Adjoint-based sensitivity for flows with shocks

S. H. Bryngelson

D. J. Bodony, J. B. Freund, C. A. Pantano

The Center for Exascale Simulation of Plasma-coupled Combustion
University of Illinois at Urbana–Champaign

Developing a consistent solution method for discontinuous adjoint flow problems is challenging. The inviscid Burgers' equation is considered as a step toward the formulation of such a method. Results are shown for linear and nonlinear numerical methods, including WENO shock capturing. Necessary and sufficient conditions for consistent and convergent solutions to the associated adjoint equation are discussed. The adjoint Euler equations are also investigated, for which no numerical schemes are provably convergent. A characteristic-based method for this problem is proposed. It transforms the adjoint equations into an uncoupled set of transport equations. These equations have the same form as the adjoint Burgers' equation, and thus inherit their proven consistency properties.

Contents

1 Introduction

1.1 Adjoint techniques

Adjoint methods determine the local sensitivity of an equations to its variables. They significantly reduce the complexity of optimization by deriving adjoint equations through partial-differential-equation-based constraints associated with the governing equations. This bypasses direct differentiation of the governing equations. Adjoint methods have been used for flow sensitivity analysis and design optimization for a long time, though typically for smooth flow variables. Flows with discontinuities like shocks, boundary layers, or reactive zones pose problems for adjoint methods. These problems are generally associated with differentiability at the discontinuity.

1.2 Flows with shocks

Hyperbolic conservation laws take the form

$$\frac{\partial \boldsymbol{u}}{\partial t} + \nabla \cdot \boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{0}. \tag{1}$$

Such equations can form discontinuities for even smooth initial data. Herein, two conservation laws with this property are considered, the inviscid Burgers' equation:

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = 0, \tag{2}$$

and the one-dimensional compressible Euler equations:

$$\frac{\partial \boldsymbol{u}}{\partial t} + \frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial x} = 0. \tag{3}$$

Shock fitting or shock capturing are usually used to solve the associated forward equations.

1.2.1 Shock fitting

Shock fitting methods explicitly introduce the shock into the solution via a Rankine–Hugoniot relation. The shock locations are an independent variable appended to the existing set of equations and treated as a surface of discontinuity. Perturbations of a discontinuous function produce a delta function, and all formulations based on variations of smooth functions (e.g. Taylor series) must be modified. This information can also be used in specialized descent algorithms. Unfortunately, this procedure is not practical for complex multidimensional flow problems (Matsuzawa and Hafez, 1997).

1.2.2 Shock capturing

Shock capturing methods express the governing equations in conservation form and evolve the discontinuities as part of the solution (in terms of a flux). No special methods are used to identify shocks. Instead, shock capturing methods attempt to damp spurious oscillations near large jumps in the solution. The most common such methods are embedded in a finite volume framework, which discretely conserves the state variables.

1.3 Adjoint solutions for flows with shocks

Adjoint equations are adjoint to the equations obtained by linearizing the nonlinear flow equations. However, linearizing about a shock is ill-posed without additional information (e.g. a Rankine–Hugoniot relation). Shock fitting methods can do this directly, though the procedure is not always trivial. Shock capturing methods do not entail a separate equation for the shock surface, so it is unclear how to impose an additional constraint.

1.3.1 Convergence via shock smearing

Giles and Ulbrich (2010a) showed that a modified Lax–Friedrichs scheme, which artificially adds dissipation as the grid is refined, is convergent and consistent for the linearized and adjoint Burgers' equation. They also asserted that artificial dissipation is required for convergence generally (Giles, 2002; Giles and Ulbrich, 2010a,b). Such dissipation could ensure that the shock is smeared across an increasing number of grid points as the grid is refined. Thus, the numerical method implicitly satisfies the adjoint interior boundary conditions at the shock location. For example, Giles and Ulbrich (2010a) showed that applying a modified Lax–Friedrichs scheme to Burgers' equation can converge to incorrect solutions if the dissipation decreases too slowly (or quickly) with grid refinement:

$$f_{i+1/2} = \frac{1}{2}(f_{i+1} + f_i) - \mu(u_{i+1} - u_i) \quad \text{where} \quad \mu = \Delta x^{\alpha-1},\ \alpha \in (2/3, 1). \tag{4}$$

Using the discrete linearized and adjoint schemes ensures that the conservative linearized equations are discretized by a conservative scheme and the nonconservative adjoint equations by a nonconservative scheme.

Note that the shock is smeared at multiple points: in the forward and reverse governing equations as a part of the numerical method, in evaluation of the cost function, in the updating of design variables, and in updating the spatial grid, if necessary, at each design iteration. It can be shown that this smearing is sometimes equivalent to modifying the cost function, and for transonic flows with weak shocks acceptable results can be obtained, as have been in the past. This is also true for optimization of integral quantities, such as lift and drag, as it has been argued that the "viscous" shock profile remains invariant, to leading order, as the shock oscillates, and the linearized shock motion remains correct.

1.3.2 Nonlinear flux-limited schemes

The proof of Giles and Ulbrich (2010a) is difficult to extend to more nonlinear numerical schemes. For example, flux limited schemes, such as MUSCL and WENO, are not proven to converge to the correct adjoint solution. This is because one of the main requirements in convergence proofs for nonlinear conservation laws with shocks is Oleinik's One-Sided Lipschitz Condition (OSLC):

$$\frac{u(x,t) - u(y,t)}{x - y} \leq \frac{1}{t}$$

and its discrete equivalent. The only flux limiter that has been proven to satisfy the OSLC is the `maxmod` limiter. Intrinsically nonlinear schemes do not involve artificial dissipation in the same way as an entropy solution and, as such, do not necessarily satisfy the OSLC. While these schemes can be modified in such a way that they satisfy the OSCL, this makes the limiter itself overly complicated (both in implementation and overhead cost), and thus are not practical. While there are no formal

convergence proofs, (Messyahu et al., 1994) suggests that TVD MUSCL schemes are bounded in Lip^+, where:

$$\|w(x)\|_{Lip^+} \equiv \operatorname*{ess\,sup}_{x \neq y} \left(\frac{w(x) - w(y)}{x - y} \right)^+, \quad (\cdot)^+ \equiv \max(\cdot, 0), \tag{5}$$

and though they are not monotonically decreasing as required by OSLC, they do converge in certain cases. Note that these cannot converge better than first order when there is a discontinuity inside a grid cell. When shocks are present, formally second-order schemes are only first order accurate in Lip'. This can be surpassed by moving the mesh so that no shock occurs in the interior of a grid cell. Then, a better error estimate will hold, and if the scheme is Lip^+-stable, will be second order accurate:

$$\|(RA - I)w\|_{Lip'} = \mathcal{O}(\Delta x^3)\|w\|_{BV} \tag{6}$$

whenever w is C^1 in the grid cell interior. Here, R is the reconstruction operator, A is the cell averaging operator, and BV is the bounded variation space:

$$\text{Continuously differentiable} \subseteq \text{Lipschitz continuous} \subseteq$$
$$\text{absolutely continuous} \subseteq \text{bounded variation} \subseteq$$
$$\text{differentiable almost everywhere.}$$

1.3.3 Non-differentiable methods

Non-differentiable optimization methods, such as those using sub-gradients, have also been used. For example, methods that do not involve gradient information, such as stochastic optimization or genetic algorithms will converge to the correct sensitivity. However, these methods involve very high computational costs compared to classical gradient descent methods. Thus, they appear intractable for large flow simulations.

2 Adjoint formulation

There are two main approaches to forming the adjoint; the differentiate-then-discretize approach, which derives the continuous adjoint equations analytically from the state equations (e.g. (26)) and then discretizes them, and the discretize-then-differentiate approach, which discretizes the state equations then derives the discrete adjoint equations based on the discretized operators. The formulation of the adjoint system is shown first.

2.1 Sensitivity: Burgers' equation

We write a non-homogeneous (26) as:

$$\mathcal{N}(u, g) = u_t + F(u)_x - g(x) = 0, \tag{7}$$

where $F(u) = u^2/2$ is the flux and $g(x)$ is a source term and our parameter of interest. The adjoint is formed as a constrained optimization problem:

$$\min_{u,g} \mathcal{J}(u, g) \quad \text{subject to} \quad \mathcal{N}(u, g) = 0, \tag{8}$$

4

where $\mathcal{J}$ is the cost functional and $\langle\cdot,\cdot\rangle$ is an inner product. Following standard practice the L_2 inner product is used:

$$\langle x,y\rangle = \int_\Omega xy\,\mathrm{d}\Omega. \tag{9}$$

We assume that the cost functional has a unique minimizer. The constrained optimization problem is transformed into an unconstrained problem via a Lagrangian functional

$$\mathcal{L}(u,g,p) = \mathcal{J}(u,g) - \langle p, \mathcal{N}(u,g)\rangle, \tag{10}$$

where p is the adjoint variable (or Lagrange multiplier). It is introduced to enforce $\mathcal{N}(u,g) = 0$. We then seek to minimize $\mathcal{L}$ instead, which is satisfied when all the corresponding partial derivatives simultaneously vanish:

$$0 = \frac{\partial\mathcal{L}}{\partial p} = -\mathcal{N}(u,g), \tag{11}$$

$$0 = \frac{\partial\mathcal{L}}{\partial u} = \frac{\partial\mathcal{J}}{\partial u} - \left(\frac{\partial\mathcal{N}}{\partial u}\right)^\dagger p, \tag{12}$$

$$0 = \frac{\partial\mathcal{L}}{\partial g} = \frac{\partial\mathcal{J}}{\partial g} - \left(\frac{\partial\mathcal{N}}{\partial g}\right)^\dagger p. \tag{13}$$

