SIAM J. Sci. Comput.
Vol. 22, No. 1, pp. 20–38

RUNGE–KUTTA SOLUTIONS OF A HYPERBOLIC CONSERVATION LAW WITH SOURCE TERM*

MARK A. AVES[†], DAVID F. GRIFFITHS[†], AND DESMOND J. HIGHAM[‡]

Abstract. Spurious long-term solutions of a finite-difference method for a hyperbolic conservation law with a general nonlinear source term are studied. Results are contrasted with those that have been established for nonlinear ordinary differential equations. Various types of spurious behavior are examined, including spatially uniform equilibria that exist for arbitrarily small time-steps, nonsmooth steady states with profiles that jump between fixed levels, and solutions with oscillations that arise from nonnormality and exist only in finite precision arithmetic. It appears that spurious behavior is associated in general with insufficient spatial resolution. The potential for curbing spuriosity by using adaptivity in space or time is also considered.

Key words. adaptivity, finite difference, nonnormality, spurious solution, steady state

AMS subject classifications. 65M06, 65M50

PII. S106482759833601X

1. Introduction.

1.1. Background.

There is a growing interest in the analysis of time-stepping methods for the long-term solution of nonlinear evolutionary equations. In particular, the propensity for widely used methods to generate spurious steady states is receiving much attention. In the context of ordinary differential equations (ODEs), many examples of spurious behavior have been identified [6, 12, 24]. Theoretical work in this area includes the study of spurious fixed points for small time-steps [10], the identification or construction of methods that guarantee to avoid spurious fixed points [7, 12], and the use of bifurcation theory to study routes to spuriosity [13, 20]. In [1] the potential for spurious behavior with an adaptive ODE algorithm was investigated, and both positive and negative results concerning the effectiveness of error control were derived.

Our aim in this work is to analyze spurious behavior arising from a discretized hyperbolic partial differential equation (PDE) of the form

$$u_t + (f(u))_x = g(u), \quad 0 < x < 1, \quad t > 0, \tag{1}$$

where the flux f and the source term g are nonlinear. Motivating this work is our desire to highlight new features that arise from the introduction of a spatial derivative, and hence to contrast the results with those that have been established for ODEs.

Difficulties arise in the numerical solution of (1) when the source term is “stiff”; that is, the time scale associated with reaction is very much shorter than that associated with advection. One of the most common problems addressed in the literature is that of incorrect speed of propagation of waves; see, for example, [3, 5, 15, 23]. This

*Received by the editors March 20, 1998; accepted for publication (in revised form) November 9, 1999; published electronically June 13, 2000. This research was supported by the Engineering and Physical Sciences Research Council of the UK under grants GR/K80228 and GR/M42206.

http://www.siam.org/journals/sisc/22-1/33601.html

[†]Department of Mathematics, University of Dundee, Dundee, DD1 4HN, Scotland (maves@mcs.dund.ac.uk, dfg@mcs.dund.ac.uk).

[‡]Department of Mathematics, University of Strathclyde, Glasgow, G1 1XH, Scotland (djh@maths.strath.ac.uk).

has lead to the development of methods, often implicit or semi-implicit, for time-accurate solutions. In this work we are concerned with the amplitude of the waves, rather than with the speed.

1.2. Model problem. To simplify the presentation, we focus on the case of linear advection, so that the model problem becomes

$$u_t + a u_x = g(u), \quad 0 < x < 1, \quad t > 0, \tag{2}$$

where $a > 0$ is constant and the source term g is nonlinear. In section 6 we indicate how our results carry through to the more general nonlinear flux case (1). For the purpose of illustration we will frequently refer to the logistic source term

$$g(u) = \alpha u(1 - u), \tag{3}$$

where $\alpha > 0$ is constant. This prototypical nonlinear function has been used in many studies of the dynamics of numerical methods.

Making the change of coordinates $s = x - at, \tau = t$ transforms (2) to

$$\frac{du}{d\tau} = g(u). \tag{4}$$

Hence, along the characteristics, where $x - at$ is constant, solutions solve the ODE (4). (The underlying ODE (4), will, of course, also play an important role in the analysis of numerical methods for (2).) For the logistic function (3), given $u_{\tau=0} = u_0$, (4) has the solution

$$u(\tau) = \frac{u_0}{u_0[1 - \exp(-\alpha\tau)] + \exp(-\alpha\tau)}.$$

Hence, when the initial condition is positive the solution tends to the stable fixed point $u \equiv 1$ as $\tau \to \infty$.

We consider both the periodic initial value problem (PIVP) and the initial boundary value problem (IBVP) for (2). In the PIVP case, we are given $u(x, 0)$ for $0 < x < 1$ and u is assumed to be periodic in space: $u(1 + x, t) = u(x, t)$. It is clear from the discussion above that with g given by (3) and positive initial data, the solution converges to the fixed point $u \equiv 1$ as $t \to \infty$.

In the IBVP case we are given $u(x, 0)$ for $0 < x < 1$ and $u(0, t)$ for $t > 0$. For simplicity, we will assume a constant boundary condition, $u(0, t) \equiv u^0$. With the logistic source term (3), if the initial and boundary data are positive, then the solution tends to a (generally) nonuniform steady state with profile

$$u(x, t) = \frac{u^0}{u^0[1 - \exp(-\alpha x/a)] + \exp(-\alpha x/a)}. \tag{5}$$

1.3. Numerical method. If the spatial derivative in (2) is approximated using first order upwind differences [21], then we obtain an ODE system of the general form

$$\mathbf{U}_t = -\tfrac{a}{\Delta x} A\mathbf{U} + \mathbf{g}(\mathbf{U}) + \tfrac{a}{\Delta x}\mathbf{b} =: S(\mathbf{U}), \tag{6}$$

where Δx is a spatial meshsize and $U_j(t) \approx u(j\Delta x, t)$. Here, $\mathbf{g}(\mathbf{U})_j = g(U_j)$. We let $N = 1/\Delta x$, so that $\mathbf{U}(t) \in \mathbb{R}^N$. For the PIVP we have

$$A = \begin{bmatrix} 1 & & & -1 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}, \quad \mathbf{b} \equiv \mathbf{0}, \tag{7}$$

while for the IBVP

$$A = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} u^0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{8}$$

To approximate the semidiscrete system (6) we consider a general explicit, two-stage, second order Runge–Kutta (RK) formula [14, p. 154] with tableau

$$\begin{array}{c|cc} 0 & 0 & \\ 1/(2\theta) & 1/(2\theta) & 0 \\ \hline & 1-\theta & \theta \end{array} .$$

Here $\theta \neq 0$ is a free parameter and we always assume that $0 < \theta \leq 1$. Common choices are $\theta = \frac{1}{2}$ and $\theta = 1$, giving the improved Euler and modified Euler methods, respectively. These time-stepping methods are widely used in the finite-difference discretization of advection problems; see, for example, [11].

Letting Δt denote the time-step, the RK method applied to (6) produces approximations $\mathbf{U}^n \in \mathbb{R}^N$ with $U_j^n \approx U(j\Delta x, n\Delta t)$ according to

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \Delta t \left[(1-\theta)S(\mathbf{U}^n) + \theta S(\mathbf{U}^n + \tfrac{\Delta t}{2\theta} S(\mathbf{U}^n))\right]. \tag{9}$$

2. Fixed points of the semidiscrete problem.

Although we are mainly concerned with the numerical solutions generated by the overall algorithm, a useful starting point is to study the fixed points of the system (6).

2.1. Initial BVP.

For the IBVP (6) has a fixed point when

$$U_{j-1} = U_j - \tfrac{\Delta x}{a} g(U_j), \qquad j = 1, 2, \dots, N, \tag{10}$$

with $U_0 = u^0$. Maps of this type have received considerable attention in the literature; see, for example, [16]. Generically, there are parameter ranges for which periodic cycles of any period may be generated.

The following results give insight into the behavior of the map (10) around a zero of g.

THEOREM 2.1. *Suppose $g \in C^1$ with $g(\beta) = 0$ and $g'(\beta) < 0$ for some $\beta \in \mathbb{R}$. Let $I \subseteq \mathbb{R}$ be an open, connected interval containing β such that $g'(u) < 0$ for all $u \in I$. Then, if $U_0 \in I$, there exists a solution sequence $\{U_j\}_{j=0}^N$ of* (10) *in which the components U_j approach β monotonically as j increases. Furthermore, this fixed point of the semidiscrete IBVP* (6) *is linearly stable.*

Proof. Consider the general case where $U_{j-1} \in I$ with $U_{j-1} \neq \beta$. Define $h_{j-1} : \mathbb{R} \mapsto \mathbb{R}$ by

$$h_{j-1}(u) = u - \frac{\Delta x}{a} g(u) - U_{j-1}. \tag{11}$$

Then, using the mean value theorem,

$$h_{j-1}(U_{j-1}) = -\frac{\Delta x}{a} g(U_{j-1}) = -\frac{\Delta x}{a}(g(U_{j-1}) - g(\beta)) = -\frac{\Delta x}{a} g'(\xi_{j-1})(U_{j-1} - \beta),$$

where $\xi_{j-1} \in I$. Hence, $h_{j-1}(\beta) h_{j-1}(U_{j-1}) = (\Delta x/a) g'(\xi_{j-1})(U_{j-1} - \beta))^2 < 0$; so there is a zero U_j of h_{j-1} between U_{j-1} and β.

