MATHEMATICS OF COMPUTATION
Volume 72, Number 241, Pages 263–288
S 0025-5718(01)01392-8
Article electronically published on December 4, 2001

A CHARACTERISATION OF OSCILLATIONS IN THE DISCRETE TWO-DIMENSIONAL CONVECTION-DIFFUSION EQUATION

HOWARD C. ELMAN AND ALISON RAMAGE

Abstract. It is well known that discrete solutions to the convection-diffusion equation contain nonphysical oscillations when boundary layers are present but not resolved by the discretisation. However, except for one-dimensional problems, there is little analysis of this phenomenon. In this paper, we present an analysis of the two-dimensional problem with constant flow aligned with the grid, based on a Fourier decomposition of the discrete solution. For Galerkin bilinear finite element discretisations, we derive closed form expressions for the Fourier coefficients, showing them to be weighted sums of certain functions which are oscillatory when the mesh Péclet number is large. The oscillatory functions are determined as solutions to a set of three-term recurrences, and the weights are determined by the boundary conditions. These expressions are then used to characterise the oscillations of the discrete solution in terms of the mesh Péclet number and boundary conditions of the problem.

1. Introduction

Convection-diffusion equations arise in mathematical models of many different processes in diverse areas of science and engineering. Analysis of the linear convection-diffusion equation

$$(1.1)\qquad \begin{aligned} -\epsilon\nabla^2 u(x,y) + \mathbf{w}\cdot\nabla u(x,y) &= f(x,y) \quad \text{in} \quad \Omega \\ u(x,y) &= g(x,y) \quad \text{on} \quad \delta\Omega, \end{aligned}$$

where the small parameter ϵ and divergence-free convective velocity field $\mathbf{w} = (w_1(x,y), w_2(x,y))$ are given, is important in this context as it can be used to gain insight into the behaviour of these more complex phenomena as well as being of interest in its own right. In this paper, we analyse the nature of the discrete solution produced when solving (1.1) using the Galerkin finite element method.

There are many textbooks which describe finite element modelling of convection-diffusion equations in detail (see for example [2], [4], [5]): we therefore present only a very brief outline here. The weak form of (1.1) has solution u in the Sobolev space $V = \mathcal{H}_0^1(\Omega)$ satisfying

$$\epsilon(\nabla u, \nabla v) + (\mathbf{w}.\nabla u, v) = (f, v) \qquad \forall\, v \in V.$$

Received by the editor March 27, 2000 and, in revised form, February 22, 2001.

2000 *Mathematics Subject Classification.* Primary 65N22, 65N30, 65Q05, 35J25.

Key words and phrases. Convection-diffusion equation, oscillations, Galerkin finite element method.

The work of the first author was supported by National Science Foundation grant DMS9972490.

The work of the second author was supported by the Leverhulme Trust.

Applying this to a finite-dimensional subspace V_h of V, we seek $u_h \in V_h$ such that

$$\epsilon(\nabla u_h, \nabla v) + (\mathbf{w}.\nabla u_h, v) = (f_h, v) \qquad \forall\, v \in V_h \tag{1.2}$$

where f_h is the $L^2(\Omega)$ orthogonal projection of f into V_h. On a grid with M interior nodes, we may choose a set of basis functions Φ_i, $i = 1, \dots, M$ for V_h and look for an approximate solution of the form

$$u_h(x, y) = \sum_{i=1}^{M} U_i \Phi_i(x, y). \tag{1.3}$$

Substituting this into (1.2) and choosing the test functions equal to the basis functions leads to a set of M linear equations

$$A\mathbf{u} = \mathbf{f} \tag{1.4}$$

where the entries of $\mathbf{u}$ are the M unknown coefficients U_i in (1.3).

It is well known that applying the Galerkin finite element method to the one-dimensional analogue of (1.1) will in certain circumstances result in a discrete solution which exhibits non-physical oscillations. For linear elements on a uniform grid, a precise statement as to exactly when such oscillations occur can be made, namely, that for a problem with mesh size h, constant advective velocity w and different values at the left and right boundaries, oscillations will occur if the mesh Péclet number

$$P_e = \frac{|w|h}{2\epsilon} \tag{1.5}$$

is greater than one (see for example [4], §1.3). The exact character of numerical approximations to the solution in this case is well understood. The same is not true, however, for problems in two or more dimensions. For example, Gresho and Sani ([2], p. 219) say of the two-dimensional case, "useful, closed-form solutions are much harder to find and are often difficult to 'interpret' even when found." We know of very little analysis which has been done in this area, although Semper [6] presents a limited discussion of oscillations in streamline diffusion finite element solutions when $\epsilon = 0$. We are not aware of any work which describes the nature of oscillations in the full two-dimensional convection-diffusion equation.

In this paper we present a mechanism for deriving useful, closed-form two-dimensional solutions to (1.1) and interpret the resulting formulae both in general and for a number of specific test problems. In particular, we will characterise the behaviour of oscillations in the direction of the flow with respect to variations in the mesh Péclet number. We begin in Section 2 with a description of the basis of our analysis. For the case where $\mathbf{w} = (0, 1)$ in (1.1) and the underlying finite element grid is uniform, we can use Fourier analysis to construct an analytic formula for the discrete solution $\mathbf{u}$. Specifically, we present a Fourier decomposition of A in (1.4) which means that blocks of the vector $\mathbf{u}$ can be explicitly expressed as a transformation of blocks of a vector $\mathbf{y}$ containing the solutions to a set of tridiagonal linear systems. This idea is the basis of some fast direct solvers for linear systems of this type [8]. Here, however, we take advantage of the fact that as these tridiagonal systems are of Toeplitz form, they can be solved analytically via a set of three-term recurrence relations: the construction and solution of these equations is described in Section 3. The result of this process is an exact analytic expression for the entries of the discrete solution vector $\mathbf{u}$ in (1.4). This takes the form of a weighted sum of three functions of the recurrence relation roots, where the weights depend on the boundary conditions of the problem.

In the remainder of the paper, we focus on the particular $\mathbf{u}$ arising from the Galerkin finite element method with bilinear basis functions on square elements. Note, however, that the analysis is not specific to the finite element method but in fact applies to any discretisation method whose matrix entries fall within a certain nine-point stencil. In Section 4, we obtain an analytic expression for the recurrence relation solution vector $\mathbf{y}$ in the bilinear case. From this, we show that for a large mesh Péclet number, certain components of the recurrence relation solution are highly oscillatory and have behaviour analogous to solutions of simple one-dimensional convection-diffusion problems. In addition, we establish the fact that there is no mesh Péclet number for which the vector $\mathbf{y}$ is always guaranteed to be oscillation free. Finally, we interpret these results in terms of the full discrete solution by examining the effects of the Fourier transform which relates the two vectors $\mathbf{y}$ and $\mathbf{u}$. By evaluating the boundary condition dependent weight functions, we can precisely characterise the behaviour of $\mathbf{u}$ for certain representative test problems.

2. Preliminary Fourier analysis

The analysis presented in this section is based on Fourier techniques. In order to use such an approach, we set $\mathbf{w} = (0, 1)$ and f=0 in (1.1) to obtain the equation

$$-\epsilon\nabla^2 u + \frac{\partial u}{\partial y} = 0 \qquad \text{in } \Omega = (0,1) \times (0,1), \tag{2.1}$$

and we apply Dirichlet boundary conditions as shown in Figure 2.1 for given functions f_t, f_r, f_b and f_l. We will discretise this problem on a uniform grid of square elements with N elements in each dimension (that is, grid parameter $h = 1/N$).

For many standard discretisation techniques with a natural ordering of the unknowns, the resulting linear system can be written in the form (1.4) where the

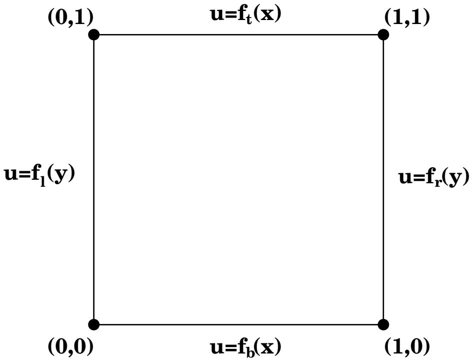


Figure 2.1. Boundary conditions.

coefficient matrix A is of order $(N-1)^2$ and has the general structure

$$A = \begin{bmatrix} M_1 & M_2 & & & 0 \\ M_3 & M_1 & M_2 & & \\ & \ddots & \ddots & \ddots & \\ & & M_3 & M_1 & M_2 \\ 0 & & & M_3 & M_1 \end{bmatrix}. \tag{2.2}$$

Here,

$$M_1 = \text{tridiag}(m_2, m_1, m_2),$$
$$M_2 = \text{tridiag}(m_4, m_3, m_4),$$

and

$$M_3 = \text{tridiag}(m_6, m_5, m_6)$$

are all tridiagonal matrices of order $N-1$. Using discrete Fourier analysis, we can obtain analytic expressions for the eigenvalues and eigenvectors of the blocks of A: the matrices M_1, M_2 and M_3 satisfy

$$\begin{array}{ll} M_1\mathbf{v}_j = \lambda_j\mathbf{v}_j & \lambda_j = m_1 + 2m_2\cos\frac{j\pi}{N} \\ M_2\mathbf{v}_j = \sigma_j\mathbf{v}_j & \sigma_j = m_3 + 2m_4\cos\frac{j\pi}{N} \\ M_3\mathbf{v}_j = \gamma_j\mathbf{v}_j & \gamma_j = m_5 + 2m_6\cos\frac{j\pi}{N} \end{array} \tag{2.3}$$

for $j = 1, \ldots, N-1$, where the eigenvectors are

$$\mathbf{v}_j = \sqrt{\frac{2}{N}}\left[\sin\frac{j\pi}{N}, \quad \sin\frac{2j\pi}{N}, \ldots, \sin\frac{(N-1)j\pi}{N}\right]^T. \tag{2.4}$$

We now introduce a decomposition of the coefficient matrix A in term of these eigenvalues and eigenvectors. First we consider the matrix V which has the vectors $\mathbf{v}_j$, $j = 1, \ldots, N-1$ as its columns and the related block diagonal matrix $\mathcal{V}$ which has V as each diagonal block. We will also use diagonal matrices $\Lambda = \text{diag}(\lambda_i)$, $\Gamma = \text{diag}(\gamma_i)$ and $\Sigma = \text{diag}(\sigma_i)$, $i = 1, \ldots, N-1$, combining them to get

$$\mathcal{V}^T A \mathcal{V} = \mathcal{T} = \begin{bmatrix} \Lambda & \Sigma & & & 0 \\ \Gamma & \Lambda & \Sigma & & \\ & \ddots & \ddots & \ddots & \\ & & \Gamma & \Lambda & \Sigma \\ 0 & & & \Gamma & \Lambda \end{bmatrix}.$$

