Applied Numerical Mathematics 156 (2020) 106–124

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Discrete comparison principles for quasilinear elliptic PDE

Sara Pollock [a,*], Yunrong Zhu [b]

[a] Department of Mathematics, University of Florida; Gainesville, FL 32611, United States of America
[b] Department of Mathematics and Statistics, Idaho State University; Pocatello, ID 83209, United States of America

ARTICLE INFO

Article history:
Received 18 July 2018
Received in revised form 26 February 2020
Accepted 17 April 2020
Available online 27 April 2020

Keywords:
Discrete comparison principle
Uniqueness
Nonlinear diffusion
Quasilinear partial differential equations
Nonmonotone problems

ABSTRACT

Comparison principles are developed for piecewise linear finite element approximations of quasilinear elliptic partial differential equations. We consider the analysis of a class of nonmonotone Leray-Lions problems featuring both nonlinear solution and gradient dependence in the principal coefficient, and a solution dependent lower-order term. Sufficient local and global conditions on the discretization are found for conforming finite element solutions to satisfy a comparison principle, which implies uniqueness of the solution. For problems without a lower-order term, our analysis shows the meshsize is only required to be locally controlled, based on the variance of the computed solution over each element. We include a discussion of the simpler semilinear case where a linear algebra argument allows a sharper mesh condition for the lower order term.

1. Introduction

We consider the finite element approximation of the quasilinear elliptic partial differential equation (PDE)

$$-\mathrm{div}(a(x,u,\nabla u)) + b(x,u) = 0 \text{ in } \Omega \subset \mathbb{R}^d, \tag{1.1}$$

where $a(x,\eta,\xi) = A(x,\eta,\xi)\xi$, for scalar-valued $A : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$, and $b : \Omega \times \mathbb{R} \to \mathbb{R}$. The domain Ω is assumed to be polygonal for $d = 2$, or an interval for $d = 1$. The boundary $\partial\Omega$ is decomposed into Dirichlet and Neumann parts, where the Dirichlet part $\Gamma_D \subseteq \partial\Omega$ has positive measure in $\mathbb{R}^{d-1}$, and the Neumann part is given by $\Gamma_N = \partial\Omega \setminus \Gamma_D$. The boundary conditions applied to (1.1) are either mixed Dirichlet/Neumann or homogeneous Dirichlet, given by

$$u = 0 \text{ on } \Gamma_D, \text{ and } a(x,u,\nabla u)\cdot n(x) = \psi(x) \text{ on } \Gamma_N, \tag{1.2}$$

for outward facing normal n. The aim of this paper is to extend the discrete comparison principle and uniqueness results recently obtained by the authors to a more general class of quasilinear elliptic equations.

Significant progress has been made on developing discrete maximum principles for divergence form quasilinear elliptic problems, as in [12,22–24,32], and developing the appropriate conditions on the angles of the mesh for these results to hold. In the nonlinear context, comparison principles rather than maximum principles for a given equation imply the uniqueness of solutions. Comparison principles also provide important information such as a natural ordering of solutions that can be useful in the analysis of numerical solutions. There are still only few results on discrete comparison principles for problem

[*] Corresponding author.
E-mail addresses: s.pollock@ufl.edu (S. Pollock), zhuyunr@isu.edu (Y. Zhu).

https://doi.org/10.1016/j.apnum.2020.04.013
0168-9274/© 2020 IMACS. Published by Elsevier B.V. All rights reserved.

(1.1), despite the significant literature on corresponding results for continuous problems, *e.g.*, [4,5,14,21,29], and [19, Chapter 10], and the references therein.

To our knowledge, the first comparison theorem which implies a global uniqueness result for a discrete version of problems in this class is that of [2], for the equation $-\mathrm{div}(\kappa(x,u)\nabla u) = f(x)$, where both a uniformly small meshsize (in an asymptotic sense for 2D) and an acuteness condition on the angles of the mesh were used. Uniqueness of solutions as the meshsize $h \to 0$ without a comparison principle was first shown in [13] and later generalized for simplicial and rectangular elements of arbitrary order under numerical quadrature [1]. The results of [1,13] while importantly establishing the computed finite element solution is a good approximation to the PDE solution, also require $\kappa(x,u)$ to be twice continuously differentiable with respect to u with bounded derivatives, and rely on the quasiuniformity of the underlying mesh (alternately, [1] also shows uniqueness of a finite element solution without the assumptions on the second derivative of κ but relying on an *a priori* H^2 norm bound on the PDE solution u). These assumptions may not be general enough for physical simulations, for example Richards' equation describing flow in partially saturated media where κ is generally Lipschitz but not differentiable with respect to u. While the requirement of an acute mesh as in [2,27] and the current presentation may increase the complexity of a simulation, acute meshing technology for general domains is available, *e.g.*, [16,18], and refinements can be performed to preserve acuteness for instance by red refinement [3]. Mesh generation and refinement of irregular domains can produce meshes that are highly nonuniform in meshsize, so for the purposes of physical modeling it is important to develop uniqueness results that can apply under these conditions, as we do here.

The meshsize assumption for P_1 elements was first relaxed in a comparison theorem framework in recent work by the current authors in [27], where the global meshsize condition was replaced by a local verifiable *a posteriori* condition on the maximum variance of the solution over each element, locally limiting the meshsize where the solution has steep gradients. As with [2], this result relied neither on the differentiability of κ nor the quasiuniformity of the mesh. The main contributions of the current manuscript are we now allow a more general diffusion coefficient, including a nonlinear dependence on the gradient; and, a (nonlinear) solution-dependent lower order term. The introduction of a lower order term does lead to a global meshsize condition, whereas without this term present the conditions are local, based on the maximum difference of nodal values in each element of the mesh. In either case, we present here a *verifiable* condition for uniqueness of a computed solution as opposed to an asymptotic one.

In the current presentation, maintaining nondegenerate ellipticity is essential to establishing the discrete comparison principle. While p-Laplacian type perturbations on the elliptic operator are allowed (for limited values of p), the elliptic operator is not permitted to degenerate, and it remains unknown whether a discrete comparison principle holds for degenerate-type elliptic operators. The current results allow the determination of whether the solution to a finite element approximation of (1.1) is unique, based only on knowledge of problem data, and accessible properties of the computed solution and the mesh. This information is useful in the analysis of adaptive algorithms (*e.g.*, [25,26]), or meshings of domains with a large variance in the relative diameters of elements, and can be used to verify the uniqueness of a discrete solution upon numerical convergence. Importantly, these results hold and can be applied based *a posteriori* estimates from a computed solution and without *a priori* knowledge of the solution to (1.1).

1.1. Problem class

The following assumptions on the diffusion coefficient, by means of the function $a_i(x,\eta,\xi) = A(x,\eta,\xi)\xi_i$, $i = 1,\dots,d$, for $x \in \Omega$, $\eta \in \mathbb{R}$, and $\xi \in \mathbb{R}^d$, are made throughout the remainder of the paper.

Assumption 1.1. Assume $a(x,\eta,\xi)$ and $b(x,\eta)$ are Carathéodory functions, C^1 in (η,ξ) (respectively, η) for almost every (a.e.) $x \in \Omega$, and measurable in x for each $(\eta,\xi) \in \mathbb{R} \times \mathbb{R}^d$, (respectively, for each $\eta \in \mathbb{R}$). Assume a is elliptic in the following sense. There is a positive constant γ_a with

$$\sum_{i,j=1}^{d} \frac{\partial a_i}{\partial \xi_j}(x,\eta,\xi)\zeta_i\zeta_j \ge \gamma_a|\zeta|^2, \tag{1.3}$$

for a.e. $x \in \Omega$, and for all $\eta \in \mathbb{R}$, $\xi \in \mathbb{R}^d$ and $\zeta \in \mathbb{R}^d$. There is a constant $K_\eta > 0$ with

$$\left|\frac{\partial A}{\partial \eta}(x,\eta,\xi)\right| \le K_\eta, \tag{1.4}$$

for a.e. $x \in \Omega$ and for all $\eta \in \mathbb{R}$ and $\xi \in \mathbb{R}^d$. Assume b is nondecreasing in η, and there is a constant $B_\eta \ge 0$ with

$$0 \le \frac{\partial b}{\partial \eta}(x,\eta) \le B_\eta, \tag{1.5}$$

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

The conditions of Assumption 1.1, used here to show a comparison theorem and uniqueness of the discrete solution, also satisfy the hypotheses of Theorem 10.7 of [19], under condition (ii), which shows a comparison theorem for the continuous problem.

Remark 1.2 *(Existence of solutions).* To understand existence of the PDE solution, it is useful to consider the Leray-Lions and coercivity conditions (see for example [9, Chapter 2]). In addition to the Carathéodory assumption above, the following conditions ensure the pseudo-monotonicity of the principal part of the elliptic operator.

(1) Growth condition: there is a function $k_0(x) \in L^q(\Omega)$ and $c_0 > 0$ with

$$|A(x,\eta,\xi)\xi_i| \le k_0(x) + c_0(|\eta|^{p-1} + |\xi|^{p-1}),\ i = 1,\dots,d,$$

with $1 < p < \infty$ and $1/p + 1/q = 1$.

(2) Monotonicity with respect to ξ: the coefficients $a_i = A\xi_i$ satisfy

$$\sum_{i=1}^{d}(A(x,\eta,\xi)\xi_i - A(x,\eta,\bar\xi)\bar\xi_i)(\xi_i - \bar\xi_i) > 0,$$

for a.e. $x \in \Omega$, all $\eta \in \mathbb{R}$, and for all $\xi, \bar\xi \in \mathbb{R}^d$ with $\xi \neq \bar\xi$.

(3) Coercivity: there is a constant $\nu > 0$ and a function $k(x) \in L^1(\Omega)$ with

$$\sum_{i=1}^{d} A(x,\eta,\xi)\xi_i^2 \ge \nu|\xi|^p - k(x),$$

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

Classes of problems satisfying the above conditions are well-studied in the literature with respect to existence of solutions and their boundedness properties. For instance, existence of solutions is shown in Chapter II.6 of [28], under the strengthened coercivity condition and additional growth condition on the lower order term

$$\sum_{i=1}^{d} A(x,\eta,\xi)\xi_i^2 \ge c_r|\xi|^p - K_r(K(x) + |\eta|^r), \quad b(x,\eta) \le K_r(k_0(x) + |\eta|^r),$$

for $k_0(x)$ from condition (1) above, some $1 \le r < p$ and $K(x) \in L^1(\Omega)$ (see [28, Lemma 6.4]).

Cases where both Assumptions 1.1 and conditions (1)-(3) above are satisfied are not uncommon. First, if in addition to Assumption 1.1, there are constants $0 < \lambda_A \le \Lambda_A$ with $\lambda_A \le A(x,\eta,\xi) \le \Lambda_A$, then conditions (1)-(3) hold with $p = q = 2$. This includes the case where $a_i(x,\eta,\xi) = A(x,\eta)\xi_i$, as in the earlier investigation [27], with $b \equiv 0$, which features applications to nonlinear heat conduction, for example [21]. More generally, these conditions hold if $A(x,\eta,\xi)$ has the form $A(x,\eta,\xi) = A_0(x,\eta) + A_1(x,\eta)f(|\xi|) + A_2(x)g(|\xi|)$, where A_0 is bounded away from zero, and $f(|\xi|)$ and $g(|\xi|)$ satisfy appropriate growth conditions. Problems of this form will be specifically considered in the discrete two dimensional case.

