

Surprising computations

Uri M. Ascher *

October 20, 2009

Abstract

In the course of simulation of differential equations, especially of marginally stable differential problems using marginally stable numerical methods, one occasionally comes across a correct computation that yields surprising, or unexpected results. We examine several instances of such computations. These include (i) solving Hamiltonian ODE systems using almost conservative explicit Runge-Kutta methods, (ii) applying splitting methods for the nonlinear Schrödinger equation, and (iii) applying strong stability preserving Runge-Kutta methods in conjunction with weighted essentially non-oscillatory semi-discretizations for nonlinear conservation laws with discontinuous solutions.

For each problem and method class we present a simple numerical example that yields results that in our experience many active researchers are finding unexpected and unintuitive. Each numerical example is then followed by an explanation and a resolution of the practical problem.

1 Introduction

The simulation of differential equations often requires complicated numerical methods. The resulting computations, even when long and complex, usually produce expected results, at least qualitatively. Such is the case, for instance, when applying “reasonable” methods for integrating parabolic PDEs, and using related procedures for solving convex optimization and numerical linear algebra problems. Of course, emphasis on efficiency and robustness in itself does not diminish the importance of corresponding numerical methods, and their careful design, analysis and implementation are crucial tasks.

Occasionally, however, one comes across a (correct, bug free) computation that yields surprising results. This may be the case when using marginally stable methods for solving marginally stable differential problems. The present paper examines several instances of such computations. These include (i) solving Hamiltonian ODE systems using almost conservative explicit Runge-Kutta (ERK) methods, (ii) applying splitting methods for the nonlinear Schrödinger (NLS) equation, and (iii) applying strong stability preserving (SSP) ERK methods in conjunction with weighted essentially non-oscillatory (WENO) semi-discretization methods for nonlinear conservation laws with discontinuous solutions. In general, the examples used to illustrate the methods and concepts examined were specifically chosen to be simple rather than complex.

*Department of Computer Science, University of British Columbia, Vancouver, Canada (ascher@cs.ubc.ca), supported in part under NSERC Discovery Grant 84306.

What can be qualified as “surprising” is of course a subjective matter. Indeed, we argue that in marginally stable situations this can occasionally be a function of trend, itself a function of human chronology, which may exhibit a swinging, pendulum-like behavior. For instance, symplectic and other symmetric methods are currently in vogue. They have been noted for their superior performance, especially for purposes involving long time integration; see, e.g., [16, 23, 2] and the many references therein. Recall that Hamiltonian systems describe the motion of frictionless, energy conserving mechanical systems. Thus, they possess *marginal* stability, which corresponding symplectic numerical schemes mimic. This “living at the edge of stability” is enabled, at least for sufficiently small (but possibly many) time steps, by the implied geometric structure that such discrete schemes conserve. On the other hand, it has long been known that conservative discretization schemes for nonlinear, nondissipative PDEs governing wave phenomena tend to become numerically unstable and exhibit other undesirable phenomena (e.g., when handling boundary conditions), hence numerical dissipation has subsequently been routinely introduced into such numerical schemes. See, e.g., [30, 15, 34, 2], which describe the seminal work of Kreiss [21] and much more. For nonlinear problems of this type, in particular, conservative difference schemes are known to occasionally yield numerical solutions which at first look fine, but at a later time may suddenly deteriorate and even explode; see for instance [3]. Consequently, until the 1980s non-dissipative schemes were discouraged, especially for long time integration. Typical work on pseudospectra, e.g., [35], when applied to stability studies of ODEs, also must assume that eigenvalues are placed off the imaginary axis and into the left half plane, so that sufficiently small circles of stability can be drawn around them: in the context of Hamiltonian systems this corresponds to using a slightly dissipative discretization scheme.

Each of the following three sections presents a numerical example that we believe to be novel, using numerical methods and demonstrating numerical phenomena that are not in themselves new but that, our experience indicates, many active researchers are finding unexpected and unintuitive. Each numerical example is then followed by an explanation and a resolution of the practical problem. Necessarily, the relevant bibliography list will be far from complete.

2 Integrating Hamiltonian systems using ode45

Surprisingly poor results can be obtained when applying the current version (numbered 7.8 and earlier) of MATLAB’s initial value ODE integrator `ode45` with default tolerances to certain Hamiltonian systems. R. McLachlan (private communication) has noticed this for the Henon-Heiles (HeHe) problem [27], where a phase plane plot that is very different from the correct one is obtained. Here we concentrate on another instance.

Example 1 A modification of the notorious *Fermi-Pasta-Ulam* (FPU) problem is presented in the introductory chapter of [16]. It consists of a chain of \hat{n} mass points connected with springs that have alternating characteristics: the odd ones are soft and nonlinear whereas the even ones are stiff and linear.

There are variables $q_1, \dots, q_{2\hat{n}}$ and $p_1, \dots, p_{2\hat{n}}$ in which the associated Hamiltonian is

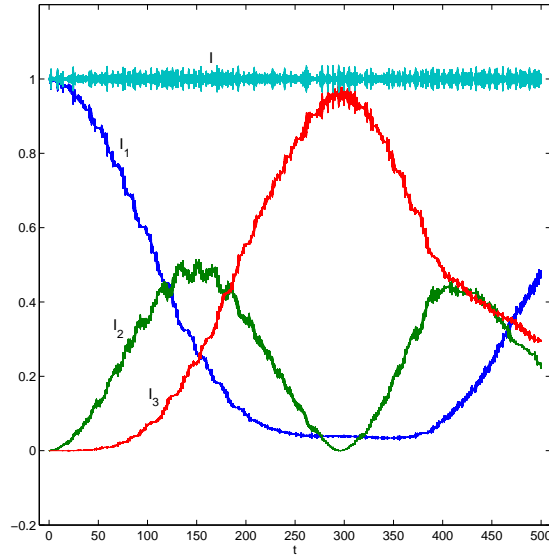


Figure 1: Oscillatory energies for the Fermi-Pasta-Ulam (FPU) problem.

written as

$$\begin{aligned}
 H(\mathbf{q}, \mathbf{p}) = & \frac{1}{4} \left[2 \sum_{i=1}^{2\hat{m}} p_i^2 + 2\omega^2 \sum_{i=1}^{\hat{m}} q_{\hat{m}+i}^2 + (q_1 - q_{\hat{m}+1})^4 + (q_{\hat{m}} + q_{2\hat{m}})^4 \right. \\
 & \left. + \sum_{i=1}^{\hat{m}-1} (q_{i+1} - q_{\hat{m}+1+i} - q_i - q_{\hat{m}+i})^4 \right]. \quad (1a)
 \end{aligned}$$

The parameter ω relates to the stiff spring constant and is large. This Hamiltonian is conserved as usual by the solution of the corresponding Hamiltonian system. In addition, denote the energy in the i th stiff spring by

$$I_i = \frac{1}{2}(p_{\hat{m}+i}^2 + \omega^2 q_{\hat{m}+i}^2). \quad (1b)$$

Then it turns out that there is an exchange of energies such that the total oscillatory energy

$$I = \sum_{i=1}^{\hat{m}} I_i,$$

is an *adiabatic invariant*, satisfying

$$I(\mathbf{q}(t), \mathbf{p}(t)) = I(\mathbf{q}(0), \mathbf{p}(0)) + O(\omega^{-1}) \quad (1c)$$

for exponentially long times t .

