Annales Univ. Sci. Budapest., Sect. Comp. 57 (2024) 55–68

DISCRETE MAXIMUM PRINCIPLES FOR THE COURANT FINITE ELEMENT SOLUTION OF SOME NONLINEAR ELLIPTIC PROBLEMS

Menghis T. Bahlibi (Budapest, Hungary)

Communicated by László Szili
(Received September 30, 2023; accepted January 7, 2024)

Abstract. The discrete maximum principle (DMP) is an important measure of the qualitative reliability of the numerical solutions for elliptic PDE models. We are motivated by known results on the DMP and nonnegativity preservation of finite element (FE) solutions under the condition that sufficiently small enough mesh size h is used. We extend the above results by explicitly looking for how much the mesh size h should be small to guarantee the qualitative properties. We determine a threshold mesh size h_0 in terms of the angle condition to ensure the validity of DMPs by Courant FEM.

1. Introduction

The maximum principle forms an important qualitative property of second-order elliptic equations [11], therefore its discrete analogs, the so-called DMPs have been studied by many researchers [1, 2, 3, 6, 12]. The DMP is an important measure of the qualitative reliability of the numerical scheme, otherwise one could get unphysical numerical solutions like negative concentrations, etc. Typical maximum principles arise either in the form

$$\max_{\overline{\Omega}} u = \max_{\Gamma_D} u$$

Key words and phrases: Nonlinear elliptic problem, discrete maximum principle, finite element method, angle condition on the mesh.
2010 Mathematics Subject Classification: 65N30, 35J60.

https://doi.org/10.71352/ac.57.055

(that is, the solution u attains its maximum on the boundary) or in the form of

$$\max_{\overline{\Omega}} u \leq \max\{0, \max_{\Gamma_D} u\}$$

(that is, the solution u can attain a nonnegative maximum only on the boundary). Analogous minimum principles are defined by reversing signs. A physically important special case is nonnegativity preservation.

We are motivated by the articles [6, 7, 8] which deal with this topic and use a sufficiently small mesh size to establish the qualitative properties, in particular, DMP and nonnegativity. Compared to this, we explicitly look for how much the step size h must be small to guarantee the qualitative properties. We determine a threshold mesh size h_0 in terms of the angle condition to ensure the validity of DMPs by Courant FEM.
We formulate and prove the corresponding DMPs. Finally, an example of a real-life problem, where the preservation of maximum principles plays an important role, is presented.

2. The problem and its discretization

Let us begin with the following nonlinear elliptic model:

$$\begin{cases} -\operatorname{div}\Big(b(x,u,\nabla u)\,\nabla u\Big) + r(x,u,\nabla u)u &= f(x) \quad \text{in } \Omega, \\ b(x,u,\nabla u)\dfrac{\partial u}{\partial \nu} &= \gamma(x) \quad \text{on } \Gamma_N, \\ u &= g(x) \quad \text{on } \Gamma_D, \end{cases} \tag{2.1}$$

where Ω is a bounded domain in $\mathbb{R}^2$ under the assumptions below:

(a) Ω has a piecewise smooth and Lipschitz continuous boundary $\partial\Omega$; Γ_N, $\Gamma_D \subset \partial\Omega$ are measurable open sets, such that $\Gamma_N \cap \Gamma_D = \emptyset$, $\overline{\Gamma}_N \cup \overline{\Gamma}_D = \partial\Omega$ and $meas(\Gamma_D) > 0$.

(b) The scalar functions $b : \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ and $r : \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ are continuous. Further, $f \in L^2(\Omega)$, $\gamma \in L^2(\Gamma_N)$ and $g = g^*{}_{|\Gamma_D}$ with $g^* \in H^1(\Omega)$.

(c) The functions b and r are bounded such that

$$0 < \mu_0 \leq b(x,\xi,\eta) \leq \mu_1, \quad 0 \leq r(x,\xi,\eta) \leq \beta \quad \forall (x,\xi,\eta) \in \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^2, \tag{2.2}$$

where μ_0 , μ_1 and β are positive constants. The weak formulation of the problem (2.1) and its unique weak solution $u \in H^1(\Omega)$ are defined as follows:

$$u = g \quad \text{on } \Gamma_D \text{ in trace sense} \quad \text{and} \tag{2.3}$$

$$\int_\Omega \left[b(x,u,\nabla u)\ \nabla u\cdot\nabla v + r(x,u,\nabla u)u\right]dx\ = \\ = \int_\Omega fv\,dx\ + \int_{\Gamma_N}\gamma v\,d\sigma\quad \forall v\in H_D^1(\Omega). \tag{2.4}$$

