SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 254–281

AN ANALYSIS OF SMOOTHING EFFECTS OF UPWINDING STRATEGIES FOR THE CONVECTION-DIFFUSION EQUATION*

HOWARD C. ELMAN† AND ALISON RAMAGE‡

Abstract. Using a technique for constructing analytic expressions for discrete solutions to the convection-diffusion equation, we examine and characterize the effects of upwinding strategies on solution quality. In particular, for grid-aligned flow and discretization based on bilinear finite elements with streamline upwinding, we show precisely how the amount of upwinding included in the discrete operator affects solution oscillations and accuracy when different types of boundary layers are present. This analysis provides a basis for choosing a streamline upwinding parameter which also gives accurate solutions for problems with non-grid-aligned and variable speed flows. In addition, we show that the same analytic techniques provide insight into other discretizations, such as a finite difference method that incorporates streamline diffusion and the isotropic artificial diffusion method.

Key words. convection-diffusion equation, oscillations, Galerkin finite element method, streamline diffusion

AMS subject classifications. 65N22, 65N30, 65Q05, 35J25

PII. S0036142901374877

1. Introduction.

There are many discretization strategies available for the linear convection-diffusion equation

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

where the small parameter ϵ and divergence-free convective velocity field $\mathbf{w} = (w_1(x,y), w_2(x,y))$ are given. In this paper, we analyze some well-known methods which involve the addition of upwinding to stabilize the discretization for problems involving boundary layers. In particular, we focus on characterizing exactly how this upwinding affects the resulting discrete solutions.

A standard discretization technique is the Galerkin finite element method (see, for example, [5], [9], [10], [11], [13]). This is based on seeking a solution u of the weak form of (1.1),

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

where the test functions v are in the Sobolev space $V = \mathcal{H}_0^1(\Omega)$. Restricting this to a finite-dimensional subspace V_h of V gives

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

where f_h is the $L^2(\Omega)$ orthogonal projection of f into V_h and h is a discretization parameter. Choosing the test functions equal to a set of basis functions for V_h (usually

[*] Received by the editors June 18, 2001; accepted for publication (in revised form) January 10, 2002; published electronically May 1, 2002.
http://www.siam.org/journals/sinum/40-1/37487.html

[†] Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (elman@cs.umd.edu). The work of this author was supported by National Science Foundation grant DMS9972490.

[‡] Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, Scotland (ar@maths.strath.ac.uk). The work of this author was supported by the Leverhulme Trust.

continuous piecewise polynomials with local support) leads to a sparse linear system whose solution can be used to recover the discrete solution u_h.

One quantity which has an important effect on the quality of the resulting discrete solution is the mesh Péclet number

$$P_e^{el} = \frac{h^{el}|\mathbf{w}|}{2\epsilon},$$

where h^{el} is a measure of element size and $|\mathbf{w}|$ represents the strength of the convective field within an element. In particular, if the mesh Péclet number is greater than one, then the discrete solution obtained from the Galerkin method may exhibit nonphysical oscillations. For the one-dimensional analogue of (2.1), this is well understood (see, for example, [10, p. 14]); for an analysis of the Galerkin discretization of the two-dimensional case, see [2]. An approach for minimizing the deleterious effects of these oscillations, especially in areas of the domain away from boundary layers, is to stabilize the discrete problem by using an upwind discretization. A particularly effective implementation of this idea is via the streamline diffusion method (see, e.g., [8], [9, sect. 9.7]). For linear or bilinear elements, the weak form (1.2) is replaced by

$$\epsilon(\nabla u_h, \nabla v) + (\mathbf{w}.\nabla u_h, v) + \sum \alpha^{el}(\mathbf{w}\cdot\nabla u_h, \mathbf{w}\cdot\nabla v)_{el} = (f_h, v) + \sum \alpha^{el}(f_h, \mathbf{w}\cdot\nabla v)_{el} \quad \forall v \in V_h, \tag{1.3}$$

where the sums are taken over all elements in the discretization. The stabilization parameters α^{el} are given by

$$\alpha^{el} = \frac{\delta^{el} h^{el}}{|\mathbf{w}|}, \tag{1.4}$$

where $\delta^{el} \geq 0$ are parameters to be chosen. Note that setting $\delta^{el} = 0$ on each element reduces (1.3) to the standard Galerkin case (1.2): this is the usual practice when $P_e^{el} < 1$. Formulation (1.3) has additional coercivity in the local flow direction, resulting in improved stability. More on the motivation behind this method can be found in [6, p. 289]. However, the best way of choosing δ^{el} for a general convection-diffusion problem is not known: for a discussion of this difficulty, see, for example, [13, Remark 3.34, p. 234].

In [2], we developed an analytic technique for characterizing the nature of oscillations in discrete solutions arising from the Galerkin discretization (1.2). More specifically, for the case of grid-aligned flow, we presented an analytic representation of the discrete solution, enabling isolation of any oscillatory behavior in the direction of the flow. Using this framework, we studied the dependence of solution behavior on the mesh Péclet number in some detail.

In this paper, we apply the tools developed in [2] to various upwinding strategies for discretizing (1.1). For the most part, we focus on the streamline diffusion method (1.3), examining the effect of stabilization on the quality of the resulting discrete solutions. In section 2, we summarize the Fourier analysis presented in [2] and derive an explicit formula for the discrete streamline diffusion solution for a model problem with constant grid-aligned flow. Section 3 contains the details of this process in the case of bilinear finite elements. The resulting formulae allow us to investigate various issues which influence the choice of stabilization parameters. In section 4, we characterize the effect of stabilization on oscillations in the discrete solution in the flow direction for three test problems whose solutions exhibit different types of

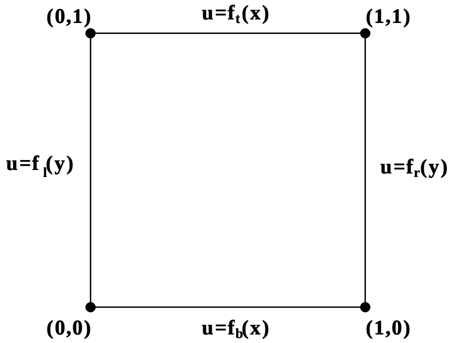


FIG. 1. *Boundary conditions.*

boundary layers. The implications of this analysis for solution accuracy are examined in section 5. In section 6, we discuss the relevance of our results for problems with non-grid-aligned and variable flow and present our recommended choice for the streamline diffusion parameters. Finally, in section 7, we illustrate how the same approach can be used to understand other discretization methods. We analyze an analogous streamline diffusion (upwind) discretization for a finite difference stencil and explain the comparative lack of effectiveness of isotropic artificial diffusion.

2. Summary of Fourier analysis. In this section, we summarize the Fourier techniques used in [2] to construct an analytic expression for the entries in the discrete solution vector $\mathbf{u}$.

Setting $\mathbf{w} = (0, 1)$ and $f = 0$ in (1.1), we obtain the "vertical wind" model problem

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

with Dirichlet boundary conditions as shown in Figure 1. Using a natural ordering of the unknowns on a uniform grid of square bilinear elements with $N = 1/h$ elements in each dimension, both (1.2) and (1.3) give rise to a linear system

$$A\mathbf{u} = \mathbf{f}, \tag{2.2}$$

where the coefficient matrix A is of order $(N-1)^2$. Denoting the coefficients of the computational molecule by

$$\begin{array}{ccccc} m_4 & & m_3 & & m_4 \\ & \nwarrow & \uparrow & \nearrow & \\ m_2 & \leftarrow & m_1 & \rightarrow & m_2 \\ & \swarrow & \downarrow & \searrow & \\ m_6 & & m_5 & & m_6 \end{array}, \tag{2.3}$$

the matrix A can be written as

$$A = \begin{bmatrix} M_1 & M_2 & & & 0 \\ M_3 & M_1 & M_2 & & \\ & \ddots & \ddots & \ddots & \\ & & M_3 & M_1 & M_2 \\ 0 & & & M_3 & M_1 \end{bmatrix}, \tag{2.4}$$

where $M_1 = \text{tridiag}(m_2, m_1, m_2)$, $M_2 = \text{tridiag}(m_4, m_3, m_4)$, and $M_3 = \text{tridiag}(m_6, m_5, m_6)$ are all tridiagonal matrices of order $N-1$. Given that the eigenvalues and eigenvectors of the blocks of A satisfy

$$\begin{aligned} M_1\mathbf{v}_j &= \lambda_j\mathbf{v}_j, \qquad & \lambda_j &= m_1 + 2m_2 \cos\tfrac{j\pi}{N}, \\ M_2\mathbf{v}_j &= \sigma_j\mathbf{v}_j, \qquad & \sigma_j &= m_3 + 2m_4 \cos\tfrac{j\pi}{N}, \\ M_3\mathbf{v}_j &= \gamma_j\mathbf{v}_j, \qquad & \gamma_j &= m_5 + 2m_6 \cos\tfrac{j\pi}{N} \end{aligned} \tag{2.5}$$

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

$$\mathbf{v}_j = \sqrt{\frac{2}{N}}\left[\sin\frac{j\pi}{N}, \quad \sin\frac{2j\pi}{N}, \quad \dots, \sin\frac{(N-1)j\pi}{N}\right]^T, \tag{2.6}$$

we may obtain the decomposition

$$A = (\mathcal{V}P)T(\mathcal{V}P)^T, \tag{2.7}$$

where $\mathcal{V} = \text{diag}(V, V, \dots, V)$ is a block diagonal matrix with each block V having the $N-1$ eigenvectors (2.6) as its columns, and P is a permutation matrix of order $(N-1)^2$. The matrix T is also block diagonal, with diagonal blocks $T_i = \text{tridiag}(\gamma_i, \lambda_i, \sigma_i)$, $i = 1, \dots, N-1$. Using this decomposition and observing that P and V are both orthogonal, (2.2) implies

$$\mathbf{u} = \mathcal{V}P\mathbf{y}, \tag{2.8}$$

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

$$T\mathbf{y} = P^T\mathcal{V}^T\mathbf{f} \equiv \hat{\mathbf{f}}. \tag{2.9}$$

As T is block diagonal, this system can be partitioned into $N-1$ independent systems of the form

$$T_i\mathbf{y}_i = \hat{\mathbf{f}}_i, \tag{2.10}$$

where T_i is defined above and $\mathbf{y}$ and $\hat{\mathbf{f}}$ are partitioned in the obvious way. 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 $\mathbf{y}_i$, $k = 1, \dots, N-1$, in (2.10). Finally, to obtain an explicit formula for the entries of $\mathbf{u}$, we permute and transform these entries via (2.8) to get

$$u_{jk} = \sqrt{\frac{2}{N}}\sum_{i=1}^{N-1} \sin\frac{ij\pi}{N} y_{ik} \tag{2.11}$$

for $j, k = 1, \dots, N-1$.

To obtain an expression for the entries y_{ik} in (2.11), we must consider the vectors $\hat{\mathbf{f}}_i$. As $f = 0$ in (2.1), the only nonzero entries in the original right-hand side vector $\mathbf{f}$ in (2.2) involve sums of certain matrix coefficients times boundary values, which are transformed and permuted to obtain $\hat{\mathbf{f}}$ in (2.9). The details of this process can be found in [2]. Here we simply state that each right-hand side vector $\hat{\mathbf{f}}_i$, $i = 1, \dots, N-1$, in (2.10) can be written as

$$\hat{\mathbf{f}}_i = \begin{bmatrix} \bar{b}_i + \bar{s}_i \\ \bar{s}_i \\ \vdots \\ \bar{s}_i \\ \bar{t}_i + \bar{s}_i \end{bmatrix}_{N-1},$$

where $\bar{b}_i$ involves data from the bottom boundary values, $\bar{t}_i$ involves data from the top boundary values, and $\bar{s}_i$ combines information from the left and right boundary values. We will make the same assumption as in [2] that the functions $f_l(y)$ and $f_r(y)$ on the left and right boundaries are constant. This simplifies the presentation of the analysis.

