VLSI DESIGN
2000, Vol. 10, No. 4, pp. 485–529
Reprints available directly from the publisher
Photocopying permitted by license only

Discretization of Anisotropic Convection-diffusion Equations, Convective *M*-matrices and their Iterative Solution

DONALD J. ROSE[a,*], HAI SHAO[a] and CRAIG S. HENRIQUEZ[b]

[a]*Department of Computer Science;* [b]*Department of Biomedical Engineering, Duke University, Durham, NC*

(Received 16 December 1998; In final form 14 December 1999)

We derive the constant-j box method discretization for the convection-diffusion equation, $\nabla j = f$, with $j = -\alpha\nabla u + \beta u$. In two dimensions, α is a 2×2 symmetric, positive definite tensor field and β is a two-dimensional vector field. This derivation generalizes the well-known Scharfetter–Gummel discretization of the continuity equations in semiconductor device simulation. We define the anisotropic Delaunay condition and show that under this condition and appropriate evaluations of α and β, the stiffness matrix, M, of the discretization is a convective M-matrix. We then examine classical iterative splittings of M and show that convection (even convection dominance) does not degrade the rate of convergence of such iterations relative to the purely diffusive ($\beta = 0$) problem under certain conditions.

Keywords: Anisotropic convection-diffusion equation, constant-j box method discretization, anisotropic Delaunay condition, convective M-matrix, curl-free condition, convective iteration

1. INTRODUCTION

We consider the two-dimensional convection-diffusion equation

$$\nabla \cdot j = f, \quad j = -\alpha\nabla u + \beta u \qquad (1.1)$$

with standard boundary conditions. Here α, β, and f are spatially dependent on the defining domain Ω; the diffusion coefficient α is a two by two symmetric and positive-definite tensor field, and the convection coefficient β is a two dimensional vector field. Equation (1.1) arises in semiconductor device simulation as the (electron and hole) continuity equations and allows the mobility and diffusivity coefficient functions a greater generality.

In §2 we derive a box method discretization for Eq. (1.1) on a meshed domain Ω_M which is a union of triangles. We use the *constant-j assumption* on triangle *edge-pairs* in order to resolve the vector

*Corresponding author. Tel.: (919) 660-6510, e-mail: djr@cs.duke.edu

current density $j=[j_1, j_2]^\top$. The use of edge-pairs is essential when diffusivity of the media is anisotropic. The edge-pair approach allows us to approximately integrate Eq. (1.1) on a "box" *via* the divergence theorem in the standard manner. The various parts of the approximate integration give a linear relation between a nodal value u_i of the unknown function u and its nodal values at the various triangle neighbors and ultimately the completely assembled linear system, in particular, the stiffness matrix M.

The use of the constant-j assumption dates from the seminal paper [29] by Scharfetter and Gummel (SG), and is well-appreciated in the semiconductor simulation community. Bank, Rose and Fichtner [5] show in §III.C, Eq. (55), that a essential ingredient of the SG constant-j derivation can be interpreted as an edge evaluation scheme for e^{ψ}. This property also allows a finite element version of the SG discretization (see [38, 16, 2, 3]). In a box method (or finite volume) scheme, it becomes natural to apply the constant-j idea directly generalizing the discussion in [5], §III.D. As mentioned above, in order to resolve both components of j, we are led to apply the constant-j assumption on *edge-pairs*. In the case where $\alpha = aI_{2\times 2}$ is a scalar, one could directly apply the constant-j idea by assuming the *projection* of j on the edge is constant. If the vector field j is needed, however, this projection assumption is insufficient. Interestingly both the edge-pair constant-j (SG) discretization and the finite element (SG) discretization give exactly the same stiffness matrix M for comparable evaluation schemes for α and β. We feel that our approach also has value when meshes are no longer triangular or tetrahedral in 3D, and we will suggest a generalization in §6.

Our box method seems peculiar, at first, since each triangle edge has two currents which may not be conservative ($I_{j,l}+I_{l,j}\neq 0$, in general, for edge (j,l)). This is due to the discretization paradigm which only requires current conservation on the whole box, and hence allows (small) deviations locally. However, there are conditions on the triangular mesh and evaluation schemes for α and β which allow the then conservative currents to be viewed edgewise. These conditions are also needed for matrix to be an *M-matrix*.

In §4 we examine, in some detail, the M-matrix nature of the system matrix M. We show, for example, that any irreducible column diagonally dominant M-matrix can be interpreted as a *generalized resistive network* where the current on edge (j,l) is defined as $I_{jl}=au_j-bu_l$, with a and b both being positive. Clearly when $a=b$, the general resistors become ordinary resistors. The quantity $c=(a-b)/2$ is identified as the discrete version of *convection*. Furthermore, the discrete analogies of the curl and divergence of the (scaled) convection vector field can represented in terms of the discrete edge conductance and edge convection as well as the incidence and cycle matrices associated with the convection-directed graph.

Our analysis in §5 shows that under certain conditions on the *scaled* convection vector field $b=\alpha^{-1}\beta$ and the vector field β itself, standard classical iterative methods converge faster when there is convection ($\beta\neq 0$) than when there is not ($\beta=0$) in Eq. (1.1). Of course, as convection begins to dominate, it becomes necessary to refine the mesh to obtain comparable accuracies of the discrete solution to the corresponding solution of (1.1). What is perhaps surprising is that, for fixed accuracy, the necessary number of iterations is *bounded* (or even slightly decreases) as convection increases.

It is historically interesting that our iterative analysis in Section 5 could be regarded as an extension of the Kahan–Varga theory of successive over-relaxation (SOR); see Varga [36], §4.4, and Kahan [22, 23]. In Kahan–Varga theory one gives up the consistent ordering assumption (Property A) and replaces it with an assumption which, in our context, follows from the discrete curl-free condition (3.13). While consistent orderings are still best in this theory, when they exist, all orderings give boundedly related rates of convergence (for Jacobi, Gauss–Seidel and SOR).

We show how these convergence rates change with convection and relate the convergence rate of a convective problem directly to the same problem with convection identically zero. The discrete curl-free condition also implies that there are orderings which put the predominance of the discrete convection in the upper or lower triangle of the stiffness matrix, but a detailed look at our analysis shows that our bounds do not depend on these convection-related orderings; instead, any ordering preference for a convective problem is simply inherited from the zero-convection (diffusive) problem. Generalizations to more sophisticated iterative methods are clearly possible.

Finally, in §6 we present briefly another applied problem from computational physiology where the anisotropy of diffusivity (2×2 nature of α) is important, and make some additional concluding remarks.

2. THE CONSTANT-j BOX METHOD

In this section, we present the constant-j box method for discretization of Eq. (1.1) in two spatial dimensions. In two (and higher) dimensions, the main difficulties for discretization are how to handle the domain geometry and how to resolve the anisotropy of diffusion. We use unstructured meshes to in order to handle arbitrary domain geometry in two dimensions. Details of mesh generation and adaptation in two dimensions can be found in [32], Chapter 5. An important feature of our box method in two dimensions is the use of edge-pair constant-j assumption which naturally resolves the anisotropy of diffusion, hence that name "constant-j box method".

In §2.6, we show the equivalence between the box method and a specific finite element method on the anisotropic convection-diffusion equations in two dimensions. Based on this equivalence, we are able to obtain an error estimate for the box method using finite element analysis as in [38].

2.1. Mesh Description

A discretization method for the model problem (1.1) in two spatial dimensions usually requires a mesh, in which the defining domain Ω is decomposed into a number of smaller simple cells. A mesh can be either structured, such as grid graphs, or unstructured, such as triangular meshes. Our box method can be applied to both types of meshes. We will, however, present the box method with the use of unstructured triangular meshes to demonstrate its greater generality. In the rest of the paper, we assume that a mesh is a triangulation over the *defining domain*, Ω, a compact and connected two-dimensional region on which the partial differential Eq. (1.1) is defined. Let $P = \{p_1, p_2, \ldots, p_{|P|}\}$ denote the set of mesh points, $E = \{e_1, e_2, \ldots, e_{|E|}\}$ the set of mesh edges, and $T = \{t_1, t_2, \ldots, t_{|T|}\}$ the set of triangles. More specifically, a point, an edge and a triangle are represented as

$$
\begin{aligned}
&p_i = (x_i, y_i) \in \mathbb{R}^2,\\
&e_j = (u_j, v_j), \quad 1 \le u_j, v_j \le n_p, \quad \text{for edge } \overline{p_{u_j} p_{v_j}},\\
&t_k = (a_k, b_k, c_k), \quad 1 \le a_k, b_k, c_k \le n_p,\\
&\qquad \text{for triangle } \Delta p_{a_k} p_{b_k} p_{c_k},
\end{aligned}
$$

respectively. We require the mesh to be a *conforming triangulation* in a sense that it preserves the boundary and interfaces of the domain Ω; that is, the domain boundary and the interfaces can be recovered by a subset of edges in E. The interfaces in the defining domain are usually used to separate areas with different materials or block flows.

The mesh serves as the base of approximation. The unknown state function u is approximated at mesh points by a vector, u_h, by means of $u(p_i)$ being approximated by u_i, the ith component of u_h. Since the values of u at the Dirichlet boundary points are already known, the approximate vector u_h needs to be defined only on the *reduced set* of the mesh points, P_r, a point set that excludes the

Dirichlet points from P. In special problems where flux blocking interfaces are included, a mesh point on a interface (end points excluded) is duplicated in P_r, with one on either side of the interface, and is treated as a zero-flux Neumann boundary point; (see [32], Appendix I).

2.2. Box Formation

To apply the box method, we need to construct a secondary mesh, a set of boxes, on top of the triangular mesh. We partition the defining domain into a set of non-overlapping boxes, $\{B_1, \ldots, B_{|P|}\}$, each around a mesh point. We tentatively form the boxes by selecting a point, called a *box vertex* in each triangle. We then connect the box vertex with the three mid-points of the triangle edges (thinner lines in Fig. 2.1) to form three *box edges* (thicker lines in Fig. 2.1). Any two of the three box edges form a *boundary patch* (shaded lines in Fig. 2.1) which will become a part of the box boundary that surrounds a vertex (p_i in Fig. 2.1) of the triangle. Typically the circumcenter (intersection of the perpendicular bisectors) or the centroid of the triangle is chosen. And the boxes formed in this way are called *Voronoi boxes* and *barycentric boxes*, respectively. An *interior box*, one that covers an interior mesh point, is bounded by box edges only, while a *boundary box*, one that covers a boundary mesh point, is bounded by both box edges and (two) boundary edges of the defining domain. Hence, the boundary boxes are also called "half boxes". Since only one equation for each unknown variable is needed, only the boxes that correspond to points in the reduced set P_r will be needed for discretization. These boxes will be referred to as the *effective boxes*.

2.3. Local Integral Form

After the boxes formed, we integrate the model Eq. (1.1) on each effective box B_i,

$$\int_{B_i} \nabla \cdot j \, da = \int_{B_i} f \, da \tag{2.1}$$

for $i = 1, \ldots, n$. The divergence theorem in vector calculus (see [24]) implies that

$$\int_{B_i} \nabla \cdot j \, da = \oint_{\partial B_i} \nu^\top j \, ds, \tag{2.2}$$

where ∂B_i is the boundary of box B_i, and ν is the unit outward normal vector of the box boundary. Therefore, we are able to replace the box area integral of the divergence of the current density by the line integral of its normal component along the box boundary, and convert the Eq. (2.1) into

$$\oint_{\partial B_i} \nu^\top j \, ds = \int_{B_i} f \, da. \tag{2.3}$$

The model problem (1.1) is now equivalently converted into a set of integral equations (2.3), or the *local integral forms*, on all effective boxes.

The area integral of the source term on the right-hand-side of Eq. (2.3) is approximated by

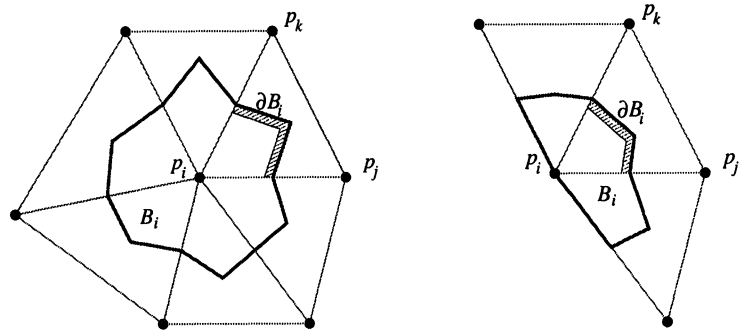

FIGURE 2.1 Box B_i for mesh point p_i as an interior point (left) and as boundary point (right).

applying the following numerical quadrature rule [10] that yields the desired degree of accuracy:

$$\int_{B_i} f\,da \approx \sum_{k\in\Lambda_i\cup\{i\}} w_{ik} f_k, \tag{2.4}$$

where Λ_i is the *index set* of the neighbors (mesh points that are directly connected to p_i) of p_i, w_{ik} is the quadrature weight which determined by the quadrature rule, and f_k is the nodal value of the source term f at p_k. Notice that for higher degree quadrature schemes, values of the integrand at points other than the mesh points (midpoints of triangle edges for example) may be needed. The *consistency condition* requires that the quadrature approximation (2.4) be *exact* when the integrand f is a constant, or

$$\sum_{k\in\Lambda_i\cup i} w_{ik} = \text{area}(B_i). \tag{2.5}$$

Often the value of w_{ik} is taken as

$$w_i = \text{area}(B_i), \tag{2.6}$$

for $i=k$, and 0 otherwise.This procedure is often referred to as the *lumping*, and is equivalent to assuming the integrand of the area integral is box-wise constant. Calculating the box areas in terms of the triangular mesh and the boxes is straightforward. Henceforth, we consider only the left-hand-side of Eq. (2.3).

2.4. Edge-pair Constant-j Assumption

Approximation of the line integral is more complex. It is carried out at two levels. First at the *edge-pair* level, we assume the *scaled current density* $\kappa=\alpha^{-1}j$ constant at a pair of edges, or an *edge-pair*, $((i,k),(i,l))$ for each triangle (i,k,l) incident to the mesh point p_i (see Fig. 2.2). Projecting onto the two edges of the edge-pair, we have

$$\begin{cases} h_{ik}^{\top}(-\nabla u + bu) = h_{ik}^{\top}\kappa_{lik}, \\ h_{il}^{\top}(-\nabla u + bu) = h_{il}^{\top}\kappa_{lik}, \end{cases} \tag{2.7}$$

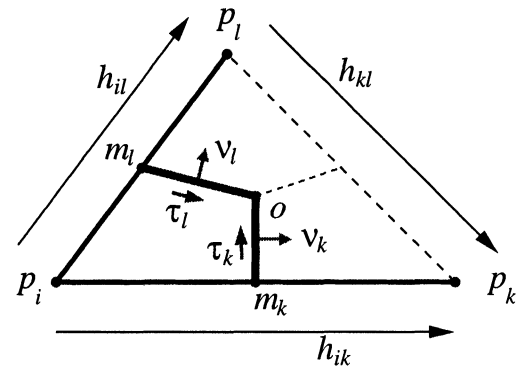

FIGURE 2.2 An edge-pair (h_1,h_2) with the boundary patch $\overline{m_1 o}\cup\overline{o m_2}$ that is contained in it.

where $b=\alpha^{-1}\beta$ is called the *scaled convection*. If we associate with each edge a constant vector approximating the scaled convection b, *e.g.*, b_{ik} for edge (i,k) and b_{il} for edge (i,l), the above set of two ordinary differential equations along the two edges can be solved exactly as

$$\begin{cases} h_{ik}^{\top}\kappa_{lik} = B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k, \\ h_{il}^{\top}\kappa_{lik} = B(-\lambda_{il})u_i - B(\lambda_{il})u_l, \end{cases} \tag{2.8}$$

where

$$\lambda_{ik} = h_{ik}^{\top}b_{ik} = h_{ik}^{\top}\alpha_{ik}^{-1}\beta_{ik},$$
$$\lambda_{il} = h_{il}^{\top}b_{il} = h_{il}^{\top}\alpha_{il}^{-1}\beta_{il},$$

are the *scaled edge convection* scalars, or *edge convection* in short; α_{ik}, β_{ik}, α_{il}, and β_{il} the evaluations of α and β at edges (i,k) and (i,l) respectively. Equation (2.8) is a direct result from solving the two one-dimensional first order ordinary differential equations in (2.7) separately. A simple calculation based on Eq. (2.8) yields the expression of the constant scaled current density as

$$\kappa_{lik} = H_{lik}^{-\top}\begin{bmatrix} B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k \\ B(-\lambda_{il})u_i - B(\lambda_{il})u_l \end{bmatrix} \tag{2.9}$$

with

$$H_{lik} = [h_{ik}, h_{il}]. \tag{2.10}$$

being the *edge-pair matrix*.

Remark 2.1 We note that the constant-j assumption is just one of many techniques that can be used to approximate the projection of (scaled) current density along an edge. For example, one may replace the Bernoulli function used in Eq. (2.8) by an appropriate function to obtain an upwinding type of discretization along an edge. As long as the function used to compute the coefficients of u in Eq. (2.8) is always nonnegative, the monotonicity of the final discretization will not depend on the convection of the problem as we shall see later.

Remark 2.2 In the case where the scaled convection is curl-free (see §3.4), we are able to solve the above set of ordinary differential equations (2.7) *exactly* without assuming the scaled convection b is constant on either of the edges. The solution is still in the form of Eq. (2.8) with

$$\lambda_{ik} = \psi(p_i) - \psi(p_k), \tag{2.11}$$

$$\lambda_{il} = \psi(p_i) - \psi(p_l), \tag{2.12}$$

where $b = -\nabla\psi$ for some potential function ψ. This result can also be extended to the box method in three dimensions.

Remark 2.3 When the model (1.1) has no convection, *i.e.*, $\beta \equiv 0$, the scaled current density is simply the negative gradient of the unknown state function, $-\nabla u$. Furthermore, the three constant vectors, namely κ_{lik}, κ_{ikl}, and κ_{kli}, approximating the negative gradient on the three edge-pairs of a triangle are identical, which means that the box method approximates the gradient by a single constant vector on each triangle. This observation is also valid in three dimensions.

Equation (2.9) enables us to approximate the edge-pair flux as

$$\int_{\overline{m_1 o} \cup \overline{o m_2}} \nu^\top j \, ds \approx \int_{\overline{m_1 o} \cup \overline{o m_2}} \nu^\top \alpha \kappa_{lik} \, ds. \tag{2.13}$$

The quasi-discrete form of Eq. (2.13) can be written in two equivalent ways:

$$\int_{\overline{m_1 o} \cup \overline{o m_2}} \nu^\top \alpha \kappa_{lik} \, ds = [\tilde{g}^{\mathrm{R}}_{lik}, \tilde{g}^{\mathrm{L}}_{lik}] \begin{bmatrix} B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k \\ B(-\lambda_{il})u_i - B(\lambda_{il})u_l \end{bmatrix}, \tag{2.14}$$

and

$$\int_{\overline{m_1 o} \cup \overline{o m_2}} \nu^\top \alpha \kappa_{lik} \, ds = e^\top \tilde{G}_{lik} \begin{bmatrix} B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k \\ B(-\lambda_{il})u_i - B(\lambda_{il})u_l \end{bmatrix}, \tag{2.15}$$

with $e = [1, 1]^\top$. Here

$$[\tilde{g}^{\mathrm{R}}_{lik}, \tilde{g}^{\mathrm{L}}_{lik}] = \int_{\overline{m_1 o} \cup \overline{o m_2}} \nu^\top \alpha H_{lik}^{-\top} \, ds, \tag{2.16}$$

is the pair of *right* and *left quasi-discrete edge conductance* scalars that are associated with the right edge (i, k) and the left edge (i, l) of the edge-pair $((i, k), (i, l))$ respectively, and

$$\tilde{G}_{lik} = \begin{bmatrix} \int_{\overline{m_1 o}} \nu^\top \alpha ds \\ \int_{\overline{m_2 o}} \nu^\top \alpha ds \end{bmatrix} H_{lik}^{-\top} \tag{2.17}$$

is the *quasi-discrete edge-pair conductance* matrix. Approximating the quasi-discrete edge conductance pair and the edge-pair conductance matrix are two equivalent approaches that allow the edge-pair flux in Eq. (2.13) to be completely discretized. However, each approach provides a different interpretation. Here, we only present the derivation of the discrete edge conductance pair approximating (2.16). For derivation of the discrete edge-pair conductance matrix approximating (2.17), see [32], §3.2.4 and also [33].

The quasi-discrete edge-pair conductance scalars (2.16) can be discretized by assuming the diffusion tensor α is constant in some manner. We may choose to use different constants approximating α for the two edge-pair conductance scalars in Eq. (2.14), or we may choose them to be equal. These two choices lead to the following two different evaluation schemes for the diffusion coefficient.

1. Per edge conductivity evaluation – we evaluate the conductivity (diffusivity) tensor α by two edge-associated constant 2×2 matrices, α_{ik} at edge (i,k) and α_{il} at edge (i,l), to approximate the right and the left quasi-discrete edge conductance scalars in Eq. (2.16) as

$$\tilde{g}^{\mathrm{R}}_{lik} \approx g^{\mathrm{R}}_{lik} = (s_l \nu_l^\top + s_k \nu_k^\top)\alpha_{ik} H_{lik}^{-\top} e_1, \quad (2.18)$$

$$\tilde{g}^{\mathrm{L}}_{lik} \approx g^{\mathrm{L}}_{lik} = (s_l \nu_l^\top + s_k \nu_k^\top)\alpha_{il} H_{lik}^{-\top} e_2, \quad (2.19)$$

where s and ν are respectively the length and the unit normal vector for a box edge; see Figure 2.2.

