Discrete maximum principles with computable mesh conditions for nonlinear elliptic finite element problems

Menghis T. Bahlibi

Supervisors: Prof. János Karátson, Prof. Ferenc Izsák

Institute of Mathematics
Faculty of Science, Eötvös Loránd University

Outline of the talk

Introduction

The maximum principle (MP) forms an important qualitative property of second-order elliptic equations [9].

- Typical MPs arise in either the following forms:

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

  i.e. the solution u attains its maximum on the boundary, or

$$\max_{\overline{\Omega}} u \le \max\{0, \max_{\partial\Omega} u\}$$

  i.e. the solution u can attain a nonnegative maximum only on the boundary.

- Analogous minimum principles (mPs) are defined by reversing signs.

- A physically important special case is nonnegativity preservation (NNP).

DMPs for the FE solution of nonlinear PDEs

The discrete analogs, the so-called discrete maximum principles (DMPs) have been studied by many researchers [1, 2, 3, 6].

Motivation: 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.

- Motivation: Similar results in [6, 7] for "small enough mesh size h".

- Achieved results: Computable conditions on the geometric characteristics of widely studied FE shapes: triangles, tetrahedra, prisms, and rectangles, and guarantee the validity of DMPs under these conditions.

Model Problem

Nonlinear elliptic PDE BVP:

$$\begin{cases} -\operatorname{div}\Big(b(x,u,\nabla u)\,\nabla u\Big) + r(x,u,\nabla u)u = f(x) & \text{in } \Omega, \\ b(x,u,\nabla u)\frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D, \end{cases} \tag{1}$$

where Ω is a bounded domain in $\mathbf{R}^d$ ($d = 2$ or 3).

Assumption 1

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

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

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

$$0 < \mu_0 \le b(x,\xi,\eta) \le \mu_1, \quad 0 \le r(x,\xi,\eta) \le \beta \qquad \forall (x,\xi,\eta) \in \overline{\Omega}\times\mathbf{R}\times\mathbf{R}^d, \tag{2}$$

where μ_0, μ_1 and β are positive constants.

FE Approximation

To find the FE solution for the model (1), consider a FE subspace V_h of first-order elements.

(B1) $0 \leq \phi_i \leq 1 \quad (\forall i = 1, \ldots, n+m)$;

(B2) $\sum_{i=1}^{n+m} \phi_i \equiv 1$,

(B3) $\phi_i(P_j) = \delta_{ij}$ for proper nodes $P_1, \ldots, P_n \in \Omega$ and $P_{n+1}, \ldots, P_{n+m} \in \partial\Omega$.

Consider Courant, tetrahedral, bilinear, and prismatic elements, for all of which the conditions (B1)-(B3) hold.

FE : $u_h \in V_h$ such that

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

$$\int_\Omega \left[b(x, u_h, \nabla u_h)\, \nabla u_h \cdot \nabla v_h + r(x, u_h, \nabla u_h) u_h v_h \right] dx = \int_\Omega f_h v_h \, dx + \int_{\Gamma_N} \gamma_h v_h \, d\sigma \tag{3}$$

Nonlinear algebraic system of equations

To find the coefficient vector $\overline{\mathbf{c}}$ of u_h, following [6], the corresponding nonlinear algebraic system of equations is given by

$$\overline{\mathbf{A}}(\overline{\mathbf{c}})\overline{\mathbf{c}} = \overline{\mathbf{b}}, \tag{4}$$

where the structure of the matrix is :

$$\overline{\mathbf{A}}(\overline{\mathbf{c}}) = \begin{pmatrix} \mathbf{A}(\overline{\mathbf{c}}) & \widetilde{\mathbf{A}}(\overline{\mathbf{c}}) \\ \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{5}$$

where $\mathbf{I}$ is an $m \times m$ identity matrix and $\mathbf{0}$ is a $m \times n$ zero matrix, further, the entries of the matrix $\overline{\mathbf{A}}(\overline{\mathbf{c}})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n+m$ are