The solution of each system (2.10) is now the solution of a three-term recurrence relation with constant coefficients whose auxiliary equation has roots

$$\mu_1(i) = \frac{-\lambda_i + \sqrt{\lambda_i^2 - 4\sigma_i\gamma_i}}{2\sigma_i}, \qquad \mu_2(i) = \frac{-\lambda_i - \sqrt{\lambda_i^2 - 4\sigma_i\gamma_i}}{2\sigma_i}. \tag{2.12}$$

The solution of this recurrence relation can be written as

$$y_{ik} = F_3(i) + [F_1(i) - F_3(i)]\, G_1(i,k) + [F_2(i) - F_3(i)]\, G_2(i,k), \tag{2.13}$$

where

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

and the functions

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

involve the coefficient matrix entries and boundary condition information (see [2] for details).

We emphasize that the functions $F_m(i)$, $m = 1, 2, 3$, in (2.13) are independent of the vertical grid index k: for fixed i, the behavior of $\mathbf{y}$ in the streamline (vertical) direction depends only on the functions $G_1(i,k)$ and $G_2(i,k)$. In addition, 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, (2.13) 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}. In the next section, we analyze the behavior of these solutions in some detail for the streamline diffusion finite element discretization (1.3) with bilinear elements.

3. Streamline diffusion discretization.

In [2], an explicit expression for (2.13) for the Galerkin finite element method with bilinear elements was derived and analyzed. Here we present the equivalent analysis for the streamline diffusion finite element discretization (1.3) with a view to precisely characterizing the effect of the extra diffusion on the oscillations that occur with the Galerkin method when $P_e^{el} > 1$. We again use bilinear elements. Note that for a uniform grid and constant grid-aligned flow, $\delta = \delta^{el}$ is constant over all elements.

3.1. The recurrence relation solution.

The coefficients in stencil (2.3) for a streamline diffusion discretization (1.3) using bilinear finite elements are given by

$$m_1 = \tfrac{4}{3}\,(\delta h + 2\epsilon), \qquad m_2 = \tfrac{1}{3}\,(\delta h - \epsilon), \qquad m_3 = -\tfrac{1}{3}\,[(2\delta - 1)\,h + \epsilon],$$

$$m_4 = -\tfrac{1}{12}\,[(2\delta - 1)\,h + 4\epsilon], \quad m_5 = -\tfrac{1}{3}\,[(2\delta + 1)\,h + \epsilon], \quad m_6 = -\tfrac{1}{12}\,[(2\delta + 1)\,h + 4\epsilon].$$

For convenience, we introduce the notation

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

and write the eigenvalues (2.5) as

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

$i = 1, \dots, N-1$. Substituting these into (2.12) gives the expressions

$$\mu_{1,2} = \frac{-2\delta - \left[\dfrac{4-C_i}{2+C_i}\right]\dfrac{1}{P_e} \pm \sqrt{1 + \dfrac{12\delta(1-C_i)}{(2+C_i)}\dfrac{1}{P_e} + \dfrac{3(5+C_i)(1-C_i)}{(2+C_i)^2}\dfrac{1}{P_e^2}}}{-2\delta + 1 - \left[\dfrac{1+2C_i}{2+C_i}\right]\dfrac{1}{P_e}} \tag{3.1}$$

for the auxiliary equation roots in (2.13).

3.2. Oscillations in the recurrence relation solution. We know from [2, Thm 5.1] that if $P_e > 1$, then the recurrence relation solution $\mathbf{y}$ and the related discrete solution $\mathbf{u}$ to the pure Galerkin problem (1.2) usually exhibit oscillations. In this section we address the question of how the streamline diffusion parameter δ can be chosen to eliminate oscillations in the recurrence relation solution $\mathbf{y}$. The issue of how this affects the resulting $\mathbf{u}$ will be discussed in section 3.3.

THEOREM 3.1. *If $P_e > 1$, then for any value of $i \in S_N \equiv \{1, \dots, N-1\}$ there exists a parameter*

$$\delta_i^c = \frac{1}{2}\left(1 - \left[\frac{1+2C_i}{2+C_i}\right]\frac{1}{P_e}\right) \tag{3.2}$$

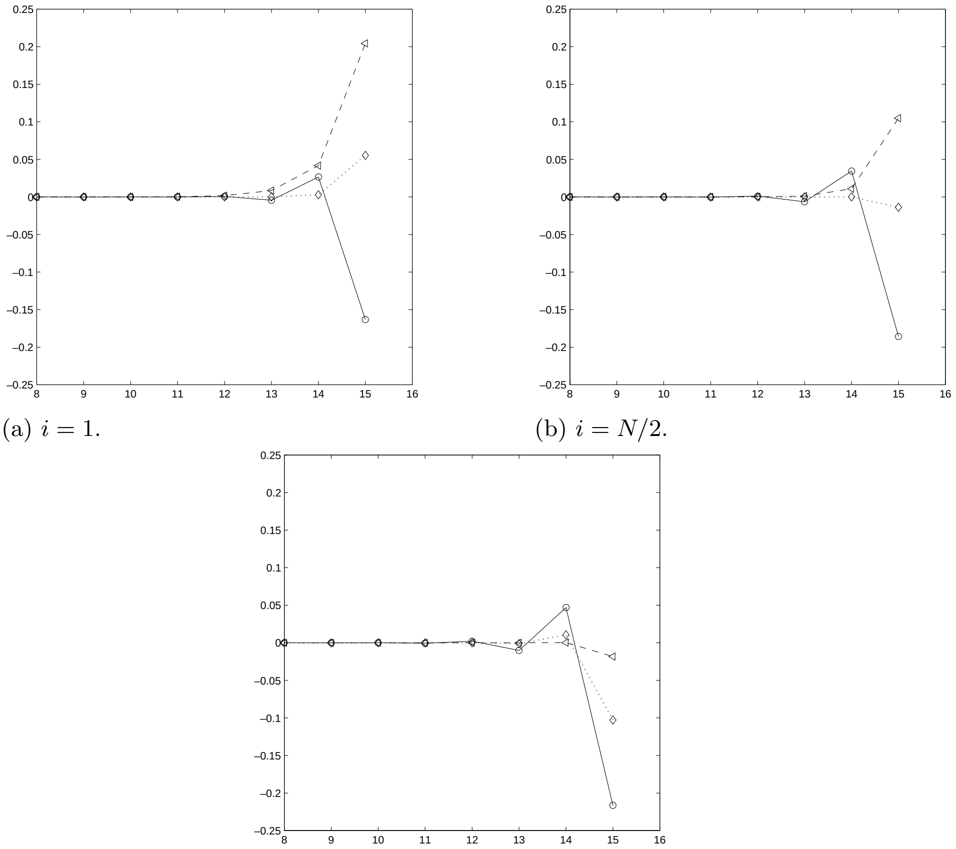
such that $\delta > \delta_i^c$ implies that $G_1(i,k)$ and $G_2(i,k)$ in (2.13) *are nonoscillatory functions of k.*

Proof. We have

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

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 $i \in S_N$. From (3.1), the numerator of μ_2 is always negative so, for δ_i^c given by (3.2), we have the conditions

$$\begin{cases} \delta > \delta_i^c \quad \Rightarrow \quad \mu_2 > 0,\ G_1(i,k) \text{ is nonoscillatory},\\ \delta < \delta_i^c \quad \Rightarrow \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory}.\end{cases}$$



(a) $i = 1$.

(b) $i = N/2$.

(c) $i = N - 1$.

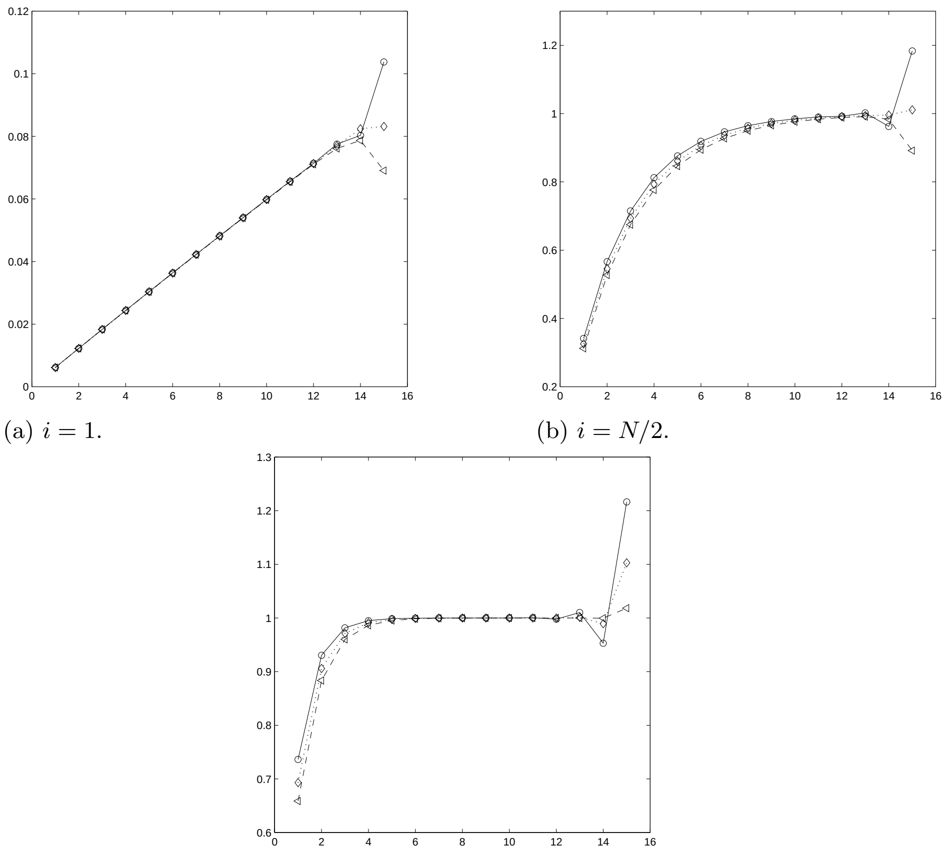
FIG. 2. *Plots of $G_1(i,k)$ against k for fixed i with $\delta = 0.2$ (solid, o), $\delta = 0.4$ (dotted, $\diamondsuit$), and $\delta = 0.6$ (dashed, $\triangle$).*

In addition, it can be shown that $0 < \mu_1 < 1$ so that if $G_1(i,k)$ is nonoscillatory, then $G_2(i,k) = (1 - \mu_1^k) - (1 - \mu_1^N)G_1(i,k)$ must also be nonoscillatory. □

Sample plots of $G_1(i,k)$ for various values of $i \in S_N$ when $N = 16$ and $P_e = 3.125$ are given in Figure 2. Only the right half of the range of k has been plotted in each case to magnify the area of interest. Each subplot shows the behavior for three distinct values of δ, namely $\delta = 0.2$ (solid line, o), $\delta = 0.4$ (dotted line, $\diamondsuit$), and $\delta = 0.6$ (dashed line, $\triangle$). Given the relevant critical values $\delta_1^c \simeq 0.34$, $\delta_{N/2}^c \simeq 0.42$, and $\delta_{N-1}^c \simeq 0.65$ for this problem, the dependence of oscillations on the value of δ is clear. For $\delta = 0.2$ (that is, $\delta < \delta_i^c$ for all $i \in S_N$), all functions $G_1(i,k)$ are oscillatory; for $\delta = 0.4$, $G_1(1,k)$ is nonoscillatory (as $\delta > \delta_1^c$) and $G_1(N/2,k)$ is only very mildly oscillatory; for $\delta = 0.6$, only $G_1(N-1,k)$ is oscillatory (as $\delta > \delta_i^c$ for $i = 1, N/2$). Analogous behavior is seen in Figure 3 for $G_2(i,k)$ with the same parameter values, although the oscillations here occur about the function $1 - \mu_1^k$ rather than zero.