The discrete problems for monotone instances of the above classes, those in which the principal coefficient is independent of η, such as the p-Laplacian, are analyzed in for instance [6,7,12], and under stronger monotonicity and Lipschitz assumptions in [10,17], exploiting the variational structure of the problem to establish uniqueness without a comparison principle. A more general approximation strategy using a Hybrid High-Order method is presented in [11]. In that setting, strong convergence of the sequence of discrete solutions is found as the meshsize goes to zero for monotone problems, but the result holds only up to a subsequence if $a(x,\eta,\xi)$ maintains its η-dependence, *i.e.*, for nonmonotone problems (see [11, Theorem 4.6]). The emphasis of this article is establishing verifiable sufficient conditions for the uniqueness of the discrete solution for the case where $a(x,\eta,\xi)$ of (1.1) maintains its η-dependence, and is not then monotone (or variational, see [21]), but rather *pseudo-monotone*, as described above.

The weak form of (1.1) is given by integration against test functions v which lie in an appropriate subspace $V_{0,D} \subset V \subseteq H^1 \cap W^{1,p}$, where $V_{0,D} = \{v \in V \mid v = 0 \text{ on } \Gamma_D\}$, and p is determined by the particular problem class, as in Remark 1.2. The reader is referred to [9, §3.2] for detailed discussion on the existence and comparison results for the continuous Dirichlet problem. Then, the weak form of the problem is: find $u \in V_{0,D}$ such that

$$\int_\Omega a(x,u,\nabla u)\cdot\nabla v + b(x,u)v\,\mathrm{d}x = \int_{\Gamma_N} \psi(x)v\,\mathrm{d}s, \quad \text{for all } v \in V_{0,D}, \tag{1.6}$$

where the Neumann data $\psi(x)$ is assumed to be bounded and measurable. For the remainder of the paper, we proceed with conditions of Assumptions 1.1, and investigate the conditions under which a discrete comparison principle holds, assuming the existence of a discrete subsolution and supersolution, as defined in the next section.

The remainder of the article is structured as follows. In §2, we state the discretization, and introduce the framework for proving the discrete comparison principle. In §3, this framework is applied to the simple case of the one dimensional problem. Then, in §4, the two dimensional problem is considered. First, additional restrictions on the discretization (angle conditions) are introduced. Then, in §4.1, useful estimates for the technical lemmas of §4.3 are reviewed. The main 2D result, Theorem 4.9, follows in §4.4. In §5 we prove a comparison principle for a simpler semilinear problem based on the previous estimates. In Theorem 5.3, we then apply a linear algebraic approach to improve the mesh condition.

2. Overview of comparison framework

We next overview the discretization and the comparison theorem framework. The subsequent sections contain the precise results and technical proofs. The cases of one and two dimensions are worked out separately to give explicit constants that can be used as criteria for verifying uniqueness of a discrete solution on a given mesh.

2.1. Discretization

Let $\mathcal{T}$ be a conforming partition of domain Ω that exactly captures the boundary of Ω, and each of Γ_D and Γ_N. In one dimension, $\mathcal{T}$ is a collection of intervals, and in two dimensions a triangulation. Let $\overline{\mathcal{D}}$ be the collection of vertices or nodes of $\mathcal{T}$, and let $\mathcal{D} = \overline{\mathcal{D}} \setminus \Gamma_D$. The nodes $a \in \mathcal{D}$ correspond to the mesh degrees of freedom. Let $\mathcal{V} := \mathcal{V}_{0,D} \subset V_{0,D}$ be the discrete space spanned by the piecewise linear basis functions $\{\varphi_j\}$ that satisfy $\varphi_i(a_j) = \delta_{ij}$ for each $a_j \in \mathcal{D}$.

For simplicity of defining the finite element solution space, the discussion assumes a homogeneous Dirichlet part under either the mixed or pure Dirichlet conditions. The method of the proof trivially generalizes to allow nonhomogeneous bounded measurable Dirichlet data, as its contribution is subtracted off as is the Neumann data, on the first step.

2.2. Discrete comparison framework

The discrete Galerkin problem for $\mathcal{V}$ is: find $u \in \mathcal{V}$ such that

$$\int_{\Omega} a(x,u,\nabla u)\cdot\nabla v + b(x,u)v \,\mathrm{d}x = \int_{\Gamma_N} \psi(x) v \,\mathrm{d}s, \quad \text{for all } v \in \mathcal{V}. \tag{2.1}$$

A *subsolution* to (2.1) is a function $u_1 \in \mathcal{V}$ with

$$\int_{\Omega} a(x,u_1,\nabla u_1)\cdot\nabla v + b(x,u_1)v \,\mathrm{d}x - \int_{\Gamma_N} \psi(x) v \,\mathrm{d}s \le 0, \tag{2.2}$$

for all $v \in \mathcal{V}^+ = \{v \in \mathcal{V} \mid v \ge 0\}$. A corresponding *supersolution* $u_2 \in \mathcal{V}$ is given by

$$\int_{\Omega} a(x,u_2,\nabla u_2)\cdot\nabla v + b(x,u_2)v \,\mathrm{d}x - \int_{\Gamma_N} \psi(x) v \,\mathrm{d}s \ge 0, \quad \text{for all } v \in \mathcal{V}^+. \tag{2.3}$$

Subtracting (2.3) from (2.2), we find

$$\int_{\Omega} (a(x,u_1,\nabla u_1) - a(x,u_2,\nabla u_2))\cdot\nabla v + (b(x,u_1)v - b(x,u_2))v \,\mathrm{d}x \le 0, \tag{2.4}$$

for all $v \in \mathcal{V}^+$. Decomposing the principal part by $a(x,u,\nabla u) = A(x,u,\nabla u)\nabla u$, and applying Taylor's theorem, it holds for $w = u_1 - u_2$ that

$$\begin{aligned}&(a(x,u_1,\nabla u_1) - a(x,u_1,\nabla u_2)) + (A(x,u_1,\nabla u_2) - A(x,u_2,\nabla u_2))\nabla u_2\\ &= \int_0^1 \frac{\partial a}{\partial \xi}(x,u_1,\nabla z(t))\nabla w \,\mathrm{d}t + \int_0^1 \frac{\partial A}{\partial \eta}(x,z(t),\nabla u_2) w \nabla u_2 \,\mathrm{d}t,\end{aligned} \tag{2.5}$$

for $z(t) = tu_1 + (1-t)u_2$. Similarly for the lower order term

$$b(x,u_1) - b(x,u_2) = \int_0^1 \frac{\partial b}{\partial \eta}(x,z(t)) w \,\mathrm{d}t. \tag{2.6}$$

Applying (2.5) and (2.6) to (2.4), and breaking the integral over the global domain into a sum of integrals over each element $T \in \mathcal{T}$, obtain

$$\int_\Omega \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) \nabla w \cdot \nabla v + \left(\frac{\partial A}{\partial \eta}\right) w \nabla u_2 \cdot \nabla v + \left(\frac{\partial b}{\partial \eta}\right) w v \, \mathrm{d}t \, \mathrm{d}x$$
$$= \sum_{T \in \mathcal{T}} \int_T \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) \nabla w \cdot \nabla v + \left(\frac{\partial A}{\partial \eta}\right) w \nabla u_2 \cdot \nabla v + \left(\frac{\partial b}{\partial \eta}\right) w v \, \mathrm{d}t \, \mathrm{d}x \le 0, \tag{2.7}$$

for all $v \in \mathcal{V}^+$. Here as in the remainder of the article the arguments may be suppressed for quantities which are ultimately bounded by constants in the analysis. The structure of $a(x, u, \nabla u) = A(x, u, \nabla u)\nabla u$ is exploited in the first term of the above decomposition to yield a quantity that is strictly positive, and in the second term to create a quantity controlled by the difference in nodal values of u_2. This factorization is a key component of the problem class that allows a condition for uniqueness similar to that in [27], dependent on the variance of the discrete solution u (or supersolution u_2) over each element.

The proof of the comparison principle follows by considering a particular test function $v \in \mathcal{V}^+$, and finding under Assumption 1.1 and additional assumptions on the discretization, that if $w > 0$ anywhere, the left hand side integration over elements of (2.7) is strictly positive, yielding a contradiction and implying $w \le 0$ everywhere, hence $u_1 \le u_2$ in Ω. Common test functions for this purpose in the continuous context include the positive part of $w = u_1 - u_2$, possibly taken to some power, as in [4,5]. In the discrete setting, the positive part of w is generally not a member of the finite element space, so a discrete version of this function can be used, as in [32]. In this case, as in [2,27], it is convenient to define a simpler test function v as follows.

Definition 2.1. Let $u_1 \in \mathcal{V}$ be a subsolution of (2.1) as in (2.2), and let $u_2 \in \mathcal{V}$ be a supersolution as in (2.3). Let $w = u_1 - u_2 \in \mathcal{V}$. Define the test function $v \in \mathcal{V}^+ \subset \mathcal{V}$ by its nodal values at each $a \in \mathcal{D}$ as

$$v(a) = \begin{cases} 1, & w(a) > 0, \\ 0, & w(a) \le 0. \end{cases} \tag{2.8}$$

If $w > 0$ anywhere on Ω, then $v(a)$ is nonzero for some $a \in \mathcal{D}$. One of the convenient properties of this test function v, is that $\nabla v = 0$ over each $T \in \mathcal{T}$ where w does not change sign. In fact, for the 1D case, an even simpler test function can be defined for which v' is supported over no more than two elements. This strategy was used in [27]; however, in this presentation we will use the same Definition 2.1 for both one and two dimensions to unify the arguments.

Partition the set $\mathcal{T}$ into subsets $\mathcal{T}_+, \mathcal{T}_-$ and $\mathcal{T}_c$, by the value of v from Definition 2.1, restricted to each element in $\mathcal{T}$.

$$\mathcal{T}_+ = \{T \in \mathcal{T} \mid v(x)\big|_T \equiv 1\}, \ \mathcal{T}_- = \{T \in \mathcal{T} \mid v(x)\big|_T \equiv 0\}, \ \mathcal{T}_c = \mathcal{T} \setminus \{\mathcal{T}_+ \cup \mathcal{T}_-\}. \tag{2.9}$$

Write the integral over Ω in (2.7) as

$$\int_\Omega = \int_{\bigcup_{T \in \mathcal{T}_+}} + \int_{\bigcup_{T \in \mathcal{T}_-}} + \int_{\bigcup_{T \in \mathcal{T}_c}} .$$

Each integral over $T \in \mathcal{T}_-$ is trivially zero. Each integral over $\mathcal{T} \in \mathcal{T}_+$ satisfies $\nabla v \equiv 0$, and the remaining lower order part is nonnegative by

$$\int_{T \in \mathcal{T}_+} \int_0^1 \frac{\partial b}{\partial \eta}(x, z(t)) w v \, \mathrm{d}t \, \mathrm{d}x = \int_{T \in \mathcal{T}_+} \int_0^1 \frac{\partial b}{\partial \eta}(x, z(t)) w \, \mathrm{d}t \, \mathrm{d}x \ge 0, \tag{2.10}$$

as $w > 0$, $v = 1$ and $\partial b / \partial \eta \ge 0$, by (1.5) of Assumption 1.1. It remains then to bound the integrals over $T \in \mathcal{T}_c$ where w changes sign. In summary, we have from (2.4), (2.7) and (2.10) that

$$0 \ge \int_\Omega (a(x, u_1, \nabla u_1) - a(x, u_2, \nabla u_2)) \cdot \nabla v + (b(x, u_1)v - b(x, u_2))v \, \mathrm{d}x$$
$$\ge \sum_{T \in \mathcal{T}_c} \int_T \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) \nabla w \cdot \nabla v + \left(\frac{\partial A}{\partial \eta}\right) w \nabla u_2 \cdot \nabla v + \left(\frac{\partial b}{\partial \eta}\right) w v \, \mathrm{d}t \, \mathrm{d}x, \tag{2.11}$$

for v given by Definition 2.1. We next develop conditions on the discretization in one and two dimensions for which the above inequality cannot hold.