2. edge-pair conductivity evaluation – we evaluate the conductivity tensor α by a single constant 2×2 matrix, α_{lik}, associated with the edge-pair $((i,k),(i,l))$ to approximate both the right and the left quasi-discrete edge conductance scalars as

$$[\tilde{g}^{\mathrm{R}}_{lik}, \tilde{g}^{\mathrm{L}}_{lik}] \approx [g^{\mathrm{R}}_{lik}, g^{\mathrm{L}}_{lik}] = (s_l \nu_{ik}^\top + s_k \nu_{il}^\top)\alpha_{lik} H_{lik}^{-\top}. \quad (2.20)$$

A convenient special case here is to let the three edge-pairs of a single triangle all give the same evaluation result, which then is called the *per triangle* evaluation scheme.

Following the derivation in [32], Appendix B, the expressions of the discrete edge conductance scalars (Eqs. (2.18)–(2.20) can be simplified as

$$g^{\mathrm{R}}_{lik} = \frac{1}{2}\frac{\det(\alpha_{\mathrm{ev}})}{\det(H_{lik})} h_{il}^\top \alpha_{\mathrm{ev}}^{-1} h_{kl}, \quad (2.21)$$

$$g^{\mathrm{L}}_{lik} = \frac{1}{2}\frac{\det(\alpha_{\mathrm{ev}})}{\det(H_{lik})} h_{ik}^\top \alpha_{\mathrm{ev}}^{-1} h_{lk}, \quad (2.22)$$

where α_{ev} can be substituted with any chosen evaluation of the diffusion tensor. The above Eqs. (2.21), (2.22) show that the discrete edge conductance scalars do *not* depend on the choice of box vertex o. They depend on only the local mesh geometry, the triangle edge vectors, and the evaluation of the diffusion tensor. However, the area integral on the right-hand-side of Eq. (2.3) still depends on the coordinates of box vertex o.

Let us define the *generalized* dot product, cross product and angle between two vectors u and v with respect to a symmetric and positive-definite μ as

$$(u \cdot v)_\mu \equiv u^\top \mu v, \quad (2.23)$$

$$(u \times v)_\mu \equiv \sqrt{\det(\mu)}\det([u,v]), \quad (2.24)$$

$$\theta_\mu(u,v) = \cot^{-1}\left(\frac{(u \cdot v)_\mu}{(u \times v)_\mu}\right), \quad (2.25)$$

which can be thought as the normal dot product, cross product and angle measured in the inner product space with the inner product (2.23). Then the edge conductance values can be interpreted as

$$g^{\mathrm{R}}_{lik} = \frac{1}{2}\sqrt{\det(\alpha_{\mathrm{ev}})}\cot(\theta^{\mathrm{R}}_{\alpha_{\mathrm{ev}}^{-1}}), \quad (2.26)$$

$$g^{\mathrm{L}}_{lik} = \frac{1}{2}\sqrt{\det(\alpha_{\mathrm{ev}})}\cot(\theta^{\mathrm{L}}_{\alpha_{\mathrm{ev}}^{-1}}), \quad (2.27)$$

where $\theta^{\mathrm{R}}_{\alpha_{\mathrm{ev}}^{-1}}$ is the *generalized angle* between edge vector h_{il} and h_{kl}, and $\theta^{\mathrm{L}}_{\alpha_{\mathrm{ev}}^{-1}}$ is the *generalized angle* between edge vector h_{ik} and h_{lk}.

With the scaled edge convection and the edge conductance well-defined, we are able to approximate the edge-pair flux in Eq. (2.13) by the *edge-pair current*

$$\begin{aligned} I_{lik} \equiv\ & g^{\mathrm{R}}_{lik}(B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k) \\ & + g^{\mathrm{L}}_{lik}(B(-\lambda_{il})u_i - B(\lambda_{il})u_l), \end{aligned} \quad (2.28)$$

which can be split into the *right edge current*

$$I^{\mathrm{R}}_{lik} = g^{\mathrm{R}}_{lik}(B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k) \quad (2.29)$$

and the *left edge current*

$$I^{\mathrm{L}}_{lik} = g^{\mathrm{L}}_{lik}(B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k). \quad (2.30)$$

2.5. Global Assembly

We approximate the left-hand-side line integral of the local integral form (2.3) (or the net outward-flux through the boundary of box B_i) by the net nodal current, the sum of edge-pair currents (2.28) for all edge-pairs incident on point p_i. Here we assume zero-flux for all Neumann boundary conditions. Since each edge-pair current consists of right and left edge currents, we can regroup the summation as

$$I_i = \sum_{((i,k),(i,l))\ni i} I_{lik} = \sum_{(i,k)\ni i} I_{ik} \tag{2.31}$$

such that the net current at p_i is expressed as the sum of *edge currents* I_{ik} for edges (i,k) incident to p_i. Based on their locations, the mesh edges can be divided into boundary and internal edges. An internal edge (i,k) is shared by two edge-pairs that are incident to point p_i (Fig. 2.3 left), and its edge current is the sum of the right edge current for the edge-pair to its left and the left edge current for the edge-pair to its right:

$$I_{ik} = I^{\mathrm{R}}_{lik} + I^{\mathrm{L}}_{kij}. \tag{2.32}$$

A boundary edge (i,k) is attached to only one edge-pair incident to point p_i (Fig. 2.3 right), and its edge current is either the right or the left edge current of the edge-pair it is attached to, depending on edge orientation with respect to the defining domain Ω:

$$I_{ik} = I^{\mathrm{L}}_{kij}. \tag{2.33}$$

Furthermore, since the scaled edge convection is computed by evaluating the diffusion and convection coefficients at edges, it is invariant on an edge regardless which edge-pair that contains the edge is being considered. As a result, we are able to write the edge current from p_i to p_k as

$$I_{ik} = g_{ik}(B(-\lambda_{ik})u_i - B(\lambda_{ik})u_k), \tag{2.34}$$

with

$$g_{ik} = g^{\mathrm{R}}_{lik} + g^{\mathrm{L}}_{kij} \tag{2.35}$$

if edge (i,k) is an internal edge and

$$g_{ik} = g^{\mathrm{L}}_{kij} \tag{2.36}$$

if it is a boundary edge with Neumann boundary condition.

By considering all the nodal net currents altogether as the box method approximation of the current density divergence over the entire defining domain Ω, we have a linear system

$$Mu_h = f_h + b \tag{2.37}$$

with the stiffness matrix M approximating the convection-diffusion operator $\nabla\cdot(-\alpha\nabla+\beta)$ in the model Eq. (1.1), the vector f_h approximating the per box area integral of the source term f, and the vector b approximating Dirichlet and Neumann boundary conditions assigned to the boundary of the defining domain.

We can construct the stiffness matrix M through edge assembly. Based on their incidence relation

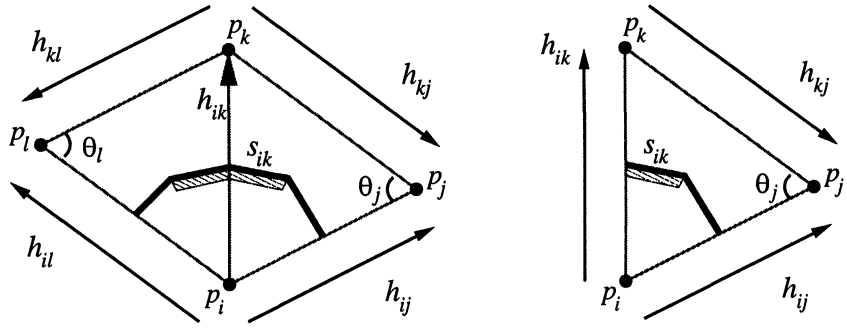

FIGURE 2.3 An internal edge (i,k) shared by two triangles (left) and a boundary edge (i,k) attached to only one triangle.

with Dirichlet boundary points, we divide the edge set into non-grounded edges E and grounded edges E_0.

- For a floating edge (j,l), both I_{jl} and I_{lj} are used for integration around point p_j and point p_l. Therefore, its contribution to the stiffness matrix M is the 2×2 edge stamp

$$s_{jl} = \begin{bmatrix} g_{jl}B(-\lambda_{jl}) & -g_{jl}B(\lambda_{jl}) \\ -g_{lj}B(\lambda_{lj}) & g_{lj}B(-\lambda_{lj}) \end{bmatrix}_{jl} \tag{2.38}$$

with use of the notation

$$\begin{bmatrix} s_{12} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}_{jl} \equiv [e_{j,n}, e_{l,n}] \begin{bmatrix} s_{12} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} e_{j,n}^\top \\ e_{l,n}^\top \end{bmatrix} \tag{2.39}$$

with $e_{i,n}$ and $e_{j,n}$ being the ith and jth columns of the $n \times n$ identity matrix respectively.

- If p_l is the Dirichlet point for grounded edge (j,l), the grounded edge (j,l) contributes to the stiffness matrix M by a 1×1 edge stamp

$$s_{jl} = g_{jl}[B(-\lambda_{jl})]_j \tag{2.40}$$

with use of the notation

$$[s]_j \equiv e_{j,n} s e_{j,n}^\top. \tag{2.41}$$

It also contributes to the right-hand-side b by

$$b_{jl} = g_{jl}B(\lambda_{jl})e_{j,n}. \tag{2.42}$$

Then it is clear that the stiffness matrix can be written in the edge assembly form as

$$M = \sum_{(j,l)\in E\cup E_0} s_{jl}, \tag{2.43}$$

and

$$b = \sum_{(j,l)\in E_0} b_{jl}. \tag{2.44}$$

The edge-assembly view of the stiffness matrix M provides a circuit interpretation of the box method discretization, and we pursue this interpretation in more detail in §3.3. There we will orient each grid edge (j,l) such that (j,l) will become an edge in the directed graph $G(M)$ with $\lambda_{jl} \geq 0$ using $\lambda_{jl} = -\lambda_{lj}$.

2.6. Equivalence to a Finite Element Method

For non-convective problems, the equivalence between the box method and the piecewise linear finite element method in one and two dimensions has been well described in [4, 18, 20]. The use of piecewise linear base functions in finite element methods can be considered exactly equivalent to the constant-j assumption of the box method (see Remark 2.3). More recently, Xu and Zikatanov [38] presented the *edge-averaged finite element* (EAFE) scheme which gives the same left-hand-side discretization as does the box method for two-dimensional *isotropic* convection-diffusion problems when the evaluation schemes for the coefficients α and β used in the two methods are consistent. See also Gatti, Micheletti and Sacco [16] for a similar discussion.

In this section, we generalize the EAFE scheme to the anisotropic EAFE scheme such that it will be equivalent to the box method for general anisotropic convection-diffusion problems again assuming consistent evaluation schemes.

Suppose we are solving the following boundary value problem

$$\nabla \cdot (-\alpha \nabla u + \beta u) = f \tag{2.45}$$

on the defining domain Ω with zero Dirichlet boundary condition. The weak formulation of the above problem is to find a function u in the Soblev space $H_0^1(\Omega)$, such that

$$a(u,v) \equiv \int_\Omega (\alpha \nabla u - \beta u)^\top (\nabla v) da = f(v) \tag{2.46}$$

for every trial function $v \in H_0^1(\Omega)$. Let V_T denote the space of piecewise linear functions on triangles in the triangular mesh T, with a set of basis functions defined as

$$v_i(p_j) = \delta_{ij}, \quad i,j = 1, \ldots, n. \tag{2.47}$$

Then the finite element formulation is to find a piecewise linear function $u_h \in V_T$ that approximates the true solution to Eq. (2.45) such that

$$a_h(u_h, v_h) = \sum_{t \in T} \int_t (\alpha \nabla u_h - \beta u_h)^\top (\nabla v_h) da = f_h \tag{2.48}$$

with $f_h = f(v_h)$, for all trial functions $v_h \in V_T$. The left-hand-side integral in (2.48) on a triangle $t = (i, k, l)$ can be written as

$$\begin{aligned} &\int_t -j^\top(u_h) \nabla v da \\ &\quad = \int_t -\kappa^\top(u_h) \alpha \nabla v da \\ &\quad = \int_t -\kappa^\top(u_h) \\ &\qquad \frac{\tilde{g}_{kl,t} h_{kl} h_{kl}^\top + \tilde{g}_{li,t} h_{li} h_{li}^\top + \tilde{g}_{ik,t} h_{ik} h_{ik}^\top}{|t|} \nabla v da, \end{aligned} \tag{2.49}$$

where

$$\tilde{g}_{kl,t} = -\frac{1}{2} \frac{\det(\alpha)}{2|t|} (h_{li}^\top \alpha^{-1} h_{ik}), \tag{2.50}$$

$$\tilde{g}_{li,t} = -\frac{1}{2} \frac{\det(\alpha)}{2|t|} (h_{ik}^\top \alpha^{-1} h_{kl}), \tag{2.51}$$

$$\tilde{g}_{ik,t} = -\frac{1}{2} \frac{\det(\alpha)}{2|t|} (h_{kl}^\top \alpha^{-1} h_{li}) \tag{2.52}$$

are the edge conductance scalars for edges h_{kl}, h_{li} and h_{ik} respectively (*cf.* Fig. 2.2), Eqs. (2.21), (2.22)). Here we have used the identity (2.19) in [2]. Using the edge average approximation Eq. (3.11) in [38], this integral (2.49) can be approximately written as

$$\sum_{e \subset t} g_{e,t} \gamma_e \delta_e (e^{-\psi_e} u_h) \delta_e (v_h), \tag{2.53}$$

where e takes edges h_{kl}, h_{li} and h_{ik}, δ_e is the edge differencing operator, and the edge-associated quantities g, ψ and γ are defined as

$$\begin{aligned} g_{kl,t} &= \frac{1}{|t|} \int_t \tilde{g}_{kl,t} da \\ &= -\frac{1}{4|t|^2} h_{li}^\top \left(\int_t \det(\alpha) \alpha^{-1} da \right) h_{ik}, \end{aligned} \tag{2.54}$$

$$\psi_{kl}(p) = \frac{1}{\|h_{kl}\|} \int_{p_k}^{p} h_{kl}^\top \alpha^{-1} \beta ds, \tag{2.55}$$

$$\gamma_{kl} = \|h_{kl}\| \left(\int_{h_{kl}} e^{-\psi_{kl}} dl \right)^{-1}, \tag{2.56}$$

for edge h_{kl}, and cyclically for edges h_{li} and h_{ik}. In reality, these integrals may have to be computed approximately. One simple approach is to assume $\alpha = \alpha_{kl}$ and $\beta = \beta_{kl}$ are constant along edge (k, l) in these integrals, which leads to the following approximation of (2.54), (2.55), (2.56):

$$g_{kl,t} = -\frac{1}{2} \frac{\det(\alpha_{kl})}{2|t|} (h_{li}^\top \alpha_{kl} h_{ik}), \tag{2.57}$$

$$\psi_{kl}(p) \approx (p - p_k)^\top \alpha_{kl}^{-1} \beta_{kl}, \tag{2.58}$$

$$\gamma_{kl} \approx B(\lambda_{kl}), \tag{2.59}$$

This assumption corresponds to the per edge evaluation scheme of the constant-j box method. Here, the edge convection $\lambda_{kl} = h_{kl}^\top \alpha_{kl}^{-1} \beta_{kl}$.

If we represent both u_h and v_h in terms of the basis functions, then the above bilinear form becomes a linear system

$$M u_h = f_h, \tag{2.60}$$

with $f_h^i = f(v_i)$ and the (i, k) entry of the stiffness matrix M in the form

$$m_{ik} = -\sum_{t \supset e = (i,k)} g_{e,t} B(\lambda_e). \tag{2.61}$$

It is easy to verify that the expression of m_{jl} above is exactly the same as the $(2, 1)$ entry of the edge stamp (3.3). For p_i not connected to a Dirichlet boundary node, we have

$$m_{ii} = \sum_{(i,k) \in E} |m_{jl}|, \tag{2.62}$$

which also coincides with the result from the box method discretization. Equivalence between the box method and the above anisotropic EAFE scheme at boundary nodes also can be easily verified. If we choose to use the finite element approach to obtain the approximation of the right-hand-side (area integral) in Eq. (2.3), then the box method and the anisotropic EAFE method are completely equivalent.

Using the same finite element analysis provided in [38], we obtain the following *a priori* error bound.

THEOREM 2.4 *Let u be the solution of the problem* (2.45). *Assume that for all* $t \in T$, α *is symmetric and positive-definite whose eigenvalues are uniformly bounded away from zero, all components of* α *and* β *are in* $W^{1,\infty}(t)$, *and the scaled current density* $\kappa \equiv -\nabla u + bu \in [W^{1,2}(t)]^2$. *Then the following estimate holds*

$$\|u - u_h\|_{1,\Omega} \leq Ch\left(\sum_{t\in\mathcal{T}} |\kappa(u)|^2_{1,2,t}\right)^{1/2} \tag{2.63}$$

for sufficiently small h.

Remark 2.5 We notice that the above result is based on the *exact* computation of the quasi-discrete quantities – edge conductance $g_{e,t}$, edge convection potential ψ_e and its exponential edge average γ_e – which are in the form of integrals of the diffusion and convection coefficients. Therefore, the error analysis needs to be interpreted carefully when the coefficients become irregular.

The above error estimator is called *a priori* because it is generally true for a class of problems in the form of Eq. (2.45) with only smoothness restrictions on the diffusion and convection coefficients. [31] presented a tight one-dimensional *a priori* error analysis on the box method for constant-coefficient convection-diffusion equations. It shows that the maximum norm of the solution error is bounded by $O(h^2|b|)$ where $b = \beta/\alpha$ is the scaled convection.

3. MATRIX PROPERTIES

In this section, we will identify a set of conditions, which depends only on the edge conductance and convection evaluation schemes as well as mesh properties, that give rise to the M-matrix properties of the discretization system of (1.1). Furthermore, we define discrete analogies of the curl and divergence of the (scaled) convection vector field in terms of the discrete edge conductance and convection. We show that certain conditions on the discrete curl and the discrete divergence imply that the stiffness matrix M is symmetrizable or row diagonally dominant. Although these convection-related properties do not affect the monotonicity of the box method discretization, they do arise in other discretization methods to ensure monotonicity, and they appear as important conditions in our convective iteration analyses in §5.

3.1. *M*-matrix

A real matrix A is said to be *nonnegative*, denoted by $A \geq 0$, if all its entries $a_{jl} \geq 0$. Let $Z^{n\times n}$ denote the class of *Z-matrices*, square matrices with nonpositive off-diagonal elements, *i.e.*,

$$Z^{n\times n} \equiv \{A = [a_{jl}] \in \mathbb{R}^{n\times n} | a_{jl} \leq 0,\ j \neq l\}.$$

A general definition of M-matrices is stated as follows.

DEFINITION 3.1 ([7] Chapter 6) A real square matrix A is called an M-matrix if and only if A can be written in the form,

$$A = sI - B, \quad s > 0, \quad B \geq 0, \tag{3.1}$$

where $s \geq \rho(B)$; $\rho(B)$ is the spectral radius of B.

The spectral radius of a square matrix is defined as the largest modulus of all its eigenvalues. A nonsingular M-matrix requires $s > \rho(B)$, hence A^{-1} exists. Equivalently, a nonsingular M-matrix is a Z-matrix with nonnegative inverse. A symmetric nonsingular Z-matrix is an M-matrix if

and only if it is positive-definite; such a matrix is called a *Stieltjes matrix*. For a Z-matrix A, there are other equivalent conditions for A to be a nonsingular M-matrix [7], such as

- There exists a positive diagonal matrix D_1 such that D_1A is strictly column diagonally dominant;
- There exists a positive diagonal matrix D_2 such that AD_2 is strictly row diagonally dominant;
- There exists positive diagonal matrices D_1 and D_2 such that D_1AD_2 is both strictly row and column diagonally dominant.

When A is irreducible (equivalently, the graph $G(A)$ defined in §4.1 is connected), then the above conditions can be changed to irreducibly row (column) diagonally dominant; see Varga [36], page 23, and Axelsson [1], Definition 4.4. An irreducible M-matrix has $A^{-1} > 0$, *i.e.*, each entry of A^{-1} is positive.

3.2. Nonnegative Column Sum and Edge Current Conservation

From the electrical circuit viewpoint, it is natural to assume that the current is conserved along the edge (j,l). Using Eq. (2.34), such an assumption corresponds to

$$I_{jl} + I_{lj} = 0 \Leftrightarrow g_{jl} = g_{lj}, \tag{3.2}$$

which we call the *edge conservation condition.*

LEMMA 3.2 *If we use the per edge or the per triangle evaluation scheme to evaluate the diffusion tensor field α when computing the edge conductance scalars, then for each edge $(i,k) \in E$, we have $g_{ik} = g_{ki}$, i.e., the edge conservation condition is satisfied.*

Proof See [32], Appendix F.

Notice that we always evaluate α and β edge-wise when computing the edge convection λ_{jl}. Then the 2×2 edge stamp that is used to assemble the stiffness matrix M in Eq. (2.43) can be simplified as

$$g_{jl}\begin{bmatrix} B(-\lambda_{jl}) & -B(\lambda_{jl}) \\ -B(-\lambda_{jl}) & B(\lambda_{jl}) \end{bmatrix}_{jl} \tag{3.3}$$

for edge (j,l), noticing that $\lambda_{jl} = -\lambda_{lj}$. Since the column sum for each edge stamp is zero, it is obvious that the column sum of the stiffness matrix is either zero when the column corresponds to a non-grounded node or positive otherwise. Finally, we show the following result.