We now define

$$\delta_* = \frac{1}{2}\left(1 - \frac{1}{P_e}\right), \qquad \delta^* = \frac{1}{2}\left(1 + \frac{1}{P_e}\right) \tag{3.3}$$



(a) $i = 1$.

(b) $i = N/2$.

(c) $i = N - 1$.

FIG. 3. *Plots of $G_2(i, k)$ against k for fixed i with $\delta = 0.2$ (solid, o), $\delta = 0.4$ (dotted, $\diamondsuit$), and $\delta = 0.6$ (dashed, $\triangle$).*

(as in [3]) so that

$$\delta_* < \delta_i^c < \delta^* \tag{3.4}$$

for all values of $i \in S_N$. If $\delta \geq \delta^*$, then $\delta > \delta_i^c$ for each $i \in S_N$ and all of the functions $G_1(i, k)$ and $G_2(i, k)$ will be nonoscillatory in terms of k. We therefore have the following corollary to Theorem 3.1.

COROLLARY 3.2. *For any value of δ such that $\delta \geq \delta^*$, the functions $G_1(i, k)$ and $G_2(i, k)$ in (2.13) are nonoscillatory functions of k for every $i \in S_N$. Hence the recurrence relation solution* $\mathbf{y}$ *is a sum of smooth functions and will not exhibit oscillations in the streamline direction.*

The case $\delta = \delta_i^c$ requires special attention. With this value, $\sigma_i = 0$ in (2.5) and the resulting matrix T_i in (2.10) is bidiagonal. This leads to a two-term recurrence relation with auxiliary equation root

$$\rho = \frac{1}{1 + \dfrac{3(1 - C_i)}{2 + C_i} \dfrac{1}{P_e}}$$

and solution

$$y_{ik} = F_3(i)\rho^k + F_2(i)(1-\rho^k). \tag{3.5}$$

As $0 < \rho < 1$ for any $i \in S_N$, y_{ik} is nonoscillatory in the streamline direction. In addition, $\rho \to 1$ as $P_e \to \infty$, giving the solution $y_{ik} = F_3(i)$. Looking ahead to section 3.3, applying transformation (2.11) gives $u_{jk} = f_b(x_j)$ (see (3.8)). This is the solution to the reduced problem (obtained by setting $\epsilon = 0$ in (2.1)) where the bottom boundary values are simply transported in the direction of the flow without any diffusion present. That is, with the choice $\delta = \delta_i^c$ for each i, the discrete solution is exact at every interior node in the limit as $P_e \to \infty$.

3.3. Oscillations in the discrete solution. In this section we consider the impact of transformation (2.11) on the recurrence relation solution $\mathbf{y}$, with a view to choosing δ to obtain an oscillation-free discrete solution $\mathbf{u}$. We begin by considering the functions $F_m(i)$, $m = 1, 2, 3$, in (2.13). Following the analysis of [2, sect. 4.4 and appendix] we can derive the following expressions

$$\begin{aligned} F_1(i) &= \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_t(x_s) \sin\frac{si\pi}{N}, \\ F_2(i) &= f_l \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin\frac{si\pi}{N}, \\ F_3(i) &= \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_b(x_s) \sin\frac{si\pi}{N} \end{aligned} \tag{3.6}$$

for the streamline diffusion weight functions in the special case where the constant left and right boundary values f_l and f_r are equal. From (2.13), we therefore have

$$\begin{aligned} y_{ik} = {} & \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_b(x_s) \sin\frac{si\pi}{N} + \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} [f_t(x_s) - f_b(x_s)] \sin\frac{si\pi}{N} G_1(i,k) \\ & + \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} [f_l - f_b(x_s)] \sin\frac{si\pi}{N} G_2(i,k) \end{aligned} \tag{3.7}$$

[2, Thm 4.2]. Note that the expressions in (3.6) hold for any stencil of the form (2.3) whose entries sum to zero. In particular, this implies that the functions in (3.6) are the same for discretizations (1.2) and (1.3).

We now apply transformation (2.11) to (3.7) to obtain an expression for the entries of the discrete solution vector $\mathbf{u}$. As in [2], for the first term we have

$$\sqrt{\frac{2}{N}} \sum_{i=1}^{N-1} \sin\frac{ij\pi}{N} \left\{ \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_b(x_s) \sin\frac{si\pi}{N} \right\} = f_b(x_j), \tag{3.8}$$

where $f_b(x)$ is the bottom boundary function in Figure 1. Applying (2.11) to the full expression (3.7) therefore gives

$$u_{jk} = f_b(x_j) + \frac{2}{N} \sum_{i=1}^{N-1} \left[a_{ij} G_1(i,k) + b_{ij} G_2(i,k) \right], \tag{3.9}$$

where

$$
\begin{aligned}
a_{ij} &= \sin\frac{ij\pi}{N}\sum_{s=1}^{N-1}[f_t(x_s) - f_b(x_s)]\sin\frac{si\pi}{N}, \\
b_{ij} &= \sin\frac{ij\pi}{N}\sum_{s=1}^{N-1}[f_l - f_b(x_s)]\sin\frac{si\pi}{N}.
\end{aligned}
\tag{3.10}
$$

That is, along a streamline (j fixed), $\mathbf{u}$ consists of the bottom boundary value on that line plus a linear combination of the functions $G_1(i,k)$ and $G_2(i,k)$ for $i \in S_N$. Note that $a_{i(N-j)} = a_{ij}$ and $b_{i(N-j)} = b_{ij}$, so that if $f_b(x)$ is symmetric about the center vertical line of the grid, then so is $\mathbf{u}$.

We can use the representation (3.9) to obtain insight into the effect of δ on the quality of the solution in the streamline direction. Recall from section 3.2 that if $\delta \geq \delta_i^c$ in (3.2), then the functions $G_1(i,k)$ and $G_2(i,k)$ are nonoscillatory in the streamline direction for that particular $i \in S_N$. It follows from Corollary 3.2 that if $\delta \geq \delta^*$ in (3.3), then (3.9) is a sum of smooth functions. We have therefore established a sufficient condition for the discrete solution to be nonoscillatory.

THEOREM 3.3. *For a streamline diffusion discretization of* (2.1) *with bilinear finite elements, the discrete solution* $\mathbf{u}$ *does not exhibit oscillations in the streamline direction when* $\delta \geq \delta^*$.

4. Analysis of boundary layer effects. In practice, it turns out that the restriction on δ given by Theorem 3.3 is too harsh, and better solutions can be obtained using values of δ smaller than δ^* due to the "smoothing" nature of transformation (2.11). The precise effect of this transformation in the context of the behavior of the Galerkin finite element solution for different mesh Péclet numbers was studied in [2]. Here we present a discussion of the effects of varying δ in the streamline diffusion method. We illustrate the ideas with three examples containing different types of boundary layers. The first two examples contain an exponential layer at the outflow and parabolic layers along the characteristic (vertical) boundaries, respectively. The third example has a Neumann boundary condition at the outflow, and we show that the analysis generalizes to this case.

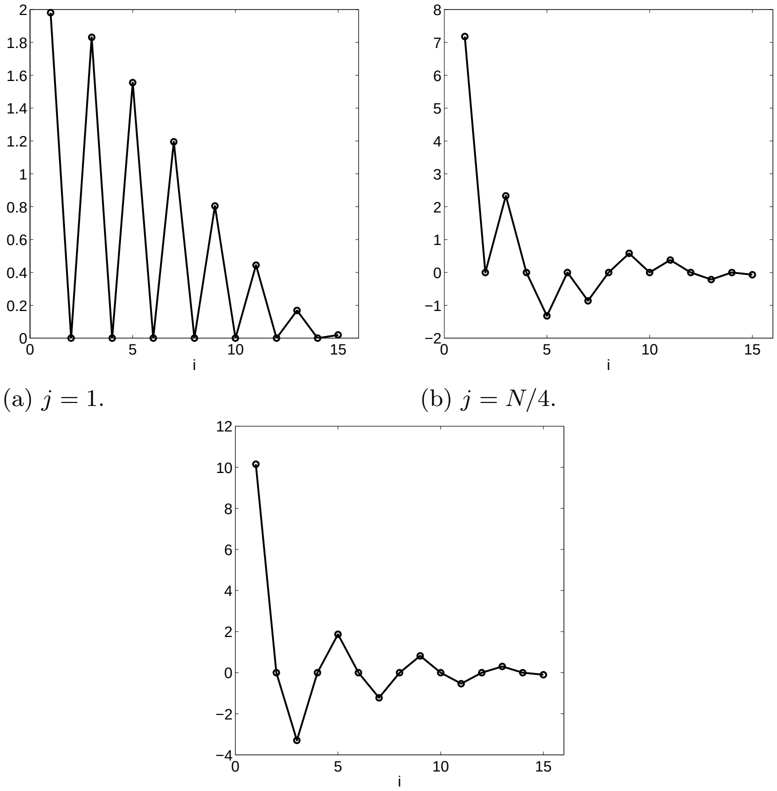
Throughout this section we will use notation based on considering u_{jk} in (3.9) as a sum of smooth and oscillatory parts. That is, letting i^* be the lowest value of $i \in S_N$ such that $\delta < \delta_i^c$, we write

$$
\begin{aligned}
u_{jk} &= f_b(x_j) + \frac{2}{N}\left(\sum_{i=1}^{i^*-1}[a_{ij}G_1(i,k) + b_{ij}G_2(i,k)] + \sum_{i=i^*}^{N-1}[a_{ij}G_1(i,k) + b_{ij}G_2(i,k)]\right) \\
&= f_b(x_j) + S_{\text{smooth}} + S_{\text{osc}}.
\end{aligned}
\tag{4.1}
$$

Note that the preceding analysis implies $S_{\text{smooth}} = 0$ when $\delta \leq \delta_*$ and $S_{\text{osc}} = 0$ when $\delta \geq \delta^*$. As δ increases from δ_*, i^* will increase so that S_{smooth} contains more and more of the terms, with the overall smoothness of $\mathbf{u}$ dependent on the relative size of the two sums S_{smooth} and S_{osc}.

Problem I. In this example we apply the Dirichlet boundary conditions

$$
f_t(x) = 1, \qquad f_b(x) = f_l(y) = f_r(y) = 0,
$$



(a) $j = 1$.

(b) $j = N/4$.

(c) $j = N/2$.

