Discrete maximum principles for the Courant finite element solution of some nonlinear elliptic problems

Supervisors: Prof. János Karátson
Prof. Ferenc Izsák

Institute of Mathematics
Faculty of Science, Eötvös Loránd University

December 7, 2023

Outline of the talk

Introduction

- The maximum principle(MP) forms an important qualitative property of second-order elliptic equations [8].
- The discrete analogs, the so-called discrete maximum principles (DMPs) have been studied by many researchers [1, 2, 3, 4, 9].

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.

Illustration: Nonnegativity preservation(NNP) for mixed boundary conditions

Let L be a linear differential operator of elliptic type:

$$\begin{cases} Lu = f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \gamma \quad \text{on } \Gamma_N, \\ u = g \quad \Gamma_D, \end{cases} \tag{1}$$

where Ω is a bounded domain in $\mathbf{R}^d$.

NNP holds:

If $f \geq 0$, $g \geq 0$ *and* $\gamma \geq 0$ *then* $u \geq 0$.

Let $\Omega = (0,1)^2$ be the unit square in $2D$, and consider the BVP

$$\begin{cases} -\Delta u = f \quad \text{in} \ \ \Omega, \\ \frac{\partial u}{\partial \nu} = \gamma \quad \text{on } \Gamma_N, \\ u = 0 \quad \Gamma_D \end{cases} \tag{2}$$

where $f(x,y) = 2x$, $\gamma(1,y) = y(1-y)$ on

$$\Gamma_N := \{(x,y) \in \partial\Omega : x = 1\}, \quad u(x,y) = xy(1-y).$$

Then

$$f \geq 0, \quad g = 0, \quad \gamma \geq 0 \quad \textit{and} \quad u \geq 0.$$

Graph NNP

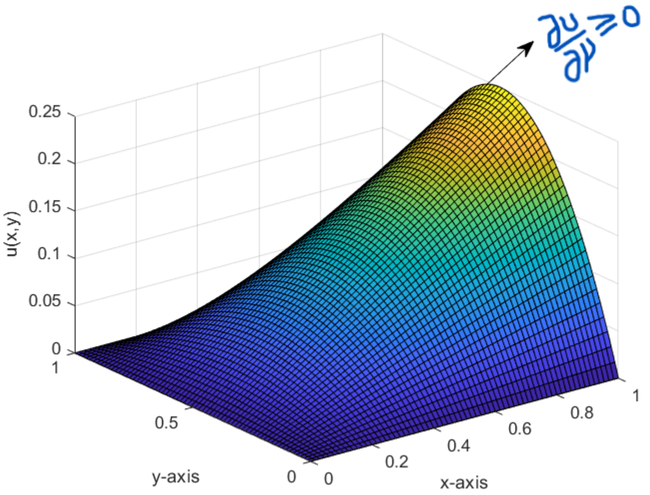


Figure: (NNP) $u(x,y) = xy(1-y)$

Continuous maximum principles

- Typical maximum principles 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

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

i.e. 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.

When does the continuous maximum principle hold? [8]

The maximum principle (MP) for elliptic operators (here $a > 0,\ q \geq 0$).
We consider Dirichlet b.c. For the mixed b.c: $\gamma \leq 0$ should also hold.

- Strong MP(SMP) for $Lu := -div(a\nabla u)$

$$f \leq 0 \Rightarrow \max_{\overline{\Omega}} u = \max_{\partial\Omega} g.$$

  i.e. the maximum is attained on the boundary.

- Weak Maximum Principle(WMP)for $Lu := -div(a\nabla u) + qu$

$$f \leq 0 \Rightarrow \max_{\overline{\Omega}} u \leq \max\{0, \max_{\partial\Omega} g\} := \max_{\partial\Omega} g^+$$

$$\max_{\overline{\Omega}} u \leq \max_{\partial\Omega} g^+$$

  (a nonnegative maximum is attained on the boundary). That is:

  - If $\max_{\partial\Omega} g \geq 0$, then

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

  - If $\max_{\partial\Omega} g \leq 0$, then

$$\max_{\overline{\Omega}} u \leq 0.$$

Continuous minimum principles

The minimum principle (mP) for elliptic operators (here $a > 0,\ q \geq 0$).
We consider Dirichlet b.c. For the mixed b.c: $\gamma \geq 0$ should also hold.

