Numer. Math. 58, 185–202 (1990)

Numerische Mathematik

Some upwinding techniques for finite element approximations of convection-diffusion equations

Randolph E. Bank [1,*], **Josef F. Bürgler** [2,**], **Wolfgang Fichtner** [3], **and R. Kent Smith** [4]

[1] Department of Mathematics, University of California at San Diego, La Jolla, CA 92093, USA
[2] Integrated Systems Laboratory, Swiss Federal Institute of Technology, CH-8092 Zürich, Switzerland
[3] Integrated Systems Laboratory, Swiss Federal Institute of Technology, CH-8092 Zürich, Switzerland
[4] AT & T Bell Laboratories, Murray Hill, NJ 07974, USA

Received September 29, 1989/April 28, 1990

Summary. A uniform framework for the study of upwinding schemes is developed. The standard finite element Galerkin discretization is chosen as the reference discretization, and differences between other discretization schemes and the reference are written as artificial diffusion terms. These artificial diffusion terms are spanned by a four dimensional space of element diffusion matrices. Three basis matrices are symmetric, rank one diffusion operators associated with the edges of the triangle; the fourth basis matrix is skew symmetric and is associated with a rotation by $\pi/2$. While finite volume discretizations may be written as upwinded Galerkin methods, the converse does not appear to be true. Our approach is used to examine several upwinding schemes, including the streamline diffusion method, the box method, the Scharfetter-Gummel discretization, and a divergence-free scheme.

Subject classifications: AMS(MOS): 65N05, 65N10, 65N20; CR: G1.8.

1 Introduction

We consider the model convection diffusion problem

$$-\nabla\cdot(\nabla u+\beta u)=0 \quad \text{in } \Omega\subset\mathscr{R}^2 \tag{1}$$
$$u=u_0 \quad \text{on } \partial\Omega_1$$
$$(\nabla u+\beta u)\cdot n=0 \quad \text{on } \partial\Omega-\partial\Omega_1.$$

* The work of this author was supported by the Office of Naval Research under contract N00014-89J-1440
** The work of this author was supported through KWF-Landis/Gyr Grant 1496, AT&T Bell Laboratories, and Cray Research

Offprint requests to: R.E. Bank

Here $\beta = \nabla \psi$ and $\psi \in \mathcal{H}^1(\Omega)$. We assume that Ω is polygonal and that $\partial\Omega_1$ is composed of one or more edges of $\partial\Omega$. The function u_0 is assumed constant on each contiguous set of Dirichlet boundary edges. The outward normal direction n is defined edgewise.

The weak form of (1) is: find $u \in \mathcal{H}_d$ such that

$$a(u, \phi) = \int_\Omega (\nabla u + \beta u) \cdot \nabla \phi \, dx \, dy = 0 \tag{2}$$

for all $\phi \in \mathcal{H}_0$, where

$$\mathcal{H}_d = \{u \in \mathcal{H}^1(\Omega) \text{ and } u = u_0 \text{ on } \partial\Omega_1\}$$

$$\mathcal{H}_0 = \{u \in \mathcal{H}^1(\Omega) \text{ and } u = 0 \text{ on } \partial\Omega_1\}.$$

Let $\mathcal{T}$ be a shape regular, although not necessarily quasi uniform, triangulation of Ω, characterized by a small parameter h indicating the size of the elements. Let $\mathcal{S}_h$ be the space of continuous piecewise linear polynomials with respect to $\mathcal{T}$, and define

$$\mathcal{S}_d = \{u \in \mathcal{S}_h \text{ and } u = u_0 \text{ on } \partial\Omega_1\}$$

$$\mathcal{S}_0 = \{u \in \mathcal{S}_h \text{ and } u = 0 \text{ on } \partial\Omega_1\}.$$

Here we are assuming that each point at which the type of boundary condition changes from Dirichlet to Neumann is a vertex in the triangulation $\mathcal{T}$. Also, we will assume that $\beta = \nabla \psi$, where $\psi \in \mathcal{S}_h$. In the practical application that we have in mind, Eq. (1) is a current continuity equation from the semiconductor device model and β is the gradient of the electrostatic potential, which itself is obtained as part of the solution of a coupled system of partial differential equations [2].

The classical Galerkin finite element method for approximating (2) is: find $u_g \in \mathcal{S}_d$ such that

$$a(u_g, \phi) = 0 \tag{3}$$

for all $\phi \in \mathcal{S}_0$. The classical method roughly corresponds to the use of centered differences in the finite difference context, and is well known to be unstable when $|\beta| h$ is large.

This has led to the use of upwind finite element techniques [4–9], which are analogous to the use of upwind differences in the finite difference arena. In this paper, we develop a uniform framework for the study of general upwinding schemes. We choose the standard weak Galerkin form (3) as the reference discretization. Then differences between other discretization schemes and the weak Galerkin form are written as artificial diffusion terms; that is, we seek to write all schemes in the form:

$$a_h(u, \phi) = a(u, \phi) + \sum_{\tau \in \mathcal{T}} \int_\tau h_\tau (\rho \nabla u) \cdot \nabla \phi \, dx \, dy = 0. \tag{4}$$

Here $\rho \equiv \rho_\tau$ is a 2×2 diffusion matrix, defined elementwise and is characteristic of the particular scheme, and h_τ is a measure of the size of τ, for example, its diameter. Normally, one might tend to think of ρ as a symmetric, positive

semidefinite matrix, but this will not be the case with many of the methods. The bilinear form $a_h(\cdot,\cdot)$ formally corresponds to the perturbed equation

$$-\nabla\cdot((I+h_\tau\rho)\nabla u+\beta u)=0$$

for $\tau\in\mathcal{T}$.

For piecewise linear triangular elements, the diffusion term $h_\tau\rho$ is contained in a four dimensional space of element diffusion matrices. Three basis matrices for this space are symmetric, rank one diffusion operators that can naturally be associated with the edges of the triangle. The fourth basis matrix is skew symmetric and is associated with a rotation by $\pi/2$.

In this paper, we will first consider the streamline diffusion method, proposed and analyzed by Hughes et al. [5, 6] and Johnson et al. [8], among others. As this is a standard approach, we do not make a formal derivation of the method, but rather refer to the existing literature.

