Chapter 2
Local and Global Error Estimates

Abstract In this chapter we analyze some simple examples, which suggest that the error quantification should take into account of the possible grow in time of the error. This observation provides a motivation for going beyond more classical local-in-time concepts of error (so-called *Local Truncation Error*).

Keywords Accuracy of schemes for balance laws · Local truncation error

It's a well-known fact that quasilinear hyperbolic equations generally admit only weak solutions, in the sense that discontinuities develop and propagate along distinguished directions (at least in one space dimension, the situation in 2D being more delicate). Their mathematical analysis is usually carried out within a Banach space of discontinuous functions with finite total variation, $BV(\mathbb{R})$, sometimes intersected with $L^1(\mathbb{R})$, in order to ensure some integrability properties (positive total mass should be preserved).

However, when it comes to derive numerical algorithms meant to approximate their (entropy) solutions on a countable lattice, people frequently evoke so-called "high-order schemes", which may seem a puzzling notion, as they rely on Taylor expansions. Besides, Kuznetsov's method indicates that, as soon as numerical viscosity is present in the numerical process, the best convergence rate (hence an upper bound for the L^1 global error) is of the order of $\sqrt{t \cdot \Delta x}$, much less than any high-order rate. So, in which sense should one understand high-order, left apart the fact that one cannot reach more than second-order accuracy when computing only local cell averages?

Hereafter, important differences between local truncation errors, a notion inherited from smooth solutions of Ordinary Differential Equations (ODE), and global errors for possibly weak solutions of Partial Differential Equations (PDE) is put at forefront. It is explained that "high-order", as a notion, essentially refers to a local in time residual, in vicinity of a point where the considered solution is as smooth as possible, which should be further bounded uniformly and integrated in order to produce a meaningful global error estimate. This sheds light on the apparent discrepancy between second-order, so-called MinMod schemes (see e.g. [3, 27, 28]) for linear advection, and the actual convergence rates (slightly over $\frac{1}{2}$, see [19]) that were obtained only recently. Local estimates are most of the times insensitive to

D. Amadori and L. Gosse, *Error Estimates for Well-Balanced Schemes on Simple Balance Laws*, SpringerBriefs in Mathematics,
DOI 10.1007/978-3-319-24785-4_2

specific features of elaborate numerical schemes [9, 11], whereas global ones may be better suited for such purposes. This is illustrated on a model of linearized shallow water equations over topography, admitting smooth solutions.

2.1 Notion of Local Truncation Error (LTE)

Let's focus onto a one-dimensional convex scalar conservation law,

$$\partial_t u + \partial_x f(u) = 0, \qquad u(t=0,\cdot) = u_0 \in BV(\mathbb{R}), \qquad (t,x) \in \mathbb{R}_*^+ \times \mathbb{R}, \quad (2.1)$$

within a Cartesian uniform computational grid, determined by a space-step Δx and a time-step Δt satisfying the standard homogeneous CFL restriction. By defining $C_k = (x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}})$ as the generic computational cell of width Δx centered on $x_k = k\Delta x$, $k \in \mathbb{Z}$, one may apply the Divergence Theorem on any rectangle $C_k^n := C_k \times (t^n, t^{n+1})$ in order to derive a mass-preserving numerical scheme for (2.1):

$$\int_{C_k} u\left(t^{n+1}, x\right) \mathrm{d}x = \int_{C_k} u\left(t^n, x\right) \mathrm{d}x - \int_{t^n}^{t^{n+1}} f\left(u\left(\tau, x_{k+\frac{1}{2}}\right)\right) - f\left(u\left(\tau, x_{k-\frac{1}{2}}\right)\right) \mathrm{d}\tau.$$

This is equivalent to the weak formulation of (2.1) with test-functions being indicator functions of C_k, denoted $\chi(C_k)$. Hereafter, as $u_k^n = \int_{C_k} u(t^n, x) \frac{\mathrm{d}x}{\Delta x}$, the observation yielding Godunov's scheme is the following: in case $u(t^n, \cdot)$ is constant on each computational cell C_k, then the boundary flux terms can be explicitly obtained by resolving all the discontinuities, that is to say, Riemann problems at both interfaces $x_{k\pm\frac{1}{2}}$. Moreover, since Riemann fans $\omega(\frac{x}{t}; u^L, u^R)$ display a self-similar structure,

$$\int_{t^n}^{t^{n+1}} f\left(u\left(\tau, x_{k+\frac{1}{2}}\right)\right) \mathrm{d}\tau = \Delta t \cdot f\left(\omega\left(0; u_k^n, u_{k+1}^n\right)\right). \quad (2.2)$$

In the context of an explicit time-marching algorithm, one may want to get rid of the Riemann solution ω, thus (2.2) rewrites as a smooth and consistent *numerical flux function*, $F : \mathbb{R}^2 \to \mathbb{R}$,

$$\forall u, v \in \mathbb{R}^2, \qquad F(u,v) = f\left(\omega(0; u, v)\right) = \begin{cases} \min\limits_{u \le \xi \le v} f(\xi) & \text{if } u \le v \\ \max\limits_{v \le \xi \le u} f(\xi) & \text{if } u > v. \end{cases} \quad (2.3)$$