The Jacobian of the ODE (6) at this fixed point is lower triangular with strictly negative diagonal entries. Hence, the fixed point is linearly stable. □

Theorem 2.1 gives the reassuring result that there are stable, smooth, fixed points of (6) that mimic those of the underlying PDE. The next result shows that unstable, oscillatory, fixed points may also exist if the spatial resolution is inadequate.

Theorem 2.2. *Suppose $g \in C^1$ with $g(\beta) = 0$ and $g'(\beta) > 2a/\Delta x$ for some $\beta \in \mathbb{R}$. Let $I \subseteq \mathbb{R}$ be an open, connected interval containing β such that $g'(u) > 2a/\Delta x$ for all $u \in I$. Then if $U_0 \in I$ and $U_0 - \Delta x g(U_0)/a \in I$ there exists a solution sequence $\{U_j\}_{j=0}^N$ of* (10) *in which the components U_j approach β as j increases, with successive components lying on opposite sides of β. Furthermore, this fixed point of the semidiscrete IBVP* (6) *is linearly unstable.*

Proof. Consider the general case where $U_{j-1} \in I$ with $U_{j-1} \neq \beta$, and let $W_{j-1} := U_{j-1} - \Delta x g(U_{j-1})/a \in I$. From the mean value theorem,

$$(W_{j-1} - \beta)(U_{j-1} - \beta) = (U_{j-1} - \beta)^2(1 - \Delta x g'(\xi_{j-1})/a) < 0,$$

where $\xi_{j-1} \in I$. Hence, U_{j-1} and W_{j-1} lie on opposite sides of β. Let h_{j-1} be defined by (11). Then $h'_{j-1}(u) < -1$ for all $u \in I$.

Now, consider the case where $U_{j-1} < \beta$. Note that $h_{j-1}(\beta) > 0$. For any $u \in I$ with $u > U_{j-1}$ we have

$$h_{j-1}(u) = h_{j-1}(U_{j-1}) + \int_{U_{j-1}}^{u} h'_{j-1}(u)\, du < h_{j-1}(U_{j-1}) - u + U_{j-1}.$$

Putting $u = W_{j-1}$ gives $h_{j-1}(W_{j-1}) < 0$. Hence, h_{j-1} has a root $U_j \in (\beta, W_{j-1})$. Also, we note that $W_j := U_j - \Delta x g(U_j)/a = U_{j-1} \in I$.

In the case where $U_{j-1} > \beta$, a similar argument shows that h_{j-1} has a root $U_j \in (W_{j-1}, \beta)$.

Linear instability follows by observing that the relevant Jacobian of the ODE (6) is lower triangular with strictly positive diagonal entries. □

For the logistic source term (3), these results lead to the following corollary.

Corollary 2.3. *Consider the semidiscrete IBVP* (6) *with logistic source term* (3).

If $U_0 \in (1/2, \infty)$, then there exists a linearly stable fixed point $\{U_j\}_{j=0}^N$ of (6) *for which the components U_j approach* 1 *monotonically as j increases.*

If $0 < \psi < 1/2$ and $(1-\psi-\sqrt{1-3\psi^2})/2 < U_0 < (1/2)-\psi$, where $\psi := a/(\alpha\Delta x)$, then there exists a linearly unstable fixed point $\{U_j\}_{j=0}^N$ of (6) *for which the components U_j approach zero as j increases with successive components lying on opposite sides of β.*

A spatially uniform fixed point (SUFP) has the form $\mathbf{U} = \beta\mathbf{e}$, where $\mathbf{e} = [1, 1, \dots, 1]^T$. It is clear from the form of (6) that such a fixed point exists only if $g(\beta) = 0$ and $u^0 = \beta$. In this case, $\beta\mathbf{e}$ is a discrete analogue of a fixed point of (4), and hence of the problem (2). (Note that strict spatial uniformity requires the appropriate boundary condition to be specified. However, as illustrated by Theorems 2.1 and 2.2, fixed points that are almost spatially uniform arise generically.)

The Jacobian of the ODE (6) at a point $\mathbf{U} = \beta\mathbf{e}$ is $g'(\beta)I - (a/\Delta x)A$. This matrix is lower bidiagonal, and hence, it has the single eigenvalue $\lambda = g'(\beta) - a/\Delta x$. This implies that a fixed point $\mathbf{U} = \beta\mathbf{e}$ is stable for

$$g'(\beta) - \tfrac{a}{\Delta x} < 0, \tag{12}$$

a condition that is automatically satisfied if $g'(\beta) \leq 0$ and $a > 0$. However, the Jacobian may be highly nonnormal, leading to extreme transient growth of perturbations in (6) and a negligible basin of attraction for the fixed point. In this case it is more profitable to study the pseudospectrum of the Jacobian [17]. Omitting the details, the pseudospectrum of this matrix in an appropriate limit is given by (see [18])

$$g'(\beta) - \tfrac{a}{\Delta x}(1 - z), \qquad z \in \mathbb{C}, \quad |z| \leq 1.$$

In order to have stability in a practical sense when N is large, it is necessary that the pseudospectrum lie in the left half of the complex plane. This reduces to the condition $g'(\beta) < 0$, which is identical to the corresponding constraint for the PIVP (see section 2.2).

2.2. Periodic IVP.

The PIVP version of (6) has a fixed point when (10) holds, with the interpretation that $U_0 \equiv U_N$. As in the IBVP case, we may study fixed points that are close to a zero of g. The arguments used in Theorems 2.1 and 2.2 remain valid, and, because of the "wrap-around" effect, the following stronger result holds.

THEOREM 2.4. *Suppose $g \in C^1$ with $g(\beta) = 0$ for some $\beta \in \mathbb{R}$.*

If $g'(\beta) < 0$, let $I \subseteq \mathbb{R}$ be the largest open, connected interval containing β such that $g'(u) < 0$ for all $u \in I$. Then the only fixed point of the semidiscrete PIVP (6) *for which $U_0 \in I$ is the linearly stable SUFP $\beta\mathbf{e}$.*

If $g'(\beta) > 2a/\Delta x$, let $I \subseteq \mathbb{R}$ be the largest open, connected interval containing β such that $g'(u) > 2a/\Delta x$ for all $u \in I$. Then the only fixed point of the semidiscrete PIVP (6) *for which $U_0 \in I$ and $U_0 - \Delta x g(U_0)/a \in I$ is the linearly unstable SUFP $\beta\mathbf{e}$.*

3. Fixed points of the overall algorithm.

3.1. Time-stepping on the underlying ODE.

We begin this section by examining fixed points of the underlying RK(θ) method on the scalar ODE $u_t = g(u)$. Introducing z to represent an intermediate stage value, a fixed point β of the RK(θ) scheme must satisfy

$$\beta + \tfrac{\Delta t}{2\theta} g(\beta) = z, \tag{13}$$

$$(1 - \theta) g(\beta) + \theta g(z) = 0. \tag{14}$$

It is clear that if $g(\beta) = 0$, then $z = \beta$ is a solution—so that fixed points of the ODE are inherited by the RK scheme. (This is true for all RK schemes [12].) However, for general nonlinear g, spurious fixed points, where $g(\beta) \neq 0$, may be admitted.

It is possible to make some general comments about spurious fixed points in (13)–(14). (Recall our assumption that $0 < \theta \leq 1$.) First, from (14), there must be at least one zero of g between z and β. Moreover, if u is such that $g(u) = 0$, $g'(u) \neq 0$, and $\beta = u + c_1\epsilon + c_2\epsilon^2 + \cdots$ for some small parameter ϵ and constants $c_1, c_2, \ldots$, then

$$(1 - \theta)\beta + \theta z = u + O(\epsilon^2).$$

This tells us that z is better than β as an approximation to u when $\theta \in (1/2, 1]$ and ϵ is small.

Fixed points arising with the logistic nonlinearity (3) have been studied by many authors. For any $0 < \theta \leq 1$, the fixed point $u = 1$ is stable for $0 < \Delta t < 2$. Letting $r := \alpha\Delta t$, the following results about spurious fixed points are known for the popular choices of $\theta = 1/2$ and $\theta = 1$, [6].

