SIAM J. NUMER. ANAL.
Vol. 49, No. 6, pp. 2386–2406

HEURISTIC MODIFIED EQUATION ANALYSIS ON OSCILLATIONS IN NUMERICAL SOLUTIONS OF CONSERVATION LAWS*

JIEQUAN LI† AND ZHICHENG YANG‡

Abstract. Oscillations are ubiquitous in numerical solutions obtained by high order or even first order schemes for hyperbolic problems and are conventionally understood as the consequence of low dissipation effects of underlying numerical schemes. Earlier analysis was done mainly through the effective discrete Fourier analysis for linear problems or the modified equation approach in smooth solution regions. In this paper, a so-called *heuristic modified equation* is derived when applied to nonlinear problems, particularly for oscillatory modes of solutions whose counterpart in linear problems are high frequency mode solutions, and the dissipation effect is distinguished as a numerical damping and a numerical diffusion. The former is reflected through the zero order term of the heuristic modified equation and the latter through the second order differential term. It turns out that the effect of dissipation is categorized as *a damping, a neutrality, and an amplification,* and that the numerical damping plays a dominant role in offsetting the oscillatory modes. When the amplification effect is taken, the numerical scheme often comes unstable.

Key words. modified equation, oscillations, total variation diminishing schemes, damping, neutrality, amplification

AMS subject classifications. Primary, 65M08, 65N08, 35L65; Secondary, 35L02, 35L40, 76M12

DOI. 10.1137/110822591

1. Introduction.

Oscillations are quite ubiquitous in numerical solutions of time dependent problems, analogous to the Gibbs phenomenon for the partial sums of the Fourier series of discontinuous functions. Lax showed in [6] that oscillations must be present when the solutions are approximated by a scheme that is more than first order accurate. Many researchers are of the impression that the oscillations only appear in the numerical solutions obtained by high order schemes. However, it was disclosed in [1, 2, 13] that even for first order total variation diminishing (TVD) schemes, including the celebrated Lax–Friedrichs scheme, local oscillations are still observed.[1] To be more precise, in [1, 2] special attention was paid to the classical Lax–Friedrichs scheme and its extensions (the Rusanov method and the second order Nessyahu–Tadmor schemes) for investigating local extrema of solutions in order to expound the source of oscillations. The conclusion there is that schemes with a large viscosity coefficient are prone to oscillations at data extrema and the analysis carried out is quite algebraic [1, 2]. Independently, the local oscillations were observed and analyzed for $2K+1$-point ($K \geq 1$) central monotone schemes from the purely computational point of view in [13]. Then the subsequent work [9] started

*Received by the editors January 28, 2011; accepted for publication (in revised form) September 2, 2011; published electronically November 22, 2011.
http://www.siam.org/journals/sinum/49-6/82259.html

†Laboratory of Mathematics and Complex System, Ministry of Education, School of Mathematical Sciences, Beijing Normal University, Beijing 100875, People's Republic of China (jiequan@bnu.edu.cn). This author was partially supported by NSFC (10971142,11031001), the Key Program from Beijing Educational Commission (KZ200910028002), and PHR (IHLB).

‡LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China (yangzhicheng@foxmail.com).

[1]The phenomenon of local oscillations does not contradict the TVD property because the latter is a global definition while the oscillations are of a local nature. The preservation of the TVD property is due to the compensation by strong decrease in solution amplitude.

with the discrete Fourier decomposition of the initial data and concluded that *local oscillations are caused as long as high frequency modes are present and no sufficient dissipation is provided to suppress them.* The approaches adopted there are both the discrete Fourier analysis and the modified equation technique [14], but just for three-point generalized Lax–Friedrichs (GLF) schemes applied basically to linear advection equations. The so-called GLF schemes are of three-point with uniformly constant coefficients of numerical viscosity. Although nonlinear cases were also touched upon in [9], the technique there is difficult to extend for general schemes (the assumption of the constant coefficient of numerical viscosity is very crucial in [9]).

In this paper we will develop a *heuristic modified equation approach*, rather than the algebraic manipulation [1, 2], to understand the oscillatory phenomenon resolved by quite generic finite difference or finite volume schemes for nonlinear conservation laws. The word *heuristic* means that the modified equation is only derived formally but is quite intuitive for further analysis of nonlinear problems. The generalization to general partial differential equations is straightforward, even with several spatial variables. The terminology *numerical damping*, distinct from the traditional numerical viscosity, is introduced to quantify the effect of dissipating the oscillatory modes.

We start with conservation laws

$$u_t + f(u)_x = 0, \tag{1.1}$$

subject to the initial data

$$u(x,0) = u_0(x) \in L_\infty(\mathbb{R}), \tag{1.2}$$

where $u = u(x,t)$ is the function of spatial variable x and the temporal variable t, and $f = f(u)$ is a smooth (say C^∞ for simplicity in presentation) convex or nonconvex flux function. We denote by τ and h respective temporal and spatial sizes of computational meshes, and denote by $\lambda = \tau/h$ their ratio. Furthermore we denote $x_j = jh$, $t_n = n\tau$, $x_{j+\frac{1}{2}} = \frac{1}{2}(x_j + x_{j+1})$, $j \in \mathbb{Z}$, and $n \in \mathbb{N}$. Then the schemes under consideration are of $(p+q+1)$-point in conservation form

$$u_j^{n+1} = H(u_{j-q}^n, \dots, u_{j+p}^n) := u_j^n - \lambda \left[F_{j+1/2}^n - F_{j-1/2}^n \right], \tag{1.3}$$

where u_j^n approximates the value of u at (x_j, t_n) or the average over the cell $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ at time $t = t_n$,

$$u_j^n = \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t_n)dx, \tag{1.4}$$

and $F_{j+1/2}^n = F(u_{j-q+1}^n, \dots, u_{j+p}^n)$ is the numerical flux, consistent with (1.1) in the sense that

$$F(u, \dots, u) = f(u). \tag{1.5}$$

In [5] the modified equation was derived for (1.3),

$$u_t + f(u)_x = h[\beta(u,\lambda)u_x]_x + \mathcal{O}(h^2), \tag{1.6}$$

where the viscosity coefficient $\beta(u,\lambda)$ is

$$\beta(u,\lambda) = \frac{1}{2\lambda} \sum_\ell \ell^2 H_\ell(u, \dots, u) - \frac{1}{2}(f'(u))^2,$$
$$H_\ell = \tfrac{\partial H}{\partial u_{j+\ell}^n}(u_{j-q}^n, \dots, u_{j+\ell}^n, \dots, u_{j+p}^n), \qquad -q \le \ell \le p.$$

For monotone schemes [3], the viscosity coefficient $\beta(u,\lambda) > 0$ plays the role of dissipation effect, i.e., the *artificial numerical viscosity* that is traced back to von Neumann and Richtmyer [12]. The philosophy of the modified equation plays a very fundamental part in understanding the mechanism of dissipation and dispersion of numerical schemes. However, it is worthwhile to point out that (1.6) is only valid implicitly in smooth regions of solutions, although its validity was verified for shock waves with monotone profiles in a later study [4]. In the language of Fourier analysis, (1.6) is only true for low frequency modes. It was shown in [9] that the presence of high frequency (oscillatory) modes is unavoidable in practice, e.g., resulting from the discretization of a square signal or numerical boundary treatment by way of the standard discrete Fourier decomposition. Furthermore, the oscillatory modes were demonstrated to cause local oscillations in numerical solutions obtained even by monotone schemes, particularly the Lax–Friedrichs scheme that is endowed with very strong numerical viscosity, as observed in [1, 2, 13]. This implies that the classical numerical viscosity, such as $\beta(u,\lambda)$ in (1.6), does not seem to be enough to offset the oscillations caused by oscillatory modes, even for linear advection equations. Hence the dissipation effect needs to be quantified deeply, and it is necessary to revisit the modified equation approach in order to understand the phenomenon of oscillations for nonlinear problems.

Let us consider the MacCormack scheme for (1.1) as an example [10]:

$$\begin{aligned} u_j^* &= u_j^n - \lambda\big[f(u_{j+1}^n) - f(u_j^n)\big], \\ u_j^{n+1} &= \frac{1}{2}\left[u_j^n + u_j^* - \lambda\big(f(u_j^*) - f(u_{j-1}^*)\big)\right]. \end{aligned} \tag{1.7}$$

The initial data is taken to be the highest frequency mode

$$u_j^0 = (-1)^j. \tag{1.8}$$

From [9] we know that the mode (1.8) is often present in the discretized data. Assume that the solution to (1.7) at $t = t_n$ is of the checkerboard mode type $u_j^n = (-1)^{j+n}b$, $b \neq 0$ if the flux function $f(u)$ is odd, $f(-u) = -f(u)$. Then it is easy to check that the solution at $t = t_{n+1}$ is

$$u_j^{n+1} = (-1)^{j+n}b\left[1 - \lambda\cdot\frac{f(b+2\lambda f(b)) - f(b)}{b}\right] = (-1)^{j+n+1}b\left[2\lambda^2\frac{f(b)}{b}f'(\xi) - 1\right],$$

where $\xi \in (b, b+2\lambda f(b))$. The amplitude of u_j^n is significantly amplified if $|2\lambda^2 \cdot f(b)\cdot f'(\xi)/b - 1| > 1$, which is possible for many numerical fluxes even though the usual CFL condition, $\lambda\max_{\xi\in\mathbb{R}}|f'(\xi)| \le 1$, is satisfied. Note also that for the linear flux $f(u) = au$ for some constant a, the MacCormack scheme is identical to the Lax–Wendroff scheme [7]. Hence, once such kinds of modes of (1.8) are present, the numerical solutions behave at least locally oscillatory [9].