We then consider the box scheme [1] and the Scharfetter-Gummel scheme [2], two finite volume discretizations. Our recasting of these schemes in the form (4) may be regarded as an extension of [1], in which only self-adjoint problems were considered. Interestingly, while finite volume discretizations may always be written as upwinded Galerkin methods, the converse does not appear to be true, since the skew symmetric elementary diffusion operator seems to have no analogue in the standard finite volume framework.

Finally, we consider the new divergence free upwinding scheme proposed by the authors [3]. In some instances, the artificial diffusion introduced by this method resembles that of the streamline diffusion method. In other cases, it can lead to very nonsymmetric and indefinite artificial diffusion matrices. In extreme cases, the overall diffusion matrix $I+h_\tau\rho$ can have one positive and one negative eigenvalue. Nevertheless, the method appears to be extremely robust and stable, and remains so even in unfavorable situations where other upwinding schemes fail [3].

The remainder of this paper is organized as follows: In Sect. 2, we describe the triangular element geometry and elemental stiffness matrix. In addition, the element diffusion matrices and there properties are presented. The next four sections are devoted to discussions of the various upwinding schemes in terms of these elemental matrices. We make some concluding remarks in the final section.

2 Preliminaries

Let $\{\phi_i\}_{i=1}^n$ denote the standard nodal basis functions for $\mathcal{S}_0$. Then the global stiffness matrix A corresponding to (3) is given by

$$A_{ij}=a_h(\phi_j,\phi_i). \tag{5}$$

The global stiffness matrix may be decomposed in terms of element stiffness matrices A_τ as

$$A=\sum_{\tau\in\mathcal{T}}A_\tau$$

where

$$(A_\tau)_{ij}=a_\tau(\phi_j,\phi_i)$$

$$a_\tau(\phi_j,\phi_i)=\int_\tau(I+h_\tau\rho)\nabla\phi_j\cdot\nabla\phi_i+\beta\phi_j\cdot\nabla\phi_i\,\mathrm{d}x\,\mathrm{d}y.$$

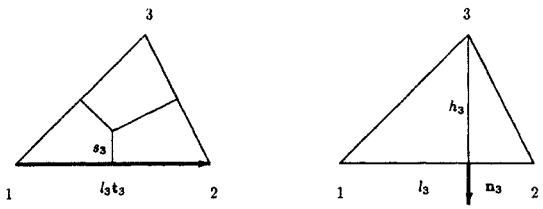

Fig. 1. Parameters associated with the triangle τ

Since there are only three nonzero basis functions on each element, we can characterize A_τ by a dense 3×3 element matrix. Without loss of generality, or by virtue of a local coordinate renumbering, we assume that our canonical element τ has vertices $v_i^t = (x_i, y_i)$, for $1 \leqq i \leqq 3$, and corresponding nodal basis functions $\{\phi_i\}_{i=1}^3$.

We define $\{n_i\}_{i=1}^3$ to be the unit outward normal vectors for τ, $\{t_i\}_{i=1}^3$ to be the unit tangent vectors for the three edges, $\{l_i\}_{i=1}^3$ to be their lengths, and $\{h_i\}_{i=1}^3$ to be the perpendicular heights (see Fig. 1). Let $\tilde{v}$ be the point of intersection for the perpendicular bisectors of the three sides of τ. Let $|s_j|$ denote the distance between $\tilde{v}$ and side j. If τ has no obtuse angles, then the s_j will be nonnegative; otherwise, the distance to the side opposite the obtuse angle will be negative.

There are many relationships among these quantities; in particular we note the following:

$$l_i h_i = 2|\tau|, \qquad 1 \leqq i \leqq 3 \tag{6}$$

$$\nabla \phi_i = -n_i / h_i, \qquad 1 \leqq i \leqq 3 \tag{7}$$

$$\phi_1 + \phi_2 + \phi_3 = 1 \tag{8}$$

$$\nabla \phi_1 + \nabla \phi_2 + \nabla \phi_3 = 0 \tag{9}$$

$$l_1 t_1 + l_2 t_2 + l_3 t_3 = 0 \tag{10}$$

$$\begin{bmatrix} l_1 t_1^t \\ l_2 t_2^t \\ l_3 t_3^t \end{bmatrix} [\nabla \phi_1 \; \nabla \phi_2 \; \nabla \phi_3] = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \tag{11}$$

$$s_1 = -|\tau| \, l_i \nabla \phi_2 \cdot \nabla \phi_3 . \tag{12}$$

Equation (12) is valid cyclically for s_2 and s_3. A hint for verifying (12) is to recall that, if the angle at vertex v_1 is θ_1, then the angle at $\tilde{v}$ between the lines joining $\tilde{v}$ to v_2 and $\tilde{v}$ to v_3 is $2\theta_1$.

The affine mapping of the reference element $\hat{\tau}$, with vertices $(\hat{x}_1, \hat{y}_1) = (0, 0)$, $(\hat{x}_2, \hat{y}_2) = (1, 0)$, and $(\hat{x}_3, \hat{y}_3) = (0, 1)$, to our canonical element τ is given by

$$\begin{bmatrix} x \\ y \end{bmatrix} = J \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} + v_1 \tag{13}$$

with

$$J = [l_3 t_3 - l_2 t_2] \tag{14}$$

and

$$J^{-t} = [\nabla \phi_2 \; \nabla \phi_3]. \tag{15}$$

The function ∇u defined on τ is transformed to $J^{-t}\hat{\nabla}\hat{u}$ defined on the reference element $\hat{\tau}$. The local basis functions on the reference element are

$$\hat{\phi}_1 = 1 - \hat{x} - \hat{y}$$
$$\hat{\phi}_2 = \hat{x}$$
$$\hat{\phi}_3 = \hat{y}.$$