Here, (11) is the constraint, (12) is the adjoint system, and (13) determines the optimality condition. The adjoint of an operator D, denoted by $D^\dagger$, is

$$\langle Dq, q^\dagger\rangle = \langle q, D^\dagger q^\dagger\rangle + b, \tag{14}$$

or

$$\underbrace{\int_\Omega (Dq)q^\dagger\mathrm{d}x}_{(a)} = \int_\Omega q(D^\dagger q)\mathrm{d}x + qq^\dagger\Big|_{x=0}^{L} \tag{15}$$

where (a) is evaluated and $D^\dagger$ and b are inferred by inspection. Traditionally $b = 0$ is chosen for simplicity. The sensitivity is given by the chain rule

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}g} = \frac{\partial\mathcal{J}}{\partial g} + \frac{\partial\mathcal{J}}{\partial u}\frac{\mathrm{d}u}{\mathrm{d}g}. \tag{16}$$

Differentiating the constraint $\mathcal{N}(u,g) = 0$ gives

$$\frac{\mathrm{d}u}{\mathrm{d}g} = \left(\frac{\partial\mathcal{N}}{\partial u}\right)^{-1}\frac{\partial\mathcal{N}}{\partial g}, \tag{17}$$

and then

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}g} = \frac{\partial\mathcal{J}}{\partial g} - \left(\frac{\partial\mathcal{N}}{\partial g}\right)^\dagger p. \tag{18}$$

For our purposes we will the cost functional

$$\mathcal{J} = \int_\Omega \frac{u(x,T)^2}{2}\,\mathrm{d}x \quad\rightarrow\quad \frac{\partial\mathcal{J}}{\partial u} = \int_\Omega u(x,T)\,\mathrm{d}x. \tag{19}$$

5

We also have

$$\frac{\partial \mathcal{N}}{\partial g} = 1 \quad \text{and} \quad \frac{\partial \mathcal{J}}{\partial g} = 0. \tag{20}$$

Thus, the adjoint system to be solved is

$$\left(\frac{\partial \mathcal{N}}{\partial u}\right)^{\dagger} p = 0 \tag{21}$$

with $p(x, T) = u(x, T)$ and the sensitivity is given by

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}g} = -\langle 1, p \rangle. \tag{22}$$

2.2 Differentiate-then-discretize

We primarily focus on the differentiate-then discretize approach, which writes the continuous adjoint equations per (21) then solves them using separate numerical methods. For (7) we have

$$\left(\frac{\partial \mathcal{N}}{\partial u}\right)^{\dagger} = (\partial_t + u\partial_x), \tag{23}$$

which is derived from the definition of the adjoint and integration by parts, where we have taken $b = 0$ and thus, for periodic boundaries $u(0, t) = u(L, t)$ the jump must match $p(0, t) = p(L, t)$. The adjoint system can then by written in continuous form as

$$p_t + u p_x = 0 \tag{24}$$

with boundary and initial (final) conditions:

$$p(0, t) = p(L, t), \ p(x, T) = u(x, T). \tag{25}$$

This system (24) can be integrated backwards in time, starting from $t = T$, with whatever numerical scheme is appropriate. Once $p(x, t)$ is solved, the sensitivity (22) is straightforward.

3 Example problem: Burgers' equation and shock initial data

3.1 Problem setup

The problem is the inviscid Burgers equation

$$u(x, t)_t + \left(\frac{u(x, t)^2}{2}\right)_x = 0, \quad x \in \Omega; \ t \in [0, T], \tag{26}$$

where $\Omega = [-2, 2]$ and $T = 1$, and we use initial condition

$$\text{(i)} \ u(x, 0) = \begin{cases} 1.5 & \text{for } x < 0 \\ -0.5 & \text{for } x \geq 0 \end{cases} \quad \text{or} \quad \text{(ii)} \ u(x, 0) = \begin{cases} 0.75 & \text{for } x < 0 \\ 0.25 & \text{for } x \geq 0. \end{cases} \tag{27}$$

The exact solution for $u(x, t)$ is advection of the shock at the shock speed, given by the Rankine–Hugoniot condition as the jump in the flux at the shock:

$$\dot{x}_s [\![u]\!] = [\![f(u)]\!] = \frac{u_L + u_R}{2} = u_s, \tag{28}$$

which gives

$$u(x, t) = \begin{cases} u_L, & x < u_s t \\ u_R, & x > u_s t \end{cases} \tag{29}$$

for both initial conditions.

3.2 Exact adjoint solution

Linear perturbations to the governing equation evolve as

$$L(u)\widetilde{u} \equiv \frac{\partial \widetilde{u}}{\partial t} + \frac{\partial}{\partial x} (a\widetilde{u}) = 0, \tag{30}$$

where $a(u) \equiv f'(u) = u$. The corresponding perturbation to the shock position satisfies the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} (\widetilde{x}_s [\![u]\!]) = [(a - \dot{x}_s)\widetilde{u}]. \tag{31}$$

We can separate the cost functional as

$$\mathcal{J} = \int_{\Omega \backslash \Gamma} G(u(x, T)) \mathrm{d}x \tag{32}$$

where $G(u) = u^2/2$ and Γ is the shock surface. The linear perturbation to the cost functional is

$$\widetilde{\mathcal{J}} = \int_{\Omega \backslash \Gamma} G'(u(x, T))\widetilde{u}(x, T) \mathrm{d}x - \widetilde{x}_s(T)[\![G]\!](T), \tag{33}$$

$$= \int_{\Omega \backslash \Gamma} p(x, 0)\widetilde{u}(x, 0) \mathrm{d}x, \tag{34}$$

where $p(x, t)$ satisfies

$$p_t + u p_x = 0 \quad \text{for} \quad x \in \Omega \backslash \Gamma \tag{35}$$

$$p(x_s(t), t) = \frac{[\![G]\!]}{[\![u]\!]}(x, T) \quad \text{for} \quad x \in \Gamma. \tag{36}$$

where the latter is called the interior boundary condition. In this case, the interior boundary condition for the adjoint equations matches the Rankine–Hugoniot condition for the forward equations at time T.

The exact forward and adjoint solutions are shown in figure 1. For time $t = 0$ the exact adjoint solution is

$$p(x, 0) = \begin{cases} u_L, & x < T\left[(u_L + u_R)/2 - u_L\right], \\ (u_L + u_R)/2, & T\left[(u_L + u_R)/2 - u_L\right] < x < T\left[(u_L + u_R)/2 - u_R\right], \\ u_R, & x > T\left[(u_L + u_R)/2 - u_R\right]. \end{cases} \tag{37}$$
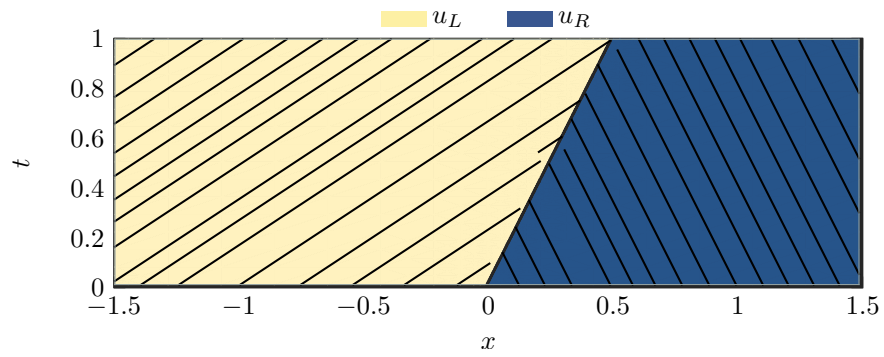
(a) Forward solution: $u(x, t)$
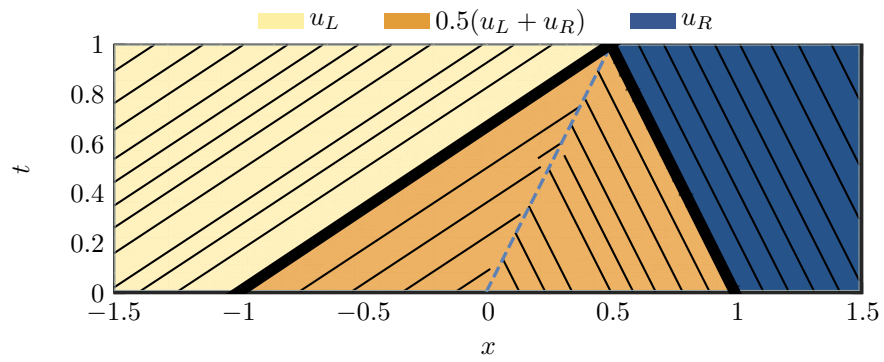
(b) Adjoint solution: $p(x, t)$

Figure 1: Characteristics for initial condition (i).

3.3 Results

3.3.1 Discrete-adjoint and upwinding flux

We first use a conservative upwinding scheme in space and forward Euler in time, following (Alauzet and Pironneau, 2012). For the forward variable this gives

$$\frac{u_i^{m+1} - u_i^m}{\Delta t} + s_i \frac{u_i^{m\,2} - u_{i-1}^{m\,2}}{2\Delta x} + (1 - s_{i+1}) \frac{u_{i+1}^{m\,2} - u_i^{m\,2}}{2\Delta x} = 0 \tag{38}$$

for $i = 1, \ldots, N$ are the uniformly spaced spatial points of step size Δx, $m = 1, \ldots, M$ the uniform time steps of size Δt, and the CFL number is $\lambda = \Delta t / \Delta x$. The s_i are the upwind indicators, which have

$$s_i = 1 \quad \text{if} \quad u_i^m > u_{i-1}^m, \ \ 0 \ \text{otherwise.}, \tag{39}$$

the discrete cost function is

$$\mathcal{J}^h = \Delta x \sum_{i=1}^{N} G(x_i, T), \tag{40}$$

and the time-consistent discrete adjoint is

$$\frac{p_i^{m-1} - p_i^m}{\Delta t} + \frac{u_i^m}{\Delta x} \left[s_i p_i^m - s_{i+1} p_{i+1}^m + (1 - s_i) p_{i-1}^m - (1 - s_{i+1}) p_i^m \right] = 0 \tag{41}$$