3. Results for one dimension

Let $\Omega = (\alpha, \beta)$, with a subdivision

$$\alpha = a_0 < a_1 < \ldots < a_{n-1} < a_n = \beta, \tag{3.1}$$

where the mesh spacing is not assumed to be uniform. Define the intervals $\mathcal{I}_k = (a_{k-1}, a_k)$, $k = 1, \ldots, n$, and let $h_k = a_k - a_{k-1}$, the length of each respective interval. Then $\mathcal{T} = \cup_{1 \le k \le n} \{\overline{\mathcal{I}}_k\}$. Let $v' = \mathrm{d}v/\mathrm{d}x$. In one dimension, for the mixed problem with Dirichlet conditions at $x = \beta$, with Neumann data $\psi(\alpha) \in \mathbb{R}$, the weak form (1.6) reduces to: find $u \in \mathcal{V} := \mathcal{V}_{0,\beta}$ such that

$$\int_\Omega a(x, u, u')v' + b(x, u)v \, \mathrm{d}x = \psi(\alpha)v(\alpha) \ \text{ for all } v \in \mathcal{V}. \tag{3.2}$$

For the pure Dirichlet problem, (1.6) reduces to: find $u \in \mathcal{V} := \mathcal{V}_0$ such that

$$\int_\Omega a(x, u, u')v' + b(x, u)v \, \mathrm{d}x = 0, \ \text{ for all } v \in \mathcal{V}. \tag{3.3}$$

Without confusion, the discrete space $\mathcal{V}$ refers to either $\mathcal{V}_{0,\beta}$, containing the piecewise linear functions that vanish at $x = \beta$ for problem (3.2); or, $\mathcal{V}_0$, containing functions that vanish at $x = \alpha$ and $x = \beta$ for problem (3.3).

Theorem 3.1 *(One dimensional comparison theorem). Let u_1 be a subsolution as in (2.2) of either the mixed problem (3.2) or the Dirichlet problem (3.3); and, let u_2 be a supersolution as in (2.3), of the same problem. Assume the conditions of Assumption 1.1, and*

$$\max_{1 \le k \le n} \left\{ |u_2(a_k) - u_2(a_{k-1})| + \left(\frac{B_\eta}{K_\eta}\right) h_k^2 \right\} < \frac{2\gamma_a}{K_\eta}. \tag{3.4}$$

Then, it holds that $u_1 \le u_2$ in Ω.

If the lower order term b is independent of u, then $B_\eta = 0$, and the condition (3.4) is similar to that in [27], for a more general diffusion coefficient. If, on the other hand, $B_\eta > 0$, a global mesh condition is introduced, as $h_k < \sqrt{2\gamma_a/B_\eta}$ for all $k = 1, \ldots, n$, is a necessary condition for satisfaction of (3.4).

The proof of Theorem 3.1 follows by using the test function v from Definition 2.1 to show the right hand side of (2.11) is strictly positive.

Proof. Assume $w = u_1 - u_2$, is positive somewhere in Ω. Then $\mathcal{T}_c$ is nonempty, and in one dimension, inequality (2.11) reduces to

$$0 \ge \sum_{\mathcal{I}_k \in \mathcal{T}_c} \int_{\mathcal{I}_k} \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) w'v' + \left(\frac{\partial A}{\partial \eta}\right) u'wv' + \left(\frac{\partial b}{\partial \eta}\right) wv \, \mathrm{d}t \, \mathrm{d}x. \tag{3.5}$$

Proceed by bounding each term on the right hand side of (3.5). On each interval $\mathcal{I}_k \in \mathcal{T}_c$, w changes sign, and by Definition 2.1 the functions w' and v' are constants with the same sign. Then, the product $w'v' = |w(a_k) - w(a_{k-1})|/h_k^2$ on $\mathcal{I}_k$, and it holds that

$$\begin{aligned} \int_{\mathcal{I}_k} \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) w'v' \, \mathrm{d}t \, \mathrm{d}x &= \frac{|w(a_k) - w(a_{k-1})|}{h_k^2} \int_{\mathcal{I}_k} \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) \mathrm{d}t \, \mathrm{d}x \\ &\ge \frac{|w(a_k) - w(a_{k-1})|}{h_k} \gamma_a, \end{aligned} \tag{3.6}$$

where γ_a is the constant from (1.3). For the second term of (3.5), it is useful to note that $\int_{\mathcal{I}_k} |w| \le |w(a_k) - w(a_{k-1})| h_k/2$, as precisely one of $w(a_k)$ and $w(a_{k-1})$ must be positive. Then

$$\begin{aligned} \int_{\mathcal{I}_k} \int_0^1 \left(\frac{\partial A}{\partial \eta}\right) u'wv' \, \mathrm{d}t \, \mathrm{d}x &\ge \frac{-K_\eta |u_2(a_k) - u_2(a_{k-1})|}{h_k^2} \int_{\mathcal{I}_k} |w| \, \mathrm{d}x \\ &\ge -\left(\frac{|w(a_k) - w(a_{k-1})|}{h_k}\right) \frac{K_\eta |u_2(a_k) - u_2(a_{k-1})|}{2}, \end{aligned} \tag{3.7}$$

where K_η is the constant from (1.4). Each integral over last term of (3.5) satisfies

$$\int_{\mathcal{I}_k}\int_0^1\left(\frac{\partial b}{\partial\eta}\right)wv\,\mathrm{d}t\,\mathrm{d}x \geq -B_\eta\int_{\mathcal{I}_k}|w|\,\mathrm{d}x \geq -|w(a_k)-w(a_{k-1})|\frac{B_\eta h_k}{2}, \tag{3.8}$$

where B_η is the constant from (1.5). Putting (3.6), (3.7) and (3.8) together into (3.5) yields

$$0 \geq \sum_{\mathcal{I}_k\in\mathcal{T}_c}\frac{|w(a_k)-w(a_{k-1})|}{h_k}\left(\gamma_a-|u_2(a_k)-u_2(a_{k-1})|\frac{K_\eta}{2}-h_k^2\frac{B_\eta}{2}\right)>0,$$

where the strict positivity in the last inequality holds under the condition (3.4). This contradiction establishes that $w = u_1 - u_2$ cannot be positive anywhere on Ω. □

As any solution u to (3.2) or (3.3) is both a subsolution and a supersolution, the uniqueness of solutions follows, under the assumption

$$\max_{1\leq k\leq n}\left\{|u(a_k)-u(a_{k-1})|+\left(\frac{B_\eta}{K_\eta}\right)h_k^2\right\}<\frac{2\gamma_a}{K_\eta}. \tag{3.9}$$

The constants γ_a, B_η and K_η are based purely on the problem data, and if they are known or can be approximated for a given problem, then (3.9) can be easily and efficiently checked, and used to determine uniqueness of a given computed solution. It is important in particular for adaptive algorithms to have such a condition which ensures the uniqueness of the discrete solution without unavailable *a priori* knowledge. As demonstrated by the counterexamples of [2] (*cf.* [27]), some conditions on the discretization are indeed necessary to ensure the uniqueness of the solution.

4. Results for two dimensions

We next establish the uniqueness of the piecewise linear finite element solution to (2.1) in two dimensions, under Assumption 1.1. The simplicial mesh is assumed to be uniformly acute, and the smallest angle to be bounded away from zero.

Assumption 4.1 *(Mesh regularity).* There are numbers $0 < t_{min} \leq t_{max} < \pi/2$, for which the interior angles $\theta_i,\ i = 1, 2, 3$, of each $T \in \mathcal{T}$ satisfy

$$t_{min} \leq \theta_i \leq t_{max},\ i = 1, 2, 3. \tag{4.1}$$

Define the quantities

$$s_{min} = \sin(t_{min}),\ \text{ and } c_{min} = \cos(t_{max}). \tag{4.2}$$

The acuteness condition which states that angles are bounded below $\pi/2$, agrees with that in [27] for the simpler case of $a(x, \eta, \xi) = A(x, \eta)\xi$. In the following analysis, the condition that the angles are bounded away from zero is used to control the maximum ratio of edge-lengths in any triangle.

The relation between the measure of each element T, and the lengths of the sides are given by standard trigonometric descriptions, is summarized below. For each $T \in \mathcal{T}$, let $|T|$ denote the two-dimensional measure, or area. For any two distinct edges e_i and e_j, the area is $|T| = |e_i||e_j|\sin\theta_k/2$, for θ_k the interior angle between edges e_i and e_j. This provides the useful formula $|e_i||e_j|/|T| = 2/\sin\theta_k$. Under Assumption 4.1, the ratio of the sines of any pair of angles in a triangle T is bounded away from zero. Define the local constants

$$c_T := \min_{i,j=1,2,3}\cos\theta_i,\quad s_T := \max_{i,j=1,2,3}\sin\theta_i,\quad r_T := \min_{i,j=1,2,3}\frac{\sin\theta_i}{\sin\theta_j}, \tag{4.3}$$

for $\theta_i,\ i = 1, 2, 3$, the angles of T. The constant r_T is used to relate the lengths of edges of triangle T by

$$r_T|e_i| \leq |e_j| \leq r_T^{-1}|e_i|,\ i, j = 1, 2, 3. \tag{4.4}$$

Each vertex corresponding to a mesh degree of freedom, $a \in \mathcal{D}$, has coordinates $a = (x_1, x_2) \in \overline{\Omega}\setminus\Gamma_D$. It is recalled from §2.1 that $\mathcal{V} := \mathcal{V}_{0,D}$ is the piecewise linear Lagrange finite element space subordinate to partition $\mathcal{T}$, that vanishes on Γ_D in the sense of the trace.

4.1. Relations between gradients of basis functions

To clarify the technical lemmas that follow, some standard notations and properties of piecewise linear finite elements in two dimensions are now reviewed. The following relations involving gradients of basis functions are used often in the analysis.