Introducing a permutation matrix P of order $(N-1)^2$, we may write

$$P^T \mathcal{T} P = T = \begin{bmatrix} T_1 & & & & 0 \\ & T_2 & & & \\ & & \ddots & & \\ & & & T_{N-2} & \\ 0 & & & & T_{N-1} \end{bmatrix}$$

where $T_i = \text{tridiag}(\gamma_i, \lambda_i, \sigma_i)$, that is, each block T_i is a tridiagonal Toeplitz matrix of order $N-1$. Hence we have the decomposition

$$A = \mathcal{V}\mathcal{T}\mathcal{V}^T = \mathcal{V}(PTP^T)\mathcal{V}^T.$$

Using this decomposition, (1.4) implies

$$\mathcal{V}(PTP^T)\mathcal{V}^T\mathbf{u} = \mathbf{f} \;\Rightarrow\; P^T\mathcal{V}^T\mathbf{u} = T^{-1}P^T\mathcal{V}^T\mathbf{f},$$

that is,

$$\mathbf{u} = \mathcal{V}P\mathbf{y} \tag{2.5}$$

where the vector $\mathbf{y}$ is the solution to the linear system

$$T\mathbf{y} = P^T\mathcal{V}^T\mathbf{f} \equiv \hat{\mathbf{f}}.$$

We recall that T is block diagonal: this system can therefore be partitioned into $N-1$ systems of the form

$$T_i\mathbf{y}_i = \hat{\mathbf{f}}_i \tag{2.6}$$

where T_i is defined above and $\mathbf{y}$ and $\hat{\mathbf{f}}$ are partitioned in the obvious way. Note that each vector $\mathbf{y}_i$ contains the ith components of the one-dimensional Fourier transforms of the blocks of $\mathbf{u}$. Because T_i is a Toeplitz matrix, each of these systems can be considered as a three-term recurrence relation which can be solved analytically to give an expression for each entry y_{ik} of $\mathbf{y}_i$, $k = 1, \ldots, N-1$. As the rows of $\mathcal{V}$ are just the eigenvectors given in (2.4), the entry of the discrete solution vector (2.5) corresponding to the grid point (jh, kh) can now be written as

$$u_{jk} = \sqrt{\frac{2}{N}}\sum_{i=1}^{N-1}\sin\frac{ij\pi}{N}y_{ik}. \tag{2.7}$$

3. Solving the recurrence relations

To see the exact form which the recurrences (2.6) take, we must consider more closely the structure of the right-hand side vector $\hat{\mathbf{f}}_i$. As there is no forcing function in (2.1), the only nonzero entries in the original right-hand side vector $\mathbf{f}$ arise from applying the Dirichlet boundary conditions. In general, each entry in this vector consists of a sum of certain matrix coefficients times boundary values. For example, suppose the $N+1$ nodes on the bottom (or top) boundary are labelled from x_0 to x_N. Now construct $(N-1)$-vectors $\mathbf{b}$ and $\mathbf{t}$ with entries given by

$$\begin{aligned} b_i &= -m_6\, f_b(x_{i-1}) - m_5\, f_b(x_i) - m_6\, f_b(x_{i+1}), \\ t_i &= -m_4\, f_t(x_{i-1}) - m_3\, f_t(x_i) - m_4\, f_t(x_{i+1}) \end{aligned}$$

for $i = 1, \ldots, N-1$ (where the values m_* are the entries of A). Similarly, if the left (right) boundary nodes are labelled from y_0 to y_N, construct vectors $\mathbf{l}$ and $\mathbf{r}$ with entries

$$\begin{aligned} l_k &= -m_6\, f_l(y_{k-1}) - m_2\, f_l(y_k) - m_4\, f_l(y_{k+1}), \\ r_k &= -m_6\, f_r(y_{k-1}) - m_2\, f_r(y_k) - m_4\, f_r(y_{k+1}) \end{aligned}$$

for $k = 1, \dots, N-1$. With each of these, further associate $(N-1)^2$-vectors

$$\hat{\mathbf{b}} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \qquad \hat{\mathbf{t}} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{t} \end{bmatrix},$$

$$\hat{\mathbf{l}} = P^T \begin{bmatrix} \mathbf{l} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{l}_1 \\ \mathbf{l}_2 \\ \vdots \\ \mathbf{l}_{N-1} \end{bmatrix}, \qquad \hat{\mathbf{r}} = P^T \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{r} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_{N-1} \end{bmatrix},$$

where $\mathbf{l}_k = [l_k, 0, \dots, 0]^T$, $\mathbf{r}_k = [0, \dots, 0, r_k]^T$ and $\mathbf{0}$ represents the zero $(N-1)$-vector. The right-hand side vector $\mathbf{f}$ can now be written as

$$\mathbf{f} = \hat{\mathbf{b}} + \hat{\mathbf{t}} + \hat{\mathbf{l}} + \hat{\mathbf{r}} = \begin{bmatrix} \mathbf{b} + \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_{N-2} \\ \mathbf{t} + \mathbf{s}_{N-1} \end{bmatrix},$$

where $\mathbf{s}_k = \mathbf{l}_k + \mathbf{r}_k$ combines the left and right boundary condition contributions. So

$$\mathcal{V}^T \mathbf{f} = \begin{bmatrix} V^T(\mathbf{b} + \mathbf{s}_1) \\ V^T \mathbf{s}_2 \\ \vdots \\ V^T \mathbf{s}_{N-2} \\ V^T(\mathbf{t} + \mathbf{s}_{N-1}) \end{bmatrix} \equiv \begin{bmatrix} \bar{\mathbf{b}} + \bar{\mathbf{s}}_1 \\ \bar{\mathbf{s}}_2 \\ \vdots \\ \bar{\mathbf{s}}_{N-2} \\ \bar{\mathbf{t}} + \bar{\mathbf{s}}_{N-1} \end{bmatrix}$$

and $\hat{\mathbf{f}} = P^T \mathcal{V}^T \mathbf{f}$ has blocks $\hat{\mathbf{f}}_i$ with entries given by

$$\hat{\mathbf{f}}_i = \begin{bmatrix} \bar{b}_i + (\bar{\mathbf{s}}_1)_i \\ (\bar{\mathbf{s}}_2)_i \\ \vdots \\ (\bar{\mathbf{s}}_{N-2})_i \\ \bar{t}_i + (\bar{\mathbf{s}}_{N-1})_i \end{bmatrix}, \tag{3.1}$$

$i = 1, \dots, N-1$. These vectors are precisely the right-hand side vectors for the systems in (2.6).

In order to simplify the analysis presented in the remainder of the paper, we will henceforth assume that the functions $f_l(\mathbf{y})$ and $f_r(\mathbf{y})$ on the left and right boundaries are constant, that is, the vectors $\mathbf{l}_k$ and $\mathbf{r}_k$ (and hence $\mathbf{s}_k$) are independent of k. In this case, the vectors $\mathbf{s}_k$ above can all be represented by a single vector $\mathbf{s}$ whose two nonzero entries are given by $s_1 = -Lf_l$ and $s_{N-1} = -Rf_r$, where f_l and f_r are the (constant) boundary values and $L = R = m_6 + m_2 + m_4$. Correspondingly, the vectors $\bar{\mathbf{s}}_k$ in (3.1) can all be replaced by the single transformed vector $\bar{\mathbf{s}}$.

Returning now to the systems (2.6), the solution of each is the solution of the equivalent three-term recurrence relation with constant coefficients

$$\gamma_i y_{i(k-1)} + \lambda_i y_{ik} + \sigma_i y_{i(k+1)} = \bar{s}_i, \qquad y_{i0} = -\frac{\bar{b}_i}{\gamma_i}, \quad y_{iN} = -\frac{\bar{t}_i}{\sigma_i} \tag{3.2}$$

(assuming γ_i, $\sigma_i \neq 0$) where the solution vector $\mathbf{y}_i$ has entries y_{ik}, $k = 1, \dots, N-1$. Such an equation can be solved to obtain an explicit expression for y_{ik} as follows. The auxiliary equation is given by

$$\sigma_i \mu^2 + \lambda_i \mu + \gamma_i = 0$$

which has solutions

$$\mu_1(i) = \frac{-\lambda_i + \sqrt{\lambda_i^2 - 4\sigma_i\gamma_i}}{2\sigma_i}, \qquad \mu_2(i) = \frac{-\lambda_i - \sqrt{\lambda_i^2 - 4\sigma_i\gamma_i}}{2\sigma_i}. \tag{3.3}$$

For notational convenience, we will frequently omit explicit reference to the i-dependence of these roots. The complementary function is

$$y_{ik} = c_1\mu_1^k + c_2\mu_2^k$$

for some constants c_1 and c_2. The right-hand side value $\bar{s}_i$ is independent of k, so the constant particular function is given by

$$y_{ik} = \frac{\bar{s}_i}{\sigma_i + \lambda_i + \gamma_i},$$

and the general solution is

$$y_{ik} = c_1\mu_1^k + c_2\mu_2^k + \frac{\bar{s}_i}{\sigma_i + \lambda_i + \gamma_i}.$$

Applying the boundary conditions gives

$$\begin{aligned} c_1 &= -\frac{\bar{b}_i}{\gamma_i} - \frac{\bar{s}_i}{\sigma_i + \lambda_i + \gamma_i} - c_2, \\ c_2 &= \frac{1}{\mu_1^N - \mu_2^N}\left[\frac{\bar{t}_i}{\sigma_i} - \frac{\bar{b}_i}{\gamma_i}\mu_1^N - \frac{\bar{s}_i}{\sigma_i + \lambda_i + \gamma_i}(\mu_1^N - 1)\right]. \end{aligned}$$

Hence the solution of the recurrence relation is

$$\begin{aligned} y_{ik} = &-\frac{\bar{t}_i}{\sigma_i}\left[\frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}\right] \\ &+ \frac{\bar{s}_i}{\sigma_i + \lambda_i + \gamma_i}\left((1 - \mu_1^k) - (1 - \mu_1^N)\left[\frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}\right]\right) \\ &- \frac{\bar{b}_i}{\gamma_i}\left(\mu_1^k - \mu_1^N\left[\frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}\right]\right). \end{aligned}$$

We will write this as

$$y_{ik} = F_1(i)G_1(i,k) + F_2(i)G_2(i,k) + F_3(i)G_3(i,k) = \sum_{m=1}^{3} F_m(i)G_m(i,k), \tag{3.4}$$

where, recalling the dependence of μ_1 and μ_2 on i,

$$\begin{aligned} G_1(i,k) &= \frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}, \\ G_2(i,k) &= (1-\mu_1^k) - (1-\mu_1^N)\left[\frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}\right], \\ G_3(i,k) &= \mu_1^k - \mu_1^N\left[\frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}\right], \end{aligned}$$

and the weight functions

$$F_1(i) = -\frac{\bar{t}_i}{\sigma_i}, \qquad F_2(i) = \frac{\bar{s}_i}{\sigma_i + \lambda_i + \gamma_i}, \qquad F_3(i) = -\frac{\bar{b}_i}{\gamma_i} \tag{3.5}$$

involve the coefficient matrix entries and boundary condition information. Note that these weight functions are independent of k: for fixed i, the behaviour of $\mathbf{y}$ in the streamline (vertical) direction depends only on the functions $G_m(i,k)$. In addition, as $G_3(i,k) = 1 - G_1(i,k) - G_2(i,k)$, we have the following result.

