On continuous and discrete maximum principles for elliptic problems with the third boundary condition

István Faragó*, Sergey Korotov†, Tamás Szabó‡

November 14, 2010

Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
H–1117, Budapest, Pázmány P. s. 1/c., Hungary
e-mail: faragois@cs.elte.hu, szabot@cs.elte.hu

BCAM - Basque Center for Applied Mathematics
Bizkaia Technology Park, Building 500, E – 48160 Derio
Basque Country, Spain
e-mail: korotov@bcamath.org

Abstract: In this work, we present and discuss some continuous and discrete maximum principles for linear elliptic problem of the second order with the third boundary condition (almost never addressed to in the available literature in this context) solved by the finite element and finite difference methods. Numerical tests are given.

Keywords: elliptic problem, the third boundary condition, maximum principle, discrete maximum principle

Mathematical Subject Classification: 35B50, 65N06, 65N30, 65N50

*The first author was supported by Hungarian National Research Fund OTKA No. K67819.

†The second author was supported by Grant MTM2008-03541 of the MICINN, Spain, the ERC Advanced Grant FP7-246775 NUMERIWAVES, and Grant PI2010-04 of the Basque Government.

‡The first and the third authors were supported by Jedlik project "ReCoMend" 2008–2011.

1 Continuous maximum principle

Consider the following boundary-value problem of elliptic type: Find a function $u \in C^2(\overline{\Omega})$ such that

$$-\Delta u + cu = f \quad \text{in} \quad \Omega \quad \text{and} \quad \alpha u + \frac{\partial u}{\partial n} = g \quad \text{on} \quad \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbf{R}^d$ is a bounded domain with Lipschitz continuous boundary $\partial\Omega$, n is the unit outward normal to $\partial\Omega$, the reactive coefficient $c(x) \geq 0$ for all $x \in \overline{\Omega}$, and the coefficient $\alpha(s) \geq 0$ for all $s \in \partial\Omega$. The boundary condition in (1) is often called the *the third boundary condition* (also known as Newton boundary condition or Robin boundary condition, see e.g. [10] for a relevant discussion of this subject). The additional assumptions on the data of the problem will be given in appropriate places of the paper later on.

First, we shall present the following key result - the continuous maximum principle (called CMP in short) for problem (1).

Theorem 1. *Assume that in* (1) *the functions* $c, f \in C(\overline{\Omega})$*, and the functions* $\alpha, g \in C(\partial\Omega)$*. In addition, let*

$$c(x) \geq c_0 > 0 \quad \textit{for all} \quad x \in \overline{\Omega} \quad \textit{and} \quad \alpha(s) \geq \alpha_0 > 0 \quad \textit{for all} \quad s \in \partial\Omega, \tag{2}$$

where c_0 *and* α_0 *are (positive) constants. Then the following (a priori) two-sided estimates for the classical solution of problem* (1) *are valid for any* $x \in \overline{\Omega}$*:*

$$\min\Big\{0, \min_{x\in\overline{\Omega}} \frac{f(x)}{c(x)}, \min_{s\in\partial\Omega} \frac{g(s)}{\alpha(s)}\Big\} \leq u(x) \leq \max\Big\{0, \max_{x\in\overline{\Omega}} \frac{f(x)}{c(x)}, \max_{s\in\partial\Omega} \frac{g(s)}{\alpha(s)}\Big\}. \tag{3}$$

Proof. First, we prove the upper estimate in (3). If $u \leq 0$ everywhere in $\overline{\Omega}$ then it is obviously valid. If u attains its positive maximum at some interior point $x_0 \in \Omega$, then all the first order partial derivatives $u'_{x_i}(x_0) = 0$, and all the second order partial derivatives $u''_{x_i x_i}(x_0) \leq 0$, therefore from the governing equation in (1) and the condition on c in (2) we observe that $u(x_0) \leq \frac{f(x_0)}{c(x_0)}$. If u attains its positive maximum at some boundary point $s_0 \in \partial\Omega$, then $\frac{\partial u}{\partial n}(s_0) \geq 0$, and therefore from the boundary condition in (1) and the condition on α in (2) we obtain, similarly to the previous case, that $u(s_0) \leq \frac{g(s_0)}{\alpha(s_0)}$. From these considerations the upper estimate in (3) follows immediately. The lower estimate in (3) can be proved just in the same way. □