LEMMA 3.3 *An irreducible square matrix with nonnegative column sums and at least one positive column sum is an M-matrix if and only if it is a Z-matrix.*

Proof If a square matrix A is a Z-matrix and has nonnegative column sums, then A is column diagonally dominant. Furthermore, if A has at least one positive column sum, A is irreducibly column diagonally dominant. Applying the M-matrix conditions given in §3.1, we have A is an M-matrix. The other part of the proof is obvious since an M-matrix is always a Z-matrix. ■

3.3. *M*-matrix and Nonnegative Edge Resistors

From classical partial differential equations theory, we know that the elliptic operator $\nabla \cdot (-\alpha \nabla + \beta)$ satisfies the *maximum principle*, which states that the maximum and the minimum of the solution u to the boundary value problem,

$$\nabla \cdot (-\alpha \nabla u + \beta u) = 0,$$

occur only at the boundary of the defining domain (*cf.* [35, 21]). The discrete analogy of such property will be that the stiffness matrix M is *inverse-monotone* or simply *monotone* (*cf.* [26]), *i.e.*,

$$Mv > 0 \Rightarrow v > 0,$$

for any vector v. A sufficient condition for M to be monotone is that M is an *M-matrix*, (*cf.* Definition 3.1). The M-matrix property of the stiffness

matrix M is equivalent to the nonnegativity of edge conductance scalars, *i.e.*,

$$g_{ik} \geq 0 \tag{3.4}$$

which is consistent with the diffusive nature of the elliptic operator $\nabla\cdot(-\alpha\nabla+\beta)$ being approximated. Since, in addition, a monotone scheme is numerically stable (no spurious oscillations), monotonicity of the stiffness matrix M is desired for discretization schemes on elliptic partial differential equations. We have seen from simulation examples that discretized systems without the M-matrix property may cause artificial spurious wave fronts in reaction-diffusion systems such as the heart models. In addition to capturing the monotonicity of the physical problem and ensuring numerical stability, many iterative methods for solving M-matrix systems are guaranteed to converge. Furthermore, detailed analyses on such convergence is often possible as we shall see in §5.

3.3.1. Main Results

We now examine conditions on the mesh and ways to evaluate conductivity tensor α that will ensure that the stiffness matrix M is an M-matrix. Both the algebra of edge-assembly (2.43) and the edge conservation condition (3.2) imply that the M-matrix property of the stiffness matrix M is completely determined by the nonnegativity of the discrete conductance scalars.

LEMMA 3.4 *Under the edge conservation condition* (3.2), *the stiffness matrix M derived by the box method is an irreducible M-matrix* (*cf. Definition* 3.1) *if and only if the edge conductance scalar g_{ik} is nonnegative for each* (*directed*) *edge $(i,k)\in E$.*

Proof Using Lemmas 3.2 and 3.3, we only need to show that the stiffness matrix M is a Z-matrix. By looking at the formulae in Eqs. (2.38)–(2.43), it is clear that the edge conductance g_{ik} is nonnegative if and only if the corresponding off-diagonal entry in the stiffness matrix M is nonpositive, due to the nonnegativity of the Bernoulli function $B(x)$ (see Appendix A). Then M is a Z-matrix by definition. ■

Remark 3.5 The stiffness matrix is irreducibly diagonally dominant if there is at least one Dirichlet node. In this case the M is a Stieltjes matrix (*cf.* §3.1) if the convection vanishes identically in Eq. (1.1).

Using the formulae for edge conductance scalars we derived in §2, we have the following results.

THEOREM 3.6 *If we use per edge evaluation scheme in the box method discretization, the stiffness matrix M is an M-matrix if and only if we have*

$$\frac{(h_{kj}\cdot h_{ij})_{\alpha_{ik}^{-1}}}{(h_{kj}\times h_{ij})_{\alpha_{ik}^{-1}}} \geq 0 \tag{3.5}$$

for each boundary or interface edge (i,k) (*cf. Fig.* 2.3 *right*), *and*

$$\frac{(h_{il}\cdot h_{kl})_{\alpha_{ik}^{-1}}}{(h_{il}\times h_{kl})_{\alpha_{ik}^{-1}}}+\frac{(h_{kj}\cdot h_{ij})_{\alpha_{ik}^{-1}}}{(h_{kj}\times h_{ij})_{\alpha_{ik}^{-1}}} \geq 0 \tag{3.6}$$

for (i,k) being other edges (*cf. Fig.* 2.3 *left*), *where α_{ik} is the evaluation of the diffusion tensor at edge (i,k).*

Proof See [32], Appendix G.

With respect to the edge conductivity evaluation α_{ik}, the inequality (3.5) is called the *anisotropic boundary Delaunay condition* for triangle $\Delta p_i p_j p_k$ that is attached to a boundary edge (i,k), and the inequality (3.6) is called the *anisotropic Delaunay condition* for the pair of triangles $\Delta p_i p_j p_k$ and $\Delta p_k p_l p_i$ that share an edge (i,k). When the diffusion tensor α becomes a scalar, we have the following known corollary [38, 3].

COROLLARY 3.7 *When the diffusion tensor α in Eq.* (1.1) *degenerates to a scalar, then the stiffness matrix of the per edge evaluation scheme is an M-matrix if and only if any two triangles (i,j,k) and (k,l,i) that share an edge (i,k) satisfy the generic Delaunay condition or simply the Delaunay*

condition

$$< p_i p_j p_k + < p_k p_l p_i \leq \pi. \tag{3.7}$$

and any triangle (i,j,k) *that is attached to a boundary or interface edge* (i,k) *satisfies the generic boundary Delaunay condition or simply the boundary Delaunay condition*

$$< p_i p_j p_k \leq \frac{\pi}{2}. \tag{3.8}$$

As of this point, we have shown that a sufficient condition for the stiffness matrix M to be an M-matrix is that (i), we evaluate the conductivity α at edges and that each pair of triangles sharing a common non-interface edge in the underlying mesh satisfies the anisotropic Delaunay condition (3.6); and that (ii), each triangle attached to a boundary or an interface or a flux barrier satisfies the anisotropic boundary Delaunay condition (3.5). Such a triangular mesh is called an *edgewise anisotropic Delaunay mesh* with respect to the edge conductivity evaluation.

A special case of the edge conductivity evaluation scheme is to assign the same conductivity value for edge (i,k) shared by triangles $\Delta p_i p_j p_k$ and $\Delta p_k p_l p_i$ (*cf.* Fig. 2.3 left), or edge (j,l) shared by triangles $\Delta p_i p_j p_l$ and $\Delta p_k p_l p_j$, for edge (i,k) or (j,l) not being an interface edge. Such evaluation scheme is also called the *per quadrilateral* conductivity evaluation scheme. A triangular mesh that satisfies the anisotropic Delaunay condition and the anisotropic boundary Delaunay condition locally and respects the quadrilateral conductivity evaluation scheme is called a *pairwise anisotropic Delaunay mesh.*

We see that the anisotropic Delaunay condition is only a condition on the "generalized angles" of the triangles, and not on their sizes. The significance lies in the fact that one is allowed to use an arbitrarily coarse mesh, if the mesh is pairwise anisotropic Delaunay, to get a qualitatively correct solution, which may be important in iterative methods where a hierarchy of meshes is used.

3.3.2. Anisotropic Delaunay Mesh

We have shown that our discretization method is guaranteed *a priori* to be "monotone" for (1.1) if the mesh satisfies the *anisotropic Delaunay condition.* While we are aware of that implementations of anisotropic meshes exist [11, 34, 8], we have also noticed that these implementations are mainly used for a *posteriori* error control of the numerical solutions; and these implementations do not always guarantee the exact anisotropic Delaunay condition (3.5) and (3.6) that is required by the constant-j box method for monotone discretization. Therefore, we chose to implement our own anisotropic Delaunay mesh generator to meet the condition for monotone discretization. Our mesh generator uses a generalized version of the Delaunay refinement algorithm which is described in [32], Chapter 5. For a given continuous diffusion tensor field on a closed and bounded domain (subdomain) with arbitrary geometry, the output mesh is guaranteed to satisfy the anisotropic Delaunay condition. The density of the mesh mostly depends on how severely the anisotropy of the diffusivity changes in the domain, and it does not depend on the strength of the anisotropy. We believe that such restriction on mesh density is likely to occur in any implementation of the anisotropic Delaunay mesh generation.

3.3.3. A Simple Monotonicity-preserving Fix

In cases where a given mesh does not have all its edges satisfying the anisotropic Delaunay condition, the stiffness matrix may lose monotonicity, a fundamental property of the elliptic operator $-\nabla \cdot \alpha \nabla$ which the discretization should always inherit. Such matrices, when applied to the heart model equations, may even cause artificial wave fronts. Here we propose the following simple post-processing procedure that will guarantee that the output stiffness matrix is an M-matrix, and therefore the discretization system is monotone:

$$\tilde{g}_{ik} = \max\{g_{ik}, 0\}. \tag{3.9}$$

By Lemma 3.4, it is obvious that the matrix assembled by discrete edge conductance $\tilde{g}_{ik}$ is an M-matrix.

3.4. Discrete Curl-free Condition and Diagonal Symmetrization

Recall that the continuity equations of semiconductor devices modeling, (1.1) take the form

$$\nabla \cdot (-D\nabla u + \mu \mathcal{E} u) = f, \quad \mathcal{E} = -\nabla \psi, \ \frac{D}{\mu} = \text{const.} \tag{3.10}$$

if the Einstein relation on mobilities and diffusivities holds; see Selberherr [30], and Fichtner, Rose and Bank [14].

We generalize the property of the electric field $\mathcal{E}$ and the Einstein relation to the curl-free condition on the *scaled convection* vector field $b = \alpha^{-1}\beta$: $\nabla \times b \equiv 0$ on the defining region Ω. Based on theorems from advanced calculus [24], we have the following equivalent conditions for b being curl-free on Ω.

- For any connected open subset Ω_0 of Ω, the line integral of $\tau^\top b$, the tangential component of b, along the closed boundary loop of Ω_0 is zero:

$$\oint_{\partial\Omega_0} \tau^\top b dl = 0;$$

- There exists a differentiable real valued function, ψ, defined on Ω, such that the vector field b is the negative gradient of the function ψ:

$$b = -\nabla \psi. \tag{3.11}$$

We consider the triangular mesh over the defining domain Ω as a planar graph $G(P, E)$. For each directed edge (j, l) in E, we assign a scalar

$$\lambda_{jl} = \int_0^{\|p_l - p_j\|} \tau^\top b ds = \psi(p_j) - \psi(p_i), \tag{3.12}$$

which is the integral of tangential component of b along the edge (j, l). We say that (G, λ), a combination of the mesh and the scalars assigned to all the mesh edges, is a discretization of the vector field b. Then the *discretized curl-free condition* is stated as follows.

CONDITION 3.8 (Discrete Curl-free) *The discretization (G, λ) of a vector field b satisfies the discretized curl-free condition if and only if for each cycle $(i_1, i_2, \ldots, i_r, i_1)$ of the graph G, we have*

$$\lambda_{i_1 i_2} + \lambda_{i_2 i_3} + \cdots + \lambda_{i_{r-1} i_r} + \lambda_{i_r i_1} = 0. \tag{3.13}$$

Remark 3.9 If the graph G is a triangulation, the discrete curl-free condition (Condition 3.8) can be equivalently stated as

$$\lambda_{ik} + \lambda_{kl} + \lambda_{li} = 0 \tag{3.14}$$

for each triangle (i, k, l) in the triangulation. This is obvious because

- A triangle (i, k, l) is a cycle in the graph G.
- Any cycle of the graph G encloses a set of triangles, and the sum of edge convection along the cycle is the same as the sum of triangle edge convection sums, which are the sums of edge convection around all triangles enclosed in the cycle.

We chose to state the discrete curl-free condition in the form of Condition 3.8 so that we would be able to apply the related results even when a non-triangular mesh is used.

There are two obvious ways to compute the edge convection λ's on a mesh so that the discretized curl-free can be satisfied when the scaled convection b in the physical problem is curl-free.

- If the convection potential function ψ is given, either analytically or discretely (with use of a background mesh), then the edge convection λ_{jl} for edge (j, l) can be computed as

$$\lambda_{jl} = \psi(p_j) - \psi(p_l), \tag{3.15}$$

where $\psi(p_j)$ and $\psi(p_l)$ can be either evaluations of some analytical function or interpolations of some discrete data (*cf.* (2.11) and (2.12)). In the former case, the edge convection is computed as the exact line integral of the scaled convection

vector field along the edge (j, l):

$$\int_0^{\|h_{jl}\|} \tau^\top b ds = \int_0^{\|h_{jl}\|} \tau^\top (-\nabla\psi) ds = \psi(p_j) - \psi(p_l) = \lambda_{jl}.$$

- If the scaled convection field b is given instead of the convection potential ψ, and if the line integral of the scaled convection vector field can be computed only by approximation, then computing the edge convection scalars b using numerical integration, say the midpoint rule, such as in (3.12) will not guarantee the discrete curl-free condition. Instead, we use the following approach:

  1. Construct a spanning tree that spans the entire mesh. For each edge (j, l) in the spanning tree with node j the parent of node l, assign to the edge the numerical integration of the scaled convection, *e.g.*,

  $$\lambda_{jl} = h_{jl}^\top b|_{((p_j+p_l)/2)}. \tag{3.16}$$

  Also let $\psi_j = \psi_l + \lambda_{jl}$. We can simply set the ψ value at the root of the spanning tree to zero.
  2. Assign an edge convection value λ to each of the remaining edges in E such that the cycle formed by adding the edge to the spanning tree satisfies the discretized curl-free condition.

Based on both the curl-free condition on b and the discrete curl-free condition on λ, it is obvious that the edge convection computed on edges that are not in the spanning tree should have same order of accuracy as those that are in the spanning tree. If the edge convection scalars in the box method discretization satisfy the discretized curl-free condition (3.13), we have the following result for the stiffness matrix M.

THEOREM 3.10 *If both the edge conservation condition* (3.2) *and the discrete curl-free condition* (3.13) *are satisfied in a box method discretization of the model Eq.* (1.1), *the stiffness matrix M can be symmetrized by a positive column scaling.*

Proof Using Eqs. (2.43) and (3.3), the stiffness matrix M can be written as

$$M = \sum_{(j,l)\in E\cup E_0} g_{jl} \begin{bmatrix} B(\psi_l - \psi_j) & -B(\psi_j - \psi_l) \\ -B(\psi_l - \psi_j) & B(\psi_j - \psi_l) \end{bmatrix}_{jl}$$

when both the edge conservation condition (3.2) and the discrete curl-free condition (3.13) are satisfied. The values of ψ are obtained through either of the methods described above. Let us define a positive diagonal matrix D_s as

$$D_s = \begin{bmatrix} e^{-\psi_1} & & \\ & \ddots & \\ & & e^{-\psi_n} \end{bmatrix}. \tag{3.17}$$

By scaling both sides from right with the positive diagonal matrix D_s, we have

$$\begin{aligned} MD_s &= \sum_{(j,l)\in E\cup E_0} g_{jl} \begin{bmatrix} B(\psi_l - \psi_j) & -B(\psi_j - \psi_l) \\ -B(\psi_l - \psi_j) & B(\psi_j - \psi_l) \end{bmatrix}_{jl} D_s \\ &= \sum_{(j,l)\in E\cup E_0} g_{jl} \begin{bmatrix} e^{-\psi_j} B(\psi_l - \psi_j) & -e^{-\psi_l} B(\psi_j - \psi_l) \\ -e^{-\psi_j} B(\psi_l - \psi_j) & e^{-\psi_l} B(\psi_j - \psi_l) \end{bmatrix}_{jl} \\ &= \sum_{(j,l)\in E\cup E_0} g_{jl} \frac{\psi_j - \psi_l}{e^{\psi_j} - e^{\psi_l}} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}_{jl}. \end{aligned}$$

It is clear that

$$S = MD_s \tag{3.18}$$

is symmetric. ∎

Using Eq. (3.18), we also see that the stiffness matrix M is diagonally similar to a symmetric matrix because

$$M_s = D_s^{-(1/2)} M D_s^{(1/2)} = D_s^{-(1/2)} S D_s^{-(1/2)} \tag{3.19}$$

is a symmetric and positive-definite matrix, where the diagonal matrix D_s is defined in Eq. (3.17).

We conclude our discussion of the discrete curl-free condition with the following remarks.

Remark 3.11 If the convection potential ψ is explicitly given in the model Eq. (1.1) as in the semiconductor device models, the symmetrized stiffness matrix S can be interpreted as the box method discretization of the symmetrized problem

$$\nabla \cdot (-e^{-\psi}\alpha \nabla v) = f, \quad v = e^{\psi}u, \tag{3.20}$$

using the third evaluation scheme for the exponential function $e^{-\psi}$ discussed in [5].

Remark 3.12 Notice that Theorem 3.10 is still true even if the sign of the edge conductance g_{jl} values or the ground conductance values in G^{g} becomes negative. In fact, the discrete curl-free condition (3.13) is exactly equivalent to the combination of the "conservation law" and the "reversibility law" defined by Parter and Young [25]. The construction of the diagonal scaling matrix $D_s^{(1/2)}$ in [25] is also equivalent to our construction of the values of ψ in the spanning tree manner we have described.

4. *M*-MATRICES AND DISCRETE CONVECTION-DIFFUSION

In this section, we present a circuit interpretation of the stiffness matrix of the constant-j box method. We examine, in some detail, the M-matrix nature of the stiffness matrix M, using graph theory to discuss the combinatorial interpretation of the curl-free condition. We also show how to interpret any convective M-matrix as a generalized resistive network.

4.1. Graph and Circuit Interpretation of Convective *M*-matrices

We define *convective M-matrices* as m-matrices which are irreducible, column diagonally dominant, and structurally symmetric. Let $M=[m_{jl}]$ be an $n \times n$ convective M-matrix. We can define an undirected graph $G(V, E)$, where

$$V = \{v_1, \ldots, v_n\} \tag{4.1}$$

is a set of n vertices, corresponding to the rows (columns) of M, and E is a set of undirected edges:

$$E = \{e_k = (j,l) | m_{jl} \neq 0,\ 1 \leq j \neq l \leq n\}. \tag{4.2}$$

Notice that E does not contain any *grounded edges*, *i.e.*, edges that are connected to nodes with Dirichlet boundary conditions. We always assume that the graph G is connected, *i.e.*, the matrix M is irreducible.

Now we assign directions to the edges. The pair (j,l) for edge e_k is oriented in such a way that

$$|m_{jl}| \geq |m_{lj}|, \tag{4.3}$$

and we say that e_k is an *out-edge* for vertex v_j and an *in-edge* for vertex v_l. See Example 4.6 below. We call this directed graph the *convection-directed* graph and often still use the notation G. Furthermore, we define the *vertex-edge incidence matrix*, or simply the *incidence matrix* (see Deo [12]), $A_{n\times m}$ as

$$A(j,k) = \begin{cases} 1 & \text{if } e_k \text{ is an out-edge for } v_j, \\ -1 & \text{if } e_k \text{ is an in-edge for } v_j, \\ 0 & \text{otherwise.} \end{cases}$$

We can order the edges, and partition the edge set E into E_T, $(n-1)$ edges in a spanning tree of the undirected graph G, and E_C, $(m-n+1)$ edges that are left, and write the incidence matrix A as

$$A = [A_T | A_C], \tag{4.4}$$

accordingly. As we know from graph theory, adding an edge in E_C to the spanning tree creates a *fundamental cycle* in the undirected graph G. Therefore, we can define the *cycle matrix* (see Rose [27]), $B_{(n-m+1)\times m}$, as

$$B = [B_0 | I], \tag{4.5}$$

and

$$B_0(i,k) = \begin{cases} 1 & \text{if edge } e_k \text{ edge in } E_T \text{ is in } ith \text{ cycle,} \\ & \text{and has the same direction as } ith \text{ edge in } E_C; \\ -1 & \text{if edge } e_k \text{ edge in } E_T \text{ is in } ith \text{ cycle,} \\ & \text{and has the opposite direction as } ith \text{ edge in } E_C \\ 0 & \text{otherwise.} \end{cases}$$

Each row of the cycle matrix corresponds to a fundamental cycle in G with nonzero entries indicating the edges forming the cycle.

We define a *generalized resistor* connecting node j and node l as a ordered pair of positive real numbers (a,b) such that the current from j to l is described as

$$I_{jl} = au_j - bu_l, \tag{4.6}$$

while the current from l to j is

$$I_{lj} = -I_{jl}. \tag{4.7}$$

Then we define a *generalized resistive network* as a connected graph where each edge represents a generalized resistor. In addition, there is at least one generalized resistor for some node $v \in V$ which is connected to a boundary set B. The graph $G_0 = (V \cup B,\ E \cup E_0)$ contains G with E_0 the set of grounded edges. Let A_0 be the incidence matrix for G_0, which adds additional columns to A. A column vector $e_{j,n}$ ($-e_{j,n}$) is added if node j is connected by an out (in) grounded edge.

THEOREM 4.1 *Any convective M-matrix is the nodal admittance matrix* [12] *for a generalized resistive network where the current I_k from node j to node l is*

$$\begin{aligned} I_{jl} &= a_k u_j - b_k u_l \\ &= d_k(u_j - u_l) + c_k(u_j + u_l) \quad \textit{see Eq.} (4.13) \\ &= g_k(B(-\lambda_k)u_j - B(\lambda_k)u_l) \quad \textit{see Eq.} (4.15) \\ &= -I_{lj}. \end{aligned} \tag{4.8}$$

Proof Consider a convective M-matrix M and its graph $G(V,E)$. For edge $e_k = (j,l)$ in E, let $a_k = |m_{jl}|$ and $b_k = |m_{lj}|$ be the absolute values of the corresponding off-diagonal elements of M, with $a_k \geq b_k$. We disassemble the matrix M by edges, *i.e.*, a 2×2 stamp (defined in Eq. (2.39))

$$\begin{bmatrix} a_k & -b_k \\ -a_k & b_k \end{bmatrix}_{jl} \equiv [e_{j,n}, e_{l,n}] \begin{bmatrix} a_k & -b_k \\ -a_k & b_k \end{bmatrix} \begin{bmatrix} e_{j,n}^\top \\ e_{l,n}^\top \end{bmatrix} \tag{4.9}$$

for edge $e_k = (j,l)$. The edge stamp is simply the nodal admittance matrix for two nodes j and l connected by a generalized resistor $[a,b]$. Now the matrix M can be written as

$$M = G^{\mathrm{g}} + \sum_{e_k=(j,l)\in E} \begin{bmatrix} a_k & -b_k \\ -a_k & b_k \end{bmatrix}_{jl}, \tag{4.10}$$

where

$$G^{\mathrm{g}} = \operatorname{diag}([g_1^{\mathrm{g}}, \ldots, g_n^{\mathrm{g}}]),$$

with

$$g_j^{\mathrm{g}} = m_{jj} - \sum_{l \neq j} |m_{lj}|$$

being the diagonal dominance for jth column. Since the convective M-matrix M is always column diagonally dominant, the ground conductance values are nonnegative. ■

Remark 4.2 Even though we always have the entry with bigger absolute value (a_k) in the lower triangle for each edge stamp, the left-hand-side of (4.9), it may appear in the upper triangle of the matrix M when $j > l$.

