SIAM J. NUMER. ANAL.
Vol. 48, No. 3, pp. 905–921

CONVERGENCE OF LINEARIZED AND ADJOINT APPROXIMATIONS FOR DISCONTINUOUS SOLUTIONS OF CONSERVATION LAWS. PART 2: ADJOINT APPROXIMATIONS AND EXTENSIONS*

MIKE GILES† AND STEFAN ULBRICH‡

Abstract. This paper continues the convergence analysis in [M. Giles and S. Ulbrich, *SIAM J. Numer. Anal.*, 48 (2010), pp. 882–904] of discrete approximations to the linearized and adjoint equations arising from an unsteady one-dimensional hyperbolic equation with a convex flux function. We consider a simple modified Lax–Friedrichs discretization on a uniform grid, and a key point is that the numerical smoothing increases the number of points across the nonlinear discontinuity as the grid is refined. It is proved that there is convergence in the discrete approximation of linearized output functionals even for Dirac initial perturbations and pointwise convergence almost everywhere for the solution of the adjoint discrete equations. In particular, the adjoint approximation converges to the correct uniform value in the region in which characteristics propagate into the discontinuity. Moreover, it is shown that the results of [M. Giles and S. Ulbrich, *SIAM J. Numer. Anal.*, 48 (2010), pp. 882–904] and the present paper hold also for quite general nonlinear initial data which contain multiple shocks and for which shocks form at a later time and/or merge.

Key words. conservation law, hyperbolic, linearized, adjoint, numerical analysis

AMS subject classifications. 65M12, 65M08

DOI. 10.1137/09078078X

1. Introduction.

In this paper we continue the analysis of discrete approximations to linearized and adjoint equations for an unsteady one-dimensional hyperbolic conservation law with a convex flux function. We consider a modified Lax–Friedrichs scheme with numerical viscosity of order $O(h^\alpha)$, $2/3 < \alpha < 1$, for spatial grid size h together with the corresponding linearized and adjoint schemes. Moreover, we analyze an output functional of tracking type and its linearization.

Part 1 [GU10] proved that for a particular form of initial data for the nonlinear equation, and for smooth initial data for the linearized equation, the linearized discrete approximation converges almost everywhere, and the corresponding discrete linearized functional also converges to the correct analytical value. In this paper we extend the analysis to Dirac initial data for the linear equation. From this it is deduced that the discrete adjoint approximation must converge to the analytic adjoint solution, as $h \to 0$, everywhere except along two characteristics, across which it is discontinuous. Moreover, the results of Part 1 [GU10] and the present paper are extended to more general initial data for the nonlinear equation.

1.1. Numerical results.

The model problem and numerical discretizations are the same as described previously in Part 1 [GU10]. Here we present some numerical results for the Burgers equation with flux $f(u) \equiv \frac{1}{2}u^2$.

*Received by the editors June 16, 2008; accepted for publication (in revised form) March 25, 2010; published electronically June 29, 2010.
http://www.siam.org/journals/sinum/48-3/78078.html
†Mathematical Institute, University of Oxford, Oxford, OX1 3LB, United Kingdom (mike.giles@maths.ox.ac.uk).
‡Department of Mathematics, TU Darmstadt, 64289 Darmstadt, Germany (ulbrich@mathematik.tu-darmstadt.de). This author's research was supported by the DFG within the SFB 568 and the SPP 1253.

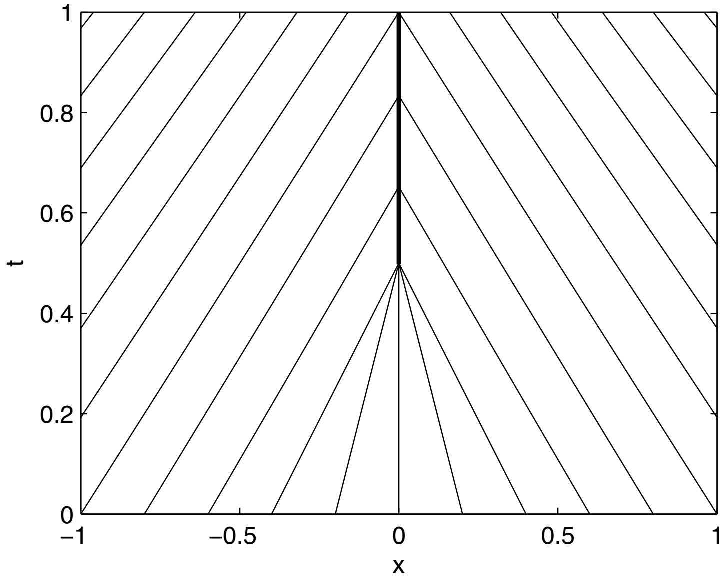

FIG. 1.1. *Characteristics of test problem, with shock formation at $t = 0.5$.*

The continuous piecewise linear initial data for the nonlinear equation is

$$u(x,0) = \begin{cases} 1 - \frac{1}{5}(x+1), & x \le -\frac{4}{9}, \\ -2x, & |x| < \frac{4}{9}, \\ -1 - \frac{1}{5}(x-1), & x \ge \frac{4}{9}. \end{cases}$$

This leads to the formation of a stationary shock at $x = 0$ at time $t = 0.5$, as shown in Figure 1.1, with $u(0^-, 1) = 1$ and $u(0^+, 1) = -1$ being the solution values on either side of the shock at the final time $t = 1$.

The initial data for the linear equation is the Normal distribution

$$\widetilde{u}(x,0) = \frac{1}{0.1\sqrt{2\pi}} \exp\left(-\frac{(x+0.1)^2}{0.02}\right).$$

The output functional

$$J = \int_{-\infty}^{\infty} \gamma(x)\, G(u(x,T))\, \mathrm{d}x$$

is defined by $\gamma(x) = 1$ on the interval $[-1, 1]$, and $G(u) \equiv u^5 - u$. Note that $[G(u)]_1$, the jump in $G(u)$ across the shock at the final time, is zero, and hence, from the analysis in the introduction to Part 1 [GU10], the analytic adjoint solution is zero in the shock region defined by the characteristics which intersect the shock.

To assess the degree to which the solutions are grid-converged, numerical results are obtained on two uniform grids with $h = 0.005, 0.01$. The corresponding timesteps are chosen to be $k = 0.2\, h^2/\varepsilon$.