For any indexes $k, n \in \mathbb{Z} \times \mathbb{N}$, the Godunov averaging furnished an approximate (formally first-order) value $u_k^n \simeq u(t^n, x_k)$. In order to locally increase its accuracy, a piecewise-linear reconstruction can be set up in each cell,

$$v_k^n : C_k \to \mathbb{R}, \qquad v_k^n(x) = u_k^n + (x - x_k)\sigma_k^n. \quad (2.4)$$

A common way to proceed is by analogy with Lax-Wendroff second-order scheme for (2.1) with linear flux, $f(u) = u$: the resulting definition of local slopes reads,

$$\sigma_k^n = \frac{u_{k+1}^n - u_k^n}{\Delta x}\phi(r_k^n), \qquad r_k^n = \frac{u_k^n - u_{k-1}^n}{u_{k+1}^n - u_k^n}.$$

The slope-limiter ϕ must meet with several constraints in order to ensure both the TVD property and formal second-order accuracy in smooth regions, see [26].

Remark 2.1 Another way to motivate MUSCL piecewise-linear reconstructions is to work out the ODE system obtained by semi-discretization in space (the "Method of Lines", evoked in [15]) in order to obtain a Local (space-) Truncation Error in Δx^2 for smooth exact solutions u: see e.g. [20, 21, 29, 30].

2.1.1 Semi-discretization in Space (Method of Lines)

Hereafter we follow the canvas of Cullen and Morton [6] in order to shed some light onto various general mechanisms of error creation/propagation (see also [4, 30]). Let a Cauchy problem for a given partial differential operator $\mathscr{L}$ be,

$$\partial_t u = \mathscr{L}u, \qquad u(t=0,\cdot) = u_0. \tag{2.5}$$

For $\Delta x > 0$ fixed and the corresponding griding of the real line, a finite-differences approximation of $\mathscr{L}$ acting on $\Delta x \cdot \mathbb{Z}$ is denoted by $\mathscr{L}_{\Delta x}$, so (2.5) reduces to an (infinite) differential system (*Method of Lines*, in [15, Chap. 17]), with $\tilde{u}(t,\cdot) \in \ell^\infty(\mathbb{Z})$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{u} = \mathscr{L}_{\Delta x}\tilde{u}, \qquad \tilde{u}(t=0,\cdot) = \mathscr{P}_{\Delta x}u_0, \tag{2.6}$$

for which one can worry about the global error $u - \tilde{u}$ at each time $t > 0$.

- One "triangulates" $u(t,\cdot) - \tilde{u}(t,\cdot)$ by inserting $\mathscr{P}_{\Delta x}u(t,\cdot)$,

$$u - \tilde{u} = (Id - \mathscr{P}_{\Delta x})u + (\mathscr{P}_{\Delta x}u - \tilde{u}) := a_{\Delta x} + e_{\Delta x},$$

  where $a_{\Delta x}$ is purely an approximation error, but $e_{\Delta x}$ stands for an evolutionary error, which may accumulate in time, and satisfies a differential equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}e_{\Delta x} = \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{P}_{\Delta x}u - \frac{\mathrm{d}}{\mathrm{d}t}\tilde{u} = \mathscr{P}_{\Delta x}\mathscr{L}u - \mathscr{L}_{\Delta x}\tilde{u}. \tag{2.7}$$

- Triangulating again, one gets

$$\frac{\mathrm{d}e_{\Delta x}}{\mathrm{d}t} = (\mathscr{P}_{\Delta x}\mathscr{L}u - \mathscr{L}_{\Delta x}\mathscr{P}_{\Delta x}u) + (\mathscr{L}_{\Delta x}\mathscr{P}_{\Delta x}u - \mathscr{L}_{\Delta x}\tilde{u}),$$

leading to,

$$\frac{\mathrm{d}}{\mathrm{d}t}e_{\Delta x} + (\mathscr{L}_{\Delta x}\tilde{u} - \mathscr{L}_{\Delta x}\mathscr{P}_{\Delta x}u) = (\mathscr{P}_{\Delta x}\mathscr{L}u - \mathscr{L}_{\Delta x}\mathscr{P}_{\Delta x}u) := L.T.E.,$$

and by substituting $\tilde{u}$ by $\mathscr{P}_{\Delta x}u - e_{\Delta x}$, we get finally:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}e_{\Delta x} + \left[\mathscr{L}_{\Delta x}(\mathscr{P}_{\Delta x}u - e_{\Delta x}) - \mathscr{L}_{\Delta x}\mathscr{P}_{\Delta x}u\right] = L.T.E.,} \tag{2.8}$$

(L.T.E. standing for Local Truncation Error). Hence, the L.T.E. is just a source term inside the differential equation (2.8) governing the scheme's evolutionary error; this was noted in [17, 20, 30], too.

In case both (2.5) and its (consistent) discrete approximation $\mathscr{L}_{\Delta x}$, are dissipative ("contractive" [21, 29], "strongly stable" in a terminology of [17]) in some norm, this source term is responsible for most of the error $e_{\Delta x}$; if, on the contrary, (2.5) happens to be accretive, for instance if $\|u(t) - v(t)\| \leq K\|u_0 - v_0\|$ with $K > 1$ like in Bressan-Glimm's theory of strictly hyperbolic systems of conservation laws [5], then both $\mathscr{L}_{\Delta x}$ and the L.T.E. can contribute to the increase of the evolutionary error.