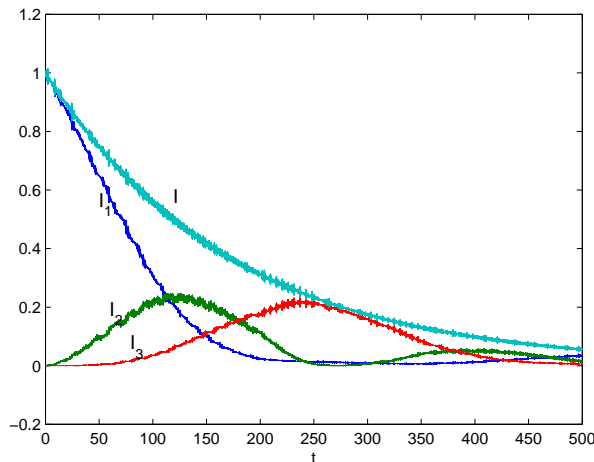


Figure 2: Oscillatory energies for the FPU problem obtained using MATLAB’s `ode45` with default tolerances. The deviation from Fig. 1 depicts a significant, non-random error.

As a variation of the example in [16] we choose $\hat{m} = 3$ (yielding an ODE system of size $m = 12$), $\omega = 100$, $\mathbf{q}(0) = (1, 0, 0, \omega^{-1}, 0, 0)^T$, $\mathbf{p}(0) = (1, 0, 0, 1, 0, 0)^T$, and integrate from $t = 0$ to $t = 500$ using the classical storage-efficient, 4th order, 4-stage ERK method, denoted RK4, with a constant step size $k = .00025$. The resulting Hamiltonian error is a mere 6.8×10^{-6} , and the oscillatory energies are recorded in Fig. 1. The curves depicted in the figure are exact as far as the eye can tell. The “noise” is not a numerical artifact. Rather, small “bubbles” of rapid oscillations occasionally flare up and quickly die away; see [16].

Next, we integrate this ODE system using `ode45` with default tolerances (which are relative tol = 10^{-3} , absolute tol = 10^{-6}). The result, depicted in Fig. 2, is a disaster. \square

The error control mechanism in `ode45`, like in all fast ODE packages, is based on local rather than global error estimates. Therefore, the above findings do not contradict any claim regarding the guaranteed reliability of this software. Nonetheless, the default error tolerances in `ode45` were undoubtedly set based on experiments that indicated that they typically work well, and the examples mentioned above are not of a freak, or highly unusual, nature (as they are, e.g., in [22, 1]). Rather, there seems to be a family of practical problems here where this software with default settings does not perform well.

The time integration method used in `ode45` is the Dormand-Prince pair of orders 4 and 5, see [10, 4]. Denoting these ERK formulas by DP4 and DP5, respectively, it is the result of the 6-stage DP5 that gets propagated from one time step to the next. This method is neither symmetric nor symplectic, so one could jump to the conclusion that the phenomenon illustrated in Example 1 (and, less fully, also in Example 6.5 of [2]) is related to that lack of structure preservation.

The crucial fact regarding Hamiltonian systems that we focus on below is that the eigen-

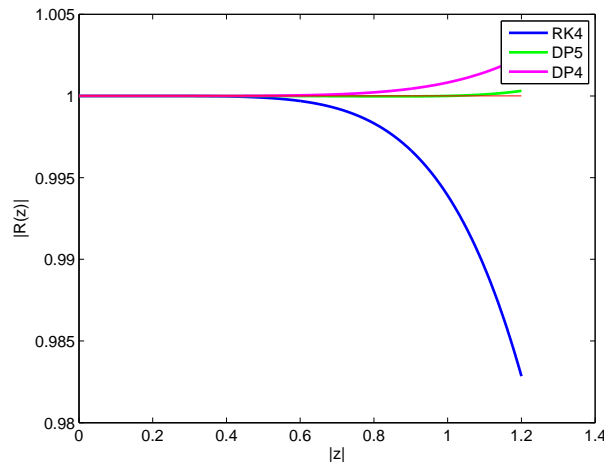


Figure 3: RK4, DP5 and DP4 amplification factors along the imaginary axis (i.e., the independent variable $z = k\lambda$ is purely imaginary).

values of the resulting Jacobian matrix, appropriately frozen, are purely imaginary. Modes of the form $e^{\lambda t}$ neither grow nor decay when λ is purely imaginary, and this gives rise to the interest in long time integration (because a quick steady state is often not a natural conclusion for the differential problem) as well as the danger that some uncontrollable perturbation would push these eigenvalues into the unstable right half-plane during a numerical calculation. Dissipative methods proposed long ago attenuate these eigenvalues towards the left half plane to ensure stability (see, e.g., Chapter 5 of [2]), at the sacrifice that the solution itself be eventually unnaturally attenuated. On the other hand, the philosophy of conservative methods, such as symplectic methods, is to not allow unnatural attenuation at all. The attenuation visible in the total energy I displayed in Fig. 2 may suggest that DP5 should not be used for this problem. We investigate this further below.

Consider the test equation

$$\frac{du}{dt} = \lambda u,$$

with λ a complex scalar, and denote $z = k\lambda$ for a numerical discretization with step size $k = \Delta t$ which can be written as

$$u_{n+1} = R(z)u_n, \quad n = 0, 1, \dots$$

It is well-known that the forward Euler method is unstable along the imaginary axis of z , i.e., $|R(z)| > 1$, unless $\lambda = 0$. Moreover, the backward Euler method is highly damping along the imaginary axis, i.e., $|R(z)|$ is significantly smaller than 1 when z is not very small. Both these methods therefore perform rather poorly for Hamiltonian systems.

The same cannot be said about the classical RK4 method. It contains a segment of the imaginary axis in its stability region and is only mildly dissipative; specifically, $|R(z)|$ is only

Problem	Method	Steps	Result good?
HeHe	ode45 def.	5,961	No
	ode45 10^{-7}	22,001	Yes
	ode45 10^{-5}	8,737	Yes
	ode45 10^{-4}	5,505	Barely
	RK4	20,000	Yes
	RK4	10,000	No
	DP5	10,000	Yes
FPU	ode45 def.	112,085	No
	ode45 10^{-4}	156,697	No
	ode45 10^{-5}	253,369	No
	ode45 10^{-6}	402,045	Yes
	RK4	1,000,000	Yes
	RK4	500,000	Barely
	RK4	200,000	No
	DP5	500,000	Yes
	DP5	200,000	Barely
	DP5	100,000	No
	Verlet	200,000	Yes
	Verlet	100,000	Barely
	Verlet	50,000	No

Table 1: Taking more time steps for the Henon-Heiles (HeHe) and Fermi-Pasta-Ulam (FPU) problems fixes the plots qualitatively. For the MATLAB function `ode45`, “def.” denotes using the default tolerances while, e.g., “ 10^{-5} ” denotes both absolute and relative tolerances set equal to 10^{-5} .

a little less than 1 for $|\mathcal{I}mz| \leq 1.2$, say. This method is therefore well-suited for integrating hyperbolic-type PDEs with smooth solutions, enjoying heavy use in practice even though it is non-conservative.

The above discussion brings up a question regarding the behavior of the DP formulae along the imaginary axis. Fig. 3 depicts the relevant amplification factors. We can see that DP5 in particular behaves very well for $\mathcal{I}mz \leq 1.1$, say. Assuming that the error control mechanism keeps z in the stable region (which Fig. 3 strongly suggests is indeed the case in view of the stability behavior of DP4), there is a rather small artificial dissipation with this method. The poor simulation depicted in Fig. 2 is not due just to the lack of symplecticity of the DP pair! Rather, it appears that the default step size selection tolerances of `ode45` are simply too permissive.