Let $\{a_1, a_2, a_3\}$ be a local counterclockwise numbering of the vertices of a simplex $T \in \mathcal{T}$. Let the corresponding edges $\{e_1, e_2, e_3\}$, follow a consistent local numbering, with edge e_i opposite vertex a_i, $i = 1, 2, 3$. Let φ_i be the basis function on element $T \in \mathcal{T}$ defined by its nodal values at the vertices of T.

$$\varphi_i(a_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}, \; i, j = 1, 2, 3.$$

The inner product between gradients of basis functions and their respective integrals over elements $T \in \mathcal{T}$, may be computed by change of variables to a reference element $\widehat{T}$, in reference domain variables $(\widehat{x}_1, \widehat{x}_2)$. Specifically, the coordinates of $\widehat{T}$ are given as $\widehat{a}_1 = (0, 0)^T$, $\widehat{a}_2 = (1, 0)^T$, $\widehat{a}_3 = (0, 1)^T$. The Jacobian of the transformation between reference coordinates $\widehat{x} = (\widehat{x}_1, \widehat{x}_2)^T$, and physical coordinates $x = (x_1, x_2)^T$, is given by $J\widehat{x} = (x - a_1)$, with $J = \begin{pmatrix} a_2 - a_1 & a_3 - a_1 \end{pmatrix}$, for which $\det J = 2|T|$, with $|T|$ the area of triangle T. The reference element $\widehat{T}$ is equipped with the nodal basis functions $\widehat{\varphi}_i$, $i = 1, 2, 3$, where $\widehat{\varphi}_1 = 1 - \widehat{x}_1 - \widehat{x}_2$, $\widehat{\varphi}_2 = \widehat{x}_1$, $\widehat{\varphi}_3 = \widehat{x}_2$. The gradients $\widehat{\nabla}$ are taken with respect to the reference domain variables $\widehat{x}_1$ and $\widehat{x}_2$, and the transformation of gradients between the physical and reference domains is given by $\nabla\varphi_i = J^{-T}\widehat{\nabla}\widehat{\varphi}_i$. The gradients of basis functions satisfy the identity $\nabla\varphi_i + \nabla\varphi_j = -\nabla\varphi_k$, for any distinct assignment of i, j and k to the integers $\{1, 2, 3\}$. This allows the representation of $\nabla\varphi_i^T\nabla\varphi_i$ in terms of edge-length $|e_i|$. The maximum interior angle $t_{\max} < \pi/2$ from Assumption 4.1 then ensures $\nabla\varphi_i^T\nabla\varphi_j < 0$, for any $i \neq j$. The inner products between gradients in each element T satisfy the following identities:

$$\nabla\varphi_i^T\nabla\varphi_i = \frac{1}{4|T|^2}|e_i|^2, \text{ and } \nabla\varphi_i^T\nabla\varphi_j = \frac{-1}{4|T|^2}|e_i||e_j|\cos\theta_k, \; i \neq j. \tag{4.5}$$

4.2. Additional assumptions for the 2D problem

We next establish estimates which demonstrate for any $T \in \mathcal{T}_c$, given by (2.9), that

$$\int_T \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) \nabla w \cdot \nabla v + \left(\frac{\partial A}{\partial \eta}\right) w \nabla u_2 \cdot \nabla v + \left(\frac{\partial b}{\partial \eta}\right) w v \, dt \, dx > 0, \tag{4.6}$$

with v the test function given by Definition 2.1. In light of (2.11), this establishes by contradiction that $w = u_1 - u_2$ is nowhere positive. To bound the leading term of (4.6) away from zero, some additional restrictions on the nonlinear diffusion coefficient A are now considered.

Assumption 4.2. Assume $A(x, \eta, \xi)$ is of the form

$$A(x, \eta, \xi) = A_0(x, \eta) + A_1(x, \eta) f(|\xi|) + A_2(x) g(|\xi|). \tag{4.7}$$

Assume there is a positive constant λ_0, and there are nonnegative Λ_1 and Λ_2, with

$$A_0(x, \eta) \geq \lambda_0, \quad 0 \leq A_1(x, \eta) \leq \Lambda_1, \text{ and } 0 \leq A_2(x) \leq \Lambda_2, \tag{4.8}$$

for a.e. $x \in \Omega$, and all $\eta \in \mathbb{R}$, and $\xi \in \mathbb{R}^2$.

Assume $f(s), g(s) \geq 0$, and f satisfies the following growth condition. There is a constant C_f with

$$s|f'(s)| \leq C_f, \text{ for all } s \geq 0. \tag{4.9}$$

Assume g satisfies one of the two following conditions.

$$s|g'(s)| \leq C_g, \text{ for all } s \geq 0, \tag{4.10}$$

$$s|g'(s)| \leq \widehat{C}_g g(s), \text{ for all } s \geq 0, \text{ with } 0 \leq \widehat{C}_g \leq c_{min}. \tag{4.11}$$

The function g, which multiplies $A_2(x)$ in (4.7), is not assumed to be either bounded or bounded away from zero, while the boundedness of f, which multiplies $A_1(x, \eta)$ is required from (1.4) of Assumption 1.1.

Under condition (4.7) of Assumption 4.2, it holds that

$$\frac{\partial}{\partial \xi} a(x, \eta, \xi) = A(x, \eta, \xi) I + \big(A_1(x, \eta) f'(|\xi|) + A_2(x) g'(|\xi|)\big) \xi \xi^T,$$

which under (4.8), and $f, g \geq 0$, maintains the strict positive definiteness condition on the principal part of the elliptic operator from (1.3) of Assumption 1.1. In the technical lemmas that follow, this is used to ensure the boundedness away from zero of terms of the form $\nabla\varphi_i^T(\partial a/\partial\xi)\nabla\varphi_i$. However, the analysis depends also on the strict negativity of the product between gradients of distinct basis functions modified by $(\partial a/\partial\xi)$, namely, $\nabla\varphi_i^T(\partial a/\partial\xi)\nabla\varphi_k$. The role of the conditions (4.9)-(4.11) on f and g, used in conjunction with Assumption 4.1 in Lemma 4.5 and Corollary 4.6, is to ensure this holds. This technicality of the discrete (finite element) analysis is also the reason for the acute angle condition of Assumption 4.1, which itself regulates the negativity of $\nabla\varphi_i^T\nabla\varphi_k$, by (4.5).

The two properties on the products between gradients of basis functions on each element of the mesh that follow from Assumptions 4.1 and 4.2,

$$\nabla\varphi_i^T(\partial a/\partial\xi)\nabla\varphi_i|_T \geq \gamma_a\nabla\varphi_i^T\nabla\varphi_i|_T, \text{ and } \nabla\varphi_i^T(\partial a/\partial\xi)\nabla\varphi_k|_T \leq -c|\nabla\varphi_i|_T|\nabla\varphi_k|_T,$$

where $c > 0$ depends on the uniform acuteness of the mesh, allow a term proportional to $|w|$ to be factored out of the first as well as the second two terms in each summand of (2.11). This is shown in Lemma 4.5 under condition (4.10), and in Corollary 4.6 under condition (4.11). Under the local and global conditions on the mesh as developed in Lemmas 4.7 and 4.8, which ensure the second two terms of (2.11) are small enough multiples of $|w|$, it becomes clear that (4.6) holds. This contradicts the earlier assumption of (2.11), and leads to the conclusion that $w = u_1 - u_2$ is nowhere positive.

Functions f and g that satisfy Assumption 4.2 are not uncommon in modeling. Some examples are given in the next remark.

Remark 4.3. Admissible functions $\phi(|\xi|)$ that satisfy (4.9), (4.10) for the gradient-dependent part of the elliptic operator include the following.

$$\phi(|\xi|) = (\kappa + |\xi|^2)^{-\alpha}, \text{ for } \kappa > 0 \text{ and } \alpha \geq 0,$$

which appears for instance as the diffusion coefficient in the equation for capillarity (see [19, Chapter 10]) as well as the equations of prescribed mean curvature (see [31]), with $\kappa = 1$ and $\alpha = 1/2$.

$$\phi(|\xi|) = 2\left(K_0 + \sqrt{K_0^2 + 4|\xi|}\right)^{-1}, \quad K_0 > 0,$$

which is numerically investigated as a specific explicit case of the more general implicitly defined coefficient used in the modeling of glacier ice, as analyzed in [20]. Other common functions that satisfy (4.10) include

$$\phi(|\xi|) = \arctan(|\xi|), \text{ and } \phi(|\xi|) = \tanh(|\xi|).$$

Unbounded functions that satisfy (4.10) include

$$\phi(|\xi|) = \log(\kappa + |\xi|^2), \ \kappa > 1,$$

which allows for $g(|\xi|)$ hence $A(x, \eta, \xi)$ to be unbounded, albeit with slow growth.

Functions satisfying (4.11) include those of p-Laplacian type, for p close to 2,

$$\phi(|\xi|) = |\xi|^{p-2}, \text{ for } |p-2| < c_{min},$$

discussed further in Example 4.11.

4.3. Technical lemmas in two dimensions

An important quantity in the analysis is the maximum variance of a function over a given element. For piecewise linear functions, this is simply the maximum difference between any two vertex values on a given triangle.

Definition 4.4. For a function $\phi \in \mathcal{V}$, and element $T \in \mathcal{T}$, define $\delta_T(\phi)$ as follows.

$$\delta_T(\phi) = \max_{i,j=\{1,2,3\}} |\phi(a_i) - \phi(a_j)|, \tag{4.12}$$

where $\{a_1, a_2, a_3\}$, are the vertices of T.

In the technical lemmas which bound each term in the expansion (4.6), the following identity is used repeatedly.

$$\begin{aligned}\nabla\phi &= \phi(a_i)\nabla\varphi_i + \phi(a_j)\nabla\varphi_j + \phi(a_k)\nabla\varphi_k \\ &= \phi(a_i)\nabla\varphi_i + \phi(a_j)\nabla(\varphi_j + \varphi_k) + (\phi(a_k) - \phi(a_j))\nabla\varphi_k \\ &= (\phi(a_i) - \phi(a_j))\nabla\varphi_i + (\phi(a_k) - \phi(a_j))\nabla\varphi_k.\end{aligned} \tag{4.13}$$

The first Lemma characterizes the strict positivity of the first term of (4.6)

Lemma 4.5. *Let Assumptions 1.1, 4.1 and 4.2 hold, with g satisfying (4.10). Let $w, u \in \mathcal{V}$, and $z(t) \in \mathcal{V}$, $0 \le t \le 1$. Let a_i, a_j and a_k be the three vertices of $T \in \mathcal{T}_c$, ordered so that $w(a_i) \ge w(a_j) \ge w(a_k)$ with $w(a_i) > 0$ and $w(a_k) \le 0$. Let v be given by Definition 2.1. Assume there is a constant $p_T > 0$, for which the constants λ_0, Λ_1 and Λ_2 of (4.8), and C_f and C_g of (4.9) and (4.10), satisfy the relation*

$$p_T := \lambda_0 \cos\theta_j - \Lambda_1 C_f - \Lambda_2 C_g > 0. \tag{4.14}$$

1. *If $w(a_j) \le 0$, namely w is positive only at the vertex a_i, it holds that*

$$\begin{aligned}&\int_T \int_0^1 \nabla w^T \left(\frac{\partial a}{\partial \xi}(x, u, \nabla z(t))\right)^T \nabla v \,\mathrm{d}t\,\mathrm{d}x \\ &\ge \frac{1}{2\sin\theta_j}\left\{(w(a_i) - w(a_j))\gamma_a r_T + (w(a_j) - w(a_k))p_T\right\}.\end{aligned} \tag{4.15}$$

2. *If $w(a_j) \ge 0$, namely w is positive at both a_i and a_j, it holds that*

$$\begin{aligned}&\int_T \int_0^1 \nabla w^T \left(\frac{\partial a}{\partial \xi}(x, u, \nabla z(t))\right)^T \nabla v \,\mathrm{d}t\,\mathrm{d}x \\ &\ge \frac{1}{2\sin\theta_j}\left\{(w(a_i) - w(a_j))p_T + (w(a_j) - w(a_k))\gamma_a r_T\right\},\end{aligned} \tag{4.16}$$

with r_T given by (4.3).