The main goal of the paper will be to construct a suitable discrete analogue of (3) (called the discrete maximum principle (or DMP in short) later on) and find practical conditions on the numerical schemes (e.g. by the finite element method (FEM) or the finite difference method (FDM)) providing its validity. To the authors' knowledge, in the available literature on CMPs / DMPs for elliptic problems, see e.g. [12, 14] and references therein, CMPs usually take a form of some implications involving certain sign-conditions. For example, in our case, it reads as follows [12, p. 680]:

$$f(x) \leq 0 \text{ in } \overline{\Omega} \quad \& \quad g(s) \leq 0 \text{ on } \partial\Omega \implies \max_{x \in \overline{\Omega}} u(x) \leq 0. \tag{4}$$

Only recently in [8], the implications with sign-conditions (like in (4)) have been generalized to (more sharp and general) two-sided a priori error estimates (similarly to DMPs used in the parabolic case, see e.g. [6, 7]) via arbitrarily given data for the reaction-diffusion problems with nonzero reactive terms. However, it was only done for a special case of homogeneous Dirichlet boundary condition. Here, we apply the approach from [8] to a more complicated case of the third boundary condition given as in (1). The elliptic equations with the third boundary condition describe some real-life problems for example in electrical engineering (heat condiction in large transformers, etc), see e.g. [17] and [19] for concrete examples in this respect.

Remark 1. We mention that DMPs, besides their practical importance for imitating the nonegativity of nonnegative physical quantities in numerical simulations, have been often used for proving stability and finding the rate of convergence of FD approximations, see e.g. [1, 2, 4], and for proving the convergence of FE approximations in the maximum norm, see e.g. [1, 5].

2 Discrete maximum principle

After discretization of problem (1) by most of popular numerical techniques (e.g. by FEM and FDM, which are considered in this work) we arrive at the problem of solving $N \times N$ system of linear algebraic equations

$$\mathbf{A}\mathbf{u} = \mathbf{F}, \tag{5}$$

where the vector of unknowns $\mathbf{u} = [u_1, \ldots, u_N]^T$ approximates the unknown solution u at certain selected points $B_1, \ldots, B_N$ of the solution domain Ω and its boundary $\partial\Omega$, and the vector $\mathbf{F} = [F_1, \ldots, F_N]^T$ approximates (in the sense related to the nature of a concrete numerical method used) the values $f(B_i)$ and $g(B_i)$ (latter - due to the considered case of the third boundary condition) for $i = 1, \ldots, N$.

In what follows, the entries of matrix $\mathbf{A}$ will be denoted by a_{ij}, and all matrix and vector inequalities appearing in the text are always understood component-wise.

Definition 1. The square $N \times N$ matrix $\mathbf{M}$ is called *monotone* if

$$\mathbf{Mz} \geq 0 \quad \Longrightarrow \quad \mathbf{z} \geq 0. \tag{6}$$

Equivalently, montone matrices are characterized as follows (cf. [2, p. 119]).

Theorem 2. *The square $N \times N$ matrix $\mathbf{M}$ is monotone if and only if $\mathbf{M}$ is nonsingular and $\mathbf{M}^{-1} \geq 0$.*

Further, if one provides the matrix $\mathbf{A}$ in the system (5) be monotone then $\mathbf{A}^{-1} \geq 0$ and using assumption that $\mathbf{F} \leq 0$ (guaranteed by the sign-conditions $f \leq 0$ and $g \leq 0$ from CMPs similar to (4), e.g. for linear FEM and FDM schemes) we immediately get that $\mathbf{u} = \mathbf{A}^{-1}\mathbf{F} \leq 0$. This observation describes the standard proof of the following DMP

$$\mathbf{F} \leq 0 \quad \Longrightarrow \quad \mathbf{u} \leq 0, \tag{7}$$

which imitates some CMPs for linear elliptic equations with homogeneous Dirichlet boundary conditions (cf. [4, 5, 15, 12]).