with final condition $p_i^M = u_i^M$.
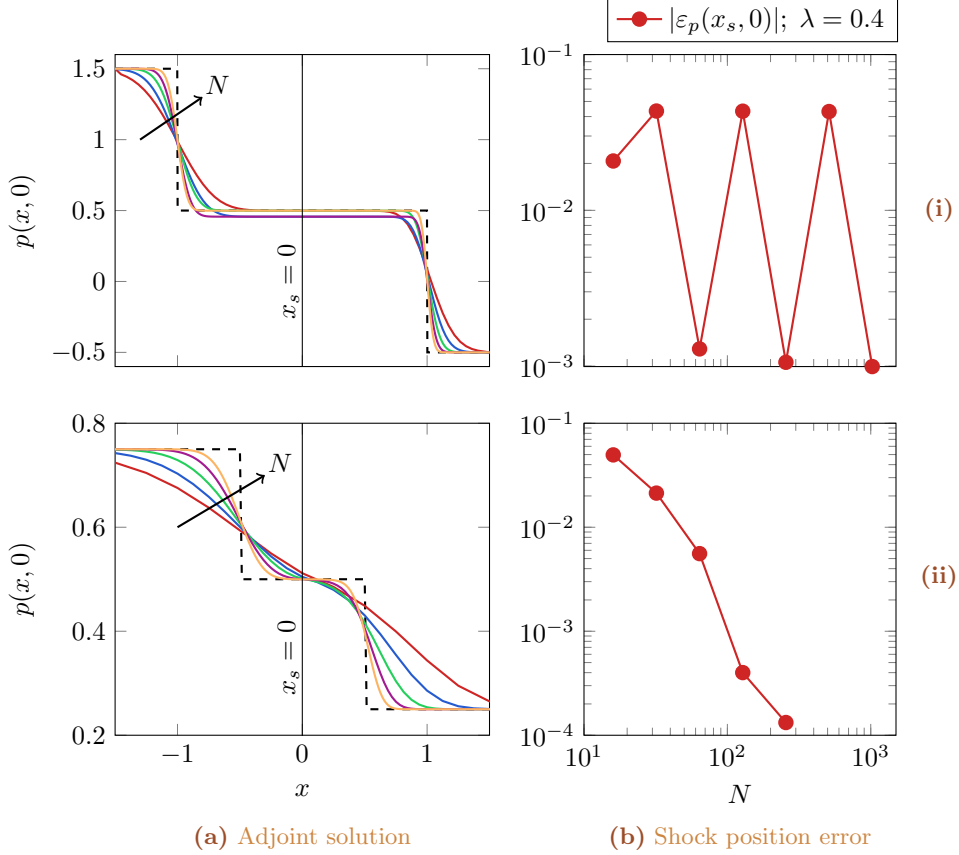
Figure 2: (a) Adjoint solution and − − − exact solution, (b) shock error for both initial conditions (i) and (ii) and the upwinding flux scheme. For initial condition (i) no convergence is seen at the shock location for increasing spatial resolution N. Convergence is clear for initial condition (ii). Both cases have the same shock speed u_s.

N	$|\varepsilon_p(x_s,0)|$	PAS	N	$|\varepsilon_p(x_s,0)|$	PAS
16	0.021	2	16	0.050	6
32	0.043	2	32	0.021	7
64	0.0013	1	64	0.0056	7
128	0.043	2	128	0.00042	9
256	0.0011	2	256	0.00013	10
(a) Initial condition (i)			**(b)** Initial condition (ii)		

Table 1: Points across the shock (PAS) for both initial conditions for the upwinding scheme, corresponding to figure 2. We see that initial condition (i) does not result in increasing PAS with increasing N, whereas (ii) does. This is the principal difference between the two cases, and corresponds to converging ε_p.

3.3.2 Continuous adjoint and WENO reconstruction

WENO solutions to the inviscid Burgers' equation are expressed as

$$u_i^{m+1} = u_i^m + \frac{\Delta t}{\Delta x}(F_{i+1/2}^m - F_{i-1/2}^m) \tag{42}$$

(a) Varying spatial resolution

(b) Varying temporal resolution

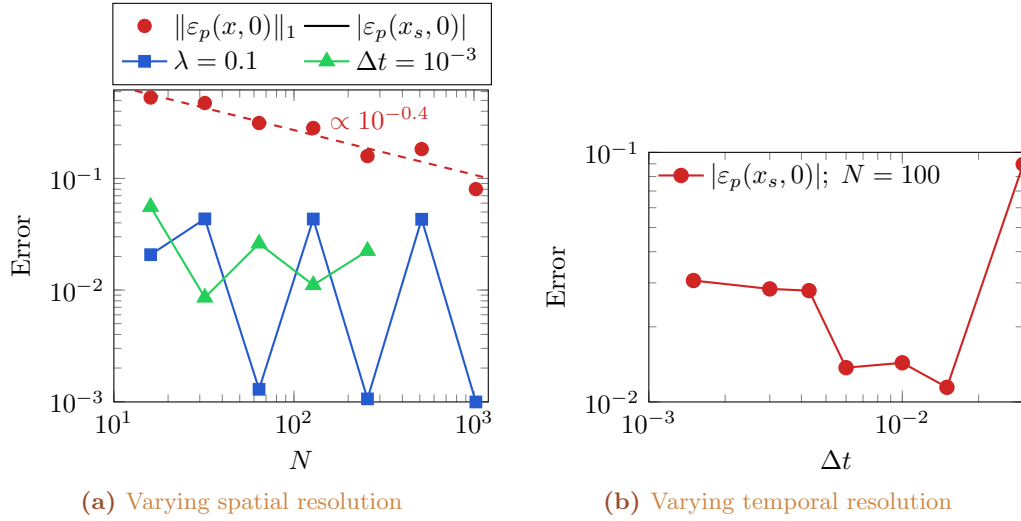**Figure 3:** Selected errors for the non-converging initial condition (i) for varying (a) Δx and (b) Δt with constant resolutions otherwise. No pattern can be seen, which suggests that the problem cannot be solved by increasing the simulation resolution in a particular way. In (a) we also show the L_1 convergence of the full solution, rather than just at the shock location; the near-monotonic character of this convergence belies the fact that such a broad measure of the error is not useful for determining convergence at the shock, which the problem is meant to demonstrate. This is because the shocks are simultaneously getting sharper.

(a) Initial condition (i)

(b) Initial condition (ii)
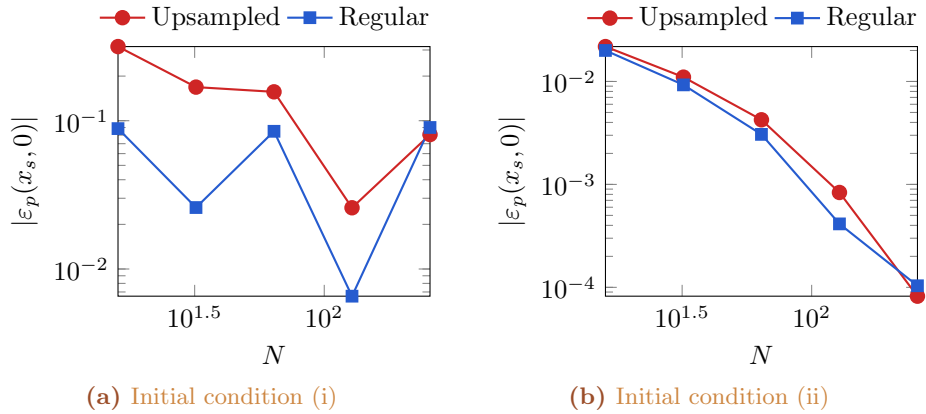
Figure 4: Shock errors are shown for both initial conditions and upsampled and regular solutions. Upsampled solutions are computed by taking the regular solution of previous figures and linearly interpolating additional points onto the forward solution. If the problem is resolution at the shock location (not enough points across the shock), then this technique can place extra grid points near this location without any expense for the forward problem. In the case of the upwinding scheme, no such advantage is found using this technique, presumably because it allows more points across the shock (more precise), but the location is not determined any more accurately.

(a) Initial condition (i)   **(a)** Initial condition (ii)
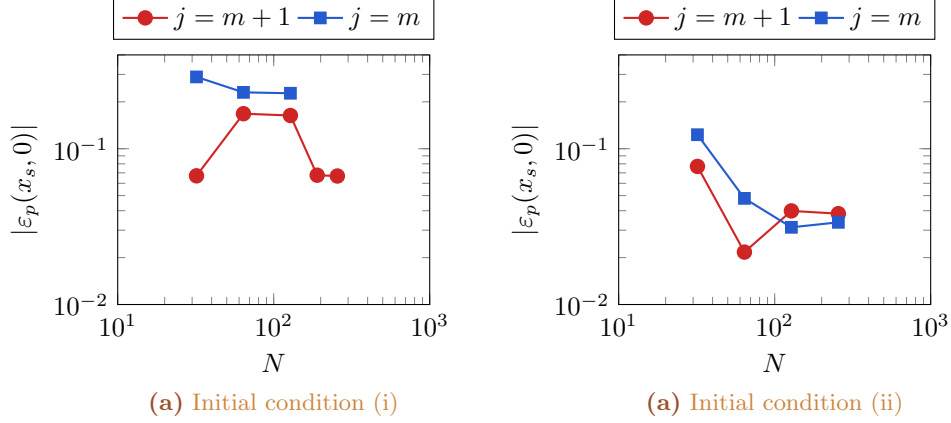
Figure 5: Adjoint shock error for the WENO scheme, $\lambda = 0.4$, varying N, both initial conditions (i) and (ii), no artificial viscosity, and both $j = m$ and $j = m + 1$. For initial condition (i) no convergence is seen, consistent with the upwinding scheme. (ii) appears to converge, though not uniformly. This is in contrast to the upwinding scheme that uniformly converged for this case.

where the numerical fluxes F_i^m are computed as

$$F_{i+1/2}^m = \frac{1}{2}(f_i^m + f_{i+1}^m) - \alpha(u_{i+1}^m - u_i^m). \tag{43}$$

Here, α is the maximum wave speed $\alpha \equiv \max_i |f'(u)|$, which gives the global Lax–Friedrichs (GLF) method. An artificial viscosity ν is sometimes added, and is denoted as appropriate.