$\theta = 1$: spurious fixed points $u = 1 + 2/r$ and $u = 2/r$ exist for all r. The former is stable for $0 < r < -1 + \sqrt{5}$, and the latter is stable for $2 < r < 1 + \sqrt{5}$.

$\theta = \frac{1}{2}$: spurious fixed points $u = [2 + r \pm \sqrt{r^2 - 4}]/(2r)$ exist for $r > 2$ and are stable for $2 < r < \sqrt{8}$.

Two points are worth emphasizing.

- In these examples, spurious fixed points either cease to exist or become arbitrarily large in modulus as $\Delta t \to 0$. Humphries [10] proved a general result in this vein—for any RK formula and any locally Lipschitz g, a spurious fixed point that persists for small Δt must blow up as $\Delta t \to 0$.
- The $\theta = 1$ case above shows that although spurious fixed points must blow up as $\Delta t \to 0$, they may remain stable for arbitrarily small Δt.

3.2. Spatially uniform fixed points of the PIVP. For the PIVP it is straightforward to confirm that $\beta\mathbf{e}$ is a fixed point of the overall algorithm if and only if β is a fixed point for the RK method on the scalar ODE $u_t = g(u)$. Hence, the question of *existence* of SUFPs has been answered in the previous subsection. *Stability*, however, is a separate issue.

Suppose first that $\beta\mathbf{e}$ is a nonspurious SUFP, so that $g(\beta) = 0$. Then we know from [12, Theorem 3] that the appropriate condition for linear stability is $|R(z)| < 1$ for every $z = \Delta t\lambda$, where λ is an eigenvalue of $-aA/\Delta x + g'(\beta)I$ and $R(z) = 1 + z + z^2/2$. The matrix A has eigenvalues $1 - \exp(i\varphi)$, where $\varphi \equiv \varphi_j = 2\pi j/N$, $j = 1, 2, \ldots, N$. Treating $\varphi \in [0, 2\pi)$ as a continuous variable (which may be justified when N is large), it follows that the spectrum of A lies on a circle centered on the negative real axis. It may then be shown that the real eigenvalues of A always dominate the stability condition and the required constraint may be written

$$0 < -\Delta t g'(\beta) < 2 - 2c. \tag{15}$$

For the logistic source term (3) with $\beta = 1$, this becomes $0 < r < 2 - 2c$, where $r = \alpha\Delta t$. This region is shown in light gray in the plots of Figure 1.

Generally, writing the map (9) as $\mathbf{U}^{n+1} = G(\mathbf{U}^n)$, after some manipulation the Jacobian of G at a point $\mathbf{U} = \beta\mathbf{e}$ may be written in the form

$$G'(\beta\mathbf{e}) = (1 + \rho - \tfrac{1}{2}\gamma^2)I + \tfrac{1}{2}(cA - \gamma I)^2, \tag{16}$$

where

$$\gamma = 1 + \tfrac{1}{2}\Delta t(g'(\beta) + g'(z)), \tag{17}$$

$$\rho = \Delta t((1-\theta)g'(\beta) + \theta g'(z) + \tfrac{1}{2}\Delta t g'(\beta)g'(z)), \tag{18}$$

with z defined in (13). We use $c := a\Delta t/\Delta x$ to denote the Courant number [21]. For the logistic source term it is possible to determine the stable parameter ranges of the spurious fixed points corresponding to each value of θ. This leads to "almost triangular" regions with a vertical edge determined by the eigenvalue $\lambda(A) = 0$ and the "hypotenuse" by $\lambda(A) = 2$, $(\varphi = \pi)$. The remaining eigenvalues reduce the regions by rounding the apex in such a way as to create a C^1 smooth boundary. The details are omitted in view of their complexity. These regions are shown in darker gray in Figure 1 for $\theta = 1/2, 3/4, 1$.

Two key points arise from Figure 1. First, for $\theta > \frac{1}{2}$ it is possible to choose parameter values where both the correct and spurious fixed points are stable. In such cases, the choice of initial data will determine which, if either, of the two steady states

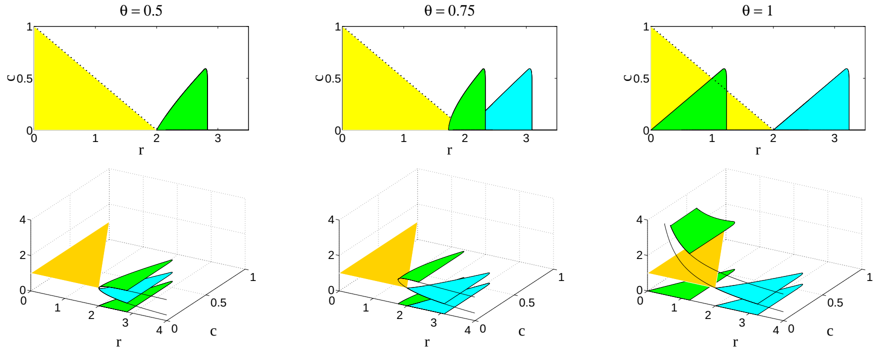


Fig. 1. *Stable parameter ranges in the* (r, c)*-plane for steady states of the RK(*θ*) methods with* $\theta = 0.5, 0.75, 1.0$ *on the PIVP with* (3)*. On the bottom row the vertical axis shows the magnitude of the steady state solution.*

is computed. A numerical illustration of this effect is given in section 3.4. The second point concerns the stability of spurious fixed points for small Δt. We know from section 3.1 that for $\theta = 1$ on the scalar ODE the spurious fixed point $u = 1 + 2/r$ is stable for $0 < r < -1 + \sqrt{5} \approx 1.2$. This stability interval is seen along the $c = 0$ axis at the base of the dark shaded $\theta = 1$ region in Figure 1. However, it is clear that if we increase c beyond zero, moving from the ODE to the PDE, then the spurious fixed point is always unstable for small r, that is, small Δt. We now develop some theory to show that this behavior is generic.

The lemma below concerns the scalar ODE.

Lemma 3.1. *Consider the RK(*θ*) method* (9) *(with* $0 < \theta \leq 1$*) applied to the scalar ODE* (4)*. Suppose that* $g : \mathbb{R} \mapsto \mathbb{R}$ *is bounded and differentiable on every bounded interval, the zeros of* g *are separated, and* $g(u)$ *is monotonic for all* $|u|$ *sufficiently large. Suppose further that for sufficiently small* Δt *there is a spurious fixed point* $\beta = \beta(\Delta t)$ *with* β *depending* C^1 *continuously upon* Δt *and with* $g(\beta)$ *bounded away from zero for small* Δt*.*

(1) *For* $\theta = 1$*, we have* $g'(\beta) < -2/\Delta t$ *for sufficiently small* Δt*.*

(2) *For* $0 < \theta < 1$*, the fixed point* β *of the method* (9) *is linearly unstable for sufficiently small* Δt*.*

Proof. Let $z = z(\Delta t)$ be defined by (13). Note that if β is bounded as $\Delta t \to 0$, then $\Delta t g(\beta) \to 0$, so that $z \to \beta$ in (13). In this case, from (14), $g(\beta) \to 0$. This contradicts the assumption that $g(\beta)$ is bounded away from zero for small Δt, and hence, we must have $|\beta| \to \infty$ as $\Delta t \to 0$.

For the remainder of the proof we assume without comment that Δt is sufficiently small for our arguments to hold. We use a dot to represent differentiation with respect to Δt so that, for example, $\dot{\beta}$ denotes $d\beta/d\Delta t$.

For $\theta = 1$ we have $g(z) = 0$ from (14). Hence, z is fixed as $\Delta t \to 0$. It follows from (13) that $\beta g(\beta) < 0$ and hence, from the monotonicity of g, that $g'(\beta) < 0$. Differentiating (13) with respect to Δt gives

$$\beta\dot{\beta}(2 + \Delta t g'(\beta)) = -\beta g(\beta) > 0. \tag{19}$$

Now β^2 increases as $\Delta t \to 0$, so $d(\beta^2)/d\Delta t < 0$; that is, $2\beta\dot{\beta} < 0$. Using this in (19) gives $g'(\beta) < -2/\Delta t$.

For $0 < \theta < 1$ it follows from (14) that $g(\beta)$ and $g(z)$ are of opposing sign.

From the boundedness and monotonicity assumptions on g, we must have $|z| \to \infty$ as $\Delta t \to 0$, with $z\beta < 0$. In (13) we then find that

$$\beta^2 + \frac{\Delta t \beta g(\beta)}{2\theta} = z\beta < 0,$$

and hence, $\beta g(\beta) < 0$. The monotonicity of g then forces $g'(\beta) < 0$.