Definition 2. The *infinity norm* $\|\cdot\|_\infty$ of the square $N \times N$ matrix $\mathbf{M}$ (with entries m_{ij}) is defined as

$$\|\mathbf{M}\|_\infty := \max_{i=1,\ldots,N} \sum_{j=1}^{N} |m_{ij}|. \tag{8}$$

Definition 3. The square $N \times N$ matrix $\mathbf{M}$ (with entries m_{ij}) is called *strictly diagonally dominant* (or SDD in short) if the values

$$\delta_i(\mathbf{M}) := |m_{ii}| - r_i > 0 \quad \text{for all} \quad i = 1, \ldots, N, \tag{9}$$

where r_i is the sum of absolute values of all off-diagonal entries in the i-th row of $\mathbf{M}$, i.e. $r_i := \sum_{j=1,\, j\neq i}^{N} |m_{ij}|$.

The following theorem has been proved in [8], see also [20, 22] for close results.

Theorem 3. *Let matrix* $\mathbf{A}$ *in system* (5) *be SDD and monotone. Then the following two-sided estimates for the entries of the solution* $\mathbf{u}$ *are valid*

$$\min\Big\{0,\ \min_{j=1,\dots,N}\frac{F_j}{\delta_j(\mathbf{A})}\Big\}\le u_i\le\max\Big\{0,\ \max_{j=1,\dots,N}\frac{F_j}{\delta_j(\mathbf{A})}\Big\},\quad i=1,\dots,N. \tag{10}$$

Remark 2. It is obvious that estimates (10) imply DMP (7) provided $\mathbf{F}\le 0$.

Remark 3. The bounds in (10) are achievable, e.g. in the case of $\mathbf{A}$ being the unit matrix.

As (10) resembles the estimates in (3), it is natural to give the following definition.

Definition 4. We say that the solution $\mathbf{u}$ of system (5) with SDD matrix $\mathbf{A}$ satisfies the *discrete maximum principle* (DMP) corresponding to CMP (3), if, first, estimates (10) are valid and if, in addition, the estimates

$$\begin{aligned}\max_{j=1,\dots,N}\frac{F_j}{\delta_j(\mathbf{A})}&\le\max\Big\{0,\max_{x\in\overline{\Omega}}\frac{f(x)}{c(x)},\max_{s\in\partial\Omega}\frac{g(s)}{\alpha(s)}\Big\},\\ \min_{j=1,\dots,N}\frac{F_j}{\delta_j(\mathbf{A})}&\ge\min\Big\{0,\min_{x\in\overline{\Omega}}\frac{f(x)}{c(x)},\min_{s\in\partial\Omega}\frac{g(s)}{\alpha(s)}\Big\},\end{aligned} \tag{11}$$

are valid.

Remark 4. In the case of earlier versions of continuous and discrete maximum principles no estimates like (11) were, in fact, needed as one dealt there with various implications involving the sign-conditions only.

Remark 5. The validity of relations (11) is important for producing controllable numerical approximations as, for example, linear FE and FD approximations do stay then within the same (a priori known from the continuous problem) bounds as the exact solutions they do approximate.

Remark 6. While the SDD-property of $\mathbf{A}$ is essentially automatically guaranteed (after discretization) by the nature of the problem under consideration (see conditions (2)), its monotonicity, required in Theorem 3, should be provided a priori (or proved separately in each concrete case). One common approach for treating this issue in FEM is to impose certain a priori geometric requirements on the FE meshes employed so that all the off-diagonal entries $a_{ij}\le 0$ (see e.g. [3, 5, 11, 12, 15] for more details on this subject). As far it concerns FDM, this property for the off-diagonal entries of $\mathbf{A}$ is often guaranteed a priori by many standard FD schemes producing the so-called M-matrices [9].