N	$|\varepsilon_p(x_s, 0)|$	Order
16	0.0640	1.49
32	0.0226	2.20
64	0.00492	3.10
128	0.000575	0.80
256	0.000330	

(a) $\lambda = 0.1$

N	$|\varepsilon_p(x_s, 0)|$	Order
16	0.0666	1.49
32	0.0236	2.26
64	0.00493	2.80
128	0.000707	0.57
256	0.000475	0.027
512	0.000467	

(b) $\lambda = 0.2$

Table 2: WENO for initial condition (ii) with artificial viscosity $\nu = \Delta x^{0.95}$ and λ as labeled. We see that both cases uniformly converge to very small errors, which are likely due to the small errors introduced by the WENO scheme ($\varepsilon = 10^{-6}$).

N	$|\varepsilon_p(x_s, 0)|$	PAS
16	0.061	2
32	0.028	2
64	0.017	2
128	0.027	2
256	0.019	2

(a) $\nu = 0$

N	$|\varepsilon_p(x_s, 0)|$	PAS
16	0.0640	6
32	0.0226	8
64	0.00492	12
128	0.000575	16
256	0.000330	18

(b) $\nu = \Delta x^{0.95}$

Table 3: WENO for initial condition (ii) with $\lambda = 0.6$ and ν as labeled. Uniform convergence is only seen for the case with finite ν, which corresponds to increasing PAS. This is consistent with our conclusions from the upwinding scheme for both initial conditions, for which only (ii) resulted in increasing PAS with increasing N.
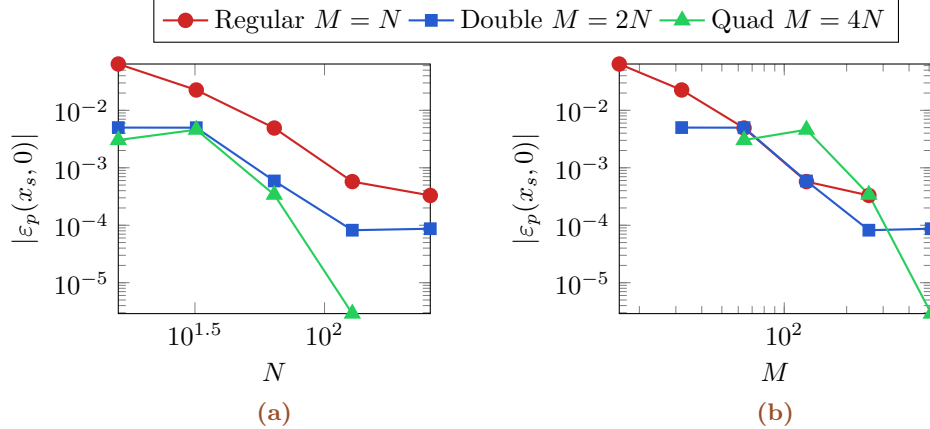
Figure 6: Adjoint shock error for upsampling WENO through linear interpolation, per figure 4; artificial viscosity is $\nu_p = \Delta x^{0.95}$ and M and N are the number of grid points in the adjoint and forward solutions respectively. The same data are shown in (a) and (b) for constant N and M. We see that controlling for constant M does not result in smaller errors at the shock, though for constant N (initial number of forward points) it does. This must be weighed in terms of relative expense of forward and adjoint solutions.

3.4   Discussion

Folowing Giles and Ulbrich (2010a), there appear to be two conditions for convergence of the adjoint solution:

1. Consistency in numerical smoothing between the forward and adjoint solutions.

2. As the grid resolution improves, the numerical smoothing varies such that the number of points across the shock increases and the shock decreases simultaneously.

In the case of the modified Lax–Friedrichs flux of Giles and Ulbrich (2010a) (see appendix A for details), the artificial dissipation is introduced consistently, satisfying condition 1., and has magnitude of the form $\nu = \Delta x^\alpha$, where $\alpha < 1$ ensures condition 2. is satisfied. From previous sections and other published results, we know that many numerical schemes do not satisfy these conditions, and thus do not converge upon the proper interior boundary condition (Alauzet and Pironneau, 2012; Fikl et al., 2016). Condition 2. was violated for initial condition (i) for the upwinding scheme considered here, but not initial condition (ii); this resulted in an inconsistent adjoint solution. The WENO solutions violated condition 2. for both initial conditions, and neither converged without additional dissipation (ν) to reinforce it. This additional dissipation was added in a way that was consistent for the forward and adjoint equations, hence the fulfillment of condition 1. as well.

3.5   Problems

The conditions of Giles and Ulbrich (2010a) are straightforward to state, though difficult to ensure *a priori*, and certainly challenging to prove, for equations more complicated than Burgers'. Further, Burgers' equation benefits from being scalar, and so there are no additional flow variables that can interact. Generally this is not the case for coupled hyperbolic system. More specifically, the artificial viscosity introduced in the WENO solution was sufficient to ensure both conditions above, though its unclear what amount of dissipation is required for more complicated systems, given the interactions present. The compressible inviscid flow equations are considered next as a demonstration of these

13

problems.

4 Compressible Euler equations

4.1 Forward problem

We analyze the one-dimensional compressible Euler equations

$$\boldsymbol{u}(x,t)_t + \boldsymbol{f}(\boldsymbol{u}(x,t))_x = 0, \quad x \in [-L, L];\ t \in [0, T], \tag{44}$$

with

$$\boldsymbol{u} = \begin{bmatrix} \rho \\ \rho v \\ E \end{bmatrix}, \quad \boldsymbol{f}(\boldsymbol{u}) = \begin{bmatrix} \rho v \\ \rho v^2 + P \\ v(E + P) \end{bmatrix}, \quad \text{and } P = (\gamma - 1)(E - \frac{\rho v^2}{2}). \tag{45}$$

We consider two problems:

1. Entropy wave, with $\gamma = 1.4$, domain length $L = 10$, Dirichlet boundary conditions, and initial conditions

$$v(x, 0) = v_\infty = 1, \quad P(x, 0) = 1/\gamma,$$

   and for density either

   - Gaussian pulse:
$$\rho(x, 0) = 3(\exp[-x^2] + 1)$$

   - Shock:
$$\rho(x, 0) = \begin{cases} 6 & \text{if } x \leq 0 \\ 3 & \text{otherwise} \end{cases}$$

   For uniform velocity and pressure initial conditions, the forward problem becomes an advection equation with velocity v_∞ for the density:

$$\rho_t + v_\infty \rho_x = 0, \tag{46}$$

   and v and p remain constant.

2. Sod shock tube, which has domain $a = -1$ and $b = 1$, $\gamma = 1.4$, initial condition

$$\{\rho, v, P\} = \begin{cases} \{1, 0, 1\} & \text{if } x \leq 0 \\ \{0.125, 0, 0.1\} & \text{if } x > 0, \end{cases} \tag{47}$$

   and Dirichlet boundary conditions. We will use $T = 0.1$ for this case; by this time a rarefaction, contact discontinuity, and shock have formed. The exact forward solution is provided for comparison by Toro (1999).

4.2   Adjoint equations

The continuous adjoint Euler equations are

$$\boldsymbol{p}_t + \boldsymbol{A}(\boldsymbol{u})^\top \boldsymbol{p}_x = \boldsymbol{0} \tag{48}$$

where

$$\boldsymbol{p} = \boldsymbol{u}^\dagger = \begin{bmatrix} \rho^\dagger \\ (\rho v)^\dagger \\ E^\dagger \end{bmatrix} \tag{49}$$

is the vector of conservative adjoint variables and

$$\boldsymbol{A}(\boldsymbol{u}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{\gamma-3}{2}\frac{u_2^2}{u_1^2} & (3-\gamma)\frac{u_2}{u_1} & \gamma-1 \\ (\gamma-1)\frac{u_2^3}{u_1^3} - \gamma\frac{u_3 u_2}{u_1^2} & \gamma\frac{u_3}{u_1} + 3\frac{1-\gamma}{2}\frac{u_2^2}{u_1^2} & \gamma\frac{u_2}{u_1} \end{bmatrix} \tag{50}$$

$$= \begin{bmatrix} 0 & 1 & 0 \\ \frac{\gamma-3}{2}v^2 & (3-\gamma)v & \gamma-1 \\ (\gamma-1)v^2 - \gamma\frac{Ev}{\rho} & \gamma\frac{E}{\rho} + 3\frac{1-\gamma}{2}v^3 & \gamma v \end{bmatrix} \tag{51}$$

is the inviscid flux Jacobian in terms of the conservative or primitive variables.