The computations use the discretizations given in the previous section, but with $\varepsilon = \mu h$, with μ held fixed as the grid is refined. Figure 1.2 shows nonlinear, linear,

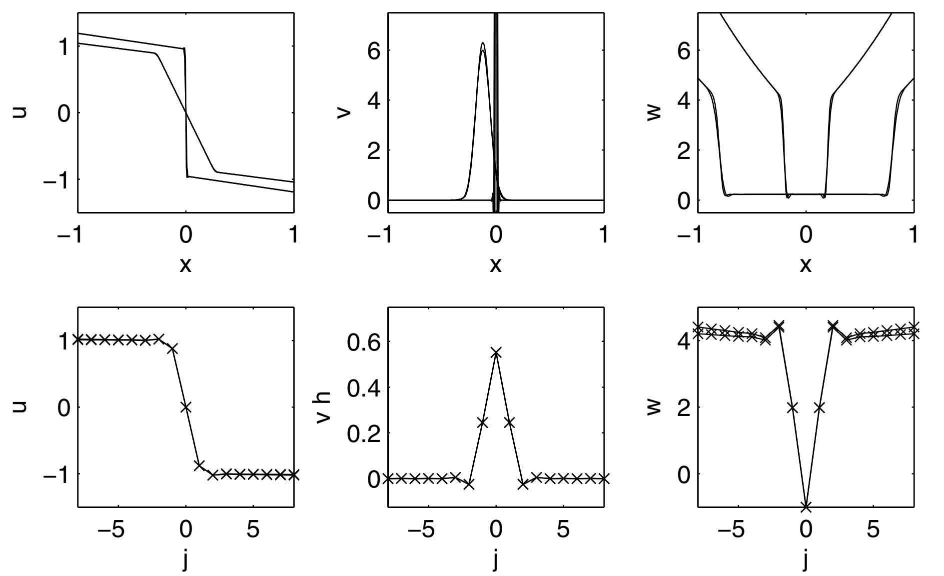

FIG. 1.2. *Nonlinear (u), linear (v), and adjoint (w) solutions using* $\mu = 0.35$*, at times* $t = 0.2, 0.8$ *and plotted versus* x *in upper plots, and at time* $t = 1$ *and plotted versus index* j *in lower plots. Results are plotted for two different grid resolutions.*

and adjoint results for $\mu = 0.35$. The top plots show results at times $t = 0.2, 0.8$ and are plotted versus the coordinate x. The bottom plots present results at the final time $t = 1$ and are plotted versus the index j, with $j = 0$ corresponding to $x = 0$. The nonlinear results on the two grids are almost identical. In the vicinity of the shock, the nonlinear solution is very close to a self-similar steady-state solution when plotted versus j, with the discrete shock profile depending solely on μ. With this level of smoothing there are very few grid points in the shock.

Similarly, the linear solution on the two grids is almost identical away from the shock and has an apparent self-similar form near the shock, when suitably scaled. For reasons that will be explained later, the linear solution is a discrete approximation to a delta function, and so its width and height are proportional to h and h^{-1}, respectively. This is the reason for multiplying v by h before plotting it versus j in the figure.

The discrete adjoint solutions also look grid-converged, and the solution is approximately constant in the shock region. However, its value is nonzero, indicating that the grid convergence is to an incorrect value.

Figure 1.3 shows results for $\mu = 1.0$. These do not appear substantially different except for the fact that there are now many grid points across the shock and therefore fairly good resolution of the differing values of $G'(u)$ for u ranging from 1 on the left of the shock to -1 on the right of the shock. Consequently, the adjoint solution is now almost perfectly zero in the shock region.

The importance of the smoothing level μ is quantified in Figure 1.4, which plots the error in the computed value for the linearized output functional. To within plotting accuracy, this is equal to the error in the adjoint solution $w(x, 0)$ in the central part of the shock region. The left-hand plot uses a log scale for the error and plots the error versus the smoothing coefficient μ. It appears from these results that for small values of μ, the error decreases exponentially with μ. The right-hand plot in Figure 1.4

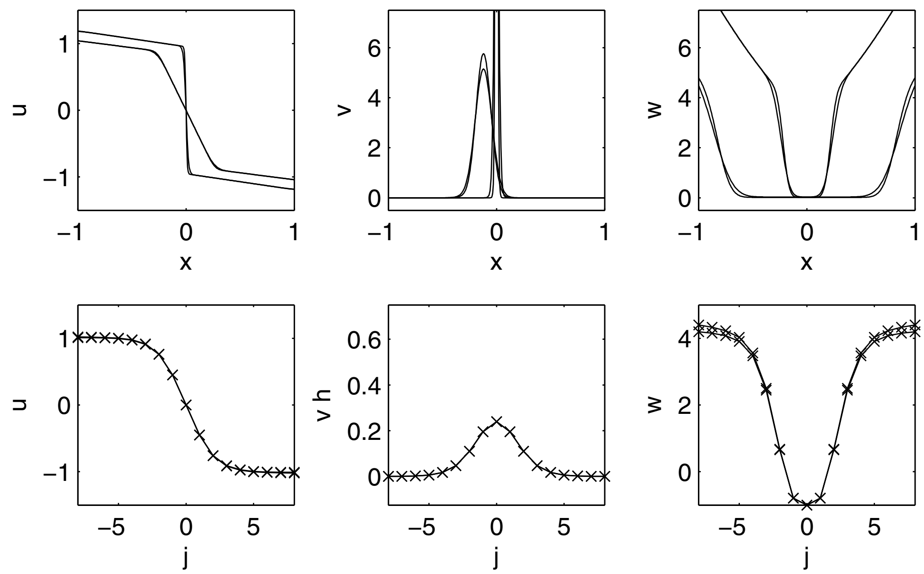

FIG. 1.3. *Nonlinear (u), linear (v), and adjoint (w) solutions using $\mu = 1.0$, at times $t = 0.2, 0.8$ and plotted versus x in upper plots, and at time $t = 1$ and plotted versus index j in lower plots. Results are plotted for two different grid resolutions.*

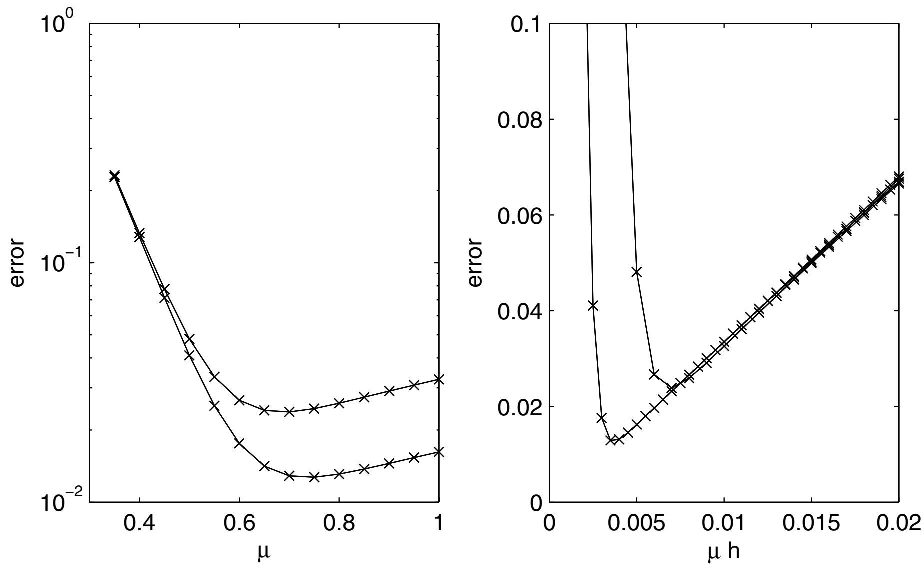

FIG. 1.4. *Error in linearized output functional as a function of numerical smoothing coefficient μ and $\varepsilon = \mu h$. Results are plotted for two different grid resolutions.*

replots the same data versus $\varepsilon \equiv \mu h$. If ε is held fixed, the numerical discretization can be viewed as a consistent approximation of the viscous Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \varepsilon\,\frac{\partial^2 u}{\partial x^2}.$$

