solid green lines coincide. The perturbation of the initial data of v is visible in the dashed blue line by its higher value to the left of the origin. In the figure on the right we see that the perturbed initial data leads to different dynamics in the shock location and for the perturbed problem the shock has travelled to the left less than for the unperturbed problem. We also see that the solid green line accurately approximates the location where the shock visibly occurs in the numerical solutions of u, v in solid red and blue and the dashed green line accurately approximates the same for the dashed red and blue graphs.

In the following we want to demonstrate via computational experiments that (i) the naive discrete tangent of the explicit Euler discretization used to compute Ξ does not converge to $\dot{\xi}$ and that (ii) a shock sensitivity algorithm based on the differentiation of the Rankine-Hugoniot condition can instead be used to get an approximation $\dot{\Xi}$ that converges to $\dot{\xi}$ as $\Delta x \to 0$. Since we do not have an analytic solution for the shock location ξ and the shock sensitivity $\dot{\xi}$ we validate the proposed method via the empirical (non)convergence of the first-order Taylor expansion error term.

By implicit differentiation of the Rankine-Hugoniot condition introduced in the next section the shock location ξ is continuously differentiable with respect to the parameter of the initial data (c.f. Proposition 23). According to the first-order Taylor expansion we have

$$\xi(t, p_0 + \dot{p}) - \xi(t, p_0) - \dot{\xi}(t, p_0) = O(\|\dot{p}\|^2) \tag{7.3.1}$$

as $\dot{p} \to 0$.

We want to check if the numerical approximation errors of the shock location Ξ and shock sensitivity $\dot{\Xi}$ converge to zero as $\Delta x \to 0$. For now, let us assume that they do converge at an $O(\Delta x^\alpha)$ rate for some $\alpha \in (0, 1)$ as $\Delta x \to 0$ and see what that implies about the error term of the first-order Taylor expansion as approximated based on the numerical solutions.

Proposition 46. Let Ξ and $\dot{\Xi}$ converge at rate $O(\Delta x^\alpha)$ for some $\alpha \in (0, 1)$, i.e.

$$\xi(t, p) - \Xi(t, p) = O(\Delta x^\alpha), \quad \dot{\xi}(t, p) - \dot{\Xi}(t, p) = O(\Delta x^\alpha)$$

as $\Delta x \to 0$ for all p in a neighborhood of p_0 and let $\Delta x = O(\|\dot{p}\|^{2/\alpha})$. Then we have

$$\Xi(t, p_0 + \dot{p}) - \Xi(t, p_0) - \dot{\Xi}(t, p_0) = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$.

Proof. With $\Delta x = O(\|\dot{p}\|^{2/\alpha})$ we have $\Delta x^\alpha = O(\|\dot{p}\|^2)$ as $\dot{p} \to 0$.

By the convergence of Ξ in a neighborhood of p_0 combined with the above we have

$$\Xi(t, p_0 + \dot{p}) - \Xi(t, p_0) - (\xi(t, p_0 + \dot{p}) - \xi(t, p_0)) = O(\Delta x^\alpha) = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$. By the convergence of $\dot{\Xi}$ we also directly have

$$\dot{\xi}(t, p) - \dot{\Xi}(t, p) = O(\Delta x^\alpha) = O(\|\dot{p}\|^2)$$

as $\dot{p} \to 0$.

Adding the two $O(\|\dot{p}\|^2)$ terms to the Taylor expansion error term of the analytic shock location ξ on the left-hand side of Equation (7.3.1) we have

$$\begin{aligned}
&\Xi(t, p_0 + \dot{p}) - \Xi(t, p_0) - \dot{\Xi}(t, p_0) \\
&\quad = \xi(t, p_0 + \dot{p}) - \xi(t, p_0) - \dot{\xi}(t, p) \\
&\qquad + \Big(\Xi(t, p_0 + \dot{p}) - \Xi(t, p_0) - (\xi(t, p_0 + \dot{p}) - \xi(t, p_0))\Big) \\
&\qquad - \Big(\dot{\Xi}(t, p_0) - \dot{\xi}(t, p_0)\Big) \\
&\quad = O(\|\dot{p}\|^2) + O(\|\dot{p}\|^2) + O(\|\dot{p}\|^2) = O(\|\dot{p}\|^2)
\end{aligned}$$

as $\dot{p} \to 0$. This concludes the proof. □

If the numerical shock location Ξ and the numerical shock sensitivities $\dot{\Xi}$ converge then we should observe a quadratic error for the first-order Taylor expansion approximation. If we do not empirically observe that quadratic error then we do not have a method converging at a $O(\Delta x^\alpha)$ rate. Technically, we are mixing nonasymptotic and asymptotic convergence results here but for the case of our example the above argument holds in a nonasymptotic sense and we observe a quadratic error empirically for the correct method presented in the next section.

7.3.1 Naive Discrete Tangent Shock Sensitivity

We now present the result of a naive discrete tangent sensitivity analysis of the shock location discretization introduced in the previous section. Just like we saw in Section 1.4.5 for a scalar conservation law the naive discrete tangent does not converge.

Example 44 (Naive discrete tangent)**.** We consider again the parametric version of the Riemann initial value problem for the isentropic gas equations from Example 42.

The discrete tangent $\dot{\Xi}$ resulting from the differentiation of the explicit Euler time step Ξ^{j+1} for the shock location as defined in the previous section is

$$\dot{\Xi}^{j+1} := \dot{\Xi}^j + \Delta t_j \Big(\mathrm{d}_{\Xi^j} \lambda_2(U^j(\Xi^j), V^j(\Xi^j)) \cdot \dot{\Xi}^j + \mathrm{d}_{U^j} \lambda_2(U^j(\Xi^j), V^j(\Xi^j)) \cdot \dot{U}^j$$

$$
\begin{aligned}
&\quad + \mathrm{d}_{V^j}\lambda_2(U^j(\Xi^j), V^j(\Xi^j)) \cdot \dot{V}^j\Big) \\
&= \dot{\Xi}^j + \Delta t_j \left(\mathrm{d}_{\Xi^j}\Big(-\sqrt{-P'(V^j(\Xi^j))}\Big) \cdot \dot{\Xi}^j + \mathrm{d}_{V^j}\Big(-\sqrt{-P'(V^j(\Xi^j))}\Big) \cdot \dot{V}^j \right) \\
&= \dot{\Xi}^j + \Delta t_j \left(\frac{1}{\sqrt{-P'(V^j(\Xi^j))}} P''(V^j(\Xi^j))\big(\dot{V}^j(\Xi^j) + \partial_x V^j(\Xi^j)\dot{\Xi}^j\big) \right)
\end{aligned}
$$

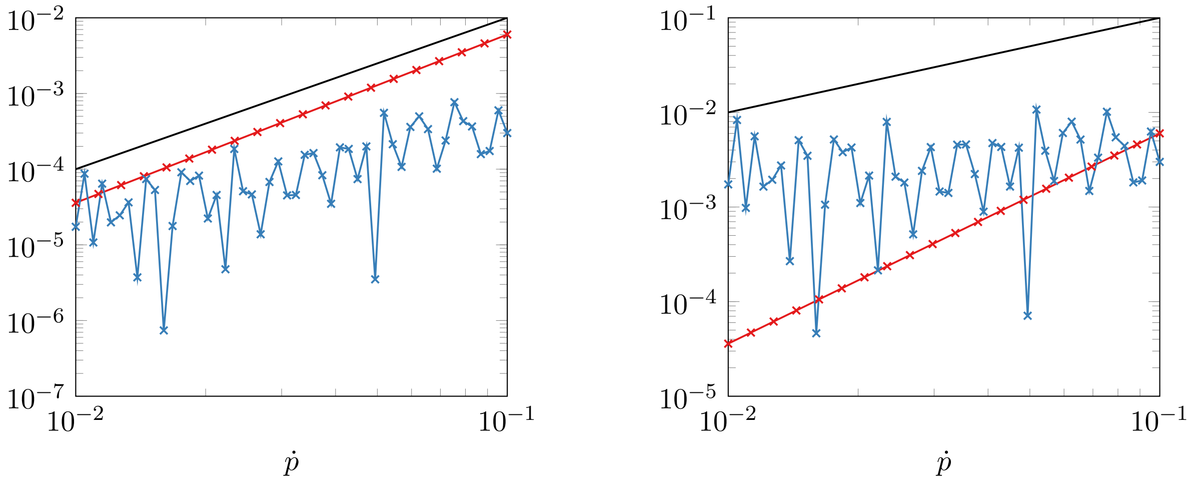
where $U^j(\cdot), V^j(\cdot)$ are linear interpolations of the numerical approximation of the weak solution, $\dot{U}^j, \dot{V}^j$ are linear interpolations of the discrete tangent of the weak solution, and Ξ^j is the shock location obtained by numerical approximation according to the method from the previous section.

Figure 7.6a shows the first-order Taylor expansion absolute error $|\Xi(p_0 + \dot{p}) - \Xi(p) - \dot{\Xi}(p_0)|$ at $p_0 = 0$ and with $\Delta x = O(\|\dot{p}\|^2)$ in blue for geometrically decaying values of $\dot{p}$. The figure also shows $\|\dot{p}\|^2$ in black and Δx in red. The blue error graph is oscillating as $\dot{p} \to 0$ and does not show clear convergence even though as $\dot{p}$ decreases we see a tendencies of the error going down. In order to have convergence for $\Delta x = O(\|\dot{p}\|^{2/\alpha})$ with $\alpha < 1$ the error would have to go down at a rate that is faster than linear.

Figure 7.6b analogously shows the divided difference absolute error

$$
\left| \frac{\Xi(p_0 + \dot{p}) - \Xi(p_0)}{\dot{p}} - \mathrm{d}_p\Xi(p_0) \right|
$$

where $\mathrm{d}_p\Xi(p_0) \equiv \dot{\Xi}/\dot{p}$ in blue. The blue error graph is oscillating as $\dot{p} \to 0$ and does not show clear convergence and as $\dot{p}$ decreases we do not see any tendency of the error going down. Therefore, there probably is no clear convergence even with $\Delta x = O(\|\dot{p}\|^{2/\alpha})$ for some $\alpha \in (0, 1)$ in the asymptote either.

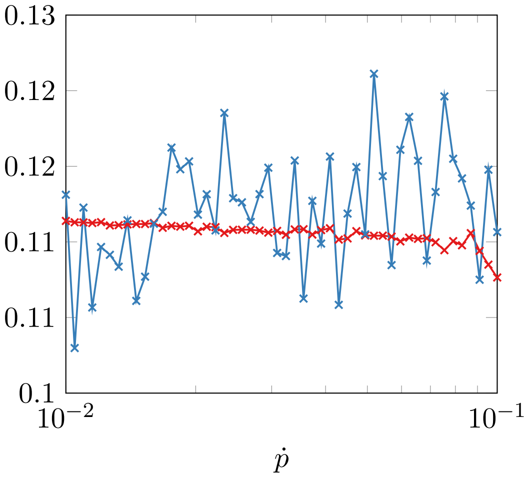


(a) $|\Xi(p_0 + \dot{p}) - \Xi(p) - \dot{\Xi}(p_0)|$ (blue), $\|\dot{p}\|^2$ (black), Δx (red)

(b) $\left|\frac{\Xi(p_0+\dot{p})-\Xi(p_0)}{\dot{p}} - \mathrm{d}_p\Xi(p_0)\right|$ (blue), $\dot{p}$ (black), Δx (red)

Figure 7.6: Naive discrete tangent oscillation and nonconvergence

Figure 7.7 shows the value of the finite divided difference in red compared to the value of the discrete tangent based derivative in blue. The red graph of the finite difference converges to the correct derivative while again the blue graph with the discrete tangent based derivative strongly oscillates as $\dot{p}$ changes and does not seem to converge to the correct value at all.



(a) $\frac{\Xi(p_0+\dot{p})-\Xi(p_0)}{\dot{p}}$ (red), $\mathrm{d}_p\Xi(p_0)$ (blue)

Figure 7.7: Naive discrete tangent oscillation and nonconvergence

The reasons behind the nonconvergence of the naive discrete tangent are the same as in the scalar setting (c.f. Figure 1.14). The shock location itself oscillates around the correct shock location due to the sliding discontinuity at the shock and the entropy condition pushing the discrete shock location trajectory towards but sometimes across the true shock location. The numerical solution and its discrete tangent are not stable to evaluate or differentiate in a region around the shock location due to the numerical viscosity and potential oscillations in the numerical solution. As a result a discrete tangent evaluated based on the discrete tangent in the shock region is oscillatory and nonconvergent as well.

7.3.2 Rankine-Hugoniot Condition Shock Sensitivity

We now show how to generalize the numerical shock sensitivity based on the Rankine-Hugoniot condition as introduced for scalar conservation laws in Section 1.4.5 and analyzed in Section 6.2.2 to the case of a hyperbolic system of conservation laws.

First, we define the averaged Jacobian and based on that the generalization of the Rankine-Hugoniot condition (see Section 3.1.4) to systems of conservation laws.

Definition 44. The *averaged Jacobian* matrix is defined as

$$\mathcal{A}(\boldsymbol{u}^+, \boldsymbol{u}^-) \equiv \int_0^1 A(\theta \boldsymbol{u}^+ + (1-\theta)\boldsymbol{u}^-)\mathrm{d}\theta$$

where $A(\boldsymbol{u})$ is the Jacobian of F at $\boldsymbol{u}$.

The averaged Jacobian at a shock location is obtained by integrating across the shock location in the sense that we consider the average of the Jacobian evaluated over all values lying between the weak solution at the left and right limit of the shock. For example for a scalar conservation law with the flux function f we can analyze the averaged Jacobian at the shock location by change of variable. By the substitution

$$v(\theta) = \theta u^+ + (1-\theta)u^-$$

with $v'(\theta) = u^+ - u^-$ we have

$$\begin{aligned} f'(u^+, u^-)(u^+ - u^-) &= \int_0^1 f'(\theta u^+ + (1-\theta)u^-)(u^+ - u^-)\mathrm{d}\theta \\ &= \int_{u^-}^{u^+} f'(v)\mathrm{d}v \\ &= f(u^+) - f(u^-) \; . \end{aligned}$$

Dividing both sides by $u^+ - u^-$ we have the averaged Jacobian as the finite difference across the shock

$$f'(u^+, u^-) = \frac{f(u^+) - f(u^-)}{u^+ - u^-}$$

as a generalized derivative concept for discontinuous functions (c.f. Equation (1.4.3) and the derivation of the adjoint at the shock in Section 5.4.2).

Based on the averaged Jacobian we have the Rankine-Hugoniot condition for the dynamics of a shock curve for a system of conservation laws.

Theorem 47 (Rankine-Hugoniot Condition)**.** Let $\boldsymbol{u} \in \mathbb{R}^N$ be the weak solution of a hyperbolic system of conservation laws such that u has a shock discontinuity along the curve ξ and is smooth on either side of ξ. Then u satisfies the Rankine-Hugoniot condition

$$\begin{aligned} \partial_t \xi(t)(\boldsymbol{u}^+ - \boldsymbol{u}^-) &= F(\boldsymbol{u}^+) - F(\boldsymbol{u}^-) \\ &= \mathcal{A}(\boldsymbol{u}^+, \boldsymbol{u}^-)(\boldsymbol{u}^+ - \boldsymbol{u}^-) \end{aligned}$$

where $\boldsymbol{u}^+ = \boldsymbol{u}(t, \xi(t)+), \boldsymbol{u}^- = \boldsymbol{u}(t, \xi(t)-)$ across the curve of discontinuity. The vector $\boldsymbol{u}^+ - \boldsymbol{u}^-$ is an eigenvector of $\mathcal{A}(\boldsymbol{u}^+, \boldsymbol{u}^-)$ and the shock speed $\partial_t \xi(t)$ is the corresponding eigenvalue.