Remark 2.2 If the approximation $\mathscr{L}_{\Delta x}$ is linear, then (2.8) simplifies into,

$$\boxed{\forall t > 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}e_{\Delta x}(t) = \mathscr{L}_{\Delta x}e_{\Delta x}(t) + \tau_u(t),}$$

where $\tau_u(t)$ stands for the L.T.E. related to (x-derivatives of) the exact solution $u(t, \cdot)$ to (2.5) at time t. Duhamel's principle yields an expression of the evolutionary error,

$$e_{\Delta x}(t) = \exp(t \cdot \mathscr{L}_{\Delta x})\left(e_{\Delta x}(t = 0) + \int_0^t \exp(-s \cdot \mathscr{L}_{\Delta x})\tau_u(s)\mathrm{d}s\right).$$

Quantities like $\exp(t \cdot \mathscr{L}_{\Delta x})$ may be estimated by "logarithmic norms", see e.g. [21].

The local truncation error is intrinsically a "low frequency" information: being a byproduct of a Taylor expansion, it tacitly assumes that higher-order derivatives of the solution get smaller. Accordingly, it was recast in the formalism of Fourier analysis in, e.g., [31, sect. 2.4]. In order to cope with oscillating, possibly high-frequency, solutions of linear transport (advection) equations, finite-difference schemes endowed with "spectral-like resolution" were devised in [12, 14, 25] for which both numerical dissipation and dispersion were carefully scrutinized. For specific problems, like the linear wave equation recast as a multi-D "div-grad" hyperbolic system, numerical schemes can be derived from the exact Kirchoff's method of spherical means: see [1] and especially the connections with the classical Lax-Wendroff formalism.

2.1.2 Local Truncation Error (LTE) and Second-Order Accuracy

Second-order accuracy in space for (2.1), or linear advection equations, was studied in [18] (see also [10, 27]). These equations are dissipative in L^1, so the former analysis yielding (2.8) indicates that the local truncation error is probably the main source of evolutionary error. For $\mathscr{L}u = -\partial_x f(u)$, it reads: $\forall k \in \mathbb{Z}$,

$$\mathscr{P}_{\Delta x}\mathscr{L}u(t,x_k) = -\frac{1}{\Delta x}\int_{x_{k-\frac{1}{2}}}^{x_{k+\frac{1}{2}}} \partial_x f(u)\mathrm{d}x = -\frac{f\left(u\left(t,x_{k+\frac{1}{2}}\right)\right) - f\left(u\left(t,x_{k-\frac{1}{2}}\right)\right)}{\Delta x},$$

by exact integration of the conservation law (2.1). Now, since high-order accuracy is only concerned with smooth exact solutions u, one approximates this expression with a second-order mid-point rule by taking advantage of $x_{k+\frac{1}{2}} = \frac{x_{k+1}+x_k}{2}$,

$$\mathscr{P}_{\Delta x}\mathscr{L}u(t,x_k) = \frac{f\left(\frac{u(t,x_{k+1})+u(t,x_k)}{2}\right) - f\left(\frac{u(t,x_k)+u(t,x_{k-1})}{2}\right) + O(\Delta x^2)}{\Delta x},$$

and so, the L.T.E. is the difference between this approximation and the numerical scheme $\mathscr{L}_{\Delta x}$ applied to the piecewise constant projection of the exact solution, $\mathscr{P}_{\Delta x}u$. Since $\mathscr{L}_{\Delta x}$ needs to be conservative and consistent with $\mathscr{L}$, we assume it is given by a (smooth) numerical flux which reads, in standard notation,

$$\tilde{F}_{k+\frac{1}{2}} = F\left(u^L_{k+\frac{1}{2}}, u^R_{k+\frac{1}{2}}\right), \qquad \mathscr{L}_{\Delta x}\mathscr{P}_{\Delta x}u(t,x_k) = \frac{\tilde{F}_{k+\frac{1}{2}}(t) - \tilde{F}_{k-\frac{1}{2}}(t)}{\Delta x},$$

where $u^{L/R}_{k+\frac{1}{2}}$ are obtained from the set of cell-centered values $\mathscr{P}_{\Delta x}u$ by means of a reconstruction like (2.4) and F is, for instance, the exact Godunov flux (2.3). Hence,

$$\boxed{L.T.E. = \frac{\left[f\left(\frac{u(t,x_{k+1})+u(t,x_k)}{2}\right) - \tilde{F}_{k+\frac{1}{2}}\right] - \left[f\left(\frac{u(t,x_k)+u(t,x_{k-1})}{2}\right) - \tilde{F}_{k-\frac{1}{2}}\right]}{\Delta x}.}$$

As the CFL condition imposes $\Delta t = O(\Delta x)$, second-order accuracy asks for,

$$\left| f\left(\frac{u(t,x_{k+1})+u(t,x_k)}{2}\right) - \tilde{F}_{k+\frac{1}{2}}(t)\right| = O(\Delta x^2),$$

which, by the smoothness of the flux functions, reduces simply to,

$$\forall t,k \in \mathbb{R}^+ \times \mathbb{Z}, \qquad \left| u^{L/R}_{k+\frac{1}{2}}(t) - \frac{u(t,x_{k+1})+u(t,x_k)}{2}\right| = O(\Delta x^2). \tag{2.9}$$