As already done in [9], the modified equation approach can be employed to investigate this oscillatory phenomenon or (nonlinear) numerical instability, which is the central goal of the present study. We recall and emphasize in passing that [9] worked basically for GLF schemes applied to linear advection equations. Our attention here is mainly on nonlinear conservation laws (1.1). Thus we consider oscillatory parts of the solutions of (1.3) and put the ansatz, as proposed in [11],

$$u_j^n = (-1)^{j+n}\tilde{u}_j^n, \tag{1.9}$$

where $\tilde{u}_j^n$ is introduced so as to dodge the complexity caused by the oscillatory behavior of u_j^n. A detailed interpretation will be given in section 2. The amplitude of

$\tilde{u}_j^n$ is the same as that of u_j^n and so it is equivalent to discuss either of them from the stability point of view. It turns out that the modified equation for $\tilde{u}$ is expressed in the form

$$\boxed{\mathcal{D}_t\tilde{u} + G(\tilde{u})_x = -2\frac{r(\tilde{u})}{\tau}\tilde{u} + \tau\cdot(\eta(\tilde{u},\lambda)\tilde{u}_x)_x + \tau\cdot\epsilon(\tilde{u},\lambda)\tilde{u}_x^2 + \mathcal{O}(\tau^2),} \tag{1.10}$$

where $\mathcal{D}_t$ is the forward difference operator in time $\mathcal{D}_t\tilde{u} = (\tilde{u}(t+\tau,\cdot) - \tilde{u}(t,\cdot))/\tau$, and G, r, η, and ϵ are all functions of $\tilde{u}$. Equation (1.10) is not a purely rigorous PDE in the form of the modified equation in [14]. However, we will show that it is very heuristic to indicate the stability of (1.3) of the oscillatory parts of solutions, as shown in the following observation. This is achieved through the term $-\frac{2r(\tilde{u})}{\tau}\tilde{u}$, which is named a *damping term* in the present paper.

PROPOSITION 1.1. *The heuristic modified equation* (1.10) *implies the criterion of the stability of the oscillatory solution of* (1.3):

(i) **Damping.** *As* $0 < r(\tilde{u}) < 1$*, the term* $-2\frac{r(\tilde{u})}{\tau}\tilde{u}$ *plays the role of numerical damping that dissipates the oscillatory modes and suppresses oscillations eventually.*

(ii) **Neutrality.** *As* $r(\tilde{u}) = 0$ *or* $r(\tilde{u}) = 1$*, the damping effect vanishes and the dissipation effect of* (1.3) *depends on the numerical viscosity (the term of second order derivative). For oscillatory modes, no dissipation effect exists and the oscillations persist.*

(iii) **Amplification.** *As* $r(\tilde{u}) < 0$ *or* $r(\tilde{u}) > 1$*, the amplitude of the oscillatory modes is amplified.*

The damping factor $r(\tilde{u})$ can be expressed explicitly for the scheme (1.3) as

$$\boxed{\begin{aligned} r(\tilde{u}) := 1 + (-1)^{n+j+1}\lambda\big[&F((-1)^{n+j-q+1}\tilde{u},\dots,(-1)^{n+j+p}\tilde{u}) \\ &-F((-1)^{n+j-q}\tilde{u},\dots,(-1)^{n+j+p-1}\tilde{u})\big]/(2\tilde{u}). \end{aligned}} \tag{1.11}$$

We show that for generic monotone schemes $0 \le r(\tilde{u}) \le 1$, and therefore oscillations cannot be amplified so that at most local oscillations are observed. For high order schemes such as the MacCormack scheme and the Zwas–Abarbanel scheme, $r(\tilde{u})$ may escape from the range $[0,1]$, and the local oscillations will be amplified significantly. Hence $r(\tilde{u})$ can be regarded as an indicator of stability for oscillatory modes when resolved by the scheme (1.3).

The plan of this paper is as follows. In section 2 we derive, though nonrigorously, the heuristic modified equation (1.10) and obtain the explicit factor $r(\tilde{u})$ in (1.11). In section 3 we show that $0 \le r(\tilde{u}) \le 1$ for monotone schemes by which oscillatory modes are therefore not amplified. Then we proceed to specify the heuristic modified equation for several important specific schemes, the generalized Lax–Friedrichs schemes, the Richtmyer scheme, the MacCormack scheme, and the Zwas–Abarbanel third order scheme in section 4. Proposition 1.1 is verified numerically for nonlinear conservation laws with convex and nonconvex fluxes, respectively, in section 5. Finally a general discussion is given in section 6.

2. Derivation of heuristic modified equation for oscillatory solutions. This section is dedicated to the derivation of the heuristic modified equation (1.10). As done in [11] the solution u_j^n can be formally decomposed into a smooth part $u_j^{n,s}$ and an oscillatory part $u_j^{n,o}$,

$$u_j^n = u_j^{n,s} + u_j^{n,o}. \tag{2.1}$$

The superscript "s" refers to "smooth" and "o" to "oscillatory." This setting is undoubtedly valid for linear problems. For nonlinear problems, we are not able to manipulate them in such a way because the superposition principle does not hold. In order to focus on the influence of the oscillatory part $u_j^{n,o}$, the solution is assumed to comprise only this part. In doing so, we set

$$u_j^n = u_j^{n,o} = (-1)^{j+n}\tilde{u}_j^n. \tag{2.2}$$

The introduction of $\tilde{u}_j^n$ can be understood in terms of Fourier analysis language heuristically as follows. Regard the oscillatory part as high frequency modes around $(-1)^{j+n}$, i.e.,

$$u_j^{n,o} = e^{i(n\omega\tau + jkh)}, \qquad i^2 = -1, \tag{2.3}$$

where both $\omega\tau$ and kh are close to π for high frequency modes. Introduce small quantities $\tilde{\omega} = o(1)$ and $\tilde{k} = o(1)$ so that $\omega\tau = \pi + \tilde{\omega}\tau$, $kh = \pi + \tilde{k}h$. Then we have

$$u_j^{n,o} = (-1)^{j+n} e^{i(n\tilde{\omega}\tau + j\tilde{k}h)}. \tag{2.4}$$

In comparing (2.2) and (2.4), it is clear that $\tilde{u}_j^n$ in (2.2) corresponds to the mode $e^{i(n\tilde{\omega}\tau + j\tilde{k}h)}$ and can be regarded as the perturbation of the checkerboard mode $(-1)^{j+n}$. Due to the smallness of $\tilde{\omega}$ and $\tilde{k}$, $e^{i(n\tilde{\omega}\tau + j\tilde{k}h)}$ resembles the usual low frequency modes. This motivates us to assume the smoothness of $\tilde{u}$ and manipulate it, as done in [5] for (1.6), to derive the modified equation that $\tilde{u}$ satisfies. Note that the amplitude of $\tilde{u}$ is the same as that of u itself. Hence it is equivalent to investigate either of them from the L_∞ stability point of view. In what follows, we refer to $u_j^{n,o}$ as *oscillatory modes* and to $\tilde{u}_j^n$ as a *perturbation.* Our attention will be just on $\tilde{u}_j^n$.

PROPOSITION 2.1. *The heuristic modified equation of* (1.3) *in terms of the perturbation* $\tilde{u}$ *is written as* (1.10).

Proof. For simplicity in presentation, we concentrate on the oscillatory modes $u_j^{n,o}$ and use u_j^n to replace $u_j^{n,o}$ in the proof as expressed in (2.2). We insert (2.2) into (1.3) to obtain

$$\begin{aligned} \tilde{u}_j^{n+1} &= -\tilde{u}_j^n - (-1)^{j+n+1}\lambda\left[\tilde{F}_{j+1/2}^{*,n} - \tilde{F}_{j-1/2}^{*,n}\right] \\ &= -\tilde{u}_j^n - \lambda\left[(-1)^{j+n+1}\tilde{F}_{j+1/2}^{*,n} - (-1)^{j+n-1}\tilde{F}_{j-1/2}^{*,n}\right], \end{aligned} \tag{2.5}$$

where $\tilde{F}_{j+1/2}^{*,n} = F((-1)^{j-q+1}\ \tilde{u}_{j-q+1}^n, \ldots, (-1)^{j+p}\tilde{u}_{j+p}^n)$. We proceed to rewrite (2.5) as

$$\begin{aligned} \tilde{u}_j^{n+1} = \ & \tilde{u}_j^n - \lambda\left[\tilde{F}_{j+1/2}^n - \tilde{F}_{j-1/2}^n\right] - 2\tilde{u}_j^n \\ & -\lambda\left[\left((-1)^{j+n+1}\tilde{F}_{j+1/2}^{*,n} - \tilde{F}_{j+1/2}^n\right) - \left((-1)^{j+n-1}\tilde{F}_{j-1/2}^{*,n} - \tilde{F}_{j-1/2}^n\right)\right], \end{aligned} \tag{2.6}$$