Proof. The proof is by divergence theorem and analogous to that of Theorem 8. See for example Bressan [Bre09, Section 2.2]. □

Just like determining the characteristic speeds required us to perform an eigendecomposition of the flux Jacobian determining the shock speed by the Rankine-Hugoniot condition requires us to perform an eigendecomposition of the averaged Jacobian. In the general case where we do not have a symbolic eigendecomposition available it is possible to perform a numerical eigendecomposition at every time step. For the isentropic gas equations and many other hyperbolic systems it is, however, possible to perform the eigendecomposition symbolically as a function of the limiting weak solution values around the shock and obtain the Rankine-Hugoniot condition directly.

Example 45 (Isentropic gas equations)**.** We now consider the eigenvalues of the averaged Jacobian for the isentropic gas equations to determine the shock speed according to the Rankine-Hugoniot condition.

By the derivation of the flux Jacobian for the isentropic gas equations in Example 40 the averaged Jacobian around a shock location, with right and left limits to the shock in the weak solution being $(u^+, v^+), (u^-, v^-)$, is

$$\begin{aligned}\mathcal{A}\big((u^+, v^+), (u^-, v^-)\big) &= \int_0^1 \begin{pmatrix} 0 & P'(\theta v^+ + (1-\theta)v^-) \\ -1 & 0 \end{pmatrix} \mathrm{d}\theta \\ &= \begin{pmatrix} 0 & \int_0^1 P'(\theta v^+ + (1-\theta)v^-)\mathrm{d}\theta \\ -1 & 0 \end{pmatrix} .\end{aligned}$$

We can derive the integral for the $\mathcal{A}_{1,2}$ element by the same change of variable method as we used for the scalar flux function f earlier. By chain rule we have

$$\mathrm{d}_\theta \frac{P(\theta v^+ + (1-\theta v^-))}{v^+ - v^-} = P'(\theta v^+ + (1-\theta v^-))$$

and by the fundamental theorem of calculus we, thus, have

$$\mathcal{A}_{1,2} = \int_0^1 \mathrm{d}_\theta \frac{P(\theta v^+ + (1-\theta v^-))}{v^+ - v^-} \mathrm{d}\theta = \frac{P(v^+) - P(v^-)}{v^+ - v^-} .$$

We obtain the eigenvalues of $\mathcal{A}$ as the roots of the characteristic polynomial

$$0 = \lambda^2 + \frac{P(v^+) - P(v^-)}{v^+ - v^-} .$$

We have the two eigenvalues

$$\lambda_1(v^+, v^-) = \sqrt{-\frac{P(v^+) - P(v^-)}{v^+ - v^-}}, \qquad \lambda_2(v^+, v^-) = -\sqrt{-\frac{P(v^+) - P(v^-)}{v^+ - v^-}}$$

as the solution of the quadratic equation (c.f. Equation (7.2.1)).

For the parametric version of the Riemann initial value problem for the isentropic gas equations we know that the shock develops in the characteristics corresponding to λ_2 and we know a priori that the Rankine-Hugoniot condition holds around the simulated shock location Ξ. Due to the entropy condition the shock speed $\lambda_2(v^+, v^-)$ lies between the left and right limiting characteristic speeds and is also real valued.

Based on the shock speed obtained as an eigenvalue of the averaged Jacobian we can now apply the scalar method for shock sensitivities from Section 6.2.2 and get a stable numerical approximation of the shock sensitivities $\dot{\Xi}$. We proceed with the method described in Section 6.5 to produce the shock sensitivity algorithm by forward mode AD.

For the example of the isentropic gas equations a discretization of the Rankine-Hugoniot condition with $\lambda_2(v^+, v^-)$ that goes far enough outside of the shock region when approximating the limits v^+ and v^- is

$$\Xi^{j+1} := \Xi^j + \Delta t \left(-\sqrt{-\frac{P(V^+) - P(V^-)}{V^+ - V^-}} \right) \tag{7.3.2}$$

with

$$V^+ = V^j(\Xi^j + C\Delta x^\alpha)$$

$$V^- = V^j(\Xi^j - C\Delta x^\alpha)$$

where $C > 0$, $\alpha \in (0,1)$ and $V^j(\cdot)$ is the linear interpolation of the weak solution approximation obtained via the Lax-Friedrichs discretization scheme. The above explicit Euler discretization of the Rankine-Hugoniot condition is stably differentiable with respect to the parameter of the initial data.

Like in the case of a scalar conservation law using the discretization of the Rankine-Hugoniot condition yields a larger discretization error than the direct explicit Euler discretization of the discontinuous ODE giving the characteristic speeds (see Section 1.4.5). Therefore we combine the shock location Ξ according to Equation (7.2.2) with the tangent of Equation (7.3.2) to get a convergent discrete shock sensitivity $\dot{\Xi}$.

The discrete tangent of Equation (7.3.2) according to the AD method of Section 6.5 is

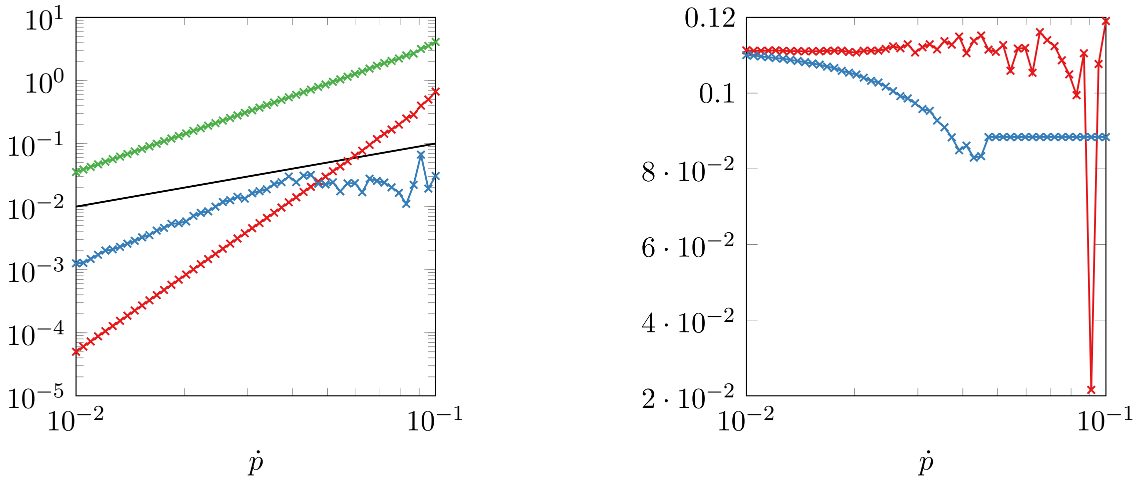
$$\begin{aligned}\dot{\Xi}^{j+1} := \dot{\Xi}^j + \Delta t \frac{1}{2}\left(-\frac{P(V^+) - P(V^-)}{V^+ - V^-}\right)^{-\frac{1}{2}} \\ \cdot \left(\frac{P'(V^+)\dot{V}^+ - P'(V^-)\dot{V}^-}{V^+ - V^-} - \frac{P(V^+) - P(V^-)}{(V^+ - V^-)^2}(\dot{V}^+ - \dot{V}^-)\right)\end{aligned}$$

with

$$\begin{aligned}\dot{V}^+ &= \dot{V}^j(\Xi^j + C\Delta x^\alpha) + \partial_x V^j(\Xi^j + C\Delta x^\alpha)\dot{\Xi}^j \\ \dot{V}^- &= \dot{V}^j(\Xi^j - C\Delta x^\alpha) + \partial_x V^j(\Xi^j - C\Delta x^\alpha)\dot{\Xi}^j\end{aligned}$$

where $V^j(\cdot)$ is the linear interpolation of the numerical solution and $\dot{V}^j(\cdot)$ is the linear interpolation of the discrete tangent of the numerical solution.

We now analyze the computational results based on the Rankine-Hugoniot method via the empirical convergence of the first-order Taylor expansion error term.

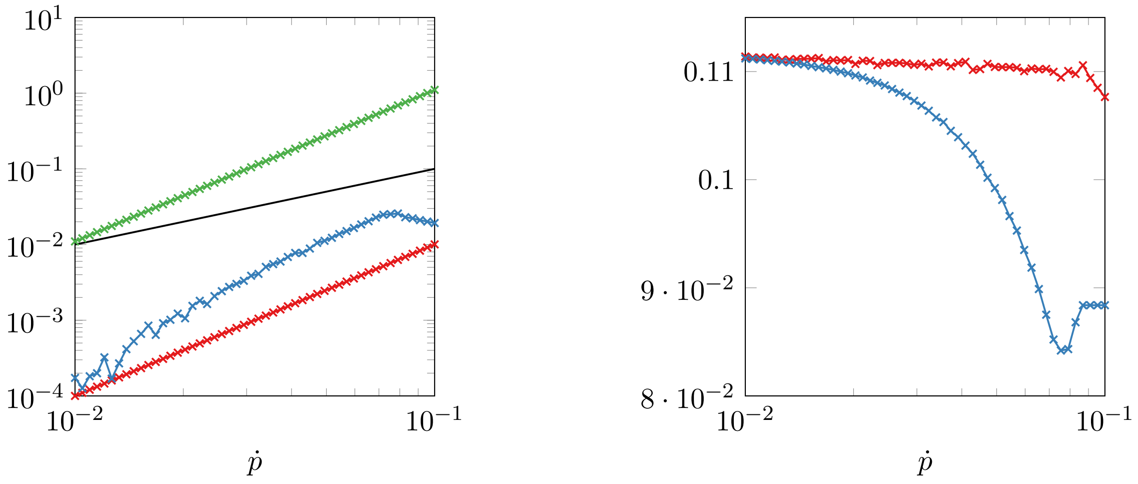


(a) $\left|\frac{\Xi(p_0+\dot{p})-\Xi(p_0)}{\dot{p}} - \mathrm{d}_p\Xi(p_0)\right|$ (blue), $\dot{p}$ (black), Δx (red), $C\Delta x^\alpha$ (green)

(b) $\frac{\Xi(p_0+\dot{p})-\Xi(p_0)}{\dot{p}}$ (red), $\mathrm{d}_p\Xi(p_0)$ (blue)

Figure 7.8: Rankine-Hugoniot shock sensitivity, $\alpha = 1/2, C = 5$

Figure 7.8 and Figure 7.9 show the empirical convergence of the discrete tangent based approximation of the derivative with respect to p for different values of α and C. The left figures are analogous to Figure 7.6b of the previous section showing the absolute error between the finite difference and the discrete tangent based derivative approximation. The left figures additionally show the shock region width (over-)estimations $C\Delta x^\alpha$ in green. The right figures are analogous to Figure 7.7 showing both the discrete tangent based derivative and the finite difference as values. The discretization accuracy used in the figures here is lower relative to $\dot{p}$ compared to the figures of the previous section in order to make the involved simulation for $\alpha = 1/2$ feasible in terms of computational time.



(a) $\left|\frac{\Xi(p_0+\dot{p})-\Xi(p_0)}{\dot{p}} - \mathrm{d}_p\Xi(p_0)\right|$ (blue), $\dot{p}$ (black), Δx (red), $C\Delta x^\alpha$ (green)

(b) $\frac{\Xi(p_0+\dot{p})-\Xi(p_0)}{\dot{p}}$ (red), $\mathrm{d}_p\Xi(p_0)$ (blue)

Figure 7.9: Rankine-Hugoniot shock sensitivity, $\alpha = 1, C = 110$

For values of Δx that are too high the distance we go outside of the shock region by $\Xi \pm C\Delta x^\alpha$ is too large to even be inside of the simulated space domain. That causes us to observe an irregular behavior of the discrete tangent for the higher values of $\dot{p}$ that should be ignored.

For the discrete tangent to be convergent in the sense of the Taylor approximation we need the error against the finite difference to decay at a $O(\|\dot{p}\|)$ rate as $\dot{p} \to 0$. For both $\alpha = 1/2$ and $\alpha = 1$ we observe a faster than linear convergence of the error. Furthermore, in the figures on the right we also observe that for small enough values of Δx the discrete tangent smoothly approaches the finite difference in contrast to the naive discrete tangent we analyzed in the previous section.

In conclusion, the Rankine-Hugoniot condition based discrete shock sensitivitiy analysis presented in this chapter converges for the shock sensitivities in the strictly hyperbolic system of isentropic gas equations. The presented method directly applies to other systems where we can perform an eigendecomposition of the averaged flux Jacobian symbolically.

An interesting direction for further work is to analyze how the method presented in this chapter can be efficiently implemented using a numerical eigendecomposition. We have performed preliminary computational experiments in that direction that give promising results.

Chapter 8

First-order Splitting Shock Sensitivities for the 2D Burgers Equation

In this chapter we show how the numerical methods for approximating shock sensitivities in 1D scalar conservation laws introduced in Chapter 6 can be used for scalar conservation laws with more than one space dimension. Specifically, we consider 2D scalar conservation laws where the numerical integrator for the weak solution is a first-order splitting scheme. The presented method extends to 3D and higher but we only discuss the 2D setting to avoid unnecessary complexity.

Example 46 (2D Burgers equation)**.** As a driving example throughout this chapter we consider a variation of the 2D Burgers equation

$$\partial_t u(t,x,y) + \partial_x f(u(t,x,y)) + \partial_y g(u(t,x,y)) = 0 \tag{8.0.1}$$

where $f(u) = \frac{1}{2}u^2$ is the Burgers flux and $g(u) = a\frac{1}{2}u^2$ is a scaled Burgers flux with $a > 0$. The conserved quantity $u(t,x,y) \in \mathbb{R}$ is a scalar field over two space dimensions.

We introduce the background on first-order operator splitting of the time stepping operator with respect to multiple space dimensions that is necessary to understand the approximation error of discrete first-order splitting numerical integrators. For a 2D scalar conservation law it is typical for a shock discontinuity to develop along a one dimensional manifold [Gli+85]. Based on the characteristics of a 2D scalar conservation law we show how to track a point that lies on a 1D shock manifold and derive a Rankine-Hugoniot condition that holds at such a point on a shock manifold. The discontinuous characteristic speeds and the Rankine-Hugoniot condition form the basis for two variants of numerical shock location and numerical shock sensitivity methods that differ in how they are combined with the first-order splitting method. Finally, we validate and compare both methods based on empirical convergence results obtained for the 2D Burgers equations with parametric initial data that produces a linear shock manifold that lies in space at different orientations. We show that the presented first-order splitting of the shock location and sensitivity methods can also be combined with other numerical integrators that are not necessarily first-order splitting schemes.

8.1 First-order Splitting

We give the necessary background on first-order splitting based numerical integrators for scalar conservation laws. Operator splitting methods are common in the literature on numerical simulation of PDEs (see e.g. [LeV02, Chapter 17], [MS16], [MQ02]). The first-order splitting analysis in this chapter is mostly based on LeVeque [LeV92, Chapter 18.2].

8.1.1 Motivation and Definition

We now give the definition of the first-order splitting time step as it is used throughout this chapter.

To motivate first-order operator splitting of the time stepping operator of a conservation law with respect to the different space dimensions we first recall that the linear scalar conservation law, i.e. the 1D scalar advection equation

$$\partial_t u(t,x) + a\partial_x u(t,x) = 0$$

has the solution

$$u(t,x) = u_0(x - at)$$

where u_0 is the initial data. Similarly, the analogous 2D linear scalar conservation law, i.e. the 2D scalar advection equation

$$\partial_t u(t,x,y) + a\partial_x u(t,x,y) + b\partial_y u(t,x,y) = 0$$

has the solution

$$u(t,x,y) = u_0(x - at, y - bt)$$

where again u_0 is the initial data [LeV92, Chapter 18.2]. The solution of the 2D scalar advection equation is a transport of the initial data by factors a, b of time where a, b determine the characteristic speed.

