Non-Normality Effects in a Discretised Nonlinear Reaction-Convection-Diffusion Equation

Desmond J. Higham * and Brynjulf Owren †

Abstract

What is the long-time effect of adding convection to a discretised reaction-diffusion equation? For linear problems, it is well known that convection may denormalise the process, and, in particular, eigenvalue-based stability predictions may be over-optimistic. This work deals with a related issue—with a nonlinear reaction term, the non-normality can greatly influence the long-time dynamics. For a nonlinear model problem with Dirichlet boundary conditions, it is shown that the basin of attraction of the "correct" steady state can be shrunk in a directionally biased manner. A normwise analysis provides lower bounds on the basin of attraction and a more revealing picture is provided by pseudo-eigenvalues. In extreme cases, the computed solution can converge to a spurious, bounded, steady-state that exists only in finite precision arithmetic. The impact of convection on the existence and stability of spurious, periodic solutions is also quantified.

1 Introduction

Traditional convergence and linear stability analysis does not apply to long term numerical simulations of nonlinear evolutionary differential equations. It is now well-established that even when a time step is chosen according to a standard linearisation, the long term numerical solution may fail to converge to a true steady state, and may, instead, diverge, or settle down to a misleading, spurious value.

Several authors have recently looked at these phenomena, using techniques from dynamical systems theory. Analysis for autonomous ordinary differential equations can be found in [5, 8, 9, 10, 22]. Other authors [2, 4] have added a space dimension, and considered the corresponding reaction-diffusion equations. The main theme of this paper is to investigate these issues in the case where, in addition to diffusion, a

*Department of Mathematics and Computer Science, University of Dundee, Dundee, DD1 4HN, Scotland. (`na.dhigham@na-net.ornl.gov`). The work of this author was supported by the Engineering and Physical Sciences Research Council of the UK and by a joint research grant from the Research Council of Norway and the British Council.

†Department of Mathematical Sciences, The University of Trondheim, N-7034 Trondheim-NTH, Norway. (`bryn@imf.unit.no`). The work of this author was supported by a joint research grant from the Research Council of Norway and the British Council.

convection term is present. Our overall aim is to look at the effect of convection on the *existence* and *stability* of the true and spurious fixed points.

The potential denormalising effect of a convection term can dramatically influence the behaviour close to a true fixed point. In particular, a superficial eigenvalue test may be inappropriate, leading to over-optimistic predictions. In section 2 we perform a "local attractivity" analysis and show that the non-normality causes the corresponding lower bound on the basin of attraction to become tiny. In section 3, we introduce some alternative analysis, based on pseudo-eigenvalues, that allows qualitative predictions to be made about the largest practical time step. The non-normality can also lead to spurious steady states that do not exist in exact arithmetic, but arise in the presence of rounding errors. Section 4 provides examples and analysis of this behaviour. In section 5, we consider spurious periodic solutions that evolve when the time step is increased beyond the linear stability limit, and we show how the convection term can influence their stability.

In the remainder of this section, we introduce the differential equation and the corresponding discrete approximation, and set up some notation. We consider the reaction-convection-diffusion equation

$$u_t + a u_x = b u_{xx} + f(u), \quad 0 < x < 1, \quad t > 0, \tag{1.1}$$

with $a, b > 0$ constant, $f(u) = u(1-u)$, $u(x,0) = \phi(x)$ given, and with boundary conditions specified at $x = 0$ and $x = 1$. The two kinds of boundary condition that we consider will be introduced below.

A finite difference discretisation with constant space step Δx and time step Δt produces approximations $U_j^n \approx u(j\Delta x, n\Delta t)$. We use central differences for the diffusion term, and forward differences (Euler's Method) for the time derivative. For the convection term, we consider three possibilities: central, forward and backward differences.

Using central differences for u_x produces the formula

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = b\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2} + f(U_j^n). \tag{1.2}$$

Setting

$$r = b\Delta t/\Delta x^2, \quad \text{and} \quad \tilde{r} = a\Delta t/\Delta x, \tag{1.3}$$

we may write (1.2) as

$$U_j^{n+1} = U_j^n - \frac{\tilde{r}}{2}(U_{j+1}^n - U_{j-1}^n) + r(U_{j+1}^n - 2U_j^n + U_{j-1}^n) + \Delta t f(U_j^n). \tag{1.4}$$

This is an explicit scheme, which computes the approximations at time $(n+1)\Delta t$ from those at the previous time level.

Approximating the convection term by backward differences (upwinding) changes the scheme from (1.4) to

$$U_j^{n+1} = U_j^n - \tilde{r}(U_j^n - U_{j-1}^n) + r(U_{j+1}^n - 2U_j^n + U_{j-1}^n) + \Delta t f(U_j^n). \tag{1.5}$$

This may be re-written as

$$U_j^{n+1} = U_j^n - \frac{\tilde{r}}{2}(U_{j+1}^n - U_{j-1}^n) + (r + \frac{\tilde{r}}{2})(U_{j+1}^n - 2U_j^n + U_{j-1}^n) + \Delta t f(U_j^n), \tag{1.6}$$

which shows that using backward differences for u_x is equivalent to using central differences and adding *artificial diffusion* to the problem.

Similarly, a forward difference (downwind) approximation to u_x leads to the formula

$$U_j^{n+1} = U_j^n - \frac{\tilde{r}}{2}(U_{j+1}^n - U_{j-1}^n) + (r - \frac{\tilde{r}}{2})(U_{j+1}^n - 2U_j^n + U_{j-1}^n) + \Delta t f(U_j^n). \quad (1.7)$$

These finite difference schemes are standard, and their local accuracy and linear stability properties are analysed in many references; see, for example [14, 20, 21]. In this work we discuss phenomena that are common to all three choices, and hence we do not add to the debate about which version is "best".

It is convenient to write the schemes in matrix-vector form, using

$$U^n := \begin{bmatrix} U_1^n \\ U_2^n \\ \vdots \\ U_N^n \end{bmatrix} \in I\!R^N$$

to denote the numerical approximation at time level n. We concentrate on two types of boundary condition; periodic and (nonhomogeneous) Dirichlet. This is done to make the analysis as clean as possible. In the final section, we briefly discuss the likely effect of changing to other boundary conditions. With periodic boundary conditions, using $\Delta x = 1/N$, the schemes have the form

$$U^{n+1} = CU^n + \Delta t f(U^n), \quad (1.8)$$

where $C \in I\!R^{N\times N}$ is a circulant matrix of the form

$$C = \mathrm{circ}(d, e, 0, 0, \ldots, 0, c) := \begin{bmatrix} d & e & 0 & \ldots & 0 & c \\ c & d & e & 0 & \ddots & 0 \\ 0 & c & d & e & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & e \\ e & 0 & \cdots & 0 & c & d \end{bmatrix}, \quad (1.9)$$

and

$$f(U) = \begin{bmatrix} f(U_1) \\ f(U_2) \\ \vdots \\ f(U_N) \end{bmatrix}.$$

Table 1.1 gives the values of c, d, e that arise for the three different discretisations.

Imposing Dirichlet boundary conditions $u(0,t) = u(1,t) = 1$, with $\Delta x = 1/(N+1)$, leads to

$$U^{n+1} = TU^n + cp + eq + \Delta t\, f(U^n), \quad (1.10)$$

where $T \in I\!R^{N\times N}$ is a tridiagonal matrix of the form

$$T = \mathrm{tridiag}(c, d, e) := \begin{bmatrix} d & e & & \\ c & d & \ddots & \\ & \ddots & \ddots & e \\ & & c & d \end{bmatrix}, \quad (1.11)$$

Table 1.1: Matrix coefficients for the discretised problem.

u_x	c	d	e
Central	$r + \tilde{r}/2$	$1 - 2r$	$r - \tilde{r}/2$
Backward	$r + \tilde{r}$	$1 - 2r - \tilde{r}$	r
Forward	r	$1 - 2r + \tilde{r}$	$r - \tilde{r}$

with c, d, e defined in Table 1.1, and

$$p = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \quad q = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Note that iterations (1.8) and (1.10) share the fixed point $U_i^* \equiv 1$ (corresponding to the steady-state $u(x,t) \equiv 1$ of the continuous problem). This work concerns the attractiveness of this fixed point, and we will show that in the case of Dirichlet boundary conditions, convection can dramatically affect the behaviour. We tacitly assume that $u(x,t) \equiv 1$ is the "correct" solution. To justify this, the appendix gives conditions on the initial data under which the solution of the Dirichlet problem converges to $u(x,t) \equiv 1$.

We conclude this section with some notation. Given $A \in \mathbb{C}^{N \times N}$, $\rho(A)$ denotes the spectral radius:

$$\rho(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}.$$

Given $x \in \mathbb{C}^N$, the Euclidean and infinity vector norms are

$$\begin{aligned} \|x\|_2 &:= \left(\sum_{i=1}^{N} |x_i|^2 \right)^{1/2}, \\ \|x\|_\infty &:= \max_{1 \le i \le N} |x_i|, \end{aligned}$$

respectively. We note, for future reference, the inequality

$$\|x\|_\infty \le \|x\|_2 \le \sqrt{N} \|x\|_\infty. \tag{1.12}$$

The matrix norms induced by these vector norms will also be denoted $\|\cdot\|_2$ and $\|\cdot\|_\infty$. We recall that $\|A\|_2 = \rho(A)$ if $A \in \mathbb{C}^{N \times N}$ is normal. Throughout this work we implicitly assume that $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are "natural" norms with which to measure vectors and matrices.