where $\tilde{F}_{j+1/2}^n$ is the numerical flux function in terms of $\tilde{u}_j^n$, i.e., $\tilde{F}_{j+1/2}^n = F(\tilde{u}_{j-q+1}, \ldots, \tilde{u}_{j+p})$. We observe that the first row of (2.6) has exactly the same form as that used to derive (1.6) in [5], and the second row is more or less a source term to dampen or amplify the oscillatory part. That is, we have

$$\tilde{F}_{j+1/2}^n - \tilde{F}_{j-1/2}^n = hf_x(\tilde{u}_j^n) - \frac{h^2}{2}\left(\sum_\ell \ell^2 \tilde{H}_\ell \tilde{u}_x\right)_x + \mathcal{O}(h^3) \tag{2.7}$$

and rewrite (2.6) as

$$\mathcal{D}_t\tilde{u}_j^n + f_x(\tilde{u}_j^n) = \frac{h^2}{2\tau}\left(\sum_\ell \ell^2 \tilde{H}_\ell \tilde{u}_x\right)_x - \frac{1}{\tau}\tilde{H}^* + \mathcal{O}(h^2), \tag{2.8}$$

where the notations $\mathcal{D}_t$, $\tilde{H}$, $\tilde{H}^*$ are

$$\begin{aligned}
\mathcal{D}_t\tilde{u}(x,t) &:= \frac{\tilde{u}(x,t+\tau)-\tilde{u}(x,t)}{\tau}, \quad \tilde{H} = \tilde{u}_j^n - \lambda\left[\tilde{F}_{j+1/2}^n - \tilde{F}_{j-1/2}^n\right], \\
\tilde{H}^*(\tilde{u}_{j-q}^n,\dots,\tilde{u}_{j+p}^n) &:= 2\tilde{u}_j^n + \lambda\Big[\left((-1)^{j+n+1}\tilde{F}_{j+1/2}^{*,n} - \tilde{F}_{j+1/2}^n\right) \\
&\qquad - \left((-1)^{j+n-1}\tilde{F}_{j-1/2}^{*,n} - \tilde{F}_{j-1/2}^n\right)\Big],
\end{aligned} \tag{2.9}$$

and $\tilde{H}_\ell$ stands for the value of the ℓth partial derivative of $\tilde{H}$, i.e., $\tilde{H}_\ell = \frac{\partial \tilde{H}}{\partial v_\ell}(v_{-q},\dots,v_p)$ at $(\tilde{u}_j^n,\dots,\tilde{u}_j^n)$, $-q \le \ell \le p$, and similarly for $\tilde{H}_\ell^*$ and $\tilde{H}_{\ell,m}^*$ below.

It is the presence of $\tilde{H}^*$ in (2.8) that makes the investigation highly involved and amplifies or dampens oscillatory modes. First we claim that $\tilde{H}^*$ can be expressed in the form

$$\begin{aligned}
\tilde{H}^*(\tilde{u}(x-qh,n\tau),\dots,\tilde{u}(x+ph,n\tau)) &= 2r(\tilde{u})\tilde{u} + h\sum_{\ell=-q}^{p}\ell\tilde{H}_\ell^*\tilde{u}_x + \frac{h^2}{2}\left(\sum_{\ell=-q}^{p}\ell^2\tilde{H}_\ell^*\tilde{u}_x\right)_x \\
&\quad - \frac{h^2}{4}\sum_{\ell,m=-q}^{p}(\ell-m)^2\tilde{H}_{\ell,m}^*(\tilde{u}_x)^2 + \mathcal{O}(h^3).
\end{aligned} \tag{2.10}$$

Indeed, we expand $\tilde{H}^*$ at $\tilde{u} := \tilde{u}_j^n$ to yield

$$\begin{aligned}
\tilde{H}^*(\tilde{u}_{j-q}^n,\dots,\tilde{u}_{j+p}^n) &= \tilde{H}^*(\tilde{u},\dots,\tilde{u}) + \sum_{\ell=-q}^{p}\tilde{H}_\ell^*(\tilde{u},\dots,\tilde{u})(\tilde{u}_{j+\ell}^n - \tilde{u}) \\
&\quad + \frac{1}{2}\sum_{\ell,m}H_{\ell,m}^*(\tilde{u},\dots,\tilde{u})(\tilde{u}_{j+\ell}^n-\tilde{u})(\tilde{u}_{j+m}^n-\tilde{u}) + \mathcal{O}(h^3) \\
&=: I + II + III + \mathcal{O}(h^3).
\end{aligned} \tag{2.11}$$

We remind the reader that the assumption $\tilde{u}_{j+\ell} - \tilde{u} = \mathcal{O}(h)$ has been used here. We now proceed by estimating I, II, and III, respectively.

First, we have for I,

$$\begin{aligned}
I = 2\tilde{u} - (-1)^{n+j}\lambda\big[&F((-1)^{n+j-q+1}\tilde{u},\dots,(-1)^{n+j+p}\tilde{u}) \\
&-F((-1)^{n+j-q}\tilde{u},\dots,(-1)^{n+j+p-1}\tilde{u})\big].
\end{aligned} \tag{2.12}$$

We define $r(\tilde{u})$ to satisfy $2\tilde{u}\cdot r(\tilde{u}) = I$ so that

$$\begin{aligned}
r(\tilde{u}) := 1 - (-1)^{n+j}\lambda\big[&F((-1)^{n+j-q+1}\tilde{u},\dots,(-1)^{n+j+p}\tilde{u}) \\
&-F((-1)^{n+j-q}\tilde{u},\dots,(-1)^{n+j+p-1}\tilde{u})\big]/(2\tilde{u}).
\end{aligned} \tag{2.13}$$

This is obviously a nontrivial term that we will clarify later. For II and III, we have

(2.14)
$$\begin{aligned} II+III &= \sum_{\ell=-q}^{p} \tilde{H}_\ell^*(\tilde{u},\tilde{u},\dots,\tilde{u})\left(\ell h\tilde{u}_x+\frac{1}{2}\ell^2h^2\tilde{u}_{xx}\right)\\ &\quad+\frac{1}{2}\sum_{\ell,m=-q}^{p}\tilde{H}_{\ell,m}^*(h^2\ell m(\tilde{u}_x)^2)+\mathcal{O}(h^3)\\ &= h\tilde{u}_x\sum_{\ell=-q}^{p}\ell\tilde{H}_\ell^*+\frac{h^2}{2}\tilde{u}_{xx}\sum_{\ell=-q}^{p}\ell^2\tilde{H}_\ell^*+\frac{h^2}{2}(\tilde{u}_x)^2\sum_{\ell,m=-q}^{p}\ell m\tilde{H}_{\ell,m}^*+\mathcal{O}(h^3)\\ &= h\tilde{u}_x\sum_{\ell=-q}^{p}\ell\tilde{H}_\ell^*+\frac{h^2}{2}\left(\sum_{\ell=-q}^{p}\ell^2\tilde{H}_\ell^*\tilde{u}_x\right)_x\\ &\quad-\frac{h^2(\tilde{u}_x)^2}{4}\sum_{\ell,m=-q}^{p}(\ell-m)^2\tilde{H}_{\ell,m}^*+\mathcal{O}(h^3). \end{aligned}$$

This completes the proof of claim (2.10).

Combining (2.8) and (2.11)–(2.14), we obtain

(2.15)
$$\begin{aligned} \mathcal{D}_t\tilde{u}+\left[f'(\tilde{u})+\lambda^{-1}\sum_\ell\ell\tilde{H}_\ell^*\right]\tilde{u}_x &= -2\frac{r(\tilde{u})}{\tau}\tilde{u}+\frac{\tau\lambda^{-2}}{2}\left(\sum_\ell\ell^2(\tilde{H}_\ell-\tilde{H}_\ell^*)\tilde{u}_x\right)_x\\ &\quad+\frac{\tau}{4}\lambda^{-2}\sum_{\ell,m}(\ell-m)^2\tilde{H}_{\ell,m}^*(\tilde{u}_x)^2+\mathcal{O}(h^2). \end{aligned}$$

We set G, η, and ϵ in (1.10) as

$$\begin{aligned} G(\tilde{u}) &= f'(\tilde{u})+\lambda^{-1}\sum_\ell\ell\tilde{H}_\ell^*,\\ \eta(\tilde{u},\lambda) &= \frac{\lambda^{-2}}{2}\sum_\ell\ell^2(\tilde{H}_\ell-\tilde{H}_\ell^*),\\ \epsilon(\tilde{u},\lambda) &= \frac{\lambda^{-2}}{4}\sum_{\ell,m}(\ell-m)^2\tilde{H}_{\ell,m}^*. \end{aligned} \tag{2.16}$$

Then we obtain the heuristic modified equation, as expressed in (1.10). ∎

Remark 2.2. Due to the high nonlinearity of the underlying problem, this modified equation cannot be written in a pure PDE form like (1.6) for smooth solutions. However, in the case that f is linear and (1.3) is a linear scheme,

$$u_j^{n+1}=H(u_{j-q}^n,\dots,u_{j+p}^n)=\sum_{\ell=-q}^{p}A_\ell u_{j+\ell}^n,\quad \sum_{\ell=-q}^{p}A_\ell=1,\quad A_\ell \text{ are constant}, \tag{2.17}$$

the quadratic term $\epsilon(\tilde{u},\lambda)$ in (1.10) vanishes, and $r(\tilde{u})=\frac{1}{2}+\frac{1}{2}\sum_{\ell=-q}^{p}A_\ell(-1)^\ell$ is a constant. Then (1.10) is written, at the highest frequency mode $kh=\pi$, as

$$\partial_t\tilde{u}+G(\pi/h)\cdot\tilde{u}_x=\frac{1}{\tau}\ln|\zeta(\pi/h)|\tilde{u}+\frac{i}{2}G'(\pi/h)\tilde{u}_{xx}+\mathcal{O}(h^2), \tag{2.18}$$

where $G(k)$ is the group velocity at the highest frequency mode $kh=\pi$, k is the wave number, $\zeta(k)=\sum_\ell A_\ell e^{i\ell kh}$ is the symbol of the scheme, and $G(k)$ is related with $\zeta(k)$

by the formula $G(k) = \frac{i}{\tau} \cdot \frac{\zeta'(k)}{\zeta(k)}$. Also ζ and r are related with

$$\zeta(\pi/h) = 2r(1) - 1. \tag{2.19}$$

We point out that (2.18) is a standard and closed (not only heuristic) modified equation of the linear scheme (2.17) depicting the propagation of oscillatory modes, and covers the result in [9] that holds just for the GLF schemes.