The idea behind the first-order splitting of the time stepping operator is that the two space directions can be transported separately and we end up at the same solution, at least for the 2D scalar advection equation.

Definition 45. A *first-order splitting time step* for Equation (8.0.1) with general scalar flux functions f, g from t_i to t_{i+1} is defined as sequentially solving two separate one dimensional conservation law problems. First, we solve

$$\partial_t u^*(t,x,y) + \partial_x f(u^*(t,x,y)) = 0$$

with initial data

$$u^*(t_i,x,y) = u(t_i,x,y)$$

for all $x, y \in \Omega, t \in [t_i, t_{i+1}]$. Second, we solve

$$\partial_t u^{**}(t,x,y) + \partial_y g(u^{**}(t,x,y)) = 0$$

with initial data

$$u^{**}(t_i,x,y) = u^*(t_{i+1},x,y)$$

for all $x, y \in \Omega, t \in [t_i, t_{i+1}]$. The solution of the second substep

$$u(t_{i+1}, x, y) \equiv u^{**}(t_{i+1}, x, y)$$

is the result of the overall first-order splitting time step.

The first-order splitting time step of the 2D scalar advection equation is exact. For the first substep of the first-order splitting we consider the equation

$$\partial_t u^*(t, x, y) + a\partial_x u^*(t, x, y) = 0$$

with initial data u_0. The first substep has the known solution of the 1D scalar advection equation

$$u^*(t, x, y) = u_0(x - at, y) \ .$$

For the second substep of the first-order splitting we consider the equation

$$\partial_t u^{**}(t, x, y) + b\partial_y u^{**}(t, x, y) = 0$$

with initial data $u^*(t, \cdot, \cdot)$. Again, the known solution of the 1D scalar advection equation for second substep is

$$u^{**}(t, x, y) = u^*(t, x, y - bt) = u_0(x - at, y - bt) \ .$$

The solution of the second substep is the solution of the first-order splitting time step and is identical to the exact solution of the 2D scalar advection equation.

8.1.2 Linear Error Analysis

We now perform an error analysis for a general first-order splitting time step of linear PDEs. The insights apply to linear conservation laws and are also helpful in understanding the approximation error of first-order splitting for nonlinear conservation laws.

Without formalizing the notion of differential operators we recall that the differentiation of a sufficiently smooth weak solution u with respect to x and y can be considered as the application of the linear operators ∂_x and ∂_y.

In the following we consider the PDE

$$\partial_t u - Au - Bu = 0 \tag{8.1.1}$$

with A, B general bounded linear operators that can be thought to include the composition with differential operators. E.g. we can have $A = a\partial_x$ and $B = b\partial_y$ to get the 2D scalar advection equation. The PDE in Equation (8.1.1), therefore, also generalizes linear conservation laws.

We also note that A and B map the whole scalar solution field $u(t, \cdot)$ to another scalar solution field of the same dimension. A potentially useful intuition is to think of A and B as the limit of matrices applied to a spacial discretization of u as the discretization grid distance goes to zero.

Due to linearity Equation (8.1.1) is equivalent to

$$\partial_t u = (A + B)u \tag{8.1.2}$$

that again due to linearity has the solution

$$u(t_{i+1},\cdot) = \exp\big(\Delta t(A+B)\big)u(t_i,\cdot) \tag{8.1.3}$$

with $\Delta t = t_{i+1} - t_i$ for a time step from t_i to t_{i+1} with initial data $u(t_i,\cdot)$.

We carefully consider the exponential factor since the argument of the exponential function is not a scalar but rather an infinite dimensional operator. Staying with the intuition of the limit of a matrix applied to a spacial discretization of u we consider the matrix exponential. The matrix exponential of a square matrix is defined by the power series

$$\exp\big(X\big) = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$$

where X^k is defined by taking the kth power of the eigenvalues in the eigendecomposition [Hig08].

The general first-order splitting of the time stepping operator for the linear PDE in Equation (8.1.1) that we consider splits the sum $A + B$ on the right-hand side of the equivalent Equation (8.1.2) into A and B. That is why we defined A and B as two separate linear operators in the first place. The corresponding first substep is the solution of the linear PDE initial value problem

$$\partial_t u^* = Au^*, \qquad u^*(t_i,\cdot) = u(t_i,\cdot)$$

and the second substep is the solution of the linear PDE initial value problem

$$\partial_t u^{**} = Bu^{**}, \qquad u^{**}(t_i,\cdot) = u^*(t_{i+1},\cdot)\ .$$

Proposition 48. The error due to sequential first-order splitting of Equation (8.1.1) with time steps of size $t_{i+1} - t_i = \Delta t$ for all $i \in \{0,\ldots,m-1\}$ up to a constant time t such that $t_0 = 0, t_m = t$ converges at a linear rate $O(\Delta t)$ as $\Delta t \to 0$.

Proof. Due to linearity for the first substep of the first-order splitting time step we have the solution

$$u^*(t_{i+1},\cdot) = \exp\big(\Delta t A\big)u(t_i,\cdot)$$

and for the second substep we have the solution

$$\begin{aligned} u^{**}(t_{i+1},\cdot) &= \exp\big(\Delta t B\big)u^*(t_{i+1},\cdot) \\ &= \exp\big(\Delta t B\big)\exp\big(\Delta t A\big)u(t_i,\cdot)\ . \end{aligned}$$

Both $\exp\big(\Delta t B\big)$ and $\exp\big(\Delta t A\big)$ are infinite dimensional operators that in many cases can intuitively be considered the limits of matrices which do not necessarily commute. Therefore, for the operator product we cannot use the rules for multiplying scalar exponentials. Instead, we multiply the two operators based on their power series definitions.

From the power series definition of the matrix exponential we have

$$\exp\big(\Delta t B\big) = \sum_{k=0}^{\infty} \frac{1}{k!} \big(\Delta t B\big)^k$$

$$
\begin{aligned}
&= \sum_{k=0}^{\infty} \frac{1}{k!} \big(\Delta t\big)^k B^k \\
&= I + \Delta t B + \frac{1}{2} \Delta t^2 B^2 + O(\Delta t^3)
\end{aligned}
$$

as $\Delta t \to 0$. Analogously, we also have

$$\exp\big(\Delta t A\big) = I + \Delta t A + \frac{1}{2} \Delta t^2 A^2 + O(\Delta t^3)$$

as $\Delta t \to 0$. For the operator product we then have

$$
\begin{aligned}
\exp\big(\Delta t B\big) \exp\big(\Delta t A\big) &= \big(I + \Delta t B + \frac{1}{2} \Delta t^2 B^2 + O(\Delta t^3)\big)\big(I + \Delta t A + \frac{1}{2} \Delta t^2 A^2 + O(\Delta t^3)\big) \\
&= I + \Delta t(B + A) + \frac{1}{2} \Delta t^2 (B^2 + A^2 + 2BA) + O(\Delta t^3)
\end{aligned}
$$

as $\Delta t \to 0$.

The exact solution $u(t_{i+1}, \cdot)$ for the time step from t_i to t_{i+1} is given by Equation (8.1.3). By the power series definition of the matrix exponential we have

$$
\begin{aligned}
\exp\big(\Delta t(A + B)\big) &= I + \Delta t(A + B) + \frac{1}{2} \Delta t^2 (A + B)^2 + O(\Delta t^3) \\
&= I + \Delta t(A + B) + \frac{1}{2} \Delta t^2 (A^2 + AB + BA + B^2) + O(\Delta t^3)
\end{aligned}
$$

as $\Delta t \to 0$. For the local error of the first-order splitting time step we then have

$$
\begin{aligned}
u(t_{i+1}, \cdot) - u^{**}(t_{i+1}, \cdot) &= \Big(\exp\big(\Delta t(A + B)\big) - \exp\big(\Delta t B\big) \exp\big(\Delta t A\big)\Big) u(t_i, \cdot) \\
&= \frac{1}{2} \Delta t^2 (AB - BA) u(t_i, \cdot) + O(\Delta t^3)
\end{aligned}
$$

as $\Delta t \to 0$. If A and B do not commute then the local error is $O(\Delta t^2)$ as $\Delta t \to 0$.

Since we perform $m = O(1/\Delta t)$ such local splitting time steps sequentially to solve up to a constant time $t = t_m = m \cdot \Delta t$ the global accumulated error is $O(\Delta t)$ as $\Delta t \to 0$. This concludes the proof. □

A similar error analysis can be performed for nonlinear operators based on their Jacobians [Gei10].

8.1.3 Discrete First-order Splitting

We now introduce a first-order splitting numerical integration scheme based on the first-order splitting time step of Definition 45 combined with the Lax-Friedrichs scheme introduced for scalar conservation laws in Section 3.2.7.

The first-order splitting time step consists of two substeps that both involve solving one dimensional initial value problems over the whole space domain. We know how to solve one dimensional scalar conservation law initial value problems via the Lax-Friedrichs scheme. So a natural idea for a numerical integrator for two dimensional problems is to replace the continuous time solution of the first-order splitting substeps by the discrete Lax-Friedrichs time stepping solution.

We use $U^{k+\frac{1}{2}}$ in order to denote the result of the first substep of the first-order splitting time step from t_k to t_{k+1} in the discrete setting. The Lax-Friedrichs first-order splitting on an equidistant space grid with $\Delta y = \Delta x$ then consists of the first substep

$$U_{i,j}^{k+\frac{1}{2}} = \frac{1}{2}(U_{i-1,j}^k + U_{i+1,j}^k) - \frac{\Delta t_k}{2\Delta x}(f(U_{i+1,j}^k) - f(U_{i-1,j}^k))$$

and the second substep

$$U_{i,j}^{k+1} = \frac{1}{2}(U_{i,j-1}^{k+\frac{1}{2}} + U_{i,j+1}^{k+\frac{1}{2}}) - \frac{\Delta t_k}{2\Delta x}(g(U_{i,j+1}^{k+\frac{1}{2}}) - g(U_{i,j-1}^{k+\frac{1}{2}}))\ .$$

Both substeps have to be performed over all cells, i.e. the substeps are performed in loops for all $i, j \in \{1, \ldots, n\}$.

The time step we use is based on the CFL condition with CFL number set to one. The CFL time step upper bound is derived from the maximum characteristic over the whole discrete solution for both the first and the second substep, i.e. we use

$$\Delta t_k = \frac{1}{\max\Big\{\max_{i,j}|f'(U_{i,j}^k)|, \max_{i,j}|g'(U_{i,j}^k)|\Big\}}\Delta x$$

which guarantees that $\Delta t_k = O(\Delta x)$ as $\Delta x \to 0$ and the numerical solution is stable according to the CFL condition (see Section 3.2.4).

In the proof of Proposition 48 we have considered the local error of the continuous first-order splitting solution operator as

$$u(t_{i+1}, \cdot) - u^{**}(t_{i+1}, \cdot) = \mathcal{S}u(t_i, \cdot) - \mathcal{S}_y\mathcal{S}_x u(t_i, \cdot)$$

where $\mathcal{S}$ is the exact solution operator and $\mathcal{S}_x$, $\mathcal{S}_y$ are the first and second substep split solution operators respectively. The result of the discrete first-order splitting can analogously be denoted as

$$U(t_{i+1}, \cdot) = \mathcal{H}_y\mathcal{H}_x U(t_i, \cdot)$$

where $\mathcal{H}_x$, $\mathcal{H}_y$ are the numerical first and second substep split solution operators that introduce additional local truncation error.

In Section 3.2.7 we saw that the local truncation error of the Lax-Friedrichs scheme for the smooth solution of a scalar conservation law is $O(\Delta x^2)$ as $\Delta x \to 0$. In that setting the local error of the first-order splitting scheme for the time step from t_i to t_{i+1} will also remain $O(\Delta x^2)$ and, hence, assuming smoothness the global error of a sequential discrete first-order splitting solution is convergent at a first-order rate $O(\Delta x)$ as $\Delta x \to 0$.

The convergence theory for discontinuous solutions is more involved and we have approximation error around discontinuities due to numerical viscosity at a rate that depends on the entropy condition. For shock discontinuities with compression we also observe a first-order rate of convergence for the sequential discrete first-order splitting solution in practice, e.g. in the next section.

8.2 Case Study: 2D Burgers Equation

We now consider a parametric initial value problem for the 2D Burgers equation obtained from Equation (8.0.1) with

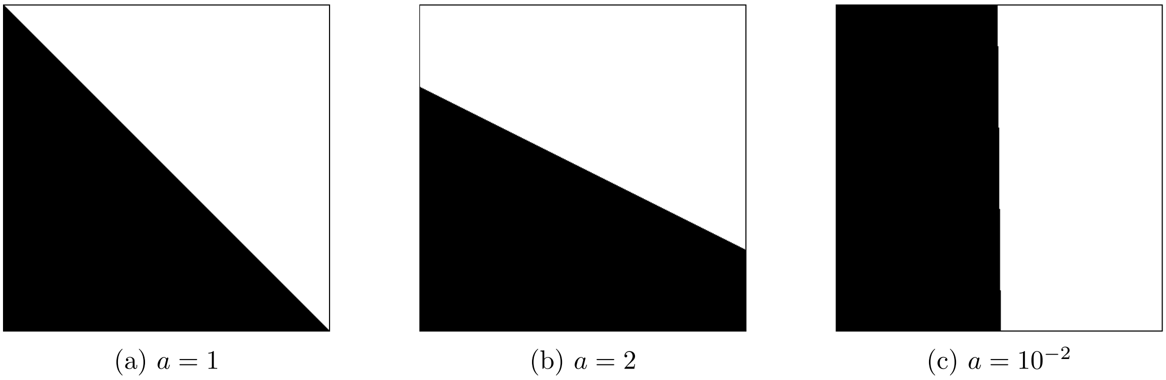
$$f(u) = \frac{1}{2}u^2, \qquad g(u) = a\frac{1}{2}u^2$$

and the initial data

$$u_0(x, y) = (1 + p)H\Big(-y - \frac{x}{a}\Big) \tag{8.2.1}$$

where p is the parameter that we consider for the sensitivity analysis and $a > 0$ is a constant.

The given initial data is a 2D generalization of the standard Riemann problem where the intial data contains a single 1D shock manifold the orientation of which depends on a. Regarding the Riemann problem for general two dimensional scalar conservation laws see e.g. Chen, Li, and Tan [CLT96]. Figure 8.1 shows a visualization of the initial data for three different values of a where the black area denotes the part of the domain for which the initial data takes the value $1 + p$ and the white area denotes the part for which the initial data takes the value 0. The 1D shock manifold takes the shape of a linear shock wave front.



(a) $a = 1$

(b) $a = 2$

(c) $a = 10^{-2}$

Figure 8.1: Shock wave front initial data contour with $p = 0$ (black is $1 + p$, white is 0).

In the following we derive an analytic weak solution for the initial value problem and use it to analyze the empirical convergence of the discrete first-order splitting solution based on the Lax-Friedrichs scheme described in the previous section. The same initial value problem is used to analyze numerical methods for tracking points on the shock manifold and their sensitivities in the next sections.

In order to understand the dynamics of the 2D Riemann problem it is helpful to realize that the characteristic speeds for the 2D burgers equations in the two spacial directions are given by the derivatives of the flux

$$f'(u) = u, \qquad g'(u) = au$$

just like for the characteristics for the separate 1D scalar conservation law steps obtained by first-order splitting. So for example with $u = 1$, e.g. in the black area of the initial data in Figure 8.1, we have the characteristic speed direction vector as

$$\partial_t \xi = \begin{pmatrix} f'(u) \\ g'(u) \end{pmatrix} = \begin{pmatrix} 1 \\ a \end{pmatrix} .$$

The characteristic lines within the black area of Figure 8.1 run in a direction that is orthogonal to the linear shock front. The characteristics in the white area are zero. Thus, we have compression at the shock front and the entropy condition is satisfied. We also have a propagation of the shock front at the speed of the average of the characteristics on both sides of the shock just like for the 1D Burgers equations and, hence, in a direction that is orthogonal to the shock front as well.