For a nonsingular $D \in \mathbb{C}^{N \times N}$ we define the vector D-norm by

$$\|x\|_D = \|D^{-1} x\|_2. \tag{1.13}$$

The corresponding induced matrix norm satisfies

$$\|A\|_D = \|D^{-1} A D\|_2. \tag{1.14}$$

Given $\delta > 0$, $U \in I\!R^N$ and a vector norm $\|\cdot\|$, we let $B_{\|\cdot\|}(\delta, U)$ denote the open ball of radius δ around U; that is,

$$B_{\|\cdot\|}(\delta, U) := \{Y \in I\!R^N : \|Y - U\| < \delta\}.$$

2 Attractivity of the True Fixed Point

Given an iteration

$$U^{n+1} = G(U^n), \quad G : I\!R^N \to I\!R^N, \tag{2.1}$$

U^* is said to be a *fixed point*, or *steady state*, if $U^* = G(U^*)$. This section concerns local attractivity of fixed points and hence we make the following definitions.

Definition 2.1 *Given a fixed point U^* of (2.1), the* ***basin of attraction*** *of U^* is the set of all points U^0 such that the sequence defined by (2.1) satisfies $U^n \to U^*$ as $n \to \infty$.*

Definition 2.2 *A fixed point U^* of (2.1) is said to be* ***locally attractive*** *if it has a basin of attraction that contains an open neighbourhood of U^*.*

We continue with a somewhat heuristic discussion in order to motivate the main analysis.

Setting $U^0 = U^* + V^0$, where $V^0 \in I\!R^N$ is small, and linearising, we find that $U^1 = U^* + V^1$, where

$$V^1 \approx G'(U^*)V^0. \tag{2.2}$$

Here G' denotes the Jacobian of G. The relation (2.2) shows that after one iteration, the displacement from the fixed point is amplified by the Jacobian matrix, $G'(U^*)$. Continuing the linearisation, after m steps we have $U^m = U^* + V^m$, where

$$V^m \approx G'(U^*)^m V^0. \tag{2.3}$$

Hence, if the linearisation is valid, the local attractivity or repulsion of the fixed point is determined by the powers of the Jacobian matrix. It is well known that for any $A \in I\!R^{N\times N}$

$$A^m \to 0 \quad \text{as} \quad m \to \infty \quad \Longleftrightarrow \quad \rho(A) < 1. \tag{2.4}$$

This suggests that if $\rho(G'(U^*)) < 1$ then U^* will attract the iterates when U^0 is sufficiently close to U^*. This can, of course, be made rigorous—see Theorem 2.1 below. However, even when $A \in I\!R^{N\times N}$ satisfies $\rho(A) < 1$, if A is not normal then it is possible for the powers A^m to become arbitrarily large for some finite m. To illustrate this, we consider the iteration (1.10) with backward differencing, using $a = 1$ and $N = 32$ (so that $\Delta x = 1/(N+1) = 1/33$). We let the diffusion coefficient, b, range over the values $\{10^{-2}, 10^{-2.5}, 10^{-3}, 10^{-3.5}\}$. For each b we choose a time step Δt so that $\rho(G'(U^*)) \leq .85$ at the fixed point $U_i^* \equiv 1$. (More precisely, we take $\Delta t = .9\Delta t_{\mathrm{lim}}$, where Δt_{lim} is the linear time step limit defined in (2.24).) In Figure 2.1 we plot the power norms $\|G'(U^*)^m\|_\infty$ against m. In the figure, the maximum height of the curve increases as b decreases. Hence, although the powers ultimately decay in each case, as the convection term becomes more dominant the level of intermediate growth becomes considerable. With $b = 10^{-3.5}$, some powers of the Jacobian are larger than 10^{14}. In such a case we should be very wary of the linearisation process. Perhaps V^0 is sufficiently small for (2.2) to be a reasonable approximation, but, after several iterations, $G'(U^*)^m V^0$ in (2.3) could be many orders of magnitude larger, and hence U^m could be far from U^*. In general, V^0 would have to be *extremely* small in order for each subsequent V^m to be small. Hence, although the condition $\rho(G'(U^*)) < 1$ guarantees local attractivity, if the powers of $G'(U^*)$ become large, then we can expect the basin of attraction to be tiny. We emphasise that this is a linear phenomenon in the sense that it is caused by the Jacobian, and does not require any particular type of nonlinearity in G.

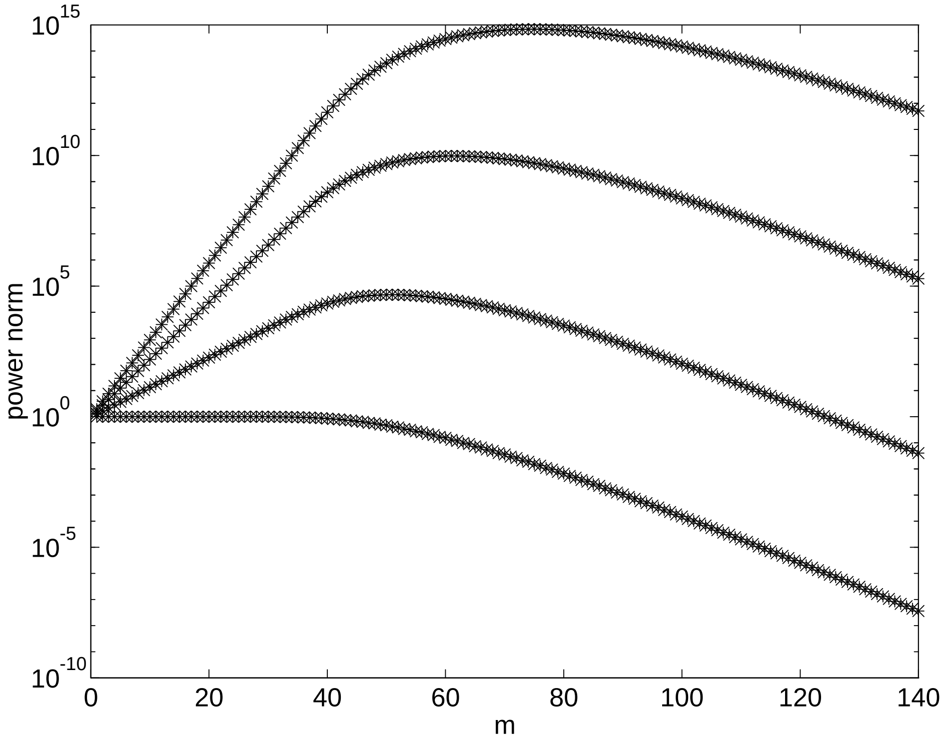

Figure 2.1: Power norms $\|G'(U^*)^m\|_\infty$ against m, on the Dirichlet problem with backward differencing. The maximum height increases as the diffusion coefficient b decreases.

The fact that convection-dominated problems can produce non-normal matrices whose powers become large has been widely observed in the context of linear stability for discretised partial differential equations. This corresponds to examining the long-term behaviour of finite difference solutions to (1.1) when $f(u) \equiv 0$ or $f(u) = -u$. Here, a condition of the form $\rho(A) < 1$ guarantees eventual decay of the solution, but, to control intermediate growth, it is also necessary for A^m to be bounded above by a reasonable constant for all $m \geq 0$. The inadequacy of the spectral radius condition has been appreciated for some time; see, for example, [1, section 10.6] and [3, 14, 20]. A thorough treatment for the central difference discretisation can be found in [7].

Trefethen and co-workers [6, 17, 18] have also examined non-normality effects. The pseudo-eigenvalue analysis developed in these references will be discussed in the next section.

We now introduce Ostrowski's Theorem, which is essentially a formalisation of the linearisation process. We include a proof of the theorem, since this will be used later. For simplicity we assume that $G'(U^*)$ can be made normal by a similarity transformation. This is the case, of course, if $G'(U^*)$ is diagonalisable. (If this assumption is removed, then the theorem remains true if the condition $\sigma + \epsilon < 1$ for (2.5) is tightened to $\sigma + 2\epsilon < 1$.)

Theorem 2.1 *(Ostrowksi) [15, Theorem 10.1.3] Let U^* be a fixed point of the iteration (2.1), and suppose that G is Fréchet-differentiable at U^*. Suppose further that a nonsingular matrix $D \in \mathbb{C}^{N\times N}$ exists for which $D^{-1}G'(U^*)D$ is normal.*

If $\sigma := \rho(G'(U^*)) < 1$, then U^* is locally attractive. Furthermore, the basin of attraction of U^* contains the ball $B_{\|\cdot\|_D}(\delta, U^*)$, where $\delta > 0$ is such that

$$\|G(U) - G(U^*) - G'(U^*)(U - U^*)\|_D \le \epsilon\|U - U^*\|_D, \quad \forall U \in B_{\|\cdot\|_D}(\delta, U^*), \tag{2.5}$$

with $\sigma + \epsilon < 1$.

Proof. Since $D^{-1}G'(U^*)D$ is normal, we have

$$\|G'(U^*)\|_D = \sigma. \tag{2.6}$$

Because G is Fréchet-differentiable at U^*, given any $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that

$$\|G(U) - G(U^*) - G'(U^*)(U - U^*)\|_D \le \epsilon\|U - U^*\|_D, \quad \forall U \in B_{\|\cdot\|_D}(\delta, U^*). \tag{2.7}$$

Hence, for any $U^0 \in B_{\|\cdot\|_D}(\delta, U^*)$,

$$\begin{aligned}
\|U^1 - U^*\|_D &= \|G(U^0) - U^*\|_D \\
&\le \|G(U^0) - G(U^*) - G'(U^*)(U^0 - U^*)\|_D + \|G'(U^*)\|_D\|U^0 - U^*\|_D \\
&\le (\sigma + \epsilon)\|U^0 - U^*\|_D.
\end{aligned}$$

Choosing ϵ so that $\sigma + \epsilon = r < 1$, and repeating this argument on each iteration, we see that

$$U^n \in B_{\|\cdot\|_D}(\delta, U^*), \quad \forall n$$

and

$$\|U^n - U^*\|_D \le r^n\|U^0 - U^*\|_D \to 0, \quad \text{as} \quad n \to \infty,$$

and the result follows. ∎

We are concerned with iterations of the form

$$U^{n+1} = MU^n + \Delta t f(U^n) + y, \tag{2.8}$$

with $M \in I\!R^{N\times N}$, $f(u) = u(1-u)$ and $y \in I\!R^N$. We suppose that there is a nonsingular diagonal matrix D such that $D^{-1}G'(U^*)D$ is normal. The following corollary applies Ostrowski's Theorem to this case.

