Numerical Algorithms (2020) 83:1007–1027
https://doi.org/10.1007/s11075-019-00713-x

ORIGINAL PAPER

A matrix analysis approach to discrete comparison principles for nonmonotone PDE

Sara Pollock[1] · **Yunrong Zhu**[2]

Received: 27 June 2018 / Accepted: 22 April 2019 / Published online: 29 April 2019

Abstract
We present a linear algebra approach to establishing a discrete comparison principle for a nonmonotone class of quasilinear elliptic partial differential equations. In the absence of a lower order term, local conditions on the mesh are required to establish the comparison principle and uniqueness of the piecewise linear finite element solution. We consider the assembled matrix corresponding to the linearized problem satisfied by the difference of two solutions to the nonlinear problem. Monotonicity of the assembled matrix establishes a maximum principle for the linear problem and a comparison principle for the nonlinear problem. The matrix analysis approach to the discrete comparison principle yields sharper constants and more relaxed mesh conditions than does the argument by contradiction used in previous work.

Keywords Discrete comparison principle · Uniqueness · Nonmonotone problems · Quasilinear partial differential equations · Monotone matrix · M-matrix

Mathematics Subject Classification (2010) 65N30 · 35J62

1 Introduction

We consider a linear algebra approach to develop a discrete comparison principle for the equation as follows:

$$-\operatorname{div}(\kappa(x,u)\nabla u) + g(x,u) = f, \text{ in } \Omega \subset \mathbb{R}^2, \tag{1.1}$$

✉ Sara Pollock
s.pollock@ufl.edu

Yunrong Zhu
zhuyunr@isu.edu

[1] Department of Mathematics, University of Florida, Gainesville, FL 32611, USA

[2] Department of Mathematics and Statistics, Idaho State University, Pocatello, ID 83209, USA

with homogeneous Dirichlet conditions $u = 0$ on $\Gamma_D = \partial\Omega$, where the domain Ω is either polygonal or a finite collection of polygons. The discrete comparison principle implies uniqueness of the discrete solution, in agreement with the comparison principle and uniqueness for the continuous problem. Here, we develop the discrete comparison principle for (1.1) by means of a discrete maximum principle for a linearized equation satisfied by the difference of a subsolution and supersolution which is demonstrated by the monotonicity of the assembled coefficient matrix.

The PDE (1.1) is both nonmonotone and nonvariational (see, e.g., [18, 29]); and, as demonstrated in [2], uniqueness of solutions to its finite element approximation can fail if the mesh is too coarse, even when the PDE solution is known to be unique. Asymptotic error estimates for a finite element approximation as the mesh size $h \to 0$ were first shown in 1975 in [11]. More recently, similar results were shown to hold under integration by quadrature in [1]. In [2], an argument by contradiction related to the approach used in the continuous case is used to establish a discrete comparison principle based on the condition that the mesh partition is globally fine enough. For the 2D case, the result is presented as an asymptotic estimate and as such does not yield a verifiable condition for uniqueness. The current authors used similar ideas in [22, 23] to demonstrate that a local verifiable condition based on the variance of the solution over each element, rather than a global mesh size condition, is sufficient for uniqueness of solutions in the absence of a lower order term. Here, we improve the constant appearing in the *a posteriori* condition and also relax the angle condition on the mesh.

The practical outcome of these results is that for the nonlinear diffusion problem with $g(x, u) = 0$, if the maximum difference between neighboring nodal solution values is small enough (less than a given constant), then the discrete solution can be verified as unique. For the reaction-diffusion problem which includes the nontrivial term $g(x, u)$, if the mesh is additionally fine enough in a global sense, then the discrete solution can be verified as unique. If these conditions are not met on a given mesh, then either refinement or remeshing may be used to attain their satisfaction.

The analytical techniques of this paper differ from those in [23] which addresses the nonlinear diffusion problem and [22] which addresses a nonlinear reaction-diffusion problem with an additional gradient-dependence in the nonlinear diffusion term. Both of the above papers develop a proof by contradiction based on element-wise integration. The current technique considers integration over patches which allows a certain relaxation on derived *a posteriori* conditions. Here, we show the comparison principle for the nonlinear problem can be established by the monotonicity of the coefficient matrix for a linearized problem. Conditions for this monotonicity are established by first controlling the maximum difference between neighboring nodal solution values to show the coefficient matrix has nonpositive off-diagonal entries, followed by a technical lemma which demonstrates two of the many established conditions for matrix monotonicity. Two independently developed approaches to matrix monotonicity are used here, one by J. Bramble and B. Hubbard from 1964 [6, 7] to establish the matrix is of "positive type," and the second by M. Fielder and V. Ptàk from 1962 [15]. A secondary result of interest is that based on the structure of the coefficient matrix, the complete set of conditions for either framework are automatically satisfied by the conditions under which the off-diagonal entries of the

coefficient matrix are nonpositive. While sufficient in this case, the nonpositivity of off-diagonal elements is not a necessary condition for matrix monotonicity.

This manuscript is motivated by the linear algebra approach used to establish a discrete maximum principle in [8], demonstrated to improve previously established constants. The current authors also showed in [22] that the maximum principle for the linear reaction-diffusion equation developed in [8, Theorem 3.8] has a direct application to a discrete comparison principle for the semilinear problem, $-\Delta u + g(x, u) = f(x)$; and, the matrix analysis approach yields an improved constant. Here, we extend the analysis to quasilinear problems. In the semilinear case, the argument follows by showing that the assembled matrix in question is monotone by showing it is a *Stieltjes* matrix, meaning it is symmetric positive definite with nonpositive off-diagonal entries. The particulars of the analysis do not apply in the quasilinear case, as the corresponding coefficient matrix for the linearization of (1.1) is nonsymmetric.

In this paper, we develop conditions under which the assembled coefficient matrix A corresponding to the PDE satisfied by the difference between a subsolution and supersolution of (1.1) is monotone. The main technical challenge is in understanding the monotonicity of the linearized coefficient matrix which resembles one arising from the discretization of a linear convection-diffusion or a reaction-convection-diffusion equation. Related ideas on the monotonicity of the coefficient matrix for a class of linear convection-diffusion problems is found in [30], where the bilinear form is altered in an edge-averaged scheme that preserves monotonicity. Here, we use a standard discretization scheme and derive local estimates sufficient to ensure the convection-like term in the linearization, which is controlled by the difference in nodal values of the solution across each edge, is controlled by the diffusion.

We proceed by developing conditions under which the assembled coefficient matrix is a Z-matrix, one with nonpositive off-diagonal entries. While not all monotone matrices are Z-matrices (see, for example [5–7]), the monotone matrices which may contain some positive off-diagonal entries are generally difficult to recognize. On the other hand, the theory of monotone Z-matrices in numerical analysis has been well-studied, largely with respect to the convergence of iterative methods [28, 31]. In [21], a collection of 40 conditions for a Z-matrix to be a nonsingular M-matrix, hence monotone, is drawn from the literature. We use one of those conditions which appears earlier in [15] to establish our results. Simultaneously, we also show under the same conditions the matrix is of "positive type," as in [6, 7], yielding the same conclusion. As we find in the sequel, the conditions used to ensure the coefficient matrix A is a Z-matrix are sufficient to assure it is monotone.

The first main contribution of this work is improving the constants in the local condition for the discrete comparison principle hence uniqueness to hold. The second main contribution is establishing the discrete comparison principle holds for problem (1.1) on meshes with at least some right angles. In previous work by the authors [22, 23], the mesh was assumed acute, meaning all interior angles were bounded below $\pi/2$. In the current results, interior angles can be no greater than $\pi/2$, and opposite angles across each edge must sum to less that π. It is well-known (see, for instance [30, Lemma 2.1]), that for the assembled matrix for the Laplacian, monotonicity holds under the condition that the mesh is Delaunay, meaning the angles opposite each edge sum to no more than π. More general geometric conditions for a

discrete maximum principle for Poisson's equation are developed in [13], in which more obtuse triangles are permitted some distance from the boundary. The stronger condition on the geometry in this work comes from the variable dependence in the principal part of the linearized problem, and the need to control the lower order convection-like terms in the linearization by the principal part.