To find the finite element solution, we solve the following problem (which is the counterpart of (2.3)–(2.4) in V_h): Find $u_h\in V_h$ such that

$$u_h\ =\ g_h\ \text{ on}\,\Gamma_D\quad \text{and}$$

$$\int_\Omega \left[b(x,u_h,\nabla u_h)\ \nabla u_h\cdot\nabla v_h +\ r(x,u_h,\nabla u_h)u_h v_h\right]dx = \\ = \int_\Omega f_h v_h\,dx\ + \int_{\Gamma_N}\gamma_h v_h\,d\sigma \tag{2.5}$$

$\forall v_h\in V_h^0$. Let the vector $\bar{\mathbf{c}}=(c_1,...,c_{n+m})^T$ contain the values of the finite element solution u_h at all the nodal points i.e. $c_i=u_h(P_i)$ and $u_h=\sum\limits_{i=1}^{n+m}c_i\phi_i$, where $\phi_1,....\phi_n$ are the interior basis functions and $\phi_{n+1},...,\phi_{n+m}$ are the boundary basis functions. Furthermore, $\bar{\mathbf{b}}=(b_1,...,b_n,g_1,...,g_m)^T$. Here, a nonlinear algebraic system of equations is obtained:

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}}=\bar{\mathbf{b}}, \tag{2.6}$$

and the structure of the matrix in (2.6) is

$$\overline{\mathbf{A}}(\bar{\mathbf{c}})=\begin{pmatrix}\mathbf{A}(\bar{\mathbf{c}}) & \widetilde{\mathbf{A}}(\bar{\mathbf{c}})\\ \mathbf{0} & \mathbf{I}\end{pmatrix}, \tag{2.7}$$

where $\mathbf{I}$ is an $m\times m$ identity matrix and $\mathbf{0}$ is a $m\times n$ zero matrix and $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is $(n+m)$ by $(n+m)$ matrix.

3. DMPs for linear systems

If equation (2.5) is in particular linear (when b and r are independent of u), then the algebraic system of the equations and the structure of the matrix are respectively

$$\bar{\mathbf{A}}\bar{\mathbf{c}}=\bar{\mathbf{b}}\ \text{ and }\ \bar{\mathbf{A}}=\begin{pmatrix}\mathbf{A} & \widetilde{\mathbf{A}}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}, \tag{3.1}$$

where the matrix $\bar{\mathbf{A}}$ has a dimension of $(n+m)$ by $(n+m)$. The DMPs for such linear systems have been studied, e.g., in [5, 9].

Definition 3.1. The matrix $\bar{\mathbf{A}}$ in (3.1) satisfies

- the *discrete weak maximum principle* (DwMP) if for any vector $\bar{\mathbf{c}} = (c_1, ..., c_{n+m})^T \in \mathbb{R}^{n+m}$ satisfying $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \leq 0,\ i = 1, ..., n$, one has

$$\max_{i=1,...,n+m} c_i \leq \max\{0, \max_{i=n+1,...,n+m} c_i\};$$

- the *discrete strict weak maximum principle* (DWMP) if for any vector $\bar{\mathbf{c}} = (c_1, ..., c_{n+m})^T \in \mathbb{R}^{n+m}$ satisfying $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \leq 0,\ i = 1, ..., n$, one has

$$\max_{i=1,...,n+m} c_i = \max_{i=n+1,...,n+m} c_i.$$

Theorem 3.1. *Let the matrix* $\bar{\mathbf{A}}$ *in* (3.1) *satisfy the following conditions, where* a_{ij} *denote the entries of* $\bar{\mathbf{A}}$*:*

(i) $a_{ij} \leq 0 \qquad (\forall i = 1, \ldots, n,\ j = 1, \ldots, n+m;\quad i \neq j)$,

(ii) $\sum_{j=1}^{n+m} a_{ij} \geq 0 \qquad (\forall i = 1, \ldots, n)$,

(iii) $\mathbf{A}$ *is positive definite.*

Then $\bar{\mathbf{A}}$ *possesses the* DwMP*. If the inequality in condition* (ii) *is replaced by equality, then* $\bar{\mathbf{A}}$ *possesses the* DWMP*.*

This theorem is proved in [6].