Corollary 2.1 *Suppose that the iteration (2.8) has a locally attractive fixed point* U^**, with* $\rho(G'(U^*)) =: \sigma < 1$*. Suppose further that there exists a nonsingular diagonal matrix* $D \in \mathbb{C}^{N\times N}$ *such that* $D^{-1}G'(U^*)D$ *is normal. Then the basin of attraction of this fixed point contains the ball* $B_{\|\cdot\|_D}(\delta, U^*)$ *for any* $\delta < (1-\sigma)/(\Delta t\sqrt{N}\|D\|_\infty)$*.*

Proof. The Jacobian of the map in (2.8) is

$$G'(U) = M + \Delta t \mathrm{diag}\,(f'(U_1), f'(U_2), \ldots, f'(U_N)). \tag{2.9}$$

In the expression $G(U) - G(U^*) - G'(U^*)(U-U^*)$, the linear part of G disappears, and we find that

$$[G(U) - G(U^*) - G'(U^*)(U-U^*)]_j = \Delta t[f(U_j) - f(U_j^*) - f'(U_j^*)(U_j - U_j^*)]. \tag{2.10}$$

Now, a Taylor expansion gives

$$f(U_j) - f(U_j^*) - f'(U_j^*)(U_j - U_j^*) = \frac{(U_j - U_j^*)^2}{2} f''(\theta_j), \tag{2.11}$$

for some θ_j between U_j and U_j^*. In our case, f is quadratic with $f'' \equiv -2$ (constant) and so, in (2.10),

$$[G(U) - G(U^*) - G'(U^*)(U-U^*)]_j = -\Delta t(U_j - U_j^*)^2. \tag{2.12}$$

The expression (2.12) is a componentwise result. In order to make use of Theorem 2.1 we must convert this to a normwise bound of the form (2.7). We use the D-norm, for which

$$\|G'(U^*)\|_D = \rho(G'(U^*)),$$

since $D^{-1}G'(U^*)D$ is normal. For convenience, write $v_j = [G(U) - G(U^*) - G'(U^*)(U-U^*)]_j$ and $w_j = U_j - U_j^*$, and let w^2 denote the vector with jth component equal to w_j^2. With this notation, (2.12) becomes $v_j = -\Delta t w_j^2$. Now, using the inequality (1.12) and exploiting the fact that D is diagonal, we have

$$\begin{aligned}
\|v\|_D &= \Delta t\|D^{-1}w^2\|_2 \\
&= \Delta t\|D(D^{-1}w)^2\|_2 \\
&\leq \Delta t\sqrt{N}\|D(D^{-1}w)^2\|_\infty \\
&\leq \Delta t\sqrt{N}\|D\|_\infty\|D^{-1}w\|_\infty^2 \\
&\leq \Delta t\sqrt{N}\|D\|_\infty\|D^{-1}w\|_2^2 \\
&= \Delta t\sqrt{N}\|D\|_\infty\|w\|_D^2.
\end{aligned}$$

It follows that (2.5) is satisfied with $\delta = \epsilon/(\Delta t\sqrt{N}\|D\|_\infty)$, where ϵ is any number smaller than $1-\sigma$. ∎

We remark that our choice of nonlinear reaction term $f(u) = u(1-u)$ leads to the simple relation (2.12). However, it is clear from (2.10) that the analysis can be adapted to any $f(u)$ for which a suitable second derivative bound is available.

The iterations (1.8) and (1.10) possess fixed points with $U_i^* \equiv 1$. For the Dirichlet case, (1.10), the Jacobian at the fixed point is tridiagonal;

$$G'(U^*) = \mathrm{tridiag}(c, d - \Delta t, e). \tag{2.13}$$

The eigenvalues are

$$d - \Delta t + 2\sqrt{ce}\cos\left(\frac{l\pi}{N+1}\right), \quad l = 1, 2, \ldots, N. \tag{2.14}$$

Letting

$$\alpha := \sqrt{c/e} \qquad \text{and} \qquad D := \mathrm{diag}(1, \alpha, \alpha^2, \ldots, \alpha^{N-1}), \tag{2.15}$$

the similarity transformation

$$G'(U^*) \mapsto D^{-1}G'(U^*)D = \mathrm{tridiag}(\sqrt{ce}, d - \Delta t, \sqrt{ce}) \tag{2.16}$$

produces a symmetric, and hence normal, matrix. Taking the values of c, d and e from Table 1.1, the corresponding value for α is displayed in Table 2.1, in terms of the *grid Péclet number*, P, defined as

$$P := \frac{a\Delta x}{2b}. \tag{2.17}$$

Note that for a fixed Δx, P increases with the convection coefficient a, but for a fixed problem, $P \to 0$ as $\Delta x \to 0$.

Table 2.1: Values of α for the Dirichlet problem.

u_x	α
Central	$\sqrt{(1+P)/(1-P)}$
Backward	$\sqrt{1+2P}$
Forward	$\sqrt{1/(1-2P)}$

Since D is diagonal in (2.15), we have $\|D\|_\infty = \max\{1, |\alpha|^{N-1}\}$. We see from Table 2.1 that for central and backward differences, $|\alpha| > 1$ for all parameter values. It is possible for $|\alpha|$ to be very large. With central differences, this happens when P is close to 1. (When $P = 1$ the Jacobian becomes defective, and Corollary 2.1 does not apply.) A similar effect in the context of linear stability is observed in [7]. With backward differences, α increases monotonically with P. In the case of forward differences, $|\alpha|$ is largest when P is close to 1/2.

When $|\alpha| \geq 1$, it follows from Corollary 2.1 that $U_i^* \equiv 1$ will attract any initial condition for which

$$\|\mathrm{diag}(1, \alpha^{-1}, \ldots, \alpha^{-N+1})(U^0 - U^*)\|_2 < (1-\sigma)/(\Delta t\sqrt{N}|\alpha|^{N-1}),$$

or equivalently

$$\|\mathrm{diag}(\alpha^{N-1}, \alpha^{N-2}, \ldots, 1)(U^0 - U^*)\|_2 < (1-\sigma)/(\Delta t\sqrt{N}). \tag{2.18}$$

If $|\alpha|$ is large, then the exponential factors in (2.18) can be very significant. In this case the inequality is most strict for components with small index. In particular, if only a single component differs from U^*, then (2.18) requires

$$|(U^0 - U^*)_1| < (1-\sigma)/(\Delta t\sqrt{N}|\alpha|^{N-1}) \tag{2.19}$$

for the first component, but only

$$|(U^0 - U^*)_N| < (1-\sigma)/(\Delta t\sqrt{N}) \tag{2.20}$$

for the last component. However, since $\Delta x = 1/(N+1)$, it follows that, for a fixed problem, $\alpha^{N-1} \to \exp(a/b)$ as $N \to \infty$ with central differences. Similarly, for either backward or forward differences, $\alpha^{N-1} \to \exp(a/(2b))$. Hence, the non-normality effect of the convection term reaches a limit as the discretisation becomes finer.

In the case of periodic boundary conditions, the Jacobian for (1.8) at $U_i^* \equiv 1$ is

$$G'(U^*) = \mathrm{circ}(d - \Delta t, e, 0, 0, \ldots, 0, c). \tag{2.21}$$

The matrix is circulant, and hence normal, so we can take D to be the identity matrix. Corollary 2.1 then shows that the basin of attraction for U^* contains the region where

$$\|U^0 - U^*\|_2 < (1-\sigma)/(\Delta t\sqrt{N}). \tag{2.22}$$

In general (2.22) is much less stringent than (2.18), and for a given σ it is independent of the convection and diffusion parameters.

It is important to note that Corollary 2.1 constructs a region that is *contained* in the basin of attraction; in other words, it provides a *lower bound* on the size of the basin of attraction. For the Dirichlet problem, the region constructed depends heavily upon the amount of convection, and in particular, its shape can be highly distorted when compared with a ball in a "natural" norm such as $\|\cdot\|_2$ or $\|\cdot\|_\infty$. For the periodic problem, the region constructed is independent of the amount of convection, and is not directionally biased.

We now present some numerical experiments to test the relevance of this analysis. We begin with the Dirichlet problem, using backward differences (upwinding) for u_x. Here $c > e > 0$ in Table 1.1 and the eigenvalues in (2.14) are real. It follows that as Δt increases away from zero, the local attractivity condition $\rho(G'(U^*)) < 1$ is first violated when an eigenvalue reaches -1. With all other parameters fixed, the time step at which $\rho(G'(U^*)) < 1$ fails is given by

$$\frac{1 - \cos(N\pi/(N+1))}{1 + \frac{2b}{\Delta x^2} + \frac{a}{\Delta x} + 2\sqrt{\frac{b}{\Delta x^2}\left(\frac{b}{\Delta x^2} + \frac{a}{\Delta x}\right)}}. \tag{2.23}$$

For large N, $\cos(N\pi/(N+1)) \approx -1$, so we follow the usual convention [14, page 43] of regarding the time step limit as

$$\Delta t_{\mathrm{lim}} := \frac{2}{1 + \frac{2b}{\Delta x^2} + \frac{a}{\Delta x} + 2\sqrt{\frac{b}{\Delta x^2}\left(\frac{b}{\Delta x^2} + \frac{a}{\Delta x}\right)}}. \tag{2.24}$$

From Ostrowski's Theorem, any time step $\Delta t \leq \Delta t_{\text{lim}}$ will therefore make U^* locally attractive.