Theorem 3.1. *The recurrence relation solution y_{ik} has the form*

$$y_{ik} = F_3(i) + [F_1(i) - F_3(i)]\, G_1(i,k) + [F_2(i) - F_3(i)]\, G_2(i,k). \tag{3.6}$$

As $F_1(i)$ is related to the top boundary values, $F_2(i)$ is related to the sum of the left and right boundary values (which have been assumed to be constant for this analysis) and $F_3(i)$ is related to the bottom boundary values, this result shows that different boundary conditions will dictate how the functions $G_1(i,k)$ and $G_2(i,k)$ combine to produce different two-dimensional recurrence relation solutions y_{ik}. This point will be investigated further in Section 5. We will use either (3.4) or (3.6) to represent the entries of $\mathbf{y}$, depending on which form is more useful at a particular point in the analysis.

Returning to (2.7), we see from (3.4) that $\mathbf{u}$ has entries

$$\begin{aligned} u_{jk} &= \sqrt{\frac{2}{N}} \sum_{i=1}^{N-1} \sin\frac{ij\pi}{N}\left(\sum_{m=1}^{3} F_m(i) G_m(i,k)\right) \\ &= \sum_{m=1}^{3}\left(\sqrt{\frac{2}{N}} \sum_{i=1}^{N-1} \sin\frac{ij\pi}{N} F_m(i) G_m(i,k)\right) \end{aligned} \tag{3.7}$$

for $j,k = 1,\ldots,N-1$. We emphasise that this is an explicit formula for the entries of the discrete solution vector $\mathbf{u}$.

4. Galerkin discretisation with bilinear elements

The analysis above is applicable to many standard discretisation methods. In this section, we will study the specific example of a Galerkin finite element discretisation of (2.1) using bilinear elements.

4.1. **The recurrence relation solution.** The entries of the coefficient matrix (2.2) in this case are

$$\begin{aligned} m_1 &= \tfrac{8}{3}\epsilon, & m_2 &= -\tfrac{1}{3}\epsilon, & m_3 &= \tfrac{1}{3}[h-\epsilon], \\ m_4 &= \tfrac{1}{12}[h-4\epsilon], & m_5 &= \tfrac{1}{3}[-h-\epsilon], & m_6 &= \tfrac{1}{12}[-h-4\epsilon]. \end{aligned} \tag{4.1}$$

For convenience, we introduce the notation

$$C_i = \cos\frac{i\pi}{N}$$

so that the eigenvalues (2.3) can be written as

$$\begin{aligned}\gamma_i &= \frac{1}{6}\left[-2\epsilon(1+2C_i) - h(2+C_i)\right]\\ \lambda_i &= \frac{2}{3}\left[\epsilon(1+2C_i) + 3\epsilon(1-C_i)\right]\\ \sigma_i &= \frac{1}{6}\left[-2\epsilon(1+2C_i) + h(2+C_i)\right]\end{aligned} \tag{4.2}$$

or

$$\gamma_i = \frac{1}{6}(-2\alpha_1 - \alpha_2), \quad \lambda_i = \frac{2}{3}(\alpha_1+\alpha_3), \quad \sigma_i = \frac{1}{6}(-2\alpha_1+\alpha_2),$$

$i = 1, \dots, N-1$, where

$$\alpha_1 = \epsilon(1+2C_i), \quad \alpha_2 = h(2+C_i), \quad \alpha_3 = 3\epsilon(1-C_i).$$

Substituting these into (3.3) gives the auxiliary equation roots

$$\mu_{1,2}(i) = \frac{-2(\alpha_1+\alpha_3) \pm \sqrt{4(2\alpha_1\alpha_3 + \alpha_3^2) + \alpha_2^2}}{\alpha_2 - 2\alpha_1}.$$

To express these roots explicitly in terms of ϵ and h, we first examine the square root term, and write

$$4(2\alpha_1\alpha_3 + \alpha_3^2) + \alpha_2^2 \quad = \quad h^2(2+C_i)^2\left\{1 + \frac{3(5+C_i)(1-C_i)}{(2+C_i)^2}\frac{1}{P_e^2}\right\},$$

where

$$P_e = \frac{\|\mathbf{w}\|h}{2\epsilon} = \frac{1}{2\epsilon N}$$

is the mesh Péclet number. Also,

$$2(\alpha_1+\alpha_3) = h\left\{\frac{1}{P_e}(4-C_i)\right\}, \qquad 2\alpha_1 - \alpha_2 = h\left\{-(2+C_i) + \frac{1}{P_e}(1+2C_i)\right\}.$$

Substituting these into $\mu_{1,2}$ gives

$$\mu_{1,2} = \frac{-\left[\dfrac{4-C_i}{2+C_i}\right]\dfrac{1}{P_e} \pm \sqrt{1 + \dfrac{3(5+C_i)(1-C_i)}{(2+C_i)^2}\dfrac{1}{P_e^2}}}{1 - \left[\dfrac{1+2C_i}{2+C_i}\right]\dfrac{1}{P_e}} \tag{4.3}$$

for the roots of recurrence relation (3.2) in the case of Galerkin approximation with bilinear elements.

4.2. **When do oscillations occur?** One question we would like to answer is, under what conditions is the recurrence relation solution (3.6) oscillatory? This point is addressed by the following theorem.

Theorem 4.1. *If* $\frac{2}{3}N < i \le N-1$, $G_1(i,k)$ *in* (3.6) *is an oscillatory function of* k *for any value of* P_e.

Proof. We have

$$G_1(i,k) = \frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N} = \left[\frac{\left(\dfrac{\mu_1}{\mu_2}\right)^k - 1}{\left(\dfrac{\mu_1}{\mu_2}\right)^N - 1} \right] \mu_2^{k-N} = \Theta(i,k)\,\mu_2^{k-N}. \tag{4.4}$$

As $|\mu_1/\mu_2| < 1$, $\Theta(i,k)$ is always positive. Hence if μ_2 is negative, $G_1(i,k)$ alternates in sign as k goes from 1 to $N-1$, that is, $G_1(i,k)$ is oscillatory for fixed i. From (4.3), the numerator of μ_2 is always negative so we have the conditions

$$\begin{cases} P_e < \phi_i \quad \Rightarrow \quad \mu_2 > 0,\ G_1(i,k) \text{ is not oscillatory} \\[2ex] P_e > \phi_i \quad \Rightarrow \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory,} \end{cases}$$

where

$$\phi_i = \frac{1+2C_i}{2+C_i}, \qquad i = 1, \dots, N-1. \tag{4.5}$$

But $P_e > 0$ by definition and $\phi_i < 0$ for $\frac{2}{3}N < i \le N-1$, so the second condition holds for these values of i independent of the value of P_e. Hence the corresponding functions $G_1(i,k)$, $\frac{2}{3}N < i \le N-1$, are oscillatory for any value of P_e. □

Corollary 4.1. *If the boundary conditions are such that the coefficient of* $G_1(i,k)$ *in* (3.6) *is nonzero, the recurrence relation solution* $\mathbf{y}$ *will exhibit oscillations in the direction of the flow for any value of* P_e.

We conclude this section with a few brief remarks concerning the above theorem.

Remark 1. Corollary 4.1 does not imply that the discrete solution $\mathbf{u}$ is oscillatory for any P_e. This issue will be explored in some detail in Section 5.

Remark 2. For other values of i, whether or not $G_1(i,k)$ exhibits oscillations depends on the value of P_e. In particular, if $P_e > 1$, then $G_1(i,k)$ is an oscillatory function of k for every $i \in \{1, \dots, N-1\}$.

Remark 3. From (4.4), the parity of the oscillations in $G_1(i,k)$ is independent of i, that is, the sign of $G_1(i,k)$ for a particular index k is the same for any i.

Remark 4. If $P_e = \phi_i$ for any i, then the eigenvalue σ_i in (4.2) is zero and (3.2) reduces to a two-term recurrence relation. This means that the analysis in Section 3 is not directly applicable, but the two-term recurrence can be solved in a similar way to obtain a formula for y_{ik}: the details are omitted here.

Remark 5. As $0 < \mu_1 < 1$, $G_2(i,k) = (1-\mu_1^k) - (1-\mu_1^N)G_1(i,k)$ will be oscillatory if and only if $G_1(i,k)$ is oscillatory, although the oscillations will occur about the function $1-\mu^k$ rather than zero in this case.

4.3. **Characterising oscillations in the recurrence relation solution.** We now use (4.3) to gain a more detailed understanding of the functions $G_m(i,k)$, $m = 1, 2$ in (3.6), with a view to characterising the oscillations which occur in the direction of the flow for large values of P_e. In particular, we highlight certain trends in the behaviour of these functions when $\epsilon \ll h$.