Proposition 49. The weak solution of Equation (8.0.1) with initial data according to Equation (8.2.1) and $f'(u) = u, g'(u) = au$ is

$$u(t,x,y) = (1+p)H\Big(-y - \frac{x}{a} + \Big(a + \frac{1}{a}\Big)\frac{1}{2}t(1+p)\Big) .$$

Proof. By weak differentiation of u we show that Equation (8.0.1) is satisfied in the weak sense.

By the chain rule the weak derivative of u with respect to time t is

$$\partial_t u(t,x,y) = \delta\Big(-y - \frac{x}{a} + \Big(a + \frac{1}{a}\Big)\frac{1}{2}t(1+p)\Big)\Big(a + \frac{1}{a}\Big)\frac{1}{2}(1+p)^2 .$$

We write the discontinuous fluxes in terms of Heaviside step function in order to make the distributional weak calculus applicable. We then have

$$\begin{aligned} f(u(t,x,y)) &= H\Big(-y - \frac{x}{a} + \Big(a + \frac{1}{a}\Big)\frac{1}{2}t(1+p)\Big)\frac{1}{2}(1+p)^2 \\ g(u(t,x,y)) &= H\Big(-y - \frac{x}{a} + \Big(a + \frac{1}{a}\Big)\frac{1}{2}t(1+p)\Big)a\frac{1}{2}(1+p)^2 . \end{aligned}$$

By the chain rule the weak derivative of the fluxes with respect to the corresponding spacial variables x and y is

$$\begin{aligned} \partial_x f(u(t,x,y)) &= -\delta\Big(-y - \frac{x}{a} + \Big(a + \frac{1}{a}\Big)\frac{1}{2}t(1+p)\Big)\frac{1}{a}\frac{1}{2}(1+p)^2 \\ \partial_y g(u(t,x,y)) &= -\delta\Big(-y - \frac{x}{a} + \Big(a + \frac{1}{a}\Big)\frac{1}{2}t(1+p)\Big)a\frac{1}{2}(1+p)^2 . \end{aligned}$$

Adding the weak derivatives of the two fluxes gives

$$\partial_x f(u) + \partial_y g(u) = -\delta\Big(-y - \frac{x}{a} + \Big(a + \frac{1}{a}\Big)\frac{1}{2}t(1+p)\Big)\Big(\frac{1}{a} + a\Big)\frac{1}{2}(1+p)^2$$

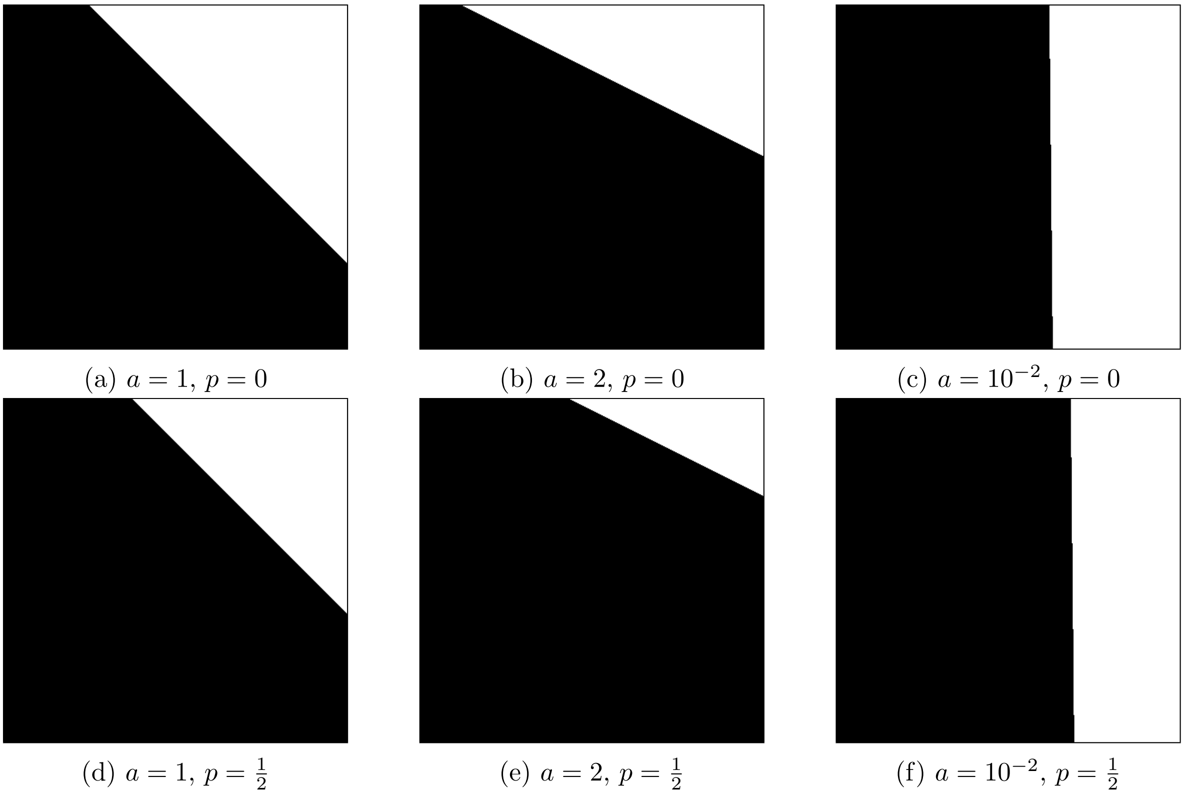
that is the negative weak derivative of u with respect to time t.

We have Equation (8.0.1) satisfied in a weak sense since

$$\int_0^t \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \big(\partial_t u + \partial_x f(u) + \partial_y g(u)\big)\phi(t,x,y)\mathrm{d}x\mathrm{d}y\mathrm{d}t = 0$$

for all smooth test functions ϕ with compact support. This concludes the proof. □

Figure 8.2 shows the analytic weak solution for the 2D Riemann initial value problem at $t = 1/2$ in a way that is analogous to Figure 8.1. The black area denotes the part of the domain for which the solution takes the value $1 + p$ and the white area denotes the part for which the solution takes the value 0. The figures in the top row show the solution for the unperturbed problem with $p = p_0 = 0$. The figures in the bottom row show the solution for the perturbed parameter $p = p_0 + \dot{p} = 1/2$.



(a) $a = 1$, $p = 0$

(b) $a = 2$, $p = 0$

(c) $a = 10^{-2}$, $p = 0$

(d) $a = 1$, $p = \frac{1}{2}$

(e) $a = 2$, $p = \frac{1}{2}$

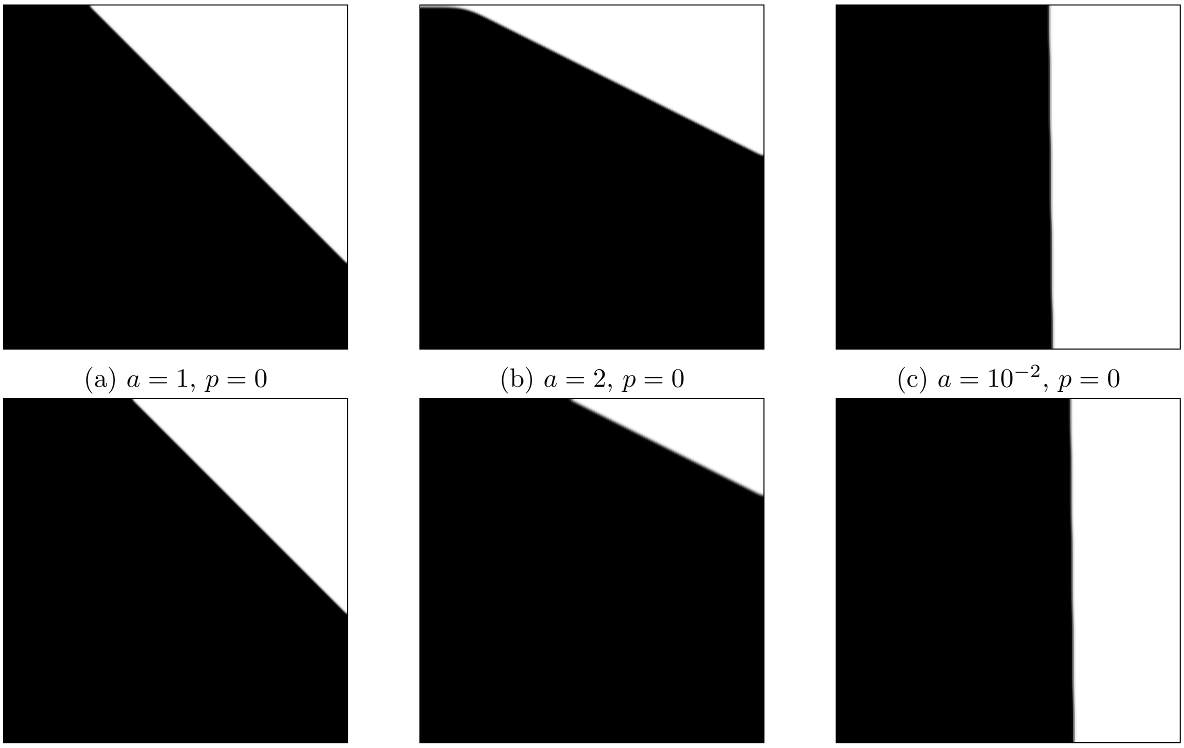
(f) $a = 10^{-2}$, $p = \frac{1}{2}$

Figure 8.2: Analytic weak solution u at $t = \frac{1}{2}$ (black is $1 + p$, white is 0).

The first observation we make is that for all the different values of the constant a the orientation of the linear shock front is parallel to that of the corresponding figure in Figure 8.1, i.e. the shock front travels in a direction that is orthogonal to the shock front itself. For all three values of a we observe that the shock front has traveled in the direction of the white area as expected from the characteristic directions compressing into the shock front. Upon closer inspection we also observe that for the higher value of $p = 1/2$ the shock has traveled further away from its initial configuation at $t = 0$ than for $p = 0$. For higher values of p we have higher values of the shock speed similar to the 1D parametric Riemann problem in Example 6.

Figure 8.3 shows a numerical weak solution for the 2D Riemann initial value problem at $t = 1/2$ in a way that is analogous to Figure 8.2. The numerical weak solution is obtained by the Lax-Friedrichs first-order splitting scheme described in the previous section using a CFL time step based on an equidistant space grid with $\Delta x = 5 \cdot 10^{-3}$.

Upon close inspection we observe a gray scale transition between the black and white regions of the figures where the shock is smeared out due to numerical viscosity. The shock smearing is barely visible to the eye with the given resolution since the discretization is already at a fairly high accuracy. A notable difference between the analytic and the numerical solution is visible when zooming in to the figure for $a = 2, p = 0$. In the numerical solution we have values of zero all along the upper boundary of the figure. This is an error that occurs because of the way the boundary conditions are handled in our implementation of the first-order splitting approach. Similar to what we described for the error of the 1D discretization in Section 6.1.1 we have errors of order $O(\Delta x)$ as $\Delta x \to 0$ at the boundary of the scalar subproblems that make up the first-order splitting steps. These errors at the boundary propagate in such a way that a whole area of the solution near the boundary takes the wrong values.



(a) $a = 1$, $p = 0$

(b) $a = 2$, $p = 0$

(c) $a = 10^{-2}$, $p = 0$

(d) $a = 1$, $p = \frac{1}{2}$

(e) $a = 2$, $p = \frac{1}{2}$

(f) $a = 10^{-2}$, $p = \frac{1}{2}$

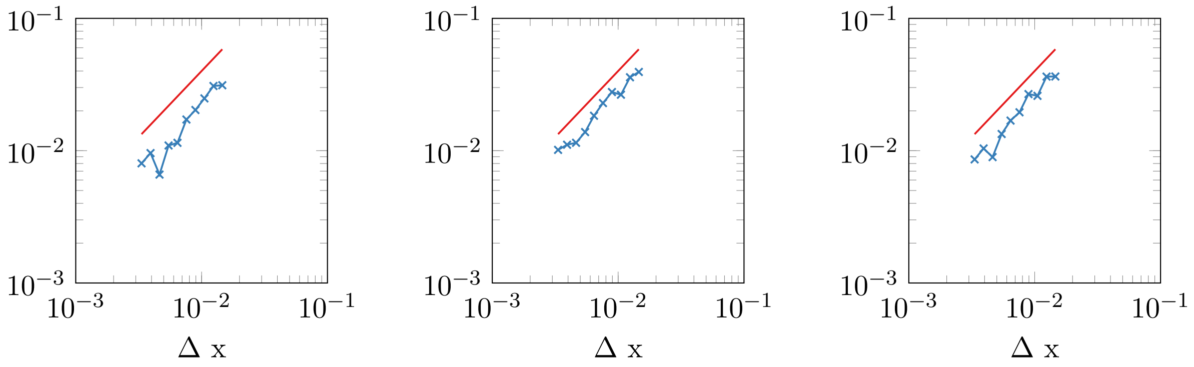
Figure 8.3: Lax-Friedrichs solution ($\Delta x = 5 \cdot 10^{-3}$) at $t = \frac{1}{2}$ (black is $1 + p$, white is 0).

In the following empirical convergence results we see that despite the errors near the boundary the Lax-Friedrichs first-order splitting converges at a first-order rate, i.e. similar to the 1D setting the errors at the boundary are only of the order $O(\Delta x)$ as $\Delta x \to 0$.

Figure 8.4 shows the empirical convergence of the numerical solution obtained via the Lax-Friedrichs based discrete first-order splitting scheme. For geometrically decaying values of Δx the blue graph shows an approximation of the L_1 norm error between the numerical and analytic solution by

$$\sum_{i=1}^{n} \sum_{j=1}^{n} |U_{i,j}^{m} - u(t_m, x_i, x_j)| \Delta x^2 = \| U(t_m, x_i, x_j) - u(t_m, x_i, x_j) \|_{L_1(\hat{\Omega})} + O(\Delta x)$$

as $\Delta x \to 0$. There are integral discretization cells where the analytic solution is discontinuous and the approximation error to the true L_1 norm integral is constant. The number of cells with constant error is $O(\Delta x)$ because the only discontinuity of the analytic solution lies on a 1D manifold. Hence, the L_1 norm approximation is accurate up to $O(\Delta x)$ and observing a decay at an $O(\Delta x)$ rate as $\Delta x \to 0$ indicates linear convergence of the L_1 norm error. For all three values of the constant a we observe empirical first-order convergence at least in the given range of Δx that we consider.



(a) $a = 1$ (b) $a = 2$ (c) $a = 10^{-2}$

Figure 8.4: Empirical convergence of Lax-Friedrichs solution at $t_m = \frac{1}{2}$ with $p = 0$, $O(\Delta x)$ (red), $\sum_{i,j} |U^m_{i,j} - u(t_m, x_i, x_j)|\Delta x^2$ (blue)

8.3 Sensitivity Analysis for Points on the Shock Manifold

In this section we show how to track an individual point on the shock manifold and compute its sensitivities with respect to the parameters of the initial data. In principle, depending on the dynamic geometry of the shock, it is possible to reconstruct the shape of the shock manifold from geometric data structures storing the neighborhood relationship of multiple such individual tracked points on the manifold. However, the challenge of such a geometric approximation is left to future work and not treated in this thesis.

We first consider the Rankine-Hugoniot condition as it applies to the analytic weak solution. The Rankine-Hugoniot condition forms the basis for the derivation of our discrete shock sensitivity methods. We then provide two variants of a numerical shock location algorithm and the two corresponding variants of the shock sensitivity algorithm derived from the Rankine-Hugoniot condition.