Our policy was to fix $N = 32$ (giving $\Delta x = 1/33$) and $a = 1$. We considered a range of diffusion coefficients with $\log_{10} b = \{-2, -2.25, -2.5, \ldots, -3.25\}$, and in each case we chose $\Delta t = .9\Delta t_{\text{lim}}$, which made the fixed point locally attractive. The spectral radius $\rho(G'(U^*))$ was less than .83 in each case. (Note that the parameters are within the range used for Figure 2.1.) For the first set of tests, we used initial conditions of the form $[1+\gamma, 1, 1, \ldots, 1]^T$. Starting with $\gamma = 10^{-16}$ we performed the iteration for 1000 steps and tested whether $U^n \to U^*$. We continued this process, increasing γ by a factor of 10 until we reached a level where the iteration failed to converge. Convergence was deemed to have occurred if

$$\|U^{1000} - U^*\|_\infty < 10^{-5}. \tag{2.25}$$

Hence, we recorded γ^1_{max}, which we define to be the perturbation beyond which (2.25) fails. Our convergence condition was determined after some experimentation. In the tests, we found that if the iterates failed to converge to U^* then they became unbounded (rather than converging to some other finite steady state). However, if b is reduced to $10^{-3.5}$ then the "cut-off" between convergence and non-convergence is less sharp. Here U^n can settle down to a state that is close to, but significantly different from U^*. This behaviour is caused by rounding errors, and will be discussed in section 4.

Table 2.2 gives the results. The table also gives the value of α in (2.15) and the bound (2.19) on $|\gamma|$ for which convergence is guaranteed. The corresponding results that arise with initial conditions of the form $[1, 1, \ldots, 1, 1+\gamma]^T$ are also presented in the table. Here, γ^N_{max} is defined in an analogous way to γ^1_{max}.

Table 2.2: Attractivity for the Dirichlet problem with backward differencing.

	$\log_{10} b$					
	-2	-2.25	-2.5	-2.75	-3	-3.25
γ^1_{max}	1e+1	1e+0	1e-3	1e-6	1e-8	1e-11
Bound from (2.19)	8e-10	5e-13	2e-16	4e-20	6e-24	9e-28
α	2.0	2.5	3.3	4.2	5.6	7.4
$\alpha^N \gamma^1_{\text{max}}$	2e+10	3e+12	8e+12	3e+13	2e+15	9e+15
γ^N_{max}	1e+1	1e+1	1e+1	1e+1	1e+1	1e+1
Bound from (2.20)	2.0	1.5	1.3	1.1	0.9	0.8

In Table 2.2 we see that the lower bounds provided by Corollary 2.1 can be pessimistic. However, the bounds do capture two important features of the actual results.

- The component U_1 is sensitive to the relative amount of convection, whilst U_N is not.

- The size of γ^1_{max} decreases approximately like α^{-N}, as does the lower bound. (The product $\alpha^N \gamma^1_{\text{max}}$ is included in the table.)

Next we consider the periodic problem, with backward differences. In this case the Jacobian (2.21) has eigenvalues

$$(c+e)\cos(\theta_l) + d - \Delta t + i(e-c)\sin(\theta_l), \quad \theta_l = \frac{l\pi}{N+1}, \quad l = 1, 2, \ldots, N. \quad (2.26)$$

Hence the lth eigenvalue, λ_l, satisfies

$$|\lambda_l|^2 = 4ce\cos^2(\theta_l) + 2(c+e)(d-\Delta t)\cos(\theta_l) + (d-\Delta t)^2 + (e-c)^2.$$

Since $ce > 0$, the maximum modulus occurs at an extreme eigenvalue, where $l = 1$ or N. As for the Dirichlet case, it is reasonable to majorise over the continuous range $\theta \in [0, \pi]$, giving the eigenvalue-based time step limit

$$\Delta t^*_{\mathrm{lim}} := \frac{2}{1 + \frac{4b}{\Delta x^2} + \frac{2a}{\Delta x}}. \quad (2.27)$$

Hence, from Ostrowski's Theorem, $\Delta t \leq \Delta t^*_{\mathrm{lim}}$ makes U^* locally attractive. We repeated the tests described above, using the same values for N, a and b, and with $\Delta t = .9\Delta t^*_{\mathrm{lim}}$. The spectral radius of the Jacobian was less than .985 in each case. We found that the analogues of γ^1_{max} and γ^N_{max} were always equal to 10; they did not vary with b. In this case, the lower bound (2.22) is within two orders of magnitude of the actual largest allowable perturbation, and captures the insensitivity to the relative amount of convection.

3 Pseudo-Eigenvalue Analysis

At the start of the previous section we showed that the local attractivity of a fixed point is closely related to the behaviour of the powers of the Jacobian. The examples that we tested involved matrices A for which $D^{-1}AD = N$, with N normal. It follows that $A^m = DN^mD^{-1}$. Letting $\sigma = \rho(A)$, we have $\|N^m\|_2 \leq \|N\|_2^m = \sigma^m$, and so

$$\|A^m\|_2 \leq \kappa_2(D)\sigma^m, \quad (3.1)$$

where $\kappa_2(D) := \|D^{-1}\|_2\|D\|_2$ is the (normwise) condition number of D. For the diagonal D in (2.15), assuming $|\alpha| \geq 1$, we have $\kappa_2(D) = |\alpha|^{N-1}$. Although (3.1) establishes that $\|A^m\|_2 \to 0$ as $m \to \infty$ when $\sigma < 1$, as a bound on $\|A^m\|_2$ it can be very pessimistic. This type of norm-based inequality lies at the heart of the analysis in section 2, and, as the numerical examples indicated, it can lead to over-conservative predictions. Overall, if we wish to know how big $\|A^m\|_2$ can become, then neither $\rho(A)$ nor $\kappa_2(D)$ can provide sharp estimates when A is far from normal. Trefethen and Reddy [17, 18] recognised this state of affairs and have put forward an alternative style of analysis, based on pseudo-eigenvalues, that helps to fill the gap between the spectral and normwise approaches. In this section we describe the relevant ideas and apply them to our iteration (1.10). We mention that in some contexts, particularly for discretisations of certain hyperbolic equations, the pseudospectral approach provides extremely clear-cut results in the limit as $\Delta x \to 0$ ($N \to \infty$). In our case, as $\Delta x \to 0$ the symmetric $b/\Delta x^2$ contribution ¿from the diffusion overwhelms the $a/\Delta x$ convection component, causing the relevant matrices to become more normal as $N \to \infty$. Consequently, our conclusions are less precise, and apply to the case where N is large, but where the convection process is still dominant.

We begin by defining the ϵ-pseudospectrum.

Definition 3.1 *Given $A \in \mathbb{C}^{N\times N}$ and $\epsilon > 0$, the **ϵ-pseudospectrum** of A is the set*

$$\Lambda_\epsilon(A) := \{z \in \mathbb{C} \,:\, z \text{ is an eigenvalue of } A+E, \text{ with } E \in \mathbb{C}^{N\times N} \text{ and } \|E\|_2 \le \epsilon\}. \tag{3.2}$$

*A number $\lambda_\epsilon \in \Lambda_\epsilon(A)$ is called an **ϵ-pseudo-eigenvalue** of A.*

Figure 3.1 plots some pseudo-eigenvalues for the Jacobian of (1.10) at $U_i^* = 1$, using backward differences with parameters $N = 32$, $a = 1$, $\Delta t = .9\Delta t_{\mathrm{lim}}$ and $b = 10^{-3}$. Note that this matrix was also used in the the power-norm plots of Figure 2.1. To produce Figure 3.1 we generated matrices $\widehat{E}$ with entries whose real and imaginary parts came from a Normal(0, 1) distribution, and then took $E = \epsilon\widehat{E}/\|\widehat{E}\|_2$. The outer 'ring' in the figure is generated from five perturbations with $\epsilon = 10^{-5}$. Similarly, the central and inner rings come from $\epsilon = 10^{-7}$ and 10^{-9}, respectively. The eigenvalues, which lie on the real axis, are marked with circles. It is clear from the figure that the spectrum is very sensitive to perturbations, and $\Lambda_{10^{-9}}(A)$ contains points outside the unit disk.

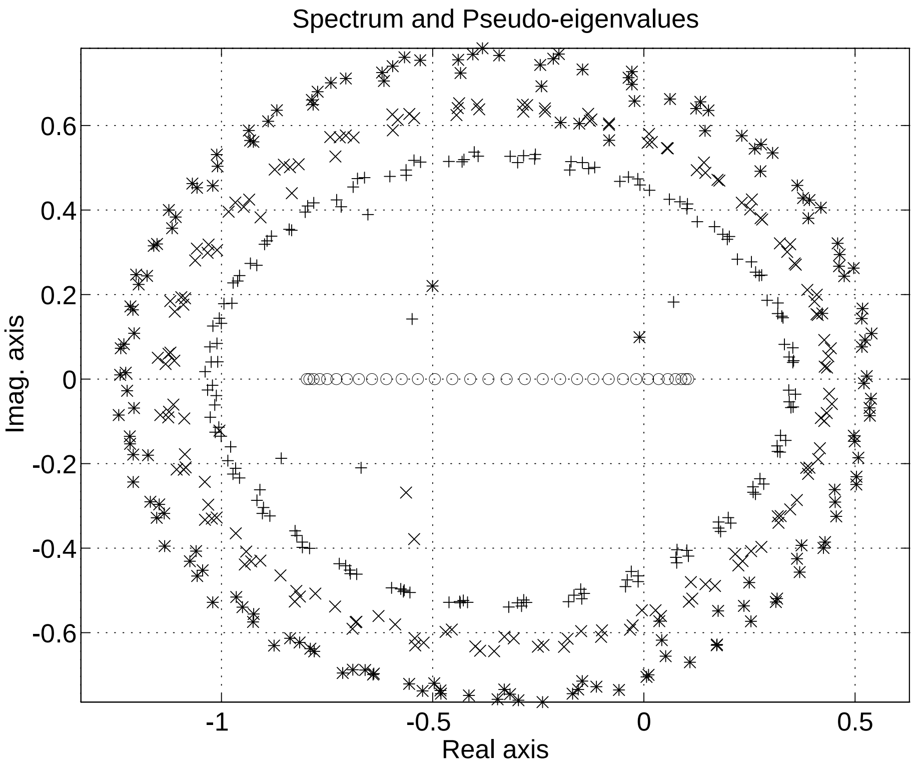

Figure 3.1: Spectrum and pseudo-eigenvalues, with $\epsilon = 10^{-5}, 10^{-7}, 10^{-9}$, of the Jacobian for a Dirichlet problem with backward differences.