Assuming the β is constant on τ, as will be the case when $\beta = \nabla\psi$ for $\psi \in \mathcal{S}_h$, the element stiffness matrix for the standard Galerkin method is given by

$$A_g = |\tau| \begin{bmatrix} \nabla\phi_1\cdot\nabla\phi_1 & \nabla\phi_1\cdot\nabla\phi_2 & \nabla\phi_1\cdot\nabla\phi_3 \\ \nabla\phi_1\cdot\nabla\phi_2 & \nabla\phi_2\cdot\nabla\phi_2 & \nabla\phi_2\cdot\nabla\phi_3 \\ \nabla\phi_1\cdot\nabla\phi_3 & \nabla\phi_2\cdot\nabla\phi_3 & \nabla\phi_3\cdot\nabla\phi_3 \end{bmatrix} + \frac{|\tau|}{3} \begin{bmatrix} \beta\cdot\nabla\phi_1 \\ \beta\cdot\nabla\phi_2 \\ \beta\cdot\nabla\phi_3 \end{bmatrix} [1\ 1\ 1]. \tag{16}$$

The first matrix on the right hand side of (16) corresponds to the contribution to A_g from the Laplace operator. This matrix is symmetric, positive semi-definite and has rank two. Its kernel is spanned by the vector $(1\ 1\ 1)^t$, a reflection of (9). The second matrix corresponds to the convection term and has rank one. Note that the column sums of both matrices are zero.

In the general setting, the contribution to the element stiffness matrix from an artificial diffusion term will be a 3×3 matrix with zero row sums and zero column sums (reflecting the fact that $\nabla c = 0$ for a constant c). It is a straightforward calculation to see that this represents five independent constraints on the nine coefficients in such a matrix. A basis for the remaining four dimensional space of element diffusion matrices is given by

$$\frac{1}{|\tau|}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix} = \frac{l_1^2}{|\tau|}\begin{bmatrix} \nabla\phi_1^t \\ \nabla\phi_2^t \\ \nabla\phi_3^t \end{bmatrix} t_1 t_1^t [\nabla\phi_1\ \nabla\phi_2\ \nabla\phi_3] \tag{17}$$

$$\frac{1}{|\tau|}\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} = \frac{l_2^2}{|\tau|}\begin{bmatrix} \nabla\phi_1^t \\ \nabla\phi_2^t \\ \nabla\phi_3^t \end{bmatrix} t_2 t_2^t [\nabla\phi_1\ \nabla\phi_2\ \nabla\phi_3] \tag{18}$$

$$\frac{1}{|\tau|}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \frac{l_3^2}{|\tau|}\begin{bmatrix} \nabla\phi_1^t \\ \nabla\phi_2^t \\ \nabla\phi_3^t \end{bmatrix} t_3 t_3^t [\nabla\phi_1\ \nabla\phi_2\ \nabla\phi_3] \tag{19}$$

$$\frac{1}{|\tau|}\begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \nabla\phi_1^t \\ \nabla\phi_2^t \\ \nabla\phi_3^t \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} [\nabla\phi_1\ \nabla\phi_2\ \nabla\phi_3]. \tag{20}$$

The 2×2 diffusion matrices

$$\hat{\rho}_i = \frac{l_i^2}{|\tau|} t_i t_i^t \tag{21}$$

for $1 \leqq i \leqq 3$, are symmetric, rank one diffusion operators which can naturally be associated with the three edges of τ. The skew symmetric operator

$$\hat{\rho}_s = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{22}$$

corresponds to a rotation by $\pi/2$.

If D is a 2×2 diffusion matrix, then we may expand D in terms of this basis as

$$D = \alpha_s \hat{\rho}_s + \sum_{i=1}^{3} \alpha_i \hat{\rho}_i \tag{23}$$

where

$$\alpha_1 = -|\tau| \nabla \phi_2 \cdot \left(\frac{D + D^t}{2} \right) \nabla \phi_3$$

(cyclically for α_2 and α_3), and

$$\alpha_s \hat{\rho}_s = \frac{D - D^t}{2} .$$

These coefficients can be computed directly using (6)–(15).

As an example, the diffusion operator corresponding to the Laplace operator $-\Delta$ is the 2×2 identity matrix, which can be decomposed as

$$I_{2 \times 2} = \sum_{i=1}^{3} L_i \hat{\rho}_i \tag{24}$$

where

$$\begin{aligned} L_1 &= -|\tau| \nabla \phi_2 \cdot \nabla \phi_3 \\ &= \frac{s_1}{l_1} . \end{aligned} \tag{25}$$

The scalars L_2 and L_3 are defined cyclically.

3 The streamline diffusion method

The streamline diffusion is one of the more widely used upwinding schemes in the finite element arena. Since derivations of the method are widely available in the literature [6, 5, 8], we will merely summarize the method within the current framework.

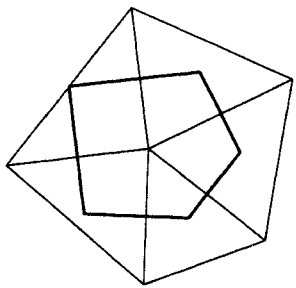

Fig. 2. The box b_i

For the streamline diffusion method, the element stiffness matrix is

$$A_s = A_g + \frac{C|\tau| h_\tau}{|\beta|} \begin{bmatrix} \beta\cdot\nabla\phi_1 \\ \beta\cdot\nabla\phi_2 \\ \beta\cdot\nabla\phi_3 \end{bmatrix} [\beta\cdot\nabla\phi_1 \ \beta\cdot\nabla\phi_2 \ \beta\cdot\nabla\phi_3] \tag{26}$$

where C is a positive constant.

The artificial diffusion term is a symmetric, positive semidefinite matrix of rank one, corresponding to the diffusion term

$$\rho_s = \frac{C}{|\beta|} \beta\beta^t. \tag{27}$$

This rank one matrix adds artificial diffusion in the streamline direction (in the direction of β).

In analogy with (24), the diffusion may be expanded in terms of only the edge diffusion matrices $\hat{\rho}_i$, $1 \leqq i \leqq 3$ as

$$\rho_s = \sum_{i=1}^{3} \alpha_i \hat{\rho}_i,$$

where

$$\alpha_1 = -\frac{C|\tau|}{|\beta|} \beta\cdot\nabla\phi_2 \ \beta\cdot\nabla\phi_3$$

and α_2 and α_3 are defined cyclically.