Entries of the matrix $\overline{\mathbf{A}}(\overline{\mathbf{c}})$

$$a_{ij}(\overline{\mathbf{c}}) = \int\limits_{\Omega_{ij}} \Big[b(x, u_h, \nabla u_h)\ \nabla\phi_i \cdot \nabla\phi_j + r(x, u_h, \nabla u_h)\ \phi_i\phi_j \Big]\, dx, \qquad (6)$$

where ϕ_i and ϕ_j are corresponding basis functions and

$$\Omega_{ij} = supp\,\phi_i \cap supp\,\phi_j\,, \qquad (7)$$

where *supp* refers to the support of a function (i.e. the closure of the set where it is nonvanishing). The vector $\overline{\mathbf{c}} = (c_1, \ldots, c_{n+m})^T$ contains the values of the FE 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 $\phi_1, \ldots.\phi_n$ are the interior basis functions and $\phi_{n+1}, \ldots, \phi_{n+m}$ are the boundary basis functions.
Furthermore, $\overline{\mathbf{b}} = (b_1, \ldots, b_n, g_1, \ldots, g_m)^T$ and $\overline{\mathbf{A}}(\overline{\mathbf{c}})$ is $(n+m)$ by $(n+m)$ matrix.

Theorem

Let V_h be any FEM subspace. The entries of the matrix $\overline{\mathbf{A}}(\overline{\mathbf{c}})$ for $i = 1, \dots, n$ and $j = 1, \dots, n+m$ are given by (6), where ϕ_i and ϕ_j are corresponding basis functions and $\Omega_{ij} = supp\,\phi_i \cap supp\,\phi_j$.
Let the general properties (B1)-(B3) hold. Then the matrix (5)–(6) satisfies

(i) $\sum_{j=1}^{n+m} a_{ij}(\overline{\mathbf{c}}) \geq 0 \qquad (\forall i = 1, \dots, n);$

(ii) $\overline{\mathbf{A}}(\overline{\mathbf{c}})$ *is positive definite.*

General Theorem

Theorem

Let the general properties (B1)-(B3) *hold. If* $a_{ij}(\bar{\mathbf{c}}) \leq 0$ $(i \neq j)$, *then* u_h *satisfies the DMP. i.e., If*

$$f(x) \leq 0 \quad (x \in \Omega) \quad \textit{and} \quad \gamma(x) \leq 0 \quad (x \in \Gamma_N), \tag{8}$$

then

$$\max_{\overline{\Omega}} u_h \leq \max\{0, \max_{\Gamma_D} g_h\}. \tag{9}$$

In particular, if $\max_{\Gamma_D} g_h \geq 0$, *then*

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h, \tag{10}$$

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

$$u_h \leq 0 \quad \text{on} \quad \overline{\Omega}. \tag{11}$$

Courant FE meshes

Definition

The family $\mathcal{F}$ of triangulations 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 $\mathcal{T}_h \in \mathcal{F}$.

Theorem

Let Assumption 1 hold and the Courant FE method be used with triangulations satisfying the Definition. Let the mesh size h satisfy

$$0 < h \leq h_0 = \left(\frac{12 \cos(\alpha_0) \mu_0}{\beta} \right)^{\frac{1}{2}}, \tag{12}$$

where α_0 is the angle that obeys the Definition, μ_0 and β are the positive constants from (2).
Then $a_{ij}(\overline{\mathbf{c}}) \leq 0, \quad i = 1, ..., n, \; j = 1, ..., n+m \quad (i \neq j)$.
Consequently, the DMP (9) holds.

Tetrahedral FE meshes

Definition

A family $\mathcal{F}$ of tetrahedral triangulations of a bounded polyhedral domain is said to be uniformly acute if there exists $\alpha_0 < \frac{\pi}{2}$ such that $\alpha_{ij}^K \leq \alpha_0$ for any angle α_{ij}^K in all $K \in \mathcal{T}_h$, and $\mathcal{T}_h \in \mathcal{F}$.

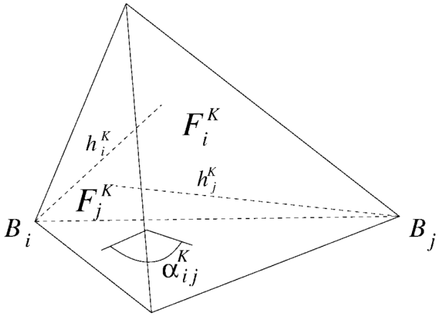

Figure: A tetrahedral cell K from [4].