In Table 1 we list the results of applying `ode45` with the indicated value for both absolute and relative tolerances, as well as applying RK4, DP5 and the symplectic Verlet method with a fixed number of steps using a constant step size. Listed are the total number of steps, as well as an indication whether the resulting plot is “qualitatively good”, meaning it looks like Fig. 1 as far as the eye can tell, or not.

For the FPU problem, using constant step size, the “failures” of RK4 and DP5 yield

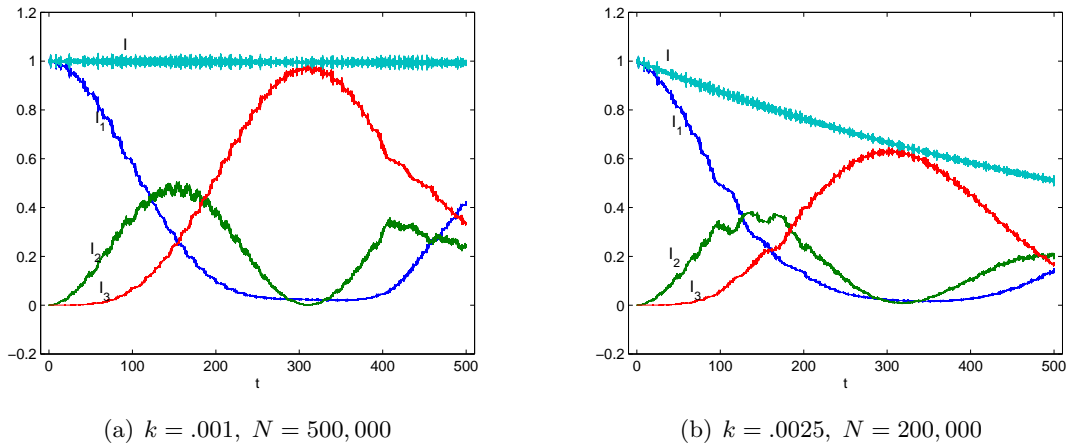


Figure 4: Integrating FPU using RK4 with a constant step size: (a) barely OK; (b) qualitatively wrong.

behavior along the pattern of Fig. 2, see Figs. 4 and 5, although the decay in I is not quite as pronounced. In contrast, the Verlet failure, depicted in Fig. 6, looks like an instability (oscillation) in the energy derivatives. As is the case for the Henon-Heiles problem, the failure especially using DP5 comes for a step size that is only about twice larger than that which causes failure of a different shape when employing Verlet's method.

It is natural to speculate on the reasons why the default tolerance values in `ode45` have been set to be so permissive as to allow the spectacular failures in the examples we have mentioned: obviously this cannot happen too often in general practice, or else the software designers would have adjusted by now.

A simple explanation is that most problems do contain a degree of dissipation or damping (assuming they are stable). The global error in any form or measure at a fixed time $t = T$ is a sum of local errors (whose formal accuracy order is one higher) propagated along the solution modes from the times where they occurred. If there is damping, therefore, then the local (or truncation) errors propagating from afar have essentially evanesced by the time they arrive at T , so the global error is essentially proportional to a sum of the local errors nearby T only. Note that $R(z) \approx e^z$ for approximation (consistency) reasons. On the other hand, for a Hamiltonian system there is no damping of any of the local errors, and the global error is therefore proportional to the sum of all local errors. The global error is therefore significantly larger than a local error, especially after many time steps, and since `ode45` controls only local errors a much larger accumulating error than what is controlled can be obtained.

For DP5 and even more so RK4, there is some numerical dissipativity, and the very small but many local error contributions thus add up to form an approximation to a perturbed ODE problem with dissipation (damping). This does not occur for the symplectic Verlet method.

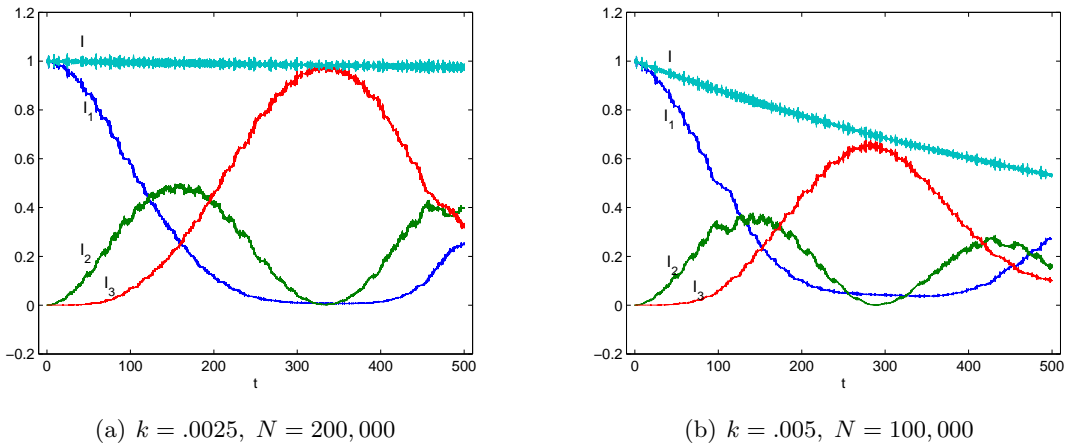


Figure 5: Integrating FPU using DP5 with a constant step size: (a) barely OK; (b) qualitatively wrong.

3 The nonlinear Schrödinger equation in 1D

The cubic nonlinear Schrödinger equation (NLS) in one space variable can be written as

$$\psi_t = \imath(\psi_{xx} + \kappa|\psi|^2\psi). \quad (2)$$

This equation arises in deep water wave modulation, in Bose-Einstein condensate theory and in nonlinear fiber optics (see, e.g., the corresponding [wikipedia](#) entry).

We know that for the pure initial value problem the solution's \mathcal{L}_2 -norm remains constant for all time, $\int \bar{\psi}(t, x)\psi(t, x)dx \equiv \|\psi(t)\|^2 = \|\psi(0)\|^2$. Moreover, this is a Hamiltonian PDE, which means that it can be written as

$$\begin{aligned} \mathbf{u}_t &= \mathcal{D} \left(\frac{\delta \mathcal{H}}{\delta \mathbf{u}} \right), \quad \text{where} \\ \mathcal{H}[\mathbf{u}] &= \int H(x, \mathbf{u}, \mathbf{u}_x, \dots) dx, \\ \int \frac{\delta \mathcal{H}}{\delta \mathbf{u}} \mathbf{v} dx &= \left(\frac{d}{d\varepsilon} \mathcal{H}[\mathbf{u} + \varepsilon \mathbf{v}] \right)_{\varepsilon=0}, \end{aligned} \quad (3)$$

with

$$\mathbf{u} = (\psi, \bar{\psi})^T, \quad \mathcal{D} = \imath, \quad H(\psi, \bar{\psi}) = \psi_x \bar{\psi}_x - \frac{\kappa}{2} \psi^2 \bar{\psi}^2.$$

The Hamiltonian \mathcal{H} is conserved in time. There is also a multisymplectic structure here [8, 6, 7].

Note that H is positive if $\kappa \leq 0$. In the case $\kappa > 0$ the solution is known to possibly exhibit instabilities, see [37] and references therein, but these are not what we see in the examples below. Explicit soliton solutions are provided in [29] for the case $\kappa > 0$.