FIG. 4. *Plots of coefficients a_{ij} against i for $N = 16$.*

as per Figure 1, so that the solution has an exponential boundary layer of width ϵ along the top boundary. For this problem, (3.7) implies

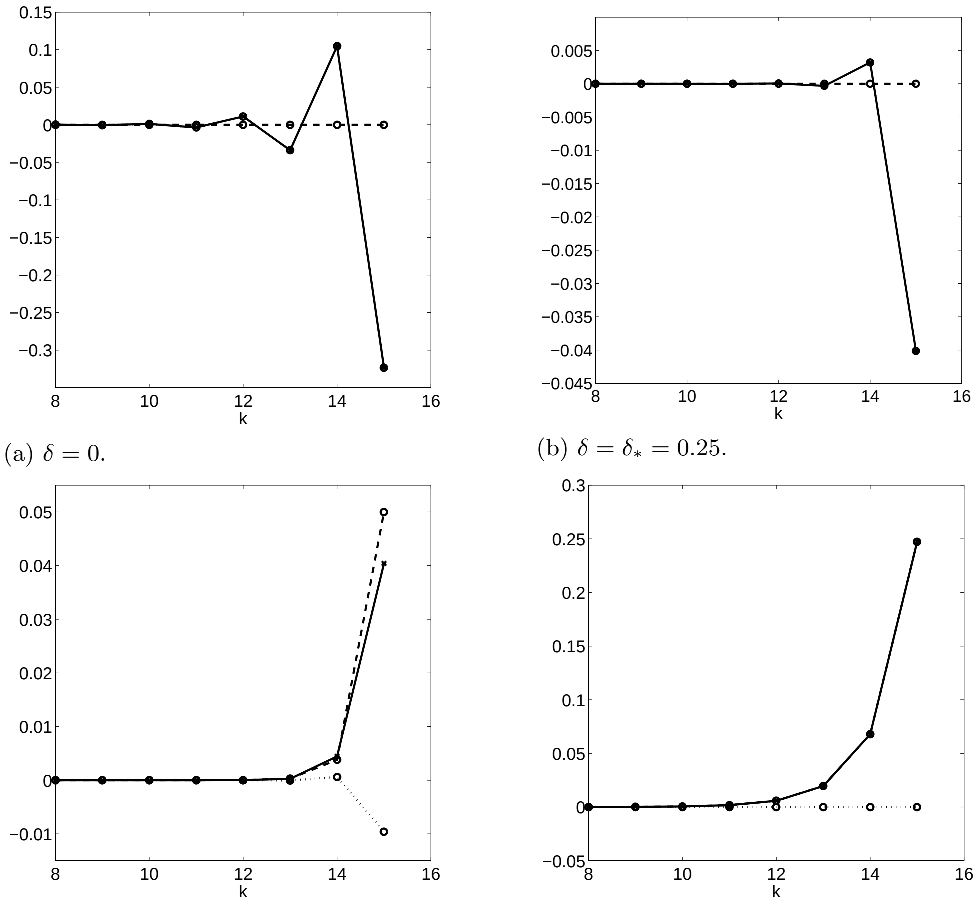
$$y_{ik} = \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin\frac{si\pi}{N} G_1(i,k) \tag{4.2}$$

so the coefficients in (3.10) simplify to

$$a_{ij} = \sin\frac{ij\pi}{N} \sum_{s=1}^{N-1} \sin\frac{si\pi}{N}, \qquad b_{ij} = 0, \tag{4.3}$$

with the magnitude of each a_{ij} decreasing rapidly as i goes from 1 to $N-1$ as shown in Figure 4 (taken from [2]). This means that the contributions to u_{jk} from the functions $G_1(i,k)$ are much larger for small indices i, so that the smoothness of $G_1(i,k)$ for small i plays a much more important role. In particular, it is not necessary for $G_1(i,k)$ to be nonoscillatory for all $i \in S_N$ in order for $|S_{\text{smooth}}|$ to dominate $|S_{\text{osc}}|$ and the resulting function $\mathbf{u}$ to be smooth.

We illustrate these ideas in Figures 5 and 6 for this example problem with $N = 16$ and $P_e = 2$. The first figure shows u_{1k} (or, equivalently, $u_{(N-1)k}$) plotted against k. This is the vertical cross-section of the solution obtained by fixing $j = 1$, which is



(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.

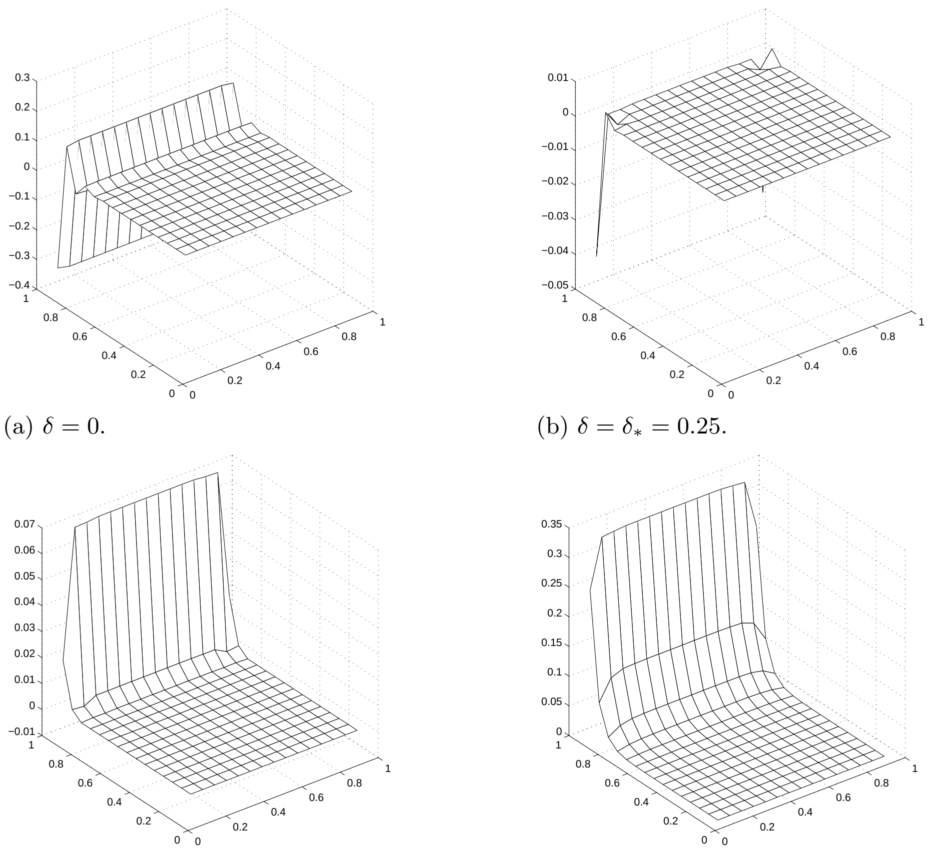
(c) $\delta = \delta_{\rm s} = 0.354$.

(d) $\delta = \delta^* = 0.75$.

FIG. 5. *Comparison of* $S_{\rm smooth}$ *(dashed line, o) and* $S_{\rm osc}$ *(dotted line, o) with* u_{1k} *(solid line, x) for Problem* I.

the most oscillatory of the vertical cross-sections for this problem. Each plot shows a comparison of $S_{\rm smooth}$ (dotted line, o) and $S_{\rm osc}$ (dashed line, o) with u_{1k} (solid line, x) for a different value of δ, where again only the right half of the range of k has been plotted to magnify the area of interest. For this example, $\delta_* = 0.25$ and $\delta^* = 0.75$. Plot (a) shows the Galerkin case ($\delta = 0$) where all of the functions $G_1(i,k)$ are oscillatory and $S_{\rm smooth}$ is zero. This is still true in plot (b), where $\delta = \delta_*$, but the magnitude and extent of the oscillations has been reduced considerably. The result of choosing $\delta = \delta^*$ according to Theorem 3.3 to guarantee an oscillation-free discrete solution by ensuring a nonoscillatory $\mathbf{y}$ is shown in plot (d). Here too much extra diffusion has been added. Plot (c) shows u_{1k} for $\delta = \delta_{\rm s} = 0.354$, which lies in the interval $(\delta_7^{\rm c}, \delta_8^{\rm c})$, that is, $i^* = 8$. This is the lowest value of i^* such that $S_{\rm smooth}$ dominates (3.9) for this problem and u_{1k} is nonoscillatory.

The corresponding full two-dimensional solutions $\mathbf{u}$ are shown in Figure 6, where the boundary values have been omitted so that the fine detail of each solution is visible. The overall behavior corresponds to that seen from the cross-sections: the severe oscillations present when $\delta = 0$ are almost eliminated by choosing $\delta = \delta_*$, and



(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.

(c) $\delta = \delta_{\rm s} = 0.354$.

(d) $\delta = \delta^* = 0.75$.

FIG. 6. *Discrete solution at interior node points for Problem* I *with* $N = 16$, $P_e = 2$.

setting $\delta = \delta^*$ gives a smooth but overly diffuse solution. For $\delta = \delta_{\rm s}$, the oscillations along the lines u_{1k} and $u_{(15)k}$ have just been eliminated to give a completely smooth solution in the flow direction.

Problem II. Next we consider the Dirichlet boundary conditions

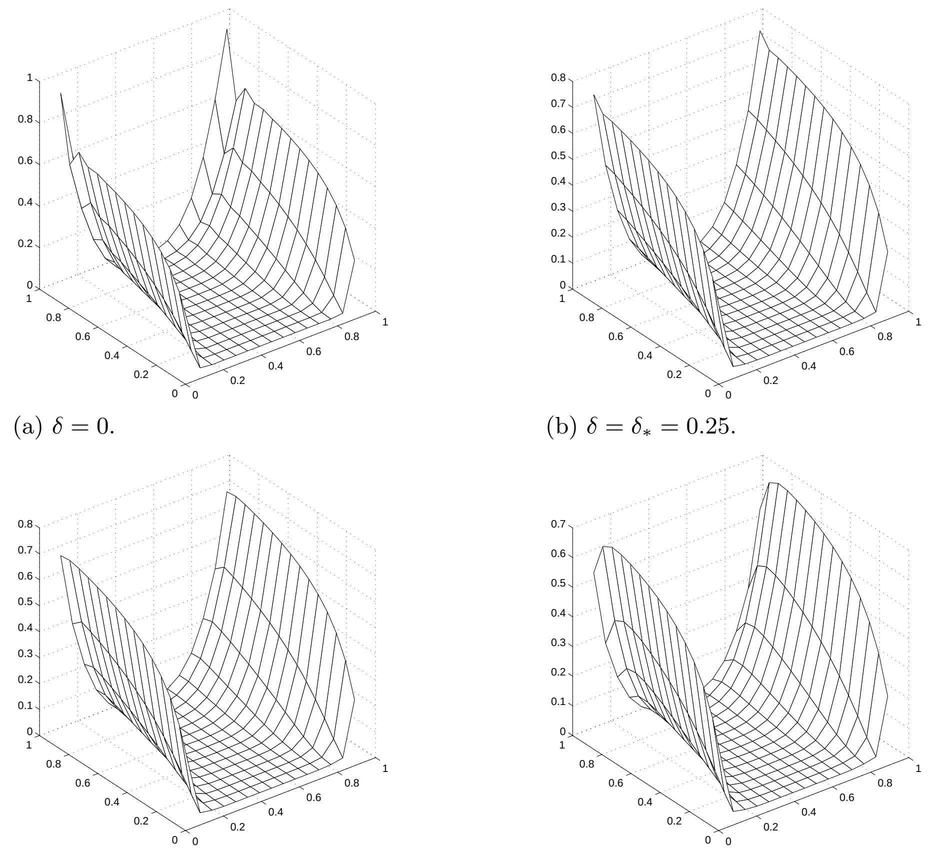
$$f_b(x) = f_t(x) = 0, \qquad f_l(y) = f_r(y) = 1,$$

which result in a solution which has parabolic layers on both vertical sides of the domain. The recurrence relation solution is

$$y_{ik} = \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N} G_2(i,k), \tag{4.4}$$

which is the same as for Problem I, except with G_2 in place of G_1 (see (4.2)). In addition, the coefficients in the full solution (3.10) are identical to those in Problem I as given by (4.3). The analysis for this problem is therefore very similar. In particular, as observed in section 3.2, G_2 is oscillatory if and only if G_1 is oscillatory, so exactly the same argument applies as to the effect of δ on solution quality.

Sample solutions for $N = 16$ with $P_e = 2$ are shown in Figure 7. These plots



(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.

(c) $\delta = \delta_{\rm s} = 0.354$.

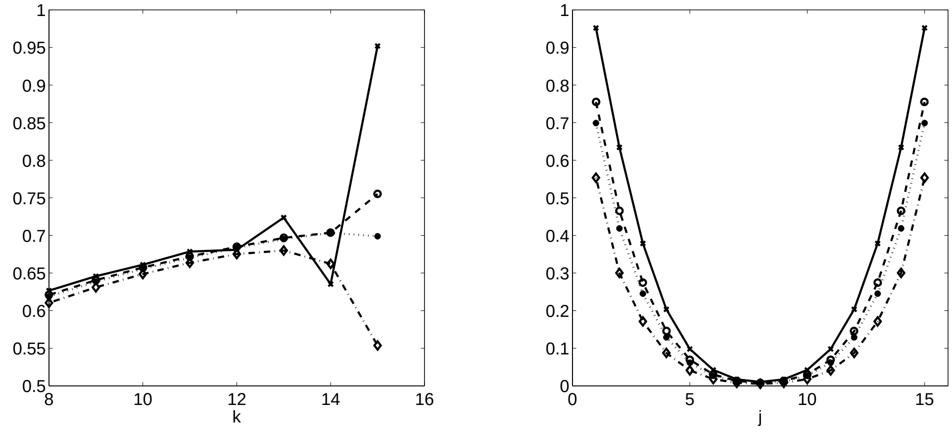
(d) $\delta = \delta^* = 0.75$.