4. DMPs for nonlinear elliptic problems

Now, we are ready to state theorems related to problem (2.1).

4.1. The general result

Definition 4.1. The family $\mathcal{F}$ of triangulations of triangular meshes of a bounded polygonal domain is said to be *uniformly acute* if there exists $\alpha_0 < \frac{\pi}{2}$ such that $\alpha_n \leq \alpha_0$ for any angle α_n in all T_k in all $\mathcal{T}_h$, where $\mathcal{T}_h \in \mathcal{F}$.

Definition 4.2. The *mesh width* h is the longest diameter occurring in the triangulation $\mathcal{T}_h$. i.e. $h := \max_{k=1,...,M} diam(T_k)$, where diam refers to diameter.

Theorem 4.1. *Let the conditions* (a)–(c) *hold and the Courant finite element method be used with triangulations satisfying Definition* 4.1. *Let the mesh size* h *satisfy*

$$0 < h \le h_0 = \Big(\frac{12\cos(\alpha_0)\mu_0}{\beta} \Big)^{\frac{1}{2}}, \tag{4.1}$$

where α_0 *is the angle that obeys Definition* 4.1, μ_0 *and* β *are positive constants from* (2.2). *Then the matrix in* (2.7) *satisfies the following properties, where* $a_{ij}(\bar{\mathbf{c}})$ *denotes the entries of* $\bar{\mathbf{A}}(\bar{\mathbf{c}})$.

(i) $a_{ij}(\bar{\mathbf{c}}) \le 0, \quad i = 1, ..., n, \ j = 1, ..., n+m \quad (i \ne j),$

(ii) $\sum_{j=1}^{n+m} a_{ij}(\bar{\mathbf{c}}) \ge 0, \quad i = 1, ..., n,$

(iii) $\mathbf{A}(\bar{\mathbf{c}})$ *is positive definite.*

Proof. Let ϕ_i and ϕ_i be any basis functions. Then the entries of the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ are:

$$a_{ij}(\bar{\mathbf{c}}) = \int_\Omega \Big[b(x, u_h, \nabla u_h)\ \nabla\phi_i \cdot \nabla\phi_j \ + r(x, u_h, \nabla u_h)\ \phi_i\phi_j \Big]\, dx. \tag{4.2}$$

We now prove the properties (i)–(iii).

(i) Let $i = 1, ..., n, \ j = 1, ..., n+m$ with $i \ne j$ and let Ω_{ij} denote the interior of $\operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j$. If $\Omega_{ij} = \emptyset$ then $a_{ij}(\bar{\mathbf{c}}) = 0$. If $\Omega_{ij} \ne \emptyset$, then to determine (4.2) we should find the bound of the following integrals:

$$\int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, dx \qquad \textit{and} \qquad \int_\Omega \phi_i\phi_j \, dx. \tag{4.3}$$

From Definition 4.1 we have the maximum angle α_0 such that $\cos(\alpha_0) := \sigma_0 \ge 0$ which is independent of i, j and h. The goal here is to find an upper bound of the stiffness matrix obtained from the first part of (4.3). Now let us begin with the inner product of the basis functions on a given triangle. For any angle α_{ij}, we have

$$\begin{aligned} \nabla\phi_i \cdot \nabla\phi_j &= |\nabla\phi_i| \cdot |\nabla\phi_j| \cos(180^0 - \alpha_{ij}) = \\ &= \frac{1}{h_i} \cdot \frac{1}{h_j}(-\cos(\alpha_{ij})) \le \frac{-\cos(\alpha_{ij})}{h^2} \le \\ &\le \frac{-\cos(\alpha_0)}{h^2} \quad \forall\, h_i, h_j \le h, \forall \alpha_{ij} \le \alpha_0. \end{aligned}$$

Therefore,

$$\nabla\phi_i \cdot \nabla\phi_j \leq -\frac{\sigma_0}{h^2} < 0. \tag{4.4}$$

Hence, using Definition 4.2 and (4.4),

$$\int_{\Omega} \nabla\phi_i \cdot \nabla\phi_j \, dx = \int_{\Omega_{ij}} \nabla\phi_i \cdot \nabla\phi_j \, dx \leq -\frac{\sigma_0}{h^2} \text{ meas}(\Omega_{ij}). \tag{4.5}$$