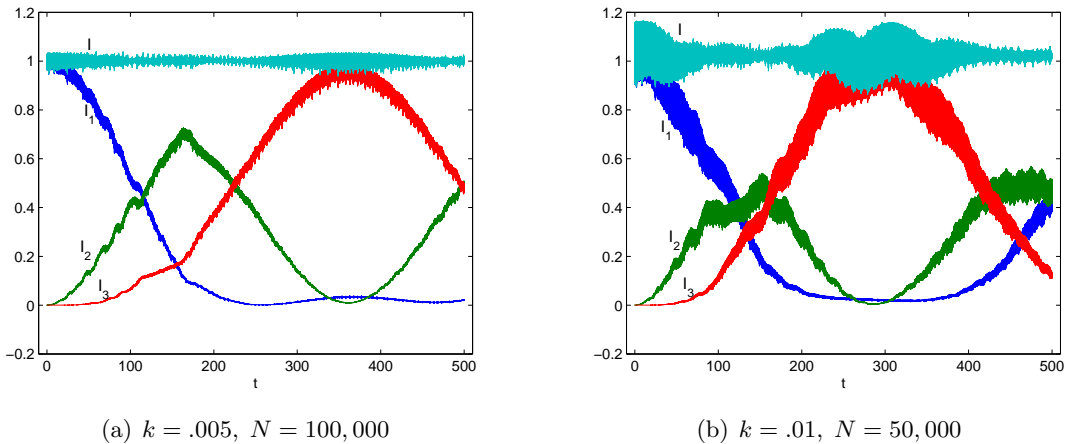


Figure 6: Integrating FPU using the symplectic Verlet method with a constant step size: (a) (perhaps not) barely OK; (b) qualitatively wrong.

Often, the scaled form

$$i\varepsilon\psi_t = -\frac{\varepsilon^2}{2}\psi_{xx} - \kappa|\psi|^2\psi, \quad x \in \mathbb{R}, t \geq 0, \quad (4)$$

with smooth prescribed initial values is considered; see for instance [26]. Of course, the constant ε is small, $0 < \varepsilon \ll 1$. Considering instead periodic BC on, say, $[-\pi, \pi]$, and limiting the time range to $0 \leq t \leq T$, where T is of moderate size, we can make the stretching transformation

$$\hat{t} = \frac{1}{\varepsilon}t, \quad \hat{x} = \frac{1}{\varepsilon}x.$$

Then $\psi_{\hat{t}} = \varepsilon\psi_t$, and in each component of x , $\psi_{\hat{x}\hat{x}} = \varepsilon^2\psi_{xx}$. Hence we have (2) in the stretched coordinates, for $0 \leq \hat{t} \leq T/\varepsilon$, $-\pi/\varepsilon \leq \hat{x} \leq \pi/\varepsilon$. Thus, (4) corresponds to our problem (2) on a large domain in both space and time.

For the numerical solution we consider some well-known splitting methods, where the right hand side of (2) is split in an obvious way into its two terms. Although there are many other methods, e.g., [9], and most numerical difficulties in practice may arise in the context of more space variables, our intention here is to examine what can happen even for well-justified and well-tested methods in a relatively simple setting. Note that the problem

$$u_t = wu_{xx} \quad (5)$$

is linear with constant coefficients, and it can be efficiently discretized in various ways to be specified below. The nonlinear part

$$w_t = i|w|^2w,$$

is really an ODE with x as a parameter, whose exact solution is

$$w(t) = w(t_0) e^{\imath(t-t_0)|w|^2},$$

with $|w|$ independent of t . Hence, stepping from t with any time step $\Delta t = k$ we have

$$w(t+k) = w(t) e^{\imath k|w|^2}.$$

So, we use the Strang splitting to compose the two resulting solution operators in a staggered, standard way to obtain an approximation for $\psi(t, x)$; see, e.g., [2].

We consider three splitting methods, depending on the discretization of (5). They are specified as follows:

1. A symmetric, compact finite difference semi-discretization of (2), using the usual D_+D_- operator with a uniform step size $\Delta x = h$ in space, yields a Hamiltonian ODE system in time. For (5) we subsequently use the (symplectic) midpoint method in time. Thus, denoting the semi-discretization for u by

$$\mathbf{v}' = \imath \Delta_h \mathbf{v},$$

we apply

$$\frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{k} = \frac{\imath}{2} [\Delta_h \mathbf{v}^{n+1} + \Delta_h \mathbf{v}^n]. \quad (6a)$$

See [37, 19, 2]. Since this method is symplectic, norm-preserving and 2nd order accurate in both t and x , so is the ensuing staggered composition with the exact solver of the ODE part of the splitting.

2. The same three-point centered scheme is used in space, and a slightly attenuated version of the midpoint method is applied in time, reading

$$\frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{k} = \frac{\imath}{2} [(1 + \varepsilon) \Delta_h \mathbf{v}^{n+1} + (1 - \varepsilon) \Delta_h \mathbf{v}^n]. \quad (6b)$$

We choose $\varepsilon = ch^2$, and set $c = 1$ in the experiments below. This method then retains 2nd order accuracy in time and space.

3. Replace the finite difference method for $u_t = uu_{xx}$ by a spectral method in both space and time. Thus, the solution to the subproblem (5) is given by

$$u(t+k) = \mathcal{F}^{-1} \left(e^{-\imath \xi^2 k} \mathcal{F}(u(t)) \right). \quad (6c)$$

This is discretized in the standard way with $u(t) \equiv \mathbf{v}^n = (v_1^n, \dots, v_j^n)$, $v_j^n \approx u(jh, kn)$, $u(t+k) \equiv \mathbf{v}^{n+1}$, and \mathcal{F} the fast Fourier transform.

The resulting method is popular in practice, see [37, 11, 25, 12] and references therein.

Example 2 To see the resulting methods in action we used an example from [19], where periodic BC were specified on the interval $[-20, 80]$, and the initial value function was

$$\psi(0, x) = e^{ix/2} \operatorname{sech}(x/\sqrt{2}) + e^{i(x-25)/20} \operatorname{sech}((x-25)/\sqrt{2}).$$

This yields two pulses for $|\psi|$ which propagate to the right at different speeds, with their shapes unchanged except when they coalesce.

We ran the methods for various values of k and h ($k = \Delta t$ and $h = \Delta x$), and at $t_f = 200$ and $t_f = 1000$ computed the *relative difference* in the discrete Hamiltonian and ℓ_2 -norm from the values at $t = 0$. Results are collected in Table 2.

t_f	method	k	h	Error-Ham	Error-norm
200	(6a)	.1	.1	3.7e-5	4.3e-13
	(6c)	.1	.1	1.5e-2	1.2e-13
	(6a)	.01	.01	3.9e-9	1.5e-11
	(6c)	.01	.01	4.3e-6	1.0e-12
1000	(6a)	.1	.1	5.2e+2	2.9e-12
	(6c)	.1	.1	8.7e+2	6.1e-13
	(6a)	.01	.1	3.3e-7	4.2e-12
	(6c)	.01	.1	2.0e+3	2.3e-12
	(6c)	.0025	.1	9.1e-8	5.6e-12
	(6a)	.01	.01	3.4e+2	7.7e-11
	(6b)	.01	.01	8.9e-4	7.3e-5
	(6c)	.01	.01	1.2e+5	8.3e-13
	(6a)	.005	.005	1.0e+2	3.0e-10
	(6c)	.005	.005	1.5e+5	1.7e-12

Table 2: Relative error indicators in Hamiltonian and ℓ_2 -norm for Example 2 measured at $t = 200$ and $t = 1000$.