FIG. 7. *Discrete solution at interior node points for Problem* II *with* $N = 16$, $P_e = 2$.

show the effect of increasing δ on the solution in the streamline direction: again, the solutions with $\delta = 0$ and $\delta = \delta_*$ exhibit oscillations while the solution with $\delta = \delta^*$ is overly diffuse. The value δ_s is the first for which the smooth part dominates to give a smooth solution. Figure 8 shows cross-sections of these plots for fixed values $j = 1$ on the left and $k = 15$ on the right.

It is known that parabolic layers such as those exhibited by the solution of this problem are wider than the exponential layers of the previous example (the widths are proportional to $\sqrt{\epsilon}$ and ϵ, respectively [13]). Oscillations transverse to the flow caused by inadequate resolution of parabolic layers will occur, but only for mesh Péclet numbers much larger than in the examples shown. However, the results given here demonstrate that *streamwise* effects also cause difficulties for problems with parabolic layers. The analysis shows that these are manifested in Problem II by the presence of G_2 in the solution and that streamline upwinding ameliorates these difficulties by making $G_2(i, \cdot)$ smoother for enough indices i. The right-hand plot in Figure 8 also shows that excessive diffusivity in the streamline direction gives the appearance of smearing of the characteristic layers.

Problem III. For this example, we replace the Dirichlet boundary condition $u = f_t(x)$ on the top boundary in Problem I by the Neumann boundary condition $\frac{\partial u}{\partial n} = 1$.



(a) $j = 1$. (b) $k = N - 1$.

FIG. 8. *Cross-sections of solutions to Problem* II *for* $N = 16$; $P_e = 2$ *for* $\delta = 0$ *(solid line,* $\times$*),* $\delta = \delta_*$ *(dashed line,* $\circ$*),* $\delta = \delta_s$ *(dotted line,* $*$*), and* $\delta = \delta^*$ *(dot-dash line,* $\diamondsuit$*).*

The other Dirichlet boundary conditions remain the same. The analysis of section 2 needs to be modified slightly to handle this case. There are now $N(N-1)$ unknowns, and the coefficient matrix A in (2.4) is replaced by

$$A^\star = \begin{bmatrix} M_1 & M_2 & & & 0 \\ M_3 & M_1 & M_2 & & \\ & \ddots & \ddots & \ddots & \\ & & M_3 & M_1 & M_2 \\ 0 & & & M_3 & M_1^\star \end{bmatrix},$$

where there are N rows of $(N-1)\times(N-1)$ blocks. For bilinear finite elements on a square mesh, $M_1^\star = \mathrm{tridiag}(m_2^\star, m_1^\star, m_2^\star)$ with entries

$$m_1^\star = \frac{1}{3}[(2\delta+1)h + 4\epsilon], \qquad m_2^\star = -\frac{1}{12}[(2\delta-1)h + 2\epsilon].$$

As the vectors $\mathbf{v}_j$ in (2.6) are eigenvectors of $M_1^\star$, we may construct a matrix $\mathcal{V}^\star$ with N copies of V on its diagonal and a permutation matrix $P^\star$ of order $N(N-1)$ such that a decomposition of type (2.7) exists. The associated block tridiagonal matrix $T^\star$ has $N-1$ diagonal blocks, each one of the form

$$T_i^\star = \begin{bmatrix} \lambda_i & \sigma_i & & & 0 \\ \gamma_i & \lambda_i & \sigma_i & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_i & \lambda_i & \sigma_i \\ 0 & & & \gamma_i & \lambda_i^\star \end{bmatrix}_{N\times N},$$

where

$$\lambda_i^\star = m_1^\star + 2m_2^\star \cos\frac{i\pi}{N}, \qquad i = 1, \ldots, N-1,$$

are the eigenvalues of $M_1^\star$. Similarly, the transformed right-hand side vector $\hat{\mathbf{f}}^\star$ can be partitioned into $N-1$ vectors of length N to give $N-1$ independent systems

$$T_i^\star \mathbf{y}_i = \hat{\mathbf{f}}_i^\star. \tag{4.5}$$

For this specific example, the vectors $\hat{\mathbf{f}}_i^\star$ are given by

$$\hat{\mathbf{f}}_i^\star = \epsilon h \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N} \end{bmatrix}_N .$$

The solution of each system (4.5) is therefore the solution of the same constant-coefficient recurrence relation as in the Dirichlet case, but with the right-hand boundary condition now of Neumann type. The roots of the auxiliary equation are given by (2.12), and the recurrence relation solution is

$$y_{ik}^\star = \epsilon h \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N} G_1^\star(i,k), \tag{4.6}$$

where

$$G_1^\star(i,k) = \frac{\mu_1^k - \mu_2^k}{(\gamma_i + \lambda_i^\star \mu_1)\mu_1^{N-1} - (\gamma_i + \lambda_i^\star \mu_2)\mu_2^{N-1}}.$$

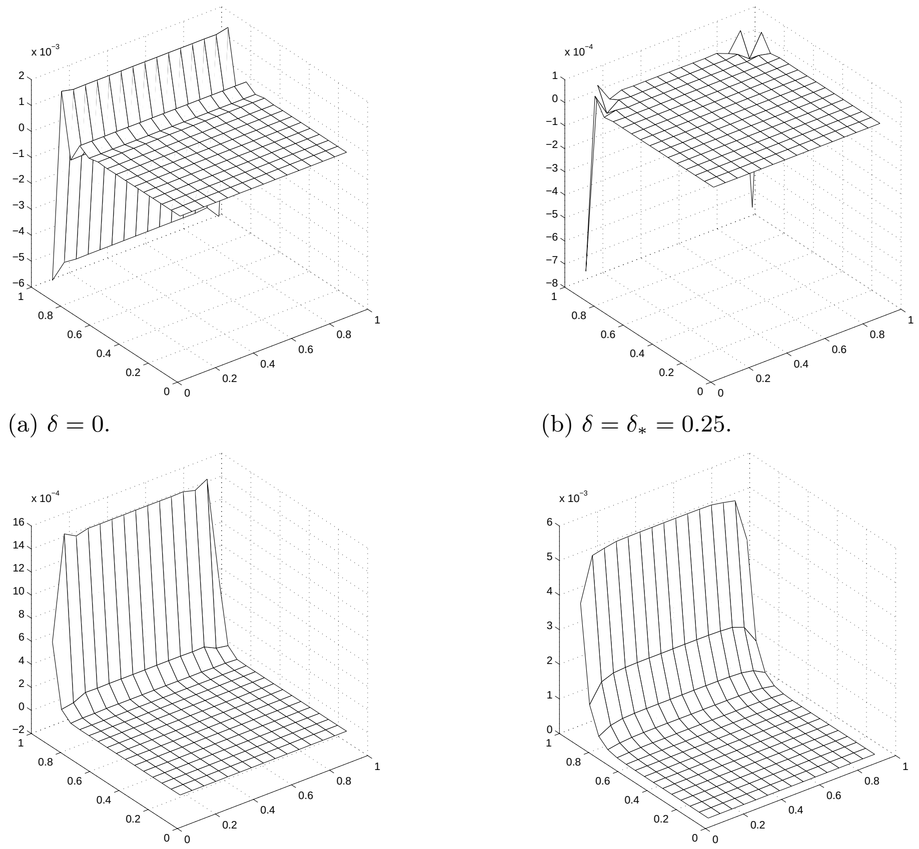
This expression compares with (4.2) in the Dirichlet case. The most significant difference is the factor of ϵh in front of the Neumann solution: this means that for this problem the oscillations will be much smaller than those in the Dirichlet case. Because of the nature of $G_1^\star$ and G_1, however, the effect of changing δ will be very similar in both cases. This is borne out by the plots of the Neumann solution shown in Figure 9 (for $N = 16$ and $P_e = 2$ so that $\epsilon h = 9.8 \times 10^{-4}$). As predicted by the analysis, these solutions are almost identical in shape to those obtained for the Dirichlet problem (see Figure 6), but any oscillations are much smaller in magnitude.

5. Solution accuracy.

We have now characterized the effect of δ on oscillations in the flow direction. One important question which remains is how the choice of δ affects the overall accuracy of the discrete solution. To investigate this, we begin with the example problems of the previous section. In each case, we compare solutions on a 16×16 grid with $\epsilon = 1/64$ (so $P_e = 2$) with a reference solution for the same value of ϵ on a 256×256 grid. On this fine grid, we use the Galerkin method ($\delta = 0$) as $P_e = 0.125 \ll 1$ and there are no oscillations. In what follows, we will denote the fine grid nodal solution vector by $\mathbf{u}_{256}$ and its associated finite element solution by u_{256}, likewise for the coarse grid solutions $\mathbf{u}_{16}^\delta$ and u_{16}^δ.

Figure 10 shows the variation with δ of the error for our test problems measured in two different norms. In all cases the norm of the error is plotted against δ for $0 \le \delta \le 1$ with the values of δ_* (o), δ_{s} ($\Diamond$), and δ^* (x) highlighted. For $P_e = 6.25$ ($\epsilon = 1/200$), $\delta_* = 0.42$, $\delta_{\mathrm{s}} = 0.468$, and $\delta^* = 0.58$. The solid line represents the discrete $L_\infty[0,1]$ norm defined by

$$\|\mathbf{u}_{256} - \mathbf{u}_{16}^\delta\|_\infty = \max_{i,j} |\mathbf{u}_{256}(x_i, y_j) - \mathbf{u}_{16}^\delta(x_i, y_j)|,$$



(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.

(c) $\delta = \delta_{\rm s} = 0.354$.

(d) $\delta = \delta^* = 0.75$.