In two dimensions, acute triangulations (all interior angles less than $\pi/2$) can be generated by standard methods, see for instance [3, 16] and the recent [14]; however, for irregular domains, the mesh partitions may not be uniform in size. The current results yield a condition that may be checked once a solution has been computed to determine whether that solution is unique. In the case that it is not, the mesh can then be refined, for instance, using a red-refinement algorithm [4] to preserve acuteness.

The remainder of the paper is structured as follows. In Section 2, the discretization and discrete comparison principle are introduced. Then, the linearized problem used to investigate the comparison principle is derived. Theorem 3 relates the monotonicity of the assembled matrix for the linearized problem to the comparison principle for the nonlinear problem. The definitions of Z-, L-, and M- matrices, matrices of "positive type,"and the relevant theorems on establishing monotonicity are recalled from the literature. Section 4 contains the technical estimates providing sufficient conditions under which the assembled matrix A is a Z-matrix, and the Appendix contains the necessary estimates related to diagonal dominance to complete the monotonicity argument.

2 Preliminaries

We make the following assumptions on the problem data $\kappa(\cdot\,,\cdot)$ and $g(\cdot\,,\cdot)$.

Assumption 1 *Assume $\kappa(x,\eta)$ and $g(x,\eta)$ are Carathéodory functions, measurable in x for each $\eta \in \mathbb{R}$, and C^1 in η for a.e. $x \in \Omega$. Assume there are constants $0 < k_\alpha < k_\beta$ with the following:*

$$k_\alpha \le \kappa(x,\eta) \le k_\beta, \tag{2.1}$$

for all $\eta \in \mathbb{R}$, and a.e. $x \in \Omega$. Assume there is a positive K_η with the following:

$$\left|\frac{\partial \kappa}{\partial \eta}(x,\eta)\right| \le K_\eta, \tag{2.2}$$

for all $\eta \in \mathbb{R}$ and a.e. $x \in \Omega$. Assume $g(x,\eta)$ is nondecreasing with respect to its second argument, and there is a constant G_η with the following:

$$0 \le \frac{\partial g}{\partial \eta}(x,\eta) \le G_\eta, \tag{2.3}$$

for all $\eta \in \mathbb{R}$ and a.e. $x \in \Omega$.

Under Assumption 1, the PDE is known to satisfy a comparison principle and have a unique solution, as demonstrated in [12, 26], and [17, Chapter 10]. Additionally,

the weak form of (1.1) is given as follows for $V = H_0^1(\Omega)$, the closed subspace of H^1 with vanishing trace on $\partial\Omega$. Find $u \in V$ such as the following:

$$\int_\Omega \kappa(x,u)\nabla u\cdot\nabla v + g(x,u)v\,\mathrm{d}x = \int_\Omega f v\,\mathrm{d}x, \text{ for all } v \in V. \tag{2.4}$$

The functional setting of the weak form can be understood in the context of the Leray-Lions conditions for pseudomonotonicity. We refer interested readers to [9, Chapters 2–3] for further details. We note that the class of problems defined by the assumptions in this section is called *nonmonotone* because the inequality as follows:

$$\int_\Omega (\kappa(x,w)\nabla w - \kappa(x,v)\nabla v)\cdot\nabla(w-v)\,\mathrm{d}x \geq 0,$$

does not in general hold for all $w, v \in V$, even for $\kappa = \kappa(u)$ if $\kappa(u)$ is nonconstant [29, Theorem 4.1].

2.1 Discretization

Let $\mathcal{T}$ be a conforming simplicial partition of domain Ω that exactly captures the boundary. Let $\bar{\mathcal{Q}}$ be the collection of vertices or nodes of $\mathcal{T}$, and let $\mathcal{Q} = \bar{\mathcal{Q}} \setminus \partial\Omega$ be the set of nodes that do not lie on the Dirichlet boundary, corresponding to the mesh degrees of freedom. Let $\mathcal{V} \subsetneq V$ be the discrete space spanned by the piecewise linear basis functions $\{\varphi_j\}$ that satisfy $\varphi_i(q_i) = 1$ and $\varphi_i(q_j) = 0$ for each $q_j \in \mathcal{Q}$ with $q_j \neq q_i$. Define the non-negative subset of $\mathcal{V}$ by $\mathcal{V}^+ := \{v \in \mathcal{V} \mid v \geq 0\}$.

Let ω_i be the support of the basis function φ_i. Define the intersection of support for any two basis functions with respect to a global numbering by $\omega_{ij} = \omega_i \cap \omega_j$. In terms of the corresponding nodes q_i and q_j, it follows that ω_{ij} is the union of elements that share both q_i and q_j as vertices.

$$\omega_{ij} = \bigcup\{T \in \mathcal{T} \mid q_i \in T \text{ and } q_j \in T\}.$$

Additional notation for the discretization is summarized as follows, and illustrated in Fig. 1.

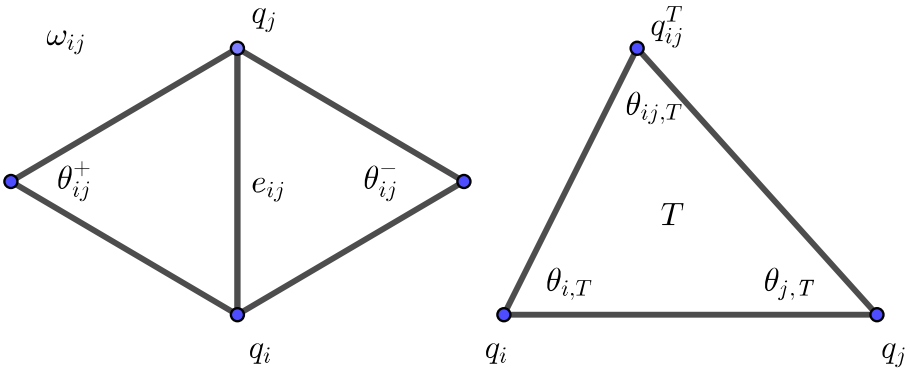

Fig. 1 Left: schematic of a patch. Right: schematic of an element

- e_{ij} denotes the edge connecting vertices q_i and q_j.
- $\delta^e_{ij}(v) = |v(q_i) - v(q_j)|$ is the absolute difference of nodal values of v across edge e_{ij}.
- $\delta_\omega(v) = \max\{\delta^e_{jk}(v) \mid q_j, q_k \in \omega\}$ denotes the maximum difference between neighboring nodal values of v in ω.
- $\delta(v) = \delta_\omega(v)$ when $\omega = \bar{\Omega}$.
- $\delta^{(i)}_\omega(v) = \max\{\delta^e_{jk}(v) \mid q_j, q_k \in \omega,\ j,k \neq i\}$ is the maximum difference between nodal values of v in ω, across edges not touching vertex q_i.
- θ^+_{ij} and θ^-_{ij} denote the two respective angles opposite edge e_{ij} in ω_{ij}.
- $\theta_{j,T}$ denotes the interior angle of triangle T at vertex q_j, and $\theta_{ij,T}$ denotes the interior angle opposite vertices q_i and q_j in triangle T.
- q^T_{ij} denotes the vertex opposite both q_i and q_j in triangle T.
- φ^T_{ij} denotes the basis function associated with q^T_{ij} in triangle T.
- $|T|$ denotes the area of triangle $T \in \mathcal{T}$.
- $|T_\mathcal{T}| = \max_{T\in\mathcal{T}} |T|$.
- $\bar{\mathcal{Q}}_i = \{q_j \in \bar{\mathcal{Q}} \mid q_j \in \omega_i,\ j \neq i\}$, denotes the set of vertices neighboring q_i, including those on Γ_D.
- $\mathcal{Q}_i = \bar{\mathcal{Q}}_i \setminus \Gamma_D$, the set of non-Dirichlet vertices neighboring q_i.

We make the following assumptions on the triangulation.

Assumption 2 *Any interior angle of the mesh satisfies $\theta \leq \pi/2$, and any two angles opposite an edge sum to less than π. In particular, there is a constant $\beta_m > 0$ for which as follows:*

$$\cot\theta^+_{ij} + \cot\theta^-_{ij} \geq \beta_m, \quad \textit{for each } \omega_{ij} \subset \mathcal{T}. \tag{2.5}$$

The mesh satisfies a smallest-angle condition over each neighborhood ω_{ij}. There is a constant $\beta_M > 0$ for which as follows:

$$\cot\theta^+_{ij} + \cot\theta^-_{ij} + \sum_{T\in\omega_{ij}} \cot\theta_{i,T} \leq \beta_M \textit{ for each } \omega_{ij} \subset \mathcal{T}. \tag{2.6}$$

For example, in an equilateral mesh $\beta_m = 2/\sqrt{3}$ and $\beta_M = 4/\sqrt{3}$.