The error indicators using all three methods (6a)–(6c) are pleasantly small at $t = 200$. The solution profile is plotted in Fig. 7(a).

But continuing on to $t_f = 1000$, using method (6a) the results were not acceptable, see Fig. 7(b). Another run, using $h = .1$ and $k = .02$, also leads to visible instability before t hits 1000. The bad effect, which is an instability in the derivative of ψ , disappeared upon using $k = h^2$ for $h = .1$.

Employing the spectral method (6c), the error indicators are again pleasantly small at $t = 200$. But continuing on to $t_f = 1000$, the results are in fact similar to and even worse than those for the finite difference scheme. Note that, keeping $h = .1$ fixed, using $k = .01$ still results in an instability here, and only a smaller value of $k = .0025$ yields decent results. See plots in Fig. 8.

If we flip to $\kappa = -1$, so that \mathcal{H} in (3) is a norm, then the solution no longer consists of a couple of moving solitons and has a wilder, varying shape. It takes longer for the same

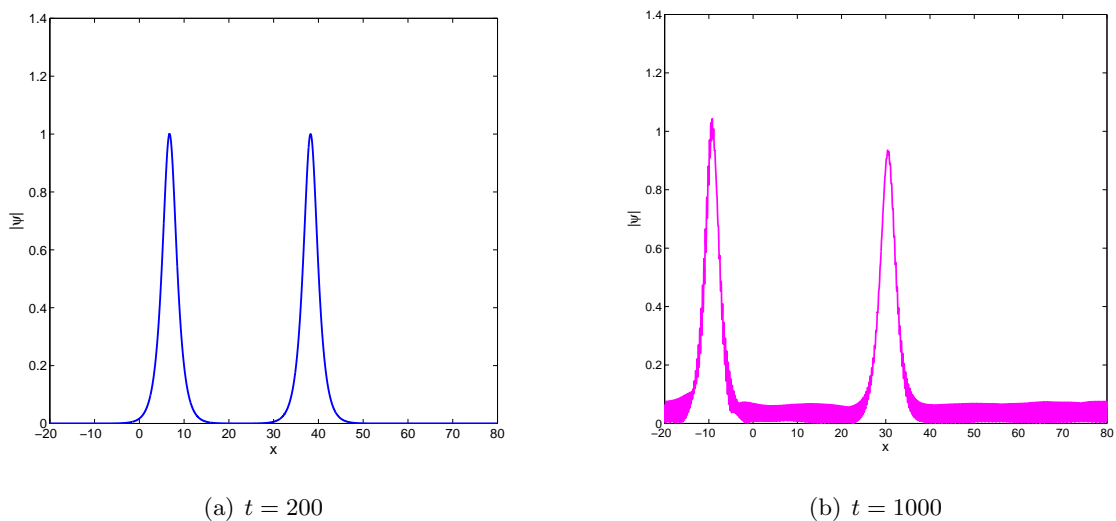


Figure 7: Solution magnitude for the Schrödinger equation (2) using a splitting symplectic 2nd order finite difference method (6a) with $k = h = .01$. The two pulses look accurate at $t = 200$. But as integration proceeds an instability in the solution derivative arises, yielding sharp oscillations that in the figure look like a thick line. See Table 2.

sort of numerical instability to set in, but the phenomenon is similar: at $t_f = 10000$ we have for $k = h = .1$, Error-Ham = $3.1e+2$; for $k = h = .01$, Error-Ham = $1.0e+4$; and for $k = .01$, $h = .1$, Error-ham = $2.9e-4$. \square

What gives rise to the poor results when using the symplectic method (6a) is the fact that the imaginary eigenvalues of the semi-discretization are $\nu O(h^{-2})$. So, when $k = O(h)$ we have in the terminology of Section 2 a very large imaginary $z = k\lambda$ stability argument. This can cause trouble (cf. [5]) in the long run in case there are unfortunate, even though small, perturbations to these large imaginary eigenvalues. Such perturbations are provided by the splitting scheme. (For this particular example, at least, the same midpoint scheme without splitting is found to be more stable.)

The error indicators in Table 2 are underestimators for the general solution error. The error in the Hamiltonian proves to be a good indicator, as it fowls up (by becoming large) where the instability sets in. In contrast, the error in the ℓ_2 -norm of the discrete solution, except for the one obtained when using (6b), is very small whether the computed solution is good or not. This should serve as a sobering example regarding “energy conserving” methods, suggesting that preserving such one property does not yield an automatic guarantee of a successful simulation.

The same symptom is seen for the spectral method. The *splitting* nature of the scheme is what provides perturbation to this conservative method for a marginally stable problem, so a better approximation to one of the split operators, which is what (6c) presumably provides for a sufficiently small h , is no guarantee against an unfavorable accumulation of

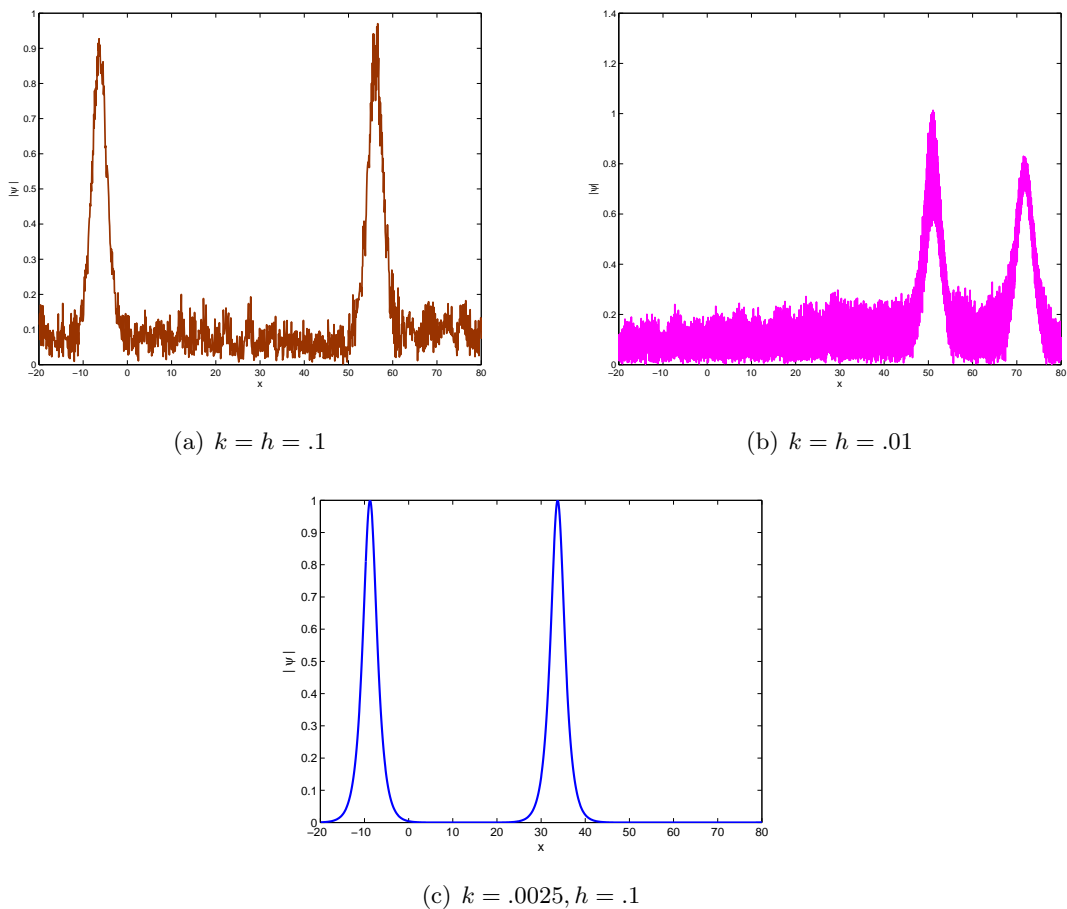


Figure 8: Solution magnitude for the Schrödinger equation (2) using a splitting spectral method (6c) at $t = 1000$. An instability develops for $k = h$ in the solution derivative, yielding sharp oscillations. See Table 2. The plot for $k = .0025$ is acceptably accurate.

roundoff errors.