To estimate the mass matrix obtained from the second part of (4.3) for general triangles, we use a reference triangle. If especially E is the reference triangle that is placed in the coordinate system with vertices (0, 0), (h, 0), and (0, h) then one can calculate

$$\int_{E} \phi_i\phi_j \, dx = \frac{h^2}{24}. \tag{4.6}$$

Based on the reference triangle, we can calculate the mass matrix for general triangles T_k using affine mappings from the reference element onto T_k such that $L_k : E \to T_k$. We also define $J_k = L_k'$. If the reference triangle E is considered with $h = 1$ in (4.6) and T_k is a fixed general triangle then

$$\int_{T_k} \phi_i\phi_j \, dx = \det(J_k) \int_{E} \tilde{\phi}_i\tilde{\phi}_j \, dx = \frac{|T_k|}{12} \tag{4.7}$$

by change of variables and using the fact that $\det(J_k) = 2|T_k|$, see [4], where $|T_k|$ is the area of the triangle, and $\tilde{\phi}_i$ and $\tilde{\phi}_j$ are respectively given by $\tilde{\phi}_i = \phi_i \circ L_k$, $\tilde{\phi}_j = \phi_j \circ L_k$. Therefore, (4.7) implies

$$\int_{\Omega_{ij}} \phi_i\phi_j \, dx = \sum_{T_k \in \Omega_{ij}} \int_{T_k} \phi_i\phi_j \, dx = \frac{1}{12} \text{ meas}(\Omega_{ij}), \tag{4.8}$$

where $\Omega_{ij} := \text{supp } \phi_i \cap \text{supp } \phi_j$. Using (2.2), (4.4) (4.5) and (4.8) in (4.2) we have

$$\begin{aligned} a_{ij}(\bar{\mathbf{c}}) &\leq \mu_0 \int_{\Omega} \nabla\phi_i \cdot \nabla\phi_j \, dx + \beta \int_{\Omega} \phi_i\phi_j \, dx \leq \\ &\leq -\frac{\sigma_0}{h^2}\mu_0 \text{ meas}(\Omega_{ij}) + \frac{\beta}{12} \text{ meas}(\Omega_{ij}) = \\ &= \text{meas}(\Omega_{ij}) \left(\frac{-\sigma_0}{h^2}\mu_0 + \frac{\beta}{12}\right). \end{aligned}$$

Let

$$a_{ij}(h) := \text{meas}(\Omega_{ij}) \left(-\frac{\sigma_0}{h^2}\mu_0 + \frac{\beta}{12} \right), \tag{4.9}$$

then

$$a_{ij}(\bar{\mathbf{c}}) \le a_{ij}(h). \tag{4.10}$$

Therefore, the sum of the terms in the bracket in (4.9) tends to $-\infty$ as $h \to 0$. (The first term goes to $-\infty$ as $h \to 0$ and the second term remains unchanged). This implies $a_{ij}(h) \le 0$ if h is small. The main task here is to find how much h should be to get the nonpositivity of (4.2). To determine the threshold $h = h_0$, the following equation must hold,

$$-\frac{\sigma_0}{h_0^2}\mu_0 + \frac{\beta}{12} = 0.$$

This implies $h_0 = \left(\frac{12\sigma_0\mu_0}{\beta} \right)^{\frac{1}{2}}$. In summary, if $0 < h \le h_0 = \left(\frac{12\sigma_0\mu_0}{\beta} \right)^{\frac{1}{2}}$, then $a_{ij}(\bar{c}) \le 0$ from (4.10).

(ii) For any $i = 1, ..., n$,

$$\begin{aligned} \sum_{j=1}^{n+m} a_{ij}(\bar{\mathbf{c}}) &= \int_\Omega b(x, u_h, \nabla u_h)\ \nabla\phi_i \cdot \nabla\Big(\sum_{j=1}^{n+m} \phi_j\Big)\, dx + \\ &+ \int_\Omega r(x, u_h, \nabla u_h)\ \phi_i \Big(\sum_{j=1}^{n+m} \phi_j\Big)\, dx = \int_\Omega r(x, u_h, \nabla u_h)\ \phi_i\, dx \ge 0, \end{aligned} \tag{4.11}$$

using the fact that $\sum_{j=1}^{n+m} \phi_j \equiv 1, 0 \le \phi_i \le 1$, and (2.2).