2.2 Comparison framework

Consider the problems: find $u_i \in \mathcal{V}$ such as the following:

$$\int_\Omega \kappa(x,u_i)\nabla u_i \cdot \nabla v + g(x,u_i)v\,\mathrm{d}x = \int_\Omega f_i v\,\mathrm{d}x, \text{ for all } v \in \mathcal{V}, \tag{2.7}$$

for $f_i \in L_2(\Omega)$, $i = 1, 2$. The discrete comparison principle for (2.7) states that whenever $f_1 \leq f_2$, (a.e. $x \in \Omega$), meaning $\int_\Omega (f_1 - f_2)v\,\mathrm{d}x \leq 0$, for each $v \in \mathcal{V}^+$, then it holds that $u_1 \leq u_2$.

The comparison principle can be restated in terms of a maximum principle for $w = u_1 - u_2$. The discrete problem satisfied by w can be understood by applying Taylor's theorem to the difference of (2.7) with $i = 1$ and $i = 2$.

$$\begin{aligned}
&\int_\Omega (\kappa(x,u_1)\nabla u_1\cdot\nabla v + g(x,u_1)v)\ \mathrm{d}x - \int_\Omega (\kappa(x,u_2)\nabla u_2\cdot\nabla v + g(x,u_2)v)\ \mathrm{d}x\\
&= \int_\Omega \kappa(x,u_1)\nabla(u_1-u_2)\cdot\nabla v - (\kappa(x,u_2)-\kappa(x,u_1))\nabla u_2\cdot\nabla v\,\mathrm{d}x\\
&\quad + \int_\Omega (g(x,u_1)-g(x,u_2))v\,\mathrm{d}x\\
&= \int_\Omega \kappa(x,u_1)\nabla w\cdot\nabla v\,\mathrm{d}x - \int_\Omega \left(\int_0^1 \frac{\partial\kappa}{\partial\eta}(x,z(t))w\,dt\right)\nabla u_2\cdot\nabla v\,\mathrm{d}x\\
&\quad + \int_\Omega \left(\int_0^1 \frac{\partial g}{\partial\eta}(x,z(t))w\,dt\right) v\,\mathrm{d}x, \qquad (2.8)
\end{aligned}$$

where $z(t) = tu_1 + (1-t)u_2$, for $0 \le t \le 1$. The equation satisfied by w is as follows:

$$\begin{aligned}
&\int_\Omega \kappa(x,u_1)\nabla w\cdot\nabla v\,\mathrm{d}x + \int_\Omega \left(\int_0^1 \frac{\partial\kappa}{\partial\eta}(x,z(t))\,dt\right) w\nabla u_2\cdot\nabla v\,\mathrm{d}x\\
&\quad + \int_\Omega \left(\int_0^1 \frac{\partial g}{\partial\eta}(x,z(t))w\,dt\right) v\,\mathrm{d}x = \int_\Omega (f_1-f_2)v\,\mathrm{d}x, \quad \text{for all } v\in\mathcal{V}. \qquad (2.9)
\end{aligned}$$

The comparison principle for u_1 and u_2 in (2.7) is then equivalent to the weak maximum principle for $w = u_1 - u_2$ in (2.9), namely $w \le 0$ whenever $f_1 - f_2 \le 0$. We now turn our attention to establishing the weak maximum principle for w.

3 Discrete maximum principle

From (2.9), the linear equation for w is a general second-order elliptic equation with convection and reaction terms as follows:

$$\int_\Omega \kappa(x,u_1)\nabla w\cdot\nabla v + \mathbf{b}(x,u_2)w\cdot\nabla v + c(x)wv\,\mathrm{d}x = \int_\Omega (f_1-f_2)v\,\mathrm{d}x, \qquad (3.1)$$

for all $v \in \mathcal{V}$, with as follows:

$$\mathbf{b}(x,u_2) := b(x)\nabla u_2(x), \text{ with } b(x) := \int_0^1 \frac{\partial\kappa}{\partial\eta}(x,z(t))\,\mathrm{d}t, \qquad (3.2)$$

$$c(x) := \int_0^1 \frac{\partial g}{\partial\eta}(x,z(t))\,\mathrm{d}t. \qquad (3.3)$$

We now consider the properties of the assembled system (3.1)–(3.3). The discrete function $w \in \mathcal{V}$ has the expansion in basis functions $w = \sum_{j=1}^n W_j\varphi_j$, where n is the number of mesh degrees of freedom. Choosing the test functions $v = \varphi_i$ for each $i = 1, \cdots, n$ in (3.1), we obtain the equivalent matrix problem $AW = F$. In particular,

letting $^\top$ denote transposition, $W = (W_1, \cdots, W_n)^\top$, $F = (F_1, \cdots, F_n)^\top$ with $A = (a_{ij})$ and F_i defined entry-wise by the following:

$$a_{ij} = \int_{\omega_{ij}} \kappa(x, u_1)\nabla\varphi_j \cdot \nabla\varphi_i \, \mathrm{d}x + \int_{\omega_{ij}} \vec{b}(x, u_2)\varphi_j \cdot \nabla\varphi_i \, \mathrm{d}x + \int_{\omega_{ij}} c(x)\varphi_j\varphi_i \, \mathrm{d}x \quad (3.4)$$

$$F_i = \int_{\omega_i} (f_1 - f_2)\varphi_i \, \mathrm{d}x. \quad (3.5)$$

The maximum principle for w is established by the monotonicity of matrix A given by (3.4). For the remainder of the section, A is assumed to be a real-valued $n \times n$ square matrix.

Definition 1 (Monotone matrix) Matrix A is monotone (in the sense of [10, Section 23]) if for all real vectors v, $Av \geq 0$ implies $v \geq 0$, where $\geq$ is the element-wise ordering.

By this definition, a monotone matrix is nonsingular because if A is a monotone matrix and x is a vector in its nullspace, then both $x \geq 0$ and $-x \geq 0$, implying $x = 0$. We mention another relevant property of monotone matrices [20].

Proposition 1 *Matrix A is monotone iff A is invertible and $A^{-1} \geq 0$.*

The next theorem summarizes how monotonicity of the assembled matrix for w implies the comparison principle, hence uniqueness of the solution.

Theorem 3 *Suppose the functions $\kappa(x, \eta)$ and $g(x, \eta)$ are each measurable with respect to the first argument, and C^1 with respect to the second. Let $u_i \in \mathcal{V}$ solve the following:*

$$\int_\Omega \kappa(x, u_i)\nabla u_i \cdot \nabla v + g(x, u_i)v \, \mathrm{d}x = \int_\Omega f_i v \, \mathrm{d}x, \text{ for all } v \in \mathcal{V}, \quad (3.6)$$

for $f_i \in L_2(\Omega)$, $i = 1, 2$, with $\int_\Omega (f_1 - f_2)v \, \mathrm{d}x \leq 0$, for all $v \in \mathcal{V}^+$. If the coefficient matrix A defined by (3.4) is monotone then $u_1 \leq u_2$. Moreover, if $f_1 = f_2$, then $u_1 = u_2$.

Proof Let the discrete function $w \in \mathcal{V}$ be given by $w = \sum_{j=1}^n W_j\varphi_j$, with $W = (W_1, \cdots, W_n)^\top$, where $n = \mathrm{card}(\mathcal{Q})$, the number of mesh degrees of freedom. The monotonicity of A implies its invertibility. By equations (2.8)–(2.9), the definition of A, and F given by (3.5), the vector W that solves $AW = F$ uniquely defines $w \in \mathcal{V}$ that satisfies $w = u_1 - u_2$.

Since A is monotone, $F \leq 0$ implies $W \leq 0$ which by the nonnegativity of basis functions implies $w \leq 0$ implying $u_1 \leq u_2$. If $f_1 = f_2$, it follows that $u_1 = u_2$. □

The consequence for uniqueness of solutions is once a discrete solution u has been computed one can determine *a posteriori* if it is the unique solution. In what follows, the conditions on the discrete solution to determine monotonicity (hence uniqueness)

are derived only in terms of the constants introduced in Assumptions 1 and 2, the mesh size $|T_{\mathcal{T}}|$ (if there is a lower order term) and *a posteriori* properties of $u_2 = u$. The subsolution u_1 does not appear explicitly.