Proof. First, expand ∇w as a linear combination of basis functions as in (4.13). For any $\nabla z \in \mathbb{R}^2$, abbreviating $\partial a(x, u, \nabla z)/\partial\xi$ as $(\partial a/\partial \xi)$, and noting the structure of a implies the symmetry of $\partial a/\partial\xi$, we have

$$\begin{aligned}\nabla w^T \left(\frac{\partial a}{\partial \xi}\right)^T \nabla v = &(w(a_i) - w(a_j))\nabla\varphi_i^T \left(\frac{\partial a}{\partial \xi}\right)\nabla v \\ &+ (w(a_k) - w(a_j))\nabla\varphi_k^T \left(\frac{\partial a}{\partial \xi}\right)\nabla v.\end{aligned} \tag{4.17}$$

In the case that w has one positive vertex, $\nabla v = \nabla\varphi_i$, and in the case that w has two positive vertices, $\nabla v = -\nabla\varphi_k$. In the first case, the ellipticity condition (1.3) implies

$$\begin{aligned}(w(a_i) - w(a_j))\nabla\varphi_i^T \left(\frac{\partial a}{\partial \xi}\right)\nabla v &\ge (w(a_i) - w(a_j))\gamma_a \nabla\varphi_i^T \nabla\varphi_i, \\ &\ge (w(a_i) - w(a_j))\gamma_a r_T \frac{|e_i||e_k|}{4|T|^2} \\ &= (w(a_i) - w(a_j))\gamma_a r_T \frac{1}{2|T|\sin\theta_j},\end{aligned} \tag{4.18}$$

where r_T defined in (4.3) is used to relate the lengths of edges e_i and e_k. In the second case, the same condition implies

$$\begin{aligned}(w(a_k) - w(a_j))\nabla\varphi_k^T \left(\frac{\partial a}{\partial \xi}\right)\nabla v &\ge (w(a_j) - w(a_k))\gamma_a \nabla\varphi_k^T \nabla\varphi_k \\ &\ge (w(a_j) - w(a_k))\gamma_a r_T \frac{1}{2|T|\sin\theta_j}.\end{aligned} \tag{4.19}$$

The above estimate for each case yields a strictly positive contribution. For the remaining term of (4.17), apply the decomposition of Assumption 4.2.

$$\begin{aligned}\nabla\varphi_i^T \left(\frac{\partial a}{\partial \xi}\right)\nabla\varphi_k &= \nabla\varphi_i^T \nabla z \left(\frac{\partial A}{\partial \xi}\right)\nabla\varphi_k + A(x, u, \nabla z)\nabla\varphi_i^T \nabla\varphi_k \\ &= (\nabla\varphi_i^T \nabla z)\left\{A_1(x, u)\frac{\partial f}{\partial \xi}(|\nabla z|) + A_2(x)\frac{\partial g}{\partial \xi}(|\nabla z|)\right\}\nabla\varphi_k \\ &+ A(x, u, \nabla z)\nabla\varphi_i^T \nabla\varphi_k.\end{aligned} \tag{4.20}$$

The Jacobian of $f(|\xi|)$ (respectively, $g(|\xi|)$) has the form

$$\frac{\partial f}{\partial \xi}(|\xi|) = f'(|\xi|)|\xi|^{-1}\xi^T.$$

The first term on the right hand side of (4.20) then satisfies

$$\begin{aligned}(\nabla\varphi_i^T\nabla z)A_1(x,u)\frac{\partial f}{\partial \xi}(|\nabla z|)\nabla\varphi_k &= A_1(x,u)(\nabla\varphi_i^T\nabla z)f'(|\nabla z|)|\nabla z|^{-1}\nabla z^T\nabla\varphi_k\\ &\le \Lambda_1(\nabla\varphi_i^T\nabla z)|f'(|\nabla z|)||\nabla\varphi_k|\\ &\le \Lambda_1|\nabla\varphi_i||\nabla\varphi_k||f'(|\nabla z|)||\nabla z|\\ &\le \Lambda_1 C_f|\nabla\varphi_i||\nabla\varphi_k|,\end{aligned} \tag{4.21}$$

where the last inequality follows from (4.9). Similarly for the second term on the right hand side of (4.20), it holds

$$(\nabla\varphi_i^T\nabla z)A_2(x)\frac{\partial g}{\partial \xi}(|\nabla z|)\nabla\varphi_k \le \Lambda_2 C_g|\nabla\varphi_i||\nabla\varphi_k|. \tag{4.22}$$

With the Assumption 4.2, it is clear that $A(x,u,\nabla z) \ge \lambda_0$. Therefore the third term on the right hand side of (4.20) satisfies

$$A(x,u,\nabla z)\nabla\varphi_i^T\nabla\varphi_k = -A(x,u,\nabla z)\frac{|e_i||e_k|\cos\theta_j}{4|T|^2} \le -\lambda_0\frac{|e_i||e_k|\cos\theta_j}{4|T|^2}. \tag{4.23}$$

Applying (4.21), (4.22) and (4.23) to (4.20), we obtain

$$-\nabla\varphi_i^T\left(\frac{\partial a}{\partial \xi}\right)\nabla\varphi_k \ge \frac{|e_i||e_k|}{4|T|^2}(\lambda_0\cos\theta_j - \Lambda_1 C_f - \Lambda_2 C_g) = \frac{p_T}{2|T|\sin\theta_j}, \tag{4.24}$$

where the sign on the left-hand side agrees with $(w(a_k)-w(a_j))$ in the case of one positive vertex where $\nabla v = \nabla\varphi_i$; and $-(w(a_i)-w(a_j))$, in the case of two positive vertices, where $\nabla v = -\nabla\varphi_k$. For the case of one positive vertex, putting (4.17) together with (4.18) and (4.24) and integrating, yields

$$\begin{aligned}&\int_T\int_0^1 \nabla w^T\left(\frac{\partial a}{\partial \xi}\right)\nabla v\,\mathrm{d}t\,\mathrm{d}x\\ &\ge \frac{1}{2\sin\theta_j}\left\{(w(a_i)-w(a_j))\gamma_a r_T + (w(a_j)-w(a_k))p_T\right\},\end{aligned} \tag{4.25}$$

establishing (4.15). Inequality (4.16) follows similarly, replacing (4.18) with (4.19). □

The next corollary shows the corresponding result if condition (4.10) on g is replaced by (4.11) in Assumption 4.2.

Corollary 4.6. *Let Assumptions 1.1, 4.1 and 4.2 hold, with g satisfying (4.11). Let $w, u \in \mathcal{V}$, and $z(t) \in \mathcal{V}$, $0 \le t \le 1$. Let a_i, a_j and a_k be the three vertices of $T \in \mathcal{T}_c$, ordered so that $w(a_i) \ge w(a_j) \ge w(a_k)$ with $w(a_i) > 0$ and $w(a_k) \le 0$. Let v be given by Definition 2.1. Assume there is a constant $p_T > 0$, for which the constants λ_0 and Λ_1 of (4.8), and C_f of (4.9), satisfy the relation*

$$p_T := \lambda_0\cos\theta_j - \Lambda_1 C_f > 0. \tag{4.26}$$

1. *If $w(a_j) \le 0$, namely w is positive only at the vertex a_i, it holds that*

$$\begin{aligned}&\int_T\int_0^1 \nabla w^T\left(\frac{\partial a}{\partial \xi}(x,u,\nabla z(t))\right)^T\nabla v\,\mathrm{d}t\,\mathrm{d}x\\ &\ge \frac{1}{2\sin\theta_j}\left\{(w(a_i)-w(a_j))\lambda_0 r_T + (w(a_j)-w(a_k))p_T\right\}.\end{aligned} \tag{4.27}$$

2. *If $w(a_j) \ge 0$, namely w is positive at both a_i and a_j, it holds that*

$$\begin{aligned}&\int_T\int_0^1 \nabla w^T\left(\frac{\partial a}{\partial \xi}(x,u,\nabla z(t))\right)^T\nabla v\,\mathrm{d}t\,\mathrm{d}x\\ &\ge \frac{1}{2\sin\theta_j}\left\{(w(a_i)-w(a_j))p_T + (w(a_j)-w(a_k))\lambda_0 r_T\right\},\end{aligned} \tag{4.28}$$

with r_T given by (4.3).

The proof is similar to Lemma 4.5, and the differences are summarized below.

Proof. The estimates (4.17)-(4.21) remain unchanged, and (4.22) is replaced by

$$\begin{aligned}(\nabla\varphi_i^T\nabla z)A_2(x)\frac{\partial g}{\partial\xi}(|\nabla z|)\nabla\varphi_k &\le A_2(x)g'(|\nabla z|)\nabla\varphi_i^T\nabla z\nabla z^T\nabla\varphi_k|\nabla z|^{-1}\\ &\le A_2(x)g'(|\nabla z|)|\nabla z|\nabla\varphi_i^T\left(\frac{\nabla z\nabla z^T}{\nabla z^T\nabla z}\right)\nabla\varphi_k\\ &\le A_2(x)\widehat{C}_g g(|\nabla z|)\frac{|e_i||e_k|}{4|T|^2}.\end{aligned}\tag{4.29}$$

The bound (4.23) is now replaced by

$$\begin{aligned}A(x,u,\nabla z)\nabla\varphi_i^T\nabla\varphi_k &= -A(x,u,\nabla z)\frac{|e_i||e_k|\cos\theta_j}{4|T|^2}\\ &\le -(\lambda_0+A_2(x)g(|\nabla z|))\frac{|e_i||e_k|\cos\theta_j}{4|T|^2}.\end{aligned}\tag{4.30}$$

Using (4.29) and (4.30) in place of (4.22) and (4.23), in (4.24) yields

$$\begin{aligned}-\nabla\varphi_i^T\left(\frac{\partial a}{\partial\xi}\right)\nabla\varphi_k &\ge \frac{|e_i||e_k|}{4|T|^2}\left(\lambda_0\cos\theta_j-\Lambda_1C_f+A_2(x)g(|\xi|)(\cos\theta_j-\widehat{C}_g)\right)\\ &\ge \frac{|e_i||e_k|}{4|T|^2}(\lambda_0\cos\theta_j-\Lambda_1C_f)=\frac{p_T}{2|T|\sin\theta_j},\end{aligned}\tag{4.31}$$

under Assumption (4.11). The remainder of the proof remains unchanged. □

The second term of (4.6) is bounded by the estimates of Lemma 4.7. These are similar to the ones found in [27], where a Lipschitz assumption replaces the bound on the derivative $\partial A/\partial\eta$. The key idea is to write $|w|$ as a multiple of $\delta_T(w)=w(a_i)-w(a_k)$, which can then be factored out of each term in the expansion (4.6). The positive part is given by the results of Lemma 4.5, and the parts that may not be assumed positive are controlled by the variance in the coefficients of u_2, which functions as a measurable control as found in Lemma 4.7; and, by the meshsize in the lower order term as given in Lemma 4.8.

Lemma 4.7. *Let Assumptions 1.1, and 4.1 hold. Let $w,u\in\mathcal{V}$, and $z(t)\in\mathcal{V}$, $0\le t\le 1$. Let a_i, a_j and a_k be the three vertices of $T\in\mathcal{T}_c$, ordered so that $w(a_i)\ge w(a_j)\ge w(a_k)$ with $w(a_i)>0$ and $w(a_k)\le 0$. Let v be given by Definition 2.1. Then, it holds that*

$$\int_T\int_0^1\frac{\partial A}{\partial\eta}(x,z(t),\nabla u)w\nabla u^T\nabla v\,dt\,dx\ge\frac{-\delta_T(w)\delta_T(u)}{2\sin\theta_j}\frac{7K_\eta}{6}(1+r_T^{-1}),\tag{4.32}$$

with r_T given by (4.3).