We let D denote the open unit disk in the complex plane,

$$D := \{z \in \mathbb{C} \,:\, |z| < 1\},$$

and, given a point $z \in \mathbb{C}$ and a set $X \subseteq \mathbb{C}$, the distance function is defined by

$$\mathrm{dist}(z, X) := \inf\{\|z - x\|_2 \,:\, x \in X\}.$$

Now we give a connection between matrix powers and pseudo-eigenvalues. This result is taken from [18]; an earlier version appears in [17]. The theorem can be regarded as a sharpened version of the Kreiss Matrix Theorem.

Theorem 3.1 *Given $A \in \mathbb{C}^{N\times N}$, let C be a constant, and consider the three conditions below;*

$$\begin{aligned} \|A^m\|_2 &\le C, \quad \forall m \ge 0, && (3.3)\\ dist(\lambda_\epsilon, D) &\le C\epsilon, \quad \forall \lambda_\epsilon \in \Lambda_\epsilon(A) \text{ and } \epsilon > 0, && (3.4)\\ \|A^m\|_2 &\le \exp(1)\min\{N, m+1\}C, \quad \forall m \ge 0. && (3.5) \end{aligned}$$

Then (3.3) $\Rightarrow$ (3.4) $\Rightarrow$ (3.5).

If we are willing to ignore the extra factor $\exp(1)\min\{N, m+1\}$ that distinguishes (3.3) from (3.5), then the theorem says that condition (3.4) is necessary and sufficient for $\|A^m\|_2$ to be bounded by C. Hence, studying the maximum size of the powers is essentially equivalent to studying $\mathrm{dist}(\lambda_\epsilon, D)/\epsilon$ for all $\epsilon > 0$. If ϵ is large, then the crude inequalities

$$\frac{\mathrm{dist}(\lambda_\epsilon, D)}{\epsilon} \le \frac{|\lambda_\epsilon|}{\epsilon} \le \frac{\|A + E\|_2}{\epsilon} \le 1 + \frac{\|A\|_2}{\epsilon}$$

provide an upper bound. So, in order for (3.4) to hold with a reasonable value of C, the behaviour of $\mathrm{dist}(\lambda_\epsilon, D)/\epsilon$ for *small* $\epsilon > 0$ provides the key. We must avoid the situation where λ_ϵ jumps outside the unit disk for small ϵ. Loosely, then, for $\|A^m\|_2$ to be controlled, it is not just the eigenvalues of A that must lie in D, but also the eigenvalues of $A + E$ for all small perturbations E. For the matrix in Figure 3.1 we see that perturbations of size 10^{-9} send eigenvalues outside the unit disk. ¿From Figure 2.1, the power norms for this matrix exceed 10^9.

The pseudo-eigenvalue characterisation in Theorem 3.1 is attractive and intuitively reasonable. However, non-normality is a difficult property to pin down, and using Theorem 3.1 to get sharp bounds is likely to be difficult in practice—what does the pseudospectrum of A look like, and what is the most relevant ϵ? Further analysis is possible when A is triangular and Toeplitz. This case arises when $e = 0$ in (1.10). We will let $S \subset \mathbb{C}$ denote the set of points

$$S := \{s \in \mathbb{C} \,:\, s = cz + d, \text{ for some } z \in \mathbb{C} \text{ with } |z| \le 1\};$$

that is, the symbol of the matrix applied to the closed unit disk. We may now quote Theorem 2.3 from [19], specialised to the bidiagonal case.

Theorem 3.2 *The family of bidiagonal Toeplitz matrices*

$$A_N = \begin{bmatrix} d & & & \\ c & \ddots & & \\ & \ddots & \ddots & \\ & & c & d \end{bmatrix} \in I\!R^{N\times N} \tag{3.6}$$

has pseudospectra that satisfy

$$\lim_{\epsilon\to 0}\lim_{N\to\infty} \Lambda_\epsilon(A_N) = S. \tag{3.7}$$

Theorem 3.2 says that for large N and small ϵ, the pseudospectrum is close to S. (Note that S is the spectrum of the corresponding Toeplitz operator.) It follows that for large N and small ϵ, the largest pseudo-eigenvalue in modulus has $|\lambda_\epsilon| \approx |c| + |d|$. For the Dirichlet iteration (1.10) with central differencing, forcing $r = \tilde{r}/2$ gives $e = 0$ in Table 1.1. In this case, as Δt is increased from zero, the limit beyond which $|c|+|d| \equiv 2r + |1-2r|$ leaves the closed unit disk is given by $r = 1/2$; that is,

$$\Delta t = \frac{\Delta x^2}{2b}. \tag{3.8}$$

The spectral radius of the Jacobian is simply $|d| \equiv |1-2r|$, so the time step limit for local attractivity occurs when $r = 1$, giving

$$\Delta t = \frac{\Delta x^2}{b}. \tag{3.9}$$

For a numerical test, we used parameters $N = 32$, $a = 1$ and $b = a\Delta x/2$, with initial data U^0 from the function $\phi(x) = 1+\gamma(1+\cos(2\pi x))$. We used a range of time steps up to the limit $\Delta x^2/b = .0606$ in (3.9). For each Δt, we we took 1000 steps and tested whether U^n converged to the steady state $U_i^* \equiv 1$. Starting with $\gamma = 10^{-17}$, we increased γ by a factor of 10, recording $\gamma_{\max}$, the largest value of γ beyond which the convergence test (2.25) failed. Figure 3.2 plots $\log_{10}\gamma_{\max}$ against Δt. It is clear that the limit of $\Delta t = .0303$ in (3.8) that comes from the pseudo-eigenvalue analysis is much more relevant than the $\Delta t = .0606$ limit in (3.9) that comes from the eigenvalues. As Δt is increased beyond .0303 the maximum perturbation that leads to convergence rapidly shrinks. The $\gamma = 10^{-17}$ level is effectively zero; since here $\phi(x) \equiv 1$ in our (double precision) finite precision arithmetic. In these cases the convergence condition was not actually satisfied; instead the iterates settled down to a spurious value of the type discussed in section 4.

We remark that Theorem 3.2 requires c and d to be the same for all N. In the test above, the theorem is clearly relevant for the particular choice of parameters. However, if we solved the same problem with smaller Δx (larger N) then c, d and e would change. This highlights a point made earlier in the section: highly non-normal matrices can arise for practical choices of grid-size, but as $\Delta x \to 0$, the non-normality effect will become less severe.

We mention that the reference [6] also looked at the impact of non-normality on the choice of time step. The authors studied linear stability of ordinary differential

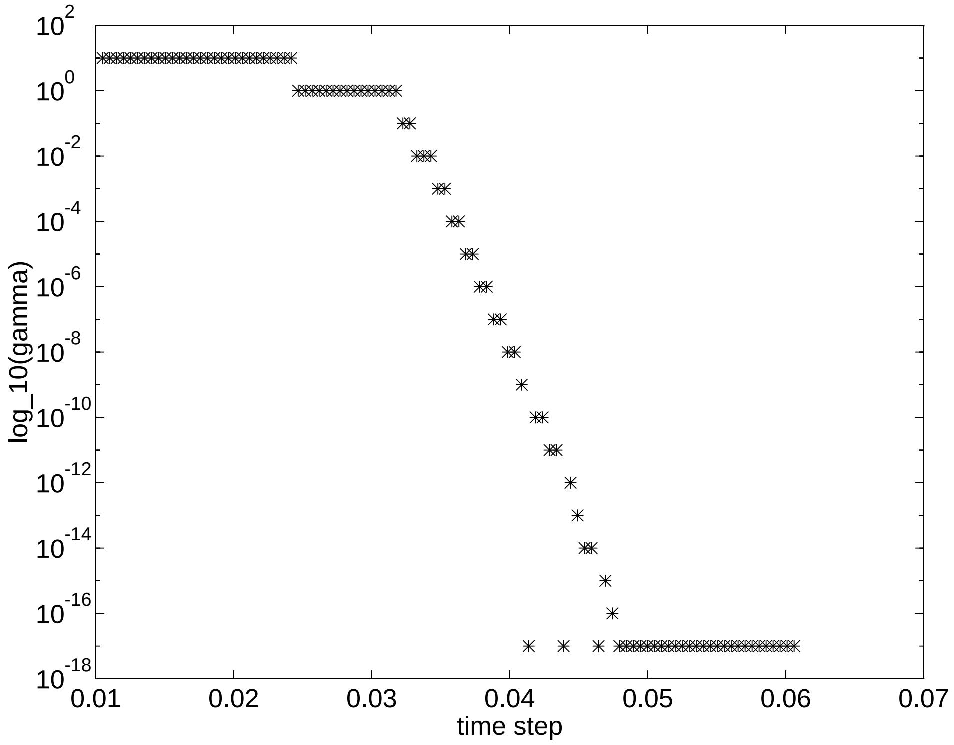

Figure 3.2: Attractivity of a fixed point where the Jacobian is bidiagonal.

equation solvers. Here, a standard condition for stability requires that $\Delta t\lambda$ must lie in the "stability region", for each eigenvalue λ of a local Jacobian. It was argued that in practice, for a non-normal Jacobian, $\Delta t\lambda_\epsilon$ must be inside the region for all small ϵ. Tests showed that the Δt values chosen by adaptive methods automatically satisfy this more stringent requirement.