In [37] there is a detailed analysis of instabilities, based on linearization around a uniform wave train, for both methods (6a) and (6c) experimented with here. First, there are instabilities in the problem itself that are mirrored, generally speaking, by the numerical methods. But the phenomenon discussed here concerns purely numerical, high frequency instabilities. For the case $\kappa > 0$ and using the spectral splitting method (6c), Weideman and Herbst derive in [37] the stability condition

$$k < \frac{h^2}{\pi}. \quad (7)$$

This condition agrees well with our experiments. But for (6a), there are no theoretical bounds in [37] corresponding to what we have observed.

Let us concentrate on the case $k = h$, $t_f = 1000$. Our explanation for the long time instability for the finite difference scheme centers around the fact that the eigenvalues of

the semi-discretization for $u_t = uu_{xx}$ are large and imaginary, so using the implicit midpoint method with $k = h$ puts us in the highly oscillatory regime for this marginally stable method as in [5]. The splitting leads to small perturbations of those imaginary eigenvalues that may push them slightly to the right half plane, so errors may accumulate in an unfavorable way. The scheme (6b) differs only little from (6a), but its attenuation improves upon the accumulation of such errors. Indeed, the results using (6b) with $k = h = .01$ are pleasing, see Table 2, even though the ℓ_2 -norm of the solution is no longer preserved to a hyper-accuracy level. The corresponding solution plot at $t = 1000$ does not differ qualitatively from Fig. 8(c). Our point here is not to promote the method (6b) as a general tool, but rather to indicate that the surprising effects depicted in Figs. 7(b) and 8(a,b) are the result of sticking with a conservative scheme to the (bitter) end. The corresponding error indicators using (6b) for $k = h = .1$ were Error-Ham = 3.0e-1 and Error-norm = 5.1e-2, which is too coarse for comfort.

We emphasize that our numerical example depicts a potential instability that becomes a practical problem only at “long times”. There may well be applications where using either one of the methods (6a) or (6c) with $k = h$ produces fine results for all intents and purposes. Hopefully, the computable error in the Hamiltonian can serve as an indicator for the potential onset of trouble for problems where the expected shape of the solution is not known in advance.

A limitation such as (7) is comparable to what we have for explicit methods, and indeed using the explicit RK4 or DP5 for the symmetrically discretized unsplit problem (2) is an alternative option under such conditions. However, the phenomenon depicted here is not that of an explicit scheme. The source is perturbations (due to the splitting) to the midpoint method in the highly oscillatory regime, and this is why the instability shows up so late in the game, and also, why introducing a slight attenuation in (6b) “fixes” it. Moreover, the implicit midpoint method provides an approximation to the matrix exponential that in the spectral method is obtained “explicitly” via the transform. Finally, if this was a simple instability for an explicit method then the results for $k = h$ would not have been acceptable at $t = 200$ and at earlier times either.

4 SSP methods

Strong stability preserving (SSP) methods were first developed in the late 1980s [32]. But they really caught fire only about a decade later; see [14] for a relatively early review and [13] for another, more general and more recent one with many relevant references.

The original development was in the context of essentially non-oscillatory (ENO) methods [17]. Consider for simplicity the scalar conservation law

$$\begin{aligned} u_t + f(u)_x &= 0, & -\infty < x < \infty, & t \geq 0, \\ u(0, x) &= u_0(x). \end{aligned} \tag{8}$$

It is well-known that discontinuities may develop in the solution $u(t, x)$ for $t > 0$ even if $u_0(x)$ is smooth. Now, the usual upwind discretization for (8), which on a uniform mesh at

$t_n = kn$ and $x_j = jh$ reads

$$v_j^{n+1} = v_j^n - \frac{k}{h} \begin{cases} [f(v_{j+1}^n) - f(v_j^n)] & \text{if } \frac{df}{du}(v_j^n) < 0 \\ [f(v_j^n) - f(v_{j-1}^n)] & \text{if } \frac{df}{du}(v_j^n) \geq 0 \end{cases}, \quad (9)$$

with $v_j^0 = u_0(x_j)$, can be viewed at time level t_n as a location-dependent semi-discretization in space followed by forward Euler in time. Further, ENO is a set of sophisticated higher order spatial semi-discretizations replacing the simple upwinding in (9). Subsequently, SSP methods are correspondingly higher order time discretizations, which preserve the non-oscillatory nature of the solution in the presence of discontinuities provided that forward Euler does so. Thus, SSP methods have generally been perceived as more accurate generalizations of the forward Euler method.

Much work was carried out in constructing such methods and in providing extensive nonlinear theory, see [13, 33, 20, 18] and references therein. There is general agreement in the relevant community that the SSP concept has yielded a bullet-proof class of time discretization methods in conjunction with ENO, even though examples where the SSP property is actually essential for performance are uncommon and even though spurious oscillations using ENO remain possible in principle.

Further, however, currently ENO methods are rarely used in practice. Instead, weighted ENO (WENO) semi-discretization methods are favored, see [31]. Thus, at each spatial mesh point a weighted combination of ENO stencils is employed, with the weights determined to maximize order of accuracy in regions where the solution is smooth. The typical setting is that from three ENO molecules of accuracy order 3 and spanning 4 mesh points each, one constructs a WENO method that has accuracy order 5, call it WENO5. Of course, the exact meaning of a high order of accuracy in the wake of a passing shock wave is subject to debate, but WENO also has other favorable properties and in any case apparently always majorizes ENO. We then ask, does the SSP property retain its meaning and importance also when the semi-discretization is WENO, rather than ENO? The present section concentrates on this question.

In the context of WENO, unlike that of ENO, there is no known theory to support SSP methods [13]. The essential reason for this lack of theory is highlighted in [36]. In fact, these authors found out both in terms of linear stability theory and in simple numerical examples that forward Euler does not do well when complementing the WENO5 semi-discretization. Thus, the method that SSP methods “want to be like” is nothing to aspire to in the WENO context; see also [2, 28].

The intuitive reason for this is as follows. In regions where the solution is smooth the WENO maximization of order makes it produce a semi-discretization that is close to being centered. Thus, the eigenvalues of the time-dependent ODE system tend to be near the imaginary axis, albeit at its stable side. But the forward Euler absolute stability region does not even get close to the imaginary axis unless the step size is rather small! Thus, unlike for typical higher order ERK discretizations (cf. Fig. 3), the forward Euler discretization produces local linear instabilities. These instabilities, for sufficiently small time steps, are automatically handled by WENO before they become too large. But then WENO is unnecessarily working on the imperfection of the time discretization scheme. The net effect is the necessity when working with forward Euler of occasionally taking much

smaller time steps than would otherwise be allowed; see also [28]. So, the SSP concept in its unmodified form may be irrelevant when a WENO semi-discretization is employed.