Proof. In the case that w has one positive vertex, $\nabla v=\nabla\varphi_i$. Applying expansion (4.13) to ∇u, followed by (4.3), one finds

$$\begin{aligned}\nabla u^T\nabla v&=(u(a_i)-u(a_j))\nabla\varphi_i^T\nabla\varphi_i+(u(a_k)-u(a_j))\nabla\varphi_k^T\nabla\varphi_i\\ &\le\delta_T(u)\frac{|e_i||e_k|}{4|T|^2}(1+r_T^{-1})\\ &=\frac{\delta_T(u)}{2|T|\sin\theta_j}(1+r_T^{-1}).\end{aligned}\tag{4.33}$$

In the case that w has two positive vertices, $\nabla v=-\nabla\varphi_k$, leading to the same result.

$$\begin{aligned}\nabla u^T\nabla v&=-(u(a_i)-u(a_j))\nabla\varphi_i^T\nabla\varphi_k-(u(a_k)-u(a_j))\nabla\varphi_k^T\nabla\varphi_k\\ &\le\frac{\delta_T(u)}{2|T|\sin\theta_j}(1+r_T^{-1}).\end{aligned}\tag{4.34}$$

Applying the bound (1.4) on $(\partial A/\partial\eta)$, then yields

$$\int_T \int_0^1 \left(\frac{\partial A}{\partial \eta}\right) w \nabla u^T \nabla v \,\mathrm{d}t\,\mathrm{d}x \le \frac{K_\eta \delta_T(u)}{2|T|\sin\theta_j}(1+r_T^{-1}) \int_T |w|\,\mathrm{d}x. \tag{4.35}$$

As shown in [27], and repeated here for convenience, the integral over T of $|w|$, can be bounded in terms of $\delta_T(w)$ making use of $\varphi_i+\varphi_j+\varphi_k=1$, and the ordering $w(a_i)\ge w(a_j)\ge w(a_k)$.

$$\begin{aligned}|w| &= |w(a_i)\varphi_i + w(a_j)\varphi_j + w(a_k)\varphi_k|\\ &= |(w(a_i)-w(a_j))\varphi_i + (w(a_k)-w(a_j))\varphi_k + w(a_j)(\varphi_i+\varphi_j+\varphi_k)|\\ &\le (w(a_i)-w(a_j))\varphi_i + (w(a_j)-w(a_k))\varphi_k + (w(a_j)-w(a_k))\\ &= (w(a_i)-w(a_j))\varphi_i + (w(a_j)-w(a_k))(1+\varphi_k).\end{aligned}$$

Applying $\int_T \varphi_i\,\mathrm{d}x = \int_T \varphi_k\,\mathrm{d}x = |T|/6$, demonstrates

$$\int_T |w|\,\mathrm{d}x \le (w(a_i)-w(a_j))\frac{|T|}{6} + (w(a_j)-w(a_k))\frac{7|T|}{6} \le \delta_T(w)\frac{7|T|}{6}. \tag{4.36}$$

Putting together (4.35) and (4.36), yields the desired result. □

Finally, we consider a bound on the third term of (4.6).

Lemma 4.8. *Let Assumptions 1.1, and 4.1 hold. Let $w, u \in \mathcal{V}$, and $z(t)\in\mathcal{V}$, $0\le t\le 1$. Let a_i, a_j and a_k be the three vertices of $T\in\mathcal{T}_c$, ordered so that $w(a_i)\ge w(a_j)\ge w(a_k)$. Suppose $w(a_k)\le 0$. Let v be given by Definition 2.1. Then, it holds that*

$$\int_T \int_0^1 \frac{\partial b}{\partial \eta}(x, z(t)) w v \,\mathrm{d}x\,\mathrm{d}t \ge -\delta_T(w)\frac{7B_\eta|T|}{6}. \tag{4.37}$$

Proof. Applying the condition (1.5) bounding $(\partial b/\partial\eta)$, and (4.36) bounded $|w|$, reveals

$$\int_T \int_0^1 \frac{\partial b}{\partial \eta}(x, z(t)) w v \,\mathrm{d}x\,\mathrm{d}t \ge -B_\eta \int_T |w|\,\mathrm{d}x \ge -\delta_T(w)\frac{7B_\eta|T|}{6}. \quad \square$$

Notably, (4.37) can be controlled by the area $|T|$ in the numerator, rather than $\delta_T(u)$ as in the result of Lemma 4.7. Effectively, this introduces a global meshsize condition as in the 1D case if the lower order term $b(x,u)$ appears in (1.1).

4.4. Comparison theorem in two dimensions

We are now ready to combine the results of Lemmas 4.5, 4.7 and 4.8 to prove a discrete comparison theorem.

Theorem 4.9 *(Two dimensional comparison theorem). Let $u_1\in\mathcal{V}$ be a subsolution of (2.1) as in (2.2), and let $u_2\in\mathcal{V}$ be a supersolution of the same problem, as in (2.3). Let $w=u_1-u_2\in\mathcal{V}$. Let Assumptions 1.1, 4.1 and 4.2 hold, with g satisfying (4.10). Assume λ_0, Λ_1, Λ_2 and C_f, C_g of Assumption 4.2, and c_{min} of (4.2) satisfy the relation*

$$\lambda_0 c_{min} - \Lambda_1 C_f - \Lambda_2 C_g > 0. \tag{4.38}$$

Define the positive constant for each $T\in\mathcal{T}$

$$p_T^* := \min\{\lambda_0 c_T - \Lambda_1 C_f - \Lambda_2 C_g, \gamma_a r_T\}, \tag{4.39}$$

with γ_a from (1.1), and c_T, r_T from (4.3). Then, the satisfaction of the condition

$$\min_{T\in\mathcal{T}}\left\{p_T^* - \delta_T(u_2)\frac{7K_\eta(1+r_T^{-1})}{6} - \frac{7B_\eta|T|s_T}{3}\right\} > 0, \tag{4.40}$$

with s_T from (4.3), implies that $u_1\le u_2$ in Ω.

Proof. Assume $w = u_1 - u_2$ is positive somewhere in Ω. This implies $w(a) > 0$ for some vertex $a \in \mathcal{D}$. Let the test function $v \in \mathcal{V}^+$ be given by Definition 2.1. Then, from (2.11), it holds that

$$\sum_{T \in \mathcal{T}_c} \int_T \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) \nabla w \cdot \nabla v + \left(\frac{\partial A}{\partial \eta}\right) w \nabla u_2 \cdot \nabla v + \left(\frac{\partial b}{\partial \eta}\right) w v \,\mathrm{d}t\,\mathrm{d}x \le 0, \tag{4.41}$$

where $\mathcal{T}_c$ defined in (2.9) is the set of all elements T where w is positive on either one or two vertices. The hypothesis (4.38) together with Lemma 4.5 implies for any $T \in \mathcal{T}_c$, it holds that

$$\int_T \int_0^1 \left(\frac{\partial a}{\partial \xi}(x, u_1, \nabla z(t)) \nabla w\right) \cdot \nabla v \,\mathrm{d}t\,\mathrm{d}x \ge \frac{\delta_T(w)}{2 \sin \theta_{T,j}} p_T^*, \tag{4.42}$$

where $\theta_{T,j}$ refers to θ_j of triangle T with respect to the local indexing, where a_i, a_j and a_k are the three vertices of T, ordered so that $w(a_i) \ge w(a_j) \ge w(a_k)$.

Lemma 4.8, together with the inequality $\sin \theta_j \le s_T$, where s_T is the sine of the maximum angle of T as in (4.3), shows for any $T \in \mathcal{T}_c$ that

$$\int_T \int_0^1 \frac{\partial b}{\partial \eta}(x, z(t)) w v \,\mathrm{d}t\,\mathrm{d}x \ge -\left(\frac{\delta_T(w)}{2 \sin \theta_{T,j}}\right) \frac{7 B_\eta |T| s_T}{3}. \tag{4.43}$$

Putting (4.42) and (4.43) and the result of Lemma 4.7 together into (2.11) yields

$$\begin{aligned} &\int_\Omega (a(x, u_1, \nabla u_1) - a(x, u_2, \nabla u_2)) \cdot \nabla v + (b(x, u_1) v - b(x, u_2)) v \,\mathrm{d}x \\ &\ge \sum_{T \in \mathcal{T}_c} \frac{\delta_T(w)}{2 \sin \theta_{T,j}} \left\{ p_T^* - \delta_T(u_2) \frac{7 K_\eta (1 + r_T^{-1})}{6} - \frac{7 B_\eta |T| s_T}{3} \right\} > 0. \end{aligned} \tag{4.44}$$

The positivity of (4.44) is in direct contradiction to the nonpositivity from (4.41), repeated from (2.11). This demonstrates that under the hypotheses of the theorem, the function v must be nowhere positive, which requires $u_1 \le u_2$ in Ω. □

Replacing Lemma 4.5 with Corollary 4.6 and 4.8 allows us to prove a second comparison result.

Corollary 4.10. *Let $u_1 \in \mathcal{V}$ be a subsolution of (2.1) as in (2.2), and let $u_2 \in \mathcal{V}$ be a supersolution of the same problem, as in (2.3). Let $w = u_1 - u_2 \in \mathcal{V}$. Let Assumptions 1.1, 4.1 and 4.2 hold, with g satisfying (4.11). Assume λ_0, Λ_1 and C_f of Assumption 4.2, and c_{min} of (4.2) satisfy the relation*

$$\lambda_0 c_{min} - \Lambda_1 > 0.$$

Define the positive constant for each $T \in \mathcal{T}$

$$p_T^* := \min\{\lambda_0 c_T - \Lambda_1 C_f, \gamma_a r_T\}, \tag{4.45}$$

with γ_a from (1.1), and c_T, r_T from (4.3). Then, the satisfaction of the condition

$$\min_{T \in \mathcal{T}} \left\{ p_T^* - \delta_T(u_2) \frac{7 K_\eta (1 + r_T^{-1})}{6} - \frac{7 B_\eta |T| s_T}{3} \right\} > 0, \tag{4.46}$$

with s_T from (4.3), implies that $u_1 \le u_2$ in Ω.

Proof. The proof follows directly by replacing Lemma 4.5 by Corollary 4.6 in (4.42) of the proof of Theorem 4.9. □

In Theorem 4.9 and Corollary 4.10, the conditions on the global meshsize depend on the u-dependence of the lower-order term $b(x, u)$ of (1.1), as developed in Lemma 4.8. The local meshsize conditions which require the difference between neighboring nodal values of the solution, $\delta_T(u)$, is small enough, as developed in Lemma 4.7, are only realized for problems in which the principal term $a(x, u, \nabla u)$ has a u-dependence. Otherwise, the angle conditions of Assumption 4.1, together with the structural assumptions of Assumptions 1.1 and 4.2 are sufficient, regardless of the meshsize, for the comparison principle of Theorem 4.9 and Corollary 4.10 to hold.

The rationale behind this is the ellipticity condition (1.3) of Assumption 1.1, together with the structural conditions of either (4.10) or (4.11) on the ξ-dependence of $A(x,\eta,\xi)$, where $a(x,\eta,\xi)_i = A(x,\eta,\xi)\xi_i$, are sufficient to determine the first term of (4.6) is bounded above a strictly positive factor of $\delta_T(w)$, via Lemma 4.5, and Corollary 4.6. The meshsize restrictions which appear with (but not without) an η-dependence in either $a(x,\eta,\xi)$ or $b(x,\eta)$ are used to control the magnitude of the remaining terms of (4.6), which also contain a factor of $\delta_T(w)$. This ensures the strictly positive term is dominant, irrespective of the magnitude of $w = u_1 - u_2$, which then implies the entire expression is positive. The case with only gradient- but not solution-dependence in the problem coefficients is illustrated in the following example.

Example 4.11. Consider a Carreau law

$$A(x,\eta,\xi) = A_0 + A_2|\xi|^{p-2}, \quad |p-2| < c_{min}, \quad \text{and } b(x,\eta) \equiv 0,$$

with $A_0, A_2 > 0$. Then $\lambda_0 = \gamma_a = A_0$, $\Lambda_1 = 0$, and $\Lambda_2 = A_2$. Corollary 4.10 applies to this case, since $g(|\xi|)$ satisfies (4.11). Then $p_T^* = \lambda_0 c_T$ in (4.45), and since $K_\eta = B_\eta = 0$, the condition (4.46) is automatically satisfied irrespective of the coarseness of the mesh.