And this meets with the definition used by Osher (see Lemma 2.1, in [18, p. 953]) and Sjogreen (see Theorem 3.9 in [24, p. 47]). A slightly different derivation of a second-order scheme for smooth solutions is given in [4, p. 53], essentially by keeping the term $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{P}_{\Delta x}u$ in (2.7) inside the expression of the L.T.E as follows:

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}\mathscr{P}_{\Delta x}u(t,\cdot) &= \lim\nolimits_{\Delta t\to 0}\left(\tfrac{\mathscr{P}_{\Delta x}u(t+\Delta t,\cdot)-\mathscr{P}_{\Delta x}u(t,\cdot)}{\Delta t}\right)\\ &= -\tfrac{F\left(u(t,\cdot+\Delta x),u(t,\cdot)\right)-F\left(u(t,\cdot),u(t,\cdot-\Delta x)\right)}{\Delta x},\end{aligned}$$

where F is the exact flux defined in (2.3). The L.T.E. is now defined like,

$$\boxed{\forall k\in\mathbb{Z},\qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{P}_{\Delta x}u(t,x_k)-\mathscr{L}_{\Delta x}\mathscr{P}_{\Delta x}u(t,x_k) = -\frac{\mathscr{F}_{k+\frac{1}{2}}(t)-\mathscr{F}_{k-\frac{1}{2}}(t)}{\Delta x},}$$

where $\mathscr{F}_{k+\frac{1}{2}}(t) = F\big(u(t,x_k+\Delta x),u(t,x_k)\big)-F_{k+\frac{1}{2}}(t)$. The scheme induced by the numerical flux $F_{k+\frac{1}{2}}$ is called second-order in space as soon as, for any smooth exact solution $u(t,\cdot)$, $\mathscr{F}_{k+\frac{1}{2}}$ is a quadratic quantity (possibly depending on $|\partial_{xx}u(t,\cdot)|$),

$$\forall t\geq 0,\qquad |\mathscr{F}_{k+\frac{1}{2}}(t)| = O(\Delta x^2). \tag{2.10}$$

Clearly, both criteria pick up variants of the (unstable) "centered scheme" which is second-order, but unstable because it lets the total variation increase strongly.

2.1.3 Illustration of Two Errors Not Controlled by LTE

The L.T.E. was defined as, given an approximate function space, the difference between the projection of the exact (smooth) solution and the outcome of the numerical process, generally after an infinitesimally small time-step $\Delta t > 0$, with identical initial values. Its shortcomings are henceforth revealed by putting in default these two main assumptions: smoothness and identical data.

Let's first present a classical example of issues which result from a lack of smoothness in the exact solution of the considered problem. Assume (2.5) is a 2×2 genuinely nonlinear system of conservation laws, like for instance the p-system. If prescribed initial data produces (after some time) a forward-going 2-shock, the numerical viscosity inherent in a standard Godunov-type scheme (2.6) breaks that shock into smaller ones, and the resulting discontinuity appears to be spread on a certain number of computational cells: see Fig. 2.1. However, the 2-shock present in the exact solution is meant to connect two states (p_L,u_L) and (p_R,u_R) belonging to the 2-shock curve in phase space, which isn't a straight line because the considered system isn't in the Temple class. As soon as it gets broken into smaller jumps by means of artificial viscosity, there's no reason for those ones to belong to the same identical curve: hence, as the time-marching scheme will advance by solving elementary Riemann

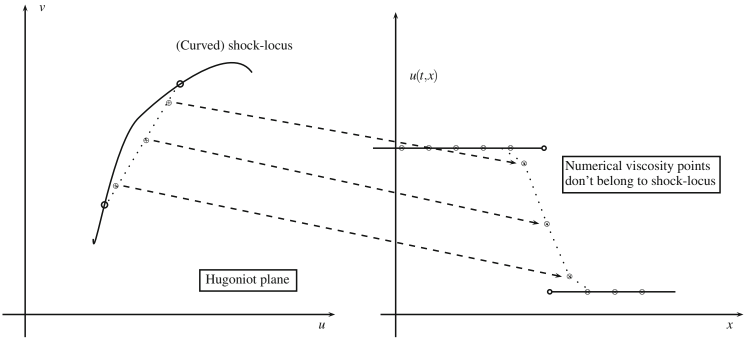
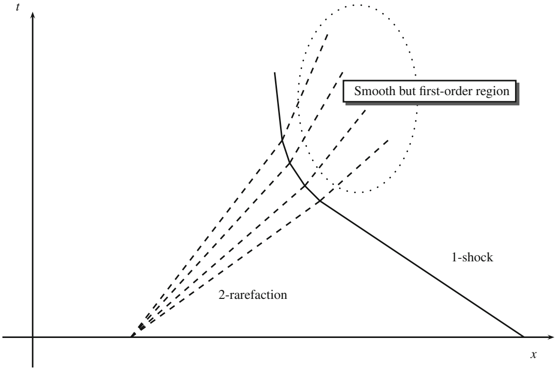


Fig. 2.1 Numerical viscosity and its effects for systems with curved shock loci

Fig. 2.2 Nonlinear 2 × 2 interaction and loss of accuracy in a smooth region



problems, spurious 1-waves are going to develop inside the numerical shock layer, as shown in e.g. [13]. Formal second-order schemes improve marginally the situation by displaying shocks spread on a reduced area, but can't eliminate this drawback of shock-capturing techniques. A thorough analysis of the deformation of shock curves in phase space was achieved by [2] and later in [23].