Remark 4.3 In general, the diagonal matrix G^{g} can be assembled as

$$G^{\mathrm{g}} = \sum_{e_k \in E_0} [f_k]_{j*}, \quad f_k = a_k \quad \text{or } f_k = b_k \tag{4.11}$$

where e_k is a grounded edge connecting node j of the graph G with node $*$ in the boundary set B.

Hence, we can write M more compactly than in (4.10) as

$$M = \sum_k \begin{bmatrix} a_k & -b_k \\ -a_k & b_k \end{bmatrix}_{jl} \tag{4.12}$$

when this causes no confusion. If M alone is specified, then only the regular resistor g_j^g is known for $e_k \in E_0$ and can be interpreted in the generalized resistor form (g_j^g, g_j^g). However, if both M and the generalized resistors on the grounded edges are specified, then G^g is further decomposed as in Eq. (4.11). This is *always* the cases when M arises from discretization.

Remark 4.4 By defining the *edge diffusion* and *convection elements* as

$$d_k = \frac{a_k + b_k}{2} > 0, \quad c_k = \frac{a_k - b_k}{2} \geq 0, \tag{4.13}$$

respectively, the edge stamp can be split into a diffusive piece and a convective piece:

$$\begin{bmatrix} d_k & -d_k \\ -d_k & d_k \end{bmatrix}_{jl} + \begin{bmatrix} c_k & c_k \\ -c_k & -c_k \end{bmatrix}_{jl}. \tag{4.14}$$

Remark 4.5 If the graph $G(V, E)$ of a convective M-matrix M can be embedded into a triangulation, called the *induced triangular mesh*, then we can construct the diffusion and convection coefficients in Eq. (1.1) so that M is exactly the stiffness matrix of the box method discretization of the constructed partial differential equation using the induced triangular mesh. This is because that for each edge in $G(V, E)$, we can define the following two quantities (g_k, λ_k) as

$$g_k = \frac{a_k - b_k}{\ln a_k - \ln b_k}, \quad \lambda_k = \ln\left(\frac{a_k}{b_k}\right) \tag{4.15}$$

when $a_k > b_k$, and

$$\begin{aligned} g_k &= a_k = b_k, \\ \lambda_k &= 0, \end{aligned} \tag{4.16}$$

when $a_k = b_k$. It is clear that if the edge convection and the edge conductance scalars take these values in the box method discretization, then the edge stamps (3.3) used to assemble the stiffness matrix are exactly the same as the edge stamps (4.9) used to assemble M, and M can be written as

$$M = \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} B(-\lambda_{jl}) & -B(\lambda_{jl}) \\ -B(-\lambda_{jl}) & B(\lambda_{jl}) \end{bmatrix}_{jl}. \tag{4.17}$$

Notice that λ_k representing a generalized resistor is always nonnegative because $e_k = (j, l)$ is oriented in this way when the convection-directed graph was constructed.

We illustrate the above results with the following example.

Example 4.6 Suppose we are given a convective M-matrix

$$M = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ -1 & 5 & -3 & -3 & 0 \\ 0 & -2 & 9 & -2 & -1 \\ 0 & -1 & -4 & 6 & -1 \\ 0 & 0 & -2 & -1 & 3 \end{bmatrix}. \tag{4.18}$$

We can define the following edge stamps based on the off-diagonal entries of the matrix M as

$$\begin{bmatrix} 2 & -1 \\ -2 & 1 \end{bmatrix}_{21}, \quad \begin{bmatrix} 3 & -2 \\ -3 & 2 \end{bmatrix}_{32}, \quad \begin{bmatrix} 3 & -1 \\ -3 & 1 \end{bmatrix}_{42}, \quad \begin{bmatrix} 4 & -2 \\ -4 & 2 \end{bmatrix}_{34}, \quad \begin{bmatrix} 2 & -1 \\ -2 & 1 \end{bmatrix}_{35}, \quad \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}_{45}$$

for the edge set

$$E = \{(2,1), (3,2), (4,2), (3,4), (3,5), (4,5)\}. \tag{4.19}$$

Also notice the grounded edge $(5, *)$ can be represented by $[1]_5$, $[1, g_*]_{5*}$ or $[g_*, 1]_{*5}$, where $*$ and g_* do not appear in M but may appear in the right-hand-side of the system of linear equations. By examining the diagonal dominance for each column of the matrix M, we have

$$G^g = \text{diag}([0, 0, 0, 0, 1]). \tag{4.20}$$

Finally the generalized resistive network is illustrated in the following Figure 4.1.

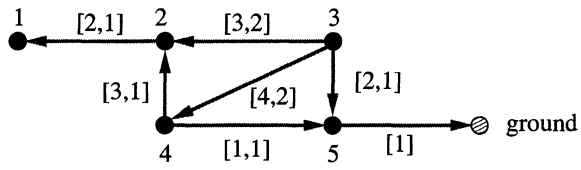

FIGURE 4.1 A generalized resistive network interpretation of the matrix M.

4.2. Discrete Curl-free Condition Revisit

Using the graph representation of a convective M-matrix that corresponds to a stiffness matrix arising from the box method discretization, we restate the *discrete curl-free condition* (3.13) as

$$B\lambda = 0, \tag{4.21}$$

where B is the cycle matrix of the graph G representing M, and

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix} \tag{4.22}$$

is the vector of the edge convection scalars as defined in Eq. (4.15). A convective M-matrix that satisfies the discrete curl-free condition is called a *conservative* convective M-matrix. Graph theory implies the following result.

THEOREM 4.7 *The discrete curl-free condition* (4.21) *holds if and only if there exists an n-dimensional vector* $\psi = [\psi_1, \ldots, \psi_n]^\top$, *such that*

$$A^\top \psi = \lambda, \tag{4.23}$$

where A is the incidence matrix for the convection-directed graph of M.

Proof Suppose there exists an n-dimensional vector $\psi = [\psi_1, \ldots, \psi_n]^\top$ and Eq. (4.23) holds. From the definitions of matrices A and B and basic graph theory, we have

$$AB^\top = A_T B_0^\top + A_C = 0, \tag{4.24}$$

$$BA^\top = B_0 A_T^\top + A_C^\top = 0. \tag{4.25}$$

Therefore, we have

$$B\lambda = BA^\top \psi = 0\psi = 0.$$

On the other hand, suppose condition (4.21) holds. We can solve the vector ψ in the following $n \times n$ system, which can be made upper triangular by reordering the edges in the spanning tree edge set E_T:

$$\begin{bmatrix} e_{1,n}^\top \\ A_T^\top \end{bmatrix} \psi = \begin{bmatrix} 0 \\ \lambda_1 \\ \vdots \\ \lambda_{n-1} \end{bmatrix} \tag{4.26}$$

Then by applying (4.25) in the first step, (4.26) in the second step, and (4.21) and (4.5) in the third step, we have

$$A_C^\top \psi = -B_0 A_T^\top \psi = -B_0 \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_{n-1} \end{bmatrix} = \begin{bmatrix} \lambda_n \\ \vdots \\ \lambda_m \end{bmatrix}. \tag{4.27}$$

Therefore, we have shown that Eq. (4.23) holds with the ψ obtained from (4.26). ■

Remark 4.8 The proof of Theorem 4.7 provides an algorithm to construct the vector ψ if the edge convection vector λ is given. The algorithm is simply backsolving Eq. (4.26), setting $\psi_0 = 0$.

COROLLARY 4.9 *The convection-directed graph defined above for a conservative convective M-matrix M is acyclic if there is at least one convective edge e_k, i.e., $\lambda_k > 0$, in each fundamental cycle.*

Proof We represent a cycle in the graph $G(M)$ by a vector c of length m in the following manner:

$$c_k = \begin{cases} 1 & \text{if edge } e_k \text{ is in the cycle and is directed counterclockwise along the cycle;} \\ -1 & \text{if edge } e_k \text{ is in the cycle and is directed clockwise along the cycle;} \\ 0 & \text{otherwise.} \end{cases} \tag{4.28}$$

Suppose there is a directed cycle in $G(M)$. Then its representation c must be either nonnegative or

nonpositive. Without loss of generality, we assume c is nonnegative. Since there is at least one convective edge in each fundamental cycle, there is at least one convective edge in the cycle that c represents, which implies that

$$c^\top \lambda > 0. \tag{4.29}$$

On the other hand, using the graph theory, we have that there exists a vector f of length p, the number of fundamental cycles, consisting of 0's and ± 1's, such that

$$c^\top = f^\top B. \tag{4.30}$$

Therefore, using Eq. (3.13), we have

$$c^\top \lambda = f^\top B\lambda = 0,$$

which contradicts (4.29), and hence the graph $G(M)$ is acyclic. ■

COROLLARY 4.10 *For a conservative convective M-matrix M, there exists a permutation P, such that the permuted matrix $[m'_{j'l'}] = PMP^\top$ has*

$$|m'_{j'l'}| \le |m'_{l'j'}| \tag{4.31}$$

for $j' < l'$, i.e., the large absolute entries are all in the lower triangle. A similar statement holds for a P' (the backward permutation) such that $[m'_{j'l'}] = P'MP'^\top$ has

$$|m'_{j'l'}| \ge |m'_{l'j'}| \tag{4.32}$$

for $j' < l'$.

Proof Since M is a conservative convective M-matrix, the convection-directed graph $G(V,E)$ is acyclic by Corollary 4.9. Using basic graph theory, there exists a topological ordering (labeling) of the nodes in the convection-directed graph G, such that, after all the nodes in G labeled according to such ordering, an edge $(j',l') \in E'$ implies that $j' < l'$. Let permutation matrix P represent the relabeling of the nodes in G, such that $Pe_j = e_{j'}$ and $Pe_l = e_{l'}$. Since the direction of an edge is defined by Eq. (4.3), which is independent of labeling, an edge $(j',l') \in E'$, or $(j,l) \in E$, is equivalent to

$$|m_{jl}| \le |m_{lj}|, \tag{4.33}$$

or Eq. (4.31). Therefore, Eq. (4.31) implies $j' < l'$.

Since the graph G is acyclic, there also exists a topological ordering of the nodes, such that an edge $(j',l') \in E'$ implies that $j' > l'$. Using this reordering, the permuted matrix has (4.31) implying $j' > l'$, or equivalently (4.32) implying $j' < l'$. ■

Based on Corollary 4.10, we say that the permutation of the original stiffness matrix $PMP^\top$ is *topologically ordered* if and only if its *upper-digraph*, the directed graph corresponding to the upper-triangle of $PMP^\top$, is *isomorphic* to the convection-directed graph G or its *transpose* $G^\top$. The transpose of a directed graph is a directed graph with the same nodes but reversed edge. Existence of topological orderings allows us to think of $j < l$ in (4.12) and (4.17), which is sometimes convenient as in §5.1 below.

Recall that consistent orderings play a role in SOR theory [36]. We give following relation between topological orderings and consistent orderings of a convective M-matrix M.

THEOREM 4.11 *If the convection-directed graph G of a convective M-matrix M is acyclic, then either all topologically ordered matrices $PMP^\top$ are consistently ordered or none of them are.*

Proof From the definition of topological ordering, it is clear that the upper-digraphs for all topological ordered matrices are either isomorphic to G or to $G^\top$. Since the cycle matrix of a directed graph is invariant under isomorphism, we can let B be the cycle matrix for the upper-digraphs isomorphic to G and $B^\top$ be the cycle matrix for those isomorphic to $G^\top$. Since the edges of $G^\top$ are exactly the reverse of those in G, we have $B^\top = -B$.

If one of the topologically ordered matrix is consistently ordered, then we have

$$B^\top e = Be = 0 \tag{4.34}$$

according to the theorem by Rose [27]. Applying the same theorem again, we have that all topologically ordered matrices are consistently ordered.

If one of the topologically ordered matrix is consistently ordered, then we have

$$-B^\top e = Be \neq 0. \tag{4.35}$$

Applying the theorem in [27] in the opposite direction, we have that no topologically ordered matrices are consistently ordered. ■

4.3. Nonnegative Divergence and Row Diagonal Dominance

At the continuous level, the convection-diffusion operator acting on a constant 1 is always nonnegative if and only if the divergence of the convection $\nabla \cdot \beta$ is nonnegative:

$$\nabla \cdot (-\alpha \nabla 1 + \beta 1) \equiv \nabla \cdot \beta \geq 0. \tag{4.36}$$

This leads us to define the following discrete version of the nonnegative divergence condition:

CONDITION 4.12 (Discrete Nonnegative Divergence)

$$d_j \equiv 2e_{j,n}^\top A_0 c \geq 0, \tag{4.37}$$

for $1 \leq j \leq n$. *Here* $c = [c_1, \ldots, c_m]^\top$, $c_k = g_k \lambda_k / 2$, *is the vector of edge convection elements defined in* (4.13).

Using (3.3) and the identity (A.4), it is easy to see that nonnegative discrete divergence (4.37) implies row diagonal dominance of M:

$$e_{j,n}^\top M e = 2e_{j,n}^\top A_0 c + \sum_{\substack{(j,l) \ni j \\ (j,l) \in E_0}} g_{jl} B(\lambda_{jl}) + \sum_{\substack{(l,j) \ni j \\ (l,j) \in E_0}} g_{lj} B(-\lambda_{lj}) \geq d_j. \tag{4.38}$$

The notation $(j,l) \ni j$ represents all (directed) edges (j,l) that are incident to node j. The discrete convection-divergence d_j in Eq. (4.37) can be interpreted as the discretization of the line integral

$$\oint_{\partial B_j} \nu^\top \beta ds \left(= \int_{B_j} \nabla \cdot \beta da \right) \tag{4.39}$$

along the boundary of box B_j, with ν the (outward) unit normal vector at the box boundary. In other words, we have

$$d_j = (\nabla \cdot \beta|_{p_j} + O(h)) \text{Area}(B_j), \tag{4.40}$$

which can be obtained using the following line of arguments (Eqs. (4.43)–(4.45)):

- Assuming $\nabla \cdot \beta$ is continuous in each box, we can apply the intermediate value theorem and have

$$\int\int_{B_j} \nabla \cdot \beta da = \nabla \cdot \beta|_{p'} \text{Area}(B_j) \tag{4.41}$$

for some $p' \in B_j$. Furthermore, if $\nabla \cdot \beta$ has finite derivatives in B_j, we have

$$\nabla \cdot \beta|_{p'} - \nabla \cdot \beta|_{p_j} = O(\text{diameter}(B_j)) = O(h), \tag{4.42}$$

or

$$\int\int_{B_j} \nabla \cdot \beta da - \nabla \cdot \beta|_{p_j} \text{Area}(B_j) = O(h) \text{Area}(B_j). \tag{4.43}$$

- Using Eq. (2.13) in Bank and Rose [4], we have

$$-\int\int_{\Omega} (\nabla v_j)^\top \beta da - \oint_{\partial B_j} \nu^\top \beta ds = O(h) \text{Area}(B_j). \tag{4.44}$$

- Using the same technique that was used to prove inequality (6.6) in Xu and Zikatanov [38], we can prove that

$$d_j + \int\int_{\Omega} (\nabla v_j)^\top \beta da = O(h) \text{Area}(B_j). \tag{4.45}$$

In general, when the convection vector field β in (1.1) fails to have nonnegative divergence, we

may rewrite the equation with a scaled variable v such that $u=\eta v$, $\eta \neq 0$, and have

$$\nabla \cdot (-\alpha \nabla(\eta v) + \beta \eta v) = f, \tag{4.46}$$

or

$$\nabla \cdot (-\tilde{\alpha} v + \tilde{\beta} v) = f \tag{4.47}$$

with

$$\tilde{\alpha} = \eta \alpha,$$
$$\tilde{\beta} = -\alpha \nabla \eta + \beta \eta.$$

Theoretically, we can find η

$$\nabla \cdot (-\alpha \nabla \eta + \beta \eta) \geq 0 \tag{4.48}$$

so that the new convection field $\tilde{\beta}$ has nonnegative divergence. In a discrete analogy, if we want to find a column scaling D for an M-matrix M such that MD is row diagonally dominant, we just need to find a vector $d > 0$ such that

$$Md \geq 0, \tag{4.49}$$

and let $D = \text{diag}(d)$.

When the scaled convection $b = \alpha^{-1}\beta$ is curl-free, then we only need to choose $\eta = e^{\psi}$. Similarly, when the convective M-matrix M satisfies the discrete curl-free condition, then we can choose $D = D_s^{-1}$ with D_s defined in (3.17).

Consider the one-dimensional model equation

$$\frac{d}{dx}\left(-\alpha \frac{du}{dx} + \beta u \right) = f \tag{4.50}$$

and suppose the curl-free condition

$$\frac{\beta}{\alpha} = -\frac{d\psi}{dx}. \tag{4.51}$$

According to Remark 3.11, Eq. (4.50) is equivalent to

$$\frac{d}{dx}\left(-e^{-\psi}\alpha \frac{dv}{dx} \right) = f, \quad v = e^{\psi} u, \tag{4.52}$$

and the discretization of (4.52) corresponds to the symmetric matrix S. The function e^{ψ} symmetrizes the differential operator.

But we can also expand (4.50) as

$$\frac{d}{dx}\left(-\alpha \frac{du}{dx} \right) + \frac{d\beta}{dx} u + \beta \frac{du}{dx} = f. \tag{4.53}$$

When $\alpha > 0$ and $(d/dx)\beta \geq 0$, the sum of the first two terms, say $\mathcal{L}_1$ of (4.53) gives rise to a symmetric and positive-definite operator which is perturbed by the first order operator $\beta(d/dx)u$. If the divergence condition $(d/dx)\beta \geq 0$ ($\nabla \cdot \beta \geq 0$, in general, see §4.3) fails, the discrete version of $\mathcal{L}_1$ can fail to be positive definite. The box discretization of (4.50) hides this, always giving rise to a convective M-matrix under the anisotropic Delaunay condition. Discretely, $\nabla \cdot \beta \geq 0$ implies that M is both irreducibly row diagonally dominant and column diagonally dominant. The $\nabla \cdot \beta \geq 0$ condition also arises as a sufficient condition to yield convective M-matrices in common "upwinding" schemes for (4.50); see (2.10) in [26].

5. CONVECTIVE ITERATION

In this section we discuss iterative methods for solving a discretized version of Eq. (1.1),

$$Mu = f, \tag{5.1}$$

with $M = [m_{jl}]$ defined in (2.43). We assume that the discretized problem has at least one Dirichlet node, and therefore, M is irreducibly column diagonally dominant, since, otherwise, the matrix M on the left-hand-side of (5.1) is singular. More generally we consider iterative methods for solving the system (5.1) when M is a convective M-matrix as defined in (4.12).

We assume that the reader is familiar with the basic theory of iterative methods of the form

$$Pu^{k+1} = Qu^k + f \tag{5.2}$$

where M is *split* as $M = P - Q$. For reference see Axelsson [1], Berman and Plemmons [7] and

Varga [36]. Recall that the *rate of convergence* of u^k in (5.2) to u of (5.1) is determined by

$$\rho(R) = \max|\lambda|, \tag{5.3}$$

where $R = P^{-1}Q$ and $Rv = \lambda v$; that is $\rho(R)$ is the maximum modulus of the set of eigenvalues of R. The sequence u^k converges to u if and only if $\rho(R) < 1$; the smaller that $\rho(R)$ is, the better the convergence. Furthermore,

$$R_\infty \equiv |\ln \rho| \tag{5.4}$$

is often called the asymptotic rate of convergence.

For any $M \in Z$ (Eq. (3.1)), M is an M-matrix if and only if any regular splitting(or weak-regular splitting) is convergent. For example, (P, Q) of (5.2) is a regular splitting, by definition, if $P^{-1} \geq 0$ and $Q \geq 0$. If $M = P - Q$ is a regular splitting of an M-matrix M, then we sometimes study the spectral radius of the matrix $H = M^{-1}Q$ which is related to $\rho(R)$ as

$$\rho(R) = \frac{\rho(H)}{1 + \rho(H)}; \tag{5.5}$$

see Varga [36], Theorem 3.13.

We will see that, under certain assumptions, it is possible to exploit convection to enhance convergence. By this statement we mean that usually $\rho(M_C) < \rho(M_D)$ where M_C corresponds to (1.1) with a nonzero convection ($\beta \neq 0$) and M_D corresponds to the same discretization but with $\beta = 0$ identically. To be more precise, we parameterize the scaled convection $b = \alpha^{-1}\beta$ in (1.1) as tb for $t \in [0, \infty)$. Therefore, the stiffness matrix M is parameterized by a single nonnegative number t which scales the edge convection scalar, λ, for all edge stamps, including the grounded ones as

$$M(t) = \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} B(-t\lambda_k) & -B(t\lambda_k) \\ -B(-t\lambda_k) & B(t\lambda_k) \end{bmatrix}_k. \tag{5.6}$$

Our analysis considerably extends, enhances, and corroborates the special case, asymptotic ($\beta \to \infty$) results in [37] by Wang and Xu. This is important since as $\beta \to \infty$, the discretization parameter h must have $h \to 0$ to insure that the discrete solution bears some resemblance to the true solution of the continuous problem (1.1). We show how the computational work W, amount of arithmetic operations, depends on the error in approximating the continuous solution by the discrete solution and the *scaled convection* $b = \alpha^{-1}\beta$, *i.e.*, we examine $W(\text{err}, b; q, d)$ as defined in (5.103). See Theorem 5.13 and Corollaries 5.14 and 5.15.