4.2.1   Characteristic-based approach

The eigen-decomposition of $\boldsymbol{A}$ is:

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}, \tag{52}$$

where $\boldsymbol{P}$ is the column matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues:

$$\boldsymbol{\Lambda} = \begin{bmatrix} v & 0 & 0 \\ 0 & v+c & 0 \\ 0 & 0 & v-c \end{bmatrix}, \tag{53}$$

$$\boldsymbol{P} = \begin{bmatrix} 1 & 1 & 1 \\ v & v+c & v-c \\ \frac{1}{2}v^2 & H+vc & H-vc \end{bmatrix}, \tag{54}$$

$$\boldsymbol{P}^{-1} = \begin{bmatrix} \frac{2(H-v^2)}{2H-v^2} & \frac{-2v}{v^2-2H} & \frac{-2}{2H-v^2} \\ \frac{v}{c}\frac{v(c+v)-2H}{4H-2v^2} & \frac{v(2c+v)-2H}{2c(v^2-4H)} & \frac{1}{2H-v^2} \\ \frac{v}{c}\frac{v(c-v)+2H}{4H-2v^2} & \frac{v(2c+v)+2H}{2c(v^2-4H)} & \frac{1}{2H-v^2} \end{bmatrix} \tag{55}$$

where c is the speed of sound

$$c = \sqrt{\gamma P/\rho}, \tag{56}$$

and H is the enthalpy

$$H = \frac{E+P}{\rho} = \frac{\gamma}{\gamma-1}\frac{P}{\rho} + \frac{1}{2}v^2. \tag{57}$$

We are interested in the transpose $\boldsymbol{A}^\top$:

$$
\begin{aligned}
\boldsymbol{A}^\top &= (\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1})^\top \\
&= (\boldsymbol{P}^{-1})^\top \boldsymbol{\Lambda}\boldsymbol{P}^\top \\
&= (\boldsymbol{P}^\top)^{-1}\boldsymbol{\Lambda}\boldsymbol{P}^\top.
\end{aligned}
\tag{58}
$$

Defining $\boldsymbol{Q} \equiv \boldsymbol{P}^\top$ we have $\boldsymbol{A}^\top = \boldsymbol{Q}^{-1}\boldsymbol{\Lambda}\boldsymbol{Q}$. Substituting into (48) yields

$$
\begin{aligned}
\boldsymbol{p}_t + \boldsymbol{Q}^{-1}\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{p}_x &= \boldsymbol{0} \\
\boldsymbol{Q}\boldsymbol{p}_t + \boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{p}_x &= \boldsymbol{0}.
\end{aligned}
$$

Introducing a change of variable $\boldsymbol{q} \equiv \boldsymbol{Q}\boldsymbol{p}$ we have, by product rule,

$$
\begin{aligned}
\boldsymbol{q}_t &= (\boldsymbol{Q}\boldsymbol{p})_t = \boldsymbol{Q}_t\boldsymbol{p} + \boldsymbol{Q}\boldsymbol{p}_t, \\
\boldsymbol{q}_x &= (\boldsymbol{Q}\boldsymbol{p})_x = \boldsymbol{Q}_x\boldsymbol{p} + \boldsymbol{Q}\boldsymbol{p}_x,
\end{aligned}
$$

and thus

$$
\begin{aligned}
\boldsymbol{q}_t + \boldsymbol{\Lambda}\boldsymbol{q}_x &= (\boldsymbol{Q}_t + \boldsymbol{\Lambda}\boldsymbol{Q}_x)\boldsymbol{p}, \\
\boldsymbol{q}_t + \boldsymbol{\Lambda}\boldsymbol{q}_x &= (\boldsymbol{Q}_t + \boldsymbol{\Lambda}\boldsymbol{Q}_x)\boldsymbol{Q}^{-1}\boldsymbol{q}.
\end{aligned}
\tag{59}
$$

Following (59), the eigenvalue decomposition decouples the derivatives of $\boldsymbol{p}$, though the equations remained coupled as $\boldsymbol{Q}(x,t)$ is both space and time dependent and $\boldsymbol{Q}$ and its inverse are full matrices.

Through fist-order time derivative operator splitting

$$
\boldsymbol{q}_t + \boldsymbol{\Lambda}\boldsymbol{q}_x = \boldsymbol{0}
\tag{60}
$$

$$
\boldsymbol{q}_t = \boldsymbol{B}\boldsymbol{q},
\tag{61}
$$

where $\boldsymbol{B} \equiv (\boldsymbol{Q}_t + \boldsymbol{\Lambda}\boldsymbol{Q}_x)\boldsymbol{Q}^{-1}$. (60) is a set of decoupled transport equations that admit the same challenges as the adjoint inviscid Burgers' equation of the previous sections in the presence of shocks. Thus, all machinery developed for the adjoint inviscid Burgers' equations can be supplanted here, including the convergence proofs of (Giles and Ulbrich, 2010a), the analysis of (Fikl et al., 2016), and the results from previous sections. (61) is problematic, as it contains both space and time derivatives of $\boldsymbol{u}$. However, these terms have already been accurately evaluated via the forward solution.

A fully general solution $\{\rho(x,t), v(x,t), p(x,t)\}$ yields

$$
\frac{\partial \boldsymbol{Q}}{\partial x} =
\begin{bmatrix}
0 & v_x & vv_x \\
0 & v_x + c_x & H_x + (vc)_x \\
0 & v_x - c_x & H_x - (vc)_x
\end{bmatrix},
\tag{62}
$$

where

$$
c_x = \sqrt{\frac{\gamma}{\rho P}}\,\frac{\rho P_x - P\rho_x}{2\rho}
\tag{63}
$$

$$
(vc)_x = \sqrt{\frac{\gamma}{\rho P}}\,\frac{\rho(vP_x + 2Pv_x) - vP\rho_x}{2\rho}
\tag{64}
$$

$$
H_x = \frac{\gamma}{\gamma - 1}\frac{\rho P_x - P\rho_x}{\rho^2} + vv_x.
\tag{65}
$$

In this case, the spatial derivatives of the primitive variables can be deduced from the derivatives of the flux $\boldsymbol{f}(\boldsymbol{u})_x$ as computed by a shock capturing scheme (e.g. WENO):

$$\boldsymbol{f}(\boldsymbol{u})_x = \begin{bmatrix} \mathcal{F}_1 \\ \mathcal{F}_2 \\ \mathcal{F}_3 \end{bmatrix} = \begin{bmatrix} \rho v \\ \rho v^2 + P \\ v(E+P) \end{bmatrix}_x = \begin{bmatrix} \rho v_x + v\rho_x \\ 2\rho v v_x + \rho_x v^2 + P_x \\ \kappa(vP_x + v_x P) + \frac{1}{2}v^2(v\rho_x + 3\rho v_x) \end{bmatrix}, \tag{66}$$

where $\kappa \equiv \gamma/(\gamma - 1)$. Solving (66) for the derivatives of the primitive variables we have:

$$\rho_x = \frac{\mathcal{F}_1(\rho v^2(3/2 - 2\kappa) + \kappa P) + \mathcal{F}_2 \kappa \rho v - \mathcal{F}_3 \rho}{v(\kappa(P - \rho v^2) + \rho v^2)}, \tag{67}$$

$$v_x = \frac{v(\mathcal{F}_1(\kappa - 1/2)v - \mathcal{F}_2 \kappa) + \mathcal{F}_3}{\kappa (P - \rho v^2) + \rho v^2}, \tag{68}$$

$$P_x = \frac{v(\mathcal{F}_1(-\kappa P - \rho v^2/2) + \rho(\mathcal{F}_2 v - \mathcal{F}_3)) + \mathcal{F}_2 \kappa \rho}{\kappa (P - \rho v^2) + \rho v^2}, \tag{69}$$

In the case of an adiabatic entropy wave we have $v = v_\infty$ and $P = 1/\gamma$ as constants, so their derivatives vanish and the solution is

$$\frac{\partial \boldsymbol{Q}}{\partial x} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & c_x & H_x + v_\infty c_x \\ 0 & -c_x & H_x - v_\infty c_x \end{bmatrix}, \tag{70}$$

with

$$c_x = -\frac{1}{2}\frac{\rho_x}{\rho^{3/2}} \qquad \text{and} \qquad H_x = \frac{1}{1 - \gamma}\frac{\rho_x}{\rho^2}. \tag{71}$$

Thus, so long as the (71) can be accurately evaluated, (59) can also be accurately and consistently evaluated. Indeed (71) follow from a shock capturing forward solution, which computes $\boldsymbol{f}(\boldsymbol{u})_x$. Finally the adjoint solution is recovered as

$$\boldsymbol{p} = \boldsymbol{Q}^{-1}\boldsymbol{q}, \tag{72}$$

which is well-posed.

4.2.2 Another way to create a conservative product

$$\boldsymbol{p}_t + \boldsymbol{A}^\top \boldsymbol{p}_x = \boldsymbol{0} \tag{73}$$

$$\boldsymbol{p}_t + (\boldsymbol{A}^\top \boldsymbol{p})_x = \boldsymbol{A}_x^\top \boldsymbol{p} \tag{74}$$

Part of the problem of creating a stable, non-oscillatory simulation is evaluating derivatives of the adjoint variable $\boldsymbol{p}$, which will also contain discontinuities. In (73), the non-conservative product of $\boldsymbol{A}^\top$ and $\boldsymbol{p}_x$ can contain discontinuities at the same spatial location, and will necessarily interact with one another. The principal way to deal with such terms is though finite volume methods, though in the case of a non-conservative product a FV scheme does not make sense. We can make the equation well posed by creating a conservative product, where derivatives of the adjoint variable $\boldsymbol{p}$ are either removed or moved into the conservative product. The spatial derivatives of $\boldsymbol{A}^\top$ are already well evaluated from the forward equations.
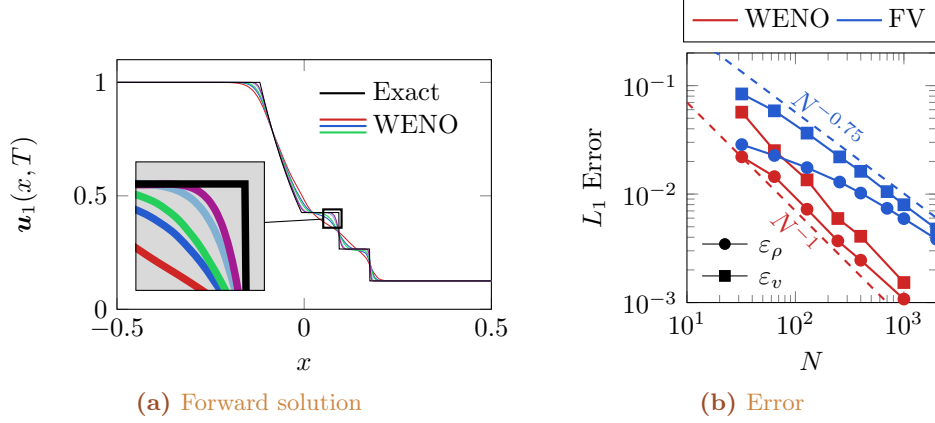
(a) Forward solution   **(b)** Error

Figure 7: (a) `WENO` solution for density $\rho(x, T)$ and various resolutions N. (b) errors for ρ and v and varying N for both `WENO` and `FV`. $\lambda = 0.9$ for all cases.