In the remainder of the paper, we develop conditions under which the assembled matrix (3.4) is monotone. Three related classes of matrices we encounter in the proof are Z-matrices, L-matrices, and M-matrices, are defined as follows.

Definition 2 (Z-matrix) $A = (a_{ij})$ is called an Z-matrix if $a_{ij} \leq 0$ when $i \neq j$ (see [15, Definition 4,1],[21]).

A Z-matrix with positive diagonal entries is called an L-matrix.

Definition 3 (L-matrix) $A = (a_{ij})$ is called an L-matrix if $a_{ii} > 0$ for all i, and $a_{ij} \leq 0$ when $i \neq j$ (see [31, Chapter 2, Definition 7.1]).

A monotone Z-matrix is an M-matrix.

Definition 4 (M-matrix) $A = (a_{ij})$ is called an M-matrix if $a_{ij} \leq 0$ for all $i \neq j$ and A is nonsingular with $A^{-1} \geq 0$ (see [28, Definition 3.22]).

This definition is equivalent to [31, Definition 7.3], and to a nonsingular M-matrix, in [21]. If the off-diagonal entries of A are nonpositive (A is a Z-matrix), then A is monotone if and only if A is an M-matrix. This is clear from Proposition 1, and the definition of the M-matrix.

In what follows, we will develop conditions under which the coefficient matrix A given by (3.4) is monotone. We first establish conditions under which A is a Z-matrix. We will see these same conditions imply A is also an L matrix. We then present two arguments to establish A is monotone, hence an M-matrix. The first uses the 1964 result of [6, 7] by which A is a matrix of "positive type," hence monotone. In the second (equivalent) approach, we construct a positive diagonal matrix D to establish A has "generalized diagonal dominance," which by a 1962 result of [15] implies A is monotone. We next review some related concepts on diagonal dominance.

Definition 5 An $n \times n$ matrix is strictly diagonally dominant (SDD) as follows:

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|, \quad \text{for all } i = 1, \ldots, n. \tag{3.7}$$

A matrix is called diagonally dominant (DD) if equality is allowed for each index i in (3.7). If matrix A is diagonally dominant with a strict inequality in (3.7) for at least one index i, we say the matrix has the DD+ property.

If A has the DD+ property and it is irreducible [27, Definition 1.15], meaning the matrix equation $Ax = b$ cannot be decoupled into two or more smaller problems, it is called "irreducibly diagonally dominant." Indeed, an irreducibly diagonally dominant L-matrix is an M-matrix [27, Corollary 3.20]. An equivalent description of

an irreducible matrix is one whose directed graph is strongly connected [27, Theorem 1.17]; referring to the counterexample of [13], the irreducibility of the Laplacian stiffness matrix can fail to hold, even on a connected mesh under the orthogonality of certain basis functions due to right angles in the mesh. The next definition, however, introduces a substantially weaker condition than irreducibility which can be used to establish a Z-matrix is an M-matrix.

Definition 6 [6, Definition 2.2] Matrix A is said to be of positive type if the following conditions hold.

1. A is a Z-matrix.
2. A is DD+. Specifically, $\sum_{j=1}^{n} a_{ij} > 0$ for $j \in J(A) \neq \emptyset$.
3. For each $i \notin J(A)$, there is a sequence of nonzero elements of the form $a_{ik_1}, a_{k_1k_2}, \ldots, a_{k_r j}$ where $j \in J(A)$.

In [6, 7], the last condition is called a "connection," in A from i to $J(A)$. In [24, 25], this is also called a "chain condition," and it is used as a sufficient condition to show a matrix is nonsingular. In those works, matrices satisfying the last two conditions of Definition 6 are referred to as "weakly chained diagonally dominant," or "chain diagonally dominant," (CDD). Our first approach depends on the following result.

Theorem 4 [7, Theorem 2.2] *If A is of positive type, then A is monotone.*

For our second approach, we rely on another variant of diagonal dominance, sometimes referred to as "generalized diagonal dominance," (GDD). Matrix A is said to have the GDD property if there is a positive matrix D for which AD is strictly diagonally dominant. In fact, [19, Theorem 3.3] shows for diagonally dominant matrices GDD is equivalent to the second two conditions of Definition 6 (GDD is equivalent to CDD). The concept of generalized diagonal dominance also appears much earlier in the literature, and we refer to the 1962 paper [15] which presents 13 equivalent statements which characterize a Z-matrix as monotone. Here, we paraphrase the two most relevant to our purposes.

Theorem 5 [15, Theorem 4, 3.4],[21, Theorem 1 (N39)] *Let A be a Z-matrix. Then, A is monotone if and only if there exists a diagonal matrix D with positive diagonal elements such that the matrix AD is strictly diagonally dominant.*

In our second approach, we construct a matrix D that satisfies Theorem 5 for matrix A given by (3.4), under the condtions for which it is a Z-matrix.

4 Properties of the assembled matrix

In this section, we develop conditions on the meshsize and the supersolution u_2 under which A, the coefficient matrix from (3.4), is a Z-matrix. Then, we will see

A satisfies the conditions of both Theorems 4 and 5, hence is monotone. Our main comparison and uniqueness result then follows from Theorem 3.

Lemma 1 *Let Assumption 1 and Assumption 2 hold. Let $A = (a_{ij})$ be given by (3.4) on satisfaction of the condition on $\delta(u_2)$ and mesh size $|T_{\mathcal{T}}|$ as follows:*

$$\frac{1}{3\beta_m}\left\{K_\eta \beta_M \delta(u_2) + G_\eta |T_{\mathcal{T}}|\right\} < k_\alpha, \tag{4.1}$$

it holds that $a_{ij} \leq 0$ for each $i \neq j$, and $a_{ij} < 0$ whenever vertices $q_i, q_j \in \mathcal{Q}$ are connected by an edge.

Proof For convenience of later calculations, we consider the entry a_{ji}. Since $a_{ji} = 0$ if there is no edge connecting q_i and q_j, we consider only edge-connected $q_i, q_j \in \mathcal{Q}$ for the remainder of the proof. By direct calculation, (see, e.g., [23, Section 3.1]), it holds for $i \neq j$ as follows:

$$\nabla\varphi_i \cdot \nabla\varphi_j|_T = \frac{-1}{2|T|}\cot\theta_{ij,T}, \tag{4.2}$$

where $\theta_{ij,T}$ is the angle opposite edge e_{ij} in triangle T (see Fig. 1). Equation (4.2) together with (2.1) and (2.5) implies the following:

$$\int_{\omega_{ij}} \kappa(x,u_1)\nabla\varphi_i\cdot\nabla\varphi_j \,\mathrm{d}x \leq k_\alpha \sum_{T\in\omega_{ij}} \int_T \nabla\varphi_i\cdot\nabla\varphi_j \,\mathrm{d}x = -\frac{k_\alpha}{2}(\cot\theta_{ij}^+ + \cot\theta_{ij}^-). \tag{4.3}$$

For the reaction term, $\int_T \varphi_i\varphi_j \,\mathrm{d}x = |T|/12$. Together with (2.3) and (3.3), this implies the following:

$$\int_{\omega_{ij}} c(x)\varphi_i\varphi_j \,\mathrm{d}x \leq \sum_{T\in\omega_{ij}} \frac{G_\eta|T|}{12}. \tag{4.4}$$

To bound the nonsymmetric term, the following decomposition is useful. Over each triangle T with vertices q_i, q_j, q_k and discrete function $v \in \mathcal{V}$ as follows:

$$\nabla v = (v(q_i) - v(q_j))\nabla\varphi_i + (v(q_k) - v(q_j))\nabla\varphi_k. \tag{4.5}$$