Remark 7. One of advantages for dealing with the property $a_{ij} \leq 0$ $(i \neq j)$ is an easy calculation of values $\delta_i(\mathbf{A})$ and establishing their relation to the coefficients c and α (see the next sections).

Later on, we shall be often using the following auxiliary inequalities.

Lemma 1. *For any real numbers λ_1, λ_2 and any real numbers $\mu_1, \mu_2 > 0$ the estimates*

$$\min\Big\{0, \frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\Big\} \leq \frac{\lambda_1 + \lambda_2}{\mu_1 + \mu_2} \leq \max\Big\{0, \frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\Big\} \tag{12}$$

are valid.

Proof. It follows from a straightforward calculation. □

3 DMPs for the finite element schemes

The standard FE scheme is based on the so-called variational formulation of (1), which reads: Find $u \in H^1(\Omega)$ such that

$$a(u, v) = \mathcal{F}(v) \qquad \forall v \in H^1(\Omega), \tag{13}$$

where

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v dx + \int_\Omega cuvdx + \int_{\partial\Omega} \alpha uvds\,,\ \mathcal{F}(v) = \int_\Omega fvdx + \int_{\partial\Omega} gvds. \tag{14}$$

The existence and uniqueness of the (weak) solution u is provided by the Lax-Milgram lemma, the Friedrichs-type inequalities, and assumptions (2), see e.g. [16, Chapt. 2]. (Actually, for the well-posedeness in above, one can require only that $c \in L^\infty(\Omega), f \in L^2(\Omega), \alpha \in L^\infty(\partial\Omega), g \in L^2(\partial\Omega)$.)

Let $\mathcal{T}_h$ be a FE mesh of $\overline{\Omega}$ with interior nodes $B_1, \ldots, B_n$ lying in Ω and boundary nodes $B_{n+1}, \ldots, B_{n+n^\partial}$ lying on $\partial\Omega$. The elements of $\mathcal{T}_h$ we will denote by the symbol T, possibly with subindices. Further, let the basis functions $\phi_1, \phi_2, \ldots, \phi_{n+n^\partial}$, associated with these nodes, have the following properties (easily met if e.g. simplicial, block or prismatic FE meshes are used):

$$\phi_i(B_j) = \delta_{ij},\ i, j = 1, \ldots, n+n^\partial,\ \phi_i \geq 0 \text{ in } \overline{\Omega},\ i = 1, \ldots, n+n^\partial,\ \sum_{i=1}^{n+n^\partial} \phi_i \equiv 1 \text{ in } \overline{\Omega}, \tag{15}$$

where δ_{ij} is the Kronecker delta. The basis functions $\phi_1, \phi_2, \ldots, \phi_{n+n^\partial}$ are spanning a finite-dimensional subspace V_h of $H^1(\Omega)$.

The FE approximation of u is defined as a function $u_h \in V_h$ such that

$$a(u_h, v_h) = \mathcal{F}(v_h) \qquad \forall v_h \in V_h, \tag{16}$$

whose existence and uniqueness are also provided by the Lax-Milgram lemma.

Remark 8. Algorithmically, $u_h = \sum_{i=1}^{n+n^\partial} u_i \phi_i$, where the coefficients u_i are the entries of the solution $\mathbf{u}$ of system (5) with $a_{ij} = a(\phi_i, \phi_j)$, $F_i = \mathcal{F}(\phi_i)$, and $N = n + n^\partial$. It is clear that, if properties (15) hold, the FE approximation u_h satisfies the bounds from (10) at each point of $\overline{\Omega}$ if all its nodal values u_i do satisfy them.