8.3.1 Rankine-Hugoniot Condition

We now consider a version of the Rankine-Hugoniot condition analogous to Theorem 8 that holds on 1D shock manifolds for 2D conservation laws. The derivation of the 2D version of the Rankine-Hugoniot condition that we use is based on the idea of splitting along the two space dimension just like the first-order splitting approach introduced earlier. The result of first-order splitting are two sequential steps that we can consider like conventional 1D problems on which the 1D Rankine-Hugoniot condition holds.

We denote the 1D shock manifold by a parametric curve $\xi(t, z) \in \mathbb{R}^2$ where t is time and $z \in Z \subseteq \mathbb{R}$ is a parameter by which we move along the manifold. We note that in this chapter the index i on ξ_i does not denote the ith shock location as in the previous sections but instead the ith component of the vector giving a point on the shock manifold. For the methods we present for 2D conservation laws in this chapter we make a fundamental nondegeneracy assumption about the orientation of the smooth shock manifold we are considering.

Assumption 11. The shock manifold is strongly nondegenerate in the sense that the tangent of the 1D shock manifold is strongly nonparallel to either the x or the y axes, i.e. we have

$$|\partial_z \xi_1(t,z)| \geq \delta, \qquad |\partial_z \xi_2(t,z)| \geq \delta$$

for some $\delta > 0$, for all $x \in \mathbb{R}$.

The assumption may seem unrealistic in the sense that it might not be satisfied in practical applications. In fact, the assumption turns out to be unnecessarily restrictive if we do not motivate the methods we use from the first-order splitting perspective. In the end of this chapter we suggest an alternative method for computing shock sensitivities that does not rely on the above assumption. Alternatively, for practical applications it may also be possible to rotate the discretization and operator splitting axes so that they are not aligned with the shock front of interest.

Proposition 50. Let u be a weak solution of Equation (8.0.1) with differentiable flux functions f, g such that u has a shock discontinuity along a smooth manifold ξ and is smooth on either side of ξ. Then u satisfies the Rankine-Hugoniot conditions

$$\begin{aligned}
\partial_t \xi_1(t,z) &= \frac{f(u(t,\xi_1(t,z)+,\xi_2(t,z))) - f(u(t,\xi_1(t,z)-,\xi_2(t,z)))}{u(t,\xi_1(t,z)+,\xi_2(t,z)) - u(t,\xi_1(t,z)-,\xi_2(t,z))} \\
\partial_t \xi_2(t,z) &= \frac{g(u(t,\xi_1(t,z),\xi_2(t,z)+)) - g(u(t,\xi_1(t,z),\xi_2(t,z)-))}{u(t,\xi_1(t,z),\xi_2(t,z)+) - u(t,\xi_1(t,z),\xi_2(t,z)-)}
\end{aligned}$$

for all $z \in Z$.

Proof. After the more extensive proof of Theorem 8 that can in principle be generalized to the 2D setting we do not use the divergence theorem here but rather make a shorter and less formal asymptotic argument. We show the derivation for $\partial_t \xi_1(t,z)$ where we denote $\xi_2 \equiv \xi_2(t,z)$. The derivation for $\partial_t \xi_2(t,z)$ is analogous.

First, we note that according to the assumptions of the proposition the piecewise smooth weak solution u in the sense of the L_1 norm is equivalent to

$$u(t,x,\xi_2) = \begin{cases} u_1(t,x,\xi_2) & \text{if } x < \xi_1(t,z) \\ u_2(t,x,\xi_2) & \text{if } \xi_1(t,z) < x \end{cases}$$

for all $x \in [\xi_L, \xi_R]$ with $\xi_L < \xi_1(t,z) < \xi_R$ and ξ_L, ξ_R close to $\xi_1(t,z)$ and where u_1, u_2 are smooth functions of x. Since u_1, u_2 are smooth they are also locally Lipschitz continuous and we have

$$\begin{aligned}
u_1(t,x,\xi_2) &= u(t,\xi_L,\xi_2) + O(|x-\xi_L|) = u(t,\xi_L,\xi_2) + O(|\xi_R - \xi_L|) \\
u_2(t,x,\xi_2) &= u(t,\xi_R,\xi_2) + O(|x-\xi_R|) = u(t,\xi_R,\xi_2) + O(|\xi_R - \xi_L|)
\end{aligned}$$

for all $x \in [\xi_L, \xi_R]$ as $\xi_R, \xi_L \to \xi_1(t,z)$. In the sense of the L_1 norm we therefore have u as

$$u(t,x,\xi_2) = H(\xi_1(t,z) - x)u(t,\xi_L,\xi_2) + H(x - \xi_1(t,z))u(t,\xi_R,\xi_2) + O(|\xi_R - \xi_L|) \tag{8.3.1}$$

for all $x \in [\xi_L, \xi_R]$ as $\xi_R, \xi_L \to \xi_1(t,z)$ by rewriting the shock discontinuity in terms of Heaviside step functions and approximating the smooth pieces by their values at the domain boundary. Similarly, we can write the flux of the discontinuous weak solution respectively as

$$f(u(t,x,\xi_2)) = H(\xi_1(t,z) - x)f(u(t,\xi_L,\xi_2)) + H(x - \xi_1(t,z))f(u(t,\xi_R,\xi_2)) + O(|\xi_R - \xi_L|)$$

$$g(u(t,x,\xi_2)) = H(\xi_1(t,z)-x)g(u(t,\xi_L,\xi_2)) + H(x-\xi_1(t,z))g(u(t,\xi_R,\xi_2)) + O(|\xi_R-\xi_L|)$$

for all $x \in [\xi_L, \xi_R]$ as $\xi_R, \xi_L \to \xi_1(t,z)$ because f and g are also differentiable and hence locally Lipschitz continuous (c.f. Example 6).

We now consider the integral of Equation (8.0.1) over the same line domain along the x axis

$$\int_{\xi^L}^{\xi^R} \partial_t u(t,x,\xi_2)\mathrm{d}x = -\int_{\xi^L}^{\xi^R} \partial_x f(u(t,x,\xi_2))\mathrm{d}x - \int_{\xi^L}^{\xi^R} \partial_y g(u(t,x,\xi_2))\mathrm{d}x\ . \tag{8.3.2}$$

We analyze each integral term and the involved weak derivatives separately.

For the weak derivative of u with respect to time according to Section 2.4 and by the chain rule we have

$$\begin{aligned}\partial_t u(t,x,\xi_2) = {} & \delta(\xi_1(t,z)-x)\cdot\partial_t\xi_1(t,z)\cdot u(t,\xi_L,\xi_2) + H(\xi_1(t,z)-x)\cdot\partial_t u(t,\xi_L,\xi_2)\\ & - \delta(x-\xi_1(t,z))\cdot\partial_t\xi_1(t,z)\cdot u(t,\xi_R,\xi_2) + H(x-\xi_1(t,z))\cdot\partial_t u(t,\xi_R,\xi_2)\\ & + O(|\xi_R-\xi_L|)\end{aligned}$$

as $\xi_R, \xi_L \to \xi_1(t,z)$. The corresponding integral evaluates to

$$\begin{aligned}\int_{\xi^L}^{\xi^R} \partial_t u(t,x,\xi_2)\mathrm{d}x = {} & \partial_t\xi_1(t,z)\cdot u(t,\xi_L,\xi_2) + (\xi_1(t,z)-\xi_L)\cdot\partial_t u(t,\xi_L,\xi_2)\\ & - \partial_t\xi_1(t,z)\cdot u(t,\xi_R,\xi_2) + (\xi_R-\xi_1(t,z))\cdot\partial_t u(t,\xi_R,\xi_2)\\ & + (\xi_R-\xi_L)\cdot O(|\xi_R-\xi_L|)\end{aligned}$$

as $\xi_R, \xi_L \to \xi_1(t,z)$. The derivatives $\partial_t u(t,\xi_L,\xi_2)$ and $\partial_t u(t,\xi_R,\xi_2)$ are bounded since ξ_L and ξ_R lie on the smooth pieces of u by assumption, so we have

$$\int_{\xi^L}^{\xi^R} \partial_t u(t,x,\xi_2) = \partial_t\xi_1(t,z)(u(t,\xi_L,\xi_2) - u(t,\xi_R,\xi_2)) + O(|\xi_R-\xi_L|)$$

as $\xi_R, \xi_L \to \xi_1(t,z)$.

For the weak derivative of $f(u)$ with respect to x we have

$$\partial_x f(u(t,x,\xi_2)) = -\delta(\xi_1(t,z)-x)\cdot f(u(t,\xi_L,\xi_2)) + \delta(x-\xi_1(t,z))\cdot f(u(t,\xi_R,\xi_2)) + O(|\xi_R-\xi_L|)$$

as $\xi_R, \xi_L \to \xi_1(t,z)$. The corresponding integral evaluates to

$$\int_{\xi_L}^{\xi_R} \partial_x f(u(t,x,\xi_2))\mathrm{d}x = f(u(t,\xi_R,\xi_2)) - f(u(t,\xi_L,\xi_2)) + O(|\xi_R-\xi_L|^2)$$

as $\xi_R, \xi_L \to \xi_1(t,z)$. We get the same result from the Leipniz rule except for the approximation error that we carried over from before.

For the weak derivative of $g(u)$ with respect to y we have

$$\partial_y g(u(t,x,\xi_2)) = H(\xi_1(t,z)-x)\partial_y g(u(t,\xi_L,\xi_2)) + H(x-\xi_1(t,z))\partial_y g(u(t,\xi_R,\xi_2)) + O(|\xi_R-\xi_L|)$$

as $\xi_R, \xi_L \to \xi_1(t,z)$. The corresponding integral evaluates to

$$\int_{\xi_L}^{\xi_R} \partial_y g(u(t,x,\xi_2))\mathrm{d}x = (\xi_1(t,z)-x_L)\partial_y g(u(t,\xi_L,\xi_2)) + (x_R-\xi_1(t,z))\partial_y g(u(t,\xi_R,\xi_2))$$

$$+ (\xi_R - \xi_L)O(|\xi_R - \xi_L|)$$

as $\xi_R, \xi_L \to \xi_1(t,z)$. The derivatives $\partial_y g(u(t,\xi_L,\xi_2))$ and $\partial_y g(u(t,\xi_R,\xi_2))$ are bounded since ξ_L and ξ_R lie on the smooth pieces of u and g is differentiable by assumption, so we have

$$\int_{\xi_L}^{\xi_R} \partial_y g(u(t,x,\xi_2))\mathrm{d}x = O(|\xi_R - \xi_L|)$$

as $\xi_R, \xi_L \to \xi_1(t,z)$.

We substitute the derived integral in Equation (8.3.2) to get

$$\partial_t\xi_1(t,z)(u(t,\xi_L,\xi_2) - u(t,\xi_R,\xi_2)) = f(u(t,\xi_L,\xi_2)) - f(u(t,\xi_R,\xi_2)) + O(|\xi_R - \xi_L|) \tag{8.3.3}$$

as $\xi_R, \xi_L \to \xi_1(t,z)$. In the limit we have the Rankine-Hugoniot condition of the proposition

$$\partial_t\xi_1(t,z) = \frac{f(u(t,\xi_1(t,z)+,\xi_2)) - f(u(t,\xi_1(t,z)-,\xi_2))}{u(t,\xi_1(t,z)+,\xi_2) - u(t,\xi_1(t,z)-,\xi_2)} \ .$$

The proof for $\partial_t\xi_2(t,z)$ is analogous by switching the x and y axes. This concludes the proof.

□

Example 47. We consider the example of the 2D Burgers equation with the initial data of the case study of the previous section and a point on the shock manifold parameterized by $z = z_0$ such that for $t = 0$ we have the origin as the initial value $\xi(0,z_0) = (0\ 0)^T$.

For some point $\xi(t,z)$ on the shock manifold we have its dynamics for the 2D Burgers fluxes f, g given by

$$\begin{aligned}\partial_t\xi_1(t,z) &= \frac{1}{2}\frac{u(t,\xi_1(t,z)+,\xi_2(t,z))^2 - u(t,\xi_1(t,z)-,\xi_2(t,z))^2}{u(t,\xi_1(t,z)+,\xi_2(t,z)) - u(t,\xi_1(t,z)-,\xi_2(t,z))} &(8.3.4)\\ &= \frac{1}{2}\big(u(t,\xi_1(t,z)+,\xi_2(t,z)) + u(t,\xi_1(t,z)-,\xi_2(t,z))\big) &(8.3.5)\end{aligned}$$

and, analogously,

$$\partial_t\xi_2(t,z) = a\frac{1}{2}\big(u(t,\xi_1(t,z),\xi_2(t,z)+) + u(t,\xi_1(t,z),\xi_2(t,z)-)\big) \ . \tag{8.3.6}$$

For $t = 0$ and the initial data u_0 according to the previous section we then have

$$\partial_t\xi_1(0,z_0) = \frac{1}{2}(1+p), \qquad \partial_t\xi_2(0,z_0) = a\frac{1}{2}(1+p) \ .$$

The resulting shock location tracks the point on the shock manifold at

$$\xi_1(t,z_0) = \frac{1}{2}(1+p)t, \qquad \xi_2(t,z_0) = a\frac{1}{2}(1+p)t \ .$$

If we compare with Figure 8.2 or the analytic weak solution we see that the point does in fact lie on the shock manifold for all $t \geq 0$.

8.3.2 Numerical Shock Location

We now consider the numerical shock location algorithm of Section 6.2.1 as it applies to a point on the shock manifold according to the idea of first-order splitting the 2D problem to obtain two separate 1D problems where we can use the established method.

Like in the previous chapters, we do not use the Rankine-Hugoniot condition for simulating the shock location since, as we argued in Section 1.4.5, a direct explicit Euler discretization of the discontinuous ODE with the characteristic speed on the right-hand side results in a higher accuracy approximation. The discontinuous ODE giving the characteristics in the case of the 2D Burgers equation is

$$\partial_t \xi(t,z) = \begin{pmatrix} u(t,\xi_1(t,z),\xi_2(t,z)) \\ a \cdot u(t,\xi_1(t,z),\xi_2(t,z)) \end{pmatrix} .$$

The Filippov differential inclusion resulting from the above discontinuous ODE system [Cor08; Ito79] is

$$\partial_t \xi(t,z) \in \bigcap_{\delta>0} \bigcap_{\mu(S)=0} \overline{\text{co}} \left\{ \begin{pmatrix} u(t,x_1,x_2) \\ a \cdot u(t,x_1,x_2) \end{pmatrix} \;\middle|\; x \in B(\xi(t,z),\delta) \setminus S \right\}$$

where $\overline{\text{co}}$ denotes the convex closure, μ is the Lebesgue measure and $B(\xi,\delta)$ is the open ball centered at ξ with radius δ. We consider a closed convex hull over points approaching $\xi(t,z)$ along curves on which the weak solution is smooth. The first set intersection is over all $\delta > 0$, i.e. observing the limit of smaller and smaller neighborhoods around ξ. The second set intersection is over all sets of Lebesgue measure zero that get excluded from the neighborhood that we consider for the convex closure, i.e. also excluding the 1D shock manifold that has measure zero.

Under our nondegeneracy assumption (Assumption 11) we can equally consider the limits to the shock manifold along the corresponding axes like we did for the Rankine-Hugoniot condition in Proposition 50. For example for the 2D Burgers case study we get

$$\begin{aligned} \partial_t \xi_1(t,z) &\in [u(t,\xi_1(t,z)-,\xi_2(t,z)), u(t,\xi_1(t,z)+,\xi_2(t,z))] \\ \partial_t \xi_2(t,z) &\in [a \cdot u(t,\xi_1(t,z),\xi_2(t,z)-), a \cdot u(t,\xi_1(t,z),\xi_2(t,z)+)] . \end{aligned}$$

Since the manifold tangent is never parallel to either of the axes all of the limits along the axes converge to $\xi(t,z)$ through smooth parts of the weak solution and since the 1D shock manifold is the only discontinuity taking the convex closure of the limit from two sides gives the full closure hull for the Filippov differential inclusion.