We begin by simplifying (4.3) using the approximation to the square root term given by

$$\left\{1 + \frac{3(5+C_i)(1-C_i)}{(2+C_i)^2}\frac{1}{P_e^2}\right\}^{1/2} \quad \simeq \quad \left\{1 + \frac{3}{2}\frac{(5+C_i)(1-C_i)}{(2+C_i)^2}\frac{1}{P_e^2}\right\}$$

which assumes that P_e^{-2} is small. The formulae for μ_1 and μ_2 then become

$$\mu_1 = \frac{1 - \left[\dfrac{4-C_i}{2+C_i}\right]\dfrac{1}{P_e} + \dfrac{3}{2}\dfrac{(5+C_i)(1-C_i)}{(2+C_i)^2}\dfrac{1}{P_e^2}}{1 - \left[\dfrac{1+2C_i}{2+C_i}\right]\dfrac{1}{P_e}},$$

$$\mu_2 = \frac{-1 - \left[\dfrac{4-C_i}{2+C_i}\right]\dfrac{1}{P_e} - \dfrac{3}{2}\dfrac{(5+C_i)(1-C_i)}{(2+C_i)^2}\dfrac{1}{P_e^2}}{1 - \left[\dfrac{1+2C_i}{2+C_i}\right]\dfrac{1}{P_e}}.$$

When these roots appear in the functions G_1 and G_2 in (3.6), they are raised to some power. Neglecting terms of order P_e^{-2} and higher, we obtain the following approximate expressions for powers of the roots:

$$\begin{aligned}
\mu_1^k &= \left(1 - \left[\frac{4-C_i}{2+C_i}\right]\frac{1}{P_e}\right)^k \left(1 - \left[\frac{1+2C_i}{2+C_i}\right]\frac{1}{P_e}\right)^{-k} \\
&\simeq \left(1 - \left[\frac{4-C_i}{2+C_i}\right]\frac{k}{P_e}\right)\left(1 + \left[\frac{1+2C_i}{2+C_i}\right]\frac{k}{P_e}\right) \\
&\simeq 1 + 3\left[\frac{C_i-1}{2+C_i}\right]\frac{k}{P_e},
\end{aligned}$$

$$\begin{aligned}
\mu_2^k &= \left(-1 - \left[\frac{4-C_i}{2+C_i}\right]\frac{1}{P_e}\right)^k \left(1 - \left[\frac{1+2C_i}{2+C_i}\right]\frac{1}{P_e}\right)^{-k} \\
&\simeq \left((-1)^k - (-1)^{k-1}\left[\frac{4-C_i}{2+C_i}\right]\frac{k}{P_e}\right)\left(1 + \left[\frac{1+2C_i}{2+C_i}\right]\frac{k}{P_e}\right) \\
&\simeq (-1)^k\left(1 + \left[\frac{5+C_i}{2+C_i}\right]\frac{k}{P_e}\right).
\end{aligned}$$

Note that although the terms including P_e^{-2} which we have omitted here may have coefficients involving powers of N $(= 1/h)$, we are working under the assumption that $\epsilon \ll h$ and so these terms are still small in size relative to those which we have retained.

We now use the results above to derive approximations to the functions $G_1(i,k)$ and $G_2(i,k)$ which appear in (3.6), simplifying the algebra by using the notation

$$\mu_1^k \simeq 1 + \psi_1\frac{k}{P_e}, \qquad \mu_2^k \simeq (-1)^k\left(1 + \psi_2\frac{k}{P_e}\right) \tag{4.6}$$

where

$$\psi_1(i) = 3\left[\frac{C_i - 1}{2 + C_i}\right] \qquad \text{and} \qquad \psi_2(i) = \frac{5 + C_i}{2 + C_i}. \tag{4.7}$$

With this notation, we have

$$\begin{aligned} G_1(i,k) = \frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N} &\simeq \frac{\left(1 + \psi_1 \dfrac{k}{P_e}\right) - (-1)^k \left(1 + \psi_2 \dfrac{k}{P_e}\right)}{\left(1 + \psi_1 \dfrac{N}{P_e}\right) - (-1)^N \left(1 + \psi_2 \dfrac{N}{P_e}\right)} \\ &= \left[(1 - (-1)^k) + (\psi_1 - (-1)^k \psi_2)\frac{k}{P_e}\right] \\ &\quad \times \left[(1 - (-1)^N) + (\psi_1 - (-1)^N \psi_2)\frac{N}{P_e}\right]^{-1} \end{aligned}$$

We consider the cases of odd and even N separately.

If N is odd, we have

$$\begin{aligned} G_1(i,k) &\simeq \left[(1 - (-1)^k) + (\psi_1 - (-1)^k \psi_2)\frac{k}{P_e}\right]\left[2 + (\psi_1 + \psi_2)\frac{N}{P_e}\right]^{-1} \\ &\simeq \frac{1}{2}\left[(1 - (-1)^k) + (\psi_1 - (-1)^k \psi_2)\frac{k}{P_e}\right]\left[1 + \frac{(\psi_1 + \psi_2)}{2}\frac{N}{P_e}\right]^{-1}. \end{aligned}$$

Thus

$$G_1 \simeq \frac{1 + \left[\dfrac{1 + 2C_i}{2 + C_i}\right]\dfrac{k}{P_e}}{1 + \left[\dfrac{1 + 2C_i}{2 + C_i}\right]\dfrac{N}{P_e}} \quad (k \text{ odd}), \qquad G_1 \simeq \frac{\left[\dfrac{C_i - 4}{2 + C_i}\right]\dfrac{k}{P_e}}{1 + \left[\dfrac{1 + 2C_i}{2 + C_i}\right]\dfrac{N}{P_e}} \quad (k \text{ even}). \tag{4.8a}$$

If N is even, we have

$$G_1(i,k) \simeq \left[(1 - (-1)^k) + (\psi_1 - (-1)^k \psi_2)\frac{k}{P_e}\right]\left[(\psi_1 - \psi_2)\frac{N}{P_e}\right]^{-1}.$$

Thus

$$G_1 \simeq \left[\frac{2 + C_i}{C_i - 4}\right]\frac{P_e}{N} + \left[\frac{1 + 2C_i}{C_i - 4}\right]\frac{k}{N} \quad (k \text{ odd}), \qquad G_1 \simeq \frac{k}{N} \quad (k \text{ even}). \tag{4.8b}$$

To characterise the behaviour of these functions for large P_e, we consider the limit as $P_e \to \infty$ (e.g., as $\epsilon \to 0$ for a fixed value of N). In this limit, we have

$$G_1(i,k) \simeq \quad \begin{array}{c|cc} & N \text{ is odd} & N \text{ is even} \\ \hline k \text{ is odd} & 1 & -\infty \\ k \text{ is even} & 0 & \dfrac{k}{N} \end{array} \tag{4.9}$$

For both odd and even N, G_1 displays highly oscillatory behaviour in the streamline direction (that is, for fixed i). The nature of these oscillations is, however, slightly different: for odd N, the oscillations remain bounded between 0 and 1 as $P_e \to \infty$, whereas for even N, the oscillations grow unboundedly.

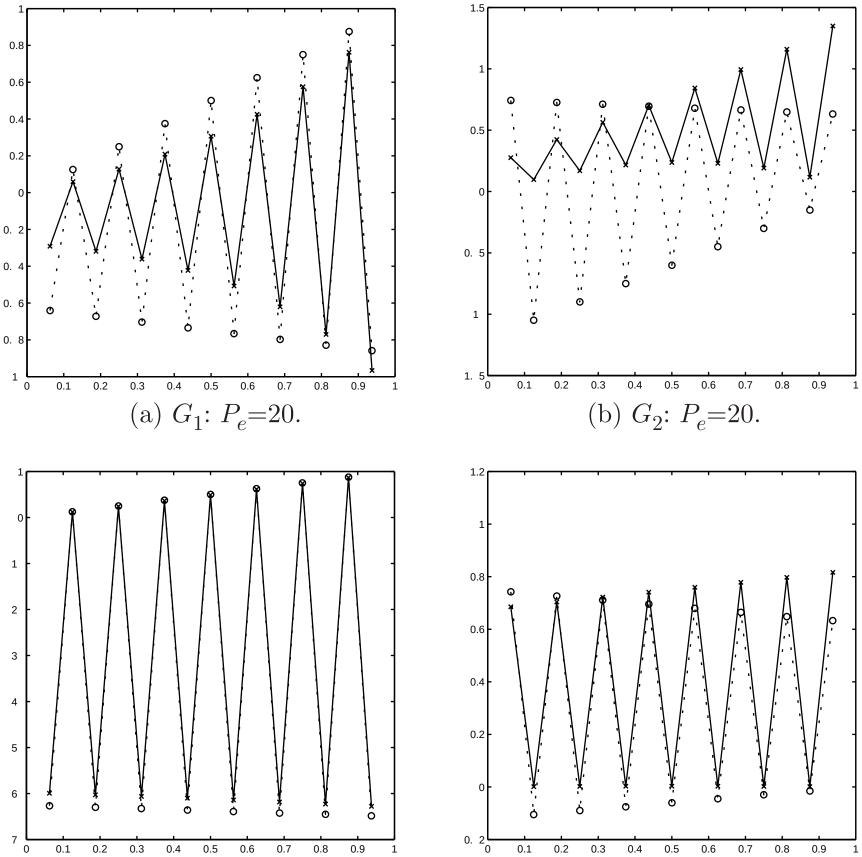
Similarly, we have

$$G_2(i,k) = (1-\mu_1^k) - (1-\mu_1^N)\left[\frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}\right] \simeq -\psi_1 \frac{1}{P_e}(k - NG_1)$$

so if N is odd,

(4.10a)

$$G_2 \simeq \frac{3\left[\dfrac{1-C_i}{2+C_i}\right]\dfrac{(k-N)}{P_e}}{1+\left[\dfrac{1+2C_i}{2+C_i}\right]\dfrac{N}{P_e}} \quad (k \text{ odd}), \quad G_2 \simeq \frac{3\left[\dfrac{1-C_i}{2+C_i}\right]\dfrac{k}{P_e}}{1+\left[\dfrac{1+2C_i}{2+C_i}\right]\dfrac{N}{P_e}} \quad (k \text{ even}),$$



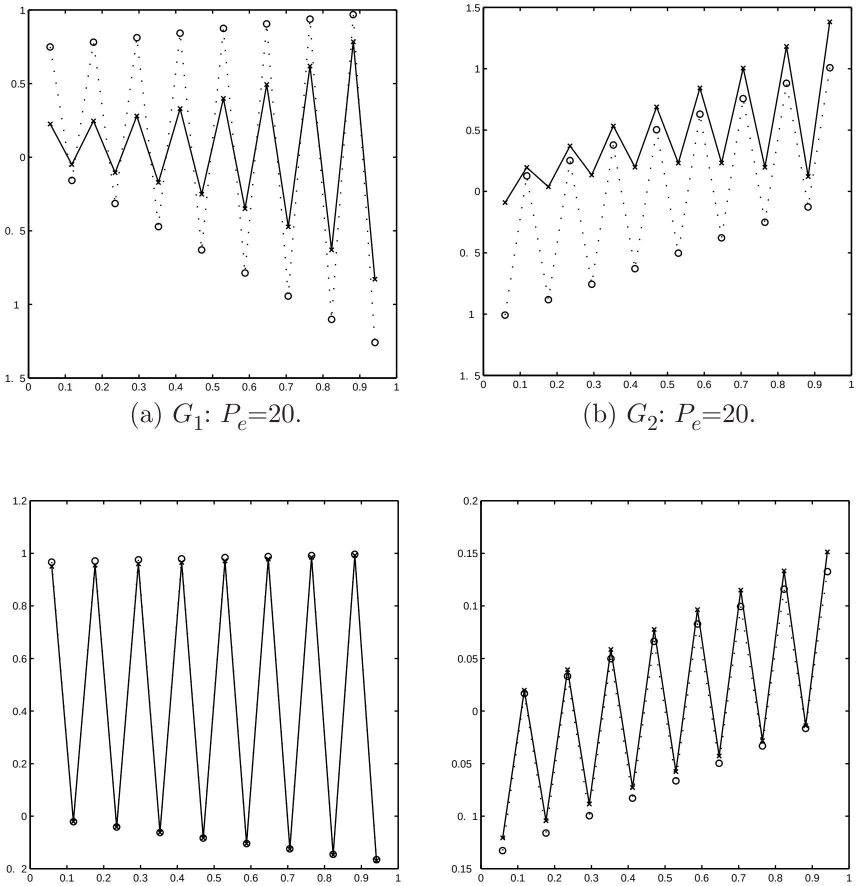
(a) G_1: P_e=20.