Now (13) and (14) give

$$2(1-\theta)z + \Delta t g(z) = 2(1-\theta)\beta. \tag{20}$$

Differentiating (13) and (20) with respect to Δt gives

$$\begin{aligned}(2\theta + \Delta t g'(\beta))\dot\beta + g(\beta) &= 2\theta \dot z,\\ (2 - 2\theta + \Delta t g'(z))\dot z + g(z) &= 2(1-\theta)\dot\beta.\end{aligned}$$

This may be written

$$\begin{bmatrix} 2\theta + \Delta t g'(\beta) & -2\theta \\ -2(1-\theta) & (2 - 2\theta + \Delta t g'(z)) \end{bmatrix} \begin{bmatrix} \dot\beta \\ \dot z \end{bmatrix} = -\begin{bmatrix} g(\beta) \\ g(z) \end{bmatrix}. \tag{21}$$

The determinant of the two-by-two matrix in (21) is 2ρ, where ρ is defined in (18). Hence,

$$2\rho \begin{bmatrix} \dot\beta \\ \dot z \end{bmatrix} = -\begin{bmatrix} (2 - 2\theta + \Delta t g'(z))g(\beta) + 2\theta g(z) \\ (2-2\theta)g(\beta) + (2\theta + \Delta t g'(\beta))g(z) \end{bmatrix}.$$

Using (14), this simplifies to

$$2\rho \begin{bmatrix} \dot\beta \\ \dot z \end{bmatrix} = -\begin{bmatrix} \Delta t g(\beta) g'(z) \\ \Delta t g(z) g'(\beta) \end{bmatrix}.$$

Thus, multiplying the first component by β gives

$$\rho \frac{d\beta^2}{d\Delta t} = -\Delta t \beta g(\beta) g'(z). \tag{22}$$

Our earlier arguments showed that $\beta g(\beta) < 0$, and since z and β tend to $\pm\infty$ in opposite directions, it follows from the monotonicity of g that $g'(z) < 0$. Hence, the right-hand side of (22) is negative. Since $d(\beta^2)/d\Delta t < 0$, we deduce that $\rho > 0$. This gives $G'(\beta) > 1$ in (16) (with the dimension set to $N = 1$ and with $A = 0$), and therefore, the spurious fixed point is unstable. □

We now use Lemma 3.1 to establish a result about the stability of spurious SUFPs of the semidiscrete PDE. The theorem below applies to a general class of semidiscrete problems, for which the PIVP (6) is a special case. (Adding this level of generality makes the theorem more widely applicable and does not complicate the proof.) The result shows that stable, spurious SUFPs occur only where there is a lack of spatial resolution.

THEOREM 3.2. *Consider the RK(θ) method* (9) *(with $0 < \theta \le 1$) applied to an ODE system of the form*

$$\mathbf{U}_t = -\tfrac{a}{\Delta x^m} A\mathbf{U} + \mathbf{g}(\mathbf{U}), \tag{23}$$

where $a, \Delta x > 0$ and $\mathbf{g}(\mathbf{U})_i \equiv g(U_i)$ with g satisfying the assumptions in Lemma 3.1. *Suppose that $A \in \mathbb{R}^{N\times N}$ satisfies $A\mathbf{e} = \mathbf{0}$ and has an eigenvalue $0 < \lambda \in \mathbb{R}$. Consider a spurious SUFP $\beta\mathbf{e}$, where $\beta = \beta(\Delta t)$ depends C^1 continuously upon Δt, with $g(\beta)$ bounded away from zero for small Δt.*

(1) *For $\theta = 1$, such a SUFP cannot be stable for small Δt if*

$$\Delta x^m < \frac{a\lambda}{g'(z)}, \tag{24}$$

where the intermediate stage value z is defined in (13).

(2) *For $0 < \theta < 1$, such a SUFP cannot be stable for small Δt.*

Proof. First, note that since $A\mathbf{e} = 0$, if $\beta\mathbf{e}$ is a SUFP for the RK(θ) method on (23), then β is a fixed point for the same method on the scalar problem (4). Hence, we may appeal to Lemma 3.1.

Stability of the SUFP $\beta\mathbf{e}$ is determined by the spectrum of the matrix $G'(\beta\mathbf{e})$ in (16)–(18), with $c := a\Delta t/\Delta x^m$. Since A has an eigenvalue $\lambda \in \mathbb{R}$, $G'(\beta\mathbf{e})$ has a real eigenvalue of the form

$$\mu = 1 + \rho + \tfrac{1}{2}c^2\lambda^2 - \gamma c\lambda.$$

Instability of the SUFP $\beta\mathbf{e}$ follows if we can show that $\mu - 1 > 0$.

Consider first the case $\theta = 1$. It is straightforward to verify that

$$\mu - 1 = \tfrac{1}{2}(\lambda c - \Delta t g'(\beta) - 2)(\lambda c - \Delta t g'(z)),$$

where, from Lemma 3.1, the first factor on the right-hand side is positive for small Δt. Consequently, $\mu - 1 > 0$ if $\lambda c > \Delta t g'(z)$, from which, using $c = a\Delta t/\Delta x^m$, we obtain (24).

The result for $0 < \theta < 1$ is almost immediate from Lemma 3.1. Since $\mu > 1$ when $c = 0$, it follows by continuity that $\mu > 1$ for small c. □

For the picture in the top right-hand corner of Figure 1 we have $\theta = 1$, $\lambda = 2$, $m = 1$, $\beta = 1 + 2/(\alpha\Delta t)$, $z = 0$, $g'(0) = \alpha > 0$. So the condition in Theorem 3.2 for no stable spurious SUFPs when Δt is small becomes $\Delta x < 2a/\alpha$; that is, $c > r/2$. Figure 1 shows this bound to be sharp.

3.3. Spatially uniform fixed points of the IBVP.

For the IBVP the PDE has a SUFP if and only if $g(u^0) = 0$. In this case, the map (9) inherits the fixed point $\beta\mathbf{e}$, with $\beta = u^0$. The Jacobian of the map at this point is lower triangular with Toeplitz form. The matrix has a single eigenvalue, which is real, and restricting this to lie in the interval $(-1, 1)$ produces the stability condition

$$0 < c - \Delta t g'(\beta) < 2. \tag{25}$$

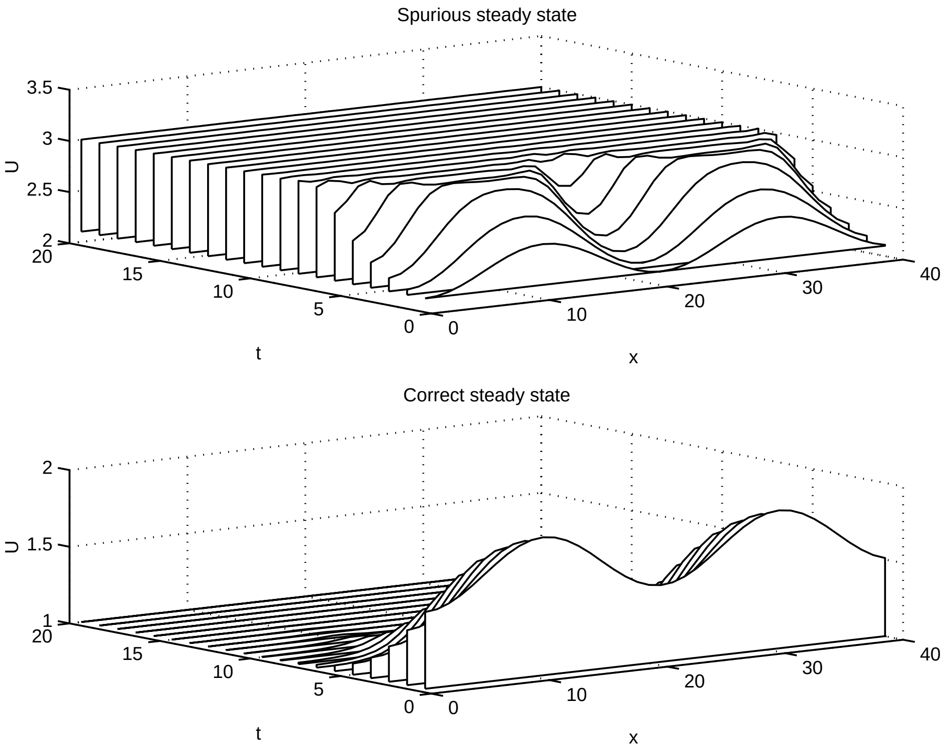
The left-hand inequality requires $\Delta x g'(\beta) < a$, which forces Δx to be sufficiently small when $g'(\beta) > 0$, but imposes no restriction when $g'(\beta) < 0$. The right-hand inequality then requires

$$\Delta t < \frac{2\Delta x}{a - \Delta x g'(\beta)}. \tag{26}$$

However, since the Jacobian is (potentially) highly nonnormal, eigenvalues may be misleading indicators of stability. This phenomenon is well known in the analysis of linear stability of PDE methods [17] and corresponds to the classical result [19] that stability of the PIVP is a necessary condition for stability of the IBVP.