Remark 2.3. Compared to the modified equation (1.6), the conspicuous term $\sum_{\ell,m=-q}^{p}(\ell-m)^2\tilde{H}^*_{\ell,m}$ does not vanish, in general, unless H is linear as in (2.17). We take three-point schemes as examples ($p = q = 1$). Then $\tilde{H}^*$ has the form (setting $j = 0$)

$$\tilde{H}^*(\tilde{u},\tilde{u},\tilde{u}) = 2\tilde{u} + \lambda\,[F(\tilde{u},-\tilde{u}) - F(-\tilde{u},\tilde{u})] \tag{2.20}$$

if n is odd, or

$$\tilde{H}^*(\tilde{u},\tilde{u},\tilde{u}) = 2\tilde{u} + \lambda\,[F(-\tilde{u},\tilde{u}) - F(\tilde{u},-\tilde{u})] \tag{2.21}$$

if n is even (recall (2.9) for the notation $\tilde{H}^*$). It is easy to check that $\sum_{\ell,m=-1}^{1}(\ell - m)^2\tilde{H}^*_{\ell,m} \neq 0$ for general numerical flux functions F.

The distinct feature of (1.10) or (2.18) from the classical modified equation (1.6) is the presence of the damping term (zero order term) $r(\tilde{u})$, which dampens oscillations or causes instability. Indeed, from the theory for the linear equation

$$v_t = \alpha v + \beta v_{xx}, \tag{2.22}$$

where α is a constant, the solution $v(t,x)$ subject to the initial data $v_0(x) = e^{i\tilde{k}x}$ is

$$v(t,x) = e^{(\alpha-\beta\tilde{k}^2)t}e^{i\tilde{k}x}. \tag{2.23}$$

Hence $v(t,x)$ exponentially decays to zero as the time increases if $\alpha < 0$ and $\tilde{k}$ is close to 0, but it is amplified dramatically to induce the instability if $\alpha > 0$. In general, the linear stability (von Neumann) condition requires

$$|\zeta(k)| \le 1 \tag{2.24}$$

for all k (when there is no source in the associated governing equation). This implies the damping effect of the source term in (2.18), $\ln|\zeta| \le 0$, which is equivalent to saying, in terms of r,

$$0 \le r(\tilde{u}) \le 1. \tag{2.25}$$

Nevertheless, for the classical Lax–Friedrichs scheme the coefficient of numerical viscosity is uniformly one, which implies $|\zeta| = 1$. Hence the damping term vanishes for high frequency modes ($\ln|\zeta| = 0$), which explains why local oscillations are observed, as justified in [9] (an alternative explanation can be found in [1]). So the usual numerical viscosity (second order term) is not sufficient to suppress the oscillations caused by oscillatory modes.

For nonlinear cases, (1.10) resembles a PDE semidiscretized in time. We simply split it to take a look at the first part (the more important part in some sense),

$$\mathcal{D}_t\tilde{u} = -2\frac{r(\tilde{u})}{\tau}\tilde{u}. \tag{2.26}$$

Obviously, whether the amplitude of $\tilde{u}(t,\cdot)$ increases or not depends on $r(\tilde{u})$:

1. $|\tilde{u}(t+\tau,\cdot)| < |\tilde{u}(t,\cdot)|$ if $0 < r(\tilde{u}) < 1$;
2. $|\tilde{u}(t+\tau,\cdot)| > |\tilde{u}(t,\cdot)|$ if $r(\tilde{u}) < 0$ or $r(\tilde{u}) > 1$;
3. $|\tilde{u}(t+\tau,\cdot)| = |\tilde{u}(t,\cdot)|$ if $r(\tilde{u}) = 0$ or $r(\tilde{u}) = 1$.

Based on this analysis, we formulate Proposition 1.1 as a criterion of the stability of oscillatory solutions. This criterion is obviously expository or heuristic. A rigorous treatment remains for the future work.

3. Dissipation property of monotone schemes on oscillatory modes. This section serves to provide the analysis of the evolution of oscillatory modes by monotone schemes [3],

$$u_j^{n+1} = H(u_{j-q}^n, \dots, u_{j+p}^n) = u_j^n - \lambda[F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n], \tag{3.1}$$

with the numerical flux $F_{j+\frac{1}{2}}^n = F(u_{j-q+1}^n, \dots, u_{j+p}^n)$. The monotonicity means that

$$\frac{\partial H}{\partial v_\ell}(v_{-q}, \dots, v_\ell, \dots, v_p) \geq 0 \tag{3.2}$$

for all $-q \leq \ell \leq p$. The generalized Lax–Friedrichs (GLF) schemes proposed in [9] belong to this class. However, the GLF schemes are just of three-point and the coefficients of numerical viscosity are assumed to be uniformly constant in [9]. Here we will discuss the general case of monotone schemes with $p+q+1$ stencils. Our conclusion is the following.

PROPOSITION 3.1. *If the schemes* (3.1) *are monotone in the sense of* (3.2)*, then there holds*

$$0 \leq r(\tilde{u}) \leq 1, \tag{3.3}$$

which implies that oscillatory modes are not amplified by the monotone schemes according to Proposition 1.1*.*

Proof. For convenience in presentation, we denote

$$\begin{aligned} &U_j^n = (u_{j-q+1}^n, \dots, u_j^n, \dots, u_{j+p}^n),\\ &\delta U_j^n = (u_{j-q+2}^n, \dots, u_{j+1}^n, \dots, u_{j+p+1}^n),\\ &\delta^{-1} U_j^n = U_{j-1}^n = (u_{j-q}^n, \dots, u_{j-1}^n, \dots, u_{j+p-1}^n), \end{aligned} \tag{3.4}$$

where δ is a translation operator. Then (3.1) is rewritten as

$$u_j^{n+1} = H(u_{j-q}^n, \dots, u_{j+p}^n) = u_j^n - \lambda\Big[F(U_j^n) - F(\delta^{-1}U_j^n)\Big]. \tag{3.5}$$

Next we focus on this $p+q+1$-point scheme. Without loss of generality we fix $j=0$ and denote

$$F_\ell(U_0^n) = \frac{\partial F(\dots, u_\ell^n, \dots)}{\partial u_\ell^n}, \quad \ell = -(q-1), \dots, p, \tag{3.6}$$

which is the partial derivative of F with respect to the ℓth variable. Then we know

$$F_\ell(\delta^{-1}U_0^n) = F_\ell(U_{-1}^n) = \frac{\partial F(U_{-1}^n)}{\partial u_{\ell-1}^n}. \tag{3.7}$$

In view of the monotonicity property (3.2), there holds

$$\frac{\partial H(u^n_{-q},\dots,u^n_p)}{\partial u^n_\ell} \ge 0, \quad \ell = -q,\dots,p \tag{3.8}$$

for all possible $(u^n_{-q},\dots,u^n_p) \in \mathbb{R}^{p+q+1}$. In other words, we have

$$H_\ell(u^n_{-q},\dots,u^n_p) = \begin{cases} 1-\lambda\Big[F_\ell(U^n_0) - F_{\ell+1}(\delta^{-1}U^n_0)\Big] \ge 0 & \text{for } \ell = 0, \\ -\lambda F_\ell(U^n_0) \ge 0 & \text{for } \ell = p, \\ \lambda F_{\ell+1}(\delta^{-1}U^n_0) \ge 0 & \text{for } \ell = -q, \\ -\lambda[F_\ell(U^n_0) - F_{\ell+1}(\delta^{-1}U^n_0)] \ge 0 & \text{for } \ell \ne -q, 0, p. \end{cases} \tag{3.9}$$

As far as oscillatory modes are concerned, we introduce a new notation $V^n_0(s) := (v^n_{-(q-1)},\dots,v^n_p)$ with v^n_ℓ,

$$v^n_\ell = \begin{cases} (-1)^{n+\ell}s & \text{for } \ell \bmod 2 \equiv 0, \\ (-1)^{n+\ell}\tilde{u} & \text{for } \ell \bmod 2 \equiv 1. \end{cases} \tag{3.10}$$

Actually, V^n_0 is a vector with the structure $(\dots,-\tilde{u},s,-\tilde{u},s,-\tilde{u},\dots)$ or $(\dots,\tilde{u},-s,\tilde{u},-s,\tilde{u},\dots)$.