Upwinding in the crosswind direction involves contributions perpendicular to the streamline direction. These terms are also symmetric and therefore involve only the edge diffusion operators. Thus both the streamline and the crosswind upwinding terms do not involve the skew symmetric operator given by (22).

4 The box method

The box method is formally derived as a finite volume approximation of (1). Assume, for the moment, that $\mathcal{T}$ is such that all triangles have interior angles that are not obtuse. This is nonessential to the definition, but will simplify our initial derivation. Indeed, once the box method has been cast in the form (4), such a restriction will obviously not be required. In any event, for each vertex v_i, we can associate a box b_i, generated by the perpendicular bisectors of the triangle edges incident on that vertex, as illustrated in Fig. 2 (although we could allow a more general definition of boxes, as in [1]).

A given triangle τ contains parts of three boxes; thus one can easily develop the concept of an element stiffness matrix for the box method. This matrix will contain the contributions to the global matrix arising from integrals on the portions of box boundaries lying within τ. See [1] for a complete discussion of this point with respect to the Laplace operator.

We now integrate Eq. (1) over the box b_i, and then apply the divergence theorem to get

$$-\int_{\partial b_i} (\nabla u+\beta u)\cdot n\,\mathrm{d}s=0 \tag{28}$$

where n is the outward normal for the box b_i, defined edgewise.

Let η_i be the index set of vertices in $\mathscr{T}$ connected via a triangle edge to vertex v_i. Then (28) is approximated by

$$\sum_{j\in\eta_i}\left\{\left(\frac{u_i-u_j}{l_{ij}}\right)s_{ij}-(\beta\cdot n_{ij})u_k\,s_{ij}\right\}=0, \tag{29}$$

where

$$k=\begin{cases} i & \text{if } \beta\cdot n_{ij}<0 \\ j & \text{if } \beta\cdot n_{ij}\geqq 0 \end{cases}$$

u_i is the approximate solution at vertex v_i, l_{ij} is the length of the triangle edge connecting vertices v_i and v_j, and s_{ij} is the length of the box edge corresponding to the perpendicular bisector of the edge connecting v_i and v_j. The normal directions n_{ij} for the box b_i correspond to (plus or minus) tangent directions for triangle edges.

To simplify our indices, we will write $\beta\cdot n_{ij}u_k$ as

$$-\beta\cdot n_{ij}u_k=\tfrac{1}{2}\{|\beta\cdot n_{ij}|-\beta\cdot n_{ij}\}u_i-\tfrac{1}{2}\{|\beta\cdot n_{ij}|+\beta\cdot n_{ij}\}u_j$$

so then (29) becomes

$$\sum_{j\in\eta_i}\left(1+\frac{l_{ij}|\beta\cdot n_{ij}|}{2}\right)\left(\frac{u_i-u_j}{l_{ij}}\right)s_{ij}-(\beta\cdot n_{ij})\left(\frac{u_i+u_j}{2}\right)s_{ij}=0. \tag{30}$$

It should be noted that the effect of the upwinding is to add a diffusion term to each triangle edge of strength $\frac{1}{2}l_{ij}|\beta\cdot n_{ij}|$.

A straightforward calculation shows that the element stiffness matrix for the box method is given by

$$\begin{aligned} A_b=&\left(1+\frac{l_1|\beta\cdot t_1|}{2}\right)\frac{s_1}{l_1}\begin{bmatrix}0&0&0\\0&1&-1\\0&-1&1\end{bmatrix}+\left(1+\frac{l_2|\beta\cdot t_2|}{2}\right)\frac{s_2}{l_2}\begin{bmatrix}1&0&-1\\0&0&0\\-1&0&1\end{bmatrix}\\ &+\left(1+\frac{l_3|\beta\cdot t_3|}{2}\right)\frac{s_3}{l_3}\begin{bmatrix}1&-1&0\\-1&1&0\\0&0&0\end{bmatrix}+\frac{s_1\beta\cdot t_1}{2}\begin{bmatrix}0&0&0\\0&-1&-1\\0&1&1\end{bmatrix}\\ &+\frac{s_2\beta\cdot t_2}{2}\begin{bmatrix}1&0&1\\0&0&0\\-1&0&-1\end{bmatrix}+\frac{s_3\beta\cdot t_3}{2}\begin{bmatrix}-1&-1&0\\1&1&0\\0&0&0\end{bmatrix}. \end{aligned} \tag{31}$$

The element stiffness matrix has zero column sums, with nonnegative diagonal and nonpositive off diagonal entries, if we assume no obtuse angles for each element. For elements with vertices on the boundary, the rows and columns corresponding to Dirichlet vertices are ignored in computing the global stiffness matrix. Thus, the global stiffness matrix will be an irreducible, diagonally dominant M-matrix with respect to its columns. This leads to a number of desirable properties, including a discrete maximum principle associated with the columns. We remark that the assumption of no obtuse angles is necessary for the condition of nonnegative diagonal and nonpositive off diagonal entries to hold element by element. It is not a necessary condition (but certainly sufficient) for the global stiffness matrix to inherit these properties [10].

The first three terms on the right hand side of (31) correspond to the Laplace term and the upwinding, which can be written as

$$|\tau|\begin{bmatrix} \nabla\phi_1^t \\ \nabla\phi_2^t \\ \nabla\phi_3^t \end{bmatrix}\left\{\sum_{i=1}^{3} L_i\hat{\rho}_i+\frac{L_i l_i|\beta\cdot t_i|\hat{\rho}_i}{2}\right\}[\nabla\phi_1\ \nabla\phi_2\ \nabla\phi_3]$$

where we have used (17)–(20), and (24)–(25).