We now show the nonexistence of spurious SUFPs.

THEOREM 3.3. *For the IBVP, when $N \geq 3$, the numerical method* (9) *does not admit a spatially uniform, spurious fixed point.*


Spurious steady state

Correct steady state

FIG. 2. *Spurious and correct steady states for the PIVP.*

Proof. If $\beta\mathbf{e}$ is a SUFP, then

$$(1-\theta)S(\beta\mathbf{e}) + \theta S\left(\beta\mathbf{e} + \frac{\Delta t}{2\theta}S(\beta\mathbf{e})\right) = 0. \tag{27}$$

From the form of S in (6), considering the jth component of (27), where $j > 2$ leads to the condition

$$(1-\theta)g(\beta) + \theta g\left(\beta + \frac{\Delta t}{2\theta}g(\beta)\right) = 0. \tag{28}$$

This shows that β must be a fixed point for the RK(θ) method applied to the underlying scalar ODE. Now, taking the second component in (27) and using (28) gives

$$u^0 - \beta = 0. \tag{29}$$

Finally, using (28) and (29) in the first component of (27) shows that

$$\frac{a\Delta t}{2\Delta x}g(\beta) = 0.$$

Hence, $g(\beta) = 0$, and the fixed point is not spurious. □

From the proof of Theorem 3.3 it is clear that if β is a spurious fixed point for the RK(θ) method applied to the underlying scalar ODE, then $\beta\mathbf{e}$ satisfies the conditions required for a fixed point on the IBVP problem, except near the boundary $j = 0$. We will see in the next section that it is possible for the method to settle to a fixed point that is close to β except near the left-hand boundary.

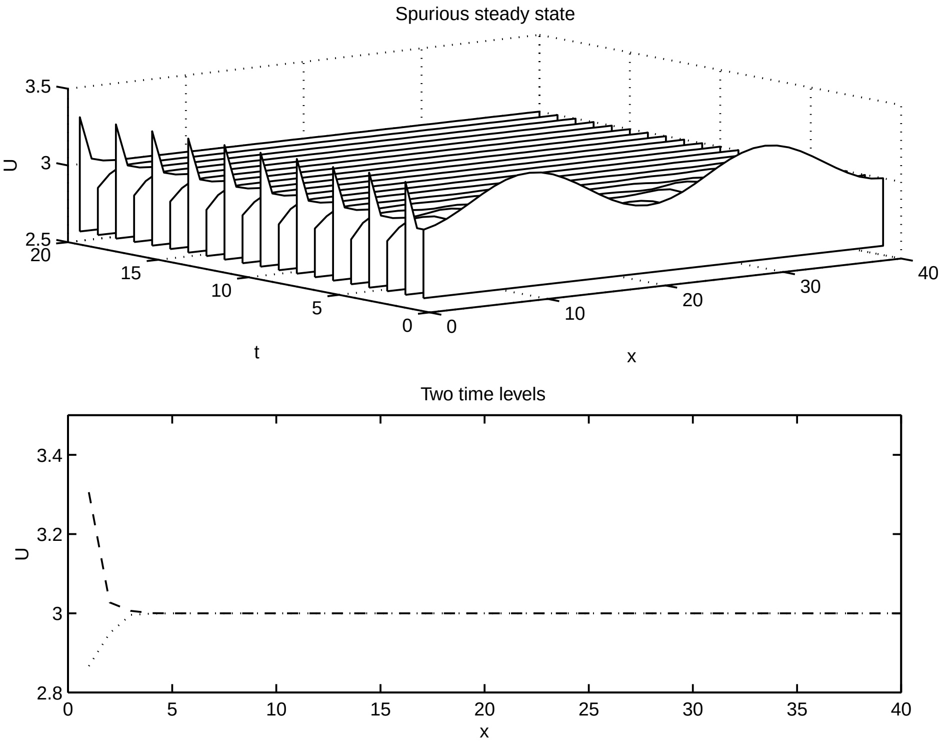


Fig. 3. *Spurious nonuniform steady state for the IBVP.*

3.4. Numerical examples and further analysis. We now present some numerical results. All computations used $a = 1$ in (2) and $\theta = 1$ in (9). We begin by illustrating that stable spurious and stable correct fixed points may coexist. Figure 2 shows the first 20 time-levels for the PIVP using $\Delta x = 1/N = 1/40$, $\Delta t = 0.1\Delta x$ and with $\alpha = 400$ in the logistic source term (3). This gives $r = 1$ and $c = 0.1$, and we see from the top right-hand picture in Figure 1 that the fixed points $\mathbf{U} = \beta\mathbf{e}$ with $\beta = \beta_{\text{corr}} := 1$ (correct) and $\beta = \beta_{\text{spur}} := 1 + 2/r = 3$ (spurious) are stable. The upper picture in Figure 2 was generated using initial data $u(x, 0) = 0.7\beta_{\text{spur}} + 0.4\sin(2\pi x)^2$, and the solution is attracted to the spurious level. The lower picture used $u(x, 0) = 0.5\beta_{\text{spur}} + 0.4\sin(2\pi x)^2$, which is seen to be in the basin of attraction of the true fixed point.

The upper picture in Figure 3 gives the result when the same parameter values as above are used on the IBVP with initial data $u(x, 0) = \beta_{\text{spur}} + 0.3\sin(2\pi x)^2$ and a boundary condition of $u^0 = \beta_{\text{spur}}$. In this case, most components have settled close to the spurious level β_{spur}, but there are oscillations at the left-hand boundary. (For clarity, the boundary value is not plotted.) The overall solution is period two in time. The lower picture in Figure 3 gives the two profiles between which the solution oscillates. Note that Theorem 3.3 shows that a spatially uniform spurious fixed point cannot be computed for the IBVP.

In the course of our experiments, we found that, on both the IBVP and PIVP, it was common for the solution to settle down to a stable steady state that consisted of disjoint "locally uniform" patches that mix together spurious and correct levels. Figure 4 gives an example. This stable steady state arose on the PIVP with the same

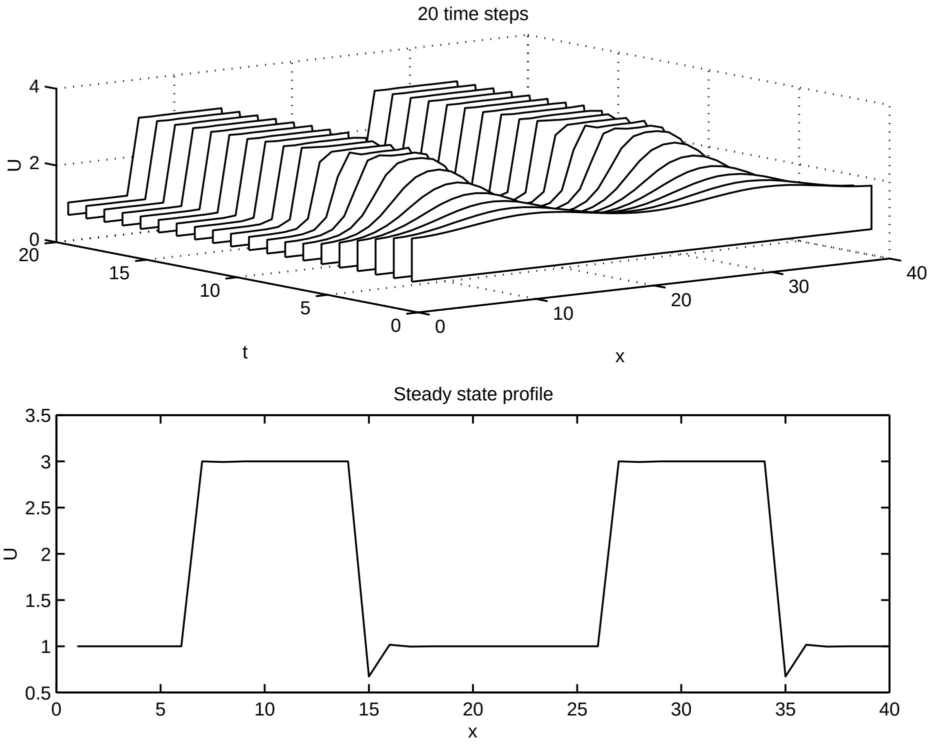


Fig. 4. *Stable steady state with jumps.*

parameters as above; we simply changed the initial condition to $u(x,0) = 0.6\beta_{\text{spur}} + 0.3\sin(2\pi x)^2$. In this example, some initial data has been attracted to the spurious level β_{spur} and some to the correct level β_{corr}. The lower picture in Figure 4 gives the profile of the steady state.