Our analysis requires some assumptions on the tensor field α and vector field β in (1.1). We require the scaled convection to satisfy $\nabla \times b = 0$ (curl$(b) = 0$) or, equivalently that $b = -\nabla\psi$ for some scalar potential function ψ. Furthermore, we require that the discretized problem preserves a discrete analog of this condition; see §3.4 Condition 3.8. Using this assumption, we have shown that the convection-directed graph $G(M)$ is acyclic (Corollary 4.9), and hence that there are (re-)orderings or permutations, P, such that the entries in the lower triangular part of PMP have special properties (Corollary 4.10). Interestingly and significantly the curl-free condition is equivalent to the existence of a diagonal matrix D_s (see (3.17)) such that $S = MD_s$ is symmetric (a Stieltjes matrix) as is (equivalently)

$$M_s = D_s^{-(1/2)} S D_s^{-(1/2)} = D_s^{-(1/2)} M D_s^{(1/2)}; \tag{5.7}$$

see Theorem 3.10. Also the nonnegative divergence condition on β (4.37) reappears in our iterative analysis, and, when both $\nabla \times b = 0$ and $\nabla \cdot \beta \geq 0$ hold, we obtain our most definitive result (Theorem 5.5).

Using Eq. (3.19) we derive a bound on $\rho(t)$ relative to $\rho(0)$ for the parameterization mentioned in (5.6), ρ being the spectral radius of the specific splitting under consideration. *This bound is dependent only on the ordering of the purely diffusive problem corresponding to* $\rho(0)$. Hence, given the curl-free condition, topological orderings seem to play no role; we have discussed them in §4.2 since they may play a more important role when M can no longer be symmetrized as above.

Our discussion on iterative solving deals almost exclusively with the Jacobi, Gauss–Seidel, and SOR point iterative splittings. We extend a block splitting example given by Wang and Xu [37]. Some generalizations to block methods, preconditioning analysis, incomplete factorization [6], multigrid interaction and other advanced iterative techniques are immediate or straightforward while others will be quite subtle. Similarly, while our analysis does not always assume $\nabla \cdot \beta \geq 0$, weakening the $\nabla \times b \equiv 0$ condition and carrying out the corresponding analyses and algorithmics are likely to be challenging.

5.1. Motivating Examples

Here we consider a simplified model problem

$$\nabla \cdot (-\nabla u + bu) = f \tag{5.8}$$

with constant b and zero Dirichlet boundary condition. The defining domain Ω is $[0,1]$ in one dimension and $[0,1] \times [0,1]$ in two dimensions.

5.1.1. Toeplitz Tridiagonal Case

Consider the box method discretization of Eq. (5.8) in one dimension. Suppose an equidistant mesh is used with $(n-1)$ points between 0 and 1. The stiffness matrix of the discretization is clearly in a Toeplitz tridiagonal form:

$$M = \frac{1}{h}\text{TriDiag}(-B^-, 2C, -B^+), \tag{5.9}$$

where $B^{\pm} = B(\pm hb)$ and $C = C(hb)$ are the Bernoulli functions (see Appendix A). Here we have implicitly used the topological ordering which is also a consistent ordering as we wrote down the stiffness matrix in the form of (5.9). Using Eq. (3.19), the stiffness matrix M can be symmetrized by

$$D_s = \text{Diag}(1, e^{-hb/2}, \ldots, e^{-(n-1)hb/2}) \tag{5.10}$$

as

$$M_s = D_s M D_s^{-1}. \tag{5.11}$$

The symmetrized stiffness matrix is a tridiagonal matrix in the form of

$$M_s = \frac{1}{h}\text{TriDiag}(-D, 2C, -D), \tag{5.12}$$

with $D = D(hb)$ defined in Appendix A. Now ordering does not matter. Suppose the length of the domain is 1, then we have $h = 1/n$. Consider the Gauss–Seidel iteration matrix $R_{\text{GS}} = P_s^{-1} Q_s$ for M_s (and also for M), with

$$P_s = \frac{1}{h}\text{TriDiag}(-D, 2C, 0),$$
$$Q_s = \frac{1}{h}\text{TriDiag}(0, 0, D).$$

Notice that the iteration matrices for M and for M_s are similar to each other. The eigenvalues and the eigenvectors of matrix R_{GS} can be computed exactly as

$$\lambda^k(R_{\text{GS}}) = [\cos(kh\pi)\text{sech}(hb/2)]^2, \tag{5.13}$$

$$w_{k,j} = [\cos(kh\pi)\text{sech}(hb/2)]^j \sin(jkh\pi) \tag{5.14}$$

by following Exercise 2.10, 2.11 in [9] or Remark 5.1. Here $w_{k,j}$ is the jth component the kth eigenvector corresponding to eigenvalue $\lambda^k(R_{\text{GS}})$. The hyperbolic function $\text{sech}(x)$ can be viewed as the ratio between $D(2x)$ and $C(2x)$, which is also called the Bernoulli E-function (see Appendix A).

Remark 5.1 The eigenvalue problem for the above Gauss–Seidel iteration matrix can be converted to the eigenvalue problem for the Gauss–Seidel iteration matrix of the standard tridiagonal matrix

$$M_0 = \text{tridiag}(-1, 2, -1)$$

using simple diagonal similarity transforms. Let $M_0 = P_0 - Q_0$ and $R_{\text{GS},0} = P_0^{-1} Q_0$ denote the Gauss–Seidel iteration for M_0. Define

$D_0 = \text{diag}([1, E(h\pi), \ldots, E^{n-1}(h\pi)])$. Then we have

$$P_s = C(hb)D_0P_0D_0^{-1},$$
$$Q_s = D(hb)E(hb)D_0Q_0D_0^{-1},$$

which implies that

$$P_s^{-1}Q_s = D_0P_0^{-1}D_0^{-1}D_0Q_0D_0^{-1}D(hb)E(hb)/C(hb)$$
$$= D_0P_0^{-1}Q_0D_0^{-1}E^2(hb). \quad (5.15)$$

Therefore, if w_0 is an eigenvector $R_{\text{GS},0}$ then D_0w_0 is an eigenvector of R_{GS}; if λ_0 is an eigenvalue of $R_{\text{GS},0}$ then $E^2\lambda_0$ is an eigenvalue of R_{GS}.

The spectral radius of the Gauss–Seidel iteration matrix R_{GS} is clearly

$$\rho(R_{\text{GS}}) = E^2(hb)\cos^2(h\pi). \quad (5.16)$$

On the other hand, the spectral radius of the Jacobi iteration matrix for a tridiagonal system is simply the square root of that of the Gauss–Seidel's:

$$\rho_{\text{J}} = E(hb)\cos(h\pi), \quad (5.17)$$

and the spectral radius of the SOR iteration matrix using the optimal over-relaxation factor is

$$\rho_{\text{SOR}} = \frac{1-\sqrt{1-\rho_{\text{J}}^2}}{1+\sqrt{1-\rho_{\text{J}}^2}}$$
$$= \frac{1-\sqrt{1-E^2(hb)\cos^2(h\pi)}}{1+\sqrt{1-E^2(hb)\cos^2(h\pi)}}. \quad (5.18)$$

We can compare the asymptotic rate convergence R_∞ defined in (5.4) as $h \to 0$ as a function of b. For Jacobi, Gauss–Seidel and SOR R_∞ increases, *i.e.*,

$$\lim_{h\to 0}\frac{R^{\text{C}}_{\infty,\text{J}}}{R^{\text{D}}_{\infty,\text{J}}} = \lim_{h\to 0}\frac{R^{\text{C}}_{\infty,\text{GS}}}{R^{\text{D}}_{\infty,\text{GS}}} = 1+\frac{b^2}{4\pi^2}, \quad (5.19)$$

$$\lim_{h\to 0}\frac{R^{\text{C}}_{\infty,\text{SOR}}}{R^{\text{D}}_{\infty,\text{SOR}}} = \sqrt{1+\frac{b^2}{4\pi^2}}, \quad (5.20)$$

where R^{C}_∞ denotes a convergence rate when solving the convective problem and R^{D}_∞ the corresponding purely diffusive problem. Notice that when the length of the defining interval is l, then simply replace b by lb in the above equations.

5.1.2. *Block Toeplitz Tridiagonal Case*

Next, we consider the box method with the use of an $(n-1)\times(n-1)$ square grid to the simplified model problem (5.8) in two dimensions with $b = [b, 0]^\top$ a constant vector field. Notice that convection only appears in horizontal edges, not in vertical edges. If we use a square grid as a mesh, then we obtain

$$M_s = T_C \otimes I + I \otimes T_D \quad (5.21)$$

if we use the crosswind blocking (vertical ordering) or

$$M_s = T_D \otimes I + I \otimes T_C \quad (5.22)$$

if we use the downwind (horizontal ordering) blocking. The $n\times n$ tridiagonal matrices T_C and T_D are defined as

$$T_C = \text{TriDiag}(-D, 2C, -D),$$
$$T_D = \text{TriDiag}(-1, 2, -1),$$

respectively. Since the convergence rates for the Gauss–Seidel and the optimal SOR methods can be derived from ρ_{J} using the consistent ordering theorem, it is sufficient that we only provide the spectral radius of the Jacobi iteration matrix. For crosswind blocking, we have for the block iterative methods

$$\rho_{\text{J}} = \frac{E(hb)\cos(h\pi)}{1+E(hb)(1-\cos(h\pi))}, \quad (5.23)$$

We again compare this with the convergence rate for corresponding purely diffusive problems, and have

$$\lim_{h\to 0}\frac{R^{\text{C}}_{\infty,\text{J}}}{R^{\text{D}}_{\infty,\text{J}}} = \lim_{h\to 0}\frac{R^{\text{C}}_{\infty,\text{GS}}}{R^{\text{D}}_{\infty,\text{GS}}} = 1+\frac{b^2}{8\pi^2}, \quad (5.24)$$

$$\lim_{h\to 0} \frac{R^{\mathrm{C}}_{\infty,\mathrm{SOR}}}{R^{\mathrm{D}}_{\infty,\mathrm{SOR}}} = \sqrt{1+\frac{b^2}{8\pi^2}}. \tag{5.25}$$

However, the downwind blocking scheme is much worse that the crosswind blocking scheme as shown in Figure 5.1, because it takes almost no advantage of the convection. We find

$$\rho(R_{\mathrm{J}}) = \frac{E(hb)\cos(h\pi)}{E(hb)+D(hb)(1-\cos(h\pi))} \tag{5.26}$$

$$\rho(R_{\mathrm{GS}}) = \left(\frac{E(hb)\cos(h\pi)}{E(hb)+D(hb)(1-\cos(h\pi))}\right)^2 \tag{5.27}$$

which implies that

$$\frac{R^{\mathrm{C}}_{\infty,\mathrm{J}}}{R^{\mathrm{D}}_{\infty,\mathrm{J}}} = \frac{R^{\mathrm{C}}_{\infty,\mathrm{GS}}}{R^{\mathrm{D}}_{\infty,\mathrm{GS}}} = \frac{R^{\mathrm{C}}_{\infty,\mathrm{SOR}}}{R^{\mathrm{D}}_{\infty,\mathrm{SOR}}} = 1. \tag{5.28}$$

Our analysis here corroborates that of Wang and Xu [37].

5.2. General Analyses

In this section, we bound the iteration spectral radius of basic iterative methods for general conservative convective M-matrix systems. We consider the parameterized edge assembly form of the parameterized stiffness matrix in Eq. (5.6). The symmetrized stiffness matrix (*cf.* (3.17) and (3.19)) is hence parameterized as

$$M_s(t) = D_s(t)^{-(1/2)} M(t) D_s(t)^{(1/2)} = \sum_{e_k\in E\cup E_0} g_k \begin{bmatrix} B(-t\lambda_k) & -D(t\lambda_k) \\ -D(t\lambda_k) & B(t\lambda_k) \end{bmatrix}_k, \tag{5.29}$$

with

$$D_s(t) = \mathrm{Diag}(e^{-t\psi_1},\ldots,e^{-t\psi_n}). \tag{5.30}$$

See §3.4 and §4.2 for construction of the convection potential ψ from the edge convection λ. Notice that $M(0)=M_s(0)$ represents the stiffness matrix for a purely diffusive problem – the model problem (1.1) with $\beta\equiv 0$ on the defining domain Ω. Also notice that $M(1)=M$ is the original stiffness matrix.

We start with some results for systems that have nonnegative convection-divergence (see Condition 4.12), and then consider two further results which relax this condition.

5.2.1. *Nonnegative Convection-divergence Case*

We shall see that under the nonnegative divergence condition (Condition 4.12), the spectral radii of both the Jacobi and the Gauss–Seidel iteration matrices decrease monotonically if β in the model Eq. (1.1) does not vanish identically, and

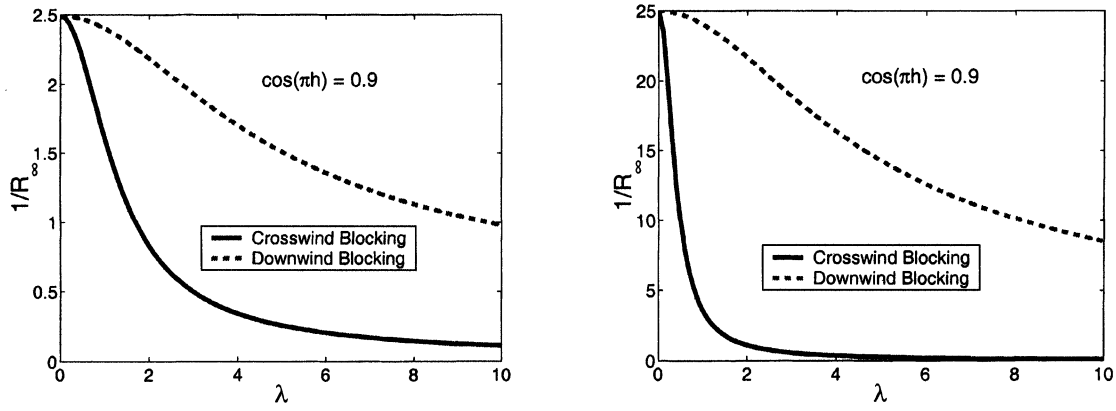

FIGURE 5.1 Comparing convergence rates for crosswind and downwind block-Gauss–Seidel.

furthermore, we derive a bound on the spectral radii that generalizes our analysis for the special Toeplitz tridiagonal case.

Let D_d denote the diagonal matrix of the discrete convection-divergence of the original stiffness matrix M:

$$D_d \equiv \mathrm{Diag}(d_1, \ldots, d_n) = \mathrm{diag}(Me); \tag{5.31}$$

cf. Eq. (4.37). The discrete convection-divergence for the parameterized stiffness matrix $M(t)$ is

$$D_d(t) = \mathrm{diag}(M(t)e) = tD_d, \tag{5.32}$$

and the edge assembly of $M(t)$ in Eq. (5.29) can be equivalently written as

$$M_s(t) = \frac{t}{2}D_d + \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} C(t\lambda_k) & -D(t\lambda_k) \\ -D(t\lambda_k) & C(t\lambda_k) \end{bmatrix}_k, \tag{5.33}$$

where the identity (A.5) is used. We further consider the edge assembly forms of the Jacobi and Gauss–Seidel splitting matrices and their derivatives with respect to the parameter t. Based on the form of M_s in (5.33), we have

$$P_{s,\mathrm{J}}(t) = \frac{t}{2}D_d + \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} C(t\lambda_k) & 0 \\ 0 & C(t\lambda_k) \end{bmatrix}_k, \tag{5.34}$$

$$Q_{s,\mathrm{J}}(t) = \sum_{e_k \in E} g_k D(t\lambda_k) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}_k; \tag{5.35}$$

$$P_{s,\mathrm{GS}}(t) = \frac{t}{2}D_d + \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} C(t\lambda_k) & 0 \\ -D(t\lambda_k) & C(t\lambda_k) \end{bmatrix}_k, \tag{5.36}$$

$$Q_{s,\mathrm{GS}}(t) = \sum_{e_k \in E} g_k D(t\lambda_k) \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}_k. \tag{5.37}$$

Using the fact that

$$\dot{C}(t\lambda) = \frac{C(t\lambda) - D^2(t\lambda)}{t}, \tag{5.38}$$

$$\dot{D}(t\lambda) = \frac{D(t\lambda)(1 - C(t\lambda))}{t}, \tag{5.39}$$

the derivative of the above splitting matrices can be written as

$$\dot{P}_{s,\mathrm{J}}(t) = \frac{D_d}{2} + \sum_{e_k \in E \cup E_0} \frac{g_k}{t} \begin{bmatrix} C(t\lambda_k) - D^2(t\lambda_k) & 0 \\ 0 & C(t\lambda_k) - D^2(t\lambda_k) \end{bmatrix}_k, \tag{5.40}$$

$$\dot{Q}_{s,\mathrm{J}}(t) = \sum_{e_k \in E} \frac{g_k}{t} D(t\lambda_k)(1 - C(t\lambda_k)) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}_k, \tag{5.41}$$

and similarly

$$\dot{P}_{s,\mathrm{GS}}(t) = \frac{D_d}{2} + \sum_{e_k \in E \cup E_0} \frac{g_k}{t} \begin{bmatrix} C(t\lambda_k) - D^2(t\lambda_k) & 0 \\ D(t\lambda_k)(C(t\lambda_k) - 1) & C(t\lambda_k) - D^2(t\lambda_k) \end{bmatrix}_k, \tag{5.42}$$

$$\dot{Q}_{s,\mathrm{GS}}(t) = \sum_{e_k \in E} \frac{g_k}{t} D(t\lambda_k)(1 - C(t\lambda_k)) \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}_k. \tag{5.43}$$

THEOREM 5.2 *If the convection-directed graph $G(M)$ for a given conservative convective M-matrix M satisfies the nonnegative discrete convection-divergence condition* (4.37), *or $D_d \geq 0$, and has at least one convective edge, say $\lambda_k \neq 0$ for some $e_k \in E \cup E_0$, then the spectral radii of the Jacobi and the Gauss–Seidel iterative methods for the parameterized system $M(t)u_h = f_h$ decrease monotonically as $t > 0$.*

Proof First notice that $M(t)$ and $M_s(t)$ are similar, and therefore, their Jacobi (Gauss–Seidel) iteration matrices have the same eigenvalues. Next, we consider the derivative of the iteration matrix as

$$\dot{R}_{s,\mathrm{J,GS}} \equiv \frac{d}{dt}(P^{-1}_{s,\mathrm{J,GS}} Q_{s,\mathrm{J,GS}}) = -P^{-1}_{s,\mathrm{J,GS}} \dot{P}_{s,\mathrm{J,GS}} R_{s,\mathrm{J,GS}} + P^{-1}_{s,\mathrm{J,GS}} \dot{Q}_{s,\mathrm{J,GS}}. \tag{5.44}$$

The derivative of its spectral radius can be written as

$$\dot{\rho}_{\mathrm{J,GS}} = \frac{y_{s,\mathrm{J,GS}}^{\top} \dot{R}_{s,\mathrm{J,GS}} x_{s,\mathrm{J,GS}}}{y_{s,\mathrm{J,GS}}^{\top} x_{s,\mathrm{J,GS}}}, \tag{5.45}$$

where $x_{s,\mathrm{J,GS}}$ and $y_{s,\mathrm{J,GS}}$ are the right and left eigenvectors for $\rho_{\mathrm{J,GS}}$ respectively for $R_{\mathrm{J,GS}}$. See Appendix B for detailed derivation of $\dot{\rho}$. From (5.44) and (5.45), we have

$$\dot{\rho}_{\mathrm{J,GS}} \leq -\frac{y_{s,\mathrm{J,GS}}^{\top} P_{s,\mathrm{J,GS}}^{-1} \dot{P}_{s,\mathrm{J,GS}} x_{s,\mathrm{J,GS}}}{y_{s,\mathrm{J,GS}}^{\top} x_{s,\mathrm{J,GS}}} \rho_{\mathrm{J,GS}}. \tag{5.46}$$

According to the Perron–Frobenius Theorem [36], Theorem 2.1 and [36], Exercise 6 on page 75, both $x_{s,\mathrm{J,GS}}$ and $y_{s,\mathrm{J,GS}}$ are vectors with all positive entries. Since $P_{s,\mathrm{J,GS}}$ is an irreducibly diagonally dominant M-matrix, $P_{s,\mathrm{J,GS}}^{-1}$ is nonnegative and has all positive diagonal entries. Since there is at least one convective edge in $G(M)$ and $D_d \geq 0$, the matrix $\dot{P}_{s,\mathrm{J,GS}}^{-1}$ is nonnegative and has at least one positive diagonal entry when $t > 0$; see Eq. (5.40) or (5.42). For $t > 0$, we see that

$$\dot{\rho}_{\mathrm{J,GS}}(t) < 0, \tag{5.47}$$

and therefore $\rho_{\mathrm{J,GS}}(t)$ decreases strictly monotonically. ■

Next, we strengthen the condition on convective edges such that each node is connected by at least one convective edges, or $\lambda^* > 0$ with λ^* defined as

$$\lambda^* \equiv \min_{1 \leq j \leq n} \max_{\substack{(j,l) \ni j \\ (j,l) \in E \cup E_0}} \lambda_{jl}. \tag{5.48}$$

THEOREM 5.3 *For a conservative convective M-matrix M, if $D_d \geq 0$ and $\lambda^* > 0$, then*

$$\rho_{J,GS}(t) \leq \rho_{J,GS}(0) C(\lambda^* t)^{-r}, \tag{5.49}$$

where

$$1 \geq r = \frac{1}{|Out|_{max} + |In|_{max}} \frac{g_{min}}{g_{max}} \frac{\lambda^*}{|\lambda|_{max}} > 0. \tag{5.50}$$

Proof It is easy to verify that

$$\dot{P}_{s,\mathrm{J,GS}} \geq \frac{1}{t} \frac{1}{|\mathrm{Out}|_{\max} + |\mathrm{In}|_{\max}} \frac{g_{\min}}{g_{\max}} \frac{C(\lambda^* t) - D^2(\lambda^* t)}{C(|\lambda|_{\max} t)} P_{s,\mathrm{J,GS}}$$

when $D_d \geq 0$. Using (5.46) and

$$\lambda_2 C(\lambda_1 t) \leq \lambda_1 C(\lambda_2 t) \tag{5.51}$$

for $\lambda_1 \geq \lambda_2 \geq 0$, we have

$$\begin{aligned} \dot{\rho}_{\mathrm{J,GS}} &\leq -\frac{1}{t} \frac{1}{|\mathrm{Out}|_{\max} + |\mathrm{In}|_{\max}} \frac{g_{\min}}{g_{\max}} \frac{C(\lambda^* t) - D^2(\lambda^* t)}{C(|\lambda|_{\max} t)} \rho_{\mathrm{J,GS}} \\ &\leq -\frac{r}{t} \frac{C(\lambda^* t) - D^2(\lambda^* t)}{C(\lambda^* t)} \rho_{\mathrm{J,GS}} \\ &= -r \frac{\dot{C}(\lambda^* t)}{C(\lambda^* t)} \rho_{\mathrm{J,GS}}. \end{aligned}$$

The proof is complete after we separate variables and integrate both sides in the above inequality. ■

Remark 5.4 Using Theorem 5.7 and Remark 5.9 below, the bound on the spectral radius $\rho_{\mathrm{J,GS}}$ can be improved as

$$\rho_{\mathrm{J,GS}}(t) \leq \min(\rho_{\mathrm{J,GS}}(0) C(\lambda^* t)^{-r}, C_{\mathrm{J,GS}}/t) \tag{5.52}$$

for some constant $C_{\mathrm{J,GS}}$ when $\lambda^* > 0$.