We consider two versions of simulating the shock location based on the explicit Euler discretization of the characteristics.

The perspective we take for the first version of the shock location algorithm is according to the first-order splitting idea and its individual 1D conservation law steps. We consider the components of the point on the shock manifold as the shock location in the corresponding 1D problems and treat it according to the method presented for 1D problems in the previous chapters. In the first version of the algorithm the discrete time steps for the two spacial components of the shock locations are interleaved with the splitting steps of the discrete time step of the conservation law corresponding to the respective axis.

Definition 46. The *shock location algorithm (1)* is defined by the time step

$$\Xi_1^{k+1} := \Xi_1^k + \Delta t_k \cdot U^k(\Xi_1^k, \Xi_2^k)$$

$$\Xi_2^{k+1} := \Xi_2^k + \Delta t_k \cdot a \cdot U^{k+\frac{1}{2}}(\Xi_1^{k+1}, \Xi_2^k)$$

where $U^k(\cdot,\cdot)$ and $U^{k+\frac{1}{2}}(\cdot,\cdot)$ are linear interpolations of the approximations obtained by the first-order splitting Lax-Friedrichs discretization.

In the second version of the shock location algorithm we just perform the explicit Euler discretization of the characteristics based on the numerical weak solution at the start of the time step.

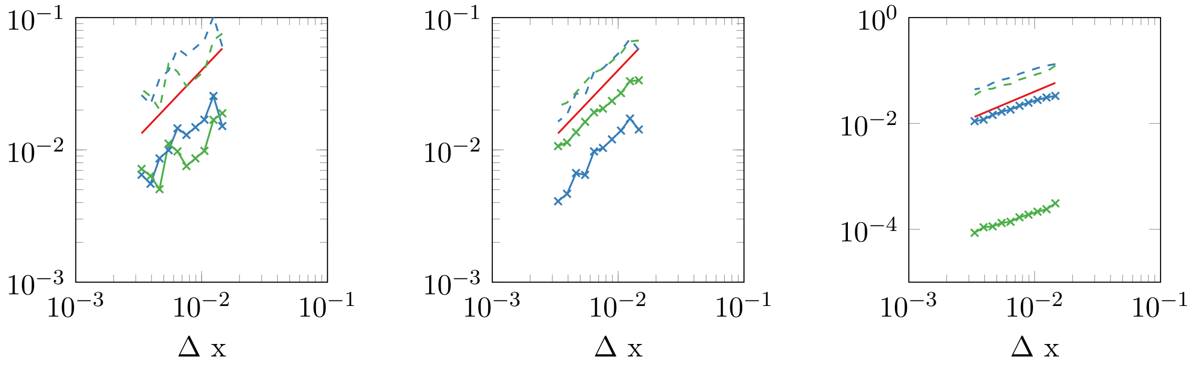
Definition 47. The *shock location algorithm (2)* is defined by the time step

$$\Xi_1^{k+1} := \Xi_1^k + \Delta t_k \cdot U^k(\Xi_1^k, \Xi_2^k)$$
$$\Xi_2^{k+1} := \Xi_2^k + \Delta t_k \cdot a \cdot U^k(\Xi_1^k, \Xi_2^k)$$

where $U^k(\cdot,\cdot)$ is a linear interpolation of the approximation obtained by the first-order splitting Lax-Friedrichs discretization.

For both methods the time step Δt_k is the same as the one used for the numerical integrator of the weak solution.

Figure 8.5 shows the empirical convergence results for the shock location algorithm (1) applied to the 2D Burgers equation case study with $p = p_0 = 0$ where the numerical weak solution is obtained via Lax-Friedrichs first-order splitting. The absolute error of the first (second) component of the shock point vector is shown in solid blue (green). The relative error of the first (second) component of the shock point vector is shown in dashed blue (green). The red line in the plots illustrates the linear rate $O(\Delta x)$ which is the rate at which we expect the shock location to converge based on what we have shown for the 1D setting. We plot both absolute and relative error because the absolute error differs so strongly between the different orientations of the shock front given by a.



(a) $a = 1$

(b) $a = 2$

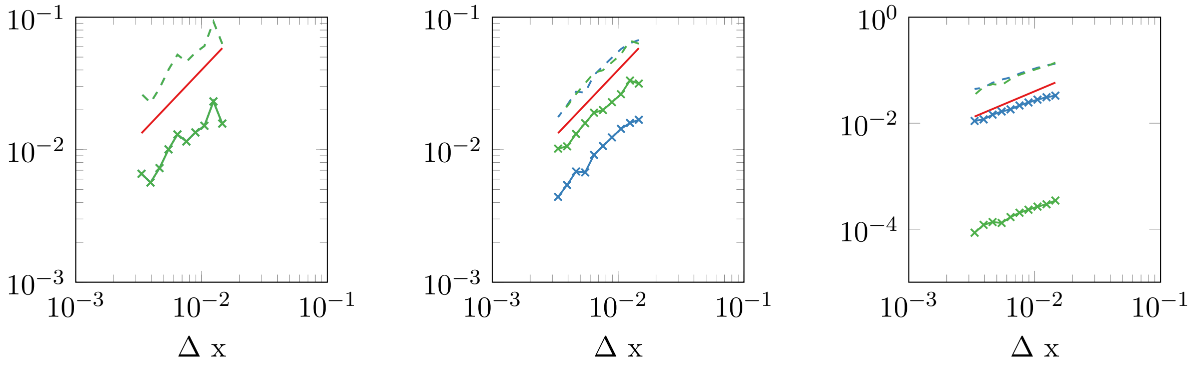
(c) $a = 10^{-2}$

Figure 8.5: Shock location algorithm (1): Empirical convergence at $t_m = \frac{1}{2}$, $O(\Delta x)$ (red), $|\Xi_1^m - \xi_1(t_m, z_0)|$ (solid blue), $|\Xi_2^m - \xi_2(t_m, z_0)|$ (solid green), $|\Xi_1^m - \xi_1(t_m, z_0)|/|\xi_1(t_m, z_0)|$ (dashed blue), $|\Xi_2^m - \xi_2(t_m, z_0)|/|\xi_2(t_m, z_0)|$ (dashed green)

For all three values of a we observe that both the absolute and relative approximation errors converge as $\Delta x \to 0$. For $a = 1$ and $a = 2$ the convergence rate seems linear although not smoothly dependent on Δx. For $a = 10^{-2}$ the convergence rate could also be sublinear. The nonsmooth dependence on Δx is explained by the oscillating behavior for the explicit Euler discretization around the discontinuity in the right-hand side of the ODE (see Section 1.4.4).

For the diagonal shock front with $a = 1$ that is symmetric between the two axes we observe that the absolute error in both components of the shock point vector is at the same order of magnitude. But for the tilted shock fronts the absolute errors in the two components differ significantly. The stronger the orientation differs from diagonal, the larger the difference in error. The difference can be explained by the fact that e.g. the smaller (larger) a, the smaller (larger) the dynamic in the y axis and, hence, the smaller (larger) the value of the y component of the shock point vector in the absolute sense. The relative approximation errors are roughly the same for all values of a.

Figure 8.6 shows the empirical convergence results of the shock location algorithm (2) analogous to Figure 8.5. In the left-most figure for $a = 1$ the green and blue lines are identical and plotted in green.



(a) $a = 1$

(b) $a = 2$

(c) $a = 10^{-2}$

Figure 8.6: Shock location algorithm (2): Empirical convergence at $t_m = \frac{1}{2}$, $O(\Delta x)$ (red), $|\Xi_1^m - \xi_1(t_m, z_0)|$ (solid blue), $|\Xi_2^m - \xi_2(t_m, z_0)|$ (solid green), $|\Xi_1^m - \xi_1(t_m, z_0)|/|\xi_1(t_m, z_0)|$ (dashed blue), $|\Xi_2^m - \xi_2(t_m, z_0)|/|\xi_2(t_m, z_0)|$ (dashed green)

For the symmetric problem with $a = 1$ we have an identical solution and consequently identical approximation errors for the two shock point vector components. Otherwise, the empirical convergence behavior is qualitatively the same for shock location algorithm (2) as what we observed for shock location algorithm (1) in Figure 8.5.

8.3.3 Numerical Shock Sensitivity

We now analyze the empirical convergence of the approximation of the sensitivities of the tracked shock point with respect to the parameter p of the initial data based on the shock sensitivity algorithm of Section 6.2.2. We consider the shock sensitivity algorithm as it applies in the sense of the first-order splitting steps for the shock location algorithm (1) and as it is motivated by the Rankine-Hugoniot condition for the 2D setting for the shock location algorithm (2).

Example 48. We consider for example the point on the shock manifold with $\xi(0, z_0) = (0\ 0)^T$ that we already considered earlier for the 2D Burgers Riemann problem case study. Based on the analytic shock point trajectory the analytic tangent sensitivities are

$$\dot{\xi}_1(t,z_0) = \partial_p \xi_1(t,z_0)\dot{p} = \frac{1}{2}t\dot{p}, \quad \dot{\xi}_2(t,z_0) = \partial_p \xi_2(t,z_0)\dot{p} = a\frac{1}{2}t\dot{p}\,.$$

A stably differentiable discrete time step for the shock point can be obtained based on an approximation of the Rankine-Hugoniot condition and using the AD approach of Section 6.5 for both shock location algorithm (1) and (2). Since we focus on the 2D Burgers equation we consider approximations of the Rankine-Hugoniot condition specifically for the Burgers equation that we derived in Equation (8.3.5) and Equation (8.3.6).

Analogous to the shock location algorithm (1) discretization of the shock point dynamics interleaved with the separate first-order splitting steps we define the explicit Euler discretization of the Rankine-Hugoniot condition as follows.

Definition 48. The *differentiable shock location step (1)* is defined by the time step

$$\begin{aligned}
\Xi_1^{k+1} &:= \Xi_1^k + \Delta t \cdot \frac{1}{2} \cdot \Big(U^k(\Xi_1^k + C\Delta x^\alpha, \Xi_2^k) + U^k(\Xi_1^k - C\Delta x^\alpha, \Xi_2^k) \Big) \\
\Xi_2^{k+1} &:= \Xi_2^k + \Delta t \cdot a \cdot \frac{1}{2} \cdot \Big(U^{k+\frac{1}{2}}(\Xi_1^{k+1}, \Xi_2^k + C\Delta x^\alpha) + U^{k+\frac{1}{2}}(\Xi_1^{k+1}, \Xi_2^k - C\Delta x^\alpha) \Big)
\end{aligned}$$

where $U^k(\cdot,\cdot)$ and $U^{k+\frac{1}{2}}(\cdot,\cdot)$ are linear interpolations of the approximations obtained by the first-order splitting Lax-Friedrichs discretization.

We apply the idea of Section 1.4.5 directly to the two steps of the first-order splitting. The left and right limits of the Rankine-Hugoniot condition are approximated by $\Xi_i \pm C\Delta x^\alpha$ for $i \in \{1,2\}$ with $\alpha \in (0,1)$ and $C > 0$ large enough to make sure we are far enough outside of the shock region to avoid the numerical viscosity.

Analogous to the shock location algorithm (2) discretization of the shock point dynamics we have the explicit Euler discretization of the Rankine-Hugoniot condition of the 2D Burgers equation defined as follows.

Definition 49. The *differentiable shock location step (2)* is defined by the time step

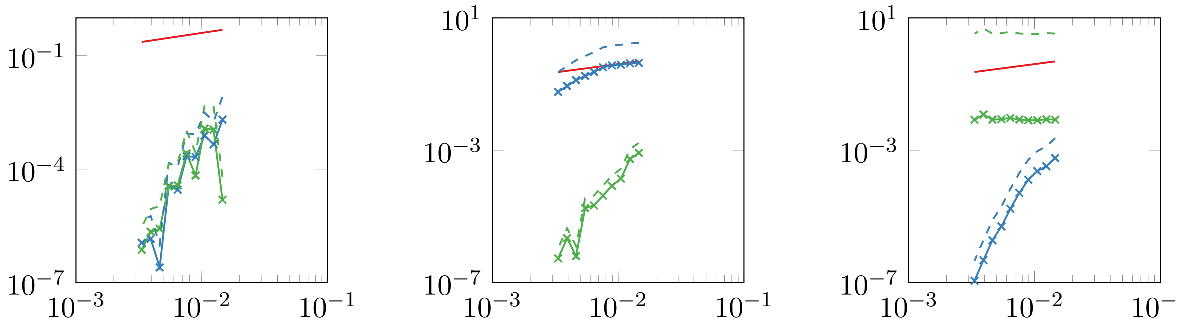
$$\begin{aligned}
\Xi_1^{k+1} &:= \Xi_1^k + \Delta t \cdot \frac{1}{2} \cdot \Big(U^k(\Xi_1^k + C\Delta x^\alpha, \Xi_2^k) + U^k(\Xi_1^k - C\Delta x^\alpha, \Xi_2^k) \Big) \\
\Xi_2^{k+1} &:= \Xi_2^k + \Delta t \cdot a \cdot \frac{1}{2} \cdot \Big(U^k(\Xi_1^k, \Xi_2^k + C\Delta x^\alpha) + U^k(\Xi_1^k, \Xi_2^k - C\Delta x^\alpha) \Big)
\end{aligned}$$

where $U^k(\cdot,\cdot)$ is a linear interpolation of the approximation obtained by the first-order splitting Lax-Friedrichs discretization.

We apply the idea of Section 1.4.5 directly to the Rankine-Hugoniot condition. Again, the left and right limits of the Rankine-Hugoniot condition are approximated by $\Xi_i \pm C\Delta x^\alpha$ for $i \in \{1,2\}$ with $\alpha \in (0,1)$ and $C > 0$.

The corresponding discrete tangents $\dot{\Xi}$ are obtained by AD methods exactly analogous to those described in Section 6.5 applied to the two differentiable shock location step versions. We do not derive the explicit formulas here since they are similar to what we already explicitly showed in Section 6.5. Crucially, we use the shock point location values of the shock location algorithms (1) and (2) in combination with the step-wise discrete tangents of the differentiable shock location steps (1) and (2). We call the resulting algorithms combining the shock location algorithm (1) and (2) with the differentiable shock location step (1) and (2) the shock sensitivity algorithm (1) and (2).

Figure 8.7 shows the empirical convergence results for the shock sensitivity algorithm (1) with $C = 1$ and $\alpha = 1/2$ applied to the case study with $p = p_0 = 0, \dot{p} = 1$ where the numerical weak solution and its discrete tangent is obtained via Lax-Friedrichs first-order splitting. The absolute error of the first (second) component of the shock sensitivity vector is shown in solid blue (green). The relative error of the first (second) component of the shock sensitivity vector is shown in dashed blue (green). The red line in the plots illustrates the sublinear rate $O(\Delta x^\alpha)$, i.e. the general asymptotic rate at which we can at best expect the numerical approximation error to converge for the shock sensitivities. We plot both absolute and relative error because the absolute error differs so strongly between the different orientations of the shock front given by a.