- Strong mP(SmP) for $Lu := -div(a\nabla u)$

$$f \geq 0 \Rightarrow \min_{\overline{\Omega}} u = \min_{\partial\Omega} g.$$

  i.e. the minimum is attained on the boundary.

- Weak Minimum Principle(WmP) for $Lu := -div(a\nabla u) + qu$

$$f \geq 0 \Rightarrow \min_{\overline{\Omega}} u \geq \min\{0, \min_{\partial\Omega} g\} := \min_{\partial\Omega} g^{-}$$

$$\min_{\overline{\Omega}} u \geq \min_{\partial\Omega} g^{-}$$

  (a nonpositive minimum is attained on the boundary).

  - If $\min_{\partial\Omega} g \leq 0$, then

$$\min_{\overline{\Omega}} u = \min_{\partial\Omega} g.$$

  - If $\min_{\partial\Omega} g \geq 0$, then

$$\min_{\overline{\Omega}} u \geq 0.$$

DMPs for the FE solution of linear PDEs

The discrete maximum principle(DMP): Analogous of the MP for the FE solution u_h.

Let us see the FE solution of a 1D reaction-diffusion problem where nonpositivity (a consequence of the MP) can fail for coarse mesh, refer to [1].

PDE BVP:

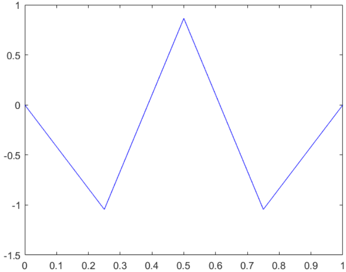
$$-\epsilon \Delta u + u = -(2x - 1)^2, \tag{3}$$

where $\epsilon = 2^{-10}, x \in (0, 1)$ and $u = 0$ on the boundary of the domain.

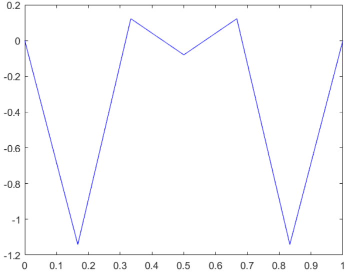
The graphs below illustrate how the numerical solutions look like, for different meshes.

FE solution of (3) for coarse meshes

Nonpositivity should hold since $f \leq 0$.
Here the numerical solution is expected to be $u_h \leq 0$, but $u_h \nleq 0$.



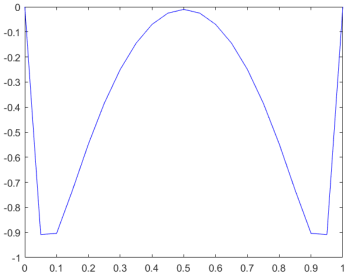
u$_h$ *for* $h = 0.25$



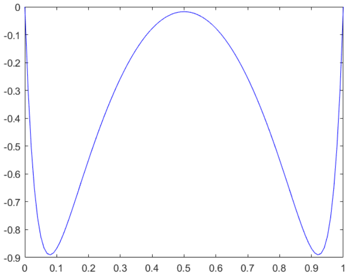
u$_h$ *for* $h = 0.17$

FE Solution of (3) for fine meshes

Here the numerical solution $u_h \leq 0$, since h is small enough.



u_h *for* $h = 0.05$



u_h *for* $h = 0.01$

Now, we extend our study to the DMPs for nonlinear models.

Note: for discrete case " h must be small enough".

DMPs for the FE solution of nonlinear PDEs

The goal of our research is to establish explicit conditions for preserving qualitative properties such as nonnegativity preservation and DMPs for nonlinear BVPs.

- Motivation: Similar results in [4, 6] for "small enough mesh size h".
- Achieved results: We have determined explicit conditions under Courant FEM for suitable mesh size in relation to angle condition for a nonlinear PDE.