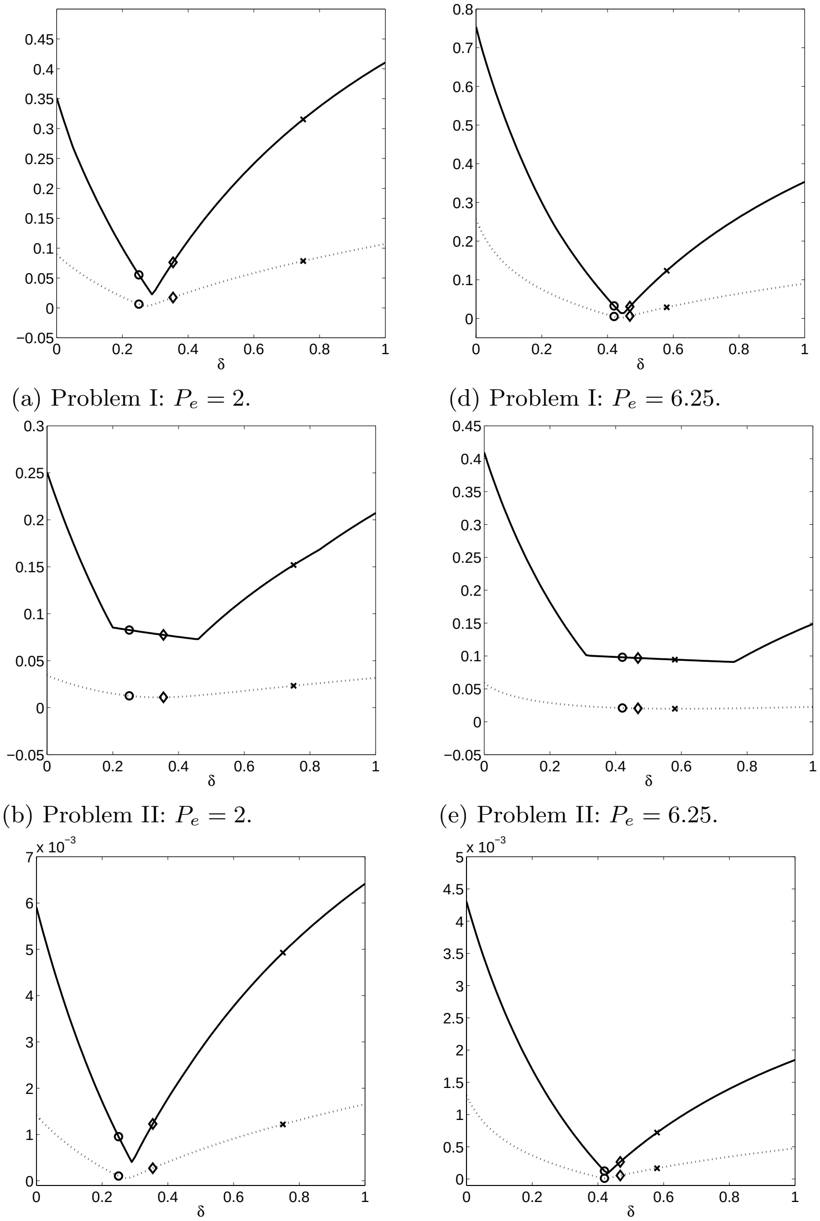
FIG. 9. *Discrete solution at interior node points for Problem* III *with $N = 16$, $P_e = 2$.*

where the points $(x_i, y_j) = (ih, jh)$ are the nodes of the 16×16 grid. When using the finite element method, it may be more natural to work with the L_2 norm

$$\|u_{256} - u_{16}^{\delta}\|_2 = \left\{ \int_{\Omega} \left(u_{256} - u_{16}^{\delta}\right)^2 \right\}^{\frac{1}{2}}. \tag{5.1}$$

However, this measure leads to misleading results for certain singular perturbation problems of this type where the overall error is heavily dominated by the error in the boundary layer, which we cannot hope to resolve on a 16×16 uniform grid using low order elements. For Problems I and III, a more meaningful measure of the error for our purposes is obtained using the L_2 norm of the error away from the boundary layer; that is, in these cases, we omit the top row of coarse grid elements from the region of integration in (5.1) and integrate over $(0, 1) \times (0, 0.9375)$ instead of $\Omega = (0, 1) \times (0, 1)$. This norm is represented by a dotted line in the error plots. We note in passing that in all of the examples, this curve is very similar to that obtained for the discrete L_2 norm defined by

$$\|\mathbf{u}_{256} - \mathbf{u}_{16}^{\delta}\|_2 = \left\{ \sum_{i,j=0}^{N} \left(u_{256}(x_i, y_j) - u_{16}^{\delta}(x_i, y_j)\right)^2 \right\}^{\frac{1}{2}},$$



(a) Problem I: $P_e = 2$.

(d) Problem I: $P_e = 6.25$.

(b) Problem II: $P_e = 2$.

(e) Problem II: $P_e = 6.25$.

(c) Problem III: $P_e = 2$.

(f) Problem III: $P_e = 6.25$.

FIG. 10. *Error variation with δ in the discrete L_∞ norm (solid) and L_2 norm (dotted) for $N = 16$.*

where (x_i, y_j) is again a node of the coarse grid.

From Figure 10, we see that the optimal choice of δ in terms of solution accuracy depends on the norm in which the error is measured, although in most cases both δ_* and $\delta_{\rm s}$ are closer than δ^* to the optimal choice. Note that setting $\delta = \delta_{\rm s}$ to produce a completely oscillation-free discrete solution $\mathbf{u}$ does not result in the most accurate solution.

6. Guidelines for choosing the streamline diffusion parameter in practice. In sections 2–4, we presented model problem analysis which enabled us to characterize the behavior of the discrete finite element solutions. Three highlighted values of δ play important roles in this analysis: δ_*, where the solution is oscillatory but the oscillations are extremely small; δ_s, which is the smallest value of δ such that the solution is found by numerical experiment to be oscillation free; and δ^*, where the solution is guaranteed to be oscillation free via Theorem 3.3. The analysis, based on Fourier techniques, is restricted to grid-aligned flow. (This is needed for the tridiagonal matrices M_1, M_2, and M_3 of (2.4) to be symmetric and have a common set of eigenvectors.) In this section, we consider several more complex problems and make some observations about choosing δ in practice.

First, we observe that although with $\delta = \delta^*$ we have a way of guaranteeing that there are no oscillations, the resulting discrete solutions are overly diffuse and inaccurate: both δ_* and δ_s are in general much better values to use. The choice $\delta = \delta_\mathrm{s}$ produces a completely oscillation-free solution but δ_s is not readily determined even for the model problems considered above. However, we know that δ_s lies between δ_* and δ^*, and the empirical results for Problems I–III suggest that the computable expression

$$\delta_\bullet = \frac{1}{2}\left(1 - \frac{0.8}{P_e}\right) \tag{6.1}$$

is a good approximation to it. Note that in the limit as $P_e \to \infty$, both δ_* and $\delta_\bullet$ tend to 0.5.

We now introduce three new test problems with non-grid-aligned or variable winds. For these problems, we use a stabilization strategy which fixes δ^{el} locally on each element by using the local element mesh Péclet number

$$P_e^{el} = \frac{h^{el}\|\mathbf{w}^{el}\|_2}{2\epsilon}$$

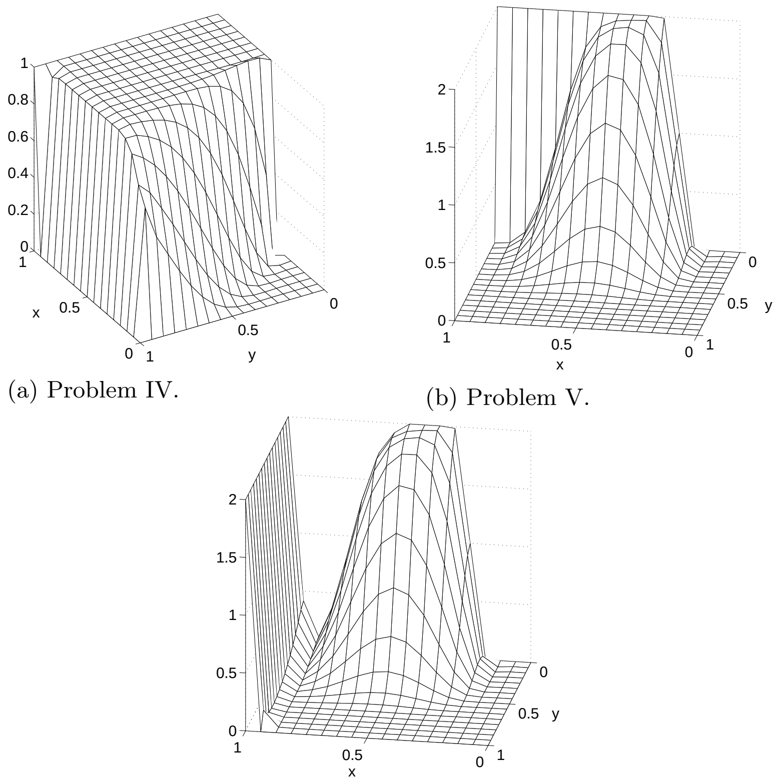
in formulae (3.3) and (6.1). This is calculated using the discrete L_2 norm of the wind value at the element center $\mathbf{w}^{el}$, with the local grid size value h^{el} taken as the distance across the element measured in the direction of the wind. In what follows, these element-based values of δ will be denoted using the superscript *el*. In all cases, the value of the stabilization parameter used is $\max(\delta^{el}, 0)$ on each element.

Problem IV. Here we impose the Dirichlet boundary conditions

$$f_b(x) = \begin{cases} 0, & 0 < x \le \frac{1}{2}, \\ 1, & \frac{1}{2} < x < 1, \end{cases} \qquad f_t(x) = f_l(y) = 0, \quad f_r(y) = 1$$

on the domain in Figure 1 and apply the wind $\mathbf{w} = (\cos 115°, \sin 115°)$ which has constant magnitude and direction but is not aligned with the grid. This problem has an exponential boundary layer on a portion of the outflow boundary and an internal layer along the characteristic caused by the discontinuity on the inflow boundary. A sample solution with $N = 16$, $\epsilon = 1/200$, and $\delta = \delta_*^{el}$ is shown in Figure 11 (a).

Error calculations carried out as described in the previous section lead to the plots in Figure 12 (a) and (d), where the values δ_*^{el} ($\circ$), $\delta^{el,*}$ ($\times$), and $\delta_\bullet^{el}$ ($\diamondsuit$) have been highlighted. When $\delta = 0$, the error is dominated by difficulties associated with the exponential layers at the outflow. As δ is increased so that these layers begin to be resolved, the error is then dominated by the effect of the discontinuity in the



(a) Problem IV.

(b) Problem V.

(c) Problem VI.

FIG. 11. *Sample solutions with $N = 16$, $\epsilon = 1/200$, and $\delta = \delta_*^{el}$.*

inflow boundary condition which is relatively insensitive to the value of δ, causing the middle of the plots to look fairly flat.

Problem V. Our two variable wind test problems are variants of the "IAHR/CEGB" test problem proposed in [14]. In this first case, we solve (1.1) on the unit square with

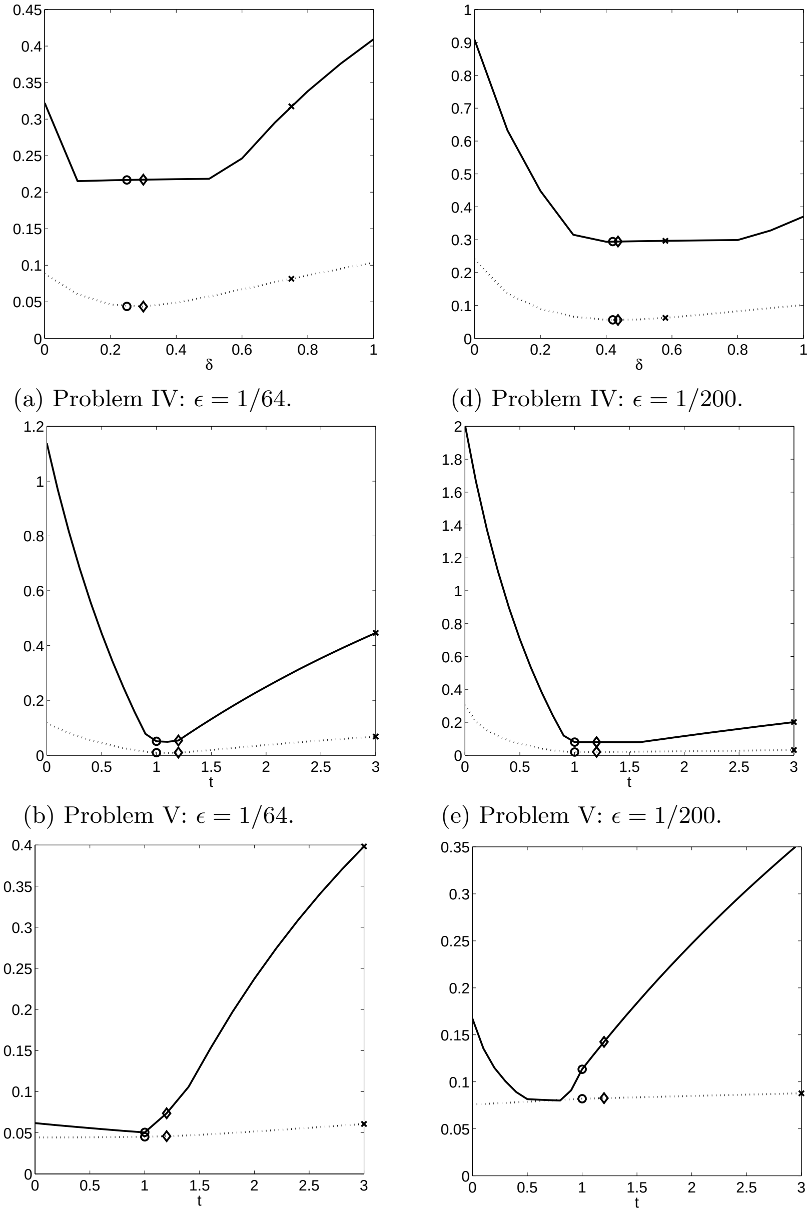
$$\mathbf{w} = (2y(1-(2x-1)^2), -2(2x-1)(1-y^2)). \tag{6.2}$$

The Dirichlet boundary conditions are given by

$$u(x,0) = 1 + \tanh[10 + 20(2x-1)] \tag{6.3}$$

on the inflow boundary (the interval $0 \le x \le 0.5$, $y = 0$) and $u(x,0) = 2$ on the outflow boundary (the interval $0.5 < x \le 1$, $y = 0$). On the remaining boundaries, we impose $f_t(x) = f_l(y) = f_r(y) = 0$. The Dirichlet boundary conditions at the bottom $y = 0$ are continuous but there is an exponential layer at the outflow portion, i.e., where $x \ge 1/2$. A sample solution for $N = 16$ and $\epsilon = 1/200$ is shown in Figure 11 (b).