In this example, since $K_\eta = B_\eta = 0$, it is sufficient to show that the first term in (4.6) is positive, rather than a positive multiple of $\delta_T(w)$. However, the acuteness condition of Assumption 4.1 is also used in Corollary 4.6 to determine that positivity, so that assumption is not relaxed in this example.

This result is reasonable, as in the limiting case of $p = 2$, the equation reduces to the linear Laplacian, for which the acuteness of the mesh is a sufficient (but not necessary) condition for a maximum principle to hold [15]. Further, as discussed in Section 1, uniqueness of discrete solutions for Example 4.11 follows without the angle conditions on the mesh, which are used here to show the stronger comparison principle.

The next example provides a slight generalization of the results of [27], which considers problems with $a(x,\eta,\xi) = A(x,\eta)\xi$, with $A(x,\eta)$ a scalar coefficient, essentially $A_0(x,\eta)$, of Assumption 4.2.

Example 4.12 *(Mildly anisotropic diffusion).* Consider equation (1.1) with $a(x,\eta,\xi) = A(x,\eta)\xi$ where $A(x,\eta)$ is a uniformly symmetric positive definite matrix with the decomposition

$$A(x,\eta) = k(x,\eta)I + B(x,\eta),$$

with scalar-valued $k(x,\eta) > \lambda_0$, and matrix-valued $B(x,\eta)$, where $\sigma_B(x,\eta)$, the largest singular value of $B(x,\eta)$ satisfies $\sigma_B(x,\eta) \le \sigma_B$ for all $\eta \in \mathbb{R}$ and almost every $x \in \Omega$, for some $\sigma_B < \lambda_0 c_{min}$.

Then the constant $p_T > 0$ in (4.14) of Lemma 4.5 is replaced by $p_T = \lambda_0 \cos\theta_j - \sigma_B$, and similarly for p_T^* of Theorem 4.9. The upper bound on the largest singular value of $B(x,\eta)$ enforces that the perturbation is small enough that terms of the form $\nabla\varphi_i^T A(x,u_1)\nabla\varphi_k$ remain negative, which allows $\delta_T(w)$ to be factored out of the first as well as the second two terms of (4.6). This enables the proof by contradiction.

Specifically, equations (4.20)-(4.24) of Lemma 4.5 are replaced by the following.

$$\begin{aligned} -\nabla\varphi_i^T\left(\frac{\partial a}{\partial\xi}\right)\nabla\varphi_k &= -\nabla\varphi_i^T A(x,u_1)\nabla\varphi_k \\ &= -\nabla\varphi_i^T\big(k(x,u_1)I + B(x,u_1)\big)\nabla\varphi_k \\ &\ge (\lambda_0\cos\theta_j - \sigma_B)|\nabla\varphi_i||\nabla\varphi_k| \\ &= \frac{p_T}{2|T|\sin\theta_j}. \end{aligned} \tag{4.47}$$

Example 4.12 illustrates the interpretation of Assumption 4.2 as allowing $A(x,\eta,\xi)$ to have anisotropic perturbations of $A_0(x,\eta)$, under conditions that don't disrupt the negativity of the product $\nabla\varphi_i^T(\partial a/\partial\xi)\nabla\varphi_k$. The hypothesis (4.38) of Theorem 4.9 (*cf.*, (4.14) of Lemma 4.5, (4.39) of Corollary 4.6) quantifies the magnitude of allowable perturbations.

Remark 4.13 *(Uniqueness of finite element solutions).* An important consequence of the comparison theorem is the uniqueness of solutions to (2.1), which as demonstrated holds in two dimensions under the hypotheses of Theorem 4.9, under the condition

$$\min_{T\in\mathcal{T}}\left\{p_T^* - \delta_T(u)\frac{7K_\eta(1+r_T^{-1})}{6} - \frac{7B_\eta|T|s_T}{3}\right\} > 0,$$

with p_T^* given either by (4.39) or (4.45). The quantities involved to verify this condition consist of global constants bounding the problem data and local quantities characterizing the triangulation and the computed solution u. The global constants are $\lambda_0, \Lambda_1, \Lambda_2$ and C_f, C_g or $\widehat{C}_g$ of Assumption 4.2 and γ_a, K_η, B_η of Assumption 1.1. The necessary triangulation data describes the area $|T|$ and the smallest and largest angles of each element $T \in \mathcal{T}$: c_T, r_T and s_T of (4.3). Finally, it is required to check the greatest difference between nodal values of the computed solution on each element, $\delta_T(u)$. All these quantities can be easily and efficiently computed in practice.

For the pure diffusion problem ($B_\eta = 0$), the meshsize condition is local rather than global, as the restriction is on the difference of neighboring nodal values, which, supposing the oscillations of the solution have been resolved, decreases as the mesh is refined. If u_1 and u_2 are both solutions of (1.1) with $b(x,u) \equiv 0$, then from (2.7), $w = u_1 - u_2$ satisfies

$$\sum_{T \in \mathcal{T}} \int_T \int_0^1 \left(\frac{\partial a}{\partial \xi}\right) \nabla w \cdot \nabla v + \left(\frac{\partial A}{\partial \eta}\right) w \nabla u_2 \cdot \nabla v \, \mathrm{d}t \, \mathrm{d}x = 0,$$

where $\left(\partial a/\partial \xi\right)$ is nondegenerate, and $\left(\partial A/\partial \eta\right)$ is bounded. This is a steady linear convection-diffusion problem for w, for which the convection term is controlled by ∇u_2 over each element. The gradient of the solution, which is piecewise constant over each element T, with vertices $\{a_i, a_j, a_k\}$, is expressed by the expansion $\nabla u_2|_T = (u_2(a_i) - u_2(a_j))\nabla \varphi_i|_T + (u_2(a_k) - u_2(a_j))\nabla \varphi_k|_T$, by which the gradient is controlled elementwise by the difference between neighboring nodal values, assuming a smallest-angle condition as in Assumption 4.1. Hence the condition that $\delta_T(u)$ must be small with respect to p_T^* can be interpreted as the statement that mesh is required to be fine where the gradient of the solution is steep. This in turn says that over each element, the diffusion coefficient for the linear equation for w, dominates the local convection coefficient.

5. A semilinear problem

In this section, we consider the discrete comparison principle for a special case of the problem class (1.1), the semilinear problem:

$$-\Delta u + b(x,u) = 0 \;\text{ in }\; \Omega \subset \mathbb{R}^d, \; u = 0 \text{ on } \partial\Omega. \tag{5.1}$$

For simplicity, we consider the homogeneous Dirichlet problem in one and two dimensions. The nonlinearity $b(x,u)$ is assumed to satisfy the requirements of Assumption 1.1. The discrete version of problem (5.1) is: Find $u \in \mathcal{V} \subset H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u^T \nabla v + b(x,u) v \mathrm{d}x = 0, \; \text{for all } v \in \mathcal{V}. \tag{5.2}$$

Based on the previous sections, we can obtain the following discrete comparison result for (5.2), which is a simplified version of Theorem 3.1 and Theorem 4.9 in the semilinear case. However, we find this technique leads to a suboptimal mesh condition. We then improve the condition with a linear algebra argument in Theorem 5.3. While the techniques of Theorem 5.3 do not apply to the quasilinear problem (1.1), they suggest sharper criteria for comparison theorems and uniqueness may be attainable. We include both approaches for the semilinear problem (5.1) for completeness.

Theorem 5.1. *Let $u_1 \in \mathcal{V}$ be a subsolution of (5.2), and let $u_2 \in \mathcal{V}$ be a supersolution of (5.2). Let $w = u_1 - u_2 \in \mathcal{V}$. Let $b(x,u)$ satisfy the Assumption 1.1, and for the 2D problem, let the partition satisfy Assumption 4.1. Under the respective conditions for the 1D and 2D problems:*

$$h_k^2 < \frac{2}{B_\eta}, \; k = 1, 2, \ldots, n, \;\; \textit{for } \; d = 1, \tag{5.3}$$

$$|T| < \frac{3}{7B_\eta} \min_{k=1,2,3} \cot \theta_{T,k}, \;\; \textit{for each } T \in \mathcal{T}, \;\; \textit{for } \; d = 2, \tag{5.4}$$

it holds that $u_1 \le u_2$ in Ω.

Proof. We proceed by contradiction. Assume $w = u_1 - u_2$ is positive on at least one vertex of $\mathcal{T}$. Then w changes signs on each $T \in \mathcal{T}_c$, which must be nonempty. Let v be defined as in Definition 2.1.

In the 1D case, similar to Theorem 3.1 on each $\mathcal{I}_k \in \mathcal{T}_c$, the product $w'v' = |w(a_k) - w(a_{k-1})|/h_k^2$. Thus by condition (5.3) we have

$$\int_{\mathcal{I}_k} w'v' + (b(x,u_1) - b(x,u_2)) v \, \mathrm{d}x \ge |w(a_k) - w(a_{k-1})| \left(\frac{1}{h_k} - \frac{B_\eta h_k}{2}\right) > 0.$$

This contradicts the condition that u_1 and u_2 are sub- and supersolutions of (5.2).

In the 2D case, on each $T \in \mathcal{T}_c$, label the vertices a_i, a_j and a_k such that $w(a_i) \geq w(a_j) \geq w(a_k)$. Then with Assumption 4.1, it holds for the case $w(a_j) \leq 0$, that

$$\begin{aligned}\int_T \nabla w^T \nabla v \, \mathrm{d}x &= \int_T (w(a_i) - w(a_j)\nabla\varphi_i^T \nabla\varphi_i + (w(a_k) - w(a_j))\nabla\varphi_k^T \nabla\varphi_i \, \mathrm{d}x \\ &= \frac{1}{2}\Big((w(a_i) - w(a_j))(\cot\theta_k + \cot\theta_j) + (w(a_j) - w(a_k))\cot\theta_j\Big) \\ &\geq \frac{1}{2}(w(a_i) - w(a_k))\cot\theta_j = \frac{1}{2}\delta_T(w)\cot\theta_j.\end{aligned}$$

As in Lemma 4.5, case $w(a_j) > 0$ follows similarly. By Lemma 4.8 and (5.4), we have

$$\int_T \nabla w^T \nabla v \, \mathrm{d}x + \int_T (b(x, u_1) - b(x, u_2))v \, \mathrm{d}x \geq \frac{1}{2}\delta_T(w)\left(\cot\theta_j - \frac{7B_\eta|T|}{3}\right) > 0.$$

Under (5.4) this yields a contradiction, establishing the result. □

A more direct linear algebraic approach to determine a discrete comparison principle which implies the uniqueness of (5.2) is next demonstrated. We can derive the discrete comparison principle by considering a discrete maximum principle for the difference $w = u_1 - u_2$, where $u_1 \in \mathcal{V}$ is a subsolution of (5.2), and $u_2 \in \mathcal{V}$ is a supersolution of (5.2). Similarly to (2.7), the piecewise linear $w \in \mathcal{V}$ satisfies

$$\int_\Omega \nabla w^T \nabla v \, \mathrm{d}x + \int_\Omega \int_0^1 \frac{\partial b}{\partial \eta}(x, z(t)) w v \, \mathrm{d}t \, \mathrm{d}x = \int_T f_\delta v \, \mathrm{d}x, \text{ for all } v \in \mathcal{V}, \tag{5.5}$$

where $z(t) = tu_1 + (1-t)u_2$, and f_δ is some nonpositive L^2 integrable function defined by the left hand side of (5.5). Clearly f_δ satisfies $\int_\Omega f_\delta v \, \mathrm{d}x \leq 0$, for all $v \in \mathcal{V}^+$. Equation (5.5) is a linear reaction-diffusion equation with a bounded, nonnegative reaction term $c(x) = \int_0^1 \partial b / \partial\eta(x, z(t)) \, \mathrm{d}t$. It is immaterial that the reaction term $c(x)$ is not explicitly available. As such, the maximum principle in §3. of [8] applies, establishing under the appropriate mesh conditions that $w \leq 0$ on Ω, hence $u_1 \leq u_2$. To make this article self-contained, the argument of [8] is summarized below.