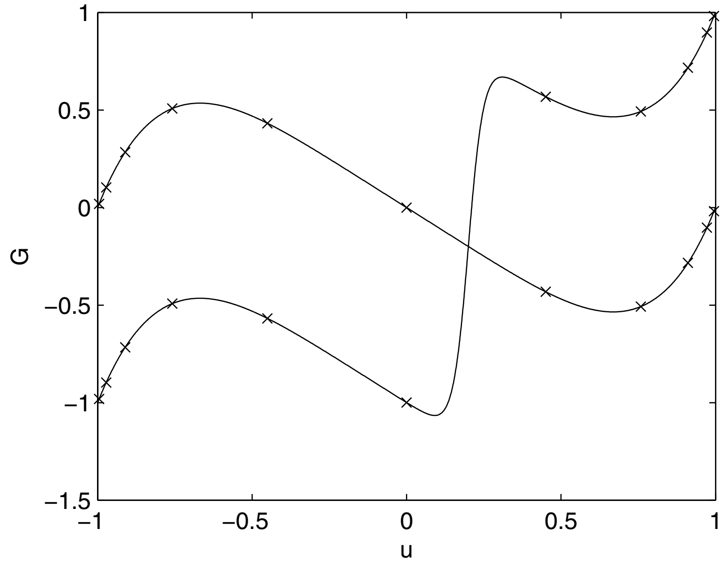

FIG. 1.5. *Two objective functions $G(u)$ with sampling points due to the discrete shock profile for $\mu = 1.0$.*

The fact that there is little difference between the two curves for the different grid resolutions for $\varepsilon > 0.01$ indicates that these are almost grid-converged results for the viscous approximation, and thus the viscous problem has a linearized functional which differs from the inviscid value by an amount which is approximately proportional to ε. This will be confirmed later by asymptotic analysis.

Combining the findings from the two plots, it appears that the error in the approximation of the linearized inviscid functional, and also the adjoint solution in the shock region, is of the approximate form

$$c_1 \exp(-c_2\, \mu) + c_3\, \mu\, h \;\equiv\; c_1 \exp(-c_2\, \varepsilon/h) + c_3\, \varepsilon$$

for appropriate constants c_1, c_2, c_3. The desire to minimize this error is the reason for the choice $\varepsilon = h^\alpha$ with α just slightly less than 1.

The clear conclusion from these numerical results is that it is necessary that, as the grid resolution improves, the numerical smoothing varies in a way which increases the number of points across the shock, while at the same time the overall width of the shock decreases. To understand why this is a fundamental requirement, we need to consider the information used in computing the linearized output functional. The analytic solution depends on the jump $[G(u)]$ across the shock at the final time $t = 1$. However, the linearized discrete approximation uses the values of $G'(u)$ for the final values of u obtained from the nonlinear calculation. This means that the discrete approximation must implicitly approximate $[G(u)]$ by approximating the integration of $G'(u)$ across the smeared shock. For this to be done accurately requires adequate resolution of the variation in $G'(u)$.

This point is illustrated in Figure 1.5. The smoother of the two curves is $G(u) = u^5 - u$, which is the objective function used in the numerical experiments. The symbols correspond to the values of u at the final time $t = 1$ computed using the central difference flux with smoothing coefficient $\mu = 1.0$, which was used for the results in Figure 1.3. The second curve is $G(u) = u^5 - u + \tanh 20(u - 0.2)$. This function has almost identical gradient values at the indicated sampling points and therefore produces numerical values for the linearized output functional which are visually indistinguishable from those in Figure 1.4. However, the analytic solution has a different jump in $G(u)$ across the shock, and so the analytic solution is quite different. This shows that for any numerical discretization with a fixed number of points across the shock, it is easy to construct a linearized output functional for which the numerical approximation will not converge to the true value.

1.2. Outline of paper. The main objective of the paper is to prove that the discrete adjoint solution converges pointwise almost everywhere to the adjoint state of the continuous problem. To this end we will show that the linear functional $\widetilde{J}_h$ converges to the analytic value $\widetilde{J}$ as $h \to 0$ for Dirac initial perturbations. In the final part of the paper we will extend these results to more general nonlinear initial data to include cases in which shocks form at a later time and/or merge.

Our analysis builds on the results of Part 1 [GU10], where approximations to both U_j^n and $\widetilde{U}_j^n$ have been constructed by using the technique of matched inner and outer asymptotic expansions. Together with discrete stability estimates to bound the errors in the asymptotic approximations, this has enabled us to show $\widetilde{J}_h \to \widetilde{J}$ as $h \to 0$ for smooth initial perturbations $\widetilde{u}_0$.

The paper is organized as follows.

- Section 2 derives stability estimates for the discrete adjoint scheme and two results concerning cumulative sums of the nonlinear and linear solutions; these are used in sections 3 and 4.
- Section 3 extends the analysis of Part 1 [GU10] to linear problems with Dirac initial data, with a particular focus on the region in which characteristics lead into the shock. From this it is concluded that the discrete adjoint approximation converges within this shock region.
- Section 4 further extends the analysis by considering more general nonlinear initial data. It explains why the special choice of nonlinear initial data in Part 1 [GU10] is not critical to the final convergence result and also extends the analysis to problems with multiple shocks.

2. Discrete stability estimates. In Part 1 [GU10] we have already derived stability estimates for the nonlinear scheme and its linearization. We derive now stability estimates for the discrete adjoint scheme and cumulative sums of U_j^n and $\widetilde{U}_j^n$.

2.1. Adjoint equations.

THEOREM 2.1. *Suppose that U_j^n is a solution of the nonlinear discrete equations*

$$U_j^{n+1} = U_j^n - \tfrac{1}{2} r \left(f(U_{j+1}^n) - f(U_{j-1}^n) \right) + \varepsilon\, d \left(U_{j+1}^n - 2U_j^n + U_{j-1}^n \right), \tag{2.1}$$

where $f(u)$ is a C^∞ function with derivative $a(u) = f'(u)$, and $r = k/h$, $d = k/h^2$, $\varepsilon = h^\alpha$, $0 < \alpha < 1$, and subject to specified initial data U_j^0 with L_∞ bound U_∞.

Furthermore, let V_j^n be an approximation to U_j^n which satisfies the equation

$$V_j^{n+1} = V_j^n - \tfrac{1}{2} r \left(f(V_{j+1}^n) - f(V_{j-1}^n) \right) + \varepsilon\, d \left(V_{j+1}^n - 2V_j^n + V_{j-1}^n \right) + k\, \tau_j^n,$$

and the same initial data U_j^0, and let U_∞ also be an upper bound on $\|V_j^n\|_\infty$.

Let W_j^n be a solution of the adjoint difference equation

$$W_j^n = W_j^{n+1} + \tfrac{1}{2} r\, a(U_j^n) \left(W_{j+1}^{n+1} - W_{j-1}^{n+1} \right) + \varepsilon\, d \left(W_{j+1}^{n+1} - 2W_j^{n+1} + W_{j-1}^{n+1} \right),$$

and let Z_j^n be an approximation to it which satisfies the equation

$$Z_j^n = Z_j^{n+1} + \tfrac{1}{2}\, r\, a(V_j^n) \left(Z_{j+1}^{n+1} - Z_{j-1}^{n+1}\right) + \varepsilon\, d \left(Z_{j+1}^{n+1} - 2Z_j^{n+1} + Z_{j-1}^{n+1}\right) + k\, \widehat{\tau}_j^n,$$

with initial data Z_j^N which may differ from W_j^N.