The second well-known example of a fallout for L.T.E. consists in setting up initial data for which there's no approximation error, for instance two Riemann problems exactly represented on the computational grid, but yielding two waves of different nature which interact in finite time. For instance, always considering the p-system, suppose that in $x = -1$ one prescribes a jump between (p_L, u_L) and (p_M, u_M) belonging to the 2-rarefaction curve, whereas in $x = 1$, another one between (p_M, u_M) and (p_R, u_R) belonging to the 1-shock curve. After some time, these two approaching waves interact and a new middle state $(\tilde{p}_M, \tilde{u}_M)$ appears with the 1-shock on the left and the 2-rarefaction on its right: see Fig. 2.2. This pattern is called a "scattering state", meaning that no wave interaction can occur anymore. Incoming waves are both the smooth 2-rarefaction, inside which a second-order scheme will actually display its formal accuracy, and the 1-shock, in the vicinity of which only first-order accuracy holds for the sake of stability. The 2-rarefaction spawn from the interaction may not be second-order accurate with respect to the exact solution because it results of the mixing of both first and second order data. This example was first presented in [8], and then studied in [7, 22] (see also [16, 32]).

2.2 Linearized Shallow Water with Topography

Hereafter, an elementary example, despite its simplicity, will reveal already a part of the specific features that govern the more complex non-linear cases when source terms are included into a system of equations. By linearizing around a static state $\bar{\rho} > 0, \bar{u} = 0$ the usual one-dimensional shallow water equations with topography,

$$\partial_t \rho + \partial_x(\rho u) = 0, \qquad \partial_t(\rho u) + \partial_x \left(\rho u^2 + \frac{\rho^2}{2}\right) = -\rho \partial_x a, \tag{2.11}$$

the following system arises,

$$\partial_t \rho + \partial_x J = 0, \qquad \partial_t J + \partial_x \rho = -\partial_x a \tag{2.12}$$

where $a = a(x)$ stands for the bottom. By linearity, solutions $\rho(t, \cdot)$, $J(t, \cdot)$ display identical integrability and smoothness as initial data. An alternative, customary way to deal with system (2.12) is to consider $a = a(x)$ as an independent variable,

$$\begin{cases} \partial_t \rho + \partial_x J = 0, \\ \partial_t J + \partial_x(\rho + a) = 0 \\ \partial_t a = 0. \end{cases} \tag{2.13}$$

2.2.1 Study of the Error Growth in Time

Characteristic speeds of system (2.12) are ±1; diagonal variables $f^\pm = \rho \pm J$ satisfy

$$(\partial_t - \partial_x) f^- = \partial_x a, \qquad (\partial_t + \partial_x) f^+ = -\partial_x a; \tag{2.14}$$

equivalently, the system (2.13) is diagonalized as follows:

$$\begin{cases} \partial_t(f^\pm + a) \pm \partial_x(f^\pm + a) = 0 \\ \partial_t a = 0. \end{cases} \tag{2.15}$$

Accordingly, two approaches coexist for approximating the solutions of (2.12):

1. The standard “centered source method” which consists in processing a set of two advection equations with a source term, (2.14).
2. The “well-balanced method”, which treats two *homogeneous* advection equations from (2.15): $\partial_t f^{\pm} \pm \partial_x (f^{\pm} + a) = 0$.

Define the Courant number ν as $\Delta t = \nu \Delta x$ with $0 < \nu \leq 1$, together with the function $a \in C^{\infty}(\mathbb{R})$ having compact support, and $f^{\pm}(t = 0, \cdot) \in C^3 \cap W^{3,\infty}(\mathbb{R})$. With respect to the preceding section, errors in both space and time will be considered.

2.2.2 Analysis of Scheme 1

In each C_j^n, the residual R_j^n of the “centered source method” is computed by plugging exact solutions into each inhomogeneous advection equation: for instance,

$$\frac{f^+(t^{n+1}, x_j) - f^+(t^n, x_j)}{\Delta t} + \frac{f^+(t^n, x_j) - f^+(t^n, x_{j-1})}{\Delta x} + k(x_j) = R_j^n,$$

which is the L.T.E. of the scheme in both space and time. By Taylor expansion,

$$|R_j^n| \leq \frac{\Delta x}{2}(1-\nu)\, \|\partial_{xx} f^{\pm}(t^n, \cdot)\|_{\infty} + \frac{\Delta t}{2} \|\partial_{xx} a(x)\|_{\infty} + C(\Delta x)^2$$

where the C above depends on both $\|\partial_{xxx} f^{\pm}(t^n, \cdot)\|_{\infty}$ and $\|\partial_{ttt} f^{\pm}(t^n, \cdot)\|_{\infty}$. By linearity, the second derivative above, $\partial_{xx} f^{\pm}(t, \cdot)$, is bounded by

$$\|\partial_{xx} f^{\pm}(t, \cdot)\|_{\infty} \leq \|\partial_{xx} f^{\pm}(t = 0)\|_{\infty} + t\|a\|_{C^3}.$$