The profile shown in Figure 4 can be analyzed further, as follows. Writing the map (9) as $\mathbf{U}^{n+1} = G(\mathbf{U}^n)$, a fixed point must satisfy $\mathbf{U} = G(\mathbf{U})$, which reduces to a set of scalar equations for the components:

$$-\frac{a}{\Delta x}[U_j - U_{j-1}] + \frac{a^2\Delta t}{2\Delta x^2}[U_j - 2U_{j-1} + U_{j-2}] - \frac{a\Delta t}{2\Delta x}[g(U_j) - g(U_{j-1})] \\ + g\left(U_j - \frac{a\Delta t}{2\Delta x}(U_j - U_{j-1}) + \frac{\Delta t}{2}g(U_j)\right) = 0,$$

for $j = 2, 3, \ldots$. Hence, the sequence $\{U_j\}_{j=0}^{N}$ satisfies a two-step recurrence that we may write as

$$q(U_{j-2}, U_{j-1}, U_j) = 0. \tag{30}$$

Note that as an equation for U_j (given U_{j-2} and U_{j-1}), (30) is implicit. On the other hand, if we regard (30) as an equation for U_{j-2} (given U_j and U_{j-1}), then the map is explicit.

With the logistic source term, to explain the upward jumps in Figure 4 we look for a solution sequence for (30) of the form

$$\ldots, 1, 1, 1 + \frac{2}{r} + c^2V_0, 1 + \frac{2}{r} + c^2V_1, 1 + \frac{2}{r} + c^2V_2, \ldots. \tag{31}$$

Since $q(1,1,1+2/r)=0$, we take $V_0=0$. The equation $q(1,1+2/r,1+\frac{2}{r}+c^2V_1)=0$ has a solution $V_1=-2/(r^2(r+2))+O(c)$. Generally, assuming that the $\{V_j\}$ are small and linearizing produces the relation

$$V_j\,(r+2+c)\,(r-c)+2cV_{j-1}\,(1+c)-c^2V_{j-2}=0. \tag{32}$$

The characteristic polynomial of this recurrence has roots $-c/(r-c)$ and $c/(r+2+c)$. Since $r>0$ and $c>0$, it follows that $V_j\to 0$ as $j\to\infty$ in (32) if $r>c$, with very rapid convergence when $r\gg c$.

Next we look for a downward jump solution of the form

$$\ldots,1+\frac{2}{r},1+\frac{2}{r},1+V_0,1+V_1,1+V_2,\ldots. \tag{33}$$

Setting $q(1+2/r,1+2/r,1+V_0)=0$ gives the quadratic

$$V_0^2r^2+V_0^2r\,(r+c-2)-4c.$$

We will take the smaller root, for which $V_0=\frac{4c}{r(r-2)}+O(c^2)$. Linearizing $q(1+2/r,1+V_0,1+V_1)$ gives

$$V_1\approx 2c\frac{r^2V_0+rcV_0-c-V_0r}{(r+c-2)\,r\,(r+rcV_0+c)}.$$

Generally, linearizing the equation $g(1+V_{j-2},1+V_{j-1},1+Vj)=0$ gives the recurrence

$$V_j\,(r+c)\,(r+c-2)-2cV_{j-1}\,(c-1+r)+c^2V_{j-2}=0, \tag{34}$$

whose characteristic polynomial has the roots $c/(r+c), c/(r+c-2)$. It follows that a sufficient condition for $V_j\to 0$ as $j\to\infty$ in (34) is $r<2-c$.

The existence of the c^2 factor in the sequence (31) suggests that upward jumps settle to the new level more rapidly than downward jumps. This agrees with the profile in Figure 4.

Next, we show that for small c it is generic to have steady states profiles of the type pictured in Figure 4, that is, profiles that jump between levels, with each level corresponding to a stable (correct or spurious) fixed point for the RK map on the underlying scalar ODE. Note that this result formalizes an argument given in [4].

THEOREM 3.4. *Suppose $g\in C^1$. Regard Δt and Δx as fixed and allow $c=a\Delta t/\Delta x$ to vary. Suppose $\mathbf{U}^*\in\mathbb{R}^N$ is such that every component of $\mathbf{U}^*$ is a stable (correct or spurious) fixed point of the RK(θ) method on the underlying scalar ODE* (4)*. Then there exists $c^*>0$ such that for all $0\le c\le c^*$ there is a stable fixed point $\mathbf{U}=\mathbf{U}(c)$ of the RK method on the PIVP* (6)*, where $\mathbf{U}(c)$ depends continuously upon c and $\mathbf{U}(0)=\mathbf{U}^*$.*

Proof. Write the RK method on the PIVP (6) as

$$\mathbf{U}^{n+1}=\mathbf{U}^n+\Delta tH(c,U^n). \tag{35}$$

We know that $H(0,\mathbf{U}^*)=0$. At $c=0$ the map (35) uncouples; the Jacobian $\partial H(0,\mathbf{U})/\partial\mathbf{U}$ is diagonal. Since each component of $\mathbf{U}^*$ represents a *stable* fixed point on the underlying scalar ODE, we have

$$\left|1+\Delta t\left(\frac{\partial H(0,\mathbf{U}^*)}{\partial\mathbf{U}}\right)_{i,i}\right|<1,\quad 0\le i\le N.$$

It follows that each element on the diagonal of $\partial H(0, \mathbf{U}^*)/\partial \mathbf{U}$ is nonzero, and hence, the Jacobian is nonsingular. The implicit function theorem may therefore be invoked, to show that there exists $c^* > 0$ such that for all $0 \leq c \leq c^*$ there is a solution $\mathbf{U}(c)$ of the nonlinear system $H(c, \mathbf{U}(c)) = 0$, where $\mathbf{U}(c)$ depends continuously upon c and $\mathbf{U}(0) = \mathbf{U}^*$. Since the eigenvalues of a matrix depend continuously upon its entries, it follows that by further reduction of c^*, if necessary, we can ensure that the eigenvalues of $I + \Delta t \partial H(c, \mathbf{U}(c))/\partial \mathbf{U}$ remain inside the unit disc for all $0 \leq c \leq c^*$. □

4. Nonnormality in the IBVP.

We now describe a type of spurious behavior that arises on the IBVP in the presence of finite precision and nonnormality. The effect is linear, in the sense that the same behavior can be seen when the logistic source term is replaced by $\alpha(1 - u)$.

Figure 5 shows the results of a computation on the IBVP with the logistic source term (3) using $\alpha = 400$, $\Delta x = 1/N = 1/40$, and $\Delta t = .97 \Delta t_{\text{lim}}$. Here, $\Delta t_{\text{lim}} := 2\Delta x/(1 + \alpha \Delta x)$ is the eigenvalue stability limit, that is, the time-step beyond which the correct steady state $\mathbf{U}^n = \mathbf{e}$ becomes linearly unstable. We took $u(x, 0) = 1 + 10^{-8}$ for the initial data and $u^0 = 1$ as the boundary condition. The upper picture in Figure 5 shows the solution evolving over the first 120 time-steps (with every fourth time level plotted). In this regime, small oscillations are growing in magnitude and moving to the right. The lower picture in Figure 5 shows time-levels $n = 4010$ to $n = 4040$. This illustrates the typical long-term behavior: there are $O(1)$ oscillations around the right-hand boundary that persist for all time.

This type of behavior arises only when (a) the time step is chosen near the linear stability limit, and (b) the initial data and boundary condition are close to the true steady state. In such circumstances the numerical solution was observed either to blow up or tend to a state of the type illustrated in Figure 5.

Spurious behavior of this form was first described on a slightly different model PDE in [9], and the following simple argument explains the broad details.

After linearizing the problem, the iteration (9) has the form $\mathbf{E}^{n+1} = B\mathbf{E}^n$, where $\mathbf{E}^n = \mathbf{U}^n - \mathbf{e}$ and $B = G'(\mathbf{e})$. The matrix B has a spectral radius of .94, and hence, in exact arithmetic we have $\mathbf{E}^n \to 0$ as $n \to \infty$. However, in the presence of rounding errors the appropriate model is $\mathbf{E}^{n+1} = \widehat{B}^n \mathbf{E}^n$, where $\widehat{B}^n$ is a perturbation of B. In our case, B is highly nonnormal and so the eigenvalues of $\widehat{B}^n$ may be very different from those of B. If $\widehat{B}^n$ has an eigenvalue close to the boundary of the unit disc, then $\mathbf{E}^n$ will not decay, and hence, $\mathbf{U}^n$ will not approach the steady state $\mathbf{e}$. In our example B has a single eigenvalue .94, but there is a perturbation of Euclidean norm 10^{-15} that produces an eigenvalue with real part beyond $+1$. What we observe in the lower picture in Figure 5 are *pseudoeigenvectors* of B; that is, eigenvectors of perturbed matrices $\widehat{B}^n$.

5. Adaptivity.

So far, we have analyzed a finite difference scheme on a uniform grid. In this section we ask whether adaption in either space or time is inimical to spuriosity.