(iii) To verify that $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is positive definite, let $d \ne 0$ be an arbitrary vector in $\mathbb{R}^n$, formed by the coefficients d_i, and let

$$v_h = \sum_{j=1}^{n} d_j \phi_j \in V_h.$$

Then $v_h \ne 0$. The vector $\bar{\mathbf{c}} \in \mathbb{R}^{n+m}$ contains the coefficients for u_h as given in (2.5)–(2.6). Then, using (2.5) and (2.2), we have

$$A(\bar{\mathbf{c}})\mathbf{d} \cdot \mathbf{d} = \sum_{i,j=1}^{n} a_{ij}(\bar{\mathbf{c}}) d_i d_j =$$

$$= \int_\Omega \Big(b(x,u_h,\nabla u_h)\nabla\Big(\sum_{i=1}^n d_i\phi_i\Big)\cdot\nabla\Big(\sum_{j=1}^n d_j\phi_j\Big)+$$

$$+r(x,u_h,\nabla u_h)\sum_{i=1}^n d_i\phi_i\sum_{j=1}^n d_j\phi_j\Big)\,dx =$$

$$= \int_\Omega \Big(b(x,u_h,\nabla u_h)|\nabla v_h|^2 + r(x,u_h,\nabla u_h)v_h^2\Big)\,dx \geq$$

$$\geq \mu_0 \int_\Omega |\nabla v_h|^2 = \mu_0 |v_h|_1^2 > 0.$$

Hence, $A(\bar{\mathbf{c}})$ is a positive definite matrix.

Altogether, equation (4.9) shows that for small enough h, we have $a_{ij}(h) \leq \leq 0$, but we gave a bound h_0 such that $a_{ij}(h) \leq 0$ for all $h \leq h_0$ and all i, j. ■

Theorem 4.2. *Under the conditions of Theorem 4.1 and letting*

$$f(x) \leq 0 \quad (x \in \Omega) \quad \textit{and} \quad \gamma(x) \leq 0 \quad (x \in \Gamma_N), \tag{4.12}$$

we have

$$\max_{\overline{\Omega}} u_h \leq \max\{0, \max_{\Gamma_D} g_h\}. \tag{4.13}$$

In particular, if $\max g_h \geq 0$, *then*

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h, \tag{4.14}$$

and if $g_h \leq 0$, *then we have the nonpositivity property*

$$\max_{\overline{\Omega}} u_h \leq 0. \tag{4.15}$$

Proof. Let $\bar{\mathbf{c}} = (c_1, ..., c_{n+m})^T \in \mathbb{R}^{n+m}$ and $\bar{\mathbf{b}} = (b_1, ..., b_n, g_1, ..., g_m)^T \in \in \mathbb{R}^{n+m}$ be the vectors that appear in (2.6). Then

$$(\bar{b})_i = \int_\Omega f\phi_i\,dx + \int_{\Gamma_N} \gamma\phi_i\,d\sigma \leq 0 \qquad (i = 1, \ldots, n) \tag{4.16}$$

owing to $f \leq 0, \gamma(x) \leq 0$ and $0 \leq \phi_i \leq 1$. Then equation (2.6) and (4.16) imply $\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} = \bar{\mathbf{b}} \leq 0$. Using these arguments and the conditions of Theorem 4.1 one can apply Theorem 3.1 to conclude that the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ possesses the DwMP, and as a result, we have

$$\max_{i=1,\ldots,n+m} c_i \leq \max\{0, \max_{i=n+1,\ldots,n+m} c_i\} \leq \max\{0, \max_{i=n+1,\ldots,n+m} g_i\}, \tag{4.17}$$

since $c_i = g_i$ for all $i = n, ..., n+m$. Owing to $0 \le \phi_i \le 1$ and $\sum_{i=1}^{n+m} \phi_i \equiv 1$, the solution vectors u_h and g_h can be estimated as follows:

$$u_h = \sum_{i=1}^{n+m} c_i \phi_i \le \max_{i=1,\dots,n+m} c_i \sum_{i=1}^{n+m} \phi_i \le \max_{i=1,\dots,n+m} c_i, \tag{4.18}$$

$$\begin{aligned} g_h = \sum_{i=n+1}^{n+m} g_i \phi_i \le \max_{i=n+1,\dots,n+m} g_i \sum_{i=n+1}^{n+m} \phi_i \;\Rightarrow \\ \Rightarrow \max_{\Gamma_D} g_h \le \max_{i=n+1,\dots,n+m} g_i. \end{aligned} \tag{4.19}$$