In order to verify (3.3), we define

$$K(s) = 1-\lambda\Bigg[\sum_{\substack{\ell=-(q-1)\\ \ell \bmod 2\equiv 0}}^{p} F_\ell(V^n_0) - \sum_{\substack{\ell=-(q-1)\\ \ell \bmod 2\equiv 1}}^{p} F_\ell(\delta^{-1}V^n_0)\Bigg]. \tag{3.11}$$

Obviously by integrating $K(s)$ from $-\tilde{u}$ to $\tilde{u}$, there holds

$$2\tilde{u}\cdot r(\tilde{u}) = \int_{-\tilde{u}}^{\tilde{u}} K(s)ds. \tag{3.12}$$

Hence we need to prove equivalently

$$0 \le K(s) \le 1. \tag{3.13}$$

For the nonnegativity of $K(s)$, we use (3.9) to assert

$$\begin{aligned} K(s) = {} & 1-\lambda\Big[F_0\big(V^n_0(s)\big) - F_1\big(\delta^{-1}V^n_0(s)\big)\Big] \\ & -\lambda F_p\big(V^n_0(s)\big)\frac{(-1)^p+1}{2} + \lambda F_{-q+1}\big(\delta^{-1}V^n_0(s)\big)\frac{(-1)^q+1}{2} \\ & -\lambda \sum_{\substack{\ell=-q+2\\ \ell \bmod 2\equiv 0,\ell\ne 0}}^{p-1} \Big[F_\ell\big(V^n_0(s)\big) - F_{\ell+1}\big(\delta^{-1}V^n_0(s)\big)\Big] \\ \ge {} & 0. \end{aligned} \tag{3.14}$$

As for the upper bound of $K(s)$ in (3.13), we want to show equivalently

$$\sum_{\substack{\ell=-(q-1)\\ \ell \bmod 2\equiv 0}}^{p} F_\ell(V^n_0) - \sum_{\substack{\ell=-(q-1)\\ \ell \bmod 2\equiv 1}}^{p} F_\ell(\delta^{-1}V^n_0) \ge 0. \tag{3.15}$$

Indeed, we use the translation operator δ introduced in (3.4),

$$\delta^{-1}V_j^n = \delta V_j^n, \quad V_j^n = \delta^{-1}\delta V_j^n \tag{3.16}$$

for $j = 0$. Then we proceed to use this property and (3.9) to obtain

$$\begin{aligned}
&\sum_{\substack{\ell=-(q-1)\\ \ell \mod 2\equiv 0}}^{p} F_\ell(V_0^n) - \sum_{\substack{\ell=-(q-1)\\ \ell \mod 2\equiv 1}}^{p} F_\ell(\delta^{-1}V_0^n)\\
&= \sum_{\substack{\ell=-(q-1)\\ \ell \mod 2\equiv 0}}^{p} F_\ell(\delta^{-1}\delta V_0^n) - \sum_{\substack{\ell=-(q-1)\\ \ell \mod 2\equiv 1}}^{p} F_\ell(\delta V_0^n)\\
&= \sum_{\substack{\ell=-(q-2)\\ \ell \mod 2\equiv 0}}^{p} F_\ell(\delta^{-1}\delta V_0^n) + F_{-(q-1)}\big(\delta^{-1}\delta V_0^n\big)\frac{(-1)^{q-1}+1}{2}\\
&\quad - \sum_{\substack{\ell=-(q-1)\\ \ell \mod 2\equiv 1}}^{p-1} F_\ell(\delta V_0^n) - F_p\big(\delta V_0^n\big)\frac{(-1)^{p+1}+1}{2}\\
&\geq - \sum_{\substack{\ell=-(q-1)\\ \ell \mod 2\equiv 1}}^{p-1} \Big[F_\ell(\delta V_0^n) - F_{\ell+1}\big(\delta^{-1}\delta V_0^n\big)\Big] \geq 0.
\end{aligned}$$

Thus we complete the verification of (3.3). □

4. The heuristic modified equation for several specific schemes. In this section we will specify the heuristic modified equation (1.10) and apply Proposition 1.1 for several schemes; the GLF schemes that were discussed in [9], the second order Richtmyer scheme, and the MacCormack scheme, as well as the third order Zwas–Abarbanel scheme [15]. The reason that we choose these schemes rather than other high resolution schemes such as ENO is that the current heuristic modified equation analysis is only concerned with the evolutional process of solutions controlled by (1.1) or the scheme (1.3), while many other high resolution schemes often use the step of data reconstruction in which extra dissipation effect is added through certain monotonicity algorithms (e.g., limiter modification). The dissipation mechanisms of those monotonicity algorithms are quantitatively unclear to us from the analysis point of view. Of course, we know that third or higher order schemes without the disposal of data reconstruction are rarely used in practice.

In the following computation of $r(\tilde{u})$, we set $j = 0$ just for notational simplicity.

4.1. The GLF schemes. The GLF schemes have already been discussed thoroughly in order to comprehend why local oscillations are present in solutions: from local data extrema [1, 2], purely numerically [13], from both discrete Fourier analysis and modified equation analysis [9]. Here we want to revisit this class of schemes using the heuristic modified equation we develop in the present paper just for the sake of completeness.

We compute $r(\tilde{u})$ for this class of schemes. The numerical flux $F^n_{j+\frac{1}{2}}$ takes the form

$$F^n_{j+1/2} = F(u^n_j, u^n_{j+1}) = \frac{f(u^n_{j+1}) + f(u^n_j)}{2} - \frac{Q}{2\lambda}(u^n_{j+1} - u^n_j), \quad 0 < Q_{\min} \leq Q \leq 1. \tag{4.1}$$

From (1.11) it is easy to calculate with $q = p = 1$

$$
\begin{aligned}
r(\tilde{u}) &= 1 + (-1)^{n+1}\lambda \left[F((-1)^n\tilde{u}, (-1)^{n+1}\tilde{u}) - F((-1)^{n-1}\tilde{u}, (-1)^n\tilde{u})\right]/(2\tilde{u}) \\
&= 1 - Q.
\end{aligned} \tag{4.2}
$$

Hence we have $0 \le r(\tilde{u}) \le 1$ if $0 \le Q \le 1$. In view of Proposition 1.1, the GLF schemes do have the damping effect around oscillatory modes if $0 < Q_{\min} \le Q < 1$ for certain $Q_{\min}$. As $Q = 1$, it is the celebrated Lax–Friedrichs scheme and the numerical viscosity that attain the maximum. For this case, $r(\tilde{u}) = 0$ and therefore the Lax–Friedrichs scheme belongs to the class of neutrality case in Proposition 1.1. That is, the numerical damping effect vanishes and the oscillatory modes persist, which explains why local oscillations in the solution obtained by the Lax–Friedrichs scheme are observed [1, 2, 13] once the initial data contains the oscillatory modes. In addition, for this class of schemes (4.1), our conclusion here is consistent with that in [2]: *"schemes with a large viscosity coefficient are prone to oscillations at data extrema."*

4.2. The Richtmyer scheme. The Richtmyer scheme is also called the two-step Lax–Wendroff scheme. This scheme and the next MacCormack scheme are identical to the Lax–Wendroff scheme for linear advection equations with constant coefficients. The numerical flux of the Richtmyer scheme takes the form

$$
\begin{cases}
F^n_{j+1/2} = F(u^n_j, u^n_{j+1}) = f(u^{n+1/2}_{j+1/2}), \\
u^{n+1/2}_{j+1/2} = \dfrac{1}{2}(u^n_j + u^n_{j+1}) - \dfrac{\lambda}{2}\left(f(u^n_{j+1}) - f(u^n_j)\right).
\end{cases} \tag{4.3}
$$

Recall the definition of $r(\tilde{u})$ in (1.11). Then we have

$$
\begin{aligned}
r(\tilde{u}) &= 1 + (-1)^{n+1}\lambda \left[F((-1)^n\tilde{u}, (-1)^{n+1}\tilde{u}) - F((-1)^{n-1}\tilde{u}, (-1)^n\tilde{u})\right]/(2\tilde{u}) \\
&= 1 + (-1)^{n+1}\lambda \left\{ f\left(-\tfrac{\lambda}{2}(f((-1)^{n+1}\tilde{u}) - f((-1)^n\tilde{u}))\right)\right. \\
&\quad \left. - f\left(-\tfrac{\lambda}{2}(f((-1)^n\tilde{u}) - f((-1)^{n-1}\tilde{u}))\right)\right\}/(2\tilde{u}).
\end{aligned} \tag{4.4}
$$

We discuss according to the parity of $f(u)$.