4.3 Numerical solutions

4.3.1 Forward solutions

The forward equations are solved using two methods:

`WENO`: A third-order WENO reconstruction on the conserved variables $\boldsymbol{u}$ and a local (GLF) or global Lax–Friedrichs (GLF) numerical flux given by

$$\hat{f}^m_{i-1/2} = \frac{1}{2}\left[f^m_i + f^m_{i-1} - \alpha_i\left(u^m_i - u^m_{i-1}\right)\right],\tag{75}$$

where

$$\alpha_i = \max_{j=i,i-1} a_j \quad\text{and}\quad a_j \equiv |v_j| + c_j,\tag{76}$$

where a_j is the local (LLF) or global (GLF) maximum wave-speed and c_j is the local (LLF) or global (GLF) sound speed. The TVD RK3 is used for time integration.

`FV`: No reconstruction. A modified Lax–Friedrichs (MLF) numerical flux is used, which matches (75) except $\alpha_i = \alpha = \Delta x^\beta/2 \max_i a_i$ where i is over all spatial points (see also appendix A). This keeps the dissipation required for stability due to the propagation of characteristics at the maximum wave speed separate from that added for slowing down the sharpening of shocks. The forward Euler method is used for time integration.

`Diag`: The characteristic-based diagonalization techniques of section 4.2.1. The forward solution is the same as `FV`.

The CFL number

$$\lambda \equiv \frac{\Delta t}{\Delta x} \max_{i,t\in(0,T)} a_i,\tag{77}$$

where max is over all spatial grid points $i = 1, \ldots, N$ and times t.

18

N	$\|\varepsilon_v(x,T)\|_1$	PAS	N	$\|\varepsilon_v(x,T)\|_1$	PAS
64	0.026	4	64	0.041	4
82	0.021	4	82	0.032	5
100	0.018	4	100	0.027	5
128	0.014	5	128	0.021	7
200	0.0083	4	200	0.013	8
300	0.0064	5	300	0.0098	9
400	0.0040	4	400	0.0060	10
(a) $\nu_p = 0$			**(b)** $\nu_p = \Delta x^{0.9}$		

Table 4: `WENO` with and without additional dissipation. PAS is points across shock. $\lambda = 0.6$ for all cases. A finite ν_p is required for the number of points across the discontinuity in v to increase with mesh refinement. Note that this was a requirement for convergence for the Burgers' equation.

4.3.2 Adjoint solutions

Again, the unmanipulated adjoint equations are:

$$\boldsymbol{p}_t + \boldsymbol{A}(\boldsymbol{u})^\top \boldsymbol{p}_x = \boldsymbol{0}. \tag{78}$$

`WENO`: Reconstruction is performed on $\boldsymbol{p}$. The spatial derivative $\boldsymbol{p}_x$ is taken as a Lax–Friedrichs flux on this reconstruction with $\alpha = 0$. The dissipation term is moved to the right hand side.

  `FV`: Similar to `WENO`, the spatial derivative is taken as Lax–Friedrichs with the dissipation part moved to the right hand side. This is the consistent formulation that goes with the forward solution using this method, such that the dissipation and $\boldsymbol{A}^\top$ do not interact.

`Diag`: The scheme of `FV` is applied to each of the independent transport equations (60).

In most cases, the shock that forms only has a few grid points across it, regardless of resolution. In previous sections we introduced artificial dissipation to ensure that the number of points across the shock increases, even as the width of the shock decreases with increasing resolution. Here, this takes the form:

$$\boldsymbol{p}_t + \boldsymbol{A}(\boldsymbol{u})^\top \boldsymbol{p}_x = \nu_p \boldsymbol{p}_{xx}, \tag{79}$$

where it is unclear exactly what ν_p should be. We discuss variations throughout.

4.3.3 `Magudi` solutions

For verification we compare our smooth solutions to the solutions of `Magudi` (Vishnampet et al., 2015). It is a compressible Navier–Stokes and Euler solver that implements high resolutions finite difference methods in space and explicit RK4 time integration. Boundary conditions are imposed via SBP–SAT energy stable schemes. Both continuous and discrete-exact adjoint solutions are available. The code has a dissipation option useful for handling weak shocks, and so we only use it for verification with always-smooth problems.

4.3.4 `Mathematica` solutions

For solutions that involve shocks, we also reference solutions from the Mathematica software package, for which we implement a high-order pseudospectral method in space:

```
#  Eqns: Governing equations. DepVars: {t,x}. Scheme: Numerical scheme.
#  Nx: number of spatial points. Order: Order of pseudospectral differentiation
Scheme[n_Integer, o_: "Pseudospectral"] := {
  "MethodOfLines",
  "SpatialDiscretization" -> {
    "TensorProductGrid",
    "MaxPoints" -> n,
    "MinPoints" -> n,
    "DifferenceOrder" -> o
  }
}
NDSolve[Eqns, DepVars, {t, 0, T}, {x, -L, L}, Method -> Scheme[Nx,Order]]
```

4.4 Results: Adjoint fields

4.4.1 Entropy wave: Gaussian initial data

4.4.2 Entropy wave: Shock initial data

$N/10^3$	$\rho_L^\dagger$	$\rho_R^\dagger$	$N/10^3$	$\rho_L^\dagger$	$\rho_R^\dagger$	$N/10^3$	$\rho_L^\dagger$	$\rho_R^\dagger$
4	11.1133	10.4378	4	11.1106	10.4308	4	11.1065	10.4204
8	11.1135	10.4383	8	11.1123	10.4350	8	11.1114	10.4329
10	11.1135	10.4383	10	11.1126	10.4360	10	11.1120	10.4343
(a) $w = 14$			**(b)** $w = 24$			**(c)** $w = 34$		

Table 5: Mathematica adjoint density results at the left and right plateaus $\rho_{L/R}^\dagger = \rho^\dagger(x_{L/R})$, with $x_L = -0.5$ and $x_R = 0.75$, following figure 9. Three levels of shock "sharpness" w are shown (a)–(c). For all w the solution appears to converge, though to different values depending upon w, as expected.

4.4.3 Sod shock tube

4.5 Higher dimensions

The 2D Euler equations are

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x + \boldsymbol{g}(\boldsymbol{u})_y = \boldsymbol{0}, \tag{80}$$

where $\boldsymbol{f}$ and $\boldsymbol{g}$ are the fluxes in the x and y directions, respectively. The corresponding linear adjoint equation is

$$\boldsymbol{p}_t + \boldsymbol{A}^\top(\boldsymbol{u})\boldsymbol{p}_x + \boldsymbol{B}^\top(\boldsymbol{u})\boldsymbol{p}_y = \boldsymbol{0} \tag{81}$$

where

$$\boldsymbol{A}(\boldsymbol{u}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \quad \text{and} \quad \boldsymbol{B}(\boldsymbol{u}) = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}. \tag{82}$$

The time derivative operator of (81) can be split:

$$\boldsymbol{p}_t + \boldsymbol{A}^\top(\boldsymbol{u})\boldsymbol{p}_x = \boldsymbol{0} \tag{83}$$

$$\boldsymbol{p}_t + \boldsymbol{B}^\top(\boldsymbol{u})\boldsymbol{p}_y = \boldsymbol{0}. \tag{84}$$