There are two error amplification mechanisms for the “centered source method”: the fact that R_j^n contains $\Delta t \|\partial_{xx} a\|_{\infty}$ which doesn’t vanish as $\nu = \frac{\Delta t}{\Delta x} = 1$ (numerical viscosity), and the linear growth of $t \mapsto \|\partial_{xx} f^{\pm}(t, \cdot)\|_{\infty}$. Yet, the pointwise error,

$$(E^{\pm})_j^n = (f^{\pm} - f^{\pm}_{\Delta x})(t^n, x_j),$$

(where $f^{\pm}_{\Delta x}$ stands for the piecewise-constant approximation of the exact solution $f^{\pm}$) satisfies a slightly modified upwind scheme that easily rewrites under the form of a convex combination plus a source term: for instance,

$$(E^+)_j^{n+1} = (E^+)_j^n \left(1 - \frac{\Delta t}{\Delta x}\right) + \frac{\Delta t}{\Delta x}(E^+)_{j-1}^n + \Delta t R_j^n.$$

It remains to take the modulus, maximize, and to sum over n all the residuals in order to derive that, for any $t \geq 0$, the following error estimate holds:

$$\|E^{\pm}(t,\cdot)\|_{L^\infty} \leq \|E^{\pm}(t=0,\cdot)\|_{L^\infty} + \Delta x \cdot t\Big[\nu\|a\|_{C^2} + (1-\nu)\left(\|f^{\pm}(t=0,\cdot)\|_{C^2} + t\|a\|_{C^3}\right)\Big] + O(1)t(\Delta x)^2 \tag{2.16}$$

where the $O(1)$ above depends on $\|f^{\pm}(t=0,\cdot)\|_{C^3}$, $\|a\|_{C^4}$ and linearly on t.

2.2.3 Analysis of Scheme 2

Oppositely, for the well-balanced method, the same LTE analysis starts with:

$$\frac{f^{+}(t^{n+1},x_j) - f^{+}(t^n,x_j)}{\Delta t} + \frac{(f^{+}+a)(t^n,x_j) - (f^{+}+a)(t^n,x_{j-1})}{\Delta x} = \tilde{R}_j^n.$$

The full upwinding of the topography function a yields a smaller residual:

$$|\tilde{R}_j^n| \leq \frac{\Delta x}{2}(1-\nu)\left\|\partial_{xx}[f^{\pm}(t,\cdot)+a(\cdot)]\right\|_{\infty} + C(\Delta x)^2,$$

with C depending on the sup norm of $\partial_{xxx}(f^{\pm}+a)$. Since diagonal variables $V = (f^{\pm}+a, a)$ are constant along their characteristics, the sup-norm of their derivatives doesn't grow in time:

$$|\tilde{R}_j^n| \leq \frac{\Delta x}{2}(1-\nu)\left\|\partial_{xx}[f^{\pm}(t=0,\cdot)+a(\cdot)]\right\|_{\infty} + C_0(\Delta x)^2,$$

for some constant C_0. Again, the scheme governing the pointwise error

$$(\tilde{E}^{\pm})_j^n = [(f^{\pm}+a) - (f^{\pm}_{\Delta x}+a)](t^n,x_j)$$

rewrites as a convex combination, so a better global error estimate follows:

$$\|\tilde{E}^{\pm}(t,\cdot)\|_{L^\infty} \leq \|\tilde{E}^{\pm}(t=0,\cdot)\|_{L^\infty} + \Delta x \cdot \frac{t}{2}(1-\nu)\|f^{\pm}(t=0,\cdot)+a\|_{C^2} + C_0 t(\Delta x)^2. \tag{2.17}$$

2.2.4 Summary and Illustration

Let's enumerate the main differences between (2.16) and (2.17):

1. The error (2.16) of the standard method is a polynomial (quadratic in the present case) function of time whereas the well-balanced estimate (2.17) is linear.

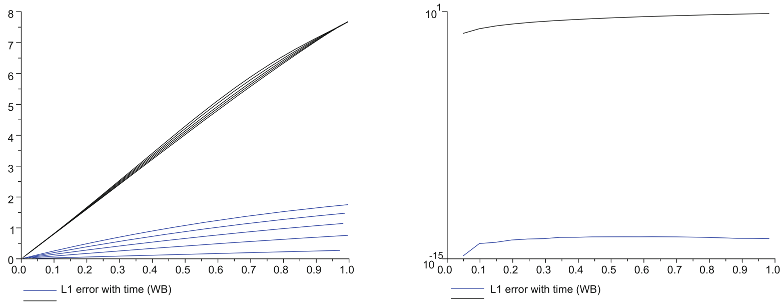


Fig. 2.3 Time-evolution of the L^∞ errors of well-balanced (*blue*) and centered source (*black*) methods for smooth $f^\pm$, a: $\nu = 0.9, 0.7, 0.5, 0.3, 0.1$ (*left*) and $\nu = 1$ (*right*)