Then, provided that $h < (2/A_\infty)^{1/(1-\alpha)}$ where $A_\infty = \sup_{|u|<U_\infty} |a(u)|$, and k is chosen so that $\varepsilon d = c$ for some constant $c < \frac{1}{2}$, the difference $\widehat{E}_j^n = Z_j^n - W_j^n$ satisfies the bound

$$\|\widehat{E}^n\|_\infty \;\le\; \|\widehat{E}^N\|_\infty + h^{-2}\,(t^N - t^n)\, B_\infty W_\infty \|\tau\|_{1,n} + \|\widehat{\tau}\|_{\infty,n},$$

where $B_\infty = \sup_{|u|<U_\infty} |a'(u)|$, W_∞ is an upper bound for $|W_j^N|$, and

$$\|\widehat{\tau}\|_{\infty,n} = k \sum_{m=n}^{N-1} \|\widehat{\tau}^m\|_\infty.$$

Proof. Defining B_j^n as

$$B_j^n = \begin{cases} \dfrac{a(V_j^n) - a(U_j^n)}{V_j^n - U_j^n}, & V_j^n \ne U_j^n, \\ a'(U_j^n), & V_j^n = U_j^n, \end{cases}$$

with bound $|B_j^n| < B_\infty$, the difference $\widehat{E}_j^n = Z_j^n - W_j^n$ satisfies the equation

$$\begin{aligned} \widehat{E}_j^n = \widehat{E}_j^{n+1} &+ \tfrac{1}{2}\, r\, a(V_j^n) \left(\widehat{E}_{j+1}^{n+1} - \widehat{E}_{j-1}^{n+1}\right) + \varepsilon\, d \left(\widehat{E}_{j+1}^{n+1} - 2\widehat{E}_j^{n+1} + \widehat{E}_{j-1}^{n+1}\right) \\ &+ \tfrac{1}{2}\, r\, B_j^n E_j^n \left(W_{j+1}^{n+1} - W_{j-1}^{n+1}\right) + k\, \widehat{\tau}_j^n. \end{aligned}$$

Since $|E_j^n| \le h^{-1} \|E^n\|_1$, it follows that

$$\|\widehat{E}^n\|_\infty \le \|\widehat{E}^{n+1}\|_\infty + r\, h^{-1} B_\infty W_\infty \|\tau\|_{1,n} + k\, \|\widehat{\tau}^n\|_\infty,$$

and therefore

$$\|\widehat{E}^n\|_\infty \;\le\; \|\widehat{E}^N\|_\infty + h^{-2}\,(t^N - t^n)\, B_\infty W_\infty \|\tau\|_{1,n} + \|\widehat{\tau}\|_{\infty,n}. \qquad \square$$

2.2. Nonlinear cumulative sums. The next result concerns the cumulative sums

$$C_j^n = h \sum_{k=-\infty}^{j} (U_k^n - U_{-\infty}),$$

which are well defined if it is known that $U_k^n = U_{-\infty}$ for k sufficiently negative.

THEOREM 2.2. *If ${}^{(1)}U_j^n$ and ${}^{(2)}U_j^n$ are two solutions of the nonlinear discretization subject to initial data with ${}^{(1)}U_j^0 = {}^{(2)}U_j^0 = U_{-\infty}$ for $j < 0$, and h and k satisfy the same conditions as in Theorem 2.1, plus the additional restriction*

$$h < \left(\frac{1-2c}{c\, A_\infty}\right)^{1/(1-\alpha)},$$

where c, A_∞ are as defined in Theorem 2.1, then ${}^{(2)}C_j^n \ge {}^{(1)}C_j^n$ $\forall j, n > 0$ if ${}^{(2)}C_j^0 \ge {}^{(1)}C_j^0$ $\forall j$.

Proof. Defining

$$\Delta U_j^n = {}^{(2)}U_j^n - {}^{(1)}U_j^n,$$

and then following the same reasoning as in the proof of [GU10, Thm. 2.1], we obtain the difference equation

$$\Delta U_j^{n+1} = \Delta U_j^n - \tfrac{1}{2}\, r \left(A_{j+1}^n \Delta U_{j+1}^n - A_{j-1}^n \Delta U_{j-1}^n \right) + \varepsilon d \left(\Delta U_{j+1}^n - 2\Delta U_j^n + \Delta U_{j-1}^n \right),$$

where

$$A_j^n = \begin{cases} \dfrac{f({}^{(2)}U_j^n) - f({}^{(1)}U_j^n)}{{}^{(2)}U_j^n - {}^{(1)}U_j^n}, & {}^{(2)}U_j^n \neq {}^{(1)}U_j^n, \\ a({}^{(1)}U_j^n), & {}^{(2)}U_j^n = {}^{(1)}U_j^n. \end{cases}$$

Summing the discrete difference equation yields

$$\Delta C_j^{n+1} = \Delta C_j^n + \left(\varepsilon\, d - \tfrac{1}{2}\, r\, A_{j+1}^n\right) h\, \Delta U_{j+1}^n - \left(\varepsilon\, d + \tfrac{1}{2}\, r\, A_j^n\right) h\, \Delta U_j^n,$$

where $\Delta C_j^n \equiv h \sum_{k=-\infty}^{j} \Delta U_k^n = {}^{(2)}C_j^n - {}^{(1)}C_j^n$.

Now, defining

$$b_j^n = \varepsilon\, d - \tfrac{1}{2}\, r\, A_{j+1}^n, \qquad c_j^n = \varepsilon\, d + \tfrac{1}{2}\, r\, A_j^n,$$

under the additional conditions on h and k, the three terms $1 - b_j^n - c_j^n$, b_j^n, and c_j^n are all strictly positive.

Making the substitution $h\Delta U_j^n = \Delta C_j^n - \Delta C_{j-1}^n$, it follows that if $\Delta C_j^n \geq 0\ \forall j$, then

$$\begin{aligned} \Delta C_j^{n+1} &= \Delta C_j^n + b_j^n(\Delta C_{j+1}^n - \Delta C_j^n) - c_j^n(\Delta C_j^n - \Delta C_{j-1}^n) \\ &= (1 - b_j^n - c_j^n)\Delta C_j^n + b_j^n \Delta C_{j+1}^n + c_j^n \Delta C_{j-1}^n \\ &\geq 0. \end{aligned}$$

Thus, if $\Delta C_j^0 \geq 0\ \forall j$, then $\Delta C_j^n \geq 0\ \forall j, n$. □

2.3. Linear cumulative sums. The final result in this section concerns the linear cumulative sums

$$\widetilde{C}_j^n = h \sum_{k=-\infty}^{j} \widetilde{U}_k^n,$$

which are well defined if $\widetilde{U}_j^n$ has compact support.

THEOREM 2.3. *If U_j^n is defined as in Theorem 2.1, $\widetilde{U}_j^n$ is a solution of the linearized difference equation*

$$\widetilde{U}_j^{n+1} = \widetilde{U}_j^n - \tfrac{1}{2}\, r \left(a(U_{j+1}^n)\, \widetilde{U}_{j+1}^n - a(U_{j-1}^n)\, \widetilde{U}_{j-1}^n \right) + \varepsilon\, d \left(\widetilde{U}_{j+1}^n - 2\widetilde{U}_j^n + \widetilde{U}_{j-1}^n \right)$$

with initial data $\widetilde{U}_j^0$, and h and k satisfy the same conditions as in Theorem 2.2, then $\widetilde{C}_j^n \geq 0\ \forall j, n > 0$ if $\widetilde{C}_j^0 \geq 0\ \forall j$.