(a) `FV`

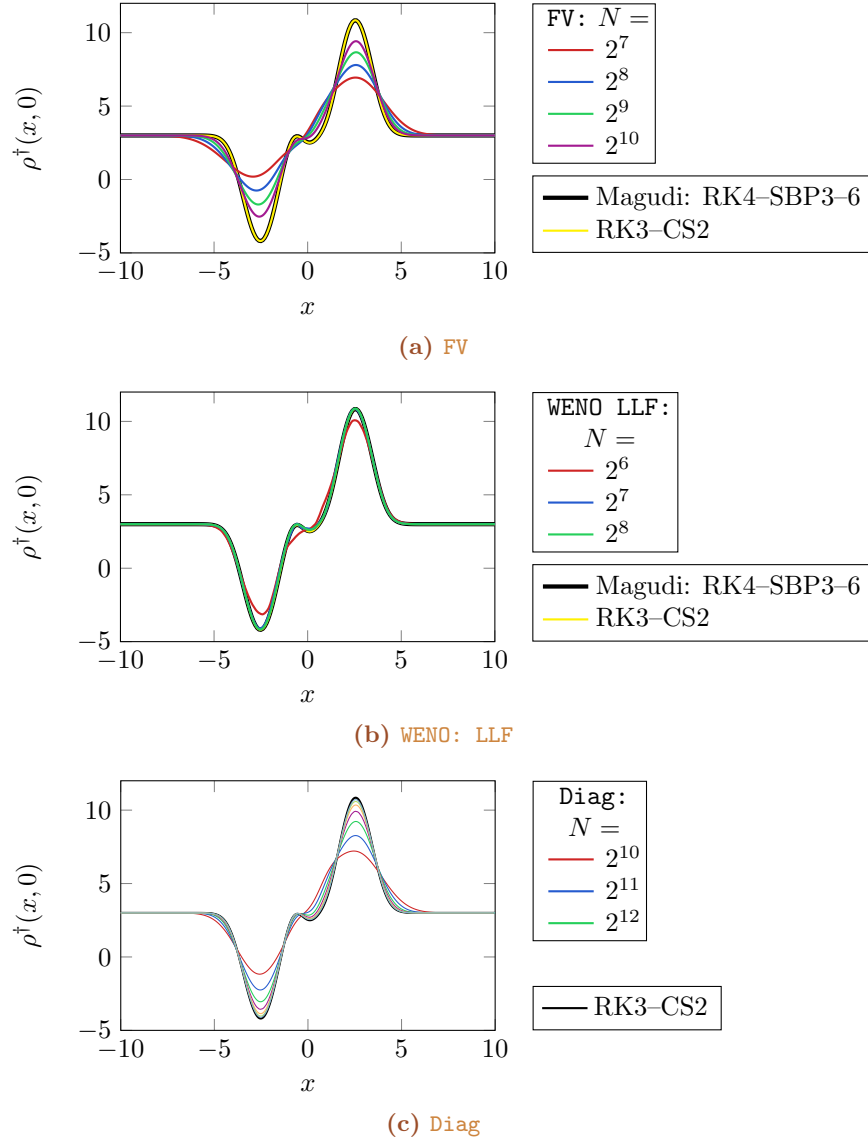**(b)** `WENO: LLF`

(c) `Diag`

Figure 8: Adjoint density fields for a Gaussian entropy wave. All solutions have $\lambda = 0.4$ and $T = 5$. `Magudi` and non-dissipative RK3-CS solutions are shown for reference. All solutions converge with increasing N. Even the entropy wave solution (advection equation forward in time) does not have an analytic adjoint solution. This is because the full coupling introduced by the transpose of the inviscid flux jacobian $\boldsymbol{A}^{\top}$, which re-couples the forward equations.
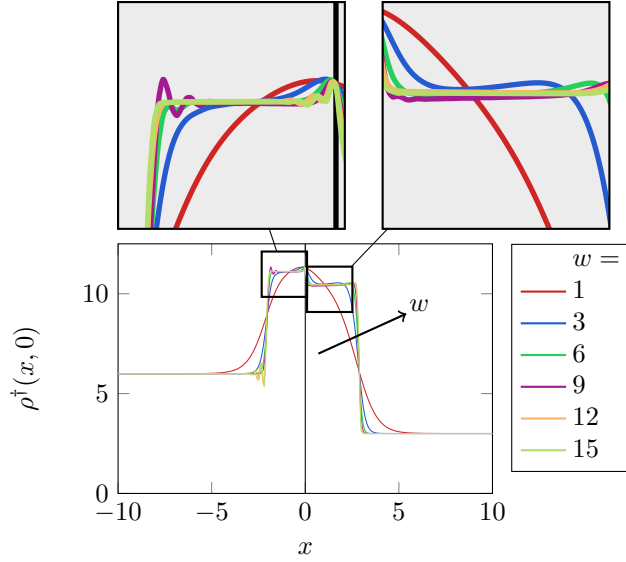
Figure 9: `Mathematica` adjoint density field for a shock initial condition, $T = 5$, and $N = 10^4$. The forward solution is the exact advection of an approximate shock $\rho(x,t) \sim \tanh(-w(x-t))$, and so the shock becomes thinner with increasing w. The solution appears to converge to two plateaus, one to the left and one to the right of $x = 0$. Spurious oscillations near the shocks are seen for $w \geq 9$ due to insufficient spatial resolution for that degree of sharpness.

$N/10^3$	$\rho_L^\dagger$	$\rho_R^\dagger$	PAS
4	11.043	10.260	9
8	11.063	10.236	10
16	11.011	10.133	11
32	11.034	10.216	14
48	11.038	10.203	20

(a) `WENO:GLF`/$\nu_p = 0$

$N/10^3$	$\rho_L^\dagger$	$\rho_R^\dagger$	PAS
4	11.149	10.491	14
8	11.163	10.541	18
16	11.127	10.462	25
32	11.123	10.455	36
48	11.121	10.452	44
64	11.120	10.450	51

(b) `WENO:GLF`/$\nu_p = 0.005\Delta x^{0.95}$

$N/10^3$	$\rho_L^\dagger$	$\rho_R^\dagger$	PAS
4	11.153	10.528	18
8	11.129	10.459	25
16	11.125	10.450	35
32	11.122	10.446	50
48	11.120	10.445	62
64	11.119	10.444	71

(c) `WENO:GLF`/$\nu_p = 0.01\Delta x^{0.95}$

$N/10^3$	$\rho_L^\dagger$	$\rho_R^\dagger$	PAS
4	11.151	10.378	73
8	11.140	10.396	103
16	11.133	10.409	145
32	11.127	10.417	204
48	11.125	10.421	249
64	11.123	10.424	288

(d) `Diag`/$\nu_p = \Delta x^{0.95}$

Table 6: Adjoint density plateaus, as defined in table 5, for (a)–(c) `WENO` with varying artificial dissipation and (d) `Diag`. The cases with the least dissipation, (a) and (b), do not converge for either plateau. (c) and (d) appear to converge, though (d) more slowly as it entails more dissipation. Note that in all cases the number of points across the shock (PAS) increases, satisfying the condition noted in previous sections, though we do not have convergence for all cases here.
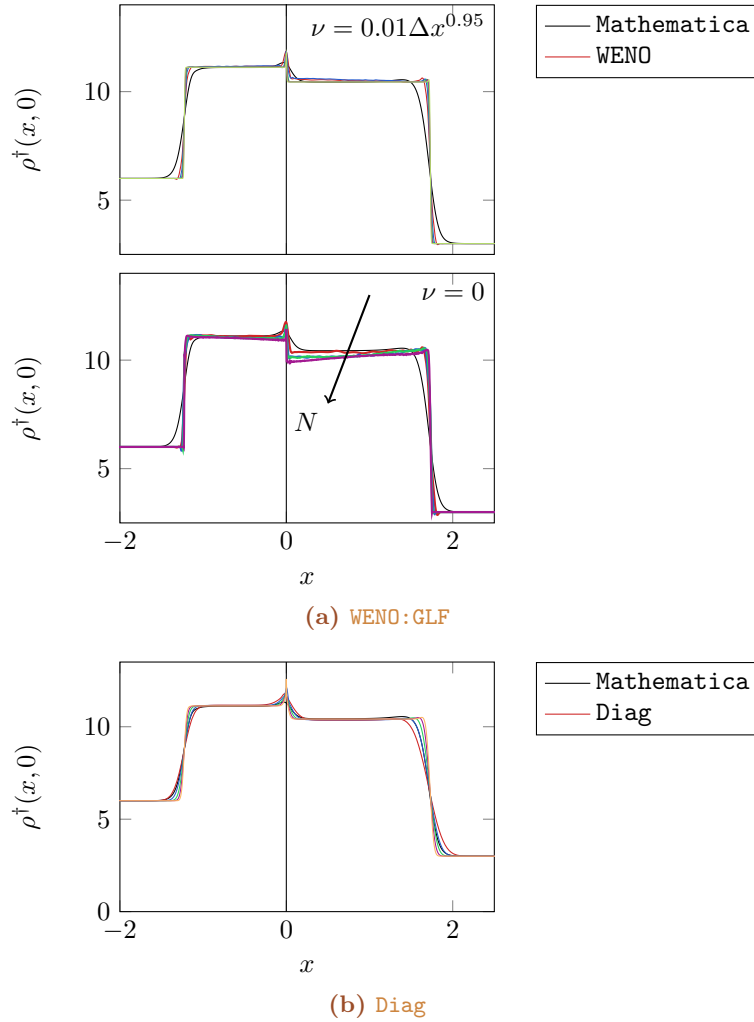
Figure 10: Adjoint density fields for two methods (a) and (b), $\lambda = 0.45$, varying resolution and two viscosities ν_p as labeled. `Diag` uses $\nu_p = \Delta x^{0.95}$. We see that the `WENO` $\nu_p = 0$ cases do not converge, while the other do. This supports an argument that a sufficient amount of dissipation is required for a stable and convergent solution, though it is unclear what this amount this. In this case the $\nu_p = 0.01\Delta x^{0.95}$ case has enough dissipation, or nearly enough, for a converging solution in this very simple case, though this determination is *ad hoc*.
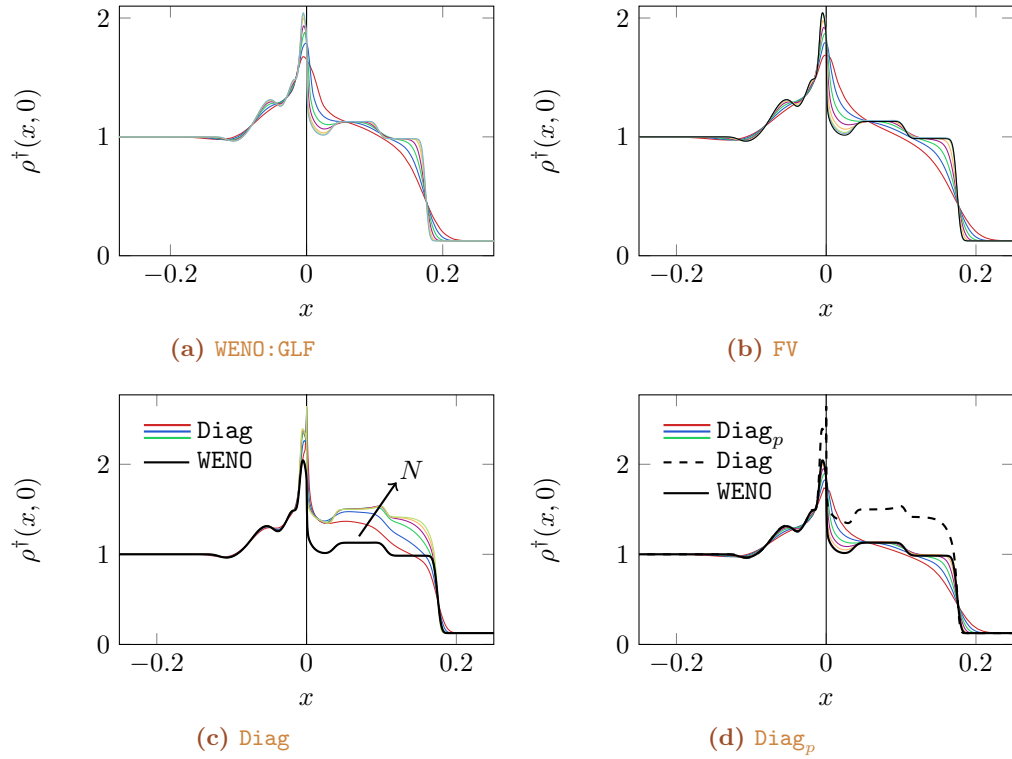
(a) `WENO:GLF`

(b) `FV`

(c) `Diag`

(d) `Diag`$_p$

Figure 11: Adjoint density field for the Sod shock tube problem with $T = 0.1$, $\lambda = 0.45$, $\nu_p = \Delta x^{0.95}$. Spatial resolutions $N = 2^k$ for $k = 10, \ldots, 16$ are shown. We see that both `WENO` and `FV` converge to the same solution. In (c) the `Diag` scheme converges to a different solution (only for $0 < x \lesssim 0.2$, however), though qualitatively similar. It is not immediately clear what solution is correct, if either. The diagonalization technique was developed such that the correct amount of dissipation is added for a consistent solution. In (d), the `Diag` scheme is modified such that the dissipation is added to the adjoint variable $\boldsymbol{p}$ rather than the diagonalized variable $\boldsymbol{q}$ (call this `Diag`$_p$). This result is consistent with the `WENO` and `FV` solutions.