In the last statement in fact equality holds because of the following arguments. The piecewise linear basis functions satisfy

$$\phi_i(P_j) = \delta_{ij} := \begin{cases} 1 & \text{if i = j,} \\ 0 & \text{if i} \ne \text{j,} \end{cases} \tag{4.20}$$

for proper nodes $P_1, \dots, P_n \in \Omega$ and $P_{n+1}, \dots, P_{n+m} \in \Gamma_D$. Let $g_k := \max_{i=n+1,\dots,n+m} g_i$. Then, using (4.20)

$$g_h(P_k) = \sum_{i=n+1}^{n+m} g_i \phi_i(P_k) = g_k,$$

since $\phi_i(P_k) = 1$ and 0 in the other nodes. That is, $g_h(P_k) = \max_{i=n+1,\dots,n+m} g_i$. Hence, we obtained the equality

$$\max_{\Gamma_D} g_h = \max_{i=n+1,\dots,n+m} g_i,$$

which implies

$$\max\{0, \max_{\Gamma_D} g_h\} = \max\{0, \max_{i=n+1,\dots,n+m} g_i\}. \tag{4.21}$$

Altogether, (4.17), (4.18) and (4.21) imply (4.13). The last two statements are simple consequences of (4.13). ■

The corresponding discrete minimum principle for the problem (2.1) can be verified in the same way by reversing signs.

Theorem 4.3. *Under the conditions of Theorem 4.1 and letting*

$$f(x) \ge 0 \quad (x \in \Omega) \qquad \textit{and} \quad \gamma(x) \ge 0 \quad (x \in \Gamma_N), \tag{4.22}$$

we have

$$\min_{\overline{\Omega}} u_h \geq \min\{0, \min_{\Gamma_D} g_h\}. \tag{4.23}$$

In particular, if $\min g_h \leq 0$*, then*

$$\min_{\overline{\Omega}} u_h = \min_{\Gamma_D} g_h, \tag{4.24}$$

and, if $g_h \geq 0$*, then we have nonnegativity property*

$$\min_{\overline{\Omega}} u_h \geq 0. \tag{4.25}$$

Remark 4.1. Let us apply a uniformly acute triangulation of a bounded polygonal domain for an elliptic problem using the Courant finite element method. For a fixed mesh, it makes sense to check whether $h \leq \left(\frac{12\cos(\alpha_0)\mu_0}{\beta}\right)^{\frac{1}{2}}$ to ensure DMP and nonnegativity preservation. If it is not satisfied then you must try with a finer mesh.

4.2. Semilinear problems

Let us consider a special case of problem (2.1) to illustrate the theorem:

$$\begin{cases} -\operatorname{div}\Big(b(x)\nabla u\Big) + q(x,u) &= f(x) \quad \text{in } \Omega, \\ b(x)\dfrac{\partial u}{\partial \nu} &= \gamma(x) \quad \text{on } \Gamma_N, \\ u &= g(x) \quad \text{on } \Gamma_D, \end{cases} \tag{4.26}$$

where Ω is a bounded polygonal domain in $\mathbb{R}^2$ and $q \in C^1(\Omega \times \mathbb{R})$. We assume that there exists $\beta > 0$ such that

$$0 \leq \frac{\partial q}{\partial \xi}(x,\xi) \leq \beta \quad \text{and} \quad 0 < \mu_0 \leq b(x) \leq \mu_1, \ \forall (x,\xi) \in (\Omega \times \mathbb{R}). \tag{4.27}$$

Let us first define a function r in terms of q :

$$r(x,\xi) := \begin{cases} \dfrac{q(x,\xi) - q(x,0)}{\xi}, & \text{if } \xi \neq 0 \\ \dfrac{\partial q}{\partial \xi}(x,0), & \text{if } \xi = 0 \end{cases} \tag{4.28}$$

then

$$r(x,\xi)\xi = q(x,\xi) - q(x,0), \qquad \forall \xi \in \mathbb{R}. \tag{4.29}$$

We need to show that $r(x,\xi)$ satisfies condition (2.2). For every $t \in [0,1]$ the first part of (4.27) implies $0 \le \frac{\partial q}{\partial \xi}(x,t\xi) \le \beta$, then by Newton–Leibniz Theorem, we have

$$0 \le \int_0^1 \frac{\partial q}{\partial \xi}(x,t\xi)\,dt \le \int_0^1 \beta\,dt \quad \Rightarrow \quad 0 \le \frac{q(x,\xi)-q(x,0)}{\xi} \le \beta$$

$$\Rightarrow 0 \le r \le \beta \quad . \tag{4.30}$$

Hence, $0 \le r(x,\xi) \le \beta \quad \text{on } \mathbb{R}$.