Theorem

Let $d = 3$ *and Assumption 1 hold, and let the tetrahedral FE method be used with triangulations satisfying the Definition. Let the mesh size h satisfy*

$$0 < h \leq h_0 = \left(\frac{20\mu_0 \cos \alpha_0}{\beta} \right)^{\frac{1}{2}}, \tag{13}$$

where α_0 *is the angle that obeys the Definition,* μ_0 *and* β *are the positive constants from (2). Then*

$$a_{ij}(\overline{\mathbf{c}}) \leq 0, \quad i = 1, ..., n, \; j = 1, ..., n+m \quad (i \neq j).$$

Consequently, the DMP (9) *holds.*

Bilinear elements

Consider a semilinear special case ($b = 1$) for problem (1), $d = 2$:

Definition

A family $\mathcal{F}$ of rectangular meshes is said to be uniformly non-narrow if there exists $\rho_0 < \sqrt{2}$ such that for any rectangle we have $\frac{H}{h} \leq \rho_0$ where H and h denote the longest and shortest side of the rectangle, respectively.

Theorem

Let Assumption 1 hold and the bilinear FE method be used with a mesh satisfying the Definition. Let the mesh size h satisfy

$$0 < h \leq h_0 = \frac{\sqrt{3\mu_0(2 - \rho_0^2)}}{\rho_0 \sqrt{\beta}} \tag{14}$$

where ρ_0 obeys the Definition, μ_0 and β are the positive constants. Then $a_{ij}(\bar{\mathbf{c}}) \leq 0, \quad i = 1, ..., n, \ j = 1, ..., n + m \quad (i \neq j)$. Consequently, the DMP (9) holds.

Example for Bilinear elements

Determine h_0 for bilinear elements.

Example: Let us apply a uniform square mesh on Ω for the following problem:

$$-\mu_0 \Delta u + \frac{u}{\lambda + \epsilon u} = f \quad \text{in} \quad \Omega \tag{15}$$

(with proper boundary conditions), which involves the rewritten form of the Michaelis-Menten nonlinearity, i.e. $\lambda, \epsilon > 0$ are given constants.

We must calculate the constants to compute h_0 in (14).

Since $\beta = \frac{1}{\lambda}$ and $\rho_0 = 1$, we obtain

$$h_0 = \sqrt{3\mu_0 \lambda}. \tag{16}$$

Prismatic Element

Consider a semilinear special case ($b = 1$) for problem (1), $d = 3$:

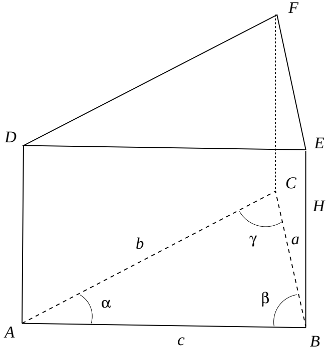

Figure: Basic notations for prismatic elements, based on [5].

Assumption 2

Let $h > 0$ be the triangular mesh parameter. There exist fixed angles

$$0 < \gamma_{min} \leq \gamma_{max} < \frac{\pi}{2}$$

such that the area $|T|$ of any triangle T satisfies

$$\frac{1}{2} h^2 \sin \gamma_{min} \leq |T| \leq \frac{1}{2} h^2 \sin \gamma_{max} .$$

Further, let γ_{med} denote a lower bound for the second largest degrees of the triangles T.