As the wind now varies in magnitude and direction from element to element, we cannot identify a single parameter δ which can be varied for the purposes of comparing errors as in the previous examples. However, we can compare various strategies for choosing δ^{el} locally within elements by considering the parameterized version of δ^{el}



(a) Problem IV: $\epsilon = 1/64$.

(d) Problem IV: $\epsilon = 1/200$.

(b) Problem V: $\epsilon = 1/64$.

(e) Problem V: $\epsilon = 1/200$.

(c) Problem VI: $\epsilon = 1/64$.

(f) Problem VI: $\epsilon = 1/200$.

FIG. 12. *Error variation with δ in the discrete L_∞ norm (solid) and L_2 norm (dotted) for* $N = 16$.

given by

$$\delta^{el} = \begin{cases} \dfrac{t}{2}\left(1 - \dfrac{1}{P_e^{el}}\right), & 0 \le t \le 1, \\[2ex] \dfrac{1}{2}\left(1 + (t-2)\dfrac{1}{P_e^{el}}\right), & 1 < t \le 3. \end{cases} \tag{6.4}$$

As t varies from 0 to 3, the value of δ^{el} on each element first increases linearly from 0 to δ_*^{el} (at $t = 1$) and then varies linearly between δ_*^{el} and $\delta^{el,*}$. The variation with t of the error for this problem for $N = 16$ with two different values of ϵ is shown in Figure 12 (b) and (e). The errors are again calculated as described in section 5. The values δ_*^{el} ($\circ$), $\delta^{el,*}$ ($\times$), and $\delta_\bullet^{el}$ ($\diamondsuit$) are highlighted. The error is dominated by problems caused by the exponential layer along the outflow boundary in a similar way to Problem I.

Problem VI. Our final test problem also has a variable wind given by (6.2) but the boundary conditions are now of mixed type. We again impose the Dirichlet condition (6.3) on the inflow boundary but now the condition imposed on the outflow boundary is a homogeneous Neumann one. The Dirichlet boundary conditions on the remaining boundaries are $f_t(x) = f_l(y) = 0$, $f_r(y) = 2$. This results in the formation of a characteristic boundary layer along the right-hand wall. A sample solution for $N = 16$ and $\epsilon = 1/200$ is shown in Figure 11 (c). Error plots for this problem with δ parameterized by t as in (6.4) are shown in Figure 12 (c) and (f). This problem features a characteristic layer, so we expect the effects of changing δ to be less pronounced, as for Problem II. This is supported by the error plots: increasing δ helps to resolve the characteristic layer until the error becomes dominated by the effects of boundary discontinuities.

The results in these experiments are essentially the same as those for the model problems. We have not displayed oscillations here, but in all of the examples, the solutions for $\delta = \delta_*$ contain slight oscillations near layers, and the choice $\delta = \delta_\bullet$ reduces but does not eliminate them in these examples. There is little difference between these values in terms of solution quality obtained, and both choices are generally better than δ^*, which adds too much diffusion. Although it is tempting to use the interpolated value $\delta_\bullet$ to produce a qualitatively smoother solution, in our view δ_* is a better choice. The oscillations it produces indicate that in fact the layers are *not* fully resolved and that mesh refinement is needed where they occur; the smoothing of these effects will be misleading (see, e.g., [4]). Streamline diffusion alone cannot completely resolve this issue, and the choice δ_* adds the right amount of diffusion to keep the errors small in most of the domain. Note that this value has previously been recommended as a good choice in [1] and was shown to be good for efficient solution of the resulting linear system by the GMRES iterative method [3]. We also remark that although the analysis of sections 2–4 does not apply to linear elements, we have performed a few experiments which indicate that δ_* yields more accurate solutions than $\delta_\bullet$ in the linear case and that the latter choice adds excessive diffusion in this setting.

7. Application to other discretizations.

To conclude, we emphasize that analysis of this type can be applied to any discretization whose stencil is of the form (2.3). We comment on two particular cases of interest here.

7.1. Finite differences with streamline diffusion.

The usual central finite difference discretization of (1.1) can also be stabilized using streamline diffusion; see, for example, [12, p. 1465]. Specifically, we apply the finite difference method to the differential equation

$$-(\epsilon\nabla^2 + \nabla \cdot D\nabla)u(x, y) + \mathbf{w} \cdot \nabla u(x, y) = f(x, y),$$

where diffusion in the streamline direction is added using

$$D = \alpha \begin{bmatrix} c^2 & cs \\ cs & s^2 \end{bmatrix}$$

with

$$c = \frac{w_1}{\|\mathbf{w}\|_2}, \qquad s = \frac{w_2}{\|\mathbf{w}\|_2},$$

and α as in (1.4). Assuming for convenience that $\|\mathbf{w}\|_2 = 1$, the full computational molecule is given by

$$\begin{tikzcd}
\dfrac{w_1 w_2 \delta}{2h} & -\dfrac{\epsilon}{h^2} + \dfrac{w_2}{2h} - \dfrac{w_2^2\delta}{h} & -\dfrac{w_1 w_2 \delta}{2h} \\
-\dfrac{\epsilon}{h^2} - \dfrac{w_1}{2h} - \dfrac{w_1^2\delta}{h} & \dfrac{4\epsilon}{h^2} + \dfrac{2\delta}{h} \arrow[ul] \arrow[u] \arrow[ur] \arrow[l] \arrow[r] \arrow[dl] \arrow[d] \arrow[dr] & -\dfrac{\epsilon}{h^2} + \dfrac{w_1}{2h} - \dfrac{w_1^2\delta}{h} \;. \\
-\dfrac{w_1 w_2 \delta}{2h} & -\dfrac{\epsilon}{h^2} - \dfrac{w_2}{2h} - \dfrac{w_2^2\delta}{h} & \dfrac{w_1 w_2 \delta}{2h}
\end{tikzcd}$$

This simplifies to a stencil of standard five-point type for our model problem (2.1) with grid-aligned flow. Using the notation of (2.3), the stencil entries are

$$m_1 = \frac{4\epsilon}{h^2} + \frac{2\delta}{h}, \qquad m_2 = -\frac{\epsilon}{h^2}, \qquad m_3 = -\frac{\epsilon}{h^2} + \frac{1}{2h} - \frac{\delta}{h},$$

$$m_4 = 0, \qquad m_5 = -\frac{\epsilon}{h^2} - \frac{1}{2h} - \frac{\delta}{h}, \qquad m_6 = 0$$

with related eigenvalues

$$\gamma_i = \frac{1}{h^2}\left[-(\epsilon + \delta h) - \frac{h}{2}\right], \qquad \lambda_i = \frac{1}{h^2}\left[2(\epsilon + \delta h) + 2\epsilon(1 - C_i)\right],$$

$$\sigma_i = \frac{1}{h^2}\left[-(\epsilon + \delta h) + \frac{h}{2}\right].$$

This results in the expressions

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

for the roots of the recurrence relation which appear in (2.13).

Here the sign of μ_2 (and hence the nature of the corresponding functions $G_1(i,k)$ and $G_2(i,k)$, $i \in S_N$) is independent of i: as the numerator of μ_2 is always negative, we simply have the conditions

$$\begin{cases} \delta > \delta_* \quad \Rightarrow \quad \mu_2 > 0,\ G_1(i,k) \text{ is nonoscillatory}, \\[2ex] \delta < \delta_* \quad \Rightarrow \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory}, \end{cases}$$

where δ_* is given by (3.3). Hence the result equivalent to Theorem 3.1 is given by the following theorem.

THEOREM 7.1. *For a streamline diffusion finite difference discretization with* $P_e > 1$, $\delta > \delta_*$ *implies that* $G_1(i,k)$ *and* $G_2(i,k)$ *in* (2.13) *are nonoscillatory functions of* k *for any value of* $i \in S_N$.

The special case $\delta = \delta_*$ leads to the two-term recurrence with auxiliary equation root

$$\rho = \frac{1}{1 + \dfrac{(1 - C_i)}{P_e}}$$

and solution (3.5). Because $\rho < 1$, this solution is nonoscillatory in the streamline direction for all $i \in S_N$ and, as in the finite element case, tends to the nodally exact solution in the limit as $P_e \to \infty$.

The fact that there is one critical parameter (independent of i) here means that there is no issue about selecting a global parameter δ as we had in the finite element case. Furthermore, the analysis of the effect of transforming from $\mathbf{y}$ to $\mathbf{u}$ (cf. section 3.3) is greatly simplified. In particular, for the same specific example problem with $f_t = 1$ and $f_b = f_l = f_r = 0$ studied in section 3.3, the equivalent expression to (4.2) using finite differences has $S_{\text{smooth}} = 0$ when $\delta < \delta_*$ and $S_{\text{osc}} = 0$ when $\delta > \delta_*$. Thus we immediately have the following theorem (cf. Theorem 3.3).

THEOREM 7.2. *For a streamline diffusion finite difference discretization of* (2.1), *the discrete solution* $\mathbf{u}$ *does not exhibit oscillations in the streamline direction when* $\delta \geq \delta_*$.

That is, in contrast to the finite element case, there is no "smoothing" introduced by the Fourier transformation: the same single parameter governs the presence of oscillations in both the recurrence relation solution $\mathbf{y}$ and the discrete two-dimensional solution $\mathbf{u}$.

7.2. Artificial diffusion. So far we have focused on adding smoothing in the streamline direction only, which is just one of the many stabilization methods available. In this section we analyze the artificial diffusion method (see, for example, [7, pp. 218–219]) with a view to comparing its smoothing effect with that of streamline diffusion. The artificial diffusion method works by adding diffusion in an isotropic way which does not take account of flow direction, and it is well known that this can result in smearing of internal layers. We can use the analytical techniques presented in this paper to confirm that the streamline diffusion method avoids this problem.