4 Rounding Error Phenomena

In the case of Dirichlet boundary conditions, the choice $\Delta t \leq \Delta t_{\text{lim}}$, where Δt_{lim} is given by (2.27), ensures that the fixed point U^* is locally attractive. In particular, if the initial condition is $U^0 = U^* = (1, 1 \ldots, 1)^T$, then, in exact arithmetic, $U^n \equiv U^*$. But experiments show that in *floating point arithmetic* this behaviour does not always arise if convection is dominating; that is, $a \gg b$. The results in Figure 4.1 were obtained with $U^0 = U^*$, $a = 1$, $b = 10^{-3}$, $N = 32$ ($\Delta x = 1/(N+1)$) using backward differencing with $\Delta t = 0.99467112840836\Delta t_{\text{lim}}$. The figure shows seven successive time levels for large n. It is clear from the figure that the long term solution is essentially period 2 in time. Choosing random values for Δt in the interval $[0.8\Delta t_{\text{lim}}, \Delta t_{\text{lim}}]$, we found that three situations typically occur

1. $U^n \to U^*$, as predicted by the theory.

2. U^n converges to a spurious state that is very close to having period two in time, such as the one in Figure 4.1.

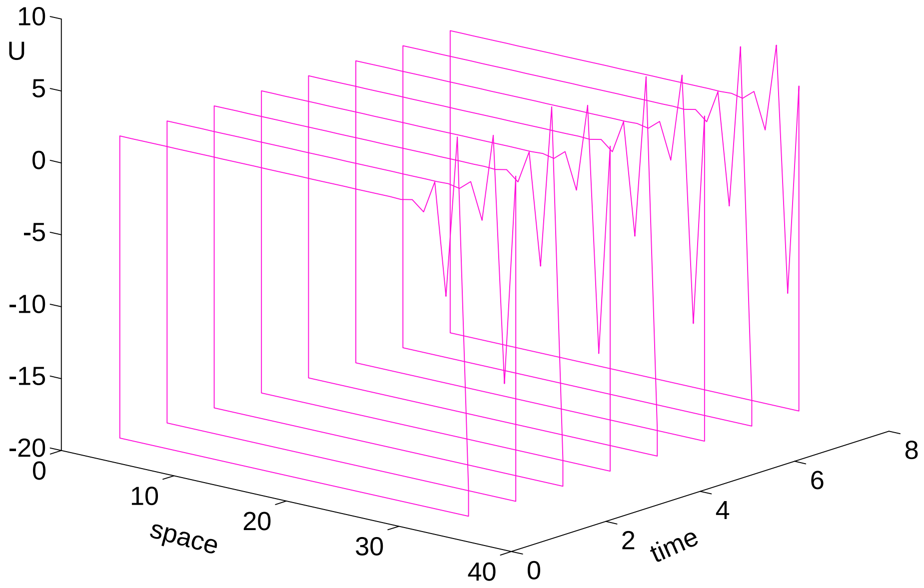

Figure 4.1: spurious periodic solution in the presence of rounding errors.

3. $\|U^n\| \to \infty$.

The third case was only observed when $\Delta t > 0.99\Delta t_{\text{lim}}$, but, otherwise, there was no obvious pattern in the way that the three possibilities occurred as Δt varied. The spurious steady state in situation 2 will henceforth be denoted $\{U^*_{s+}, U^*_{s-}\}$. ¿From the experiments we have observed that $U^*_{s\pm}$ is roughly of the form

$$U^*_{s\pm} \approx U^* \pm \epsilon\,(1, \beta, \beta^2, \ldots, \beta^{N-1})^T, \tag{4.1}$$

where ϵ is related to the machine precision and $\beta < 0$ depends on Δt. In our experiments, using Matlab [13], $\epsilon \approx 1.1 \times 10^{-16}$. The magnitude of β increases as Δt approaches Δt_{lim}.

A rigorous analysis of these phenomena would involve a model of floating point arithmetic. Related work is performed in [11]. Here, we give a more heuristic, machine-independent explanation of why solutions like the one seen in Figure 4.1 may exist in the presence of roundoff error.

This spurious steady state phenomenon is caused by non-normality rather than nonlinearity. Therefore, to simplify the analysis, we replace the nonlinear term $f(u) = u(1-u)$ by the linear approximation $1-u$. Experiments show that this does not alter the qualitative results. With this modification (1.10) becomes

$$U^{n+1} = TU^n + cp + eq + (U^* - U^n), \tag{4.2}$$

where $U^* = [1, \ldots, 1]^T$ is the true fixed point. Notice that this linearization does not change the value of Jacobian at the fixed point U^*. By looking at (1.10) we see that the true fixed point $U^* = [1, \ldots, 1]^T$ satisfies

$$U^* = TU^* + cp + eq, \tag{4.3}$$

Subtracting (4.3) from (4.2) and setting $V^n = (U^n - U^*)$ we obtain

$$V^{n+1} = (T - \Delta t I)V^n. \tag{4.4}$$

To allow for rounding errors in (4.4), we may write

$$V^{n+1} = (T - \Delta t I + E)V^n,$$

where E has small elements. As discussed in the previous section, if T is non-normal then $T - \Delta t I + E$ may have eigenvalues that are very different from those of $T - \Delta t I$. This suggests that a steady state of the form (4.1) may arise when a vector $V_s^* = (U_{s\pm}^* - U^*)$ is a pseudo-eigenvector of $T - \Delta t I$, with a corresponding pseudo-eigenvalue $\lambda_\epsilon = -1$. Assuming that the pseudo-eigenvectors of $T - \Delta t I$ are roughly of the form $(1, \beta, \beta^2, \ldots, \beta^{N-1})$ we can approximate $\beta = \beta(\Delta t)$ as a root of the quadratic equation

$$e\beta^2 + (d+1)\beta + c = 0, \tag{4.5}$$

where $T - \Delta t I = \text{tridiag}(c, d, e, 0, \ldots, 0)$. For our example, there are two real roots of (4.5) for each $\Delta t \in [0.8\Delta t_{\text{lim}}, \Delta t_{\text{lim}}]$. The root with the largest magnitude results in a steady state with numbers exceeding the maximum floating point number on our system. The smallest root corresponds closely to the observed value from our experiments. Figure 4.2 compares the observed and computed values for β in terms of $\Delta t/\Delta t_{\text{lim}}$. The deviation between observed and computed values of β near $\Delta t = \Delta t_{\text{lim}}$ could be caused by the linearization in (4.2), since the size of the last few components of $U_{s\pm}^*$ is quite significant in this case.

5 Spurious Periodic Solutions

Griffiths and Mitchell [4] study stable periodic bifurcations of the discrete system (1.8) in the case where the matrix defined by (1.9) is symmetric; more precisely, when $c = e = r$ and $d = 1 - 2r$, corresponding to $a = 0$ in (1.1). By applying the same techniques as Griffiths and Mitchell, we will find conditions under which stable spurious periodic solutions to (1.8) exist for arbitrary c and e, where

$$d = 1 - c - e. \tag{5.1}$$

This condition is necessary for the vector $e \equiv (1, \ldots, 1)^T$ to be a fixed point of (1.8). We may look for a solution to (1.8) that is period p in time and period q in space, where q divides N. Such a solution can be represented by a set of distinct vectors $V^1, \ldots, V^p$ satisfying

$$G(V^m) = V^{(m+1)\ mod\ p} \quad and \quad V_{j+q}^m = V_j^m, \tag{5.2}$$

with q minimal. We refer to this as a period-(p, q) solution. Condition (5.1) ensures that $(p, 1)$ states are independent of c and e, since if $U^n = v_n e$ we obtain from (1.8)

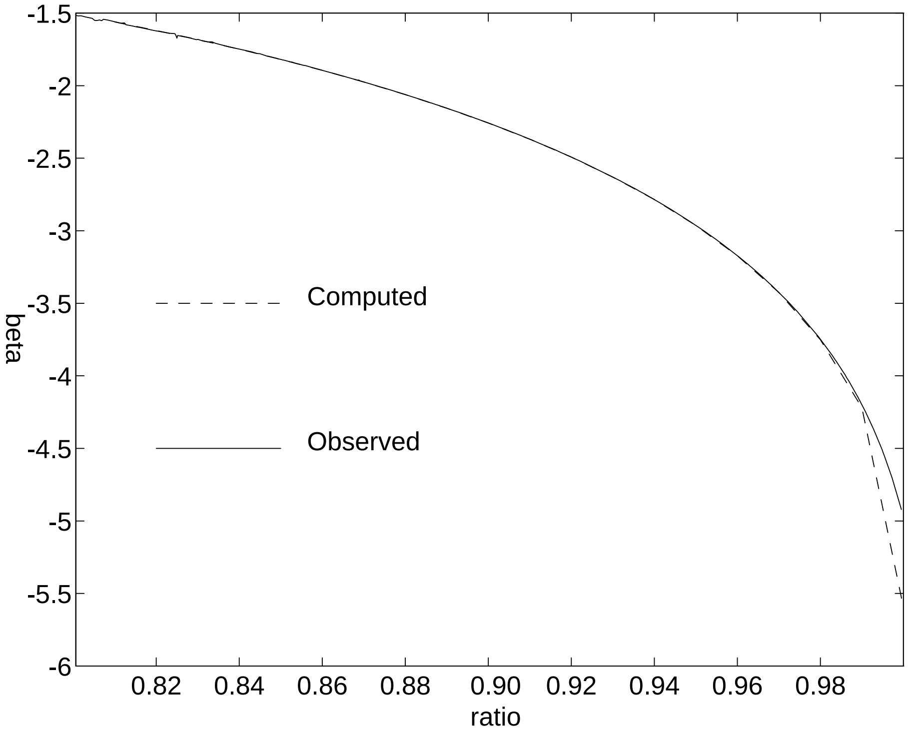

Figure 4.2: Observed and computed β in terms of $\Delta t/\Delta t_{\text{lim}}$.

that $v_{n+1} = v_n(1 + \Delta t(1 - v_n))$. However, stability of such states will depend on c and e.