Let n_{dof} be the cardinality of $\mathcal{D}$, the number of interior vertices of $\mathcal{T}$. The approximation $w \in \mathcal{V}$ is a linear combination of basis functions given by $w = \sum_{i=1}^{n_{dof}} w(a_i)\varphi_i$, with $W = (W_1, \dots W_{n_{dof}})^T$ the corresponding vector of coefficients. The discrete form of the problem (5.2) is recovered by the solution to the linear algebra system

$$AW = F, \text{ with } A = S + M, \quad A = (a_{ij}), \quad M = (m_{ij}), \quad F = (f_j), \tag{5.6}$$

for stiffness matrix S and mass matrix M defined entrywise by

$$\begin{aligned} s_{ij} = s_{ji} &= \sum_{T\in\mathcal{T}} \int_T \nabla\varphi_i^T \nabla\varphi_j \, \mathrm{d}x, \\ m_{ij} = m_{ji} &= \sum_{T\in\mathcal{T}} \int_T \int_0^1 \frac{\partial b}{\partial\eta}(x, z(t))\varphi_i\varphi_j \, \mathrm{d}t \, \mathrm{d}x. \end{aligned} \tag{5.7}$$

The load vector is given by $f_j = \sum_{T\in\mathcal{T}} \int_T f_\delta \varphi_j \, \mathrm{d}x$. Each f_j is nonpositive, from (5.5).

From (5.6), it is sufficient to show that A^{-1} is entrywise nonnegative, to establish that each W_j is nonpositive, from which it follows that $w \leq 0$ and $u_1 \leq u_2$. This is established by showing A is a *Stieltjes matrix*, meaning A is symmetric positive definite with nonpositive off-diagonal entries (see for example [30, Definition 3.23]).

Remark 5.2. As mentioned in [8, §3.], it is easier to show A is a Stieltjes matrix than an M matrix, as it is not necessary to show irreducibility. It also makes the current argument unsuitable for the full quasilinear problem (1.1), as the resulting linearized equation would induce a nonsymmetric matrix.

The next theorem is a restatement of [8, Theorem 3.7], reframed in the present context.

Theorem 5.3. *Let $u_1 \in \mathcal{V}$ be a subsolution of (5.2), and let $u_2 \in \mathcal{V}$ be a supersolution of (5.2). Let $w = u_1 - u_2 \in \mathcal{V}$. Let $b(x, u)$ satisfy the Assumption 1.1, and for the 2D problem, let the partition satisfy Assumption 4.1. Then A as given in (5.6)-(5.7) is a Stieltjes matrix, under the respective conditions for the 1D and 2D problems.*

$$|T| \le \frac{6}{B_\eta} \min_{k=1,2,3} \cot\theta_{T,k}, \;\; \textit{for each } T \in \mathcal{T}, \;\; \textit{for } d = 2, \tag{5.8}$$

$$h_k^2 \le \frac{6}{B_\eta}, \; k = 1, 2, \ldots, n, \;\; \textit{for } d = 1. \tag{5.9}$$

Proof. In d dimensions $\int_T \varphi_i \varphi_j \,\mathrm{d}x = |T|/(d+1)(d+2)$ for $i \neq j$, so the summands of the off-diagonal entries of M satisfy

$$\int_T \int_0^1 \frac{\partial b}{\partial \eta}(x, z(t))\varphi_i \varphi_j \,\mathrm{d}t\,\mathrm{d}x \le \frac{B_\eta |T|}{(d+1)(d+2)}, \; i \neq j, \tag{5.10}$$

and the diagonal entries of M are nonnegative. By (4.5) in 2D (trivially in 1D). The diagonal entries of S are positive. The off-diagonal entries of S constructed by (5.7), satisfy $\int_T \nabla\varphi_i^T \nabla\varphi_j \,\mathrm{d}x = -\cot\theta_{T,i,j}/2$ in 2D, where $\theta_{T,i,j}$ is the angle of triangle T between edges e_i and e_j. In the 1D case, $\int_{\mathcal{I}_k} \varphi_k' \varphi_{k-1}' \,\mathrm{d}x = -1/h_k$.

Under the conditions (5.8) (respectively, (5.9)), and the construction (5.6)-(5.7), the matrix A is symmetric positive definite with positive diagonal and nonpositive off-diagonal entries. Hence it is a Stieltjes matrix. □

It follows directly from Theorem 5.3 and (5.5) that the solution W to $AW = F$ is nonpositive, so that $w = u_1 - u_2 \le 0$. This method of proof is preferred for the semilinear problem (5.1), as it gives an improved constant in the mesh condition. While it does not apply directly to the quasilinear problem (1.1), a variant using an M-matrix or otherwise nonsymmetric monotone matrix may be applicable.

6. Conclusion

In this paper, we proved comparison theorems in 1D and 2D for elliptic quasilinear diffusion problems discretized by standard P_1 finite elements, significantly extending the results of [27]. We found the discrete comparison principles hold based on conditions relating the given problem data, information about the area and angles of the mesh, and the variance of the computed solution over each mesh element. The proofs are more complicated than the comparison theorem for the continuous problem, the main setback being the positive part in the difference of two solutions does not lie in the finite element space. There remains a significant gap between the class of problems for which comparison principles hold for the PDE and for the corresponding discrete problem. For the class of problems investigated here, the discrete comparison principle implies the uniqueness of the solution to the discrete problem, based on efficiently computable conditions. These results are useful for h-adaptive algorithms, where the mesh presumably remains coarse away from steep gradients in the solution or (near) singularities in the data.

Acknowledgements

SP was supported in part by NSF DMS 1719849 and NSF DMS 1852876. YZ was supported in part by NSF DMS 1319110. SP would like to acknowledge Wright State University where a substantial portion of the writing was completed. Both authors would like to thank an anonymous referee for providing useful feedback which helped to clarify the results.

References

[1] A. Abdulle, G. Vilmart, A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems, Numer. Math. 121 (2012) 397–431.

[2] N. André, M. Chipot, Uniqueness and nonuniqueness for the approximation of quasilinear elliptic equations, SIAM J. Numer. Anal. 33 (1996) 1981–1994.

[3] R.E. Bank, A.H. Sherman, A. Weiser, Refinement algorithms and data structures for regular local mesh refinement, in: Scientific Computing, IMACS/North-Holland Publishing Company, Amsterdam, 1983, pp. 3–17.

[4] G. Barles, A.P. Blanc, C. Georgelin, M. Kobylanski, Remarks on the maximum principle for nonlinear elliptic pdes with quadratic growth conditions, Ann. Sc. Norm. Super. Pisa, Cl. Sci. Ser. IV 28 (1999) 381–404.

[5] G. Barles, F. Murat, Uniqueness and the maximum principle for quasilinear elliptic equations with quadratic growth conditions, Arch. Ration. Mech. Anal. 133 (1995) 77–101.

[6] J.W. Barrett, W.B. Liu, Finite element approximation of the p-Laplacian, Math. Comput. 61 (1993) 523–537.

[7] L. Belenki, L. Diening, C. Kreuzer, Optimality of an adaptive finite element method for the p-Laplacian equation, IMA J. Numer. Anal. 32 (2012) 484–510.

[8] J.H. Brandts, S. Korotov, M. Křížek, The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, Linear Algebra Appl. 429 (2008) 2344–2357. Special Issue in honor of Richard S. Varga.

[9] S. Carl, V.K. Le, D. Motreanu, Nonsmooth Variational Problems and Their Inequalities: Comparison Principles and Applications, Springer Monographs in Mathematics, Springer Science+Business Media, New York, 2007.

[10] S. Congreve, T.P. Wihler, Iterative Galerkin discretizations for strongly monotone problems, J. Comput. Appl. Math. 311 (2017) 457–472.

[11] D.A. Di Pietro, J. Droniou, A hybrid high-order method for Leray-Lions elliptic equations on general meshes, Math. Comput. 86 (2017) 2159–2191.

[12] L. Diening, C. Kreuzer, S. Schwarzacher, Convex hull property and maximum principle for finite element minimisers of general convex functionals, Numer. Math. 124 (2013) 685–700.

[13] J. Douglas, T. Dupont, A Galerkin method for a nonlinear Dirichlet problem, Math. Comput. 131 (1975) 689–696.

[14] J. Douglas, T. Dupont, J. Serrin, Uniqueness and comparison theorems for nonlinear elliptic equations in divergence form, Arch. Ration. Mech. Anal. 42 (1971) 157–168.

[15] A. Drăgănescu, T.F. Dupont, L.R. Scott, Failure of the discrete maximum principle for an elliptic finite element problem, Math. Comput. 74 (2005) 1–23.

[16] H. Erten, A. Üngör, Computing triangulations without small and large angles, in: 2009 Sixth International Symposium on Voronoi Diagrams, pp. 192–201.

[17] E.M. Garau, P. Morin, C. Zuppa, Convergence of an adaptive Kačanov FEM for quasi-linear problems, Appl. Numer. Math. 61 (2011) 512–529.

[18] J.L. Gerver, The dissection of a polygon into nearly equilateral triangles, Geom. Dedic. 16 (1984) 93–106.

[19] D. Gilbarg, N.S. Trudinger, Elliptic Partial Differential Equations of Second Order, Grundlehren der mathematischen Wissenschaften, vol. 224, Springer-Verlag, Berlin, New York, 1983.

[20] R. Glowinski, J. Rappaz, Approximation of a nonlinear elliptic problem arising in a non-newtonian fluid flow model in glaciology, Modél. Math. Anal. Numér. 37 (2003) 175–186.

[21] I. Hlaváček, M. Křížek, J. Malý, On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type, J. Math. Anal. Appl. 184 (1994) 168–189.

[22] A. Jüngel, A. Unterreiter, Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations, Numer. Math. 99 (2005) 485–508.

[23] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer. Math. 99 (2005) 669–698.

[24] J. Karátson, S. Korotov, M. Křížek, On discrete maximum principles for nonlinear elliptic problems, Math. Comput. Simul. 76 (2007) 99–108.

[25] S. Pollock, An improved method for solving quasilinear convection diffusion problems, SIAM J. Sci. Comput. 38 (2016) A1121–A1145.

[26] S. Pollock, Stabilized and inexact adaptive methods for capturing internal layers in quasilinear PDE, J. Comput. Appl. Math. 308 (2016) 243–262.

[27] S. Pollock, Y. Zhu, Uniqueness of discrete solutions of nonmonotone PDEs without a globally fine mesh condition, Numer. Math. 139 (2018) 845–865.

[28] R.E. Showalter, Monotone Operators in Banach Space and Nonlinear Partial Differential Equations, Mathematical Surveys and Monographs, vol. 49, American Mathematical Society, Providence, RI, 1997.

[29] N.S. Trudinger, On the comparison principle for quasilinear divergence structure equations, Arch. Ration. Mech. Anal. 57 (1974) 128–133.

[30] R.S. Varga, On recurring theorems on diagonal dominance, Linear Algebra Appl. 13 (1976) 1–9.

[31] R. Verfürth, A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations, Math. Comput. 62 (1994) 445–475.

[32] J. Wang, R. Zhang, Maximum principles for $P1$-conforming finite element approximations of quasi-linear second order elliptic equations, SIAM J. Numer. Anal. 50 (2012) 626–642.