Theorem

Let Assumption 2 hold, and let us fix a constant δ_1 *such that*

$$0 < \delta_1 < \frac{4 \cot \gamma_{max}}{\sin \gamma_{max}}. \tag{17}$$

If the mesh parameters satisfy the following conditions, where μ_0 *and* β_0 *are constants from (2) :*

$$h^2 \leq \frac{3\mu_0\delta_1}{\beta_0}, \tag{18}$$

$$\frac{\cot \gamma_{med} + \cot \gamma_{min}}{\sin \gamma_{min}} + \frac{1}{2}\delta_1 \leq \left(\frac{h}{H}\right)^2 \leq \frac{4 \cot \gamma_{max}}{\sin \gamma_{max}} - \delta_1 . \tag{19}$$

Then

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

Consequently, the DMP (9) *holds.*

Necessity of fine enough mesh

We illustrate the above theoretical results with an experiment for the bilinear FE solution of a 2D reaction-diffusion problem (Michaelis-Menten nonlinearity) by Murry [8], where nonnegativity can fail for a too-coarse mesh.

$$\begin{cases} -\mu_0 \Delta u + \frac{u}{1+\epsilon u} = f \quad \text{in } \Omega := [0,1]^2, \\ u = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{20}$$

Numerical Experiments

- In the experiment $\mu_0 = 10^{-5}$ and $\epsilon = 10^{-3}$ are constants given by Keller, see in [8].
- $f(x, y) := (2x - 1)^6 \geq 0$ describes a source function mostly concentrated near two sides of the square domain.

The graphs below illustrate the numerical solutions for five different meshes.

FE solution of (16) for coarse mesh

The NN of the numerical solution fails. i.e., $\min u_h < 0$.

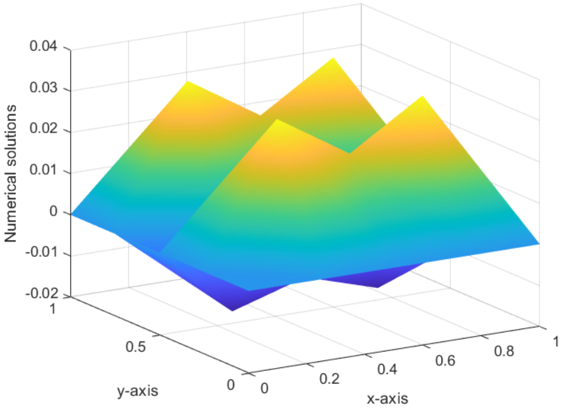

Figure: FE solution for $h = 0.25$: $\min u_h = -0.0170$.

FE solution of (16) for coarse mesh

The NN of the numerical solution fails. i.e., $\min u_h < 0$.

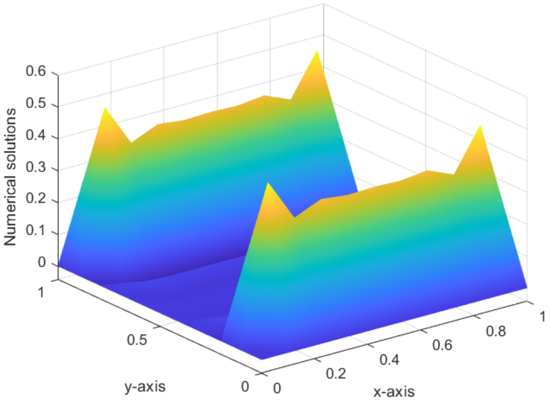

Figure: FE solution for $h = 0.1$: $\min u_h = -0.0421$.

FE solution of (16) for fine mesh

The NN of the numerical solution fails. i.e., $\min u_h < 0$.

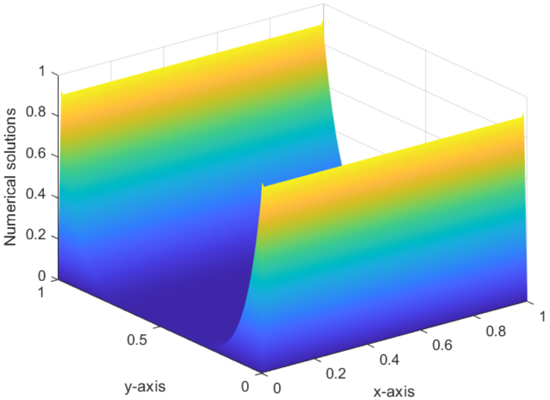

Figure: FE solution for $h = 0.0075$: $\min u_h = -8.8156e - 14$.

FE solution of (16) for fine mesh

The NN of the FE solution holds. i.e., $u_h \geq 0$ only for sufficiently small mesh sizes h.