A solution (5.2) is said to be *locally attractive*, or more loosely, *stable*, if there exist open neighbourhoods around the V^m such that the iterates will converge to the solution as $n \to \infty$ whenever U^0 is in one of these neighbourhoods. The solution will be locally attractive if

$$\rho(J_p) < 1, \quad \text{with} \quad \mathrm{J}_p := J(V^p)\dots J(V^1). \tag{5.3}$$

Here,

$$J(U) := C + \Delta t(I - 2\mathrm{diag}(U)) \tag{5.4}$$

is the Jacobian of the map in (1.8). The condition (5.3) can be derived from first principles (by linearising the map), or, alternatively, it can be found by applying Ostrowski's Theorem to the composite map $X \to G^p(X)$, for which V^1 is a fixed point.

5.1 Period-$(2, 1)$ solutions

We consider solutions of the form

$$U_j^{2n} = \xi, \quad U_j^{2n+1} = \eta$$

to the scheme

$$U_j^{n+1} = cU_{j-1}^n + dU_j^n + eU_{j+1}^n + \Delta t U_j^n(1 - U_j^n)$$

where $d = 1 - c - e$. Letting n be even and odd respectively we obtain the two equations

$$\eta = \xi(1 + \Delta t(1 - \xi)), \quad \xi = \eta(1 + \Delta t(1 - \eta)).$$

We ignore the trivial solution $\xi = \eta$ and obtain

$$\xi = \left(2 + \Delta t + \sqrt{\Delta t^2 - 4}\right)/(2\Delta t), \ \eta = \left(2 + \Delta t - \sqrt{\Delta t^2 - 4}\right)/(2\Delta t), \tag{5.5}$$

where ξ and η can be interchanged. Hence, period-$(2,1)$ solutions exist for $\Delta t > 2$.

To analyse the stability of these solutions we compute the eigenvalues of the product of Jacobians

$$J_2 = J(\xi e)J(\eta e).$$

By substituting the expressions (5.5) for ξ and η we obtain

$$\lambda_j(J_2) = (\lambda_j(C) - 2)^2 + 4 - \Delta t^2.$$

The eigenvalues of C have the form

$$\lambda_j(C) = (c + e)\cos(\theta_j) + d + i(e - c)\sin(\theta_j), \quad \theta_j = \frac{2\pi j}{N}.$$

Thus, the eigenvalues lie on an ellipse centered at $(d, 0)$ with semi-axes $|c + e|$ and $|c - e|$. The necessary and sufficient conditions for the eigenvalues of J_2 to lie in the unit disk depend on N. A sufficient condition for stability could be obtained by requiring the image of of the above ellipse under the transformation $z \to (z - 2)^2 + 4 - \Delta t^2$ to be contained in the unit disk. Rather then pursuing this we begin by looking at the case where N is even and we require that $\lambda_N(J_2)$ as well as $\lambda_{N/2}(J_2)$ be bounded by one in absolute value. We obtain the condition

$$\max\{4, 4 + 4(d-1)(d-2)\} \leq \Delta t^2 \leq \min\{6, 6 + 4(d-1)(d-2)\}. \tag{5.6}$$

This interval is nonempty for $\frac{1}{2}(3 - \sqrt{3}) \leq d \leq \frac{1}{2}(3 + \sqrt{3})$. Similarly, if 4 divides N, then we can consider the eigenvalues $\lambda_{N/4}(J_2)$ and $\lambda_{3N/4}(J_2)$ and the corresponding condition is

$$|(c - e)(1 + c + e)| \leq \frac{1}{2}. \tag{5.7}$$

All three choices of discretisation in Table 1.1 satisfy $|c - e| = \tilde{r}$, thus period $(2,1)$ solutions will become unstable for large enough values of the convection coefficient.

5.2 Period-$(2,2)^*$ solutions

Period-$(2,2)$ solutions can generally occur according to the diagram

	j even	j odd
n even	ξ_1	η_1
n odd	ξ_2	η_2

We consider the special case where $\xi_1 = \eta_2 := \xi$, $\eta_1 = \xi_2 := \eta$ called $(2,2)^*$ solutions by Griffiths and Mitchell [4]. Let us assume that N is even. We get the two equations

$$\eta d = \xi(d + \Delta t(1-\xi)), \quad \xi d = \eta(d + \Delta t(1-\eta)),$$

which have the solution

$$2\Delta t\xi = \Delta t + 2d + \sqrt{\Delta t^2 - 4d^2}, \quad 2\Delta t\eta = \Delta t + 2d - \sqrt{\Delta t^2 - 4d^2},$$

where again ξ and η can be interchanged. Notice that the periodic state only depends on the diagonal element d. In this way we obtain two alternating spatial patterns $V^1 = (\xi, \eta, \xi, \eta, \dots, \xi, \eta)^T$ and $V^2 = (\eta, \xi, \eta, \xi \dots, \eta, \xi)^T$.

We first consider the case $N = 2$. The product of the two Jacobians gives

$$J_2 = \begin{pmatrix} X + (c+e)^2 & 2A(c+e) \\ 2B(c+e) & X + (c+e)^2 \end{pmatrix}, \; X = AB, \; A = d + \Delta t(1-2\xi), \; B = d + \Delta t(1-2\eta).$$

It is easily verified that $X = 5d^2 - \Delta t^2$. The eigenvalues of J_2 are found to be

$$\lambda = (c + e \pm \sqrt{X})^2. \tag{5.8}$$

Note at this point that if central differencing is used for the u_x term, the stability characteristics will be the same as in the diffusion/reaction case. Now, if $X = 5d^2 - \Delta t^2 \geq 0$, the eigenvalues are real. This leads to the condition

$$0 \leq \Delta t^2 \leq \begin{cases} 4d^2 & \text{if} \quad 0 \leq d \leq 1 \\ 4(d^2 + d - 1) & \text{if} \quad 1 < d \leq 2 \end{cases}$$

In [4] $d = 1 - 2r \leq 1$ for $r \geq 0$, hence only the upper inequality is necessary. If $X < 0$, the eigenvalues are complex, and we obtain the inequality $\Delta t^2 \leq 2d(2d+1)$, so stability of the $(2,2)^*$ solution is ensured if

$$4 \cdot \max\{d^2, d^2 + d - 1\} \leq \Delta t^2 \leq 2d(2d+1) \text{ for } 0 \leq d \leq 2. \tag{5.9}$$

For general (even) N we compute $J(V^2)J(V^1)$ and obtain the block circulant matrix

$$J_2 = \begin{pmatrix} D & E & & & & F \\ F & D & E & & & \\ & F & D & & & \\ & & & \ddots & & \\ & & & F & D & E \\ E & & & & F & D \end{pmatrix},$$

where D, E, F are 2×2 matrices

$$D = \begin{pmatrix} X + 2ce & 2Be \\ 2Ac & X + 2ce \end{pmatrix}, \; E = \begin{pmatrix} e^2 & 0 \\ 2Ae & e^2 \end{pmatrix}, \; F = \begin{pmatrix} c^2 & 2Bc \\ 0 & c^2 \end{pmatrix}.$$

We will write the eigenvectors of J_2 in the form $(v_1, \ldots, v_{N/2})^T$ with $v_j \in \mathbb{C}^2$. This gives the recursion

$$F\, v_{j-1} + (D - \lambda I)\, v_j + E\, v_{j+1} = 0.$$

By writing

$$v_j = e^{i\theta j}\gamma, \quad j = 1, \ldots, N/2,$$

where $\theta = \theta_m = 2\pi m/(\frac{1}{2}N)$, we get

$$(e^{-i\theta}F + (D - \lambda I) + e^{i\theta}E)\gamma = 0.$$

We then require the determinant of this matrix to vanish, which gives

$$(X + q^2 - \lambda)^2 = 4Xq^2 \quad \Longleftrightarrow \quad \lambda = (q \pm \sqrt{X})^2, \tag{5.10}$$

with $q = e\,\exp(i\theta/2) + c\,\exp(-i\theta/2)$. Consistency with the Mitchell and Griffiths case is seen by observing that when $c = e = r$ we get $q = 4r^2\cos^2(\theta/2)$. For the eigenvalues corresponding to $\theta = 0$, (5.10) reduces to (5.8) so a necessary condition for stability is (5.9). If we assume that 4 divides $N/2$, we can consider the eigenvalues corresponding to $m = \frac{1}{4}(N/2)$. When $X = 5d^2 - \Delta t^2 \geq 0$ the eigenvalues are complex and we find

$$|\lambda| = (e - c)^2 + X \leq 1 \quad \Rightarrow \quad 5d^2 + (e - c)^2 - 1 \leq \Delta t^2 \leq 5d^2.$$

If $-X \geq 0$, the eigenvalues corresponding to $m = N/8$ are purely imaginary numbers and we find that we must impose the condition

$$5d^2 \leq \Delta t^2 \leq 5d^2 + (1 - |e - c|)^2 \quad \text{and} \quad |e - c| \leq 1.$$

In conclusion, in the case that 8 divides N we have the necessary conditions

$$5d^2 + (e - c)^2 - 1 \leq \Delta t^2 \leq 5d^2 + (1 - |e - c|)^2 \quad \text{and} \quad |e - c| \leq 1.$$

Note in particular that stable $(2, 2)^*$ solutions cannot exist if $|e - c|$ exceeds one.