5.1. Adaptivity in space.

First, we consider a "static" spatial equidistribution algorithm of the type described in [2, 22]. The algorithm proceeds as follows. Suppose that at time-level n we have an approximation $\mathbf{U}^n$ on a spatial mesh given by $\Delta x_i^n := x_{i+1}^n - x_i^n$. Then the finite difference scheme (9) (with appropriate modifications of the spatial differences to accommodate the nonuniform mesh) is used to produce an

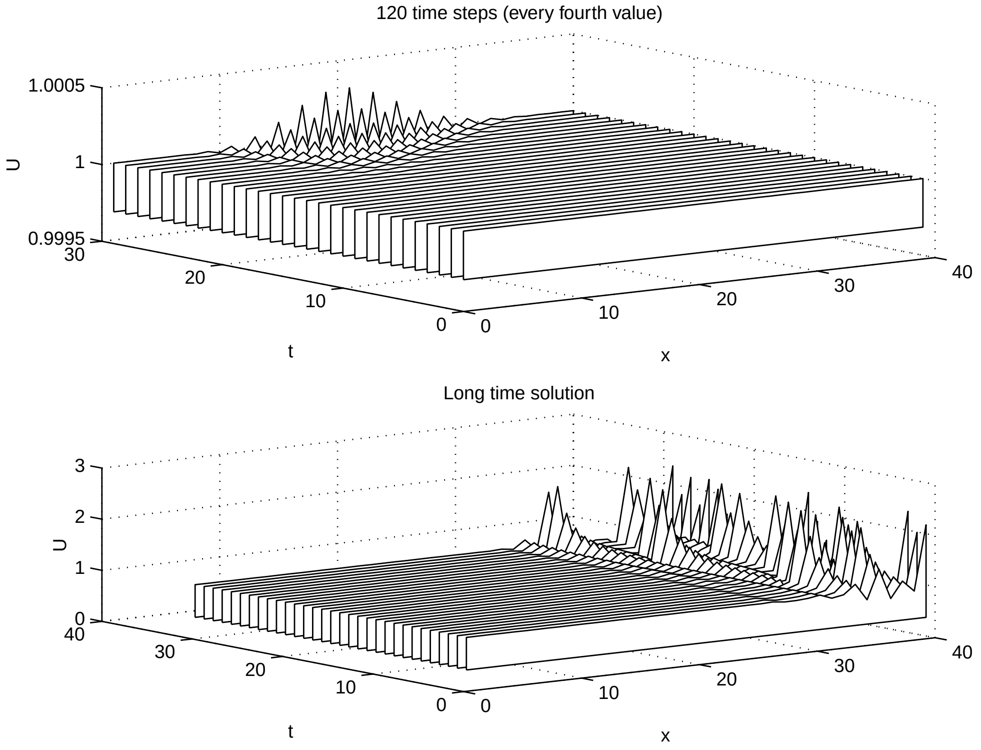


Fig. 5. *Spurious long-term behavior on the IBVP caused by nonnormality.*

approximation $\widehat{\mathbf{U}}^{n+1}$ at time-level $n+1$. We then form the weights

$$w_i^{n+1} := \sqrt{1-\mu+\mu(d_i^{n+1})^2},$$

which involve the gradient approximations

$$d_i^{n+1} := \frac{\widehat{U}_{j+1}^{n+1} - \widehat{U}_j^{n+1}}{\Delta x_i^n}$$

and the parameter $\mu \in [0,1]$. The new mesh $\{\Delta x_i^{n+1}\}$ is found by setting $w_i \Delta x_i^{n+1}$ equal for all i. This represents a linear system to be solved for the new mesh. The final numerical solution $\mathbf{U}^{n+1}$ is obtained by piecewise cubic interpolation from the original data $\{x_i^n, \widehat{\mathbf{U}}^{n+1}\}$ to the new mesh $\{x_i^{n+1}\}$. The motivation for this algorithm is that the weights measure the "sharpness" of the solution, and so the sharpness per unit interval is fixed. In this way, the algorithm aims to place more meshpoints in regions of x where there is rapid spatial change in the solution. The choice $\mu = \frac{1}{2}$ equidistributes a discretization of the arclength, while $\mu = 0$ gives a uniform mesh.

In experimenting with this algorithm, we fixed $a = 1$ and at each time-level chose a value $0.1 \min_i\{\Delta x_i^n\}$ for the time-step. In this way the Courant number is bounded above by 0.1.

It is immediately clear that the equidistribution algorithm described above cannot eliminate spatially uniform, spurious fixed points. (Inserting a SUFP $\mathbf{U}^n = \beta \mathbf{e}$, with $\Delta x_i^n \equiv 1/N$, we see that the solution, the spatial mesh, and the time-step are unchanged at the next time-level.) Analyzing the *stability* of such a fixed point is a complicated process, since the solution, the spatial mesh, and the time-step may be perturbed. Also, the stability potentially depends upon the fine-details of the algorithm, such as the choice of μ and the type of interpolation process used.

In our tests, we found that some of the spatially uniform spurious fixed points that we identified for a fixed spatial mesh in section 3.2 were stable as solutions for the equidistribution algorithm. Hence, the algorithm can be made to compute spurious fixed points. However, because of the Courant number control, this behavior only arose when the initial data was close to the uniform spurious level. For other types of initial data, spatial variation of the solution caused a reduction in $\min_i\{\Delta x_i^n\}$, and hence, in the time-step, and this in turn forced the numerical solution away from the spurious level.

We also found, as we intuitively expected, that the equidistribution algorithm eliminated spurious solutions with “jumps” of the form illustrated in Figures 3 and 4.

In summary, although the spatial mesh movement offered an improvement, it did not completely eliminate the potential for spatially uniform spurious fixed points.

5.2. Adaptivity in time. We now consider adaptation of the time-step as a means to eliminate spurious behavior. In the context of ODE solvers, this issue has been looked at in [1], and positive results have been obtained about the benefit of error control.

We will analyze the standard approach of using an embedded pair of formulas for the system (6). We assume that the spatial mesh remains uniform. The second order RK formula in (9) is coupled with Euler’s method, which is first order, to give

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \Delta t^n \left[(1-\theta)S(\mathbf{U}^n) + \theta S\left(\mathbf{U}^n + \frac{\Delta t^n}{2\theta}S(\mathbf{U}^n)\right)\right], \tag{36}$$

$$\mathbf{U}_{\mathrm{E}}^{n+1} = \mathbf{U}^n + \Delta t^n S(\mathbf{U}^n). \tag{37}$$

The secondary formula (37) is used only in the estimation of the error and selection of a new time-step.

The step is regarded as acceptable if

$$\frac{\| \mathbf{U}^{n+1} - \mathbf{U}_{\mathrm{E}}^{n+1} \|_\infty}{\Delta t^n} \leq \tau, \tag{38}$$

where $\tau > 0$ is the error tolerance. If the error criterion (38) is not satisfied, then the step is retaken with a smaller Δt^n. The details of how the time-step is altered after successful or rejected steps will not affect our conclusions. Also, we mention that changing the norm in (38) or using the error-per-step alternative $\| \mathbf{U}^{n+1} - \mathbf{U}_{\mathrm{E}}^{n+1} \|_\infty \leq \tau$ would not impact the results greatly.

We first prove a result that holds when the RK pair (36)–(37) is used to solve a general ODE system. We show that any fixed point cannot be genuinely spurious, in the sense that the residual is bounded by the error tolerance.

Lemma 5.1. *Suppose that the pair* (36)–(37) *subject to the error criterion* (38) *is used to solve a general ODE system* $\mathbf{U}_t = S(\mathbf{U})$*. If* $\mathbf{U}^n = \mathbf{U}^{n+1} = \mathbf{U}^*$*, then* $\| S(\mathbf{U}^*) \|_\infty \leq \tau$*.*

Proof. If $\mathbf{U}^n = \mathbf{U}^{n+1} = \mathbf{U}^*$, then from (36)

$$S(\mathbf{U}^*) = \theta \left[S(\mathbf{U}^*) - S\left(\mathbf{U}^* + \frac{\Delta t^n}{2\theta}S(\mathbf{U}^*)\right)\right]. \tag{39}$$

However, the error criterion (38) ensures that

$$\left\| \theta \left[S(\mathbf{U}^*) - S\left(\mathbf{U}^* + \frac{\Delta t^n}{2\theta} S(\mathbf{U}^*)\right)\right]\right\|_\infty \leq \tau. \tag{40}$$

Combining (39) and (40) gives the result. □

This result is closely related to those proved for regular RK pairs in [8]. In fact, (36)–(37) is a very special case of a regular pair and the result is therefore stronger than those in [8]. (In particular the residual bound does not involve a Lipschitz constant of S.)