- Odd flux case. As the flux function $f(u)$ is odd, $f(-u) = -f(u)$, (4.4) becomes

$$
r(\tilde{u}) = 1 - \lambda \frac{f(\lambda f(\tilde{u}))}{\tilde{u}}. \tag{4.5}
$$

- Even flux case. As f is even, $f(-u) = f(u)$, we find that $F(-\tilde{u}, \tilde{u}) = F(\tilde{u}, -\tilde{u}) = f(0)$. It turns out that

$$
r(\tilde{u}) = 1. \tag{4.6}
$$

- General cases. We decompose the flux function $f(u)$ as a sum of odd and even functions, i.e.,

$$
f(u) = f_{odd}(u) + f_{even}(u) := \frac{f(u) - f(-u)}{2} + \frac{f(u) + f(-u)}{2}. \tag{4.7}
$$

Then we obtain

$$r(\tilde{u}) = 1 - \lambda \frac{f_{odd}(\lambda f_{odd}(\tilde{u}))}{\tilde{u}}. \tag{4.8}$$

In view of Proposition 1.1, we have the following conclusion. (i) As f is linear ($f(u) = au$) the CFL condition $|\lambda a| \le 1$ implies $0 \le r < 1$ ($\lambda a \ne 0$). (ii) As $f(u)$ is even, we obtain a neutrality case. The oscillations resulting from oscillatory modes will persist. (iii) As $f(u)$ is a nonlinear odd function, the situation is much more involved: the damping factor $r(\tilde{u})$ may escape from the $[0, 1]$ range and nonlinear instability may be caused.

4.3. The MacCormack scheme. The numerical flux of the MacCormack scheme takes the form

$$\begin{cases} F^n_{j+1/2} = F(u^n_j, u^n_{j+1}) = \dfrac{f(u^n_{j+1}) + f(u^*_j)}{2}, \\ u^*_j = u^n_j - \lambda \left(f(u^n_{j+1}) - f(u^n_j) \right). \end{cases} \tag{4.9}$$

The damping factor $r(\tilde{u})$ is

(4.10)

$$\begin{aligned} r(\tilde{u}) &= 1 + (-1)^{n+1}\lambda \left[F((-1)^n\tilde{u}, (-1)^{n+1}\tilde{u}) - F((-1)^{n-1}\tilde{u}, (-1)^n\tilde{u}) \right] / (2\tilde{u}) \\ &= 1 + (-1)^{n+1} \frac{\lambda}{4\tilde{u}} \left\{ f((-1)^{n+1}\tilde{u}) - f((-1)^n\tilde{u}) \right. \\ &\quad + f\left[(-1)^n\tilde{u} - \lambda \left[f((-1)^{n+1}\tilde{u}) - f((-1)^n\tilde{u}) \right] \right] \\ &\quad \left. - f\left[(-1)^{n-1}\tilde{u} - \lambda \left[f((-1)^n\tilde{u}) - f((-1)^{n-1}\tilde{u}) \right] \right] \right\}. \end{aligned}$$

We also compute $r(\tilde{u})$ by three cases.

- Odd flux case. As $f(u)$ is odd, the factor $r(\tilde{u})$ is expressed as

$$r(\tilde{u}) = 1 + \frac{\lambda}{2} \left[f(\tilde{u}) - f(\tilde{u} + 2\lambda f(\tilde{u})) \right] / \tilde{u}. \tag{4.11}$$

- Even flux case. As $f(u)$ is even, the factor $r(\tilde{u})$ is uniformly one,

$$r(\tilde{u}) = 1. \tag{4.12}$$

- General cases. In general, the factor $r(\tilde{u})$ is

$$r(\tilde{u}) = 1 + \frac{\lambda}{2} \left[f_{odd}(\tilde{u}) - f_{odd}(\tilde{u} + 2\lambda f_{odd}(\tilde{u})) \right] / \tilde{u}. \tag{4.13}$$

Hence the MacCormack scheme behaves almost the same as the Richtmyer scheme.

4.4. The Zwas–Abarbanel third order scheme. The third order scheme adopted here is due to Zwas and Abarbanel [15],

$$\begin{aligned} u^{n+1}_j = u^n_j &- \lambda \left[\frac{1}{2}(f^n_{j+1} - f^n_{j-1}) - \frac{\gamma_1}{12}(f^n_{j+2} - 2f^n_{j+1} + 2f^n_{j-1} - f^n_{j-2}) \right] \\ &+ \frac{\lambda^2}{2} \left[a^n_{j+\frac{1}{2}}(f^n_{j+1} - f^n_j) - a^n_{j-\frac{1}{2}}(f^n_j - f^n_{j-1}) \right] \\ &+ \frac{\gamma_2\lambda^3}{12} \left[a^n_{j+1}(f^n_{j+2} - f^n_j) - 2a^n_j(f^n_{j+1} - f^n_{j-1}) + a^n_{j-1}(f^n_j - f^n_{j-2}) \right], \end{aligned} \tag{4.14}$$

where $f_j^n = f(u_j^n)$, $a_{j+\frac{1}{2}}^n = f'(u_{j+\frac{1}{2}}^n)$, $u_{j+\frac{1}{2}}^n = \frac{1}{2}(u_j^n + u_{j+1}^n)$, and $a_j^n = f'(u_j^n)$. As $\gamma_1 = \gamma_2 = 1$, it is third order accurate. And as $\gamma_1 = \gamma_2 = 0$, it is again identical to the (second order) Lax–Wendroff scheme. The numerical flux has the form

(4.15)
$$F_{j+1/2}^n = \frac{1}{2}(f_{j+1}^n + f_j^n) - \frac{1}{12}(f_{j+2}^n - f_{j+1}^n - f_j^n + f_{j-1}^n) - \frac{\lambda}{2}a_{j+1/2}(f_{j+1} - f_j^n) - \frac{\lambda^2}{12}\left[a_{j+1}^n(f_{j+2}^n - f_j^n) - a_j^n(f_{j+1}^n - f_{j-1}^n)\right].$$

Then as $u_j^n = (-1)^{n+j}\tilde{u}_j^n$, the numerical flux becomes,

(4.16)
$$F_{j+1/2}^n((-1)^{n+2}\tilde{u}, (-1)^{n+1}\tilde{u}, (-1)^n\tilde{u}, (-1)^{n-1}\tilde{u}) = \begin{cases} (-1)^n\lambda a(0)f(\tilde{u}), & f \text{ is odd,} \\ f(\tilde{u}), & f \text{ is even.} \end{cases}$$

So, just as in the previous subsections, we have the following cases.

- Odd flux case. As f is odd, the factor $r(\tilde{u})$ is

  (4.17) $$r(\tilde{u}) = 1 - \lambda^2 a(0)f(\tilde{u})/\tilde{u}.$$

- Even flux case. As f is even, the factor $r(\tilde{u})$ is

  (4.18) $$r(\tilde{u}) = 1.$$

- General cases. We have

  (4.19) $$r(\tilde{u}) = 1 - \lambda^2 a(0)f_{odd}(\tilde{u})/\tilde{u}.$$

Again, the behavior of this third order scheme is analogous to how the Richtmyer scheme and the MacCormack scheme behave. Hence we end this section by making the following statements.

1. As $f(u)$ is even, there is no damping effect on the oscillations caused by oscillatory modes so that some local oscillations persist and are observable once the data contain oscillatory modes. We also notice that $r(\tilde{u}) \equiv 0$ for the Lax–Friedrichs scheme while $r(\tilde{u}) \equiv 1$ for the Richtmyer, MacCormack, and Zwas–Abarbanel schemes although they all fall into the category of neutrality in Proposition 1.1. Starting from the same initial data $u_j^0 = (-1)^j$, the Lax–Friedrichs scheme produces a checkerboard solution
   $$u_j^n = (-1)^{j+n}.$$
   In contrast, the Richtmyer, MacCormack, and Zwas–Abarbanel schemes produce a solution
   $$u_j^n = (-1)^j,$$
   which is stagnant.
2. As $f(u)$ is odd, the factor $r(\tilde{u})$ may escape from the stability interval $(0, 1)$ unless $f(u)$ is linear, $f(u) = au$, for which the CFL condition can guarantee $r(\tilde{u}) \in (0, 1)$. We take

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

as a nonlinear example. Then $f'(u) = \frac{1}{(1+u^2)^{\frac{3}{2}}}$. As the solution subject to the initial data around the checkerboard mode $u_j^0 = (-1)^j$ is considered, the CFL constraint for the first step is

$$\lambda \max_j |f'(u_j^0)| = \lambda \max_j \left| \frac{1}{[1+(u_j^0)^2]^{\frac{3}{2}}} \right| \le 1. \tag{4.21}$$

Then, for instance, the ratio λ can be taken to be 2. However, for the Richtmyer scheme, $r(\tilde{u}) = 1 - \frac{\lambda^2}{\sqrt{1+\tilde{u}^2+\lambda^2\tilde{u}^2}}$. Obviously, as $\tilde{u}$ is close to 1, $r(\tilde{u}) < 0$ (the third order scheme has the same property). Hence the usual stability condition such as (4.21) does not guarantee stability. In contrast, for the MacCormack scheme, the factor $r(\tilde{u})$ is still in the damping region, i.e., $r(\tilde{u}) \in (0,1)$, as $\tilde{u}$ is close to 1. Figures 4 and 5 in the next section support our statement here.

3. In general, we can take into account the odd part of the flux function to understand the behavior of solutions once the solution is polluted with oscillatory or checkerboard-type modes.