(b) G_2: P_e=20.

(c) G_1: P_e=200.

(d) G_2: P_e=200.

FIGURE 4.1. Plots of G_1 and G_2 (solid line, x) and approximations (dotted line, o) with $N = 16$ and $i = 8$.



(a) G_1: P_e=20.

(b) G_2: P_e=20.

(c) G_1: P_e=200.

(d) G_2: P_e=200.

FIGURE 4.2. Plots of G_1 and G_2 (solid line, x) and approximations (dotted line, o) with $N = 17$ and $i = 9$.

and if N is even,

$$\begin{aligned} G_2 &\simeq \left[\frac{(1+2C_i)(1-3C_i)}{(C_i-1)(2+C_i)}\right]\frac{k}{N} + \left[\frac{C_i-1}{2+C_i}\right] \ (k \text{ odd}), \\ G_2 &\simeq 3\left[\frac{1-C_i}{2+C_i}\right]\frac{(k-N)}{P_e} \ (k \text{ even}). \end{aligned} \tag{4.10b}$$

In the limit as $P_e \to \infty$ we have

(4.11)

		N is odd	N is even
$G_2(i,k) \simeq$	k is odd	0	$\left[\frac{C_i-1}{2+C_i}\right] + \left[\frac{(1+2C_i)(1-3C_i)}{(C_i-1)(2+C_i)}\right]\frac{k}{N}$
	k is even	0	0

This time the two cases have different characters: if N is odd, the oscillations in $G_2(i,k)$ will die out as P_e increases, whereas if N is even, the oscillations remain, although their size is bounded independent of P_e.

We point out here that the type of oscillations in the streamline direction which are caused by (4.8) and (4.10) are in each case of the same character as oscillations in certain discrete solutions of the one-dimensional convection-diffusion equation

$$-\epsilon u'' + wu' = f, \qquad u(0) = \xi_0, \quad u(1) = \xi_N \tag{4.12}$$

where w is a positive constant and f, ξ_0 and ξ_N are given. Specifically, consider the Galerkin method with linear elements applied to (4.12) with a uniform discretisation on $N+1$ points in $[0,1]$. The function $G_1(i,k)$ (for fixed i) exhibits similar asymptotic behaviour to the discrete solution of (4.12) with $f=0$, $\xi_1=0$, and $\xi_2=1$, and $G_2(i,k)$ (for fixed i) exhibits asymptotic behaviour like the discrete solution of (4.12) with $f=N/P_e$, $\xi_1=0$ and $\xi_2=0$ (where P_e is the mesh Péclet number (1.5) of the one-dimensional problem).

Representative plots of the exact (solid line, x) and approximate (dotted line, o) expressions for $G_1(i,k)$ and $G_2(i,k)$ for fixed i are plotted in Figures 4.1 and 4.2. Note that these plots have different scales: the oscillations in G_1 are in general of greater magnitude than those in G_2. The approximations (4.8) and (4.10) clearly capture the nature of the exact functions.

4.4. **The weight functions.** Recall from (3.6) that the functions $F_m(i)$, $m=1,2,3$ regulate how $G_1(i,k)$ and $G_2(i,k)$ combine in y_{ik}. They therefore play an important role in dictating the nature of oscillations. These weight functions come from transforming the right-hand side vector containing the boundary value and matrix coefficient information as described in Section 3. As shown in the Appendix, we can derive the expressions

$$F_1(i) = \sqrt{\frac{2}{N}}\sum_{p=1}^{N-1} f_t(x_p)\sin\frac{pi\pi}{N}, \qquad F_3(i) = \sqrt{\frac{2}{N}}\sum_{p=1}^{N-1} f_b(x_p)\sin\frac{pi\pi}{N}, \tag{4.13a}$$

and, for the special case where the constant left and right boundary values f_l and f_r are equal,

$$F_2(i) = f_l\sqrt{\frac{2}{N}}\sum_{p=1}^{N-1}\sin\frac{pi\pi}{N}. \tag{4.13b}$$

We may combine (4.13) and (3.6) to obtain the following result.

Theorem 4.2. *If f_l=f_r, the recurrence relation solution y_{ik} has the form*

$$\begin{aligned} y_{ik} \;&=\; \sqrt{\frac{2}{N}}\sum_{p=1}^{N-1} f_b(x_p)\sin\frac{pi\pi}{N} + \sqrt{\frac{2}{N}}\sum_{p=1}^{N-1}[f_t(x_p)-f_b(x_p)]\sin\frac{pi\pi}{N}G_1(i,k) \\ &\quad + \sqrt{\frac{2}{N}}\sum_{p=1}^{N-1}[f_l - f_b(x_p)]\sin\frac{pi\pi}{N}G_2(i,k). \end{aligned} \tag{4.14}$$

We emphasise the important difference between results (3.6) and (4.14): the latter expression involves the actual Dirichlet boundary functions themselves, as opposed to the more complicated weight functions $F_m(i)$ given by (3.5). The effect of each boundary condition on y_{ik} is therefore much more readily seen in (4.14). If there is a difference between the bottom and top boundary functions, this will introduce oscillations via $G_1(i,k)$; similarly, a difference between the bottom boundary function and the left (right) boundary function will result in oscillations coming from $G_2(i,k)$.

5. The full two-dimensional solution

We have established that the recurrence relation solution $\mathbf{y}$ in (4.14) is influenced by two functions G_1 and G_2 which each look like the solution to a particular one-dimensional convection-diffusion problem. In this section we explore the implications of this for the final two-dimensional solution $\mathbf{u}$, that is, we examine the effect of the Fourier transformations (2.7).

5.1. The discrete solution u: An outflow boundary layer example.

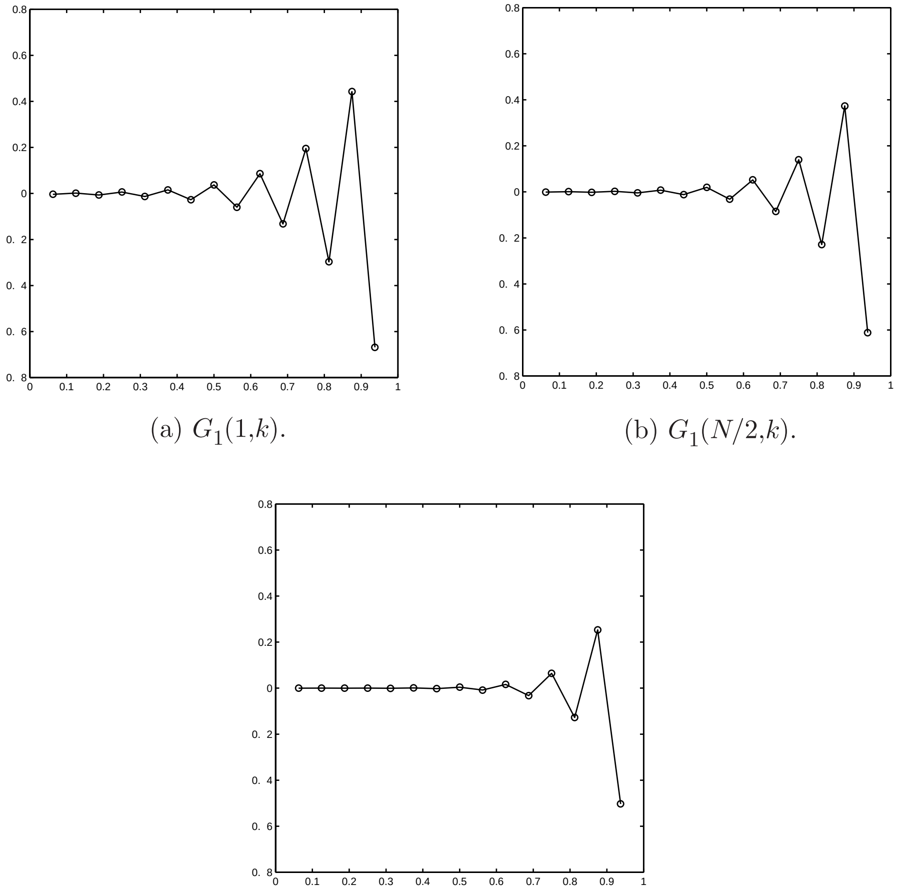
Ideally, we would like to obtain an expression equivalent to (4.14) for $\mathbf{u}$. As (4.14) is a sum, we may examine the effect of transformation (2.7) term by term. For the first term, this is straightforward: due to the orthogonality of the eigenvectors $\mathbf{v}_j$ in (2.4), we have

$$\begin{aligned}
&\sqrt{\frac{2}{N}}\sum_{i=1}^{N-1}\sin\frac{ij\pi}{N}\left\{\sqrt{\frac{2}{N}}\sum_{p=1}^{N-1}f_b(x_p)\sin\frac{pi\pi}{N}\right\}\\
&\qquad=\sum_{p=1}^{N-1}f_b(x_p)\left\{\sum_{i=1}^{N-1}\sqrt{\frac{2}{N}}\sin\frac{ji\pi}{N}\sqrt{\frac{2}{N}}\sin\frac{pi\pi}{N}\right\}\\
&\qquad=\sum_{p=1}^{N-1}f_b(x_p)\mathbf{v}_j^T\mathbf{v}_p\\
&\qquad=f_b(x_j).
\end{aligned}\tag{5.1}$$

That is, the first term in $\mathbf{u}$ is just the bottom boundary function $f_b(\mathbf{x})$.

Unfortunately, it is nontrivial to repeat this for the other terms in (4.14) due to the effect of the sine transform in (2.7) on each $G_m(i,k)$. We can, however, make a number of useful observations. Firstly, there is the obvious point that if G_1 or G_2 has a zero coefficient in (4.14) due to equality in the relevant boundary functions, then the resulting $\mathbf{u}$ will also have no contribution from that function. To illustrate a more subtle point, we look at the case where $f_t = 1$ and $f_b = f_l = f_r = 0$, so that u_{jk} consists of a contribution from the function G_1 alone. Consider the solution $\mathbf{u}$ along a specified vertical grid line, that is, fix j. From (2.7), (4.13) and (4.14) we have

$$\begin{aligned}
u_{jk} &= \sqrt{\frac{2}{N}}\sum_{i=1}^{N-1}\sin\frac{ij\pi}{N}\left\{\sqrt{\frac{2}{N}}\sum_{p=1}^{N-1}\sin\frac{ip\pi}{N}G_1(i,k)\right\}\\
&= \frac{2}{N}\left\{\sum_{i=1}^{N-1}\sin\frac{ij\pi}{N}\sum_{p=1}^{N-1}\sin\frac{ip\pi}{N}G_1(i,k)\right\}\\
\Rightarrow u_{jk} &= \frac{2}{N}\left\{\sum_{i=1}^{N-1}d_{ij}G_1(i,k)\right\},
\end{aligned}\tag{5.2}$$



(a) $G_1(1,k)$.