Model problem

Let us consider the following nonlinear elliptic model:

$$\begin{cases} -\operatorname{div}\Big(b(x,u,\nabla u)\,\nabla u\Big) + r(x,u,\nabla u)u = f(x) \quad \text{in } \Omega, \\ b(x,u,\nabla u)\frac{\partial u}{\partial \nu} = \gamma(x) \quad \text{on } \Gamma_N, \\ u = g(x) \quad \text{on } \Gamma_D, \end{cases} \tag{4}$$

where Ω is a bounded domain in $\mathbf{R}^2$.

Assumptions

(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 $\overline{\Gamma}_N \cup \overline{\Gamma}_D = \partial\Omega$.

(b) The scalar functions $b : \overline{\Omega} \times \mathbf{R} \times \mathbf{R}^2 \to \mathbf{R}$ and $r : \overline{\Omega} \times \mathbf{R} \times \mathbf{R}^2 \to \mathbf{R}$ are continuous functions. 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

$$0 < \mu_0 \leq b(x, \xi, \eta) \leq \mu_1, \quad 0 \leq r(x, \xi, \eta) \leq \beta \tag{5}$$

$$\forall (x, \xi, \eta) \in \overline{\Omega} \times \mathbf{R} \times \mathbf{R}^2,$$

where μ_0 , μ_1 and β are positive constants.

Finite element approximation

Courant FEM:

The obtained nonlinear algebraic system of equations is:

$$\overline{\mathbf{A}}(\overline{\mathbf{c}})\overline{\mathbf{c}} = \overline{\mathbf{b}}, \tag{6}$$

where the structure of the matrix is :

$$\overline{\mathbf{A}}(\overline{\mathbf{c}}) = \begin{pmatrix} \mathbf{A}(\overline{\mathbf{c}}) & \widetilde{\mathbf{A}}(\overline{\mathbf{c}}) \\ \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{7}$$

- In (7), $\mathbf{I}$ is an $m \times m$ identity matrix, $\mathbf{0}$ is a $m \times n$ zero matrix and $\overline{\mathbf{A}}(\overline{\mathbf{c}})$ $(n+m)$ by $(n+m)$ matrix.
- The vector $\overline{\mathbf{c}} = (c_1, ..., c_{n+m})^T$ contains the values of the finite element solution u_h at all the nodal points. i.e. $c_i = u_h(P_i)$ and $u_h = \sum_{i=1}^{n+m} c_i\phi_i$, where $\phi_1, ...\phi_n$ are the interior basis functions and $\phi_{n+1}, ..., \phi_{n+m}$ are the boundary basis functions.
  We use the theory from [5] on linear systems.

The Definition and Theorem below are from [5]

Definition

The matrix $\bar{A}$ in (7) satisfies the *discrete weak maximum principle (DwMP)* if for any vector $\bar{c} = (c_1, ..., c_{n+m})^T \in \mathbf{R}^{n+m}$ satisfying $(\bar{A}\bar{c})_i \leq 0, \;\; i = 1, ..., n$, one has

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

Theorem

Let the matrix $\bar{A}$ in (7) satisfy the following conditions, where a_{ij} denote the entries of $\bar{A}$:

(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) *A is positive definite. Then $\bar{A}$ possesses the DwMP.*

Theorem 2

Angle condition on the mesh:

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 α_n in all T_k in all $\mathcal{T}_h$, where $\mathcal{T}_h \in \mathcal{F}$.

For the proof of our main result, we need the following Theorem.

Theorem

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

$$0 < h \leq h_0 = \left(\frac{12\cos(\alpha_0)\mu_0}{\beta} \right)^{\frac{1}{2}}, \qquad (8)$$

where α_0 is the angle that obeys the Definition, μ_0 and β are positive constants from (5). Then the matrix in (7) satisfies the following:

The matrix in (7) satisfies

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

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

(iii) **A** is positive definite.