5. Numerical verifications. In this section, we will numerically verify the observation implied by the heuristic modified equation we proposed previously. We choose three different fluxes, $f(u) = \frac{u^2}{2}$, $\frac{u}{\sqrt{1+u^2}}$ (resp., $\frac{u}{1+u^2}$), and $\frac{u^2}{u^2+b(1-u)^2}$, to stand for the even, odd, and the general flux functions, respectively. The first is the classical Burgers equation, the last is the Buckley–Leverett equation simulating of oil-reservoir [8], where b is a parameter. We take $b = 1$ in the present paper. The schemes used are the generalized Lax–Friedrichs schemes, the Richtmyer scheme, the MacCormack scheme, and the Zwas–Abarbanel third order scheme already presented in the last section. In all examples, the initial data are distributed over the interval $[0, 10]$, 100 grid points are used to designate the discrete data, and the periodic boundary condition is adopted just for simplicity. Some of the results are just displayed over the interval $[3, 7]$ in order to focus on the oscillatory phenomenon under consideration. The following CFL condition

$$\frac{\tau}{h} \max_{1\le j\le 100} |f'(u_j^n)| \le 0.8 \tag{5.1}$$

is used for each $n \ge 0$. We recall that the entropy solution of scalar conservation laws (1.1) satisfies the maximum principle

$$\min_{x\in[0,10]} u_0(x) \le u(x,t) \le \max_{x\in[0,10]} u_0(x). \tag{5.2}$$

Our examples are devised in accordance with Proposition 1.1: (i) damping; (ii) neutrality; and (iii) amplification. We will clarify how oscillatory modes are suppressed, preserved, and amplified through the heuristic modified equation (1.10). Before doing so, we quote from [9] the fact that the discrete initial data may contain the highest frequency mode via the discrete Fourier decomposition, which implies the necessity of the study of oscillatory modes.

We consider the initial data of a single square signal,

$$u_0(x) = \begin{cases} 1, & 0 < x_1 < x < x_2 < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

In the discretization of such an initial datum, we can use an odd or an even number of grid points to designate the interval $[x_1, x_2]$, respectively. It was shown in [9] that if an odd number of grid points is used, then the discrete data contain the notorious highest frequency mode $(-1)^j$,

$$u_j^0 = (-1)^j h + \sum_{kh \neq \pi} A_k e^{jkh}; \tag{5.4}$$

otherwise, if an even number of grid points is used, no highest frequency mode is present in the discrete data

$$u_j^0 = 0 \times (-1)^j h + \sum_{kh \neq \pi} A_k e^{jkh}, \tag{5.5}$$

where A_k are the standard constant coefficients of discrete Fourier series; the reader can be referred to [9] for details.

The set of single square signals can serve as a basis to form arbitrary (e.g., L^∞) initial data. Hence the above analysis shows that the discrete data may contain the highest frequency mode in practice. In the following examples, we use two and three grid points to express the datum "1" of the single square single, respectively,

$$D_{even}: \quad u_j^0 = \begin{cases} 1, & j = 50, 51, \\ 0 & \text{otherwise}, \end{cases} \tag{5.6}$$

and

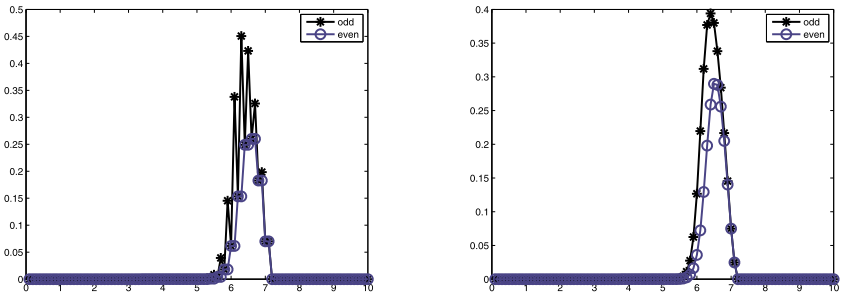
$$D_{odd}: \quad u_j^0 = \begin{cases} 1, & j = 49, 50, 51, \\ 0 & \text{otherwise}. \end{cases} \tag{5.7}$$

Therefore, in view of (5.4) and (5.5), the data D_{odd} contain the highest frequency mode but the data D_{even} do not.

5.1. Oscillations in numerical solutions. In Figure 1 we use the GLF schemes (4.1) to obtain the solution of (1.1) with the flux function $f(u) = \frac{u}{\sqrt{1+u^2}}$, subject to the single square signal initial data. In Figure 1 (a) the classical Lax–Friedrichs scheme ($Q = 1.0$) produces a local oscillatory solution for the data D_{odd} while in Figure 1 (b) we observe that the oscillations are suppressed when the generalized Lax–Friedrichs scheme with $Q = 0.9$ is used.

The oscillations in Figure 1(a), although observed and analyzed in [1, 2, 13], are counterintuitive because the Lax–Friedrichs scheme is endowed with the largest numerical viscosity. However, such numerical viscosity is in allusion to smooth flows (low frequency modes). Hence the classical modified approach does not work here. We will make our analysis in line with Proposition 1.1 in the subsequent subsections.

5.2. The effect of numerical damping. Once the highest frequency mode is present, the damping mechanism indeed needs to be introduced in order to suppress the resulting oscillations. Let us again look at the GLF schemes. When the highest frequency mode $u_j^0 = (-1)^j$ is considered purely as the initial data, the solution is of the checkerboard type $u_j^n = (1 - 2Q)^n (-1)^j$. Hence this mode is damped with the



(a) The L–F scheme with $Q = 1.0$ (b) The generalized L–F scheme with $Q = 0.9$

FIG. 1. *The propagation of a single square signal.*

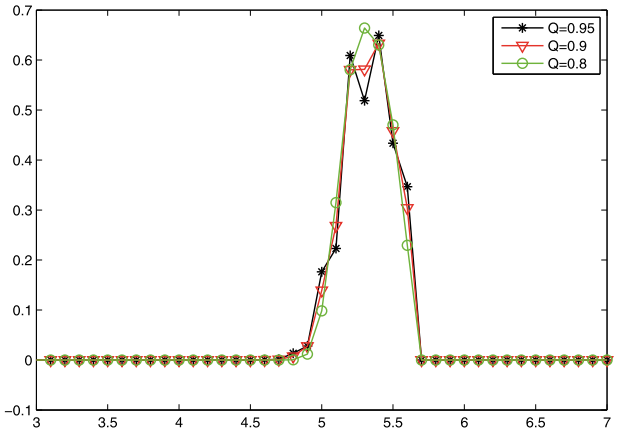


FIG. 2. *The damping effect of the GLF schemes on oscillatory modes.*

rate $|1 - 2Q| < 1$ if $0 < Q_{min} \le Q < 1$. More generally, when we return to the single square signal initial data, the solution is also damped. We observe from Figure 2 clearly how the highest frequency mode affects the solution and how it is dampened.

In terms of Proposition 1.1, $r(\tilde{u}) \equiv 1 - Q$ and $0 < r < 1$ if $0 < Q < 1$. Hence a damping effect is exerted on the highest frequency mode so that oscillatory modes are suppressed gradually.

5.3. Neutrality. Neutrality means that the oscillatory modes are neither dampened nor amplified, and that the factor $r(\tilde{u}) = 0$ or 1 according to Proposition 1.1. The classical aforementioned Lax–Friedrichs scheme is such an example, $r(\tilde{u}) = 0$. Also as the flux function $f(u)$ is even, the high order schemes we discussed in the last section exhibit neutrality $r(\tilde{u}) = 1$. Starting with the highest frequency mode $(-1)^j$, the solution obtained by the Richtmyer scheme, MacCormack scheme, or Zwas–Abarbanel scheme keeps invariant $u_j^n = (-1)^j$, which is different from the checkerboard mode $(-1)^{j+n}$ obtained by the Lax–Friedrichs scheme. It is evident that the amplitudes remain fixed although the propagation speeds are substantially different. The example

in Figure 3 shows the neutrality of the schemes under consideration, for which the initial data are the perturbation of the highest frequency modes,

$$u_j^0 = C_j + 0.0001 \times (D_{odd})_j, \tag{5.8}$$

where C_j is the highest frequency mode $C_j = (-1)^j$ and $(D_{odd})_j$ is the data given in (5.4).