(b) $G_1(N/2,k)$.

(c) $G_1(N{-}1,k)$.

FIGURE 5.1. Functions $G_1(i,k)$ for $N = 16$ and $P_e = 5$.

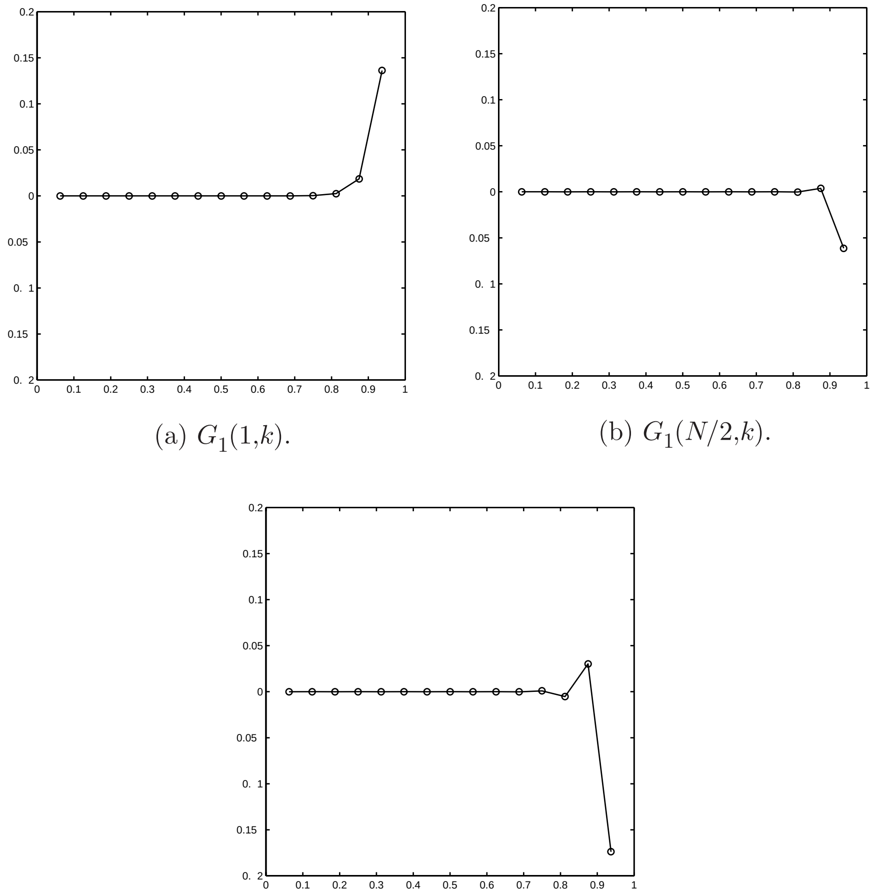
where

$$d_{ij} = \sin\frac{ij\pi}{N}\sum_{p=1}^{N-1}\sin\frac{ip\pi}{N}. \tag{5.3}$$

That is, u_{jk} is a linear combination of the functions $G_1(i,k)$ for $i = 1,\ldots,N-1$. Examples of these functions for $N = 16$ are plotted in Figure 5.1 for $P_e = 5$ and in Figure 5.2 for $P_e = 0.75$.

We can use the representation (5.2) to obtain insight into the quality of the solution. The identity

$$\sum_{p=1}^{N-1}\sin\frac{ip\pi}{N} = \frac{\sin\dfrac{i\pi}{2}\sin\dfrac{(N-1)i\pi}{2N}}{\sin\dfrac{i\pi}{2N}}$$



(a) $G_1(1,k)$.

(b) $G_1(N/2,k)$.

(c) $G_1(N{-}1,k)$.

FIGURE 5.2. Functions $G_1(i,k)$ for $N = 16$ and $P_e = 0.75$.

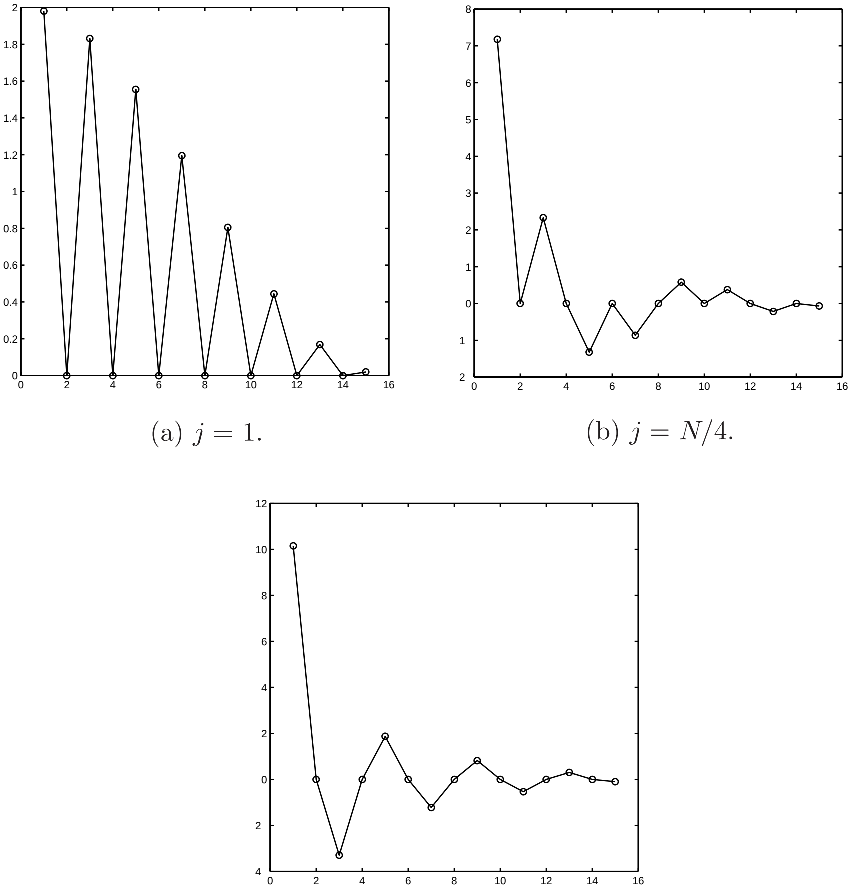
([7], 19.40) leads to the simplified expression for the coefficients d_{ij},

$$d_{ij} = \left(\sin^2 \frac{i\pi}{2}\right) \left(\sin \frac{ij\pi}{N}\right) \left(\frac{\cos \frac{i\pi}{2N}}{\sin \frac{i\pi}{2N}}\right). \tag{5.4}$$

It follows immediately that $d_{ij} = 0$ for even values of i. For the case $j = 1$, the nonzero values are

$$d_{i1} = 2\cos^2 \frac{i\pi}{2N}\,,$$

which are all positive. Now recall from the proof of Theorem 4.1 that if $P_e > \phi_i$ in (4.5) then $G_1(i,k)$ is an oscillatory function of k. As we observed in subsection 4.2, if $P_e > 1$, then $G_1(i,k)$ is oscillatory for every i, with oscillations of the same parity in each case (see Figure 5.1). Thus, for $j = 1$, u_{jk} is an oscillatory function of k, and we have established the following result:



(a) $j = 1$.

(b) $j = N/4$.

(c) $j = N/2$.

FIGURE 5.3. Coefficients d_{ij} for $N = 16$.

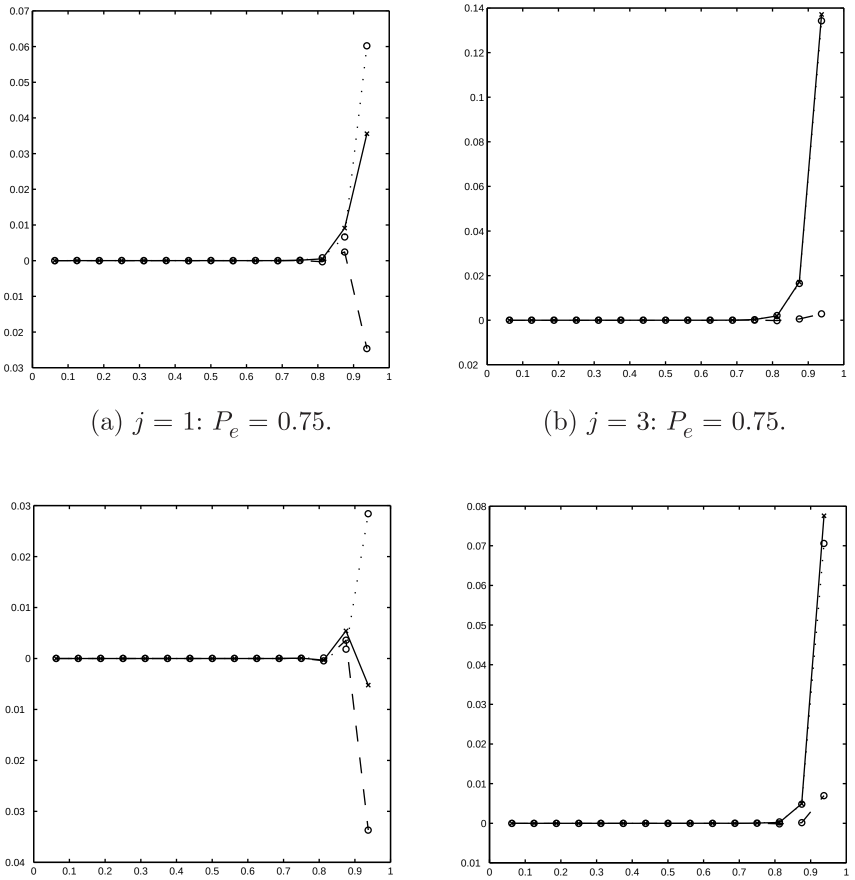
Theorem 5.1. *If $P_e > 1$ in the Galerkin discretisation of* (2.1) *with bilinear elements, then for some choice of boundary conditions the solution* **u** *exhibits oscillations.*

This corresponds to the well-known restriction in the one-dimensional convection-diffusion problem cited in the introduction.