The proof of (i):

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

$$a_{ij}(\bar{c}) = \int_\Omega \Big[b(x, u_h, \nabla u_h)\; \nabla\phi_i \cdot \nabla\phi_j + r(x, u_h, \nabla u_h)\; \phi_i\phi_j \Big]\, dx. \tag{9}$$

To estimate (9) we calculate the bounds of the following integrals in terms of the mesh size and angle condition:

$$\int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, dx \qquad and \qquad \int_\Omega \phi_i\phi_j \, dx \tag{10}$$

Stiffness matrix

From the Definition we have the maximum angle α_0, and $\sigma_0 > 0$ such that $cos(\alpha_0) = \sigma_0$ 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 (10).
The inner product of the basis functions: for any acute angle α, we have

$$\begin{aligned}\nabla\phi_i \cdot \nabla\phi_j &= |\nabla\phi_i|.|\nabla\phi_j| cos(180^0 - \alpha)\\ &= \frac{1}{h_i}.\frac{1}{h_j}(-cos(\alpha)) \leq \frac{-cos(\alpha)}{h^2}\\ &\leq \frac{-cos(\alpha_0)}{h^2} \quad \forall\, h_i,\ h_j \leq h,\ \forall\alpha \leq \alpha_0.\\ &\Rightarrow \nabla\phi_i \cdot \nabla\phi_j \leq -\frac{\sigma_0}{h^2} < 0 \qquad (11)\\ &\Rightarrow \int_\Omega \nabla\phi_i \cdot \nabla\phi_j\, dx \leq -\frac{\sigma_0}{h^2} meas(\Omega_{ij}). \qquad (12)\end{aligned}$$

Mass matrix

To estimate the mass matrix for general triangles, we use a reference triangle.
If E is the reference triangle with vertices (0, 0), (h, 0), and (0, h) then one can calculate

$$\int_E \phi_i \phi_j \, dx = \frac{h^2}{24} . \qquad (13)$$

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 $L_k : E \to T_k$.

Mass matrix

We also define $J_k = L'_k$. If the reference triangle E is considered with $h = 1$ in (13) and T_k is a fixed general triangle then

$$\int_{T_k} \phi_i \phi_j \, dx = det(J_k) \int_E \tilde{\phi}_i \tilde{\phi}_j \, dx = \frac{|T_k|}{12} \tag{14}$$

by change of variables and using the fact that $det(J_k) = 2|T_k|$, where $|T_k|$ is the area of the triangle, and $\tilde{\phi}_i$ and $\tilde{\phi}_j$ are respectively given by $\tilde{\phi}_i = \phi_i o L_k$, $\tilde{\phi}_j = \phi_j o L_k$. Therefore, (14) implies

$$\int_{\Omega_{ij}} \phi_i \phi_j \, dx = \sum_{T_k \in \Omega_{ij}} \int_{T_k} \phi_i \phi_j \, dx = \frac{1}{12} meas(\Omega_{ij}). \tag{15}$$

Nonpositivity

where $\Omega_{ij} := supp\ \phi_i \cap supp\ \phi_j$. Using (5),(12), and (15) in (9), we have

$$a_{ij}(\bar{c}) \leq \mu_0 \int\limits_{\Omega} \nabla\phi_i \cdot \nabla\phi_j \, dx + \beta \int\limits_{\Omega} \phi_i\phi_j \, dx$$

$$\leq -\frac{\sigma_0}{h^2}\mu_0 \,\mathrm{meas}\,(\Omega_{ij}) + \frac{\beta}{12}\ meas(\Omega_{ij}) = meas(\Omega_{ij})\left(\frac{-\sigma_0}{h^2}\mu_0 + \frac{\beta}{12}\right).$$

Let

$$a_{ij}(h) := meas(\Omega_{ij})\left(-\frac{\sigma_0}{h^2}\mu_0 + \frac{\beta}{12}\right) \tag{16}$$

then

$$a_{ij}(\bar{c}) \leq a_{ij}(h). \tag{17}$$

This implies $a_{ij}(h) \leq 0$ if h is small enough.

Choice of h

The main task here is to find how much h should be to get the nonpositivity.
To determine the optimal $h = h_0$, 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 $0 < h \leq h_0 = \left(\frac{12\sigma_0\mu_0}{\beta}\right)^{\frac{1}{2}}$, then $a_{ij}(\bar{c}) \leq 0$ from (17).

Theorem 3

In summary, the mesh size h is crucial to ensure the DMP of the proposed problem. With this, we state the main result.