We again consider a vertical wind model problem using bilinear finite elements on a uniform grid. The idea of the artificial diffusion method is to replace equation (2.1) with

$$-(\epsilon + \delta h)\nabla^2 u + \frac{\partial u}{\partial y} = 0 \qquad \text{in } \Omega = (0,1) \times (0,1), \tag{7.1}$$

with δ once again a stabilization parameter to be chosen. When $P_e < 1$, we set $\delta = 0$ as before. Galerkin discretization using bilinear finite elements results in a matrix of the form (2.4), which is therefore covered by our analysis. The stencil entries in this case are given by

$$m_1 = \frac{8}{3}(\delta h + \epsilon), \qquad m_2 = -\frac{1}{3}(\delta h + \epsilon), \qquad m_3 = -\frac{1}{3}[(\delta - 1)h + \epsilon],$$

$$m_4 = -\frac{1}{12}[(4\delta - 1)h + 4\epsilon], \quad m_5 = -\frac{1}{3}[(\delta + 1)h + \epsilon], \quad m_6 = -\frac{1}{12}[(4\delta + 1)h + 4\epsilon],$$

so the roots (2.12) of the corresponding recurrence relation are given by

$$\mu_{1,2} = \frac{-\left(2\delta + \frac{1}{P_e}\right)\left[\frac{4 - C_i}{2 + C_i}\right] \pm \sqrt{1 + \frac{3(1 - C_i)(5 + C_i)}{(2 + C_i)^2}\left(2\delta + \frac{1}{P_e}\right)^2}}{1 - \left(2\delta + \frac{1}{P_e}\right)\left[\frac{1 + 2C_i}{2 + C_i}\right]}. \tag{7.2}$$

First we briefly consider the issue of oscillations in the streamline direction. Here, as in section 3.2, the sign of μ_2 (and hence the presence of oscillations in the recurrence relation solution) depends on the value of $i \in S_N$. Defining the new critical value

$$\tilde{\delta}_i^c = \frac{1}{2}\left(\left[\frac{2 + C_i}{1 + 2C_i}\right] - \frac{1}{P_e}\right),$$

we have different conditions for two sets of i values, namely

$$1 \le i \le \tfrac{2}{3}N : \quad \begin{cases} \delta > \tilde{\delta}_i^c & \Rightarrow \quad \mu_2 > 0,\ G_1(i,k) \text{ is nonoscillatory}, \\ \delta < \tilde{\delta}_i^c & \Rightarrow \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory}, \end{cases}$$

$$\tfrac{2}{3}N < i \le N - 1 : \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory}.$$

Notice that this is different from the streamline diffusion case (cf. Theorem 3.1) in that there is no choice of δ which will make the recurrence relation solution oscillation free, as some of the contributing functions $G_1(i,k)$ are always oscillatory. However, it can be seen using an argument of the type presented in section 3.3 that the transformed solution is again dominated by contributions from functions pertaining to lower values of i. Hence, despite the fact that $G_1(i,k)$ is always oscillatory for large i, it is still possible to obtain a nonoscillatory discrete solution $\mathbf{u}$. Note that inequality (3.4) is satisfied with δ_c^i replaced by $\tilde{\delta}_c^i$. For the particular (i-independent) choice $\delta = \delta_*$ from (3.3), equation (7.1) (and hence the artificial diffusion solution) is independent of ϵ.

To gain insight into the main difference between this method and the streamline diffusion technique, we must examine solution behavior in the "crosswind" direction, that is, perpendicular to the direction of the flow. To fix ideas, we will use the discontinuous boundary conditions

$$f_b(x) = \begin{cases} 0, & x < 0.5, \\ 1, & x \ge 0.5, \end{cases} \qquad f_r(y) = 1, \quad f_t(x) = f_l(y) = 0$$

so that the solution has an internal layer along $x = 0.5$ as well as a boundary layer along the right half of the top boundary. The internal layer derives from propagation of the bottom boundary condition through the domain and, as $\epsilon \to 0$, the width of this layer tends to zero. Ideally, this phenomenon should be reproduced by a discretization method, that is, we would like to obtain a set of discrete solutions $\mathbf{u}$ in this limit whose variation from the bottom boundary function is independent of j for fixed k. We now show that while the streamline diffusion method has this property, the artificial diffusion method does not.

Consider the recurrence relation solution vector $\mathbf{y}$ for this problem. From (2.13), its entries are given by

$$y_{ik} = F_3(i)\,(1 - G_1(i,k)) + [F_2(i) - F_3(i)]\,G_2(i,k) \tag{7.3}$$

with

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

[2, appendix] and $F_3(i)$ as in (3.6). As the functions $F_2(i)$ and $F_3(i)$ are the same for both discretizations, any difference in solution behavior must come from a difference in the behavior of the functions $G_1(i,k)$ and $G_2(i,k)$ associated with the two methods. We therefore now focus on how these functions vary with $i \in S_N$ as $\epsilon \to 0$ ($P_e \to \infty$) for $k \in S_N$ fixed. To simplify the presentation of this analysis, we will assume that δ is fixed independent of P_e, with $\delta \neq 0, 0.5$.

With the streamline diffusion discretization, neglecting terms of $O(P_e^{-1})$ and higher in (3.1) gives the approximations

$$\mu_1 \simeq 1, \qquad \mu_2 \simeq \frac{2\delta+1}{2\delta-1} \equiv \beta$$

so that

$$\begin{aligned} G_1(i,k) &= \frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N} \simeq \frac{1-\beta^k}{1-\beta^N} \equiv G_1^a(k), \\ G_2(i,k) &= (1-\mu_1^k) - (1-\mu_1^N)G_1(i,k) \simeq 0. \end{aligned}$$

Thus, in the limit as $P_e \to \infty$, both functions are independent of i. We then have

$$y_{ik} \simeq F_3(i)(1 - G_1^a(k));$$

hence, using (2.8),

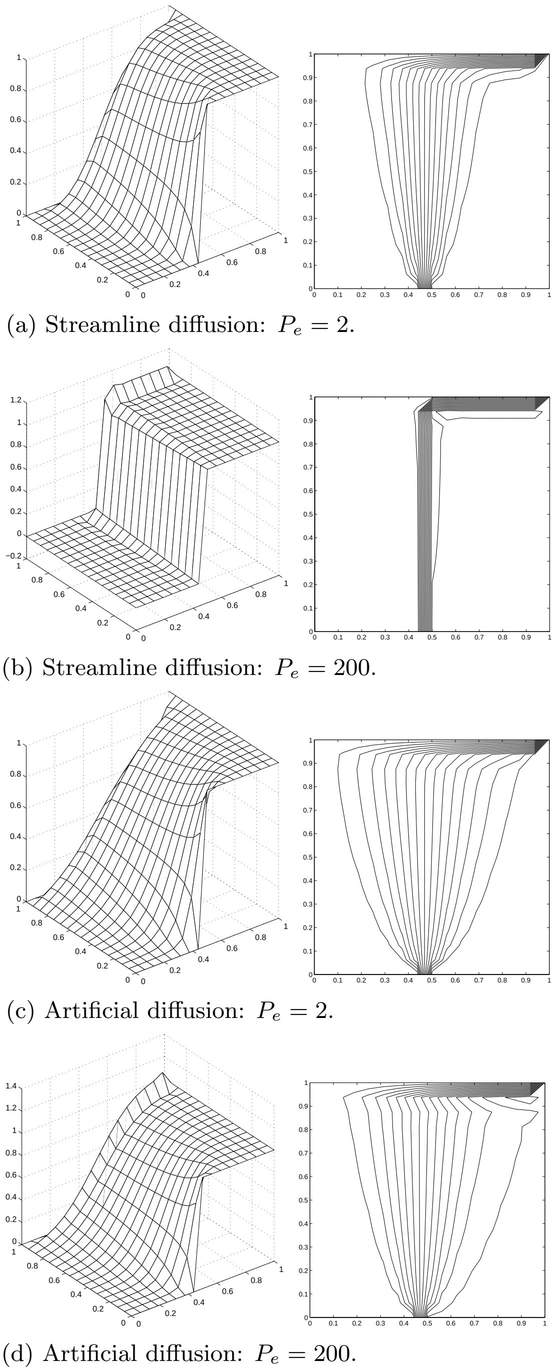
$$u_{jk} \simeq f_b(x_j)(1 - G_1^a(k)).$$

That is, the variation of u_{jk} from the bottom boundary function is independent of j in this limit. For the artificial diffusion discretization, however, neglecting terms of $O(P_e^{-1})$ and higher in (7.2) gives

$$\mu_{1,2} \simeq \frac{-2\delta(4-C_i) \pm \sqrt{4(1+15\delta^2) + 4(1-12\delta^2)C_i + (1-12\delta^2)C_i^2}}{2(1-\delta) + (1-4\delta)C_i},$$

leading to approximations for $G_1(i,k)$ and $G_2(i,k)$ which depend on i through C_i. From (7.3) the solution is therefore

$$u_{jk} \simeq f_b(x_j) - \sqrt{\frac{2}{N}}\sum_{i=1}^{N-1} \sin\frac{ij\pi}{N}\left(F_3(i)G_1(i,k) - [F_2(i) - F_3(i)]\,G_2(i,k)\right).$$



(a) Streamline diffusion: $P_e = 2$.

(b) Streamline diffusion: $P_e = 200$.

(c) Artificial diffusion: $P_e = 2$.

(d) Artificial diffusion: $P_e = 200$.

FIG. 13. *Solutions and contour plots for $\delta = 0.4$ and $N = 16$.*

This has a j-dependence which the continuous solution in this limit does not.

This fundamental difference between the discretizations is demonstrated pictorially in Figure 13, which shows streamline and artificial diffusion approximations (and associated contour plots) for this example problem with two values of ϵ, $\delta = 0.4$, and $N = 16$. Plots (a) and (b) show that the streamline diffusion method captures the narrowing of the internal layer exhibited by the continuous solution as $\epsilon \to 0$ ($P_e \to \infty$). The equivalent artificial diffusion approximation does not, as shown in plots (c) and (d).

8. Summary. In this study, we have performed a Fourier analysis of model problems with grid-aligned flow that identifies the effects of upwinding in discretizations of the convection-diffusion equation. Our emphasis is on streamline-diffusion discretization with bilinear elements, where we show how the choice of streamline diffusion parameter affects the qualitative behavior of the solution with respect to oscillations. This analysis gives theoretical justification for the choice

$$\delta = \delta_* = \frac{1}{2}\left(1 - \frac{1}{P_e^{el}}\right).$$

Our analysis also shows that δ_* is the optimal choice for finite difference discretizations, provides insight into the method of isotropic artificial diffusion, and yields qualitatively good solutions in a variety of computational experiments.

REFERENCES

[1] A. Brooks and T. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[2] H.C. Elman and A. Ramage, *A characterisation of oscillations in the discrete two-dimensional convection-diffusion equation,* Math. Comp., to appear.

[3] B. Fischer, A. Ramage, D.J. Silvester, and A. J. Wathen, *On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems,* Comput. Methods Appl. Mech. Engrg., 179 (1999), pp. 179–195.

[4] P.M. Gresho and R.L. Lee, *Don't suppress the wiggles—they're telling you something,* in Finite Element Methods for Convection Dominated Flows, AMD 34, T.J.R. Hughes, ed., ASME, New York, 1979, pp. 37–61.

[5] P.M. Gresho and R.L. Sani, *Incompressible Flow and the Finite Element Method*, John Wiley and Sons, Chichester, UK, 1999.

[6] M.D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Comput. Sci. Sci. Comput., Academic Press, New York, 1989.

[7] W. Hackbusch, *Multi-grid Methods and Applications*, Springer-Verlag, New York, 1980.

[8] T.J.R. Hughes and A. Brooks, *A multidimensional upwind scheme with no crosswind diffusion*, in Finite Element Methods for Convection Dominated Flows, AMD 34, T.J.R. Hughes, ed., ASME, New York, 1979, pp. 120–131.

[9] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.

[10] K.W. Morton, *Numerical Solution of Convection-Diffusion Problems,* Chapman and Hall, London, 1996.

[11] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations.* Springer-Verlag, New York, 1994.

[12] H.-G. Roos, *Necessary convergence conditions for upwind schemes in the two-dimensional case,* Internat. J. Numer. Methods Engrg., 21 (1985), pp. 1459–1469.

[13] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.

[14] R.M. Smith and A.G. Hutton, *The numerical treatment of advection—a performance comparison of current methods,* Numer. Heat Trans., 5 (1982), pp. 439–461.