Finally for nonnegative convection divergence, we generalize our analysis on the special Toeplitz tridiagonal case by the following theorem.

THEOREM 5.5 *For a conservative convective M-matrix M, if its discrete convection-divergence is nonnegative, $D_d \geq 0$, then its Jacobi or Gauss–Seidel iteration spectrum radius is bounded as*

$$\rho_{J,GS}(t) \leq \rho_{J,GS}(0) E(|\lambda|_{min} t), \tag{5.53}$$

for $t \in [0, \infty)$. See Figure A.1 for a plot of the Bernoulli E-function.

Proof We first consider the derivative of the iteration matrix $R_{s,\mathrm{J}} = P_{s,\mathrm{J}}^{-1} Q_{s,\mathrm{J}}$ for the Jacobi method and the derivative of $R_{s,\mathrm{GS}} = P_{s,\mathrm{GS}}^{-1} Q_{s,\mathrm{GS}}$ for the Gauss–Seidel method. From Eqs. (5.40) and (5.41), we have

$$\dot{P}_{s,\mathrm{J}}(t) \geq \frac{1}{t}\left(1 - \frac{D^2(|\lambda|_{\min})}{C(|\lambda|_{\min})}\right) P_{s,\mathrm{J}}(t), \tag{5.54}$$

$$\dot{Q}_{s,\mathrm{J}}(t) \leq \frac{1 - C(t|\lambda|_{\min})}{t} Q_{s,\mathrm{J}}(t), \tag{5.55}$$

where $|\lambda|_{\min} \equiv \min_{e_k \in E \cup E_0} |\lambda_k|$. Similarly, from Eqs. (5.42) and (5.43), we have

$$\dot{P}_{s,\mathrm{GS}}(t) \geq \frac{1}{t}\left(1 - \frac{D^2(|\lambda|_{\min})}{C(|\lambda|_{\min})}\right) P_{s,\mathrm{GS}}(t), \tag{5.56}$$

$$\dot{Q}_{s,\mathrm{GS}}(t) \leq \frac{1 - C(t|\lambda|_{\min})}{t} Q_{s,\mathrm{GS}}(t). \tag{5.57}$$

Applying the form of $\dot{R}_{s,\mathrm{J,GS}}$ in Eq. (5.44), we have for both Jacobi and Gauss–Seidel that

$$\dot{R}_{s,\mathrm{J,GS}} \leq -\left[\frac{1}{t}\left(1 - \frac{D^2(|\lambda|_{\min} t)}{C(|\lambda|_{\min} t)}\right) + \frac{C(|\lambda|_{\min} t) - 1}{t}\right] R_{s,\mathrm{J,GS}}. \tag{5.58}$$

Furthermore, we apply the expression for the derivative of the spectral radius of the iteration matrix in Eq. (5.45) and have

$$\dot{\rho}_{\mathrm{J,GS}}(t) \leq -\left[\frac{1}{t}\left(1 - \frac{D^2(|\lambda|_{\min} t)}{C(|\lambda|_{\min} t)}\right) + \frac{C(|\lambda|_{\min} t) - 1}{t}\right] \rho_{\mathrm{J,GS}}(t) \tag{5.59}$$

for $t \geq 0$. Using

$$\frac{d}{dt}(\ln C(\lambda t)) = \frac{1}{t}\left(1 - \frac{D^2(\lambda t)}{C(\lambda t)}\right),$$

$$\frac{d}{dt}(\ln D(\lambda t)) = -\frac{C(\lambda t) - 1}{t},$$

which can be easily derived from Eqs. (5.38) and (5.39), we are able to separate variables (ρ and t) on both sides on inequality (5.59), then integrate both sides, and finally complete the proof. ■

Notice that the *bound* given by the above theorem decreases like the Bernoulli E-function if and only if $|\lambda|_{\min} > 0$, or all edges in the corresponding convection-directed graph are convective. If the convection vector field β does not vanish anywhere in the defining domain Ω, then a mesh can be constructed (adjusted) such that no mesh edges are orthogonal to the convection vector field, and hence the discretized system will meet this condition.

Further notice that the spectral radius bound for the Gauss–Seidel iteration method is not as tight as that for the Jacobi iteration method because the positive off-diagonal entries in $\dot{P}_{\mathrm{GS}}$ were simply discarded in the above proof. In the best situation, where the consistent orderings exist and are used, as we have seen in the Toeplitz and block Toeplitz cases, the spectral radius of the Gauss–Seidel iteration matrix decreases as $E^2(\lambda t)$. In more general situations, one may use the Jacobi bound provided by Theorem 5.5 in combination of Theorem 4.8 in Varga [36] to derive possibly a tighter bound for the Gauss–Seidel case.

5.2.2. *No Assumption on Divergence*

We start by proving the bounds on the 2-norm of the structural matrix of an undirected graph $G(V, E)$. Let Q_0 be the strictly upper-triangular matrix such that

$$Q_0(j,l) = \begin{cases} 1 & \text{if } (j,l) \in E \quad j < l, \\ 0 & \text{otherwise.} \end{cases} \tag{5.60}$$

Then we assign directions to the edges in G to make it the upper-digraph of $Q_0^{\top} + Q_0$, *i.e.*, assign edge (j,l) the direction from j to l if $j < l$; otherwise, assign the direction from l to j. Let Out_j and In_j denote the set of nodes that are connected with j by its out-edges and in-edges respectively.

LEMMA 5.6

$$\|Q_0^{\top} + Q_0\|_2 \leq \max_j (|Out_j| + |In_j|), \tag{5.61}$$

and

$$\|Q_0\|_2 \le \max_j(|Out_j|)\max_j(|In_j|). \tag{5.62}$$

Proof Since $Q_0^\top + Q_0$ is symmetric,

$$\|Q_0^\top + Q_0\|_2 = \rho(Q_0^\top + Q_0). \tag{5.63}$$

Then the first bound can be easily obtained by directly applying the Gershgorin Theorem [1].

For the second bound, we have

$$\|Q_0\|_2 = [\rho(Q_0^\top Q_0)]^{(1/2)}. \tag{5.64}$$

The (j,l) entry of the matrix $(Q_0^\top Q_0)$ is $|\mathrm{Out}_j|$ when $j=l$, and

$$\sum_{l\in \mathrm{Out}_j}(|\mathrm{In}_l| - 1). \tag{5.65}$$

Notice that each node in Out_j has at least one in-edge (j,l). Therefore, all entries of $(Q_0^\top Q_0)$ are non-negative, and the jth row sum can be written as

$$\begin{aligned}|\mathrm{Out}_j| + \sum_{l\in \mathrm{Out}_j}(|\mathrm{In}_l| - 1) &= \sum_{l\in \mathrm{Out}_j}|\mathrm{In}_l| \\ &\le \sqrt{\max_j(|\mathrm{Out}_j|)\max_j(|\mathrm{In}_j|)}.\end{aligned} \tag{5.66}$$

The second bound of the lemma is also proved by applying the Gershgorin Theorem. ∎

We show in the following that as the parameter t is sufficiently large, the convergence rate decreases as the magnitude of convection decreases uniformly in the domain. Here, we still assume curl-free condition on both the continuous and the discrete convection.

THEOREM 5.7 *Let M be a conservative convective M-matrix. We have*

$$\rho_J \le \min\left(1, C_J\frac{D(t|\lambda|_{min})}{F^*(t)}\kappa(M_s)\right) \tag{5.67}$$

$$\rho_{GS} \le \min\left(1, C_{GS}\frac{D(t|\lambda|_{min})}{F^*(t)}\kappa(M_s)\right), \tag{5.68}$$

where $F^*(t)$ is defined as

$$F^*(t) = \min_{1\le j\le n}\left(\max_{e_k\ni j} g_k F(t\lambda_k)\right), \tag{5.69}$$

and

$$F(t\lambda_k) = \begin{cases} C(\lambda_t) - D(\lambda_t) & \text{if } e_k \text{ is a non-grounded edge,} \\ C(\lambda_t)/2 & \text{if } e_k \text{ is a grounded edge.}\end{cases} \tag{5.70}$$

is yet another member of the family of Bernoulli functions (see Appendix A). The constants C_J and C_{GS} depend only on the mesh.

Proof Let H_J and H_{GS} denote $M_s^{-1}(Q_s^\top + Q_s)$ and $M_s^{-1}Q_s$ respectively. We have

$$\rho(H_J) \le \|H_J\|_2 \le \|M_s^{-1}\|_2\|Q_s^\top + Q_s\|_2. \tag{5.71}$$

We will give bounds on $\|M_s^{-1}\|_2$ and $\|Q_s^\top + Q_s\|_2$ respectively as follows. First we get

$$\|M_s^{-1}\|_2 \le \frac{\kappa(M_s)}{F^*(t)} \tag{5.72}$$

from the definition of the *condition number* of a matrix

$$\kappa(M_s) = \|M_s\|_2\|M_s^{-1}\|_2, \tag{5.73}$$

and

$$\begin{aligned}\|M_s\|_2 &\ge \frac{e^\top M_s e}{e^\top e} \\ &\ge \frac{1}{n}\sum_{e_k\in E\cup E_0} g_k e^\top \begin{bmatrix} B(-t\lambda_k) & -D(t\lambda_k) \\ -D(t\lambda_k) & B(t\lambda_k)\end{bmatrix}_k e \\ &= \frac{1}{n}\sum_{j=1}^{n}\sum_{e_k\ni j} g_k F(t\lambda_k) \\ &\ge g_{\min}F^*(t).\end{aligned} \tag{5.74}$$

Applying Lemma 5.6, we have

$$\begin{aligned}\|Q_s^\top + Q_s\|_2 &\le \left\|\sum_{e_k\in E} g_k D(t\lambda_k)\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}_k\right\|_2 \\ &\le g_{\max}D(t|\lambda|_{\min})(\mathrm{Out}_{\max} + \mathrm{In}_{\max}).\end{aligned} \tag{5.75}$$

Therefore, we can bound the spectral radius of H_J as

$$\rho(H_J) \leq \frac{g_{\max} D(t|\lambda|_{\min})(|\mathrm{Out}_{\max}| + |\mathrm{In}_{\max}|)}{g_{\min} F^*(t)} \kappa(M_s). \tag{5.76}$$

Assuming that the diffusion tensor is uniformly symmetric and positive definite, we see that

$$C_J = \frac{g_{\max}(|\mathrm{Out}_{\max}| + |\mathrm{In}_{\max}|)}{g_{\min}}, \tag{5.77}$$

can be controlled by the mesh. This bound also applies to ρ_J since

$$\rho_J = \frac{\rho(H_J)}{1 + \rho(H_J)} \leq \rho(H_J). \tag{5.78}$$

Applying the second part of Lemma 5.6, we obtain for the Gauss–Seidel iteration that

$$C_{GS} = \frac{\sqrt{|\mathrm{Out}_{\max}||\mathrm{In}_{\max}|}}{|\mathrm{Out}_{\max}| + |\mathrm{In}_{\max}|} C_J, \tag{5.79}$$

which certainly can be controlled by the mesh. The proof is complete by using the fact that both Jacobi and Gauss–Seidel splittings and regular splittings, and hence $\rho_J \leq 1$ and $\rho_{GS} \leq 1$. ■

Remark 5.8 When $t[\lambda_k]_{\min} \gg 1$, M_s tends to a diagonal matrix due to the Bernoulli D-function, and the condition number is roughly bounded by the ratio between the largest and the smallest diagonal elements of M_s.

$$\kappa(M_s) \approx \frac{|\mathrm{Out}_{\max}||\lambda|_{\max}}{\lambda^*}, \tag{5.80}$$

with λ^* defined in (5.48).

Remark 5.9 If we assume that all edge convection scalars uniformly bounded away from zero:

$$|\lambda_k| \geq \lambda_{\min} > 0, \tag{5.81}$$

then for $t \in [0, \infty)$, the spectral radius of the Jacobi or Gauss–Seidel iteration matrix for $M(t)$ decays exponentially as $t \to \infty$. This follows easily from the properties of the Bernoulli functions. Even if there are some zero-convection edges but $\lambda^* > 0$, the bound given by the above theorem still decays as $O(1/t)$ according to the property of function F.

We next study the behavior of the spectral radius of Jacobi or Gauss–Seidel related iteration matrices H_J and H_{GS} as $t \to 0$. First, we consider the derivative of the Gauss–Seidel spectral radius H_{GS} at $t = 0$. Using (5.33) and the fact that $\dot{C}(0) = \dot{D}(0) = 0$, we have

$$\dot{M}_s(0) = \frac{1}{2} D_d, \tag{5.82}$$

$$\dot{Q}_s(0) = 0, \tag{5.83}$$

which yields

$$\begin{aligned} \dot{H}_{GS}(0) &= -M_s^{-1}(0)\dot{M}_s(0)H_{GS}(0) + M_s^{-1}(0)\dot{Q}_s(0) \\ &= -\frac{1}{2} M_s^{-1}(0) D_d H_{GS}(0); \end{aligned} \tag{5.84}$$

$$\begin{aligned} \dot{\rho}(H_{GS}, 0) &= \frac{y^\top \dot{H}_{GS}(0) x}{y^\top x} \text{ (see Appendix B)} \\ &= -\frac{1}{2} \rho(H_{GS}, 0) y^\top M_s^{-1}(0) D_d x / (y^\top x). \end{aligned} \tag{5.85}$$

It is clear that when the discrete convection-divergence d_j is nonnegative for $1 \leq j \leq n$ and nonzero at least at one node, then $\dot{\rho}(H_{GS}) < 0$. However, when d_j are negative for all $1 \leq j \leq n$, then $\rho(H_{GS})$ increases as t increases from 0. Since, we know the spectral radius will eventually reach the exponential decay region when t is sufficiently large, there must exists a positive number t_0 such that $\rho(t) < \rho(0)$ for all $t > t_0$, since $\rho(t)$ is a continuous function. We will give a bound on the value of t_0 which will depend on the divergence of convection, the condition number of the stiffness matrix, and the mesh properties.

We first note the following lemma; the proof is straightforward.

LEMMA 5.10 *Let M_1 and M_2 be two nonsingular M-matrices with the same size. Then $M_1^{-1} \geq M_2^{-1}$ if and only if $M_1 \leq M_2$.*

THEOREM 5.11 *Suppose $\|\beta\|$ is uniformly bounded away from zero in the defining domain Ω. Then $H_{GS}(t) \leq H_{GS}(0)$ when $t \geq t_0$ with*

$$t_0 = max\left\{0, C_m \frac{(-\nabla \cdot \beta|)_{max} + O(h)}{\|\beta^*\|^2_{\alpha^{*-1}}}\right\},$$

where C_m is a mesh-controlled quantity, and α^* and β^* are evaluations of the diffusivity tensor and the convection vector field that are used to compute λ^* defined in (5.48).

Proof It is sufficient to find t_0 such that when $t \geq t_0$, $Q_s(t) \leq Q_s(0)$ and $M_s(t) \geq M_s(0)$, which will imply that $H_{GS}(t) \leq H_{GS}(0)$ using Lemma 5.10. First, we see that $Q_s(t) \leq Q_s(0)$ for any $t > 0$ since

$$Q_s(t) = \sum_{e_k \in E} g_k D(t\lambda_k) \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \leq Q_s(0). \quad (5.86)$$

For M_s, we have

$$\begin{aligned} M_s(t) &\geq \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} B(-t\lambda_k) & -1 \\ -1 & B(t\lambda_k) \end{bmatrix} \\ &= \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \\ &\quad + \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} B(-t\lambda_k) - 1 & 0 \\ 0 & B(t\lambda_k) - 1 \end{bmatrix} \\ &\geq M_s(0) + \mathcal{E}(t) \end{aligned}$$

with

$$\mathcal{E}(t) = \sum_{e_k \in E \cup E_0} g_k \begin{bmatrix} \frac{t\lambda_k}{2} + \frac{t^2\lambda_k^2}{12} & 0 \\ 0 & -\frac{t\lambda_k}{2} \end{bmatrix},$$

where we have used the Taylor's expansion of the Bernoulli function (see Appendix A). It is easy to verify that for each entry of the diagonal matrix $\mathcal{E}$ to be nonnegative, *i.e.*,

$$\mathcal{E}(j,j) = \frac{t}{12}\left[\sum_{l \in \text{In}_j \cup \text{Out}_j} 6 g_{jl}\lambda_{jl} t + \sum_{l \in \text{Out}_j} g_{jl}\lambda_{jl}^2 t^2\right] \geq 0, \quad (5.87)$$

it suffices to take

$$t \geq \frac{-6\sum_{l \in \text{In}_j \cup \text{Out}_j} g_{jl}\lambda_{jl}}{\sum_{l \in \text{Out}_j} g_{jl}\lambda_{jl}^2}. \quad (5.88)$$

Using Eq. (4.40), we can write the numerator of the right-hand-side fraction as

$$[-6\nabla \cdot \beta|_{p_j} + O(h)]\text{Area}(B_j). \quad (5.89)$$

On the other hand, the denominator can be bounded as

$$\sum_{l \in \text{Out}_j} g_{jl}\lambda_{jl}^2 t^2 \geq g_{\min}\lambda^{*2}, \quad (5.90)$$

where $\lambda^* = h^{*\top}\alpha^{*-1}\beta^*$ is defined in (5.48). Notice that

$$\begin{aligned} \lambda^{*2} &= ((\alpha^{*-(1/2)}h^*)^\top(\alpha^{*-(1/2)}\beta^*))^2 \\ &= \|h^*\|^2_{\alpha^{*-1}}\|\beta^*\|^2_{\alpha^{*-1}}\cos^2(\theta_{\alpha^{*-1}}(h^*, \beta^*)), \end{aligned} \quad (5.91)$$

where $\|\cdot\|_{\alpha^{*-1}}$ and $\theta_{\alpha^{*-1}}(*,*)$ are the generalized norm and generalized angle with respect to the induced inner product

$$(u, v)_{\alpha^{*-1}} \equiv u^\top \alpha^{*-1} v. \quad (5.92)$$

Therefore, for $M_s(t) \geq M_s(0)$, it suffices to take

$$t \geq \frac{(6(-\nabla \cdot \beta|_{p_j}) + O(h))\text{Area}(B_j)}{g_{\min}\|h^*\|^2_{\alpha^{*-1}}\|\beta^*\|^2_{\alpha^{*-1}}\cos^2(\theta_{\alpha^{*-1}}(h^*, \beta^*))} \quad (5.93)$$

for $1 \leq j \leq n$. It is clear that when the mesh is quasiuniform (the triangular mesh satisfies certain angle bounds), we have $0 < C_1 \leq \text{Area}(B_j)/\|h^*\|^2_{\alpha^{*-1}} \leq C_2$ for some constants C_1 and C_2.

Then it suffices to take

$$t \geq C_m \frac{(-\nabla \cdot \beta|)_{\max} + O(h)}{\|\beta^*\|^2_{\alpha^{*-1}}}, \tag{5.94}$$

where $C_m = 6C_2/(g_{\min} \cos^2(\theta_{\alpha^{*-1}}(h^*, \beta^*)))$ is a quantity that can be controlled by the mesh. Of course, if the $(-\nabla \cdot \beta|)_{\max}$ turns out to be sufficiently negative, we only need to take $t \geq 0$. ■

Notice that the value of t_0 largely depends on the characters of the diffusion and the convection coefficients. For $\nabla \cdot \beta \geq 0, t_0 = 0$. The value γ can be controlled by meshing techniques so it should be close to 2. Obviously, when the nonnegativity of the convection-divergence is given, then the spectral radius of the Gauss–Seidel iteration matrix decreases monotonically for $t \geq 0$. Furthermore, we have the following analysis for $\dot{H}_{GS}$ and $\dot{\rho}(H_{GS})$.