The last three terms of (31) correspond to the centered difference approximation to the convective term by the finite volume method. To analyze these terms, we begin by defining

$$\beta_i=\frac{1}{|\tau|}l_i s_i t_i\cdot\beta \tag{32}$$
$$=\frac{l_i^2}{|\tau|}L_i t_i\cdot\beta.$$

With this definition, we have, from (24) and (25)

$$\beta=\sum_{i=1}^{3}\beta_i t_i. \tag{33}$$

This decomposes β into components lying along the tangent directions of each edge of τ. Using (11), we next observe that

$$\begin{bmatrix} -1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}=(e_2-e_1)(e_1+e_2)^t$$
$$=\begin{bmatrix} \nabla\phi_1^t \\ \nabla\phi_2^t \\ \nabla\phi_3^t \end{bmatrix} l_3 t_3(e_1+e_2)^t,$$

where e_i is the i-th column of the 3×3 identity matrix.

Thus, the last three terms of (31) can be written as

$$\frac{|\tau|}{2}\begin{bmatrix}\nabla\phi_1^t\\ \nabla\phi_2^t\\ \nabla\phi_3^t\end{bmatrix}\{\beta_1 t_1(e_2+e_3)^t+\beta_2 t_2(e_3+e_1)^t+\beta_3 t_3(e_1+e_2)^t\}.$$

Our next task is to compute the form of the artificial diffusion associated with the box method, and then to recast the box method in the form (4). We begin by finding the matrix corresponding to the upwinding (relative to the standard Galerkin method) given by $A_b - A_g$. To simplify the resulting expressions, we will need

$$\left(\frac{e_2+e_3}{2}\right)-\left(\frac{e_1+e_2+e_3}{3}\right)=\frac{1}{6}(e_2-e_1+e_3-e_1) \tag{34}$$

$$=\frac{1}{6}\begin{bmatrix}\nabla\phi_1^t\\ \nabla\phi_2^t\\ \nabla\phi_3^t\end{bmatrix}(l_3 t_3-l_2 t_2)$$

$$\equiv\frac{1}{2}\begin{bmatrix}\nabla\phi_1^t\\ \nabla\phi_2^t\\ \nabla\phi_3^t\end{bmatrix}d_1.$$

The vectors d_2 and d_3 are defined cyclically. Thus using (32)–(34), as well as (24), we have

$$A_b-A_g=\frac{|\tau|}{2}\begin{bmatrix}\nabla\phi_1^t\\ \nabla\phi_2^t\\ \nabla\phi_3^t\end{bmatrix}\left\{\sum_{i=1}^{3} l_i|\beta_i|\,t_i t_i^t+\beta_i t_i d_i^t\right\}[\nabla\phi_1\ \nabla\phi_2\ \nabla\phi_3]. \tag{35}$$

Note that $\beta_i = O(|\beta|)$ and $d_i = O(h_\tau)$. Thus

$$h_\tau \rho_b=\tfrac{1}{2}\sum_{i=1}^{3} l_i|\beta_i|\,t_i t_i^t+\beta_i t_i d_i^t \tag{36}$$

is the artificial diffusion term (4) for the box method. Note that there are two types of terms on the right hand side of (36). The first type comes from upwinding along a single edge; these terms contribute symmetric, positive semidefinite artificial diffusion terms to $h_\tau \rho_b$. The second set of terms arise from the differences in approximating the convection term using centered differences; the box method considers only approximations along each edge, while the standard Galerkin method develops approximations within the triangle as a whole. This generally contributes a nonsymmetric artificial diffusion term to the overall upwinding.

Having defined the form of the artificial diffusion, we can now interpret the box method as a finite element method, which remains well defined even when some elements have obtuse angles, and when β is no longer assumed to be constant on each element.

5 The Scharfetter-Gummel method

A second finite volume scheme, similar to the box method of Sect. 4, but making explicit use of the assumption that $\beta = \nabla \psi$, is the Scharfetter-Gummel discretization. Originally proposed for the one dimensional discretization of the current continuity equation in the semiconductor device model, it has been generalized to two dimensions, and is a widely used discretization in contemporary device simulators [2]. The Scharfetter-Gummel discretization is an exponential upwinding scheme which will produce the exact values at the vertices for a one dimensional problem in the special case where ψ is linear.

We define the Bernoulli function $\mathscr{B}(x)$ by

$$\mathscr{B}(x) = \frac{x}{e^x - 1}. \tag{37}$$

We will use the identity

$$\mathscr{B}(-x) = \mathscr{B}(x) + x$$

in the forms

$$\mathscr{B}(x) = \mathscr{C}(x) - \frac{x}{2}$$

$$\mathscr{B}(-x) = \mathscr{C}(x) + \frac{x}{2},$$

where

$$\mathscr{C}(x) = \frac{\mathscr{B}(x) + \mathscr{B}(-x)}{2} = \mathscr{B}(|x|) + \frac{|x|}{2}. \tag{38}$$

Along the triangle edge connecting vertices v_1 and v_2 in element τ, the flux term

$$-(\nabla u + \beta u) \cdot n = -e^{-\psi} \nabla (e^{\psi} u) \cdot n$$

is approximated along the box boundary by

$$e^{-\tilde{\psi}} \left(\frac{e^{\psi_1} u_1 - e^{\psi_2} u_2}{l_3} \right), \tag{39}$$

where $\psi_i \equiv \psi(v_i)$. The value of $\tilde{\psi}$ is given by [2]

$$e^{-\tilde{\psi}} = \left(\frac{\int_{\psi_2}^{\psi_1} e^{\psi} \, d\psi}{\psi_1 - \psi_2} \right)^{-1} = \frac{\psi_1 - \psi_2}{e^{\psi_1} - e^{\psi_2}}.$$

This allows us to write (39) as

$$\frac{u_1 \mathscr{B}(\psi_1 - \psi_2) - u_2 \mathscr{B}(\psi_2 - \psi_1)}{l_3}. \tag{40}$$

Assuming that ψ is linear, we have

$$\psi_2 - \psi_1 = \beta \cdot t_3 \, l_3.$$

Setting $\mathscr{C}_3 \equiv \mathscr{C}(\beta \cdot t_3 \, l_3)$, our flux approximation becomes

$$\mathscr{C}_3\left(\frac{u_1-u_2}{l_3}\right)+\beta \cdot t_3\left(\frac{u_1+u_2}{2}\right). \tag{41}$$

Notice that the second term in (41) is identical to the corresponding term for the box method.