Each of (84) can be diagonalized following an analogous procedure to section 4.2.1:

$$\boldsymbol{A}^{\top} = \boldsymbol{Q}_{(x)}^{-1} \boldsymbol{\Lambda}_{(x)} \boldsymbol{Q}_{(x)}, \tag{85}$$

$$\boldsymbol{B}^{\top} = \boldsymbol{Q}_{(y)}^{-1} \boldsymbol{\Lambda}_{(y)} \boldsymbol{Q}_{(y)}, \tag{86}$$

where the $(\cdot)$ subscripts indicate the matrices in that coordinate direction, not partial derivatives. We then have

$$\boldsymbol{p}_t + \boldsymbol{Q}_{(x)}^{-1} \boldsymbol{\Lambda}_{(x)} \boldsymbol{Q}_{(x)} \, \boldsymbol{p}_x = \boldsymbol{0} \tag{87}$$

$$\boldsymbol{p}_t + \boldsymbol{Q}_{(y)}^{-1} \boldsymbol{\Lambda}_{(y)} \boldsymbol{Q}_{(y)} \, \boldsymbol{p}_y = \boldsymbol{0}. \tag{88}$$

Now each of (88) can be solved through a similar though a distinct change in variable:

$$\boldsymbol{q}_t^{(x)} + \boldsymbol{\Lambda}_{(x)} \boldsymbol{q}_x^{(x)} = (\boldsymbol{Q}_t^{(x)} + \boldsymbol{\Lambda}_{(x)} \boldsymbol{Q}_x^{(x)}) \boldsymbol{Q}_{(x)}^{-1} \boldsymbol{q}^{(x)} \tag{89}$$

$$\boldsymbol{q}_t^{(y)} + \boldsymbol{\Lambda}_{(y)} \boldsymbol{q}_y^{(y)} = (\boldsymbol{Q}_t^{(y)} + \boldsymbol{\Lambda}_{(y)} \boldsymbol{Q}_y^{(y)}) \boldsymbol{Q}_{(y)}^{-1} \boldsymbol{q}^{(y)} \tag{90}$$

where $\boldsymbol{q}^{(x,y)} = \boldsymbol{Q}_{(x,y)} \boldsymbol{p}$. Each of (90) can be solved for $\boldsymbol{q}_t^{(x,y)}$, again following section 4.2.1. We can then compute $\boldsymbol{p}_t$ through the backwards transform for each stage and sum the two. This is more costly than the one-dimensional problem, since the forward and backward transforms must be computed at every time step, though not prohibitively so compared to the similar decompositions computed for WENO schemes.

5 Conclusions

We considered the adjoint Burgers' and 1D compressible Euler equations, which are scalar and vector hyperbolic conservation laws, respectively, that can develop discontinuities for even smooth initial conditions. The goal was to identify consistent and convergent numerical methods for these problems. This is particularly challenging in the presence of shocks, for which special numerical methods are required for even just the forward equations. In this case, the most common techniques are shock-fitting and shock-capturing. Shock fitting introduces the kinematic condition at the shock, treating it independently from the continuous smooth equations. This method can be prohibitively complicated: the shock conditions must be imposed at the shock location, which requires tracking new shocks as they form and disappear. Further, a separate set of perturbation equations are required for the adjoint problem. This method was not considered further. Instead, we investigated shock-capturing, finite volume methods in particular, as a route to solve for the flow.

In the case of Burgers' equation, we found that for a consistent and convergent adjoint solution two conditions must be satisfied: dissipation keeps the number of grid points along the shock increasing more quickly than the shock decreases in width as the grid is refined, and that the dissipation is applied in a consistent way. This result was proven by Giles and Ulbrich (2010a) for a modified Lax–Friedrichs scheme that satisfied these properties. Here, we showed that this result holds more generally through empirical observation, and that the listed properties are sufficient for convergence to a consistent solution.

Actual flow equations are more complex than the Burgers' equation, and there are no proven results to this end for even the compressible 1D Euler equations. This problem is more complicated primarily because of the full-rank inviscid flux Jacobian that appears as a front-factor to the spatial derivative of the adjoint variables, which fully couples all variables of the adjoint equation and forward solution.

We aimed to demonstrate empirically if the above properties (consistent dissipation and increasing points along the shock) were sufficient for convergence in this case. Unfortunately, exact solutions to the adjoint equations are not available, even for cases without shocks. As a result, we first considered the convergence of the solution, regardless if it was the correct one. In the case of an advecting adiabatic entropy wave with discontinuous initial data, we showed that some dissipation was required for a converging solution, and that the amount utilized for Burgers' equation was sufficient. We also considered a Sod shock tube initial condition, which displays discontinuities in all flow variables. Again, we saw that with sufficient dissipation a converging solution could be found. For this problem we also introduced a characteristic-based formulation of the adjoint, which aimed to solve a set of uncoupled transport equations along the forward solution characteristics. These transport equations were qualitatively no different than the adjoint Burgers' equation, and so the same consistent and converging numerical methods should be able to be applied. Upon using this method for the Sod shock tube, we saw a different converging solution than the solution involving full coupling. Ultimately, we showed that the way the dissipation is added, even if it is of the same magnitude, can effect the solution. This disagreement occurred even though both conditions stipulated by Giles and Ulbrich (2010a) for consistent and convergent solutions to the adjoint Burgers' equation were satisfied. The characteristic-based scheme was also extended to two-dimensions.

A Lax–Friedrichs fluxes

The hyperbolic conservation law can be written in semi-discrete form as

$$(\boldsymbol{u}_t)_i + \frac{\hat{\boldsymbol{f}}_{i+1/2} - \hat{\boldsymbol{f}}_{i-1/2}}{\Delta x} = 0 \tag{91}$$

The Lax–Friedrichs numerical flux takes the form

$$\hat{\boldsymbol{f}}_{i-1/2} = \frac{1}{2} \left[\boldsymbol{f}_i + \boldsymbol{f}_{i-1} - \alpha \left(\boldsymbol{u}_i - \boldsymbol{u}_{i-1} \right) \right], \tag{92}$$

where

$$\text{Global Lax–Friedrichs (GLF):} \qquad \alpha \equiv \frac{\Delta x}{\Delta t} \tag{93}$$

$$\text{Modified Lax–Friedrichs (MLF):} \qquad \alpha \equiv \frac{2\nu}{\Delta x}; \ \nu \equiv \Delta x^\beta \tag{94}$$

With (91) we have

$$\boldsymbol{u}_t + \frac{\boldsymbol{f}_{i+1} - \boldsymbol{f}_{i-1}}{2\Delta x} = \frac{\alpha}{2} \left[\frac{\boldsymbol{u}_{i+1} - 2\boldsymbol{u}_i + \boldsymbol{u}_{i-1}}{\Delta x} \right] \tag{95}$$

which takes the form

$$\boldsymbol{u}_t + \boldsymbol{f}_x = \epsilon \boldsymbol{u}_{xx} \quad \text{where} \quad \epsilon \equiv \frac{\Delta x \alpha}{2}, \tag{96}$$

and $\boldsymbol{f}_x$ and $\boldsymbol{u}_{xx}$ are approximated with the usual second-order central difference schemes. With (93) and (94) this gives

$$\text{GLF:} \qquad \epsilon = \frac{1}{2} \frac{\Delta x^2}{\Delta t} \tag{97}$$

$$\text{MLF:} \qquad \epsilon = \Delta x^\beta \tag{98}$$

The GLF scheme looks like $\epsilon = \Delta x / 2\lambda$ where $\lambda = \Delta t / \Delta x$ is a CFL number that is typically fixed, and so $\epsilon = c\Delta x$ upon mesh refinement, whereas the MLF scheme has $\beta < 1$, and thus the dissipation becomes smaller more slowly when decreasing Δx than the GLF. This ensures that shocks are dissipated across an increasing number grid points as the mesh is refined.

References

Alauzet, F. and Pironneau, O. (2012). Continuous and discrete adjoints to the Euler equations for fluids. *J. Num. Methods Fluids*, 70:135–157.

Fikl, A., Le Chenadec, V., Sayadi, T., and Schmid, P. J. (2016). A comprehensive study of adjoint-based optimization of non-linear systems with application to Burgers' equation. volume 46. AIAA Aviation Forum.

Giles, M. B. (2002). Adjoining approximations with shocks. In *Proceedings of the Ninth International Conference on Hyperbolic Problems*, pages 185–194.

Giles, M. B. and Ulbrich, S. (2010a). Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 1: Linearized approximations and linearized output functionals. *SIAM J. Num. Anal.*, 48:882–904.

Giles, M. B. and Ulbrich, S. (2010b). Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoining approximations and extenions. *SIAM J. Num. Anal.*, 48:905–921.

Matsuzawa, T. and Hafez, M. (1997). Optimum shape design using adjoint equations for compressible flows with shock waves. *AIAA J.*, 97(2078).

Messyahu, H., Tadmor, E., and Tassa, T. (1994). The convergence rate of Godunov type schemes. *SIAM J. Num. Anal.*, 31:1–16.

Toro, E. F. (1999). *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*. Applied mechanics: Researchers and students. Springer.

Vishnampet, R., Bodony, D. J., and Freund, J. B. (2015). A practical discrete-adjoint method for high-fidelity compressible turbulence simulations. *J. Comp. Phys.*, 285:173–192.