Applying (4.2) and (4.5) with $v = u_2$ and $q_k = q_{ij}^T$ for each $T \in \omega_{ij}$ yields the following:

$$\begin{aligned}
\int_{\omega_{ij}} b(x)\nabla u_2\cdot\nabla\varphi_j\varphi_i \,\mathrm{d}x &= (u_2(q_i) - u_2(q_j)) \int_{\omega_{ij}} \nabla\varphi_i\cdot\nabla\varphi_j\, b(x)\varphi_i \,\mathrm{d}x \\
&\quad + \sum_{T\in\omega_{ij}} (u_2(q_{ij}^T) - u_2(q_j)) \int_T \nabla\varphi_{ij}^T\cdot\nabla\varphi_j\, b(x)\varphi_i \,\mathrm{d}x \\
&\leq \frac{K_\eta}{6}|u_2(q_i) - u_2(q_j)|(\cot\theta_{ij}^+ + \cot\theta_{ij}^-) \\
&\quad + \frac{K_\eta}{6}\sum_{T\in\omega_{ij}} |u_2(q_{ij}^T) - u_2(q_j)|\cot\theta_{i,T},
\end{aligned} \tag{4.6}$$

where the inequality follows from (2.2) of Assumption 1 and (3.2). Then,

$$\left| \int_{\omega_{ij}} \vec{b} \cdot \nabla \varphi_j \varphi_i \, \mathrm{d}x \right| \leq \frac{K_\eta}{6} \left(\delta^e_{ij}(u_2)(\cot\theta^+_{ij} + \cot\theta^-_{ij}) + \delta^{(i)}_{\omega_{ij}}(u_2) \sum_{T \in \omega_{ij}} \cot\theta_{i,T} \right). \tag{4.7}$$

Applying (4.3), (4.4), and (4.7) to (3.4), we have the following:

$$\begin{aligned} a_{ji} &\leq \frac{1}{2} \left(-k_\alpha + \frac{K_\eta}{3} \delta^e_{ij}(u_2) \right) (\cot\theta^+_{ij} + \cot\theta^-_{ij}) && (4.8) \\ &\quad + \frac{1}{2} \sum_{T \in \omega_{ij}} \left(\frac{K_\eta}{3} \delta^{(i)}_{\omega_{ij}}(u_2) \cot\theta_{i,T} + \frac{G_\eta |T|}{6} \right) && (4.9) \\ &\leq \frac{1}{2} \left\{ -k_\alpha \beta_m + \frac{K_\eta \beta_M}{3} \delta_{\omega_{ij}}(u_2) + \frac{G_\eta |T_{\mathcal{T}}|}{3} \right\}, && (4.10) \end{aligned}$$

where the last inequality follows from the application of both angle conditions (2.5) and (2.6) from Assumption 2. The conclusion then follows under condition (4.1). □

In the next lemma, we show the diagonal entries of A are positive, under the given condition which bounds the difference of nodal values across each edge in the mesh. The local condition (4.11) for each a_{ii} to be positive is weaker than (4.1), implying that matrix A is an L-matrix when it is a Z-matrix.

Lemma 2 *Let Assumption 1 and Assumption 2 hold. Let $A = (a_{ij})$ be given by (3.4). Then, under the condition as follows:*

$$\delta(u_2) < \frac{3k_\alpha}{K_\eta}, \tag{4.11}$$

it holds that $a_{ii} > 0$, for each i.

Proof First, consider the diffusion term. Summing integrals over each $\omega_{ij} \subset \omega_i$ integrates twice over ω_i. Applying the identity $\nabla\varphi_i \cdot \nabla\varphi_i = -\nabla\varphi_j \cdot \nabla\varphi_i - \nabla\varphi^T_{ij} \cdot \nabla\varphi_i$, over each element $T \in \omega_i$ with nodal indices q_i, q_j, q^T_{ij}, and combining like terms to integrate each product once per element, we have the following:

$$\begin{aligned} \int_{\omega_i} \kappa(x, u_1) \nabla\varphi_i \cdot \nabla\varphi_i \, \mathrm{d}x &= \frac{1}{2} \sum_{\omega_{ij} \subset \omega_i} \int_{\omega_{ij}} \kappa(x, u_1) \nabla\varphi_i \cdot \nabla\varphi_i \, \mathrm{d}x \\ &= - \sum_{\omega_{ij} \subset \omega_i} \int_{\omega_{ij}} \kappa(x, u_1) \nabla\varphi_j \cdot \nabla\varphi_i \, \mathrm{d}x \\ &\geq \frac{k_\alpha}{2} \sum_{\omega_{ij} \subset \omega_i} \left(\cot\theta^-_{ij} + \cot\theta^+_{ij} \right), \end{aligned} \tag{4.12}$$

where the last inequality follows from (2.1) and Assumption 2. Next, consider the nonsymmetric term. Summing integrals over each $\omega_{ij} \subset \omega_i$ then combining like terms as above, we have the following:

$$\begin{aligned}\int_{\omega_i} b(x)\nabla u_2 \cdot \nabla\varphi_i\varphi_i \,\mathrm{d}x &= \frac{1}{2}\sum_{\omega_{ij}\subset\omega_i}\int_{\omega_{ij}} b(x)\nabla u_2\cdot\nabla\varphi_i\varphi_i \,\mathrm{d}x \\ &= \sum_{\omega_{ij}\subset\omega_i}(u_2(q_j)-u_2(q_i))\int_{\omega_{ij}}\nabla\varphi_j\cdot\nabla\varphi_i b(x)\varphi_i \,\mathrm{d}x \\ &\geq -\frac{K_\eta}{6}\sum_{\omega_{ij}\subset\omega_i}|u_2(q_j)-u_2(q_i)|\left(\cot\theta_{ij}^+ + \cot\theta_{ij}^-\right), \quad (4.13)\end{aligned}$$

where the last inequality follows from (2.2), Assumption 2, and the integration of φ_i over each element. From (3.2) and (4.13), we have the following:

$$\int_{\omega_i}\mathbf{b}(x,u_2)\cdot\nabla\varphi_i\varphi_i \,\mathrm{d}x \geq -\frac{K_\eta}{6}\sum_{\omega_{ij}\subset\omega}\delta_{ij}^e(u_2)\left(\cot\theta_{ij}^+ + \cot\theta_{ij}^-\right). \quad (4.14)$$

The lowest order term from (3.4) satisfies $\int_{\omega_i} c(x)\varphi_i^2 \,\mathrm{d}x \geq 0$, for $c(x) \geq 0$ as in (3.3) under the condition (2.3). Putting together (4.12) and (4.14) into (3.4), we have under Assumption 1 as follows:

$$\begin{aligned}a_{ii} &\geq \frac{1}{2}\sum_{\omega_{ij}\subset\omega_i}\left(\cot\theta_{ij}^+ + \cot\theta_{ij}^-\right)\left(-\frac{K_\eta}{3}\delta_{ij}^e(u_2)+k_\alpha\right) \\ &\geq \frac{\beta_m}{2}\left(-\frac{K_\eta}{3}\delta_{ij}^e(u_2)+k_\alpha\right), \quad (4.15)\end{aligned}$$

from which the result follows under condition (4.11). □

Based on the results of Lemmas 1 and 2, the matrix $A^\top$ can be seen to be diagonally dominant, with positive row sums corresponding to each vertex that neighbors the Dirichlet boundary.

Lemma 3 *Let Assumption 1 and Assumption 2 hold. Let $A = (a_{ij})$ be given by (3.4). Assume condition (4.1) of Lemma 1 holds true. Then, $A^\top$ has the DD+ property given in Definition 5, with positive row sums for each index i such that q_i neighbors the boundary.*

The proof of Lemma 3 is delayed until after Lemma 4 in the Appendix, from which it easily follows.

Theorem 6 *Let Assumption 1 and Assumption 2 hold. Assume condition (4.1) of Lemma 1 holds true. Then, matrix $A = (a_{ij})$ given by (3.4) is monotone.*

We prove this in two ways, first by Theorem 4, then by Theorem 5.

Proof For our first proof, we show that $A^\top$ is of positive type in accordance with Definition 6, which implies its monotonicity by Theorem 4. The first condition

that $A^\top$ is a Z-matrix has been established by Lemma 1. The second condition is demonstrated in the Appendix argument, Lemma 3, in which it is shown that $A^\top$ is diagonally dominant and each vertex q_l that neighbors the boundary corresponds to an index $l \in J(A^\top)$, meaning the lth row sum is positive. The third condition follows again from Lemma 1 where imposing the strict inequality implies that $a_{ij} \neq 0$ for edge-connected nodes q_i and q_j. Hence, under the Dirichlet boundary condition even if the mesh is not connected (and it need not be), there is a chain of nonzero entries $a_{ij}, a_{j,k_1}, \ldots, a_{k_p,l}$, connecting a_{ij} that corresponds to edge-connected vertices q_i, q_j to some $a_{k_p l}$ where q_l neighbors the boundary, meaning $l \in J(A^\top)$. More simply, the strict inequality in (4.1) implies the connectivity of the directed graph of A agrees with the connectivity of the mesh.