Example 3 The *Buckley-Leverett flux* [24] for water is given by

$$f(u) = \frac{u^2}{u^2 + a(1-u)^2}. \quad (10)$$

Set $a = .5$ and choose periodic boundary conditions on $[-1, 1]$.

The WENO5 semi-discretization as in [31, 36] was employed in space. In time, we looked at the following ERK methods, denoted (s,p) where s is the number of stages and p is the order:

1. Forward Euler (which is SSP(1,1)).
2. Explicit trapezoidal (which is SSP(2,2)).
3. The popular SSP(3,3) method of [32].
4. The “optimal” SSP(5,3) method of [33].
5. The classical RK4 (which is a storage-efficient non-SSP (4,4) method).
6. The storage-efficient SSP(10,4) method of [20], which is implemented within the software package *Clawpack*, see <http://www.amath.washington.edu/~claw>.

Only uniform meshes are considered, as before.

At first, consider the initial value profile

$$u_0(x) = .25 + .5 \sin(\pi x). \quad (11)$$

The qualitatively exact solution is depicted in Fig. 9(c). The SSP(2,2) result in Fig. 9(b) is not quite clean, but lowering the step size to $k = .0004$ (not shown) does provide a clean profile for all times $0 \leq t \leq 1.1$. The SSP(1,1) method in Fig. 9(a) is a disaster, although it does not blow up. For $h = .01$, $k = .001$, forward Euler also yields oscillatory results, but for $h = .01$, $k = .0001$, a qualitatively correct solution profile is recovered.

The SSP(3,3) method performs similarly to SSP(2,2) in terms of time step size “allowed” (meaning, still providing a qualitatively correct solution profile). The SSP(5,3) method allows for a step size that is less than 1.5 times as large as that of SSP(3,3), so it is slightly inferior for this example.

The SSP(10,4) method allows for a step size as large as $k = .0015$. Dividing by the number of function evaluations per time step, this is comparable to RK4.

So, for this example, we pick three winners and one loser. The winners are the workhorse RK4 and the 10-stage-4th-order low-storage SSP method. The SSP(2,2) method allows the largest time step per function evaluation, but of course it is only 2nd order accurate, and its stability region is more susceptible to perturbations near the imaginary axis. The loser is forward Euler: its allowed time step is much smaller than those of any of the other methods.

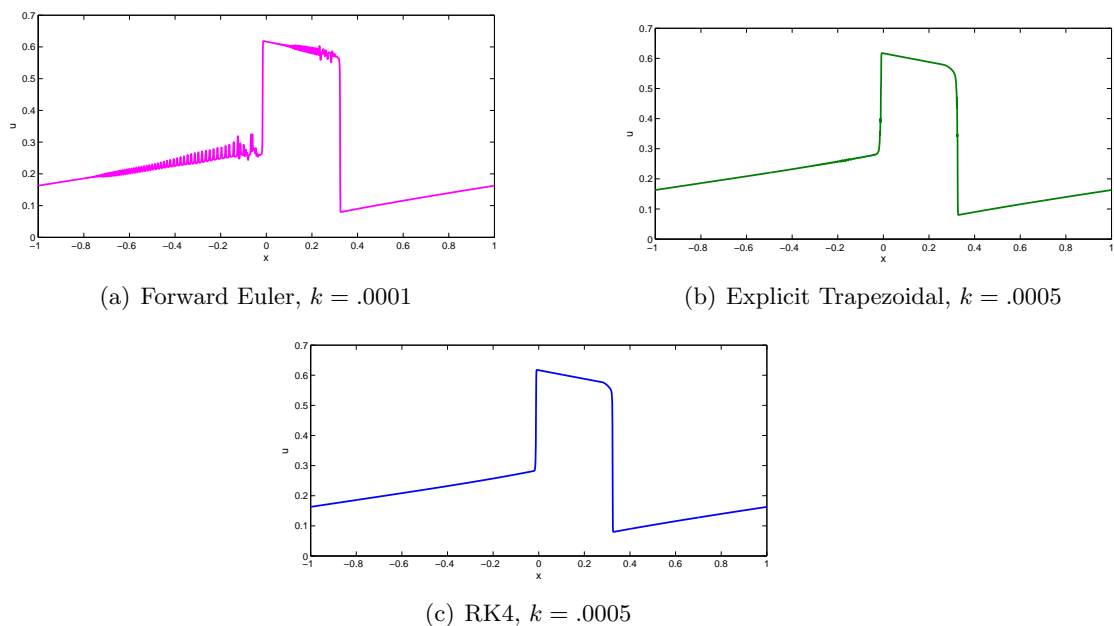


Figure 9: Solution profiles at $t = 1.1$ for the Buckley-Leverett conservation law with $a = .5$ and initial profile given by (11). The spatial step size in all cases is $h = .001$.

Next, consider the initial value profile

$$u_0(x) = \sin(10\pi x)e^{-10x^2}. \quad (12)$$

This yields several solution discontinuities.

Some results are depicted in Fig. 10. In all cases, $h = .001$, although corresponding results for $h = .01$ were run as well, yielding no further insight. With forward Euler a step size of $k = .0001$ is small enough to yield a qualitatively acceptable solution, unlike that in Fig. 10(a). The step sizes in Fig. 10(b) and Fig. 10(c) are close to the largest they are allowed to be for good quality solutions throughout $0 \leq t \leq 1$. These two 4th order methods again perform roughly similarly, while forward Euler for this case is not that much behind either (only by a factor of about 2, when counting function evaluations). \square

Not much can be concluded with certainty from one or two examples, although we have run some tests also for the Burgers equation. However, the inadequacy of forward Euler is already clear enough. We are then faced with the situation where some SSP methods do perform rather well with WENO, but others do not, and it is unclear whether the property of being SSP is an important, “defining” one, or it is just another property that happens to hold for some good methods.

The SSP(10,4) time discretization [20] is impressive by the mere fact that a 10-stage method can be competitive. But its bottom-line performance is not breathtakingly better. Such appears to be the overall impression regarding the use of SSP methods in the WENO context.

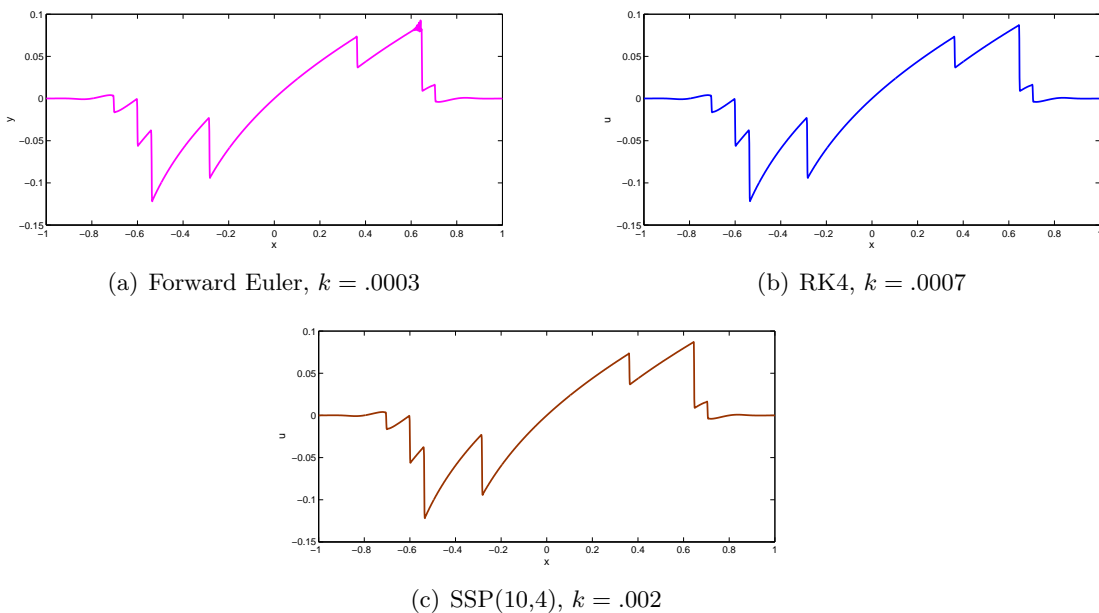


Figure 10: Solution profiles at $t = 1$ for the Buckley-Leverett conservation law with initial profile given by (12). The spatial step size in all cases is $h = .001$.