Proof. Summing the linear discrete equations yields

$$\widetilde{C}_j^{n+1} = \widetilde{C}_j^n + \left(\varepsilon\, d - \tfrac{1}{2}\, r\, a(U_{j+1}^n)\right) h\, \widetilde{U}_{j+1}^n - \left(\varepsilon\, d + \tfrac{1}{2}\, r\, a(U_j^n)\right) h\, \widetilde{U}_j^n.$$

Defining

$$b_j^n = \varepsilon\, d - \tfrac{1}{2}\, r\, a(U_{j+1}^n), \qquad c_j^n = \varepsilon\, d + \tfrac{1}{2}\, r\, a(U_j^n),$$

under the conditions of Theorem 2.1, b_j^n and c_j^n are both strictly positive. Furthermore, the new additional restriction on h ensures that $rA_\infty < 1 - 2c$, and hence

$$b_j^n + c_j^n = 2c - \tfrac{1}{2}\, r \left(a(U_{j+1}^n) - a(U_j^n)\right) \le 1.$$

Making the substitution $h\widetilde{U}_j^n = \widetilde{C}_j^n - \widetilde{C}_{j-1}^n$, it follows that if $\widetilde{C}_j^n \ge 0\ \forall j$, then

$$\begin{aligned}\widetilde{C}_j^{n+1} &= \widetilde{C}_j^n + b_j^n(\widetilde{C}_{j+1}^n - \widetilde{C}_j^n) - c_j^n(\widetilde{C}_j^n - \widetilde{C}_{j-1}^n)\\ &\ge (1 - b_j^n - c_j^n)\widetilde{C}_j^n\\ &\ge 0.\end{aligned}$$

Thus, if $\widetilde{C}_j^0 \ge 0\ \forall j$, then $\widetilde{C}_j^n \ge 0\ \forall j, n$. ∎

COROLLARY 2.4. *If* ${}^{(1)}\widetilde{U}_j^n$, ${}^{(2)}\widetilde{U}_j^n$, ${}^{(3)}\widetilde{U}_j^n$ *are three solutions of the linear equations for the same nonlinear solution* U_j^n, *and* h *and* k *satisfy the conditions in Theorem* 2.3, *and*

$${}^{(1)}\widetilde{C}_j^0 \le {}^{(2)}\widetilde{C}_j^0 \le {}^{(3)}\widetilde{C}_j^0 \quad \forall j,$$

then

$${}^{(1)}\widetilde{C}_j^n \le {}^{(2)}\widetilde{C}_j^n \le {}^{(3)}\widetilde{C}_j^n \quad \forall j, n.$$

Proof. The proof follows immediately due to linearity. ∎

3. Dirac linear initial data and convergence of the discrete adjoint. In this section we consider the extension of the analysis in Part 1 [GU10] to the situation in which the linear problem has Dirac initial data. As explained in the introduction to Part 1, the value of the adjoint solution W_j^0 at the initial time $t = 0$ is identically equal to the value of the linearized output functional at the final time in response to Dirac initial data at point j. Thus, proving the convergence of the linearized output functional due to Dirac initial data is equivalent to proving the pointwise convergence of the discrete adjoint solution.

As in Part 1 [GU10], the initial data is assumed to satisfy the following conditions:

(A1) Apart from a discontinuity at $x_s(0)$, $u_0(x)$ is C^∞ with all derivatives having a finite L_1 norm over $(-\infty, x_s(0))$ and $(x_s(0), \infty)$.

(A2) The discontinuity has finite strength for the entire time interval $[0, T]$, and no other discontinuity is formed during this time interval.

These assumptions will be relaxed considerably in section 4.

We recall that in Part 1 [GU10] we used matched asymptotic analysis to construct functions V and $\widetilde{V}$, such that $V_j^n = V(x_j, t^n)$ and $\widetilde{V}_j^n = \widetilde{V}(x_j, t^n)$ are approximate solutions of the nonlinear and linearized schemes, respectively. The structure of V and $\widetilde{V}$ has then allowed us to conclude that $V \to U$ and $\widetilde{V} \to \widetilde{U}$ almost everywhere and $\widetilde{J}_h \to \widetilde{J}$ as $h \to 0$. Our matched asymptotic analysis breaks the domain into three overlapping regions:

A: $x_s - x > \varepsilon^\beta$.

B: $|x - x_s| < 2\varepsilon^\beta$.

C: $x - x_s > \varepsilon^\beta$.

Here, $\frac{2}{3} < \alpha < 1$, $\varepsilon = h^\alpha$, and $\beta < 1$ is sufficiently large. A lower bound on β was determined in [GU10, Thm. 4.2].

Note that [GU10, Thm. 4.3] used carefully constructed initial data $U_j^0 = V(x_j, 0)$ for the nonlinear discrete equations, and this remains key to the analysis in this section. Section 4 will extend the analysis to include other initial data, including the more natural choice $u_0(x_j)$.

3.1. Convergence in the shock region. In this section we are specifically concerned with Dirac initial data at a point x_0 lying within the shock region interval $[x_l, x_r]$ from which characteristics enter the shock.

THEOREM 3.1. *If the nonlinear initial data is as defined in* [GU10, Thm. 4.3], *and $\widetilde{J}_h$ is the discrete linear functional due to Dirac initial data at x_0, then $\widetilde{J}_h - \widetilde{J} = O(\varepsilon)$.*

Proof. The proof is based on Corollary 2.4, letting ${}^{(2)}\widetilde{U}_j^n$ be the solution corresponding to the Dirac initial data at x_0. Defining ${}^{(1)}D(x)$ to be a nonnegative C^∞ function with compact support in the interval $[x_l, x_0]$ and unit integral, ${}^{(1)}\widetilde{U}_j^n$ is taken to be the numerical solution with initial data

$$ {}^{(1)}\widetilde{U}_j^0 = \left(h \sum_k {}^{(1)}D(x_k) \right)^{-1} {}^{(1)}D(x_j), $$

corresponding to analytic initial data ${}^{(1)}\widetilde{U}(x,0) = {}^{(1)}D(x)$. Note that because of the Euler–Maclaurin error formula [SB80] the normalization factor

$$ \left(h \sum_k {}^{(1)}D(x_k) \right)^{-1} $$

is equal to $1 + o(h^q)$ $\forall q > 0$.