For our second proof, we show that $A^\top$ is a GDD Z-matrix, or in other words, satisfies Theorem 5. Again by Lemma 1, A (hence $A^\top$) is a Z-matrix. Then, by the construction of Appendix argument Lemma 4, there is a positive diagonal matrix D for which $A^\top D$ is strictly diagonally dominant.

In both cases, the monotonicity of A follows from the monotonicity of $A^\top$ by Proposition 1. □

The motivation to establish conditions under which matrix A of (3.4) is monotone is to establish the discrete comparison theorem and uniqueness result for the piecewise linear finite element solution. These results are summarized in the following corollary.

Corollary 1 *Let Assumption 1 and Assumption 2 hold. Assume condition* (4.1) *of Lemma 1 holds true for some* u_2. *Then, A given by* (3.4) *is monotone, and the comparison principle of Theorem 3 holds. Moreover, if condition* (4.1)— *given on the supersolution* u_2 *is satisfied by any solution* $u \in \mathcal{V}$ *to* (3.6), *then u is the unique solution to* (3.6).

Proof Apply the results of Theorem 6 to Theorem 3. □

The conditions given in Lemma 4 which imply the comparison theorem and uniqueness of the solution improve the conditions found in previous work by the authors. Other results in the literature [1, 2, 11] regarding uniqueness of discrete solution to (1.1) yield only asymptotic estimates making direct comparison difficult.

Remark 1 To illustrate the improved constant in the case where $g(x, u) = 0$, consider an equilateral mesh. The minimum ratio of sines is equal to one, and the cosine of each angle is $c_T = 1/2$. To put the result in the current notation, the Lipschitz constant L_0 is taken as K_η. Then, the condition for unqiueness found in [23, Theorem 3.4] for the 2D case reduces to the following:

$$\delta(u) < \frac{3k_\alpha}{14K_\eta}.$$

The conditions found in this investigation for the same problem (1.1) without the lower order term on an equilateral mesh are found with $\beta_m = 2/\sqrt{3}$, and $\beta_M = 4/\sqrt{3}$. Then, the requirement for uniqueness given by (4.1) is as follows:

$$\delta(u) < \frac{3k_\alpha}{2K_\eta},$$

which sharpens the constant by a factor of 7.

5 Conclusion

In this article, we established a discrete comparison theorem for (1.1), a quasilinear PDE with a solution-dependent lower order term. We established sufficient local and global conditions for the monotonicity of the assembled coefficient matrix for the PDE corresponding to the difference of two solutions. The monotonicity then implies uniqueness of the finite element solution under the given conditions which are seen to improve upon those in previous work. This argument is also seen to relax the angle conditions to allow some right triangles in the mesh, so long as the sum of angles opposite each edge remains bounded below π. Considering the elements of the assembled matrix rather than the integral over each individual element further allows a sharper local condition on the maximum difference between neighboring nodal values of a computed solution. As in previous work, we find the mesh should be globally fine if the PDE contains a lower order solution-dependent nonlinearity. Otherwise, the mesh is required to be fine where the gradient of the solution is steep, as the absolute difference in neighboring nodal values rather than the mesh size itself requires control. This is a condition that can easily be checked *a posteriori* once a discrete solution has been computed.

Acknowledgements SP would like to acknowldege Wright State University where a substantial portion of the writing was completed. Both authors would like to acknowledge the anonymous reviewers for suggestions that improved the clarity of the results.

Funding information SP was supported in part by NSF DMS 1719849 and 1852876. YZ was supported in part by NSF DMS 1319110.

Appendix

Here, we include the detailed arguments on the diagonal dominance requirements of matrix $A^\top$, with A given by (3.4). In Lemma 4 we show $A^\top$ is GDD in accordance with Theorem 5. From the computations therein, we show in Lemma 3 that $A^\top$ is also DD+, in accordance with Theorem 4, with positive row sums for each index i such that vertex q_i neighbors the Dirichlet boundary. Both GDD and DD+ properties hold on satisfaction of the conditions of Lemma 1.

We proceed to construct a positive diagonal matrix D_ε for which all row sums of $A^\top D_\varepsilon$ are positive. The diagonal elements d_i of D_ε are defined as an increasing sequence based on their distance from Γ_D. First, we require a notion of distance from the boundary.

Definition 7 Let p_i denote the length of the shortest path to a neighborhood of the Dirichlet boundary from vertex q_i. In particular, if $\bar{\mathcal{Q}}_i \setminus \mathcal{Q}_i \neq \emptyset$, then $p_i = 0$. Otherwise, p_i is defined to be the least number of edges traversed between q_i and any vertex q_j with $p_j = 0$.

This notion of distance to the boundary is well-defined regardless of the number of connected components comprising domain Ω.

Lemma 4 *Let Assumption 1 and Assumption 2 hold. Let $A = (a_{ij})$ be given by (3.4), and let $A^\top = (\alpha_{ij})$. Assume condition (4.1) of Lemma 1 holds true, and for some $\bar{\varepsilon} > 0$ it holds as follows:*

$$\frac{K_\eta \beta_M \delta(u_2)}{3\beta_m} < k_\alpha - \bar{\varepsilon}. \tag{6.1}$$

Let D_ε be the diagonal matrix with entries d_i given by the following:

$$d_i = 1 - \varepsilon_{p_i}, \quad \varepsilon_{p_i} = \varepsilon_{p_i - 1} - r^{p_i - 1}\delta_0, \ p_i \geq 1, \tag{6.2}$$

for p_i given by Definition 7, fixed $0 < \varepsilon_0 < 1$, and $0 < r, \delta_0 < 1$, to be defined below. Then, matrix $A^\top D_\varepsilon$ is strictly diagonally dominant, and condition (6.1) relaxes to condition (4.1) for A to be a Z-matrix, as $\bar{\varepsilon} \to 0$.

Proof First, it is noted that the sequence $\{\varepsilon_j\}$ from (6.2) is a strictly decreasing sequence. By summing the geometric terms in (6.2), we also see the sequence $\{\varepsilon_j\}$ is strictly positive if $\varepsilon_0 > \delta_0/(1-r)$, which will be assured as $\bar{\varepsilon} \to 0$ for fixed ε_0. As a result, the coefficients d_i are ordered by the distance of each q_i to the boundary according to Definition 7, and $d_i \to 1 + \delta_0/(1-r) - \varepsilon_0$, for increasing p_i.

By the positivity of the diagonals, and the nonpositivity of the off-diagonals, we require for each row i of the product AD_ε as follows:

$$d_i \alpha_{ii} + \sum_{j=1, j\neq i}^{n} d_j \alpha_{ij} > 0. \tag{6.3}$$

By slight abuse of notation, let $j \in \mathcal{Q}_i$ mean index j such that $q_j \in \mathcal{Q}_i$. Let $n = \text{card}(\mathcal{Q})$, the number of mesh degrees of freedom. For conciseness, $\kappa(x, u)$ will be

denoted $\kappa(u)$ in the remainder of the proof. Expanding (6.3) by (3.4) and rearranging terms yields the following:

$$\begin{aligned} d_i\alpha_{ii} + \sum_{j=1, j\neq i}^{n} d_j\alpha_{ij} \;=\; & d_i \int_{\omega_i} \kappa(u_1)\nabla\varphi_i\cdot\nabla\varphi_i \,\mathrm{d}x + \sum_{j\in\mathcal{Q}_i} d_j \int_{\omega_{ij}} \kappa(u_1)\nabla\varphi_i\cdot\nabla\varphi_j \,\mathrm{d}x \\ & + d_i \int_{\omega_i} \mathbf{b}(x,u_2)\cdot\nabla\varphi_i\varphi_i \,\mathrm{d}x + \sum_{j\in\mathcal{Q}_i} d_j \int_{\omega_{ij}} \mathbf{b}(x,u_2)\cdot\nabla\varphi_j\varphi_i \,\mathrm{d}x \\ & + d_i \int_{\omega_i} c(x)\varphi_i^2 \,\mathrm{d}x + \sum_{j\in\mathcal{Q}_i} d_j \int_{\omega_{ij}} c(x)\varphi_j\varphi_i \,\mathrm{d}x. \end{aligned} \tag{6.4}$$

The contribution to the total sum from the last line of (6.4) is strictly nonnegative and need not be considered further. Each of the first two lines of (6.4) is now considered with respect to the membership of each $q_j \in \bar{\mathcal{Q}}_i$ in one of three sets.