5 Conclusions and further comments

We have briefly considered three popular problem areas and corresponding popular numerical methods, and have shown that conventional wisdom may run into snags using such methods, even for rather simple examples.

Although the topics of Sections 2, 3 and 4 are rather different, there are clearly uniting themes in our observations and explanations. Essentially, we are advocating special awareness when treating marginally stable differential problems. For such problems it may be easier and relevant to prove theorems regarding conservation properties for numerical methods that attempt to reproduce important dynamical system and exact solution features. But this may also be the underlying cause for unwelcome surprises in practical computation. Ironically, some geometric integration methods are nowadays making their way into the toolboxes of computer graphics simulation experts, perhaps simply because such methods may look “different” in certain contexts. Such a generally positive development from the numerical analyst’s point of view thus also dictates maintaining a heightened sense of alert.

Acknowledgment I would like to thank Drs. David Ketcheson, Christian Lubich, Robert McLachlan and Steve Ruuth for fruitful discussions.

References

- [1] U. Ascher. On symmetric schemes and differential-algebraic equations. *SIAM J. Scient. Comput.*, 10:937–949, 1989.
- [2] U. Ascher. *Numerical Methods for Evolutionary Differential Equations*. SIAM, Philadelphia, PA, 2008.
- [3] U. Ascher and R. I. McLachlan. On symplectic and multisymplectic schemes for the KdV equation. *J. Sci. Computing*, 25:83–104, 2005.
- [4] U. Ascher and L. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, Philadelphia, PA, 1998.
- [5] U. Ascher and S. Reich. The midpoint scheme and variants for hamiltonian systems: advantages and pitfalls. *SIAM J. Scient. Comput.*, 21:1045–1065, 1999.
- [6] T. Bridges and F. Laine-Pearson. Multi-symplectic relative equilibria, multi-phase wavetrains and coupled nls equations. *Studies in Appl. Math.*, 107:137–155, 2001.
- [7] T. J. Bridges and S. Reich. Multi-symplectic integrators: numerical schemes for Hamiltonian PDEs that conserve symplecticity. *Phys. Lett. A*, 284 (4-5):184–193, 2001.
- [8] C. J. Budd and M. D. Piggott. Geometric integration and its applications. *Handbook of Numerical Analysis vol. XI*, pages 35–139, 2003. P. G. Ciarlet and F. Cucker (eds.), North-Holland, Amsterdam.
- [9] M. Dahlby and B. Owren. Plane wave stability of some conservative schemes for the cubic Schrödinger equation. *M2AN*, 43:677–687, 2009.
- [10] J. R. Dormand and P. J. Prince. A family of embedded runge-kutta formulae. *J. Comp. Appl. Math.*, 6:19–26, 1980.
- [11] B. Fornberg. *A Practical Guide to Pseudospectral Methods*. Cambridge Press, 1998.
- [12] L. Gauckler and C. Lubich. Splitting integrators for nonlinear Schrödinger equations over long times. *Found. Comp. Math.*, 2009. to appear.
- [13] S. Gottlieb, D. Ketcheson, and C.-W. Shu. High order strong stability preserving time discretizations. *J. Scient. Comput.*, 38:251, 2009.
- [14] S. Gottlieb, C.-W. Shu, and E. Tadmor. High order time discretization methods with the strong stability property. *SIAM Review*, 43:89–112, 2001.
- [15] B. Gustafsson, H.-O. Kreiss, and J. Olinger. *Time Dependent Problems and Difference Methods*. Wiley & Sons, New York, 1995.
- [16] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*. Springer, 2002.

- [17] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes, iii. *J. Comp. Phys.*, 71:231–303, 1987.
- [18] I. Higuera. On strong stability preserving time discretization methods. *J. Scient. Comput.*, 21:193–223, 2004.
- [19] W. Hundsdorfer and J. G. Verwer. *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer, 2003.
- [20] D. Ketcheson. Highly efficient strong stability preserving runge-kutta methods. *SIAM J. Scient. Comput.*, 30:2113–2136, 2008.
- [21] H.-O. Kreiss. On difference approximations of the dissipative type for hyperbolic differential equations. *Comm. Pure Applied Math.*, 17:335–353, 1964.
- [22] H.-O. Kreiss. Centered difference approximation to singular systems of ODEs. In *Symposia Mathematica X*. Inst. Nazionale di Alta Math, 1972.
- [23] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2004.
- [24] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge, 2002.
- [25] C. Lubich. On splitting methods for Schrödinger-Poisson and cubic nonlinear Schrödinger equations. *Math. Comp.*, 77:2141–2153, 2008.
- [26] P. Markowich, P. Pietra, and C. Pohl. Numerical approximation of quadratic observables of Schrödinger-type equations in the semi-classical limit. *Numerische Mathematik*, 81:595–630, 1999.
- [27] R. I. McLachlan and G. R. W. Quispel. Geometric integrators for odes. *J. Physics A*, 39:5251–5285, 2006.
- [28] M. Motamed, C. MacDonald, and S. Ruuth. On the linear stability of the fifth-order WENO discretization. 2009. manuscript.
- [29] A. Polyanin and V. Zaitsev. *Handbook of Nonlinear Partial Differential Equations*. Chapman & Hall/CRC, 2004.
- [30] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial-Value Problems*. Wiley, 1967.
- [31] C.-W. Shu. Essentially non-oscillatory and weighted essentially nonoscillatory schemes for hyperbolic conservation laws. In *Advances in Numerical Approximation of Nonlinear Hyperbolic Equations*, pages 325–432. Springer Lecture notes in Math 1697, 1998.
- [32] C.-W. Shu and S. Osher. Efficient implementation of of essentially non-oscillatory shock capturing schemes. *J. Comp. Phys.*, 77:439–471, 1988.
- [33] R. Spiteri and S. Ruuth. A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM J. Numer. Anal.*, 40:469–491, 2002.

- [34] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. SIAM, 2004. 2nd Edition.
- [35] L. N. Trefethen. Pseudospectra of linear operators. *SIAM Review*, 39:383–406, 1997.
- [36] R. Wang and R. Spiteri. Linear instability of the fifth order WENO method. *SIAM J. Numer. Anal.*, 45(5):1871–1901, 2007.
- [37] J.A.C. Weideman and B.M. Herbst. Split-step methods for the solution of the nonlinear Schrödinger equation. *SIAM J. Numer. Anal.*, 23:485–507, 1986.