${}^{(3)}\widetilde{U}_j^n$ is defined similarly using a function ${}^{(3)}D(x)$ with compact support in the interval $[x_0, x_r]$. By construction, the cumulative sums of these three solutions satisfy the conditions of Corollary 2.4, and therefore

$$ {}^{(1)}\widetilde{C}_j^N \le {}^{(2)}\widetilde{C}_j^N \le {}^{(3)}\widetilde{C}_j^N \quad \forall j. $$

Since ${}^{(1)}\widetilde{U}_j^N$ and ${}^{(3)}\widetilde{U}_j^N$ correspond to smooth initial data, the earlier asymptotic analysis in Part 1 [GU10, Thms. 4.4–4.6] is applicable. Let t_P be the critical time at which the last of the characteristics coming from the compact support of the initial data for the two solutions reaches the shock. Beyond this time, the outer approximations will be identically zero; see the proof of [GU10, Thm. 4.4]. Moreover, the proof of [GU10, Thm. 4.5] shows that the inner approximations satisfy the same equations with the same homogeneous boundary conditions and must have the same values for $\widetilde{x}_s(t)$. Hence, ${}^{(1)}\widetilde{V}_j^N$ and ${}^{(3)}\widetilde{V}_j^N$ are identical, and therefore

$$ \left\| {}^{(3)}\widetilde{C}_j^N - {}^{(1)}\widetilde{C}_j^N \right\|_\infty \le \left\| {}^{(3)}\widetilde{U}_j^N - {}^{(1)}\widetilde{U}_j^N \right\|_1 = O(\varepsilon), $$

from which it follows that

$$\left\| {}^{(2)}\widetilde{C}_j^N - {}^{(1)}\widetilde{C}_j^N \right\|_\infty = O(\varepsilon).$$

Using summation by parts, we obtain

$$\begin{aligned}
{}^{(2)}\widetilde{J}_h - {}^{(1)}\widetilde{J}_h &= h \sum_j \gamma(x_j)\, G'(U_j^N) \left({}^{(2)}\widetilde{U}_j^N - {}^{(1)}\widetilde{U}_j^N \right) \\
&= \sum_j - \left(\gamma(x_{j+1})\, G'(U_{j+1}^N) - \gamma(x_j)\, G'(U_j^N) \right) \left({}^{(2)}\widetilde{C}_j^N - {}^{(1)}\widetilde{C}_j^N \right) \\
&= \sum_j - \frac{\gamma(x_{j+1}) + \gamma(x_j)}{2} \left(G'(U_{j+1}^N) - G'(U_j^N) \right) \left({}^{(2)}\widetilde{C}_j^N - {}^{(1)}\widetilde{C}_j^N \right) \\
&\quad + \sum_j - \left(\gamma(x_{j+1}) - \gamma(x_j) \right) \frac{G'(U_{j+1}^N) + G'(U_j^N)}{2} \left({}^{(2)}\widetilde{C}_j^N - {}^{(1)}\widetilde{C}_j^N \right).
\end{aligned}$$

The total variation of U_j^N is uniformly bounded because of the TVD property of the monotone discretization and the bounded variation of the initial data. Also, the weighting function γ is specified to have bounded variation. Hence,

$$\left| {}^{(2)}\widetilde{J}_h - {}^{(1)}\widetilde{J}_h \right| \;\le\; (c_1 + c_2) \left\| {}^{(2)}\widetilde{C}_j^N - {}^{(1)}\widetilde{C}_j^N \right\|_\infty \;=\; O(\varepsilon),$$

where

$$c_1 = \|\gamma\|_\infty \max_{|u| \le U_\infty} |G''(u)|\, TV(u_0),$$

and

$$c_2 = TV(\gamma) \max_{|u| \le U_\infty} |G'(u)|.$$

Since the analytic values ${}^{(1)}\widetilde{J}$, ${}^{(2)}\widetilde{J}$, and ${}^{(3)}\widetilde{J}$ are all equal, and the earlier asymptotic analysis in [GU10, Thm. 5.1] proves that ${}^{(1)}\widetilde{J}_h - {}^{(1)}\widetilde{J} = O(\varepsilon)$, we obtain the final result that

$${}^{(2)}\widetilde{J}_h - {}^{(2)}\widetilde{J} = O(\varepsilon),$$

so the error in the discrete approximation to the linearized functional due to Dirac initial data in the shock region is $O(\varepsilon)$. □

Extending the asymptotic analysis in [GU10, Thm. 4.5] to higher levels of expansion in powers of h, the inner approximations for ${}^{(1)}\widetilde{U}_j^n$ and ${}^{(3)}\widetilde{U}_j^n$ after time t_P must be identical at each level of the expansion since, due to [GU10, (1.8)],

$$h \sum_{j=-\infty}^{\infty} {}^{(1)}\widetilde{U}_j^n = h \sum_{j=-\infty}^{\infty} {}^{(3)}\widetilde{U}_j^n = 1.$$

Therefore, at the final time T, ${}^{(1)}\widetilde{U}_j^N - {}^{(3)}\widetilde{U}_j^N = o(h^q)$ for any $q > 0$. Consequently, ${}^{(1)}\widetilde{J}_h - {}^{(3)}\widetilde{J}_h = o(h^q)$, and hence the discrete adjoint solution within the shock region (more precisely, within any subdomain bounded away from its two bounding characteristics) is constant to within $o(h^q)$ for any $q > 0$, in addition to being equal to the analytic value to within $O(\varepsilon)$.

3.2. Convergence outside the shock region. Now we consider the case in which the Dirac initial data is specified at a point x_0 outside the shock region.

We can split the weighting function $\gamma(x)$ into two C^∞ components, $\gamma_1(x)$, which has compact support in a small neighborhood of the shock, bounded away from the characteristic extending from x_0, and $\gamma_2(x)$, which is zero in a small neighborhood of the shock.

Considering γ_1 first, one can follow an argument similar to that in the previous section, with ${}^{(1)}\widetilde{U}_j^N$ and ${}^{(3)}\widetilde{U}_j^N$ being chosen to correspond to smooth initial data with compact support in a small neighborhood of x_0, so that the asymptotic approximation to their linear functionals is zero because the approximate solution is zero along the characteristics leading into the compact support of γ_1. From this one can conclude with [GU10, Thm. 5.1] as in the previous section that the error in the linear functional for the Dirac initial data is $O(\varepsilon)$.

Considering γ_2 next, in this case the analytic adjoint solution is smooth, and zero in the neighborhood of the shock. One can construct asymptotic approximations to both the nonlinear and adjoint solutions and use Theorem 2.1 to deduce that the L_∞ error in the adjoint approximation, and hence the error in the linear functional for the Dirac initial data, is $O(\varepsilon)$.

Adding the contributions from the two components, one reaches the conclusion that the error in the approximation of the original linear functional due to the Dirac initial data is $O(\varepsilon)$, and hence the error in the adjoint solution is $O(\varepsilon)$.

3.3. On the boundary of the shock region. The analytic adjoint solution is in general discontinuous across the characteristic which defines the boundary of the shock region. From standard results on the error analysis of contact discontinuities in convection/diffusion problems, one should expect a numerical boundary layer with width $O(\sqrt{\varepsilon})$ in the neighborhood of this characteristic.

4. Extension to more general nonlinear initial data. The main analysis in Part 1 [GU10] and in the previous section considered a very particular form of initial data for the nonlinear discretization with a single shock in the analytic initial data. In this section, we extend the analysis to much more general nonlinear initial data. The approach we use is similar to the lower and upper bounds employed in the proof of Theorem 3.1 and makes use of the result in Theorem 2.2.