2. In (2.16), the source term always contributes at the same rate regardless to the value of the Courant number ν. Oppositely, fixing $\nu = 1$ implies that for the well-balanced method, the initial error (2.17) stands still,

$$\|\tilde{E}^\pm(t,\cdot)\|_{L^\infty} \simeq \|\tilde{E}^\pm(t=0,\cdot)\|_{L^\infty}.$$

3. The dependence on a and its derivates of the two estimates (2.16) and (2.17) is much different: for small times (at the first order in t), (2.16) depends on

$$\nu\|a\|_{C^2} \;+\; (1-\nu)\,\|f^\pm(t=0,\cdot)\|_{C^2}$$

while (2.17) depends on

$$(1-\nu)\|f^\pm(t=0,\cdot)+a\|_{C^2}.$$

That is, compensation between a and the initial data is allowed. Moreover, as observed before, the quantity in the above formula vanishes for $\nu = 1$.

On Fig. 2.3, both these theoretical aspects are illustrated: the graphic on the left shows that the well-balanced pointwise error $\tilde{E}$ grows more strongly when $\nu \to 0$ (the numerical viscosity increases) and the right one reveals that $\tilde{E}$ remains constant in time when $\nu = 1$ but this nice property doesn't hold for a conventional method.

Remark 2.3 One may wonder if the global error estimate for the time-splitting strategy may be improved, in particular its dependence with respect to the derivatives of $a(x)$. A simple computation, relying on elementary stability theory of position-dependent ODE, reveals that in general this cannot be the case: the space dependence of a (more precisely, on $k = a'$) usually introduces additional variation in the unknown, when the time-split update is applied. Intuitively, this occurs because the differential equation in time, at each time-step, is sensitive to the changing values of k. Let $g : \mathbb{R} \to \mathbb{R}$ be bounded and Lipschitz continuous and pick $k_1, k_2 \in \mathbb{R}^2$, corresponding solutions $y_1(t)$, $y_2(t)$

$$\dot{y}_1 = k_1 \cdot g(y_1), \qquad \dot{y}_2 = k_2 \cdot g(y_2), \qquad y_1(0) = y_2(0),$$

clearly satisfy[1]

$$|y_1(t) - y_2(t)| \leq \|g\|_\infty \, |k_2 - k_1| \, t. \tag{2.18}$$

For a scalar law with source term $\partial_t u + \partial_x f(u) = k(x)g(u)$, see next Chap. 3, a time splitting approach typically asks for solving $\partial_t u = k(x)g(u)$; from (2.18), a local amplification of the approximation $u^{\Delta x}$, of the order of $t \cdot \mathrm{TV}\,(k)$, is induced. Most of the available error estimates for such equations contain $\mathrm{TV}\,(u^{\Delta x})$, thus one may expect global errors of time-splitting strategies to grow with $t \cdot \mathrm{TV}\,(k)$, especially if $g'(u)$ is nonnegative. Oppositely, well-balanced strategies can be proved to obey error bounds depending only on the L^1 norm of k: assuming that $k \in L^1 \cap BV(\mathbb{R})$ has compact support, say (a, b), there exists an optimal constant of Poincaré's inequality,

$$\boxed{\|k\|_{L^1(a,b)} \leq \frac{b-a}{2} \, \mathrm{TV}\,(k),}$$

a quite interesting property for applications where k displays steep gradients, like for instance (2.11) with a topography term, or periodic forcing as well.

2.3 Preliminary Conclusions

The L.T.E., as recalled in the beginning of this chapter, is too weak as a notion to reliably quantify the errors generated by a numerical scheme approximating one-dimensional (systems of) balanced because it doesn't take into account for several things, like for instance the time-propagation of incomplete initial data approximations. More specifically, the loss of smoothness in a compressive shock can induce peculiar issues in the form of spurious waves of different characteristic families, which may react in the presence of an accretive source term. Those issues are still not perceived by the L.T.E. which generally isn't well suited for dealing with accretive problems. Error growth taking place in nonlinear wave interactions displays similar features, namely it combines local reduction of accuracy close to a shock wave with data mixing. It therefore follows that a more relevant object of interest appears to be the global error, considered as a function depending on both the time and grid-size. Especially, it perceives a possible time-amplification of errors already present in the approximation of initial data. Moreover, it is able to fully account for limited smoothness of solutions in case it doesn't rely on Taylor expansions. This is the main tool which is about to be exploited in the sequel of this book.