Define the sets as follows:

$$\mathcal{Q}_i^{-1} = \bar{\mathcal{Q}}_i \setminus \mathcal{Q}_i, \quad \text{and} \quad \mathcal{Q}_i^{p} := \{q_j \in \bar{\mathcal{Q}}_i \mid p_j = p\}, \;\; p \geq 0. \tag{6.5}$$

A first key observation for the following analysis is for each vertex $q_j \in \bar{\mathcal{Q}}_i$, q_j is in exactly one of $\mathcal{Q}_i^{p_i-1}$, $\mathcal{Q}_i^{p_i}$, $\mathcal{Q}_i^{p_i+1}$. A second key observation is at least one q_j in $\bar{\mathcal{Q}}_i$ is in $\mathcal{Q}_i^{p_i-1}$, meaning at least one neighbor of q_i is closer in the sense of Definition 6.5 to Γ_D. We now partition the terms of (6.4) into sums over each of these three sets. Again, let $j \in \mathcal{Q}_i^p$ mean index j for which $q_j \in \mathcal{Q}_i^p$, and for simplicity of notation, let p denote p_i. If $p = 0$, meaning vertex q_i neighbors the Dirichlet boundary, the contribution from the first line on the RHS of (6.4) is as follows:

$$-d_i \sum_{j\in\mathcal{Q}_i^{-1}} \int_{\omega_{ij}} \kappa(u_1)\nabla\varphi_i\cdot\nabla\varphi_j \,\mathrm{d}x + \sum_{j\in\mathcal{Q}_i} (d_j - d_i) \int_{\omega_{ij}} \kappa(u_1)\nabla\varphi_i\cdot\nabla\varphi_j \,\mathrm{d}x. \tag{6.6}$$

For $p > 0$, meaning vertex q_i is without neighbors on Γ_D, we have the following:

$$\sum_{j\in\mathcal{Q}_i} (d_j - d_i) \int_{\omega_{ij}} \kappa(u_1)\nabla\varphi_i\cdot\nabla\varphi_j \,\mathrm{d}x. \tag{6.7}$$

For $p = 0$, the contribution from the second line of (6.4) can be written as follows:

$$\begin{aligned} & d_i \sum_{j\in\mathcal{Q}_i^{-1}} (u_2(q_j) - u_2(q_i)) \int_{\omega_{ij}} \nabla\varphi_j\cdot\nabla\varphi_i b(x)\varphi_i \,\mathrm{d}x \\ & + \sum_{j\in\mathcal{Q}_i} (d_j - d_i)(u_2(q_i) - u_2(q_j)) \int_{\omega_{ij}} \nabla\varphi_i\cdot\nabla\varphi_j b(x)\varphi_i \,\mathrm{d}x \\ & + \sum_{j\in\mathcal{Q}_i} d_j \sum_{T\in\omega_{ij}} (u_2(q_{ij}^T) - u_2(q_j)) \int_{T} \nabla\varphi_{ij}^T\cdot\nabla\varphi_j b(x)\varphi_i \,\mathrm{d}x. \end{aligned} \tag{6.8}$$

Let $\mathcal{K}_{ij}^T := (u_2(q_{ij}^T) - u_2(q_j)) \int_T \nabla\varphi_{ij}^T \cdot \nabla\varphi_j b(x)\varphi_i \, \mathrm{d}x$. The third line of (6.8) can be expanded as follows:

$$\left(d_i \sum_{j\in\bar{\mathcal{Q}}_i} \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T \right) + \left(\sum_{j\in\mathcal{Q}_i} (d_j - d_i) \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T \right) - \left(d_i \sum_{j\in\bar{\mathcal{Q}}_i\setminus\mathcal{Q}_i} \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T \right)$$
$$= \left(\sum_{j\in\mathcal{Q}_i} (d_j - d_i) \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T \right) - \left(d_i \sum_{j\in\bar{\mathcal{Q}}_i\setminus\mathcal{Q}_i} \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T \right), \tag{6.9}$$

as the first term in the left of (6.9) is zero because the $\mathcal{K}_{ij}^T$ terms cancel pairwise when summed over the entire patch $\bar{\mathcal{Q}}_i$.

For $p > 0$, the contribution from the second line of (6.4) can be written as follows:

$$\sum_{j\in\mathcal{Q}_i} (d_j - d_i)(u_2(q_i) - u_2(q_j)) \int_{\omega_{ij}} \nabla\varphi_i \cdot \nabla\varphi_j b(x)\varphi_i \, \mathrm{d}x + \sum_{j\in\mathcal{Q}_i} d_j \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T, \tag{6.10}$$

where similarly to (6.9) but with $\bar{\mathcal{Q}}_i \setminus \mathcal{Q}_i = \emptyset$, the last sum over $j \in \mathcal{Q}_i$ can be written.

$$\sum_{j\in\mathcal{Q}_i} (d_j - d_i) \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T. \tag{6.11}$$

For the case $p = 0$, applying expansions (6.6), (6.8), and (6.9) to (6.4), we have the following:

$$d_i\alpha_{ii} + \sum_{j=1, j\neq i} d_j\alpha_{ij} \geq -d_i \sum_{j\in\mathcal{Q}_i^{-1}} \mathcal{J}_{ij} + \sum_{j\in\mathcal{Q}_i} (d_j - d_i)\mathcal{J}_{ij}, \tag{6.12}$$

with $\mathcal{J}_{ij} < 0$ given by the following:

$$\mathcal{J}_{ij} := \int_{\omega_{ij}} (\kappa(x, u_1) + (u_2(q_i) - u_2(q_j))b(x)\varphi_i)\nabla\varphi_i \cdot \nabla\varphi_j \, \mathrm{d}x + \sum_{T\in\omega_{ij}} \mathcal{K}_{ij}^T$$
$$\leq \frac{1}{2}\left(-k_\alpha\beta_m + \frac{K_\eta\beta_M\delta_{\omega_{ij}}(u_2)}{3} \right) < -\frac{\beta_m\bar{\varepsilon}}{2} < 0, \tag{6.13}$$

where the first inequality follows by the angle conditions of Assumption 2 and the second by condition (6.1), as in (4.7)–(4.8) without the lower order term. It is also important to note that taking into consideration the upper bound on κ given in (2.1), we have finite numbers $\mathcal{J}_L, \mathcal{J}^U$ for which as follows:

$$\frac{\beta_m\bar{\varepsilon}}{2} \leq \mathcal{J}_L \leq |\mathcal{J}_{ij}| \leq \mathcal{J}^U. \tag{6.14}$$

For $p = 0$, the sum over $\mathcal{Q}_i^{-1}$ contains at least one term and $d_i = 1 - \varepsilon_0$, whereas $d_j - d_i$ is either zero or $\varepsilon_0 - \varepsilon_1$. Let m be the maximum number of neighbors of any given vertex. A fixed maximum m is implied by the smallest-angle condition

(2.6). Together with theses observations, applying (6.13) and (6.14) to (6.12) yields the following:

$$d_i\alpha_{ii} + \sum_{j=1, j\neq i}^{n} d_j\alpha_{ij} \geq \frac{(1-\varepsilon_0)\beta_m\bar{\varepsilon}}{2} - (m-1)(\varepsilon_0-\varepsilon_1)\mathcal{J}^u. \tag{6.15}$$

From (6.2), $\varepsilon_0 - \varepsilon_1 = \delta_0$, so setting is as follows:

$$\delta_0 := \frac{\bar{\varepsilon}\beta_m(1-\varepsilon_0)}{2m\mathcal{J}^U}, \tag{6.16}$$

forces the row sum in (6.15) to be strictly positive for the case $p = 0$.