In contrast to the one-dimensional case, the converse of Theorem 5.1 is not true in two dimensions. To see this, let i^* be the lowest value of $i \in \{1, \dots, N-1\}$ such that $P_e > \phi_i$ and write

$$\begin{aligned} u_{jk} &= \frac{2}{N} \sum_{i=1}^{i^*-1} d_{ij} G_1(i,k) + \frac{2}{N} \sum_{i=i^*}^{N-1} d_{ij} G_1(i,k) \\ &= S_{\text{smooth}} + S_{\text{osc}}, \end{aligned}$$

where S_{smooth} and S_{osc} are sums of smooth and oscillatory functions respectively. The overall behaviour of the solution u_{jk} for a particular j is determined by the relative size of these two terms: if S_{smooth} dominates, u_{jk} will be smooth and if



(a) $j = 1$: $P_e = 0.75$.

(b) $j = 3$: $P_e = 0.75$.

(c) $j = 1$: $P_e = 0.85$.

(a) $j = 3$: $P_e = 0.85$.

FIGURE 5.4. Comparison of S_{smooth} (dotted line, o) and S_{osc} (dashed line, o) with u_{jk} (solid line, x).

S_{osc} dominates, u_{jk} will be oscillatory. As P_e decreases, i^* increases so that most of the functions $G_1(i, k)$ are smooth and S_{smooth} contains most of the terms in (5.2). Furthermore, the magnitude of the nonzero coefficients d_{ij} decreases rapidly as i goes from 1 to $N-1$: sample plots of d_{ij} against i for $N = 16$ are shown in Figure 5.3 (note that we show values for j between 1 and $N/2$ only as $d_{i(N-j)} = d_{ij}$). Thus, for small enough values of P_e, S_{smooth} will dominate. In this sense, the Fourier transformations in (2.7) have a 'smoothing' effect on the oscillatory recurrence relation solution $\mathbf{y}$.

Figure 5.4 shows a comparison of S_{smooth} (dotted line, o) and S_{osc} (dashed line, o) with u_{jk} (solid line, x) for $N = 16$ with two values of P_e which lead to different behaviour. In plots (a) and (b), $P_e = 0.75$ and S_{smooth} is larger in magnitude than S_{osc}, producing a smooth u_{jk}. We have observed that the size of S_{osc} decreases relative to that of S_{smooth} as j goes from 1 to $N/2$, so that S_{osc} is most influential near the left and right boundaries ($j = 1$ and $j = N - 1$). For j as small as 3 (Figure 5.4(b)) it is already difficult to distinguish between the plots of S_{smooth} and

u_{3k} as S_{osc} is very small. Plots (c) and (d) show the equivalent data for $P_e = 0.85$. In this case, $|S_{\text{osc}}|$ is larger than $|S_{\text{smooth}}|$ when $j = 1$, resulting in an oscillatory u_{1k}. The discrete solution $\mathbf{u}$ is therefore oscillatory, although $P_e < 1$.

5.2. **Further examples and the effects of characteristic boundary layers.** To complete this discussion, we present some illustrations of the behaviour of $\mathbf{u}$ using a set of examples chosen to highlight some common features of convection-diffusion problems.

Problem I. Suppose we have the uniform boundary conditions

$$f_r(\mathbf{y}) = f_l(\mathbf{y}) = f_t(\mathbf{x}) = f_b(\mathbf{x}) = 1$$

which gives the exact solution $u = 1$ everywhere. This perfectly smooth solution can be recovered by setting $f_t = f_b = f_l = 1$ in (4.14) to get

$$y_{ik} = \sqrt{\frac{2}{N}} \sum_{p=1}^{N-1} \sin \frac{pi\pi}{N}.$$

Note that G_1 and G_2 are not present. Under transformation (2.7) we obtain $u_{jk} = 1$, as in (5.1).

Problem II. We return to the example considered in subsection 5.1 and set

$$f_t(\mathbf{x}) = 1; \qquad f_b(\mathbf{x}) = f_r(\mathbf{y}) = f_l(\mathbf{y}) = 0,$$

so that the solution has an exponential boundary layer of width ϵ along the top boundary. As f_b and f_l are zero, the recurrence relation solution (4.14) is

$$y_{ik} = \sqrt{\frac{2}{N}} \sum_{p=1}^{N-1} \sin \frac{pi\pi}{N} G_1(i,k), \tag{5.5}$$

with resulting full solution (5.2) which is highly oscillatory for large P_e. Representative discrete solutions $\mathbf{u}$ for $N = 16$ are illustrated in Figures 5.5 (a) and (d). The propagation of oscillations away from the boundary layer is caused by the behaviour of G_1. This behaviour is typical of any problem where the top and bottom boundary conditions differ.

Problem III. For this example, we choose

$$f_b(\mathbf{x}) = f_t(\mathbf{x}) = 0; \qquad f_l(\mathbf{y}) = f_r(\mathbf{y}) = 1,$$

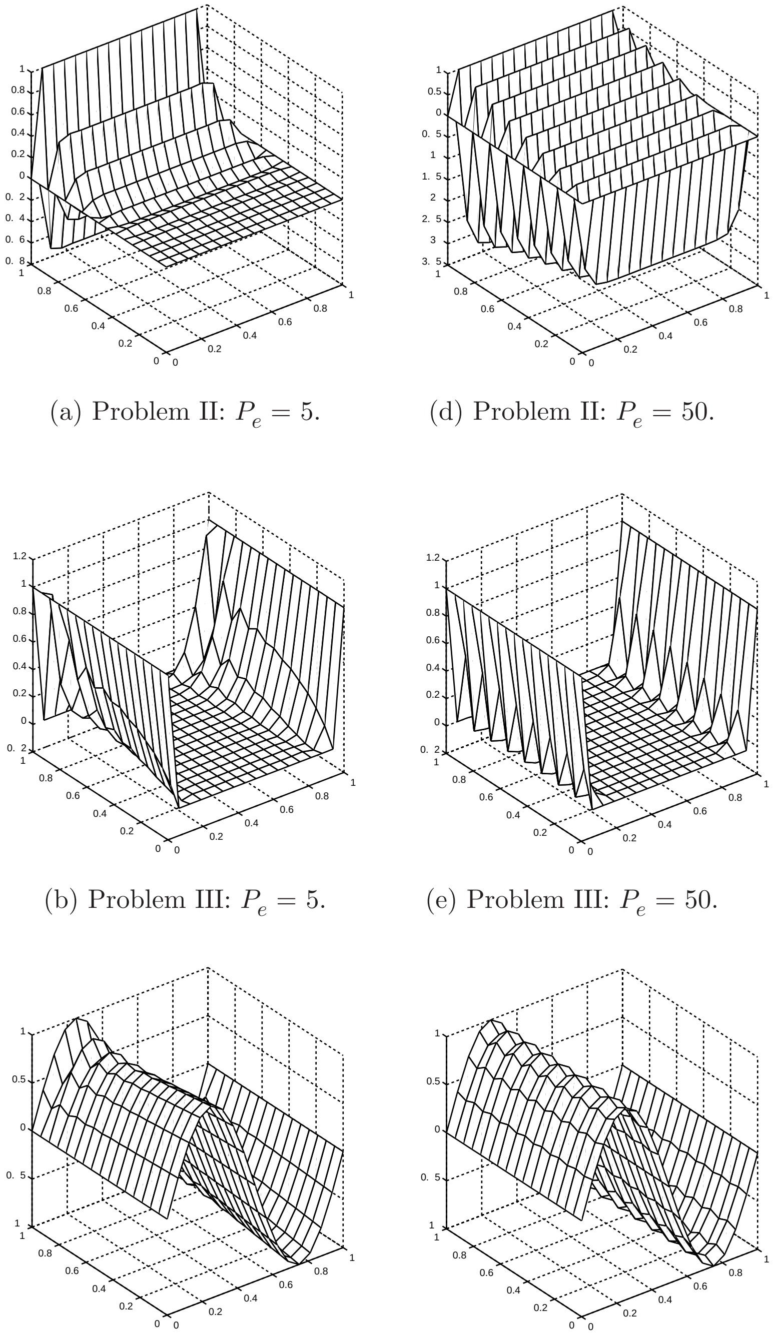
so that the solution has characteristic layers on both sides of the domain. In this case, recurrence relation solution (4.14) becomes

$$y_{ik} = \sqrt{\frac{2}{N}} \sum_{p=1}^{N-1} \sin \frac{pi\pi}{N} G_2(i,k).$$

Note that this is the same as the recurrence relation solution (5.5) of Problem II, except with G_2 in place of G_1. The related full two-dimensional solution is

$$u_{jk} = \frac{2}{N} \left\{ \sum_{i=1}^{N-1} d_{ij} G_2(i,k) \right\} \tag{5.6}$$

(cf. (5.2)) where the coefficients d_{ij} are again given by (5.3). We may therefore apply the analysis of subsection 5.1 in the same way to Problem III, replacing G_1 by G_2 throughout. In particular, u_{jk} is a sum of smooth and oscillatory components.



(a) Problem II: $P_e = 5$.

(d) Problem II: $P_e = 50$.

(b) Problem III: $P_e = 5$.

(e) Problem III: $P_e = 50$.

(c) Problem IV: $P_e = 5$.

(f) Problem IV: $P_e = 50$.

FIGURE 5.5. Galerkin finite element solutions with $N = 16$.

Moreover, G_2 is oscillatory if and only if G_1 is oscillatory (see Remark 5 at the end of subsection 4.2), so that the critical value of P_e in terms of the onset of oscillations is the same for both problems. The conclusion here is that there are oscillations in the streamline direction caused solely by the fact that the characteristic boundary conditions are different from those at the bottom (inflow) boundary; the precise effect of this difference can be seen in the last term of (4.14).

Sample solutions for $N = 16$ with $P_e = 5$ and $P_e = 50$ are shown in Figures 5.5 (b) and (e). It can be seen that the oscillations are much larger near the characteristic boundary layers than in the interior of the domain. Some insight into this fact can be obtained from closer consideration of the components of (5.6). In particular, since the oscillations of $G_1(i,k)$ with respect to k have the same parity, those of G_2 will augment one another when their multipliers d_{ij} have the same sign, but they will tend to cancel when the signs vary. From (5.4), we see that for $j = 1$ or $N-1$ (i.e., for the grid lines next to the left and right boundaries) the coefficients d_{ij} are all positive, whereas they alternate in sign for the other values of j shown in Figure 5.3, causing cancellation. These trends are further accentuated by the fact that $G_2(i,k)$ is larger for large i, but except for indices j near 1 (and $N-1$), the corresponding multipliers d_{ij} decrease in size dramatically as i increases. Thus, the largest oscillations are near the layers, when $j = 1$ and $N-1$. We have confirmed numerically, however, that the solutions for $P_e > 1$ oscillate along all vertical lines.

The pictures in Figure 5.5 also show that oscillations propagate away from the outflow boundary as P_e increases. We know from the analysis of G_2 in subsection 4.3 that if N is odd, these oscillations will die away as $P_e \to \infty$, although for any finite N they are always present: for example, with P_e=50, the largest oscillations are of magnitude 0.3 when $N = 17$ as opposed to 0.5 when $N = 16$ as in Figure 5.5 (e).