4.1. Single shock. First, suppose that U_j^0 is initial data of the special form considered previously, and consider the perturbed initial data

$$U_j^0 + \varepsilon^\beta \, (D_j - D_{j-1}),$$

where $D_j \equiv D(x_j + \frac{1}{2}h)$ with $D(x)$ being C^∞ and zero everywhere except on a small open interval around the shock, with the characteristics emerging from the two ends meeting at a point P (at time t_P) on the shock, as shown in Figure 4.1. The usual asymptotic analysis in [GU10, Thms. 4.1–4.3] shows that the effect of the perturbation on the asymptotic solution is confined to the characteristics coming out of the open interval on which $D(x)$ is nonzero, the perturbation region in the figure. The fact that

$$\sum_{j=-\infty}^{\infty} (D_j - D_{j-1}) = 0$$

ensures that there is no net perturbation to the shock location after P. By continuing the asymptotic analysis to higher orders, noting that the outer solution is

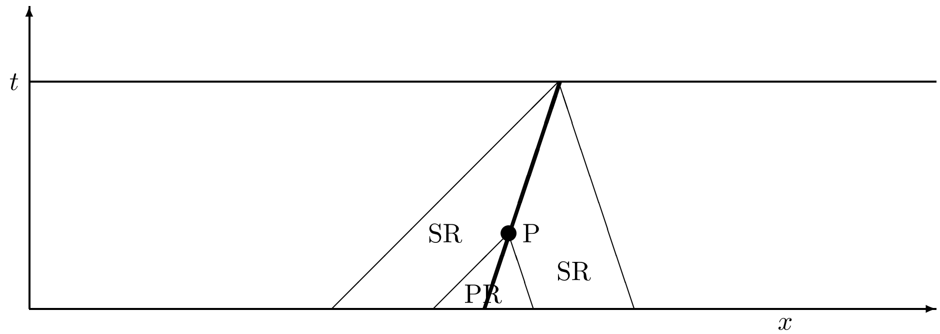

FIG. 4.1. *Illustration of shock perturbation, showing characteristics bounding the shock region (SR) and perturbation region (PR) and point P marking the limit of the influence of the perturbation on the asymptotic approximation.*

identically zero outside the perturbation region, it can be shown that the effect of the perturbation outside this region is $o(h^q)\ \forall\ q > 0$.

Next, suppose that $\overline{U}_j^0$ is another set of initial data which differs from U_j^0 only in the region previously labeled as region B, within a distance $2\varepsilon^\beta$ of the shock. We impose the additional restrictions that $\overline{U}_j^0 - U_j^0 = O(1)$ within this region, and the shock location in the definition of U_j^0 is chosen so that

$$\sum_{k=-\infty}^{\infty} U_k^0 - \overline{U}_k^0 = 0.$$

We now define two new sets of initial data ${}^{(1)}U_j^0$ and ${}^{(2)}U_j^0$ as

$${}^{(m)}U_j^0 = U_j^0 + a_m \varepsilon^\beta \left(D_j - D_{j-1}\right),$$

where D_j is as defined before, but with the additional restriction that it is nonnegative. Defining

$$C_j^n = h \sum_{k=-\infty}^{j} \left(U_k^n - U_{-\infty}\right),$$

$$\overline{C}_j^n = h \sum_{k=-\infty}^{j} \left(\overline{U}_k^n - U_{-\infty}\right),$$

$${}^{(m)}C_j^n = h \sum_{k=-\infty}^{j} \left({}^{(m)}U_k^n - U_{-\infty}\right),$$

we have

$${}^{(m)}C_j^0 = C_j^0 + a_m \varepsilon^\beta D_j$$

and can choose the a_m such that ${}^{(1)}C_j^0 < \overline{C}_j^0 < {}^{(2)}C_j^0$. We can now appeal to the result in Theorem 2.2 to prove that, under the conditions of that theorem, ${}^{(1)}C_j^n < \overline{C}_j^n < {}^{(2)}C_j^n\ \forall$ timesteps n.

However, the discussion at the beginning of this section proved that ${}^{(1)}C_j^n - {}^{(2)}C_j^n = o(h^q)$ after time t_P. Hence, $\overline{C}_j^n - C_j^n = o(h^q)$ after this time as well, and so the difference in initial data between $\overline{U}_j^0$ and U_j^0 has negligible consequence after time t_P.

Turning now to the linearized problem, if we have smooth initial data with compact support inside the shock region, but outside the perturbation region, then the asymptotic form of the linear solution will be the same, regardless of whether the underlying nonlinear solution is due to initial data $\overline{U}_j^0$ or U_j^0. The linearized functional will also be unaffected, and the bounding argument of the previous section can again be used to deduce that the linearized functional is also unaffected for smooth initial data inside the perturbation region. It also follows from the linear bounding argument that the linearized functionals due to Dirac initial data are unaffected, and hence the adjoint approximation is constant within the shock region to within $o(h^q)\ \forall\ q > 0$ and equal to the analytic value to within $O(\varepsilon)$.

4.2. Multiple shocks. The analysis so far has considered initial data giving solutions with a single shock extending from the initial time. We now extend this to a much larger class of solutions by replacing assumptions (A1) and (A2) with the following assumptions:

(A3) $\sup_x |u_0(x)| < \infty$.

(A4) Defining X_0 to be the set of points x_0 for which the characteristics propagate up to the final time T without entering a shock, i.e.,

$$X_0 = \{x : u(x + a(u_0(x))t, t) = u_0(x),\ 0 < t < T\},$$

there exists an open set X_1 containing $\overline{X_0}$ such that $u_0(x)$ is C^∞ on $\overline{X_1}$, all of its derivatives have a finite L_1 norm on $\overline{X_1}$, and

$$a'(u_0(x))\, u_0'(x) \ > \ -T^{-1} \qquad \forall x \in \overline{X_1}. \tag{4.1}$$

The first assumption gives a uniform bound on the whole solution, while the second assumption ensures that no new shocks form at time T, pre-existing shocks have a smooth behavior in an open neighborhood of T, and between the shocks the solution $u(x, T)$ is smooth.

In outlining the analysis, we will consider the situation illustrated in Figure 4.2 in which there is a single shock at the final time T; the extension to multiple shocks at time T is straightforward. The points x_L and x_R are chosen to lie within X_1 so that the solution is smooth outside the region bounded by the two characteristics which meet at P, and on either side of the shock.

Following the approach of the previous section, we want to construct two different sets of initial data which will lead to upper and lower bounds. The weak form of the nonlinear PDE gives

$$\int_{x_L}^{x_0} u_0(x)\, \mathrm{d}x = \int_0^{t_P} f(u(x_0 + \lambda t, t)) - \lambda\, u(x_0 + \lambda t, t)\, \mathrm{d}t - \int_0^{t_P} f(u_0(x_L)) - a(u_0(x_L))\, u_0(x_L)\, \mathrm{d}t,$$

where x_0 is any point in the interval $[x_L, x_R]$ and the first integral on the right-hand side is along the straight line linking $(x_0, 0)$ to P, for which $x = x_0 + \lambda t$, and