Using this approximation to the flux, the element stiffness matrix for the Scharfetter-Gummel discretization can be found in a fashion, completely analogous to (28)–(31), to be

$$\begin{aligned} A_{sg} = |\tau| \begin{bmatrix} \nabla \phi_1^t \\ \nabla \phi_2^t \\ \nabla \phi_3^t \end{bmatrix} \left\{ \sum_{i=1}^{3} L_i \mathscr{C}_i \tilde{\rho}_i \right\} [\nabla \phi_1 \; \nabla \phi_2 \; \nabla \phi_3] \\ + \frac{|\tau|}{2} \begin{bmatrix} \nabla \phi_1^t \\ \nabla \phi_2^t \\ \nabla \phi_3^t \end{bmatrix} \{ \beta_1 \, t_1 (e_2+e_3)^t + \beta_2 \, t_2 (e_3+e_1)^t + \beta_3 \, t_3 (e_1+e_2)^t \}. \end{aligned} \tag{42}$$

Similarly, the upwinding operator ρ_{sg} can be found, by forming $A_{sg}-A_g$, to be

$$h_\tau \rho_{sg} = \sum_{i=1}^{3} L_i (\mathscr{C}_i - 1) \hat{\rho}_i + \beta_i t_i d_i^t, \tag{43}$$

where the d_i are defined as in (34).

We point out here that the term $\mathscr{C}_i - 1$ is formally of order $\mathcal{O}(h_\tau^2 |\beta|^2)$ as $h_\tau \beta \to 0$, whereas the corresponding term in the standard box method is $\mathcal{O}(h_\tau |\beta|)$. Thus, while in some regimes we can expect the two discretizations to behave quite similarly, there can be cases where are significant differences.

6 Divergence-free upwinding

Our new discretization [3] is defined in terms of a single element τ and the corresponding element stiffness matrix A_d. Let the *current* $\mathscr{J}$ be defined by

$$\mathscr{J} = \nabla u + \beta u \tag{44}$$

so that (2) becomes

$$\int_\Omega \mathscr{J} \cdot \nabla \phi \, \mathrm{d}x \, \mathrm{d}y = 0 \tag{45}$$

for all $\phi \in \mathscr{H}_0$.

Since $\beta = \nabla \psi$, we may write (44) as

$$\mathscr{J} = e^{-\psi} \nabla (e^{\psi} u). \tag{46}$$

For the case $\psi \in \mathscr{S}_h$, we can replace $e^{\pm\psi}$ by $e^{\pm \beta \cdot v}$, where $v^t = (x \; y)$.

For our approximation, we seek a discrete current $\mathcal{J}_h$ in the form

$$\mathcal{J}_h = e^{-\psi} \nabla (e^{\psi} \eta) \tag{47}$$
$$= \nabla \eta + \beta \eta,$$

where η is a linear polynomial in τ. Over all of Ω, η will be a discontinuous piecewise linear polynomial.

The consistency of our approximation is determined by the *edge conditions*

$$\int_{v_i}^{v_j} e^{\psi} \mathcal{J} \cdot \mathrm{d}s \equiv \int_{v_i}^{v_j} e^{\psi} \mathcal{J}_h \cdot \mathrm{d}s \tag{48}$$

where v_i and v_j are two vertices of τ. Since the integrations can be carried out exactly, we may write (48) as

$$e^{\psi(v_j)} u(v_j) - e^{\psi(v_i)} u(v_i) = e^{\psi(v_j)} \eta(v_j) - e^{\psi(v_i)} \eta(v_i). \tag{49}$$

Although there are three edge conditions, only two represent independent constraints on η. In any event, the edge conditions imply that

$$\eta = u_h + \alpha \mathcal{I}(e^{-\psi}), \tag{50}$$

where u_h is the finite element solution, α is a scalar, and $\mathcal{I}(e^{-\psi})$ is the linear polynomial interpolating $e^{-\psi}$ at the vertices of τ. Note that since $u_h \in \mathcal{S}_h$, the discontinuities in η can arise only from α having different values in different elements.

The scalar α, and the stability of the discretization, is determined by the *divergence condition*

$$\nabla \cdot \mathcal{J}_h = 0 \tag{51}$$

on τ, which implies, for $\psi \in \mathcal{S}_h$,

$$\alpha = -\frac{\beta \cdot \nabla u_h}{\beta \cdot \nabla \mathcal{I}(e^{-\psi})}. \tag{52}$$

Setting $z = \mathcal{I}(e^{-\psi})$, we have

$$\mathcal{J}_h \cdot \nabla \phi = (\nabla \eta + \beta \eta) \cdot \nabla \phi \tag{53}$$
$$= (\nabla u_h + \beta u_h) \cdot \nabla \phi - \frac{\beta \cdot \nabla u_h}{\beta \cdot \nabla z} (\nabla z + \beta z) \nabla \phi$$
$$= (\nabla u_h + \beta u_h) \cdot \nabla \phi + \nabla u_h \cdot (\beta d^t) \nabla \phi,$$

where

$$d = \frac{\nabla z + \beta z}{-\beta \cdot \nabla z}. \tag{54}$$

The first term on the right hand side of the last line in (53) corresponds to the standard Galerkin method; thus the artificial diffusion for the divergence-free upwinding scheme is

$$h_\tau \rho_d = \beta d^t \tag{55}$$

which is a generally nonsymmetric, rank one diffusion matrix.

By noting that

$$\nabla e^{-\psi} + \beta e^{-\psi} = 0$$

we can set

$$\begin{aligned} \varepsilon &= e^{-\psi} - \mathcal{I}(e^{-\psi}) \\ &= e^{-\psi} - z \end{aligned}$$

and write (54) as

$$d = -\frac{\nabla \varepsilon + \beta \varepsilon}{\beta \cdot \beta \varepsilon^{-\psi} + \beta \cdot \nabla \varepsilon}. \tag{56}$$

Since ε is the interpolation error for linear interpolation of $e^{-\psi}$, we can see (formally) that $|d| = \mathcal{O}(h_\tau)$.