Problem (4.26) can be written as:

$$\begin{cases} -\operatorname{div}\Big(b(x)\nabla u\Big) + q(x,u) - q(x,0) &= f(x) - q(x,0) \quad \text{in } \Omega, \\ b(x)\dfrac{\partial u}{\partial \nu} &= \gamma(x) \quad \text{on } \Gamma_N, \\ u &= g(x) \quad \text{on } \Gamma_D. \end{cases} \tag{4.31}$$

This implies

$$\begin{cases} -\operatorname{div}\Big(b(x)\nabla u\Big) + r(x,u)u &= \tilde{f}(x) \quad \text{in } \Omega, \\ b(x)\dfrac{\partial u}{\partial \nu} &= \gamma(x) \quad \text{on } \Gamma_N, \\ u &= g(x) \quad \text{on } \Gamma_D, \end{cases} \tag{4.32}$$

where $r(x,u)u = q(x,u) - q(x,0)$ from (4.29) and $\tilde{f}(x) = f(x) - q(x,0)$. Then we can see that (4.32) is a special case of problem (2.1) from Theorem 4.3.

Corollary 4.1. *Let us consider problem* (4.32) *with assumptions of* (4.27) *and the conditions of Theorem* 4.1 *that is Definition* 4.1 *must hold, and* h *must satisfy*

$$0 < h \le h_0 = \Big(\frac{12\sigma_0\mu_0}{\beta}\Big)^{\frac{1}{2}}. \tag{4.33}$$

Then:

- *if* $\tilde{f} \le 0$ *and* $\gamma \le 0$, *then*

$$\max_{\overline{\Omega}} u_h \le \max\{0, \max_{\Gamma_D} g_h\}. \tag{4.34}$$

In particular, if $\max g_h \ge 0$, *then*

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h, \tag{4.35}$$

and if $g_h \le 0$, *then we have the nonpositivity property*

$$\max_{\overline{\Omega}} u_h \le 0. \tag{4.36}$$

- *if* $\tilde{f} \ge 0$ *and* $\gamma \ge 0$, *then we have*

$$\min_{\overline{\Omega}} u_h \ge \min\{0, \min_{\Gamma_D} g_h\}. \tag{4.37}$$

In particular, if $\min g_h \le 0$, *then*

$$\min_{\overline{\Omega}} u_h = \min_{\Gamma_D} g_h, \tag{4.38}$$

and if $g_h \ge 0$, *then we have the nonnegativity property*

$$\min_{\overline{\Omega}} u_h \ge 0. \tag{4.39}$$

This corollary is illustrated by the example stated below.

4.3. Example: diffusion-kinetics enzyme problem

A diffusion-kinetics equation governing the steady-state concentration u of some substrate in an enzyme-catalyzed reaction has the following form, see in [10], where we included mixed boundary conditions:

$$\begin{cases} \operatorname{div}(D(x)\nabla u) = q(x,u) & \text{in } \Omega, \\ D\dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma_N, \\ u = u_0 & \text{on } \Gamma_D, \end{cases} \tag{4.40}$$

where Ω is a bounded domain in $\mathbb{R}^2$, $D(x)$ is the positive molecular diffusion coefficient of the substrate in a medium containing some continuous distribution of bacteria, q is the rate of the enzyme-substate reaction. In this example, the DMP and nonnegativity of the solution are obtained for a particular case for the reaction rate by Michaelis–Menten theory:

$$q(x,\xi) = \frac{\epsilon^{-1}\xi}{\xi + k} \qquad for \quad \xi \ge 0, \tag{4.41}$$

where $k > 0$ is the Michaelis constant and $\epsilon > 0$. The condition of $D(x)$ is given by $0 < \mu_0 \le D(x) \le \mu_1$, where μ_0 and μ_1 are positive constants. Further, $u_0 \ge 0$ and $\beta = \frac{1}{\epsilon k}$. First, rewrite the equation in the form (4.32). Then $\tilde{f}(x) = 0$ since $q(x,0) = 0$ and $r(x,\xi) := \frac{\epsilon^{-1}}{\xi+k}$ and it satisfies (4.30) because $0 \le r \le \frac{1}{\epsilon k}$. We can also extend $q(x,\xi)$ from $\xi \ge 0$ to $\xi \in \mathbb{R}$, by the formula $q(x,-\xi) = -q(x,\xi)$ for $u \le 0$ [6].