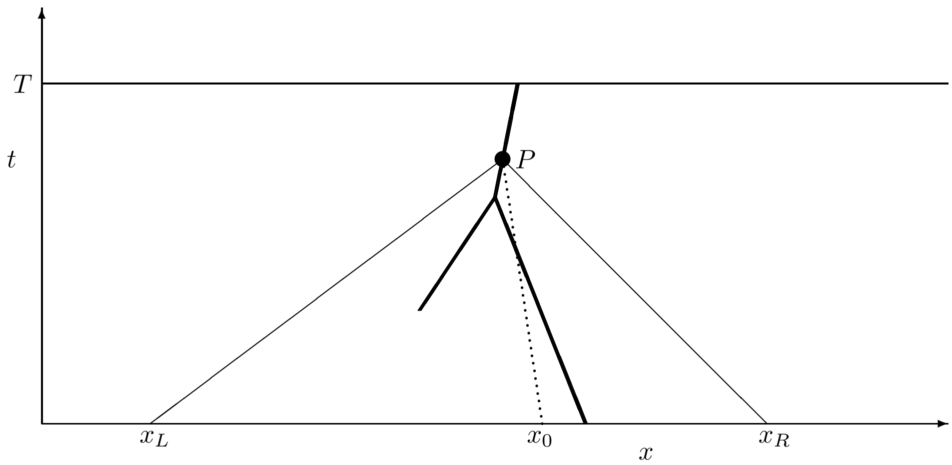

FIG. 4.2. *Illustration of a solution with multiple shocks (thick lines), one forming after $t = 0$, and the characteristics (thin lines) and construction line (dotted) used in the analysis.*

the second integral is along the characteristic linking $(x_L, 0)$ to P. Because of the convexity of the flux function, $f(u(x,t)) - \lambda u(x,t)$ is a minimum when $a(u(x,t)) = \lambda$, which corresponds to the construction line being a characteristic. Hence,

$$\int_{x_L}^{x_0} u_0(x)\,\mathrm{d}x \geq \int_{x_L}^{x_0} u_1(x)\,\mathrm{d}x, \tag{4.2}$$

where $u_1(x)$ is the initial data for which $a(u_1(x))$ varies linearly giving a compression fan leading to a shock forming at time t_P. Note also that when $x_0 = x_R$ we get

$$\int_{x_L}^{x_R} u_0(x)\,\mathrm{d}x = \int_{x_L}^{x_R} u_1(x)\,\mathrm{d}x. \tag{4.3}$$

Outside the interval $[x_L, x_R]$ we define $u_1(x)$ to equal $u_0(x)$.

We also define initial data $u_2(x)$ to equal $u_0(x)$ outside the interval $[x_L, x_R]$, while within it

$$u_2(x) = \begin{cases} \sup_x |u_0(x)|, & x < x_s, \\ -\sup_x |u_0(x)|, & x \geq x_s, \end{cases}$$

with x_s chosen so that

$$\int_{x_L}^{x_R} u_2(x)\,\mathrm{d}x = \int_{x_L}^{x_R} u_0(x)\,\mathrm{d}x. \tag{4.4}$$

By construction this gives

$$\int_{x_L}^{x_0} u_2(x)\,\mathrm{d}x \geq \int_{x_L}^{x_0} u_0(x)\,\mathrm{d}x, \tag{4.5}$$

and it leads to an expansion-fan/shock/expansion-fan combination resulting in a shock at P.

These two sets of initial data $u_1(x)$ and $u_2(x)$ need to be slightly modified before the earlier analysis can be applied. $u_2(x)$ is not smooth at x_L and x_R; this is easily remedied by a small local adjustment at each end without affecting either (4.4) or (4.5). With $u_1(x)$, the main problem is that the initial data does not contain a shock. This is remedied by adding a small perturbation so that

$$a(u_1(x)) = a\,x + b - c\,H\left(x - \tfrac{1}{2}(x_L + x_R)\right) \exp\left(-\left(x - \frac{1}{2}(x_L + x_R)\right)\right)$$

for $0 < c \ll 1$, with $H(\cdot)$ being the Heaviside function. This introduces a very small shock at $\frac{1}{2}(x_L + x_R)$ which will increase in strength until it achieves its full strength at P. Small local adjustments can then be made in the neighborhood of x_L and x_R to ensure the initial data is smooth there, while satisfying (4.2), (4.3), and (4.1).

We are now able to approximate the modified initial data $u_1(x)$ and $u_2(x)$ so that the corresponding discrete initial data ${}^{(1)}U_j^0$ and ${}^{(2)}U_j^0$ have the properties ${}^{(1)}C_j^0 < C_j^0 < {}^{(2)}C_j^0$, where ${}^{(m)}C_j^0$ are as defined in the previous section. Since $u_1(x)$ and $u_2(x)$ give rise to the same analytic solution after time t_P, it follows from the previous analysis that ${}^{(1)}U_j^n - U_j^n = o(h^q)\ \forall q > 0$ after that time, too. It then follows, as before, that the response to smooth initial data for the linearized problem with compact support within the shock region (bounded by the two characteristics which reach the shock at time T) but outside $[x_L, x_R]$ is the same, to $o(h^q)$, regardless of whether the nonlinear initial data comes from ${}^{(1)}U_j^0$ or U_j^0. Hence, to within $o(h^q)$, the linearized functions are identical for both smooth initial data and Dirac initial data, and therefore the discrete adjoint solutions also differ by $o(h^q)$ for any $q > 0$.

5. Conclusions. This paper has continued the convergence analysis in [GU10] of approximate linear and adjoint solutions for a class of convex flux functions using a particular modified Lax–Friedrichs discretization. We have shown that in the case of a single shock, the linear discrete output functional $\widetilde{J}_h$ converges to the correct value $\widetilde{J}$ even for Dirac initial perturbations located in the strict interior or exterior of the shock funnel. From this it follows that the adjoint approximation also converges pointwise everywhere except along the two characteristics at which it is discontinuous. In the final part of the paper, the convergence of the linear output functional and of the adjoint solution is extended to cases with multiple shocks and shock formation/interaction.

To obtain these results we have relied on the facts that (1) the linear discretization is a linearization of the conservative nonlinear discretization; (2) the adjoint discretization is a discrete adjoint of the linear discretization; and (3) the number of mesh points across the smeared shock increases as $h \to 0$. The numerical results indicate that the error arising from the shock region decays exponentially with the number of grid points across which the shock is spread, and we believe that the first two elements are essential for a numerical discretization to have this feature.

As already pointed out in [GU10] the modified Lax–Friedrichs discretization considered in this paper is not practical, since it provides only $O(h^\alpha)$ convergence for $0 < \alpha < 1$. However, adaptive smoothing could be used by reducing the magnitude of ε or using a high order method in the smooth regions on either side of the shocks, while the proposed method is used on an adaptively refined grid in the vicinity of the shock. This combination should give $O(\overline{h}^2)$ convergence for linearized functionals and adjoint solutions, with $\overline{h}$ being the average grid spacing. However, extending the analysis in this paper to such a scheme is likely to be extremely challenging. Therefore, we have preferred to gain basic insight by using a simplified discretization.

As a final comment, a possible direction for future research is to apply the approach in this paper to the analysis of the linear and adjoint discretizations which arise when there is a discontinuous nonlinear source term, as arises in meteorological applications [TC87, Xu96, ZZA01]. It may be possible to prove the convergence of the linear and adjoint approximations when the discontinuity is spread over an increasing number of grid points using a suitable regularization.

REFERENCES

[GU10] M. Giles and S. Ulbrich, *Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part* 1: *Linearized approximations and linearized output functionals*, SIAM J. Numer. Anal., 48 (2010), pp. 882–904.

[SB80] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[TC87] O. Talagrand and P. Courtier, *Variational assimilation of meteorological observations with the adjoint vorticity equation,* I, *Theory*, Q. J. R. Meteorol. Soc., 113 (1987), pp. 1311–1328.

[Xu96] Q. Xu, *Generalized adjoint for physical processes with parameterized discontinuities. Part* I: *Basic issues and heuristic examples*, J. Atmospheric Sci., 33 (1996), pp. 1123–1142.

[ZZA01] S. Zhang, X. Zou, and J. E. Ahlquist, *Examination of numerical results from tangent linear and adjoint of discontinuous nonlinear models*, Monthly Weather Rev., November (2001), pp. 2791–2804.