An interesting special case occurs whenever β is perpendicular to one of the edges of τ. Then d and β are parallel vectors, and the divergence-free upwinding scheme is similar to the streamline diffusion method, in terms of the added artificial diffusion. However, unlike the streamline diffusion method, there is no constant to be adjusted; in effect, the constant was chosen to satisfy the divergence condition.

For the case $\psi \in \mathcal{S}_h$, the element stiffness matrix for the divergence-free upwinding scheme is given by

$$A_d = A_g + |\tau| \begin{bmatrix} \nabla \phi_1 \cdot d \\ \nabla \phi_2 \cdot d \\ \nabla \phi_3 \cdot d \end{bmatrix} [\beta \cdot \nabla \phi_1 \ \beta \cdot \nabla \phi_2 \ \beta \cdot \nabla \phi_3]. \tag{57}$$

An important consideration for the divergence-free upwinding scheme is the question of whether it is always well defined. In particular, we must examine conditions under which $\beta \cdot \nabla z = 0$, since this term is in the denominator of (54). We can begin by observing that

$$\begin{aligned} -\beta \cdot \nabla z &= \beta \cdot \beta e^{-\psi} + \beta \cdot \nabla \varepsilon \\ &= |\beta|^2 e^{-\psi} + O(|\beta|^2 h_\tau e^{-\psi}) \\ &> 0 \quad \text{as} \quad h_\tau \to 0 \end{aligned}$$

so that the method is certainly well defined for h sufficiently small. On the other hand, it is possible on a coarse mesh, with proper element geometry and a certain element orientation with respect to β, that $-\beta \cdot \nabla z \leqq 0$.

To see how this can occur, assume for the moment that our element τ has vertices $v_1^t=(0\,0)$, $v_2^t=(1\,0)$, $v_3^t=(\bar{x}\,\bar{y})$, and that $\psi\in\mathcal{S}_h$. The Jacobian matrix J for this element is

$$J=\begin{bmatrix}1 & \bar{x}\\ 0 & \bar{y}\end{bmatrix}$$

$$J^{-1}=\frac{1}{\bar{y}}\begin{bmatrix}\bar{y} & -\bar{x}\\ 0 & 1\end{bmatrix}$$

$$J^{-1}J^{-t}=\frac{1}{\bar{y}^2}\begin{bmatrix}\bar{x}^2+\bar{y}^2 & -\bar{x}\\ -\bar{x} & 1\end{bmatrix}.$$

Let

$$J^t\beta=\begin{bmatrix}q_2\\ q_3\end{bmatrix}.$$

Then

$$z=\phi_1+e^{-q_2}\phi_2+e^{-q_3}\phi_3$$

and

$$\nabla z=(e^{-q_2}-1)\nabla\phi_2+(e^{-q_3}-1)\nabla\phi_3.$$

Without loss of generality, assume that $q_2\geqq q_3\geqq 0$ and $q_2>0$. Then let

$$r=\frac{q_3}{q_2}$$

$$s=\frac{e^{-q_3}-1}{e^{-q_2}-1}.$$

Clearly

$$0\leqq r\leqq s\leqq 1$$

and

$$-\nabla z\cdot\beta=-\frac{q_2(e^{-q_2}-1)}{\bar{y}^2}\,[1\;r]\begin{bmatrix}\bar{x}^2+\bar{y}^2 & -\bar{x}\\ -\bar{x} & 1\end{bmatrix}\begin{bmatrix}1\\ s\end{bmatrix}.$$

The condition $-\nabla z\cdot\beta=0$ implies

$$\bar{x}^2+\bar{y}^2-\bar{x}(r+s)+rs=0$$

or

$$\bar{y}^2+\left(\bar{x}-\frac{r+s}{2}\right)^2=\left(\frac{s-r}{2}\right)^2, \tag{58}$$

Equation (58) is the equation of a circle with center $((r+s)/2, 0)$ and radius $(s-r)/2$.

The properties of this upwinding scheme have a nice geometrical interpretation as illustrated in Fig. 3. The outer circle C_1 separates acute from obtuse triangles. All triangles with $(\bar{x}, \bar{y})$ lying outside this circle are acute, those with $(\bar{x}, \bar{y})$ inside are obtuse, and those with $(\bar{x}, \bar{y})$ lying on C_1 are right triangles.

The inner circle C_0, corresponding to (58), always lies inside the circle C_1, and separates triangles of positive and negative $-\beta\cdot\nabla z$. Triangles with $(\bar{x}, \bar{y})$

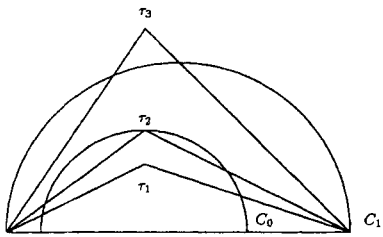

Fig. 3. Geometrical interpretation of upwinding term

lying outside this circle have $-\beta \cdot \nabla z > 0$. Clearly, $-\nabla z \cdot \beta \leqq 0$ requires τ to have an obtuse angle.

For triangles with $(\bar{x}, \bar{y})$ lying on this circle, $\beta \cdot \nabla z = 0$, and the discretization is not defined. The chance of this condition being met in practice is very small. Indeed, we don't even check for this in our code, since roundoff error will almost certainly produce nonzero values of $\beta \cdot \nabla z$.

On coarse meshes containing many badly shaped elements, it may be possible to have triangles with $(\bar{x}, \bar{y})$ lying inside this circle, in which case, $-\beta \cdot \nabla z < 0$. When this occurs, it is analogous to *subtracting* a one dimensional artificial diffusion from the system, which seems rather counter intuitive (and dangerous). In particular, the eigenvalues of the 2×2 diffusion matrix $I + \beta d^t$ are

$$\lambda = 1$$

and

$$\lambda = \frac{\beta \cdot \beta z}{-\beta \cdot \nabla z}$$

so that the overall diffusion term ceases to be elliptic whenever $-\beta \cdot \nabla z < 0$.