Mitchell and Griffiths [4] use perturbation theory to establish the existence of stable period-$(4, 1)$ solutions to (1.8) when $c = e = r$ in a limited region of the $(r, \Delta t)$ plane. By a continuity argument it is clear that such solutions will also exist in the presence of a small convection term. We have confirmed this by numerical experiments.

In conclusion, we have seen that the stable spurious periodic solutions found by Griffiths and Mitchell [4] in the discretised reaction diffusion equation continue to exist when a moderate convection term is introduced.

6 Concluding Remarks

The phenomena outlined in this work are likely to arise from any discrete map $U^{n+1} = G(U^n)$ where the Jacobian of G is non-normal. In the Dirichlet version of model problem (1.1) we fixed the reaction term and the boundary conditions so that $u \equiv 1$ is a steady state. This is convenient for the analysis. However, we emphasise that the underlying principles apply more generally. In particular, the Neumann boundary condition $u_x \equiv 0$ could be imposed at $x = 0$ and/or $x = 1$. This preserves the steady-state $u \equiv 1$. The corresponding linear stability analysis in [7] (for the case of central differences) shows that the non-normality effect is still relevant. More general Dirichlet or mixed boundary conditions could also be considered, but these may give rise to steady states that are not spatially uniform, which complicates the analysis considerably. The precise form of the logistic reaction term $f(u) = u(1-u)$, which is commonly used in mathematical biology, did not play a central role in the analysis. It is clear from equation (2.10) that Corollary 2.1 is readily adapted to any f for which f'' can be bounded.

Acknowledgements
We thank David Griffiths for many useful discussions, and we thank Brian Sleeman for outlining a proof of the result in the appendix.

References

[1] K. Dekker and J.G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, 1984.

[2] A.R. Gardiner and A.R. Mitchell, *Bifurcation studies in long time solutions of discrete models in reaction diffusion*, Tech. Report NA/136, University of Dundee, 1992.

[3] D.F. Griffiths, I. Christie, and A.R. Mitchell, *Analysis of error growth for explicit difference schemes in conduction-convection problems*, Int. J. Numer. Meth. Eng., 15 (1980), pp. 1075–1081.

[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, 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.

[6] D.J. Higham and L.N. Trefethen, *Stiffness of ODEs*, BIT, 33 (1993), pp. 285–303.

[7] A.C. Hindmarsh, P.M. Gresho, and D.F. Griffiths, *The stability of explicit Euler-time integration for certain finite difference approximations of the multi-dimensional advection-diffusion equation*, Int. J. Num. Meth. Fluids, 4 (1984), pp. 853–897.

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

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

[10] 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.

[11] K.R. Jackson and B. Owren, *Stepsize reduction in explicit Runge-Kutta methods when solving stiff constant-coefficient linear ODEs having nonnormal coefficient matrices.* Unpublished Manuscript, 1992.

[12] D.S. Jones and B.D. Sleeman, *Differential Equations and Mathematical Biology*, George Allen and Unwin, 1983.

[13] The MathWorks, Inc., *MATLAB User's Guide*, Natick, Massachusetts, 1992.

[14] A.R. Mitchell and D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*, Wiley, 1985.

[15] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[16] M.H. Protter and H.F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall (Re-issued by Springer-Verlag), 1967.

[17] S.C. Reddy and L.N. Trefethen, *Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues*, Comp. Math. Appl. Mech. Eng., 80 (1990), pp. 147–164.

[18] ——, *Stability of the method of lines*, Numer. Math., 62 (1992), pp. 235–267.

[19] L. Reichel and L.N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Lin. Alg. Applics, 162–164 (1992), pp. 153–185.

[20] R.D. Richtmeyer and K.W. Morton, *Difference Methods for Initial Value Problems*, Wiley, 1967.

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

[22] 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. Comp. Phys., 97 (1991), pp. 249–310.

7 Appendix: Theoretical Result

The following theorem gives conditions on the initial data under which solutions to the Dirichlet problem converge to $u(x,t) \equiv 1$ as $t \to \infty$.

Theorem 7.1 *Consider the reaction-convection-diffusion equation*

$$u_t + au_x = bu_{xx} + u(1-u), \qquad \text{in } (0,1)\times(0,\infty),$$

with a, b constant, $b > 0$, subject to the initial and boundary conditions

$$u(x,0) = \phi(x) \in C^1, \quad u(0,t) = u(1,t) = 1.$$

If the initial data $\phi(x)$ satisfies either

- $0 \le \phi(x) \le 1$ *for all* $0 \le x \le 1$, *or*
- $\phi(x) \ge 1$ *for all* $0 \le x \le 1$,

then the solution $u(x,t)$ is bounded for all $t > 0$ and converges pointwise to the steady-state $u(x,t) \equiv 1$ as $t \to \infty$.

The proof relies on the following two theorems.

Theorem 7.2 *(Comparison Theorem) Given constants $T > 0$ and A, B, and a continuously differentiable function $f : IR \to IR$, let $u(x,t)$ and $v(x,t)$ be bounded functions satisfying*

$$u_t + au_x - bu_{xx} - f(u) \ge v_t + av_x - bv_{xx} - f(v), \qquad \text{in } (0,1)\times(0,T),$$

with a, b constant, $b > 0$, together with

$$B \ge u(0,t) \ge v(0,t) \ge A, \qquad B \ge u(1,t) \ge v(1,t) \ge A, \qquad \text{for } 0 \le t \le T,$$

and

$$B \ge u(x,0) \ge v(x,0) \ge A, \qquad \text{for} \quad 0 < x < 1.$$

Then $u(x,t) \ge v(x,t)$ in $[0,1]\times[0,T]$.

Proof. For $a = 0$, the result is proved in [12, Theorem 10.4c, page 236]. With $a \ne 0$, we can use the technique of proof in [12], appealing to the more general version of the strong maximum principle theorem in [16, Theorem 7, page 174]. ∎

Theorem 7.3 *Consider the linear reaction-convection-diffusion equation*

$$w_t + aw_x = bw_{xx}, \qquad \text{in } (0,1)\times(0,\infty),$$

with a, b constant, $b > 0$, subject to the initial and boundary conditions

$$w(x,0) = \psi(x) \ge 0, \quad \psi(x) \in C^1, \quad w(0,t) = w(1,t) = 0.$$

Then $w(x,t)$ is bounded for all t and converges pointwise to the steady-state $w(x,t) \equiv 0$ as $t \to \infty$.

Proof. The result can be proved by elementary techniques, such as separation of variables. ∎

Proof of Theorem 7.1:

Consider the case $0 \leq \phi(x) \leq 1$ for all $0 \leq x \leq 1$. Given any $T > 0$, assume for the moment that $u(x,t)$ is bounded for $0 \leq t \leq T$. The first step is to show that $0 \leq u(x,t) \leq 1$ in $[0,1] \times [0,T]$.

The function $w(x,t) \equiv 0$ trivially satisfies

$$w_t + aw_x - bw_{xx} + w(1-w) = 0, \qquad in\ (0,1) \times (0,T),$$

together with

$$w(x,0) = 0, \quad w(0,t) = w(1,t) = 0.$$

Hence, the Comparison Theorem gives $u(x,t) \geq w(x,t) \equiv 0$ in $[0,1] \times [0,T]$. Similarly, using $w(x,t) \equiv 1$ shows that $u(x,t) \leq 1$ in $[0,1] \times [0,T]$.

The validity of the assumption that $u(x,t)$ is bounded for $0 \leq t \leq T$ can be confirmed by contradiction. If $u(x,t)$ is not bounded then there exists a $0 \leq t^* \leq T$ such that

$$\max{(x,t) \in [0,1] \times [0,t^*]} |u(x,t)| = 2.$$

But the comparison arguments above can be applied to show $0 \leq u(x,t) \leq 1$ in $[0,1] \times [0,t^*]$, giving the contradiction.

Next, let $u = 1 - v$. The function v satisfies $0 \leq v(x,t) \leq 1$ in $[0,1] \times [0,T]$, and hence

$$0 = v_t + av_x - bv_{xx} + v(1-v) \geq v_t + av_x - bv_{xx}, \qquad in\ (0,1) \times (0,T).$$

Letting $\widehat{v}$ satisfy

$$\widehat{v}_t + a\widehat{v}_x - b\widehat{v}_{xx} = 0, \qquad in\ (0,1) \times (0,T),$$

and

$$\widehat{v}(x,0) = 1 - \phi(x) \geq 0, \quad \widehat{v}(0,t) = \widehat{v}(1,t) = 0,$$

we see that the Comparison Theorem (with $f \equiv 0$) can be applied to give $\widehat{v}(x,t) \geq v(x,t)$ in $[0,1] \times [0,T]$. Also, Theorem 7.3 gives $\widehat{v}(x,t) \to 0$ as $t \to \infty$, so that $v(x,t) \to 0$, as required.

The proof for the case where $\phi(x) \geq 1$ for all $0 \leq x \leq 1$ is similar. Comparison with $w(x,t) \equiv 1$ shows that $u(x,t) \geq 1$. Writing $v = u - 1$ leads to

$$0 = v_t + av_x - bv_{xx} + v(1+v) \geq v_t + av_x - bv_{xx}.$$

Letting $\widehat{v}$ satisfy

$$\widehat{v}_t + a\widehat{v}_x - b\widehat{v}_{xx} = 0, \qquad in\ (0,1) \times (0,T),$$

and

$$\widehat{v}(x,0) = \phi(x) - 1 \geq 0, \quad \widehat{v}(0,t) = \widehat{v}(1,t) = 0,$$

we see that the Comparison Theorem can be applied to give $\widehat{v}(x,t) \geq v(x,t)$ in $[0,1] \times [0,T]$. Theorem 7.3 gives $\widehat{v}(x,t) \to 0$ as $t \to \infty$, so that $v(x,t) \to 0$, as required. ∎