(a) $a = 1$ (b) $a = 2$ (c) $a = 10^{-2}$

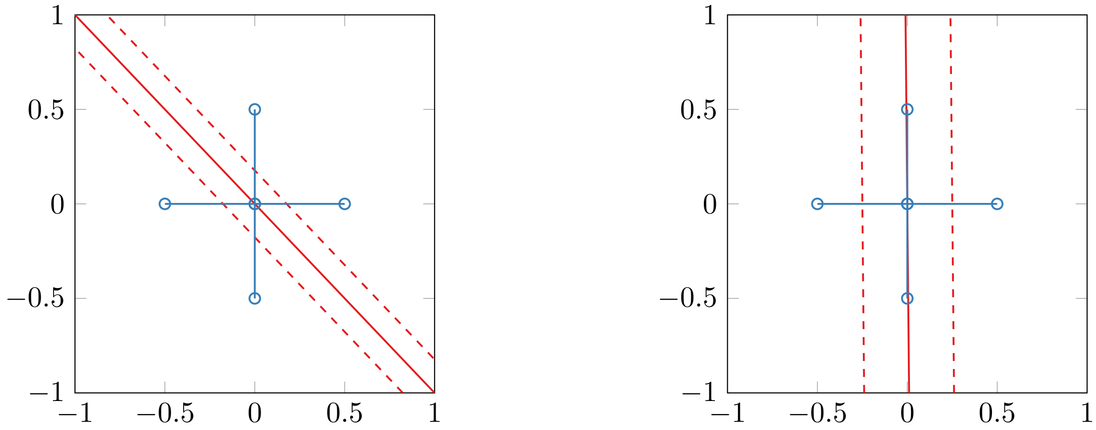
Figure 8.7: Shock sensitivity algorithm (1): Empirical convergence at $t_m = \frac{1}{2}$, $C = 1$, $\alpha = \frac{1}{2}$, $O(\Delta x^\alpha)$ (red), $|\dot{\Xi}_1^m - \dot{\xi}_1(t_m, z_0)|$ (solid blue), $|\dot{\Xi}_2^m - \dot{\xi}_2(t_m, z_0)|$ (solid green), $|\dot{\Xi}_1^m - \dot{\xi}_1(t_m, z_0)|/|\dot{\xi}_1(t_m, z_0)|$ (dashed blue), $|\dot{\Xi}_2^m - \dot{\xi}_2(t_m, z_0)|/|\dot{\xi}_2(t_m, z_0)|$ (dashed green)

For the symmetric and strongly nondegenerate case (with $|\partial_z \xi_1(t, z_0)| \gg 0$ and $|\partial_z \xi_2(t, z_0)| \gg 0$) of the shock front orientation given by $a = 1$ we observe a fast convergence of both the absolute and relative error of the shock sensitivity components at the same order of magnitude. That convergence is actually faster than the sublinear rate $O(\Delta x^\alpha)$ here is due to the fact that the weak solution is piecewise constant and, hence, the approximation error from being too far out from the shock region does not increase when approximating the left and right limits to the shock point of the weak solution and its tangent senstivities.

For $a = 2$ we observe a fast converging and low absolute and relative error for the second component of the shock sensitivity corresponding to the y axis that is more aligned with the direction of shock travel. But for the first component corresponding to the x axis both absolute and relative error are much higher. This error behavior is actually contrary to the $a = 2$ figure of Figure 8.5 where the first component had much lower error. The difference in absolute and relative error are explained by the phenomenon of overall lower absolute sensitivity values. But the difference between the components being such that the error for the first component is much higher is caused by how we approximate the limits to the shock on the axis corresponding to that component. Because the shock front is more aligned with the x axis the method converges slower for the shock sensitivity vector component corresponding to the x axis. The approximation of the limits takes much longer to leave the shock region with numerical viscosity.

A crucial insight from the empirical convergence analysis and the reason for Assumption 11 is the following: As long as the shock manifold is strongly nondegenerate the limit approximation leaves the numerical viscosity region in finite time and the shock sensitivity converges.

The effect of almost degeneracy can be observed for $a = 10^{-2}$ where the shock front is almost parallel to the y axis. The error of the second component of the shock sensitivity (in green) does not show any convergence for the range of Δx values that we consider. Figure 8.8 illustrates the issue of approximating the limits to the shock point at points that are at $(\Xi_1 \pm C\Delta x^\alpha, \Xi_2)$ and $(\Xi_1, \Xi_2 \pm C\Delta x^\alpha)$, i.e. on points where one of the x, y components is the same as the shock point.



(a) Strongly nondegenerate

(b) Almost degenerate

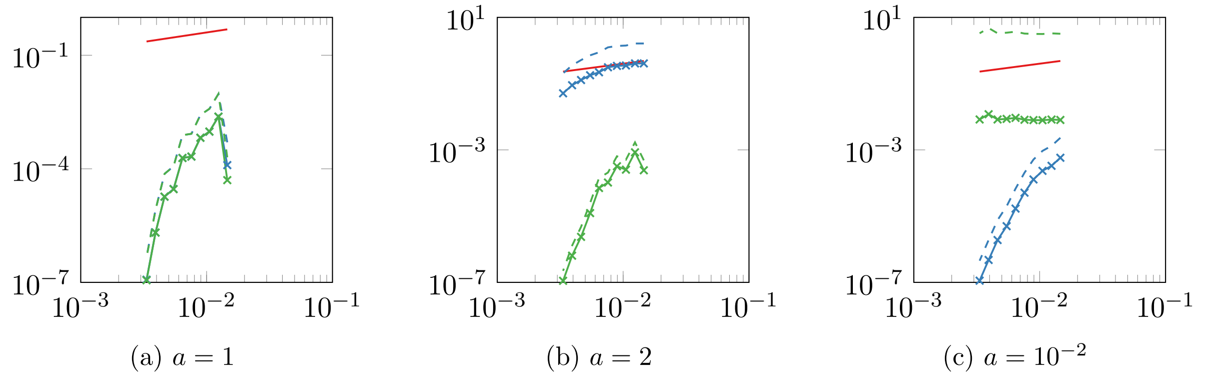
Figure 8.8: Visualization of the shock front degeneracy issue causing non-convergence

The solid red line illustrates the linear shock manifold and the dashed red lines illustrate the shock region in which the numerical weak solution contains numerical viscosity. The figure on the left shows the strongly nondegenerate case where we evaluate the limit approximations outside of the shock region. The figure on the right shows the almost degenerate case where we evaluate the limit approximations of one component inside the shock region unless we go far away from the shock. That is the behavior that effectively leads to the nonconvergence for $a = 10^{-2}$ in our example.

Figure 8.9 shows the empirical convergence results for the shock sensitivity algorithm (2) analogous to Figure 8.7. In the left-most figure for $a = 1$ the green and blue lines are almost identical and the green plot overrides the blue at most points.

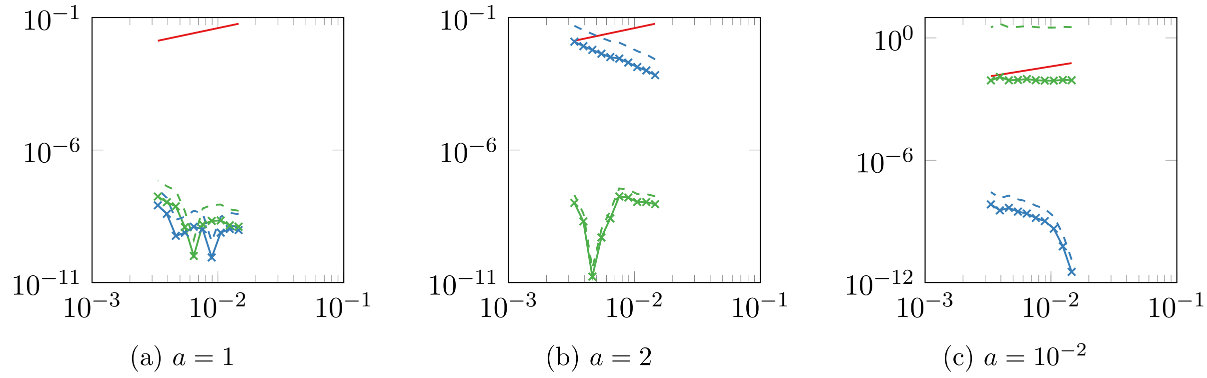
For the symmetric problem with $a = 1$ we have an almost identical solution and consequently almost identical approximation errors for the two shock sensitivity vector components just like we observed for the shock point approximation using shock location algorithm (2). Otherwise, the empirical convergence behavior is qualitatively the same for shock sensitivity algorithm (2) as for shock sensitivity algorithm (1) in Figure 8.7.

Figure 8.10 and Figure 8.11 show the empirical convergence results for shock sensitivity algorithms (1) and (2) respectively analogous to Figure 8.7 and Figure 8.9 but with $C = 20$ and $\alpha = 1$ instead. In the left-most figure of Figure 8.11 for $a = 1$ the green and blue lines are identical and plotted in green.



(a) $a = 1$ (b) $a = 2$ (c) $a = 10^{-2}$

Figure 8.9: Shock sensitivity algorithm (2): Empirical convergence at $t_m = \frac{1}{2}$ $C = 1$, $\alpha = \frac{1}{2}$, $O(\Delta x^\alpha)$ (red), $|\dot{\Xi}_1^m - \dot{\xi}_1(t_m, z_0)|$ (solid blue), $|\dot{\Xi}_2^m - \dot{\xi}_2(t_m, z_0)|$ (solid green), $|\dot{\Xi}_1^m - \dot{\xi}_1(t_m, z_0)|/|\dot{\xi}_1(t_m, z_0)|$ (dashed blue), $|\dot{\Xi}_2^m - \dot{\xi}_2(t_m, z_0)|/|\dot{\xi}_2(t_m, z_0)|$ (dashed green)



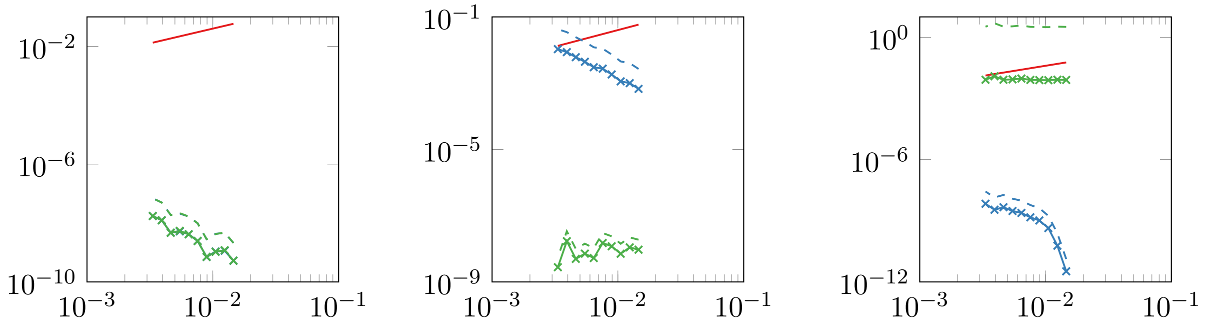
(a) $a = 1$ (b) $a = 2$ (c) $a = 10^{-2}$

Figure 8.10: Shock sensitivity algorithm (1): Empirical convergence at $t_m = \frac{1}{2}$ $C = 20$, $\alpha = 1$, $O(\Delta x^\alpha)$ (red), $|\dot{\Xi}_1^m - \dot{\xi}_1(t_m, z_0)|$ (solid blue), $|\dot{\Xi}_2^m - \dot{\xi}_2(t_m, z_0)|$ (solid green), $|\dot{\Xi}_1^m - \dot{\xi}_1(t_m, z_0)|/|\dot{\xi}_1(t_m, z_0)|$ (dashed blue), $|\dot{\Xi}_2^m - \dot{\xi}_2(t_m, z_0)|/|\dot{\xi}_2(t_m, z_0)|$ (dashed green)

For $\alpha = 1$ we actually observe the divergence of the shock sensitivity approximation error at rates that differ somewhat between the components for the different values of a. For $a = 1$ the approximation error starts out low because C is large enough but does not improve because the approximation does not move further out of the numerical viscosity region as $\Delta x \to 0$ with $\alpha = 1$. For $a = 2$ and $a = 10^{-2}$ we can see that it is crucial to have $\alpha < 1$. Considering the approximation structure illustrated in Figure 8.8 we see that in order to have a chance of moving out of the shock region even for the nondegenerate case where there is some alignment to the axis the distance of the evaluation points e.g. at $(\Xi_1, \Xi_2 \pm C\Delta x^\alpha)$ from the shock location point (Ξ_1, Ξ_2) needs to increase relative to the width of the shock region where numerical viscosity occurs.

Again, the only difference between shock sensitivity algorithm (1) and (2) with $\alpha = 1$ is that for the symmetric case of $a = 1$ the error in both shock sensitivity components is identical for shock sensitivity algorithm (2). Otherwise, the empirical convergence behavior is qualitatively the same.

We have demonstrated that the discrete sensitivity analysis of shock locations based on the Rankine-Hugoniot condition extends to points on a 1D shock manifold for 2D conservation laws.



(a) $a = 1$

(b) $a = 2$

(c) $a = 10^{-2}$

Figure 8.11: Shock sensitivity algorithm (2): Empirical convergence at $t_m = \frac{1}{2}$ $C = 20$, $\alpha = 1$, $O(\Delta x^{\alpha})$ (red), $|\dot{\Xi}_1^m - \dot{\xi}_1(t_m, z_0)|$ (solid blue), $|\dot{\Xi}_2^m - \dot{\xi}_2(t_m, z_0)|$ (solid green), $|\Xi_1^m - \dot{\xi}_1(t_m, z_0)|/|\dot{\xi}_1(t_m, z_0)|$ (dashed blue), $|\Xi_2^m - \dot{\xi}_2(t_m, z_0)|/|\dot{\xi}_2(t_m, z_0)|$ (dashed green)

When evaluating the Rankine-Hugoniot approximation on an orthogonal grid aligned with the spacial axes the method only works if the shock is not aligned with one of the axes, i.e. in order to have convergence in finite time we need to make a strong nondegeneracy assumption. In the next chapter we suggest an idea for further work that breaks away from the perspective of first-order splitting along the axes and removes the requirement for the nondegeneracy assumption.

Chapter 9

Conclusions and Further Work

The goal of this thesis was to derive a discrete tangent and adjoint sensitivity analysis for discontinuous solutions of hyperbolic conservation laws approximated by shock capturing numerical schemes. We proposed numerical methods for the simulation of shock locations and their tangent and adjoint sensitivities with respect to a parameter of the initial data.

Our proposed shock sensitivity approximations empirically converge for a convex and a non-convex flux function in one dimensional scalar conservation laws, for a convex flux function in a two dimensional scalar conservation law and in a system of hyperbolic conservation laws in computational experiments with two different shock capturing numerical schemes.

We also proposed numerical methods utilizing the numerical shock sensitivities to obtain the discrete adjoint sensitivies of integral objective functionals of the weak solution and to obtain a discrete generalized tangent approximation in the sense of Bressan and Marson [BM95b]. For one dimensional scalar conservation laws we proved the approximation error convergence rate $O(\Delta x^\alpha)$ with $0 < \alpha < 1$ for all of the proposed tangent and adjoint sensitivity approximations, as $\Delta x \to 0$, i.e. as the discretization grid is refined. For the convergence result we made the assumptions that the discontinuous weak solution is piecewise smooth and that the shock capturing numerical scheme is convergent at the same rate outside of the shock regions with a width at most $O(\Delta x^\alpha)$.

With a shock detection approach our proposed approach can be used when shocks are not yet present at $t = 0$ but only appear at a later time. The original initial value problem can then be separated into two initial value problems at the time of shock detection where the second initial value problem starts with discontinuous initial data to which the numerical methods proposed in this thesis apply.

Further work regarding the simulation of shock sensitivities for 2D conservation laws without the strong nondegeneracy assumption can be based on the idea that we do not have to evaluate the limits to the point on the shock discontinuity by moving along the axes. An alternative condition can be derived with limits taken in the direction of movement of the shock, i.e. moving outside of the shock region along the direction that is orthogonal to the shock manifold.