Corollary 4.2. *Let us consider the problem* (4.40). *If the geometric condition in Theorem* 4.1 *holds and the mesh h satisfies*

$$h \leq h_0 = \Big(12cos(\alpha_0)\mu_0\epsilon k \Big)^{\frac{1}{2}}, \tag{4.42}$$

then the finite element solution u_h *of* (4.40) *is bounded by:*

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} u_{0h} \quad \textit{and} \quad \min_{\overline{\Omega}} u_h \geq 0.$$

Proof. (4.40) is a special case of equation (4.26) for $g := u_0 \geq 0$, $f(x) = 0$, $\gamma(x) = 0$, and the mesh size h when $\beta = \frac{1}{\epsilon k}$ in Theorem 2 equation (4.1) is:

$$h \leq h_0 = \Big(\frac{12cos(\alpha_0)\mu_0}{(\epsilon k)^{-1}} \Big)^{\frac{1}{2}} = \Big(12cos(\alpha_0)\mu_0\epsilon k \Big)^{\frac{1}{2}}.$$

Both (4.34) and (4.37) are true, since $f(x) = 0$, $\gamma(x) = 0$. Moreover, (4.35) and (4.39) are satisfied because $g := u_0 \geq 0$. Therefore, $\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} u_{0h}$ and $\min_{\overline{\Omega}} u_h \geq 0 \quad \forall x \in \Omega$. ■

Altogether, we are able to guarantee the nonnegativity and determine the maximum of the solution of the PDE without knowing the numerical solution, because the maximum of the solution is attained at the boundary.

Acknowledgements. I would like to express my deep gratitude to my research supervisor professor János Karátson for his patience, guidance, enthusiastic encouragement, and useful critiques of my research work. I thank the anonymous referee for the useful comments on the manuscript.

References

[1] **Brandts, J., S. Korotov and M. Křížek,** The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, *Linear Algebra Appl.*, **429** (2008), 2344–2357. https://doi.org/10.1016/j.laa.2008.06.011

[2] **Ciarlet, P.G.,** Discrete maximum principle for finite-difference operators, *Aequationes Math.*, **4** (1970), 338–352. https://doi.org/10.1007/BF01844166

[3] **Drăgănescu, A., T.F. Dupont and L.R. Scott,** Failure of the discrete maximum principle for an elliptic finite element problem, *Math. Comp.* 74 (2005), no. 249, 1–23. https://doi.org/10.1090/S0025-5718-04-01651-5

[4] **Elman, H.C., D.J. Silvester and A.J. Wathen,** *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

[5] **Faragó, I.,** Matrix and Discrete Maximum Principles, in: *LSSC 2009*, LNCS 5910, 563–570 (2010).

[6] **Karátson J. and S. Korotov,** Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, *Numer.Math.*, **99** (2005), 669–698
https://doi.org/10.1007/s00211-004-0559-0

[7] **Karátson J. and S. Korotov,** Discrete maximum principles for FEM solutions of some nonlinear elliptic interface problems, *Int. J. Numer. Anal. Modeling.*, **6(1)** (2008), 1–16.

[8] **Karátson J., S. Korotov and M. Křížek,** On discrete maximum principles for nonlinear elliptic problems, *Math. Comput. Simul.*, **76** (2007), 99–108.
https://doi.org/10.1016/j.matcom.2007.01.011

[9] **Karátson J. and S. Korotov,** Some discrete maximum principles arising for nonlinear elliptic finite element problems, *Computers and Mathematics with Applications*, **70** (2015), 2732–2741.
https://doi.org/10.1016/j.camwa.2015.06.036

[10] **Keller, H.B.,** Elliptic boundary value problems suggested by nonlinear diffusion processes, *Archive for Rational Mechanics and Analysis*, **35** (1969), 363–381.
https://doi.org/10.1007/BF00247683

[11] **Protter, M.H. and H.F. Weinberger,** *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984.

[12] **Vejchodsky, T.,** The discrete maximum principle for Galerkin solutions of elliptic problems, *Cent. Eur. J. Math.*, **10(1)** (2012), 25–43.
https://doi.org/10.2478/s11533-011-0085-0

Menghis T. Bahlibi
Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
Budapest
Hungary
menghis.teweldebrhan@gmail.com