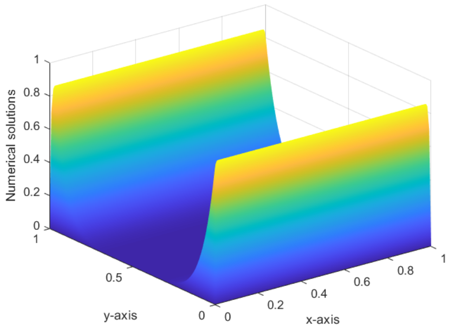

Figure: FE solution for $h = 0.005$: $\min u_h = 0$.

From (16) $h \leq h_0 = 0.0054$ (Theoretical results), and in the runs, we obtained nonnegative minima for $h \leq 0.0074$ (Experimental results).

FE solution of (16) for sufficient small mesh

The NN of the FE solution holds. i.e., $u_h \geq 0$ only for sufficiently small mesh sizes h.

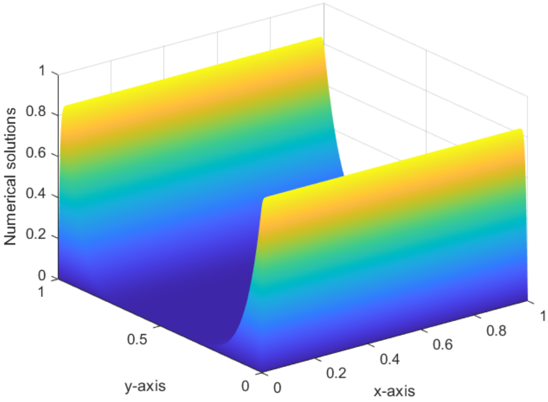

Figure: FE solution for $h = 0.001$: $\min u_h = 0$.

Summary

The summary of the above experiments for different mesh sizes h and the corresponding minima of numerical solutions u_h are given in the following table.

h	0.25	0.1	0.01	0.0075	0.005	0.001
min u_h	-0.017	-0.04	-8.3×10^{-11}	-8.8×10^{-14}	0	0

Table: Minima of the FE solutions min u_h for some values of h.

Conclusion

- We have been able to determine threshold mesh sizes for h using the computable conditions on the geometric characteristics of widely studied FE shapes: triangles, tetrahedra, prisms, and rectangles, and thus ensure the validity of DMPs for nonlinear elliptic PDEs.

References

[1] Jan H Brandts, Sergey Korotov, and Michal Křížek. "The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem". In: *Linear Algebra and its Applications* 429.10 (2008), pp. 2344–2357.

[2] Philippe G Ciarlet. "Discrete maximum principle for finite-difference operators". In: *Aequationes mathematicae* 4.3 (1970), pp. 338–352.

[3] Andrei Drăgănescu, Todd Dupont, and LR Scott. "Failure of the discrete maximum principle for an elliptic finite element problem". In: *Mathematics of computation* 74.249 (2005), pp. 1–23.

[4] István Faragó, Róbert Horváth, and Sergey Korotov. "Discrete maximum principles for FE solutions of nonstationary diffusion-reaction problems with mixed boundary conditions". In: *Numerical Methods for Partial Differential Equations* 27.3 (2011), pp. 702–720.

[5] Antti Hannukainen, Sergey Korotov, and Tomáš Vejchodskỳ. "Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes". In: *Journal of computational and applied mathematics* 226.2 (2009), pp. 275–287.

[6] János Karátson and Sergey Korotov. "Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions". In: *Numerische Mathematik* 99.4 (2005), pp. 669–698.

[7] János Karátson, Sergey Korotov, and Michal Křížek. "On discrete maximum principles for nonlinear elliptic problems". In: *Mathematics and Computers in Simulation* 76.1-3 (2007), pp. 99–108.

[8] Herbert B Keller. "Elliptic boundary value problems suggested by nonlinear diffusion processes". In: *Archive for Rational Mechanics and Analysis* 35.5 (1969), pp. 363–381.

[9] Murray H Protter and Hans F Weinberger. *Maximum principles in differential equations*. Springer Science & Business Media, 2012.

Thank you for your attention!