Remark 5.12 These bounds we obtained in Theorem 5.5 and Theorem 5.7 can be applied to the SOR methods in a similar manner using

$$\omega_b - 1 \leq \rho_{SOR} \leq \sqrt{\omega_b - 1} \tag{5.95}$$

with

$$\omega_b = \frac{2}{1 + \sqrt{1 - \rho_J^2}}. \tag{5.96}$$

See Varga [36], Theorem 4.9. For example, the bounds on ρ_{SOR} using Eq. (5.53) in Theorem 5.5 can be derived as

$$\rho_{SOR} \leq \left[\frac{1 - \sqrt{1 - \rho_J(0) E(|\lambda|_{\min} t)}}{1 + \sqrt{1 - \rho_J(0) E(|\lambda|_{\min} t)}} \right]^{(1/2)}. \tag{5.97}$$

5.3. Operation Count

We count the total number of operations, such as additions and multiplications, that are needed for (iteratively) solving linear systems arising from discretization of convection-diffusion equations. It is clear that the total number of operations is the product of the number of iterations and the work per iteration, or

$$W = mk. \tag{5.98}$$

The number of iterations k is roughly the reciprocal of the asymptotic convergence rate $R_\infty = -\ln \rho$ for given tolerance, *i.e.*,

$$k \propto \frac{|\ln \text{err}|}{|\ln \rho(b, h)|}. \tag{5.99}$$

The number of the operations per iteration is proportional to the number of nonzeros in the stiffness matrix, which, in turn, is proportional to the number of edges in the mesh, *i.e.*,

$$m \propto h^{-d} \tag{5.100}$$

with d the spatial dimensionality.

Since the discretization error depends on both the mesh size h and the scaled convection b (*cf.* [31], Appendix B), it is not appropriate to fix the mesh size while solving a spectrum of problems with different scaled convection. Our analysis that follows is with respect to a given tolerance that bounds both the discretization error and the iterative solving error. We first assume that the discretization error has the form of

$$\text{err} = h^q (C_1 + C_2 \|b\|), \tag{5.101}$$

for some positive constant numbers C_1, C_2 and q. We further assume that $q \geq 1$. Then in order to ensure that the discretization error is less than the tolerance we need to set the mesh size as

$$h = \left(\frac{\text{err}}{C_1 + C_2 \|b\|} \right)^{(1/q)}. \tag{5.102}$$

Therefore, we can express the total amount of work W in terms of err, b, q and d as

$$W(\text{err}, b; q, d) = \frac{|\ln \text{err}|}{|\ln \rho(\text{err}, b; q)|} \left(\frac{C_1 + C_2 b}{\text{err}} \right)^{(d/q)}, \tag{5.103}$$

while treating h as an implicit variable. If we consider the work for solving a convective problem relative to solving a corresponding diffusive problem, we have

$$\frac{W(\text{err}, \|b\|; q, d)}{W(\text{err}, 0; q, d)} = \frac{|\ln \rho(\text{err}, 0; q)|}{|\ln \rho(\text{err}, \|b\|; q)|} \left(1 + \frac{C_1}{C_2}\|b\|\right)^{(d/q)}. \tag{5.104}$$

It is obvious that the convective-diffusive ratio of the work is the product of the ratio of the number of iterations $k = C_k / |\ln(\rho)|$ and the ratio of the number of operations per iteration m. Given (5.102), the increase of $m(b)/m(0)$ is inevitable as b increases. The main advantage of using iterative methods for solving convective problems is stated in the following theorem and corollaries. Here, we only consider the Jacobi and the Gauss–Seidel iterative methods since similar results can be obtained for the SOR method using Remark 5.12.

THEOREM 5.13 *Assume the conditions in Theorem 5.5. The ratio of number of iterations $k(b)/k(0)$ is bounded above by a number which does not depend on the spatial dimensionality when $q \geq 1$ in* (5.101).

Proof Since

$$\frac{k(b)}{k(0)} = \frac{|\ln \rho(\text{err}, 0; q)|}{|\ln \rho(\text{err}, \|b\|; q)|} \tag{5.105}$$

and the numerator depends only on err and q, it is sufficient for us to show that $\rho(\text{err}, \|b\|; q)$ is uniformly bounded above by one for given err and q.

We see that the iteration spectral radius $\rho(h, b)$ depends on both the mesh size and the magnitude of convection. Using the bound given in Theorem 5.5, we have

$$\rho(h(\text{err}, \|b\|; q), b) \leq \rho(h(\text{err}, \|b\|; q), 0) E(C_h h(\text{err}, \|b\|; q)\|b\|),$$

where we assume that $|\lambda|_{\min} \geq C_h h \|b\|$ for some constant C_h. Notice that both terms on the right-hand-side of the above equation are always between zero and one. We pick an arbitrary number $b_0 > 0$. When $0 \leq \|b\| \leq b_0$, we have

$$\rho(h(\text{err}, \|b\|; q), b) \leq \rho(h(\text{err}, \|b\|; q), 0) \leq \max_{h_1 \leq h \leq h_2} \rho(h, 0)$$

where

$$h_1 = \left(\frac{\text{err}}{C_1 + C_2 b_0}\right)^{(1/q)},$$
$$h_2 = \left(\frac{\text{err}}{C_1}\right)^{(1/q)}.$$

It is clear that the bound $\max_{h_1 \leq h \leq h_2} \rho(h, 0)$ is bounded above by one, and it only depends on the choice of b_0 and the iteration spectral radius of the diffusive problem as well as err and q. When $\|b\| \geq b_0$, we have

$$\rho(h(\text{err}, \|b\|; q), b) \leq E(C_h h(\text{err}, \|b\|; q)\|b\|) \leq E(C_h h(\text{err}, b_0; q) b_0) \tag{5.106}$$

since $h(\text{err}, \|b\|; q)\|b\|$ monotonically increases as $\|b\|$ increases for $q \geq 1$. From the property of the Bernoulli-E function, it is clear that $E(C_h h(\text{err}, b_0; q)b_0)$ is strictly less than one, and it only depends on b_0, err, and q. Since the spectral radius (ρ) does not depend on the spatial dimensionality, it is clear that the bound does not either. ■

Examining (5.106) in view of the properties of the Bernoulli-E function gives a more precise result as follows.

COROLLARY 5.14 *Assume the condition of Theorem 5.5. We have*

$$\frac{k(b)}{k(0)} \leq C\|b\|^{(1/q)-1}, \tag{5.107}$$

for $q > 1$ and $\| b \| \geq b_0$ with

$$b_0 = max\left(\frac{C_1}{C_1 + C_2}, \left(\frac{2 \ln 2}{C_h} \right)^{(q/(q-1))} \left(\frac{C_1 + C_2}{err} \right)^{(1/(q-1))} \right). \quad (5.108)$$

The constant C in (5.107) depends only on err, q and the iteration spectral radius of the diffusive problem. In other words, the convective-diffusive ratio of number of iterations decreases (to zero) as $\| b \|$ increases (to infinity) for $q > 1$.

Proof From Eq. (5.102), we have

$$h \geq \left(\frac{\text{err}}{C_1 + C_2} \right)^{(1/q)} \| b \|^{-(1/q)}, \quad (5.109)$$

when $\| b \| \geq b_0 \geq C_1/(C_1 + C_2)$ as in (5.108). Hence, we have

$$\begin{aligned} \rho(h(\text{err}, \|b\|; q), b) &\leq E(C_h h(\text{err}, \|b\|; q) \|b\|) \\ &\leq E\left(C_h \left(\frac{\text{err}}{C_1 + C_2} \right)^{(1/q)} \|b\|^{1-(1/q)} \right) \end{aligned} \quad (5.110)$$

Since $E(x) \leq 2e^{-x/2} \leq 1$ for $x \geq 2 \ln 2$, and $\| b \| \geq b_0$ as in (5.108), we have from (5.110)

$$\begin{aligned} |\ln \rho(h(\text{err}, \|b\|; q), b)| &\geq \frac{C_h}{2} \left(\frac{\text{err}}{C_1 + C_2} \right)^{(1/q)} \|b\|^{1-(1/q)} - \ln 2 \\ &\geq C'(\text{err}, q) \|b\|^{1-(1/q)} \end{aligned} \quad (5.111)$$

with

$$C'(\text{err}; q) = \frac{C_h}{2} \left(\frac{\text{err}}{C_1 + C_2} \right)^{(1/q)} + \frac{\ln 2}{b_0}. \quad (5.112)$$

Therefore, using (5.106) the constant in Eq. (5.107) is written as

$$C = \frac{|\ln \rho(h(\text{err}, 0; q), 0)|}{C'(\text{err}; q)}, \quad (5.113)$$

which clearly depends only on err, q, and $\rho(h(\text{err}, 0; q), 0)$, the iteration spectral radius of the diffusive problem. ■

COROLLARY 5.15 *For $q > 1$ and $\| b \| \geq b_0$ as in (5.108), the convective-diffusive ratio of the total work can be bounded as*

$$\frac{W(err, \|b\|; q, d)}{W(err, 0; q, d)} \leq C_W \|b\|^{((d+1)/q)-1}, \quad (5.114)$$

where C_W depends only on err, q, and $\rho(h(err, 0; q), 0)$.

Proof Using (5.104) and (5.107), we have

$$\begin{aligned} \frac{W(\text{err}, \|b\|; q, d)}{W(\text{err}, 0; q, d)} &\leq C \|b\|^{1-(1/q)} \left(1 + \frac{C_1}{C_2} \|b\| \right)^{(d/q)} \\ &\leq C \|b\|^{1-(1/q)} C'' \|b\|^{(d/q)}, \end{aligned} \quad (5.115)$$

where $C'' = 1/b_0 + C_1/C_2$ for $\| b \| \geq b_0, b_0$ defined in (5.108). It is clear that $C_W = CC''$, which only depends on err, q, and $\rho(h(\text{err}, 0; q), 0)$. ■

Remark 5.16 If sparse Gaussian elimination techniques are used to factor $M = LU$, symmetric reordering theory applies because M is structurally symmetric and column diagonally dominant (pivoting not necessary). As $\| b \|$ increases $n(\| b \|)/n(0) = O(\| b \|^{1/q})$ from (5.102). This implies that in two spatial dimensions, $W_D(\| b \|)/W_D(0)$, W_D the work of the sparse LU factorization, must be at least $W_D(\| b \|)/W_D(0) = O(\| b \|^{3/2})$ in the best case $q = 2$ for grid-like graphs. The iterative result of Corollary 5.15 is $W(\| b \|)/W(0) = O(\| b \|^{1/2})$. See Rose [28], Hoffman, Martin and Rose [19], and George and Liu [17].

For the following plots on the convective-diffusive ratios of k, m and W, we assume that

- The discretized system satisfies the discrete curl-free condition (Condition 3.8);
- The discretized system satisfies the nonnegative discrete divergence condition (Condition 4.12);

- All edges are convective in the corresponding convection-directed graph:

$$|\lambda|_{\min} = C_h h \|b\|; \tag{5.116}$$

- The spectral radii of the iteration matrices for a purely diffusive discretized system are

$$\rho_{J,0} = \cos(2\pi h), \tag{5.117}$$

$$\rho_{GS,0} = \cos^2(2\pi h). \tag{5.118}$$

Therefore, we may use the bound given by Theorem 5.5 and Eq. (5.97) and have

$$\rho_J(\text{err}, b; q) \le \cos(2\pi h(\text{err}, b; q)) \; E(C_h h(\text{err}, b; q)\|b\|) \tag{5.119}$$

$$\rho_{GS}(\text{err}, b; q) \le \cos^2(2\pi h(\text{err}, b; q)) \; E(C_h h(\text{err}, b; q)\|b\|) \tag{5.120}$$

$$\rho_{SOR}(\text{err}, b; q) \le \left[\frac{1-\sqrt{1-\cos(2\pi h(\text{err},b;q))E(C_h h(\text{err},b;q)\|b\|)}}{1+\sqrt{1-\cos(2\pi h(\text{err},b;q))E(C_h h(\text{err},b;q)\|b\|)}}\right]^{(1/2)}. \tag{5.121}$$

The following figures illustrate the dependencies of k, m, and W on the norm of the scaled convection $\| b \|$ at different values of err, q, and d, where we set the constants $C_h = 0.1$ and $C_1 = C_2 = 1$. Note that Figure 5.2 corroborates Theorem 5.13 and Corollary 5.14, and Figures 5.3 to 5.5 corroborate Corollary 5.15.

6. CONCLUDING REMARKS

In this section, we comment briefly on some aspects of our work which flow beyond the mainstream of our results.

Remark 6.1 The constant-j box method has been successfully extended to the investigation of wavefront propagation in anisotropic cardiac tissue; see [33]. By using a triangulated mesh that is formed with consideration of the underlying material properties (anisotropic Delaunay triangulation), the M-matrix properties are preserved, avoiding the possibility of spurious spikes or any non-monotone behavior associated with traditional finite differencing on quadrilateral meshes with variable anisotropy. The governing equation for modeling wavefront conduction is typically formulated by considering the current flow within the tissue, either intra-cellular or extra-cellular, as arising only from spatial gradients in the potential (∇u) created by local changes in membrane ionic fluxes. In other words, only the drift term is considered in the description of current flow, namely

$$j = -\alpha \nabla u \tag{6.1}$$

as any variation of ion concentration inside or outside the cell is considered to be negligibly small. For diseased myocardium, it is possible to have local accumulation of ions near an injury leading to spatial gradients in ion concentrations. These concentration variations will also contribute to the overall ion flux and will need to be explicitly taken into account in the model. The advantage of using the box scheme described in this work is that any reformulation of the model equations to handle both diffusion and drift of ions can be incorporated in a straightforward manner with no loss in computational performance.

The drift-diffusion equations arise more naturally when modeling ion fluxes at the individual channel level; see Eisenberg [13], and Gardner, Jerome and Eisenberg [15]. In conventional heart modeling, the channel currents are lumped to create a macroscopic decryption of the current kinetics. There are, however, a number of pharmacological or even genetic therapies for wavefront anomalies (arrhythmias) that will need to target at the level of the ion channel. The effects of these drugs on the overall macroscopic models may be hard to predict unless the models themselves are initially constructed from a channel perspective. This construction process may involve solving both the Poisson and drift-diffusion equations in three dimensions over the relevant

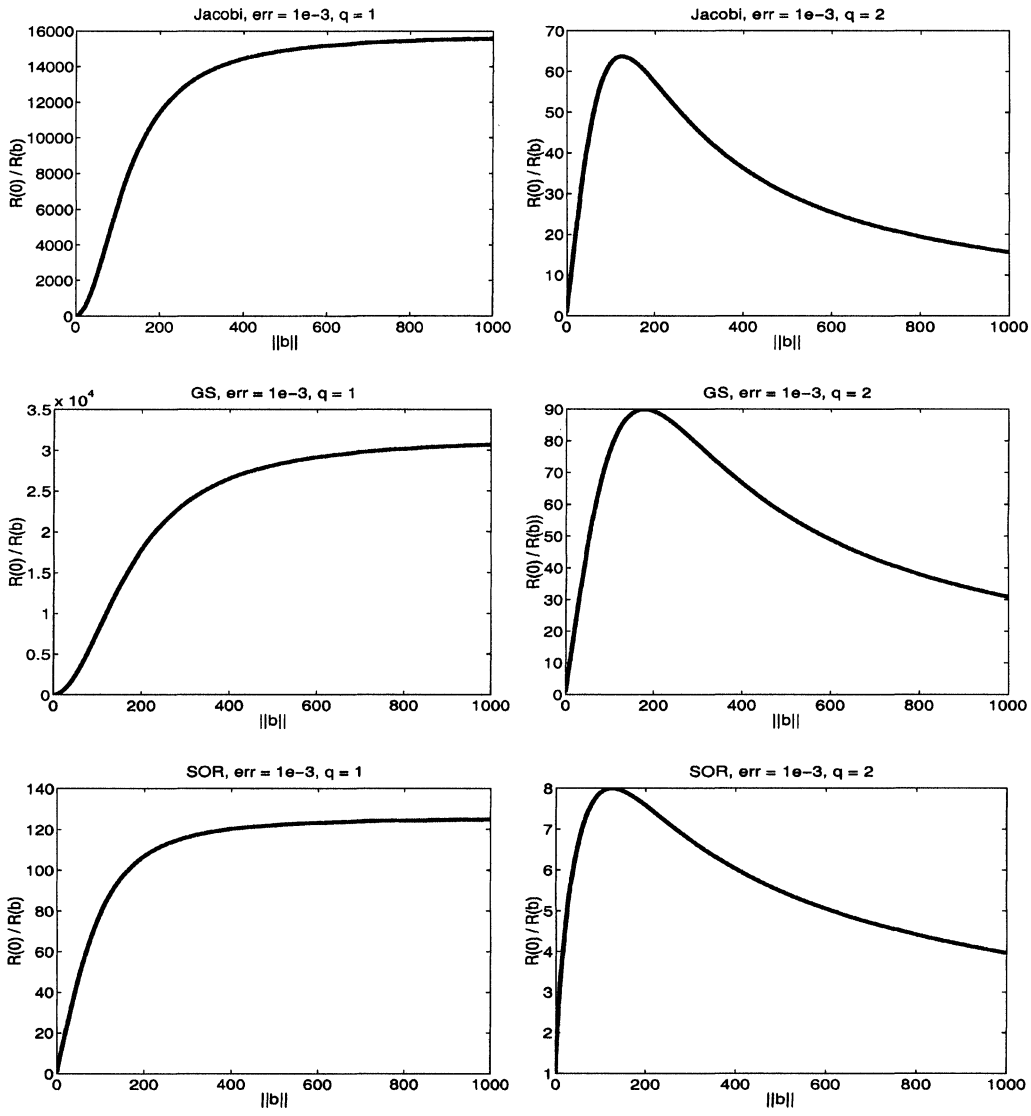

FIGURE 5.2 Upper bounds on convective-diffusive ratio of the numbers of iterations k, $R = |\ln(\rho)|$.

portion of a cell. The governing equations are analogous to those used in modeling semiconductor devices. The results from this work suggest a robust numerical approach that will enable the design of a drug or gene therapy and its action on a given cell much in the same way that one might design a specific portion of a semiconductor element. This "re-engineering" of the cardiac myocyte may lead to the equivalent of creating a semiconductor heart.

Remark 6.2 We note that the constant-j box method discretization paradigm presented here is not limited to triangular meshes. Indeed, the edge-pair constant-j assumption is easily seen to apply to quadrilateral meshes, for example. Even the "terminating line" quadrilateral meshes as in Selberherr [30], page 177, present no fundamental difficulties. For example, once the box is chosen as in Figure 6.1 (adapted from [30], Fig. 6.2–3), edge-pairs are then chosen to determine how to

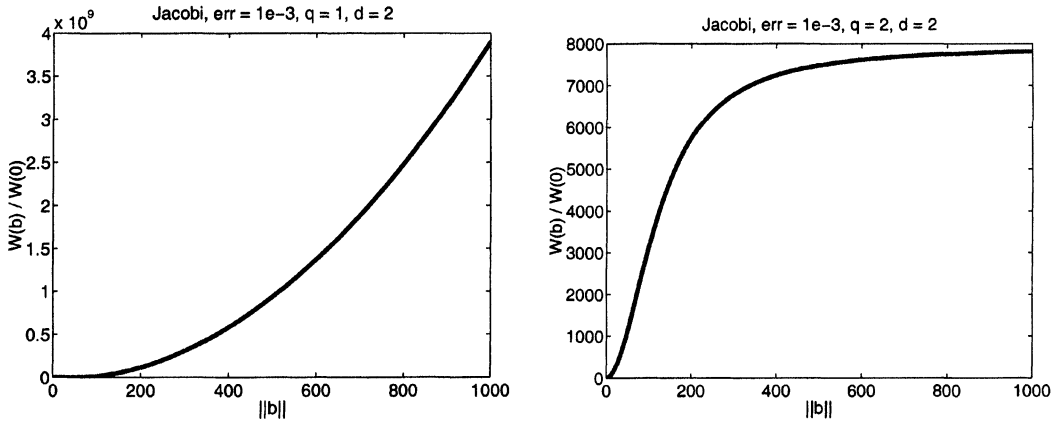

FIGURE 5.3 Convective-diffusive ratio of total operations W, Jacobi.

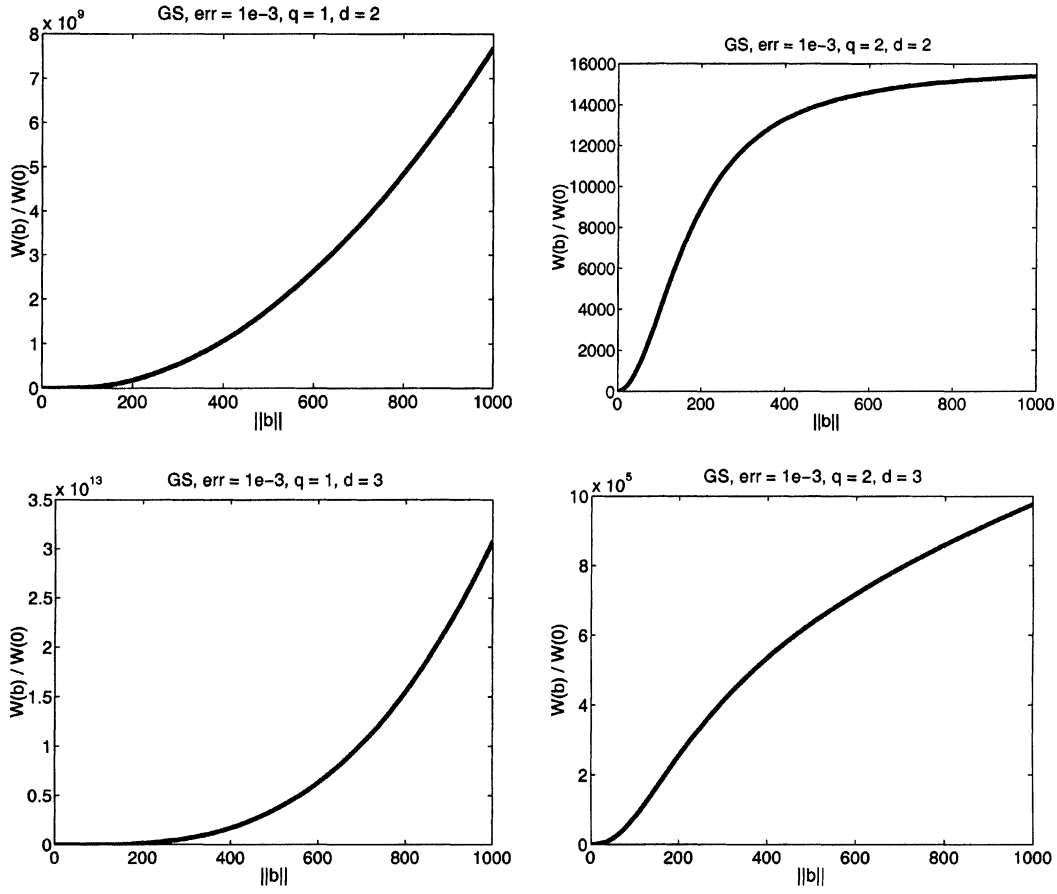

FIGURE 5.4 Convective-diffusive ratio of total operations W, Gauss–Seidel.