[1] Indeed, let $Y(t)$ satisfy $Y' = g(Y)$, $Y(0) = y_1(0) = y_2(0)$, so that $y_j(t) = Y(k_j t)$, $j = 1, 2$, and then

$$|y_1(t) - y_2(t)| = \left| \int_{k_2 t}^{k_1 t} g(Y(\tau)) \, d\tau \right| \leq \|g\|_\infty \, |k_2 - k_1| \, t.$$

References

1. B. Alpert, L. Greengard, T. Hagstrom, An integral evolution formula for the wave equation. J. Comput. Phys. **162**(2), 536–543 (2000)
2. M. Arora, P.L. Roe, On postshock oscillations due to capturing schemes in unsteady flows. J. Comput. Phys. **130**, 25–40 (1997)
3. C. Berthon, C. Sarazin, R. Turpault, Space-time generalized Riemann problem solvers of order k for linear advection with unrestricted time step. J. Sci. Comput. **55**, 268–308 (2013)
4. F. Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources*, Frontiers in Mathematics Series (Birkhäuser, Boston, 2004). ISBN 3-7643-6665-6
5. A. Bressan, *Hyperbolic systems of conservation laws—the one-dimensional Cauchy problem*, vol. 20, Oxford Lecture Series in Mathematics and Its Applications (Oxford University Press, Oxford, 2000)
6. M.J.P. Cullen, K.W. Morton, Analysis of evolutionary error in finite element and other methods. J. Comput. Phys. **34**, 245–267 (1980)
7. G. Efraimsson, G. Kreiss, A remark on numerical errors downstream of slightly viscous shocks. SIAM J. Numer. Anal. **36**, 853–863 (1999)
8. B. Engquist, B. Sjögreen, The convergence rate of finite difference schemes in the presence of shocks. SIAM J. Numer. Anal. **35**, 2464–2485 (1998)
9. H. Gilquin, Une famille de schémas numériques T.V.D. pour les lois de conservation hyperboliques. RAIRO—Model. Math. Anal. Numer. **20**, 429–460 (1986)
10. J.B. Goodman, R.J. LeVeque, A geometric approach to high resolution TVD schemes. SIAM J. Numer. Anal. **25**, 268–284 (1988)
11. L. Gosse, MUSCL reconstruction and Haar wavelets. Commun. Math. Sci. **13**, 1501–1514 (2015)
12. Z. Haras, S. Ta'asan, Finite difference schemes for long-time integration. J. Comput. Phys. **114**(2), 265–279 (1994)
13. S. Jin, J.-G. Liu, The effects of numerical viscosities. I. Slowly moving shocks. J. Comput. Phys. **126**, 373–389 (1996)
14. S. Lele, Compact finite difference schemes with spectral-like resolution. J. Comput. Phys. **103**(1), 16–42 (1992)
15. R.J. LeVeque, *Numerical Methods for Conservation Laws* (Birkhauser, ETH Zurich, Basel, 1992)
16. R. Menikoff, Errors when shock waves interact due to numerical shock width. SIAM J. Sci. Comput. **15**, 1227–1242 (1994)
17. K.W. Morton, On the analysis of finite volume methods for evolutionary problems. SIAM J. Numer. Anal. **35**, 2195–2222 (1998)
18. S. Osher, Convergence of generalized MUSCL schemes. SIAM J. Numer. Anal. **22**, 947–961 (1985)
19. B. Popov, O. Trifonov, Order of convergence of second order schemes based on the MINMOD limiter. Math. Comput. **75**, 1735–1753 (2006)
20. J.M. Sanz-Serna, J.G. Verwer, Convergence analysis of one-step schemes in the method of lines. Appl. Math. Comput. **31**, 183–196 (1989)
21. J.M. Sanz-Serna, J.G. Verwer, Stability and convergence at the PDE/stiff ODE interface. Appl. Numer. Math. **5**, 117–132 (1989)
22. M. Siklosi, G. Efraimsson, Analysis of first order errors in shock calculations in two space dimensions. SIAM J. Numer. Anal. **43**, 672–685 (2005)
23. M. Siklosi, B. Batzorig, G. Kreiss, An investigation of the internal structure of shock profiles for shock capturing schemes. J. Comput. Appl. Math. **201**, 8–29 (2007)
24. Sjögreen, B.: Lecture notes. http://www.math.fsu.edu/~sussman/Bjorn_Sjogreen_Notes.pdf
25. B. Swartz, B. Wendroff, The relative efficiency of finite difference and finite element methods. I: hyperbolic problems and splines, SIAM J. Numer. Anal. **11**, 979–993 (1974)

26. P.K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws. SIAM J. Numer. Anal. **21**, 995–1011 (1984)
27. E.F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*, 3rd edn. (Springer, New York, 2009)
28. B. Van Leer, Towards the ultimate conservative difference scheme, V. A second order sequel to Godunov's method. J. Comput. Phys. **32**, 101–136 (1979)
29. J.G. Verwer, Contractivity in locally one-dimensional splitting methods. Numer. Math. **44**, 247–259 (1984)
30. J.G. Verwer, J.M. Sanz-Serna, Convergence of method of lines approximations to partial differential equations. Computing **33**, 297–313 (1984)
31. R. Vichnevetsky, J.B. Bowles, Fourier analysis of numerical approximations of hyperbolic equations. SIAM J. Applied Math. **5** (1982)
32. Zaide, D.W.-M.: Numerical shock-wave anomalies. Ph.D. thesis, University of Michigan (2012)

http://www.springer.com/978-3-319-24784-7

Error Estimates for Well-Balanced Schemes on Simple Balance Laws
One-Dimensional Position-Dependent Models
Amadori, D.; Gosse, L.
2015, XV, 110 p. 24 illus., 15 illus. in color., Softcover
ISBN: 978-3-319-24784-7