The extension of the numerical methods proposed in this thesis to optimal control problems and other more general parameterization is possible through standard AD approaches and poses another interesting challenge for further work.

Bibliography

[Hop50] E. Hopf. “The partial differential equation $u_t + uu_x = \mu_{xx}$”. In: *Communications on Pure and Applied mathematics* 3.3 (1950), pp. 201–230.

[Dou52] A. Douglis. “Some existence theorems for hyperbolic systems of partial differential equations in two independent variables”. In: *Communications on Pure and Applied Mathematics* 5.2 (1952), pp. 119–154.

[CL55] E. A. Coddington and N. Levinson. *Theory of ordinary differential equations.* Tata McGraw-Hill Education, 1955.

[Lax57] P. D. Lax. “Hyperbolic systems of conservation laws II”. In: *Communications on pure and applied mathematics* 10.4 (1957), pp. 537–566.

[Ole57] O. A. Oleinik. “Discontinuous solutions of non-linear differential equations”. In: *Uspekhi Matematicheskikh Nauk* 12.3 (1957), pp. 3–73.

[Rud64] W. Rudin. *Principles of mathematical analysis.* Vol. 3. McGraw-hill New York, 1964.

[CS66] E. Conway and J. Smoller. “Global solutions of the Cauchy problem for quasi-linear first-order equations in several space variables”. In: *Communications on Pure and Applied Mathematics* 19.1 (1966), pp. 95–105.

[Bal70] D. P. Ballou. “Solutions to nonlinear hyperbolic Cauchy problems without convexity conditions”. In: *Transactions of the American Mathematical Society* 152.2 (1970), pp. 441–460.

[Kru70] S. N. Kružkov. “First order quasilinear equations in several independent variables”. In: *Mathematics of the USSR-Sbornik* 10.2 (1970), p. 217.

[Sch73] D. G. Schaeffer. “A regularity theorem for conservation laws”. In: *Advances in Mathematics* 11.3 (1973), pp. 368–386.

[Bau74] F. L. Bauer. “Computational graphs and rounding error”. In: *SIAM Journal on Numerical Analysis* 11.1 (1974), pp. 87–96.

[Jen74] G. Jennings. “Discrete shocks”. In: *Communications on pure and applied mathematics* 27.1 (1974), pp. 25–37.

[Daf77] C. M. Dafermos. “Generalized characteristics and the structure of solutions of hyperbolic conservation laws”. In: *Indiana University Mathematics Journal* 26.6 (1977), pp. 1097–1119.

[DiP79] R. J. DiPerna. “Uniqueness of solutions to hyperbolic conservation laws”. In: *Indiana University Mathematics Journal* 28.1 (1979), pp. 137–188.

[Ito79] T. Ito. “A Filippov solution of a system of differential equations with discontinuous right-hand sides”. In: *Economics Letters* 4.4 (1979), pp. 349–354.

[CM80] M. G. Crandall and A. Majda. “Monotone difference approximations for scalar conservation laws”. In: *Mathematics of Computation* 34.149 (1980), pp. 1–21.

[Col85] J. F. Colombeau. *Elementary introduction to new generalized functions.* Vol. 113. 1985.

[Gli+85] J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv. “Front tracking and two-dimensional Riemann problems”. In: *Advances in Applied Mathematics* 6.3 (1985), pp. 259–290.

[Hal87] P. R. Halmos. *Finite-dimensional vector spaces.* Springer, 1987.

[Fil88] A. F. Filippov. *Differential equations with discontinuous righthand sides.* Kluwer Academic Publishers, 1988.

[Pre+88] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C.* 1988.

[DL92] A. Dontchev and F. Lempio. “Difference methods for differential inclusions: a survey”. In: *SIAM review* 34.2 (1992), pp. 263–294.

[LeV92] R. J. LeVeque. *Numerical methods for conservation laws.* Vol. 132. Springer, 1992.

[BM95a] A. Bressan and A. Marson. “A maximum principle for optimally controlled systems of conservation laws”. In: *Rendiconti del Seminario Matematico della Universita di Padova* 94 (1995), pp. 79–94.

[BM95b] A. Bressan and A. Marson. “A variational calculus for discontinuous solutions of systems of conservation laws”. In: *Communications in partial differential equations* 20.9 (1995), pp. 1491–1552.

[DLM95] G. Dal Maso, P. G. Lefloch, and F. Murat. “Definition and weak stability of nonconservative products”. In: *Journal de mathematiques pures et appliquees* 74.6 (1995), pp. 483–548.

[Jam95] A. Jameson. “Optimum aerodynamic design using control theory”. In: *Computational Fluid Dynamics Review* 3 (1995), pp. 495–528.

[CLT96] G.-Q. Chen, D. Li, and D. Tan. “Structure of Riemann solutions for 2-dimensional scalar conservation laws”. In: *Journal of differential equations* 127.1 (1996), pp. 124–147.

[GJU96] A. Griewank, D. Juedes, and J. Utke. “Algorithm 755: ADOL-C: a package for the automatic differentiation of algorithms written in C/C++”. In: *ACM Transactions on Mathematical Software (TOMS)* 22.2 (1996), pp. 131–167.

[IS96] A. Iollo and M. D. Salas. “Contribution to the optimal shape design of two-dimensional internal flows with embedded shocks”. In: *Journal of Computational Physics* 125.1 (1996), pp. 124–134.

[BG97] A. Bressan and G. Guerra. “Shift-differentiabilitiy of the flow generated by a conservation law”. In: *Discrete & Continuous Dynamical Systems-A* 3.1 (1997), p. 35.

[CHS97] E. M. Cliff, M. Heinkenschloss, and A. R. Shenoy. “An optimal control problem for flows with discontinuities”. In: *Journal of Optimization Theory and Applications* 94.2 (1997), pp. 273–309.

[Sab97] F. Sabac. “The optimal convergence rate of monotone finite difference methods for hyperbolic conservation laws”. In: *SIAM journal on numerical analysis* 34.6 (1997), pp. 2306–2318.

[TT97] T. Tang and Z.-H. Teng. “Viscosity methods for piecewise smooth solutions to scalar conservation laws”. In: *Mathematics of computation* 66.218 (1997), pp. 495–526.

[TZ97] Z.-H. Teng and P. Zhang. “Optimal L1-rate of convergence for the viscosity method and monotone scheme to piecewise constant solutions with shocks”. In: *SIAM journal on numerical analysis* 34.3 (1997), pp. 959–978.

[BJ98] F. Bouchut and F. James. “One-dimensional transport equations with discontinuous coefficients”. In: *Nonlinear Analysis* 32.7 (1998), p. 891.

[CHS98] E. M. Cliff, M. Heinkenschloss, and A. R. Shenoy. “Adjoint-based methods in aerodynamic design-optimization”. In: *Computational methods for optimal design and control.* Springer, 1998, pp. 91–112.

[Eva98] L. C. Evans. *Partial differential equations.* American Mathematical Society, 1998.

[BJ99] F. Bouchut and F. James. “Differentiability with Respect to Initial Data for a Scalar Conservation Law”. In: *Hyperbolic Problems: Theory, Numerics, Applications; Seventh International Conference in Zürich, February 1998.* Vol. 1. Springer Science & Business Media. 1999, p. 113.

[Bre99] A. Bressan. “Hyperbolic systems of conservation laws”. In: *Revista matemática complutense* 12.1 (1999), pp. 135–200.

[GR99] E. Godlewski and P.-A. Raviart. “The linearized stability of solutions of nonlinear hyperbolic systems of conservation laws: A general numerical approach”. In: *Mathematics and Computers in Simulation* 50.1-4 (1999), pp. 77–95.

[Bre00] A. Bressan. *Hyperbolic systems of conservation laws: the one-dimensional Cauchy problem.* Vol. 20. Oxford University Press on Demand, 2000.

[DHV00] A. L. Dontchev, W. W. Hager, and V. M. Veliov. “Second-order Runge–Kutta approximations in control constrained optimal control”. In: *SIAM Journal on Numerical Analysis* 38.1 (2000), pp. 202–226.

[GP01] M. B. Giles and N. A. Pierce. “Analytic adjoint solutions for the quasi-one-dimensional Euler equations”. In: *Journal of Fluid Mechanics* 426 (2001), p. 327.

[LeV02] R. J. LeVeque. *Finite volume methods for hyperbolic problems.* Vol. 31. Cambridge university press, 2002.

[MQ02] R. I. McLachlan and G. R. W. Quispel. “Splitting methods”. In: *Acta Numerica* 11 (2002), p. 341.

[Ulb02] S. Ulbrich. “A sensitivity and adjoint calculus for discontinuous solutions of hyperbolic conservation laws with source terms”. In: *SIAM journal on control and optimization* 41.3 (2002), pp. 740–797.

[BP03] C. Bardos and O. Pironneau. “Derivatives and control in the presence of shocks”. In: *Computational Fluid Dynamics Journal* 11.4 (2003), pp. 383–391.

[Gil03] M. B. Giles. “Discrete adjoint approximations with shocks”. In: *Hyperbolic problems: theory, numerics, applications.* Springer, 2003, pp. 185–194.

[Tra03] J. F. Traub. *Information-based complexity.* John Wiley and Sons Ltd., 2003.

[Ulb03] S. Ulbrich. “Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws”. In: *Systems & Control Letters* 48.3-4 (2003), pp. 313–328.

[GGD05] M. B. Giles, D. Ghate, and M. C. Duta. “Using automatic differentiation for adjoint CFD code development”. In: (2005).

[MC05] J.-D. Müller and P. Cusdin. “On the performance of discrete adjoint CFD codes using automatic differentiation”. In: *International journal for numerical methods in fluids* 47.8-9 (2005), pp. 939–945.

[DR06] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler.* Springer-Verlag, 2006.

[Wal07] A. Walther. “Automatic differentiation of explicit Runge-Kutta methods for optimal control”. In: *Computational Optimization and Applications* 36.1 (2007), pp. 83–108.

[Cor08] J. Cortes. “Discontinuous dynamical systems”. In: *IEEE Control systems magazine* 28.3 (2008), pp. 36–73.

[GW08] A. Griewank and A. Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation.* Vol. 105. Siam, 2008.

[Hig08] N. J. Higham. *Functions of matrices: theory and computation.* SIAM, 2008.

[Bre09] A. Bressan. “Lecture notes on hyperbolic conservation laws”. In: *Department of mathematics, Penn State University, University park* (2009), pp. 1–85.

[DL09] L. Dieci and L. Lopez. “Sliding motion in Filippov differential systems: theoretical results and a computational approach”. In: *SIAM Journal on Numerical Analysis* 47.3 (2009), pp. 2023–2051.

[Bet10] J. T. Betts. *Practical methods for optimal control and estimation using nonlinear programming.* Vol. 19. Siam, 2010.

[GZ10] S. J. Galiano and M. U. Zapata. “A new TVD flux-limiter method for solving nonlinear hyperbolic equations”. In: *Journal of computational and applied mathematics* 234.5 (2010), pp. 1395–1403.

[Gei10] J. Geiser. “Consistency of iterative operator-splitting methods: Theory and applications”. In: *Numerical Methods for Partial Differential Equations: An International Journal* 26.1 (2010), pp. 135–158.

[GU10a] M. B. Giles and S. Ulbrich. “Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 1: Linearized approximations and linearized output functionals”. In: *SIAM Journal on numerical analysis* 48.3 (2010), pp. 882–904.

[GU10b] M. B. Giles and S. Ulbrich. “Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoint approximations and extensions”. In: *SIAM Journal on Numerical Analysis* 48.3 (2010), pp. 905–921.

[Han+10] R. Hannemann, W. Marquardt, U. Naumann, and B. Gendler. “Discrete first-and second-order adjoints and automatic differentiation for the sensitivity analysis of dynamic models”. In: *Procedia Computer Science* 1.1 (2010), pp. 297–305.

[SA10] D. E. Stewart and M. Anitescu. “Optimal control of systems with discontinuous differential equations”. In: *Numerische Mathematik* 114.4 (2010), pp. 653–695.

[JMC11] D. Jones, J.-D. Müller, and F. Christakopoulos. “Preparation and assembly of discrete adjoint CFD codes”. In: *Computers & Fluids* 46.1 (2011), pp. 282–286.

[LLN11] J. Lotz, K. Leppkes, and U. Naumann. “dco/c++-derivative code by overloading in C++”. In: *Aachener Informatik Berichte (AIB-2011-06)* (2011).

[SH11] I. Stakgold and M. J. Holst. *Green’s functions and boundary value problems.* Vol. 99. John Wiley & Sons, 2011.

[Fra12] J. N. Franklin. *Matrix theory.* Courier Corporation, 2012.

[Nau12] U. Naumann. *The art of differentiating computer programs: an introduction to algorithmic differentiation.* Vol. 24. Siam, 2012.

[SA12] H. Schiller and M. Arnold. “Convergence of continuous approximations for discontinuous ODEs”. In: *Applied Numerical Mathematics* 62.10 (2012), pp. 1503–1514.

[Bre13] A. Bressan. “Hyperbolic conservation laws: an illustrated tutorial”. In: *Modelling and optimisation of flows on networks.* Springer, 2013, pp. 157–245.

[Che13] W. Cheney. *Analysis for applied mathematics.* Vol. 208. Springer Science & Business Media, 2013.

[GR13] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws.* Vol. 118. Springer Science & Business Media, 2013.

[GV13] G. H. Golub and C. F. Van Loan. *Matrix computations.* Fourth. JHU Press, 2013.

[Gra13] J. Gratus. “Colombeau Algebra: A pedagogical introduction”. In: *arXiv preprint arXiv:1308.0257* (2013).

[HP13] L. Hascoet and V. Pascual. “The Tapenade automatic differentiation tool: Principles, model, and specification”. In: *ACM Transactions on Mathematical Software (TOMS)* 39.3 (2013), pp. 1–43.

[WK13] Y. Wang and C.-Y. Kao. “Central schemes for the modified Buckley–Leverett equation”. In: *Journal of Computational Science* 4.1-2 (2013), pp. 12–23.

[STN14] A. Sen, M. Towara, and U. Naumann. “A discrete adjoint version of an unsteady incompressible solver for OpenFOAM using algorithmic differentiation”. In: *6th European Conference on Computational Fluid Dynamics, 5014.* Vol. 5023. 2014.

[MS16] S. MacNamara and G. Strang. “Operator splitting”. In: *Splitting methods in communication, imaging, science, and engineering.* Springer, 2016, pp. 95–114.

[Sch+16] T. Schilden, W. Schröder, S. R. C. Ali, A.-M. Schreyer, J. Wu, and R. Radespiel. “Analysis of acoustic and entropy disturbances in a hypersonic wind tunnel”. In: *Physics of Fluids* 28.5 (2016), p. 056104.

[Sch+19] P. Schäfer Aguilar, J. M. Schmitt, S. Ulbrich, and M. Moos. “On the numerical discretization of optimal control problems for conservation laws”. In: *Control & Cybernetics* 48.2 (2019).

[SS19] T. Schilden and W. Schröder. “Inclined slow acoustic waves incident to stagnation point probes in supersonic flow”. In: *Journal of Fluid Mechanics* 866 (2019), pp. 567–597.

[TLN19] M. Towara, J. Lotz, and U. Naumann. “Discrete Adjoint Approaches for CHT Applications in OpenFOAM”. In: *Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences.* Springer, 2019, pp. 163–178.

[CHA20] J. I. Cardesa, L. Hascoët, and C. Airiau. “Adjoint computations by algorithmic differentiation of a parallel solver for time-dependent PDEs”. In: *Journal of computational science* 45 (2020), p. 101155.

[Her+21] M. Herty, J. Hüser, U. Naumann, T. Schilden, and W. Schröder. “Algorithmic differentiation of hyperbolic flow problems”. In: *Journal of Computational Physics* 430 (2021), p. 110110.