Lemma 5.1 translates immediately into a result about SUFPs of the PIVP. However, the following result is more widely applicable.

THEOREM 5.2. *Suppose that the pair* (36)–(37) *subject to the error criterion* (38) *is used for either the IBVP or PIVP. If, for some* $j \geq 2$, $U_{j-1}^n = U_j^n = U_j^{n+1} = \beta$, *then* $|g(\beta)| \leq \tau$.

Proof. Since $U_j^{n+1} = U_j^n$, we have

$$(1-\theta)S(\mathbf{U}^n)_j + \theta S\left(\mathbf{U}^n + \frac{\Delta t^n}{2\theta}S(\mathbf{U}^n)\right)_j = 0.$$

Hence,

$$S(\mathbf{U}^n)_j = \theta\left[S(\mathbf{U}^n)_j + S\left(\mathbf{U}^n + \frac{\Delta t^n}{2\theta}S(\mathbf{U}^n)\right)_j\right].$$

So the error criterion (38) ensures that

$$|\, S(\mathbf{U}^n)_j \,| \leq \tau. \tag{41}$$

Since $U_j^n = U_{j-1}^n = \beta$ and $j \geq 2$, using the form of S we find

$$S(\mathbf{U}^n)_j = g(\beta). \tag{42}$$

The result follows from (41) and (42). □

Note that this result requires only a very weak type of spatial uniformity and spuriosity: a single component must be constant locally across space and time. Hence, in addition to ruling out SUFPs on the PIVP, this type of error control is also guaranteed to eliminate "jagged" spurious fixed points of the type illustrated in Figure 4.

In our experiments with the tolerance set to $\tau = \Delta x$, no spurious fixed points were encountered.

We remark that the design of error control strategies for time-dependent PDEs involves many considerations, in particular with regard to the balance of spatial and temporal errors. Theorem 5.2 illustrates that a simple ODE-based approach offers immediate advantages from a spuriosity perspective. We are currently investigating the use of more sophisticated adaptive procedures.

6. Nonlinear flux.

We briefly discuss how the results obtained in the previous sections extend to the more general conservation law (1) with nonlinear flux. The semidiscrete system (6) takes the form

$$\mathbf{U}_t = -\tfrac{1}{\Delta x} A\,\mathbf{f}(\mathbf{U}) + \mathbf{g}(\mathbf{U}) + \tfrac{1}{\Delta x}\mathbf{b},$$

where $\mathbf{f}(\mathbf{U})_j = f(U_j)$. The matrix A given by (7) or (8) is based on upwind differencing and it is therefore appropriate to assume that

$$\infty > f'(u) \geq a > 0.$$

The assumption that $f'(u)$ is bounded strictly away from zero is imposed to avoid degeneracy at fixed points β, where $f'(\beta) = 0$. The results of section 2 on existence of SUFPs may then be generalized by replacing g by the composition $ag \circ f^{-1}$ (so that g' becomes ag'/f').

The Jacobian of the fully discrete system may, in this more general setting, be defined by

$$G'(\beta\mathbf{e}) = I + \Delta t\left((1-\theta)F'(\beta) + \theta F'(z)\right) + \frac{1}{2}\Delta t^2 F'(\beta)F'(z),$$

where $F'(u) = g'(u)I - (1/\Delta x)f'(u)A$. This reduces to (16) when $f(u) = au$. Our main stability result, Theorem 3.2, then extends naturally and the condition (24) is replaced by

$$\Delta x^m < \frac{\lambda f'(z)}{g'(z)}.$$

7. Summary. This work concerns spurious behavior of finite difference schemes, with emphasis on the new features that arise on moving from an ODE to a hyperbolic PDE model.

For the case of uniform meshes, the main results may be summarized as follows:

- Theorem 3.2 for the PIVP shows that, unlike in the ODE case, as the mesh is refined in a natural manner, spatially uniform spurious fixed points are generically unstable.
- The computations and analysis in section 3.4 show that nonuniform spurious fixed points arise naturally. In particular, Theorem 3.4 and the accompanying analysis highlights the existence of stable nonsmooth steady states that jump between levels.
- The discussion in section 4 shows how a different type of spurious solution arises on the IBVP, as a consequence of nonnormality. This behavior exists only in the presence of finite precision arithmetic.

Section 5 gives results for adaptive algorithms, where the mesh is varied to take account of the solution. An equidistribution algorithm for the spatial grid offers some help in the avoidance of spurious behavior but does not completely eliminate spatially uniform spurious fixed points. In the case of a standard ODE-based time-step control, Theorem 5.2 provides a rigorous bound on the level of spuriosity.

REFERENCES

[1] M. A. Aves, D. F. Griffiths, and D. J. Higham, *Does error control suppress spuriosity?*, SIAM J. Numer. Anal., 34 (1997), pp. 756–778.

[2] C. J. Budd, G. P. Koomullil, and A. M. Stuart, *On the solution of convection-diffusion boundary value problems using equidistributed grids*, SIAM J. Sci. Comput., 20 (1998), pp. 591–618.

[3] B. Engquist and B. Sjögreen, *Robust Difference Approximation of Stiff Inviscid Detonation Waves*, Tech. Rep. CAM 91-03, Department of Mathematics, UCLA, Los Angeles, CA, 1991.

[4] D. F. Griffiths and A. R. Mitchell, *Stable periodic bifurcations of an explicit discretization of a nonlinear partial differential equation in reaction diffusion*, IMA J. Numer. Anal., 8 (1988), pp. 435–454.

[5] D. F. Griffiths, A. M. Stuart, and H. C. Yee, *Numerical wave propagation in an advection equation with a nonlinear source term*, SIAM J. Numer. Anal., 29 (1992), pp. 1244–1260.

[6] D. F. Griffiths, P. K. Sweby, and H. C. Yee, *On spurious asymptotic numerical solutions of explicit Runge-Kutta methods*, IMA J. Numer. Anal., 12 (1992), pp. 319–338.

[7] E. Hairer, A. Iserles, and J. M. Sanz-Serna, *Equilibria of Runge-Kutta methods*, Numer. Math., 58 (1990), pp. 243–254.

[8] D. J. Higham, *Regular Runge–Kutta pairs*, Appl. Numer. Math., 25 (1997), pp. 229–241.

[9] D. J. Higham and B. Owren, *Non-normality effects in a discretised, nonlinear, reaction-convection-diffusion equation*, J. Comput. Phys., 124 (1996), pp. 309–323.

[10] A. R. Humphries, *Spurious solutions of numerical methods for initial value problems*, IMA J. Numer. Anal., 13 (1993), pp. 263–290.

[11] W. Hundsdorfer, B. Koren, M. van Loon, and J. G. Verwer, *A positive finite-difference advection scheme*, J. Comput. Phys., 117 (1995), pp. 35–46.

[12] A. Iserles, *Stability and dynamics of numerical methods for nonlinear ordinary differential equations*, IMA J. Numer. Anal., 10 (1990), pp. 1–30.

[13] A. Iserles, A. T. Peplow, and A. M. Stuart, *A unified approach to spurious solutions introduced by time discretisation. Part* I: *Basic theory*, SIAM J. Numer. Anal., 28 (1991), pp. 1723–1751.

[14] J. D. Lambert, *Numerical Methods for Ordinary Differential Systems*, Wiley, New York, 1991.

[15] R. J. Leveque and H. C. Yee, *A study of numerical methods for hyperbolic conservation laws with stiff source terms*, J. Comput. Phys., 86 (1990), pp. 187–210.

[16] T.-Y. Li and J. A. Yorke, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985–992.

[17] S. C. Reddy and L. N. Trefethen, *Stability of the method of lines*, Numer. Math., 62 (1992), pp. 235–267.

[18] L. Reichel and L. N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162–164 (1992), pp. 153–185.

[19] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial Value Problems*, 2nd ed., Wiley, New York, 1967.

[20] O. Stein, *Bifurcations of hyperbolic fixed points for explicit Runge–Kutta methods*, IMA J. Numer. Anal., 17 (1997), pp. 151–175.

[21] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth and Brooks/Cole, Belmont, CA, 1989.

[22] P. K. Sweby and H. C. Yee, *On the dynamics of some grid-adaption schemes*, in Proceedings of the Fourth International Conference on Numerical Grid Generation in CFD and Related Fields, Swansea, Wales, 1994, pp. 467–478.

[23] H. C. Yee, *A Class of High-Resolution Explicit and Implicit Shock-Capturing Methods*, NASA Technical Memorandum 101088, NASA Ames Research Center, Moffett Field, CA, 1989.

[24] H. C. Yee, P. K. Sweby, and D. F. Griffiths, *Dynamical systems approach study of spurious steady state numerical solutions of nonlinear differential equations.* 1. *The ODE connection and its implications for algorithm developments in computational fluid dynamics*, J. Comput. Phys., 97 (1991), pp. 249–310.