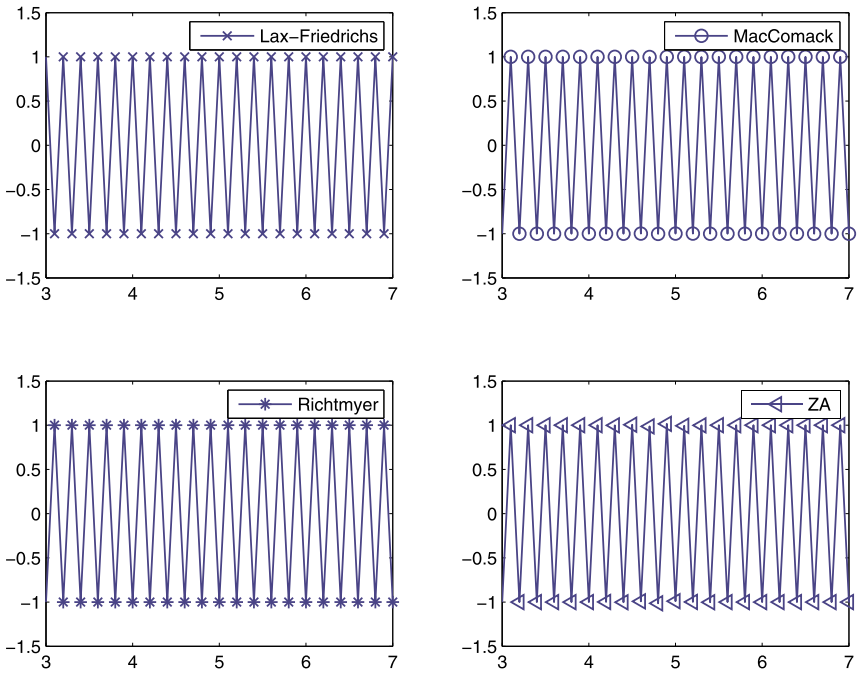


FIG. 3. *The neutrality of schemes on oscillatory modes. The conservation law is the Burgers equation with* $f(u) = u^2/2$. *The number of computation steps is nine.*

5.4. Amplification of oscillatory modes. Once the factor $r(\tilde{u})$ escapes from the interval $[0, 1]$, oscillatory modes are amplified, which may result in instability. As we consider the flux $f = \frac{u}{\sqrt{1+u^2}}$, i.e., an odd flux function, some of the above four schemes are unstable for the highest frequency mode. When we set the initial data as in the last subsection, i.e., (5.8), we can compute $r(\tilde{u})$ for these schemes for the first step (regard the initial data as the perturbation around $(-1)^{j+n}$, then correspondingly $\tilde{u}_j^n \equiv 1$),

$$\begin{cases} r_{\mathrm{LxF}}(1) = 0, \\ r_{\mathrm{Mac}}(1) = 0.6994, \\ r_{\mathrm{Rich}}(1) = -0.9188, \\ r_{\mathrm{ZA}}(1) = -2.6204. \end{cases} \tag{5.9}$$

In view of Proposition 1.1, we conclude that the Richtmyer scheme and the Zwas–Abarbanel scheme amplify oscillatory modes, while the MacCormack scheme dampens

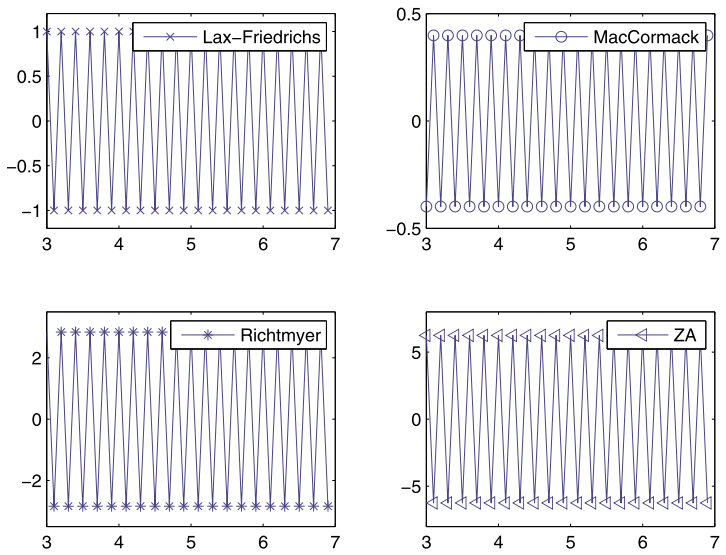


FIG. 4. *One step computation: The flux function is* $f(u) = \dfrac{u}{\sqrt{1+u^2}}$.

the modes and the Lax–Friedrichs scheme exhibits the neutral property. We show in Figure 4 the amplification effect of these schemes in one step. We further present the amplification or damping trend in four steps in Figure 5. Therefore Proposition 1.1 indicates the effect of the factor $r(\tilde{u})$ on oscillatory modes very heuristically.

6. Discussion. In this paper we use one-dimensional nonlinear scalar conservation laws to investigate the dissipation mechanism of finite difference or finite volume schemes on oscillatory modes and propose a heuristic dissipation criterion, as summarized in Proposition 1.1. In particular, we discuss several well-known schemes as specific examples: the generalized Lax–Friedrichs schemes, the MacCormack scheme, the Richtmyer scheme, and the third order Zwas–Abarbanel scheme. We also show that the oscillatory modes cannot be amplified by monotone schemes (3.1).

This criterion is believed to be valid in a general setting. For example, we can extend the criterion to multidimensional cases or hyperbolic systems. Even more, we can extend it to general time-dependent partial differential equations of the form

$$\mathbf{u}_t = \mathcal{P}(\partial_x, x, \mathbf{u})\mathbf{u}, \tag{6.1}$$

where $\mathcal{P}$ is a differential operator depending on x and u. In fact, the discretization of (6.1) takes the (explicit or implicit) form

$$A\mathbf{u}^{n+1} = B\mathbf{u}^n, \tag{6.2}$$

where A, B are matrices, and A is invertible. The ansatz (2.1) can be adopted and oscillatory modes can be analyzed analogously.

Obviously, (1.10) shows the inconsistency of (1.3) with (1.1) in regions of oscillatory modes. Hence once the oscillatory modes are present, special care should

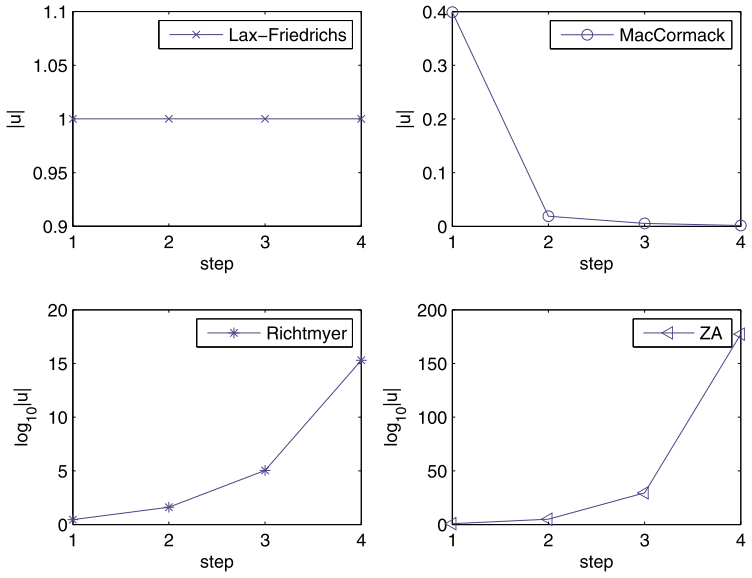


FIG. 5. *Amplification or damping trend of oscillatory modes.*

be taken in order to ensure the stability of underly schemes, and effective damping mechanisms need to be introduced, which leaves room for future study.

Acknowledgments. Both authors deeply appreciate illuminating discussions with Professor Huazhong Tang, who showed us in particular the factor $r(\tilde{u})$ for the MacCormack scheme. They also thank the anonymous referees for their very careful suggestions and criticisms, which substantially polished the present work.

REFERENCES

[1] M. BREUSS, *The correct use of the Lax-Friedrichs method*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 519–540.
[2] M. BREUSS, *An analysis of the influence of data extrema on some first and second order central approximations of hyperbolic conservation laws*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 965–993.
[3] M. CRANDALL AND A. MAJDA, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.
[4] J. GOODMAN AND A. MAJDA, *The validity of the modified equation for nonlinear shock waves*, J. Comput. Phys., 58 (1985), pp. 336–348.
[5] A. HARTEN, J. M. HYMAN, AND P. D. LAX, *On finite-difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), pp. 297–322.
[6] P. D. LAX, *Gibbs phenomena*, J. Sci. Comput., 28 (2006), pp. 445–449.
[7] P. D. LAX AND B. WENDROFF, *Systems of conservation laws*, Comm. Pure Appl. Math., 13 (1960), pp. 217–237.
[8] R. J. LEVEQUE, *Numerical Methods for Conservation Laws*, 2nd ed., Birkhäuser Verlag, 1992.
[9] J. LI, H. TANG, G. WARNECKE, AND L. ZHANG, *Local oscillations in finite difference solutions of hyperbolic conservation laws*, Math. Comp., 78 (2009), pp. 1997–2018.
[10] R. MACCORMACK, *The effect of viscosity in hypervelocity impact cratering*, AIAA paper, 1969, pp. 69–354.

[11] K. W. Morton and D. F. Mayers, *Numerical Solution of Partial Differential Equations*, 2nd ed., Cambridge University Press, Cambridge, UK, 2005.
[12] J. von Neumann and R. D. Richtmyer, *A method for the numerical calculation of hydrodynamic shocks*, J. Appl. Phys., 21 (1950), pp. 232–237.
[13] H. Z. Tang and G. Warnecke, *A note on* $(2K+1)$*–point conservative monotone schemes*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 345–357.
[14] R. F. Warming and B. J. Hyett, *The modified equation approach to the stability and accuracy analysis of finite difference methods*, J. Computational Phys., 14 (1974), pp. 159–179.
[15] G. Zwas and S. Abarbanel, *Third and fourth order accurate schemes for hyperbolic equations of conservation law form*, Math. Comp., 25 (1971), pp. 229–236.