If we scale τ to be an element with the same geometry but with diameter h, we note that q_2 and q_3 will scale to be of size $|\beta| h$. Thus, as $h \to 0$, $r \to s$ and the (relative) radius of the circle C_0 tends to zero, which is consistent with our earlier remarks. Also, $r = s$ if $q_3 = 0$; this implies that β is perpendicular to one side of τ. In general, if β is perpendicular to any side of τ, then $-\nabla z \cdot \beta \neq 0$, since then ∇z is in the direction β as in the streamline diffusion method.

Given the above comments, one might naturally approach this method with a great deal of skepticism with respect to its usefulness in general and its stability in particular (we certainly did). At present, we do not have any a priori error estimates for the method, except in the case when it reduces to the streamline diffusion method and existing estimates for that method apply. Nevertheless, the method is extremely stable, even under unfavorable geometric conditions. This stability comes from the divergence condition, as can be seen from the following line of reasoning. Let ϕ_i be the piecewise linear nodal basis function associated with vertex v_i in the triangulation. Then, using integration by parts, element by element, we have from (4)

$$\int_\Omega \mathscr{J}_h \cdot \nabla \phi_i \, dx = \sum_{e_{ij}} \int_{e_{ij}} \{\mathscr{J}_h \cdot n_{ij}\} \phi_i \, ds = 0,$$

where e_{ij} is the triangle edge connecting vertices v_i and v_j and $\{\mathscr{J}_h \cdot n_{ij}\}$ is the jump in the normal component of $\mathscr{J}_h$ across e_{ij}. By simple geometry, it seems

clear that in order to have a massive overshoot or undershoot (a "spike") at v_i, the sum of the normal components of these jumps must be correspondingly large in magnitude, a circumstance which is prohibited by the divergence condition.

In effect, the divergence condition prevents the creation of any numerical sources or sinks within element interiors. The edge conditions guarantee good approximation along element edges, in particular at the vertices. The situation is entirely analogous to the finite element approximation of the Laplacian using piecewise linear elements; there $\Delta u = 0$ within each element and it is the jumps in the normal components of ∇u across the triangle edges that support the approximation. Thus we can expect, at least with hindsight, that this method will provide a stable and accurate approximation to (2).

We end this section by noting that this method and its derivation remain well defined for three dimensional meshes based on tetrahedral elements. Indeed, it was our desire to have an upwinding procedure for tetrahedral meshes that remains stable even in the presence of unfavorable element geometries, which motivated our current work.

7 Summary

A uniform framework is developed for the study of general upwinding schemes. The standard finite element weak Galerkin discretization is chosen as the reference. Differences between other discretization schemes and the weak Galerkin form are written as artificial diffusion terms. These artificial diffusion terms are spanned by a four dimensional space of element diffusion matrices. Three basis matrices are symmetric, rank one diffusion operators which can naturally be associated with the edges of the triangle. The fourth basis matrix is skew symmetric and is associated with a rotation by $\pi/2$.

The streamline diffusion method is one of the more widely used upwinding schemes in the finite element arena. Both the streamline and the crosswind upwinding terms are symmetric, positive semidefinite matrices of rank one and involve only the edge diffusions operators.

Two finite volume discretizations, the box method and the Scharfetter-Gummel method, are then analyzed. Finite volume methods involve only approximations along each triangle edge, while the standard Galerkin method uses approximations within the triangle as a whole. Discretizations of convection diffusion problems give rise to two types of contributions to the element stiffness matrices. The first type corresponds to the upwinding terms, which contribute symmetric, positive semidefinite artificial edge diffusion terms. The second type arises from the centered difference approximation of the convective term. When viewed as a finite element method, these terms contribute nonsymmetric artificial diffusion upwinding terms. While finite volume discretizations may always be written as upwinded Galerkin methods, the converse does not appear to be true, since the skew symmetric elementary diffusion operator seems to have no analogue in the standard finite volume framework.

Finally, the divergence-free upwinding scheme is analyzed. In general, the artificial diffusion introduced by this method leads to both symmetric and nonsymmetric diffusion terms. However, whenever the velocity is perpendicular to one of the triangle edges, the streamline diffusion method is recovered. In some

extreme cases, the overall diffusion matrix has both positive and negative eigenvalues. Nevertheless, the method appears to be extremely robust and stable, and remains so even in unfavorable situations where other upwinding schemes fail.

References

1. Bank, R.E., Rose, D.J.: Some error estimates for the box method. SIAM J. Numer. Anal. **24**, 777–787 (1987)
2. Bank, R.E., Rose, D.J., Fichtner, W.: Numerical methods for semiconductor device simulations. IEEE Trans. Electr. Dev., ED-**30**, 1031–1041 (1983)
3. Bürgler, J.F., Bank, R.E., Fichtner, W., Smith, R.K.: A new discretization for the semiconductor continuity equations. IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems **8**, 479–489 (1989)
4. Caussignac, P., Touzani, R.: Linear conforming and nonconforming upwind finite elements for the convection-diffusion equation. IMA J. Numer. Anal. **8**, 85–103 (1988)
5. Hughes, T.J.R., Brooks, A.: Streamline-upwind Petrov-Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier-Stokes equations. Comput. Methods Appl. Mech. Eng. **32**, 199–259 (1982)
6. Hughes, T.J.R., Brooks, A.: A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: Application to the streamline upwind procedure. In: Gallagher, R.H., Norrie, D.H., Oden, J.T., Zienkiewicz, O.C. (eds.) Finite elements in fluids, Vol. 4, pp 47–66 New York: Wiley 1984
7. Ikeda, T.: Maximum principle in finite element models for convection-diffusion phenomena. North-Holland Mathematics Studies, Vol. 76, (Lecture Notes in Numerical and Applied Analysis, Vol. 4). Amsterdam: North-Holland 1983
8. Johnson, C., Navert, U., Pitkaranta, J.: Finite element methods for linear hyperbolic problems. Comput. Methods Appl. Mech. Eng. **45**, 285–312 (1984)
9. Ohmori, K., Ushijima, T.: A technique of upstream type applied to a linear conforming finite element approximation of convection-diffusion equations. RAIRO Anal. Numer. **18**, 309–332 (1984)
10. Strang, G., Fix, G.: An Analysis of the finite element method. Prentice-Hall, Englewood Cliffs, New Jersey 1973