For the case $p > 0$, applying expansions (6.7), (6.10), and (6.11) to (6.4), and $\mathcal{J}_{ij} < 0$ from (6.13), we have the following:

$$\begin{aligned} d_i\alpha_{ii} + \sum_{j=1, j\neq i} d_j\alpha_{ij} &= \sum_{j\in\mathcal{Q}_i^{p-1}} (d_j - d_i)\mathcal{J}_{ij} + \sum_{j\in\mathcal{Q}_i^{p+1}} (d_j - d_i)\mathcal{J}_{ij} \\ &= \sum_{j\in\mathcal{Q}_i^{p-1}} (\varepsilon_p - \varepsilon_{p-1})\mathcal{J}_{ij} + \sum_{j\in\mathcal{Q}_i^{p+1}} (\varepsilon_p - \varepsilon_{p+1})\mathcal{J}_{ij} \\ &\geq (\varepsilon_{p-1} - \varepsilon_p)\mathcal{J}_L - (m-1)(\varepsilon_p - \varepsilon_{p+1})\mathcal{J}^U, \end{aligned} \tag{6.17}$$

where the first sum must be nonempty because at least one vertex $q_j \in \mathcal{Q}_i$ must be closer to the boundary than q_i. From the construction (6.2), (6.17) then implies the following:

$$d_i\alpha_{ii} + \sum_{j=1, j\neq i} d_j\alpha_{ij} \geq r^{p-1}\delta_0\mathcal{J}_L - (m-1)r^p\delta_0\mathcal{J}^U > 0, \tag{6.18}$$

for $r < \mathcal{J}_L/((m-1)\mathcal{J}^U)$. In particular, inequality (6.18) is satisfied for the following:

$$r := \frac{\mathcal{J}_L}{m\mathcal{J}^U} < 1. \tag{6.19}$$

It is finally noted for the sequence $\{\varepsilon_i\}$ given by (6.2) as follows:

$$\varepsilon_i > \varepsilon_0 - \frac{\delta_0}{1-r} = \varepsilon_0\left(1 + \frac{\bar{\varepsilon}\beta_m}{2(m\mathcal{J}^U - \mathcal{J}_L)}\right) - \frac{\bar{\varepsilon}\beta_m}{2(m\mathcal{J}^U - \mathcal{J}_L)} > 0,$$

for $\bar{\varepsilon}$ small enough.

These arguments together show the matrix $A^\top D_\varepsilon$ is strictly diagonally dominant for diagonal matrix D_ε defined by (6.2) with δ_0 given by (6.16) and r given by (6.19). The remainder of the result follows by sending $\bar{\varepsilon} \to 0$. □

Finally, we note that setting $d_i = 1$ for each i in the proof of Lemma 4 rather than rescaling the row sums as performed above provides a proof of Lemma 3 that $A^\top$ has the DD+ property required for Theorem 4, with positive row sums for each index i such that q_i neighbors the boundary.

Proof (Lemma 3) Indices i of $A^\top$ for which q_i neighbors the Dirichlet boundary are seen to have positive row sums by setting $d_i = d_j = 1$ in (6.12), and noting $\mathcal{J}_{ij} < 0$ under the given hypotheses. Indices i of $A^\top$ for which q_i does not neighbor the

boundary are seen to be zero by setting $d_i = d_j = 1$ in the first line of (6.17). This establishes $A^\top$ is DD+ under the conditions that A ($A^\top$) is a Z-matrix. □

References

1. Abdulle, A., Vilmart, G.: A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems. Numer. Math. **121**(3), 397–431 (2012)
2. André, N., Chipot, M.: Uniqueness and nonuniqueness for the approximation of quasilinear elliptic equations. SIAM J. Numer. Anal. **33**(5), 1981–1994 (1996)
3. Baker, B.S., Grosse, E., Rafferty, C.S.: Nonobtuse triangulation of polygons. Discrete Comput. Geom. **3**, 147–168 (1988)
4. Bank, R.E., Sherman, A.H., Weiser, A.: Refinement algorithms and data structures for regular local mesh refinement. In: Scientific Computing, pp. 3–17. IMACS/North-Holland Publishing Company, Amsterdam (1983)
5. Bouchon, F.: Monotonicity of some perturbations of irreducibly diagonally dominant M-matrices. Numer. Math. **105**(4), 591–601 (2007)
6. Bramble, J.H., Hubbard, B.E.: New monotone type approximations for elliptic problems. Math. Comp. **18**(87), 349–367 (1964)
7. Bramble, J.H., Hubbard, B.E.: On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type. J. Math. Phys. **43**(1–4), 117–132 (1964)
8. Brandts, J.H., Korotov, S., Křížek, M.: The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem. Linear Algebra Appl. **429**(10), 2344–2357 (2008). Special Issue in honor of Richard S. Varga
9. Carl, S., Le, V.K., Motreanu, D.: Nonsmooth variational problems and their inequalities: comparison principles and applications. Springer monographs in mathematics New York: Springer Science+Business Media (2007)
10. Collatz, L.: Functional Analysis and Numerical Mathematics. Translated from the German by Hansjörg Oser. Academic Press, New York (1966)
11. Douglas, J., Dupont, T.: A Galerkin method for a nonlinear Dirichlet problem. Math. Comput. **131**, 689–696 (1975)
12. Douglas, J., Dupont, T., Serrin, J.: Uniqueness and comparison theorems for nonlinear elliptic equations in divergence form. Arch. for Ration. Mech. Anal. **42**(3), 157–168 (1971)
13. Drăgănescu, A., Dupont, T.F., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comp. **74**(249), 1–23 (2005)
14. Erten, H., Üngör, A.: Computing triangulations without small and large angles. In: 2009 Sixth International Symposium on Voronoi Diagrams, pp. 192–201 (2009)
15. Fiedler, M., Pták, V.: On matrices with non-positive off-diagonal elements and positive principal minors. Czechoslov. Math. J. **12**(3), 382–400 (1962)
16. Gerver, J.L.: The dissection of a polygon into nearly equilateral triangles. Geom. Dedicata. **16**, 93–106 (1984)
17. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order Grundlehren Der Mathematischen Wissenschaften: 224. Springer, Berlin (1983)
18. Hlaváček, I., Křížek, M., Malý, J.: On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type. J. Math. Anal. Appl. **184**(1), 168–189 (1994)
19. Li, W.: On Nekrasov matrices. Linear Algeba Appl. **281**, 87–96 (1998)
20. Mangasarian, O.L.: Characterizations of real matrices of monotone kind. SIAM Rev. **10**(4), 439–441 (1968)
21. Plemmons, R.J.: M-matrix characterizations. I–nonsingular M-matrices. Linear Algebra Appl. **18**, 175–188 (1977)
22. Pollock, S., Zhu, Y.: Discrete comparison principles for quasilinear elliptic PDE. Submitted (2017)
23. Pollock, S., Zhu, Y.: Uniqueness of discrete solutions of nonmonotone PDEs without a globally fine mesh condition. Numer. Math. **139**(4), 845–865 (2018)
24. Shivakumar, P.N., Chew, K.H.: A sufficient condition for nonvanishing of determinants. Proc. Amer. Math. Soc. **43**(1), 63–66 (1974)

25. Shivakumar, P.N., Williams, J.J., Ye, Q., Marinov, C.A.: On two-sided bounds related to weakly diagonally dominant M-matrices with application to digital circuit dynamics. SIAM J. Matrix Anal. Appl. **17**(2), 298–312 (1996)
26. Trudinger, N.S.: On the comparison principle for quasilinear divergence structure equations. Arch. for Ration. Mech. and Anal. **57**(2), 128–133 (1974)
27. Varga, R.S.: On recurring theorems on diagonal dominance. Linear Algebra Appl. **13**(1), 1–9 (1976)
28. Varga, R.S.: Matrix Iterative Analysis, Springer Series in Computational Mathematics, Expanded Edn., vol. 7. Springer, Berlin (2000)
29. Vejchodský, T.: On the nonmonotony of nonlinear elliptic operators in divergence form. Adv. Math. Sci. Appl. **14**(1), 25–33 (2004)
30. Xu, J., Zikatanov, L.: A monotone finite element scheme for convection-diffusion equations. Math. Comput. **68**(228), 1429–1446 (1999)
31. Young, D.: Iterative Solution of Large Linear Systems. Academic Press Inc., New York (1971)