Problem IV. Here we apply the boundary conditions

$$f_b(\mathbf{x}) = f_t(\mathbf{x}) = \sin(2\pi\mathbf{x}); \quad f_r(\mathbf{y}) = f_l(\mathbf{y}) = 0.$$

In the limit as $\epsilon \to 0$, the solution of this problem is the smooth function $\mathbf{u} = \sin(2\pi\mathbf{x})$ everywhere. As $f_b(\mathbf{x}) = f_t(\mathbf{x})$ and $f_l = 0$, (4.14) gives

$$y_{ik} = \sqrt{\frac{2}{N}}\sum_{p=1}^{N-1} \sin\frac{2p\pi}{N}[1 - G_2(i,k)]\sin\frac{pi\pi}{N}.$$

As in the previous example, there is no contribution from $G_1(i,k)$ here, but there are oscillations in $\mathbf{u}$ for $P_e > 1$ due to the presence of $G_2(i,k)$. Looking now at the full solution, we have

$$\begin{aligned} u_{jk} &= \sum_{i=1}^{N-1}\sin\frac{ij\pi}{N}[1-G_2(i,k)]\left\{\sum_{p=1}^{N-1}\sqrt{\frac{2}{N}}\sin\frac{2p\pi}{N}\sqrt{\frac{2}{N}}\sin\frac{pi\pi}{N}\right\} \\ &= \sum_{i=1}^{N-1}\sin\frac{ij\pi}{N}[1-G_2(i,k)]\mathbf{v}_2^T\mathbf{v}_i \\ &= \sin\frac{2j\pi}{N}[1-G_2(2,k)]. \end{aligned}$$

The oscillations in this case are therefore relatively small as $G_2(2,k)$ is small for all k. Note, however, that choosing $f_b(\mathbf{x}) = f_t(\mathbf{x}) = \sin(m\pi\mathbf{x})$ for an integer $2 < m < N-1$ increases the size of the oscillations: the solution would involve $G_2(m,k)$ and, as noted above, $G_2(i,k)$ is larger for large i.

In contrast to the previous example, the solution here is not dramatically worse next to the characteristic boundary layers; here the oscillatory part of the solution is almost independent of j.

6. Summary

In this paper, we have derived closed form expressions for discrete solutions to the two-dimensional convection-diffusion equation on a square that show how oscillations arise and are affected by boundary conditions. Using bilinear finite elements as a concrete example, the analysis gives rigorous justification to the general belief that a mesh Péclet number greater than one leads to oscillatory solutions. The analysis is applied to a problem with a unidirectional flow that is aligned with the grid. However, even for this simple model, the results show several significant differences from the one-dimensional case. In particular, it is possible for the discrete solution to contain oscillations even if the mesh Péclet number is less than one. Moreover, the oscillations are affected by two-dimensional phenomena. As with one-dimensional problems, oscillations may arise because Dirichlet boundary conditions are different at the inflow and outflow, but two-dimensional effects, in particular, nearness to the side boundaries, also influence the size and quality of the oscillations: the solutions tend to be rougher near the side boundaries than in the middle of the domain. In addition, oscillations may arise solely from characteristic boundary effects when the boundary conditions at the inflow differ from those along the sides, even if the inflow and outflow boundary conditions are the same.

Finally, we note that one way to reduce the extent of oscillations is to add artificial diffusion in the streamline direction, using the so-called streamline-diffusion method ([3], §9.7.2). The methodology introduced in this paper can be used to develop insight into this discretisation technique. This topic is treated in the companion paper [1].

Appendix A. The weight functions

We derive formulae for the boundary condition dependent functions $F_m(i)$, $i = 1, 2, 3$ in (3.6) for a bilinear Galerkin finite element approximation.

We begin with

$$F_1(i) = -\frac{\bar{t}_i}{\sigma_i}.$$

Using the notation of Section 3, the vector $\mathbf{t}$ has entries

$$t_i = \begin{cases} -m_3\, f_t(x_1) - m_4\, f_t(x_2), & \\ -m_4\, f_t(x_{i-1}) - m_3\, f_t(x_i) - m_4\, f_t(x_{i+1}), & i \in \{2, \dots, N-2\} \\ -m_4\, f_t(x_{N-2}) - m_3\, f_t(x_{N-1}) & \end{cases}$$

and so the ith entry of $\bar{\mathbf{t}} = V^T\mathbf{t}$ is

$$\bar{t}_i = \sqrt{\frac{2}{N}} t_i^*,$$

where

$$\begin{aligned}
t_i^* &= [-m_3\, f_t(x_1) - m_4\, f_t(x_2)] \sin\frac{i\pi}{N} \\
&\quad + \sum_{p=2}^{N-2} [-m_4\, f_t(x_{p-1}) - m_3\, f_t(x_p) - m_4\, f_t(x_{p+1})] \sin\frac{pi\pi}{N} \\
&\quad + [-m_4\, f_t(x_{N-2}) - m_3\, f_t(x_{N-1})] \sin\frac{(N-1)i\pi}{N} \\
&= -m_3 \sum_{p=1}^{N-1} f_t(x_p) \sin\frac{pi\pi}{N} - m_4 \sum_{p=1}^{N-2} f_t(x_{p+1}) \sin\frac{pi\pi}{N} \\
&\quad - m_4 \sum_{p=2}^{N-1} f_t(x_{p-1}) \sin\frac{pi\pi}{N} \\
&= \sum_{p=2}^{N-2} f_t(x_p) \left\{ -m_3 \sin\frac{pi\pi}{N} - m_4 \sin\frac{(p-1)i\pi}{N} - m_4 \sin\frac{(p+1)i\pi}{N} \right\} \\
&\quad + f_t(x_1) \left\{ -m_3 \sin\frac{i\pi}{N} - m_4 \sin\frac{2i\pi}{N} \right\} \\
&\quad + f_t(x_{N-1}) \left\{ -m_3 \sin\frac{(N-1)i\pi}{N} - m_4 \sin\frac{(N-2)i\pi}{N} \right\} \\
&= \sum_{p=2}^{N-2} f_t(x_p) \sin\frac{pi\pi}{N} \left\{ -m_3 - 2m_4 \cos\frac{i\pi}{N} \right\} \\
&\quad + f_t(x_1) \sin\frac{i\pi}{N} \left\{ -m_3 - 2m_4 \cos\frac{i\pi}{N} \right\} \\
&\quad + f_t(x_{N-1}) \sin\frac{(N-1)i\pi}{N} \left\{ -m_3 - 2m_4 \cos\frac{i\pi}{N} \right\} \\
\Rightarrow t_i^* &= -\sum_{p=1}^{N-1} f_t(x_p) \sin\frac{pi\pi}{N} \left\{ m_3 + 2m_4 \cos\frac{i\pi}{N} \right\}.
\end{aligned}$$

Recalling from (2.3) that

$$\sigma_i = m_3 + 2m_4 \cos\frac{i\pi}{N},$$

we have

$$F_1(i) = -\sqrt{\frac{2}{N}}\frac{t_i^*}{\sigma_i} = \sqrt{\frac{2}{N}} \sum_{p=1}^{N-1} f_t(x_p) \sin\frac{pi\pi}{N}.$$

By a similar argument,

$$F_3(i) = -\frac{\bar{b}_i}{\gamma_i} = \sqrt{\frac{2}{N}} \sum_{p=1}^{N-1} f_b(x_p) \sin\frac{pi\pi}{N}.$$

Finally, consider

$$F_2(i) = \frac{\bar{s}_i}{\gamma_i + \lambda_i + \sigma_i}.$$

Assuming that the left and right boundary functions are constant, that is,

$$f_l(y_{k-1}) = f_l(y_k) = f_l(y_{k+1}) \equiv f_l, \quad f_r(y_{k-1}) = f_r(y_k) = f_r(y_{k+1}) \equiv f_r,$$

we have

$$s_1 = -(m_2 + m_4 + m_6)f_l = \epsilon f_l, \quad s_{N-1} = -(m_2 + m_4 + m_6)f_r = \epsilon f_r.$$

Hence the vector $\bar{\mathbf{s}} = V^T\mathbf{s}$ has entries

$$\bar{s}_i = \sqrt{\frac{2}{N}}\epsilon\left(f_l \sin\frac{i\pi}{N} + f_r \sin\frac{(N-1)i\pi}{N}\right) = \sqrt{\frac{2}{N}}\epsilon\left(f_l + (-1)^{i+1}f_r\right)\sin\frac{i\pi}{N}.$$

As

$$\gamma_i + \lambda_i + \sigma_i = (m_1 + m_3 + m_5) + 2(m_2 + m_4 + m_6)C_i = 2\epsilon(1 - C_i),$$

we obtain the expression

$$F_2(i) = \sqrt{\frac{2}{N}}\frac{\left[f_l + (-1)^{i+1}f_r\right]\sin\dfrac{i\pi}{N}}{2\left(1 - \cos\dfrac{i\pi}{N}\right)}.$$

For the special case where $f_l = f_r$, we have (see [7], 19.40)

$$F_2(i) = f_l\sqrt{\frac{2}{N}}\frac{[1 + (-1)^{i+1}]\sin\dfrac{i\pi}{N}}{2\left(1 - \cos\dfrac{i\pi}{N}\right)} = f_l\sqrt{\frac{2}{N}}\sum_{p=1}^{N-1}\sin\frac{pi\pi}{N}.$$

References

1. H.C. Elman and A. Ramage, *An analysis of smoothing effects of upwinding strategies for the convection-diffusion equation*, Tech. Report UMCP-CSD:CS-TR-4160, University of Maryland, College Park MD 20742, 2000.
2. P.M. Gresho and R.L. Sani, *Incompressible flow and the finite element method*, John Wiley and Sons, Chichester, 1999.
3. C. Johnson, *Numerical solutions of partial differential equations by the finite element method*, Cambridge University Press, Cambridge, 1987. MR **89b:**65003a
4. K.W. Morton, *Numerical solution of convection-diffusion problems*, Chapman and Hall, London, 1996. MR **98b:**65004
5. H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical methods for singularly perturbed differential equations*, Springer-Verlag, Berlin, 1996. MR **99a:**65134
6. B. Semper, *Numerical crosswind smear in the streamline diffusion method*, Comput. Methods Appl. Mech. Engrg **113** (1994), 99–108. MR **94k:**76060
7. M.R. Spiegel, *Mathematical handbook of formulas and tables*, Schaum's outline series, McGraw-Hill, New York, 1990.
8. P.N. Swarztrauber, *The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle*, SIAM Review **19** (1977), 490–501. MR **55:**11639

Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742

E-mail address: elman@cs.umd.edu

Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, Scotland

E-mail address: alison@maths.strath.ac.uk