Theorem

Under the conditions of Theorem 2 and letting $f \leq 0$ and $\gamma \leq 0$ we have

$$\max_{\overline{\Omega}} u_h \leq \max\{0, \max_{\Gamma_D} g_h\}. \tag{18}$$

In particular, if $\Gamma_D \neq \emptyset$ and $g \geq 0$, then

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h, \tag{19}$$

and if $\Gamma_D \neq \emptyset$ and $g \leq 0$, or if $\Gamma_D = \emptyset$, then we have the nonpositivity property

$$\max_{\overline{\Omega}} u_h \leq 0. \tag{20}$$

The main idea of the proof:

- Theorem 3 (the main theorem) is proved using the consequence of Theorem 2, Theorem 1 (which deals with the DMPs for the coordinates), and the effect of the right-hand side of the problem (4).
  - Since $f \leq 0, \gamma \leq 0$ and $0 \leq \phi_i \leq 1$, we obtain

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

This implies DMP for the coordinates.That is,

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

Goal:

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

The figure below illustrates the finite element solution u_h at the nodal points in 1D.

The finite element solution u_h at all the nodal points

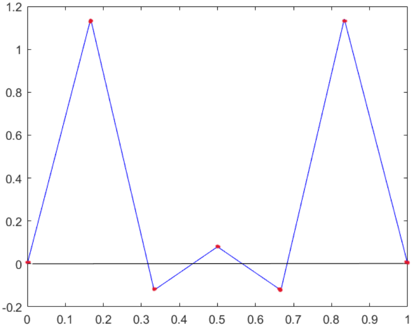


Figure: $u_h(P_i) = c_i$

Thus, using the fact that $0 \leq \phi_i \leq 1$ and $\sum_{i=1}^{n+m} \phi_i = 1$,

$$\max_{i=1,\ldots,n+m} c_i = \max_{\overline{\Omega}} u_h \quad and \quad \max_{i=n+1,\ldots,n+m} c_i = \max_{\Gamma_D} g_h.$$

Hence DMP for the solution itself holds.

Consequence

- As a consequence of the main theorem the corresponding discrete minimum principle and, as a special case, discrete non-negativity for system (4) can be verified in the same way by reversing signs.

Example

A special case of problem (4): Steady-state concentration u of some substrate in an enzyme-catalyzed reaction

$$\begin{cases} \operatorname{div}(D(x)\nabla u) = q(x,u) \quad \text{in} \quad \Omega, \\ \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma_N, \\ u = u_0 \quad \text{on} \quad \Gamma_D, \end{cases} \tag{21}$$

Michaelis-Menten theory

- Reaction rate by Michaelis-Menten theory:

$$q(x,\xi) = \frac{\epsilon^{-1}\xi}{\xi + k} \qquad \textit{for} \quad \xi \geq 0, \tag{22}$$

  where $k > 0$ is the Michaelis constant and $\epsilon > 0$ [7].
- The condition of $D(x)$: $0 < \mu_0 \leq D(x) \leq \mu_1$, where μ_0 and μ_1 are positive constants. Further, $u_0 \geq 0$ and $\beta = \frac{1}{\epsilon k}$.
  $q(x,\xi) = r(x,\xi)\xi$, where $r(x,\xi) = \frac{\epsilon^{-1}}{\xi+k}$ and $0 \leq r \leq \frac{1}{\epsilon k}$.
- Bounds of the FE solution under the conditions of Theorem 3:

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

  since $u_{0h} \geq 0$.

Conclusion

- We have been able to determine the threshold mesh size h using the acute angle condition and thus ensure the validity of DMPs for Courant FEM for suitably small mesh size 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] 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.

[5] János Karátson and Sergey Korotov. "Some discrete maximum principles arising for nonlinear elliptic finite element problems". In:

Computers & Mathematics with Applications 70.11 (2015), pp. 2732–2741.

[6] 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.

[7] 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.

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

[9] Tomáš Vejchodský. “The discrete maximum principle for Galerkin solutions of elliptic problems”. In: *Open Mathematics* 10.1 (2012), pp. 25–43.

Thank you for your attention!