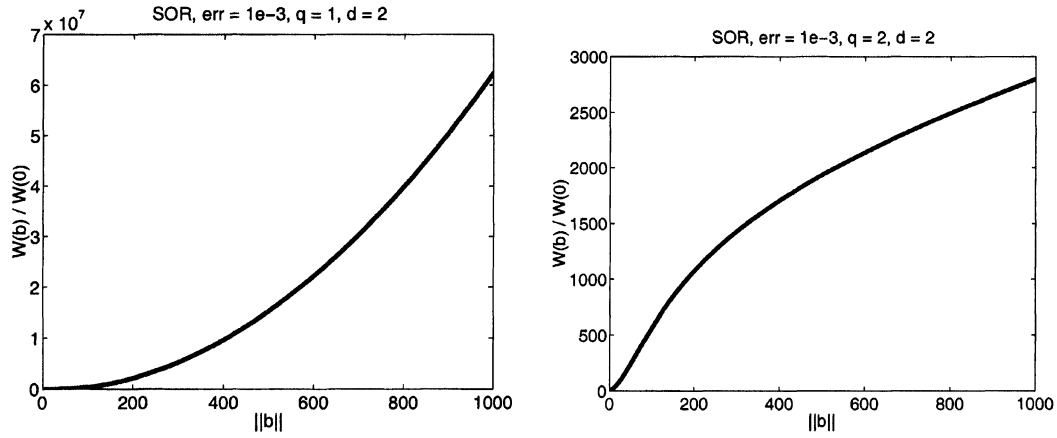

FIGURE 5.5 Convective-diffusive ratio of total operations W, SOR.

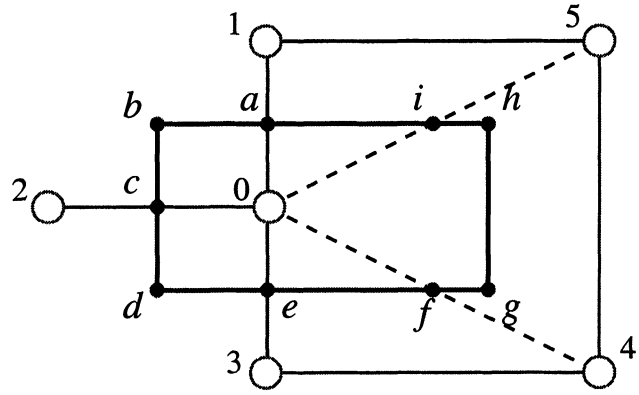

FIGURE 6.1 Terminating line mesh point with two standard edge-pairs (and box segments) and three non-standard edge-pairs (and box segments).

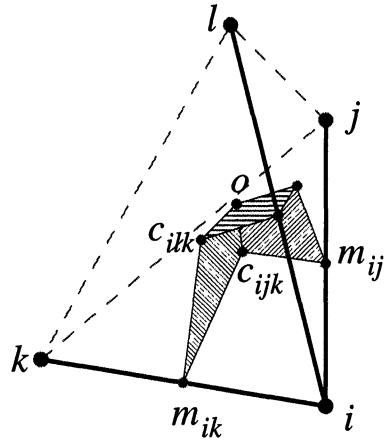

FIGURE 6.2 An edge-triple $((i,j),(i,k),(i,l))$, with m's the edge midpoints, c's the triangle circumcenters and o the tetrahedral circumcenter.

compute j on various parts of the box boundary. For example, edge-pair $((0,1),(0,2))$ is used to compute flux through boundary part $\overline{ab} \cup \overline{bc}$, edge-pair $((0,2),(0,3))$ for boundary part $\overline{cd} \cup \overline{de}$, edge-pair $((0,3),(0,4))$ for boundary part $\overline{ef}$, edge-pair $((0,4),(0,5))$ for boundary part $\overline{fg} \cup \overline{gh} \cup \overline{hi}$, and edge-pair $((0,5),(0,1))$ for boundary part $\overline{ia}$. (Some of the edges of these edge-pairs, *e.g.*, $(0,4)$ and $(0,5)$, may not be part of the original quadrilateral mesh.) This provides an alternative to using finite difference and allows flexibility in both the choice of the box for an unknown and the corresponding edge-pairs. Extension to 3D meshes is also straightforward in principle. Of course, the special properties of the stiffness matrix are no longer guaranteed.

Remark 6.3 Formally, extending the constant-j box method to three dimensions is straightforward where the two-dimensional edge-pairs are replaced by edge-triples (Fig. 6.2) which will be used to derive the local constant-j's. In three dimensions, box vertices need to be specified for all edges, triangles, and tetrahedra in the mesh. The edge conductance values will depend on local tetrahedron vertices and box vertices, as well as evaluation of the 3×3 diffusion tensor. Unlike the two-dimensional case, the box-vertex dependency does not naturally vanish in the expression for edge conductance. Finding a clean mesh condition for the M-matrix discretization property is unresolved when diffusion is anisotropic and varies over space. As in 2D, these conditions may be needed to serve as guidelines for mesh generation in three dimensions. However, in an isotropic case, *i.e.*, $\alpha = a(x,y,z)I_{3\times 3}$, the M-matrix property of the stiffness matrix can be obtained when the mesh is a Delaunay tetrahedralization.

References

[1] Axelsson, O., *Iterative Solution Methods*. Cambridge University Press, 1996.

[2] Bank, R. E., Bügler, J. F., Fichtner, W. and Smith, R. K. (1990). Some upwinding techniques for finite element approximations of convection-diffusion equations. *Numerische Mathematik*, **58**, 185–202.

[3] Bank, R. E., Coughran, W. M. and Cowsar, L. C. (1998). Analysis of the finite volume Scharfetter–Gummel method for steady convection-diffusion equations. *Computing and Visualization in Science*, **1**(3), 123–136.

[4] Bank, R. E. and Rose, D. J. (1987). Some error estimates for the box method. *SIAM J. Numerical Analysis*, **24**(4), 777–787.

[5] Bank, R. E., Rose, D. J. and Fichtner, W. (1983). Numerical methods for semiconductor simulation. *SIAM J. Scientific and Statistical Computing*, **4**, 416–435.

[6] Bank, R. E. and Smith, R. K. (1999). The incomplete factorization multigraph algorithm. *SIAM J. Scientific Computing*, **20**(4), 1349–1367.

[7] Berman, A. and Plemmons, R. J., *Nonnegative matrices in the mathematical sciences*. Computer Science and Applied Mathematics. Academic Press, 1979.

[8] Bossen, F. J. and Heckbert, P. S., A pliant method for anisotropic mesh generation. In: *Proc. 5th International Meshing Roundtable*, pp. 63–76. Sandia National Laboratories, 1996.

[9] Briggs, W. L. (1987). A Multigrid Tutorial. *SIAM*.

[10] Ciarlet, P. G., *The finite element method for elliptic problems*. Studies in mathematics and its applications. North-Holland, 1978.

[11] D'Azevedo, E. F., Optimal triangular mesh generation by coordinate transformation. *SIAM J. Scientific and Statistical Computation*, **12**(4), 755–786, July, 1991.

[12] Deo, N., *Graph Theory with Applications to Engineering and Computer Science*. Prentice-Hall, 1974.

[13] Eisenberg, R. S. (1996). Computing the field in proteins and channels. *J. Membrane Biology*, **150**, 1–25.

[14] Fichtner, W., Rose, D. J. and Bank, R. E. (1983). Semiconductor device simulation. *SIAM J. Scientific and Statistical Computing*, **4**, 391–415.

[15] Gardner, C. L., Jerome, J. W. and Eisenberg, R. S., *Electrodiffusion model of rectangular current pauses in ionic channels of cellular membranes*. Manuscript, 1999.

[16] Gatti, E., Micheletti, S. and Sacco, R. (1998). A new galerkin framework for the drift-diffusion equation in semiconductors. *East-West J. Numer. Math.*, **6**, 101–135.

[17] George, A. and Liu, J. W., *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, 1981.

[18] Hackbusch, W. (1989). On first and second order box schemes. *Computing*, **41**, 277–296.

[19] Hoffman, A. J., Martin, M. S. and Rose, D. J. (1973). Complexity bounds for regular finite difference and finite element grids. *SIAM J. Numerical Analysis*, **10**, 364–369.

[20] Idelsohn, S. R. and Oñate, E. (1994). Finite volumes and finite elements: two 'good friends'. *International J. Numerical Methods in Engineering*, **37**, 3323–3341.

[21] Ikeda, T., Maximum Principle in Finite Element Models for Convection-Diffusion Phenomena, Volume 4 of *Lecture Notes in Numerical and Applied Analysis*. North-Holland/Kinokuniya, 1983.

[22] Kahan, W. (1957). The rate of convergence of the extrapolated Gauss–Seidel iteration. *J. ACM*, **4**, 521–522.

[23] Kahan, W., Gauss–Seidel methods for solving large systems of linear equations. *Ph.D. Thesis*, University of Toronto, 1958.

[24] Marsen, J. E. and Tromba, A. J., *Vector Calculus*. Freeman, W. H. and Company, 3rd edition, 1988.

[25] Parter, S. V. and Youngs, J. W. T. (1962). The symmetrization of matrices by diagonal matrices. *J. Mathematical Analysis and Applications*, **4**, 102–110.

[26] Roose, H. G., Stynes, M. and Tobiska, L., *Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion and Flow Problems.* Springer Series in Computational Mathematics. Springer Verlag, 1996.
[27] Rose, D. J., A note on consistent ordering and zero circulation. *J. Association for Computing Machinery*, **18**(4), 573–575, October, 1971.
[28] Rose, D. J., A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations. In: Read, R. C. Ed. *Graph Theory and Computing.* Academic Press, New York, 1972.
[29] Scharfetter, D. and Gummel, H. (1973). Large-signal analysis of a silicon read diode oscillator. *IEEE Solid-state Electron*, **16**, 64–77.
[30] Selberherr, S., *Analysis and Simulation of Semiconductor Devices.* Springer-Verlag, Wien, New York, 1984.
[31] Shao, H., Meshing and discretization for divergence form partial differential equations in two spatial dimensions. *Master's Thesis*, Department of Computer Science, Duke University, 1997.
[32] Shao, H., *Numerical Analysis of Meshing and Discretization for Anisotropic Convection-Diffusion Equations. Ph.D. Thesis*, Department of Computer Science, Duke University, 1999.
[33] Shao, H., Henriquez, C. S. and Rose, D. J. (1999). The box method for discretization of anisotropic elliptic operators in heart modeling equations. *Computing and Visualization in Science, To be submitted.*
[34] Simpson, R. B. (1994). Anisotropic mesh transformations and optimal error control. *Applied Numerical Mathematics*, **14**, 183–198.
[35] Strikwerda, J. C., *Finite Difference Schemes and Partial Differential Equations.* Chapman & Hall, 1989.
[36] Varga, R. S., *Matrix Iterative Analysis.* Prentice-Hall, 1962.
[37] Wang, F. and Xu, J. (1999). A cross-wind-block iterative method for convection-dominated problems. *SIAM J. Scientific Computing*, To appear.
[38] Xu, J. and Zikatanov, L. (1999). A monotone finite element scheme for convection-diffusion equations. *Mathematics of Computation*, Published electronically on May 20, 1999.

APPENDIX A: BERNOULLI FUNCTIONS

The one-variable real-valued *Bernoulli function* is defined as

$$B(t) = \begin{cases} \frac{t}{e^t - 1} & \text{if } t \neq 0, \\ 1 & \text{if } t = 1. \end{cases} \tag{A.1}$$

Notice that the range of the Bernoulli function is $(0, \infty)$ and that of its derivative is $(-1, 0)$. We denote the arithmetic and geometric averages of the Bernoulli function and its image with respect to the y-axis by

$$C(t) = \frac{B(t) + B(-t)}{2} = \frac{t}{2}\coth\left(\frac{t}{2}\right), \tag{A.2}$$

and

$$D(t) = \sqrt{B(t)B(-t)} = \frac{t}{2}\operatorname{csch}\left(\frac{t}{2}\right). \tag{A.3}$$

The following identities have been frequently used in previous sections.

$$B(t) + t = B(t)e^t = B(-t), \tag{A.4}$$

$$B(t) + \frac{t}{2} = B(-t) - \frac{t}{2} = C(t), \tag{A.5}$$

$$B(t)e^{t/2} = B(-t)e^{-t/2} = D(t). \tag{A.6}$$

The derivatives of the Bernoulli function can be computed recursively by using the following relation:

$$\frac{B(t)}{dt} = \begin{cases} -\dfrac{B(t)(B(-t) - 1)}{t} & \text{if } t \neq 0, \\ -\dfrac{1}{2} & \text{if } t = 1. \end{cases} \tag{A.7}$$

We define the function $E(t)$ as the ratio between the $D(t)$ and $C(t)$ functions, which turns out to be a hyperbolic function.

$$E(t) \equiv \frac{D(t)}{C(t)} \equiv \operatorname{sech}\left(\frac{t}{2}\right). \tag{A.8}$$

The second derivatives for $E(t)$ and $E^2(t)$ can be written as

$$E'' = \frac{1}{4}\operatorname{sech}^3\left(\frac{t}{2}\right)\left(\sinh^2\left(\frac{t}{2}\right) - 1\right), \tag{A.9}$$

$$(E^2)'' = \operatorname{sech}^4\left(\frac{t}{2}\right)\left(\sinh^2\left(\frac{t}{2}\right) - \frac{1}{2}\right), \tag{A.10}$$

which implies that the two functions have inflection points (zero second derivative point) at

$$t_{c,\mathrm{J}} = \ln(1 + \sqrt{2}), \quad E(t_{c,\mathrm{J}}) = \frac{\sqrt{2}}{2}, \tag{A.11}$$

$$t_{c,\mathrm{GS}} = \ln\left(\sqrt{\frac{1}{2}} + \sqrt{\frac{3}{2}}\right), \quad E^2(t_{c,\mathrm{GS}}) = \frac{2}{3}. \tag{A.12}$$

Finally, we give fifth-order Taylor's expansions for the family of Bernoulli functions as follows

$$B(t) = 1 - \frac{t}{2} + \frac{t^2}{12} - \frac{t^4}{720} + O(t^6), \tag{A.13}$$

$$C(t) = 1 + \frac{t^2}{12} - \frac{t^4}{720} + O(t^6), \tag{A.14}$$

$$D(t) = 1 - \frac{t^2}{24} + \frac{7t^4}{5760} + O(t^6), \tag{A.15}$$

$$E(t) = 1 - \frac{t^2}{8} + \frac{5t^4}{384} + O(t^6). \tag{A.16}$$

APPENDIX B: DERIVATIVE OF SPECTRAL RADIUS

Let $A(t)$ be a (parameterized) nonnegative square matrix, ρ be its spectral radius, and x and y be its right and left Perron eigenvectors respectively.

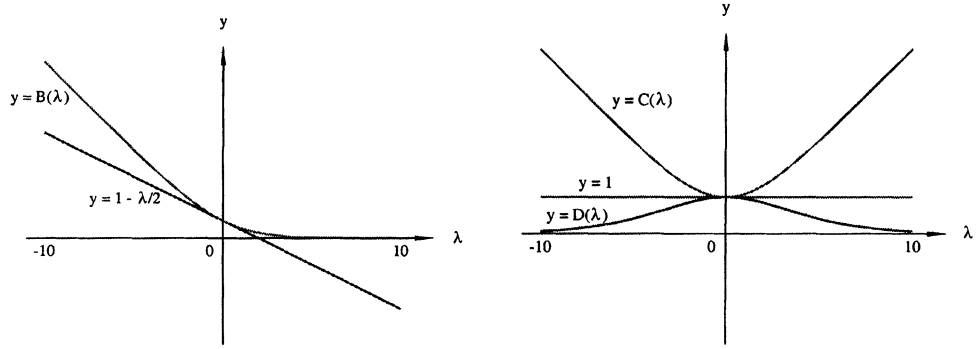

FIGURE A.1 Bernoulli functions: $B(x) = (x/(e^x - 1))$, $C(x) = (1/2)(B(x) + B(-x))$ and $D(x) = (B(x)B(-x))^{(1/2)}$.

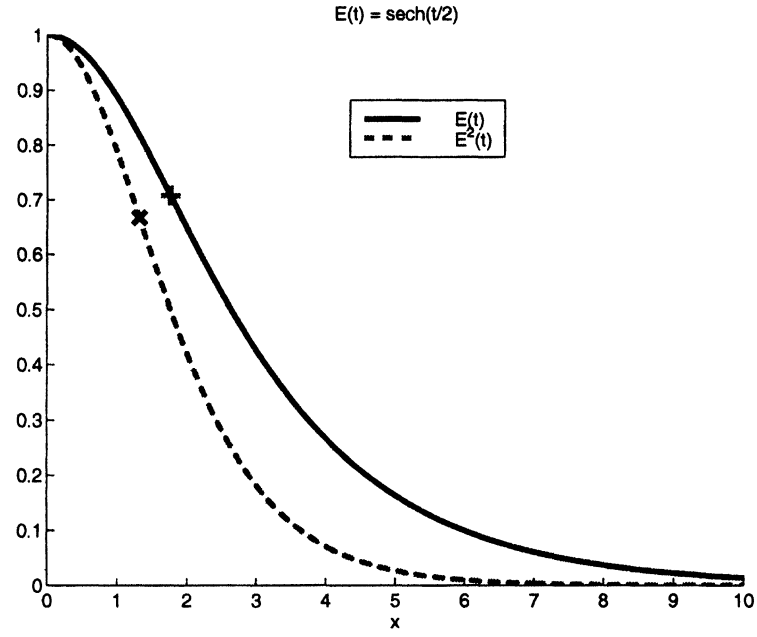

FIGURE A.2 The Bernoulli-E functions marked with inflection points.

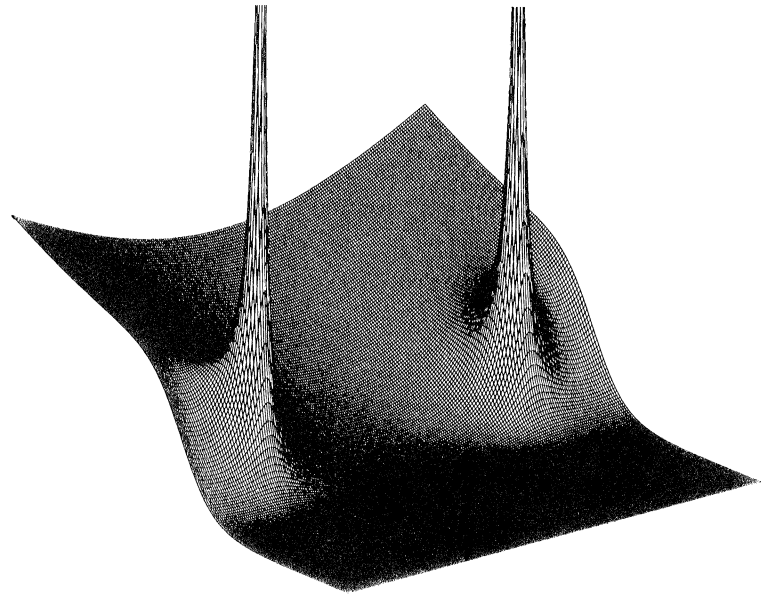

FIGURE A.3 The Ski-Slope function ($|B(z)|, z \in \mathbb{C}$).

Then we have

$$Ax = \rho x. \tag{B.1}$$

Taking the derivative with respect to t on both sides, we have

$$\dot{A}x + A\dot{x} = \rho\dot{x} + \dot{\rho}x, \tag{B.2}$$

or

$$y^\top \dot{A}x + y^\top A\dot{x} = y^\top \rho\dot{x} + y^\top \dot{\rho}x. \tag{B.3}$$

Since y is the left Perron eigenvector of A, we have $y^\top A = y^\top \rho$. Therefore, we have

$$\dot{\rho} = \frac{y^\top \dot{A}x}{y^\top x}. \tag{B.4}$$

Authors' Biographies

Donald J. Rose received the B.A. degree in mathematics from the University of California at Berkeley in 1966, and the M.A. and Ph.D. degrees in applied mathematics from Harvard University, Cambridge, MA, in 1967 and 1970, respectively. From 1970 to 1984, Dr. Rose held research positions at the University of Denver, Harvard University, Vanderbilt University, and AT&T Bell Laboratories. In 1984, he joined the Computer Science Department at Duke University, Durham, NC, where he investigates problems involving numerical analysis and scientific computation. Dr. Rose is a member of SIAM, AMS, MAA, ACM and IEEE.

Hai Shao attended Nanjing University in Nanjing, China and Gordon College in Wenham, Massachusetts and received the B.S. degrees in Physics and Mathematics from the latter institute in 1994. He then attended Duke University in Durham, North Carolina and received the M.S. and Ph.D. degrees in Computer Science in 1997 and 1999, respectively. He is affiliated with ACM, IEEE and SIAM.

Craig Henriquez received both his BSE in Biomedical and Electrical Engineering and Ph.D. in Biomedical Engineering from Duke University in 1981 and 1989, respectively. He joined the faculty

in Biomedical Engineering at Duke University in 1991 and is currently an Associate Professor. His research focuses on developing large scale computer models to study both the genesis of arrhythmias in the heart and neuronal signaling in cortical networks of the brain. Dr. Henriquez heads the Computational Modeling Thrust for the Center for Emerging Cardiovascular Technologies at Duke.