Lemma 2. *Assume that problem* (1) *(under condition on coefficients* (2)*) is solved by FEM with the basis functions having properties* (15)*. In addition, let matrix* $\mathbf{A}$ *in the resulting matrix equation* $\mathbf{Au} = \mathbf{F}$ *be such that* $a_{ij} \le 0$ $(i \ne j)$*. Then* $\mathbf{A}$ *is SDD and the estimates* (10) *are valid.*

Proof. Clearly, from (14) and (2), it follows that $a_{ii} = a(\phi_i, \phi_i) > 0$ for all $i = 1, \dots, n + n^\partial$. If $a_{ij} \le 0$ $(i \ne j)$, we observe for $i = 1, \dots, n + n^\partial$ that

$$\delta_i(\mathbf{A}) = \sum_{j=1}^{n+n^\partial} a_{ij} = a(\phi_i, \sum_{j=1}^{n+n^\partial} \phi_j) = a(\phi_i, 1) = \int_\Omega c\phi_i dx + \int_{\partial\Omega} \alpha\phi_i ds > 0, \tag{17}$$

where the last (strict) inequality holds due to (2). (We notice that, in fact, $\delta_i(\mathbf{A}) = \int_\Omega c\phi_i dx$ if $i \in \{1, \dots, n\}$.) Thus, the matrix $\mathbf{A}$ is always SDD for our type of problems. Moreover $\mathbf{A}$ is the Minkowski matrix, and therefore it is monotone (cf. [2, p. 119]). Hence, estimates (10) are valid, due to Theorem 3, with $\delta_i(\mathbf{A})$ computed as in (17). □

In general, the proofs of estimates (11) strongly depend on how we compute a_{ij} and F_j in real FEM calculations. However, we consider in detail only the following quite representative case.

Theorem 4. *Assume that the coefficients* c *and* α *are (positive) constant and the functions* f *and* g *are such (e.g. piecewise polynomials) that all entries* a_{ij} *and* F_j *in system* (5) *are computed exactly. Then estimates* (11)*, and therefore DMP, corresponding to* (3)*, are valid provided* $a_{ij} \le 0$ $(i \ne j)$*.*

Proof. We see immediately that, if $F_i \le 0$ for all $i = 1, \dots, n + n^\partial$, then the upper estimate in (11) holds. Let now $F_{i_0} > 0$ for some index $i_0 \in \{1, \dots, n\}$. Then due to (14) and (17) we observe

$$\frac{F_{i_0}}{\delta_{i_0}(\mathbf{A})} = \frac{\int_\Omega f\phi_{i_0} dx}{\int_\Omega c\phi_{i_0} dx} \le \frac{\int_\Omega \max\{0,\ \max_{\xi\in\overline{\Omega}} f(\xi)\}\, \phi_{i_0} dx}{c\int_\Omega \phi_{i_0} dx} = \max\Big\{0, \max_{\xi\in\overline{\Omega}} \frac{f(\xi)}{c}\Big\}.$$

Let now $F_{i_0} > 0$ for some index $i_0 \in \{n+1, \ldots, n+n^\partial\}$. Then, in view of (17), (14), (12), and (2), we get

$$\frac{F_{i_0}}{\delta_{i_0}(\mathbf{A})} = \frac{\int_\Omega f\phi_{i_0}dx + \int_{\partial\Omega} g\phi_{i_0}ds}{\int_\Omega c\phi_{i_0}dx + \int_{\partial\Omega} \alpha\phi_{i_0}ds} \leq$$

$$\leq \frac{\int_\Omega \max\{0,\ \max\limits_{\xi\in\overline{\Omega}} f(\xi)\}\,\phi_{i_0}dx + \int\limits_\Omega \max\{0,\ \max\limits_{\xi\in\partial\Omega} g(\xi)\}\,\phi_{i_0}ds}{c\int_\Omega \phi_{i_0}dx + \alpha\int_{\partial\Omega}\phi_{i_0}ds} \leq$$

$$\leq \max\Big\{0, \max_{\xi\in\overline{\Omega}} \frac{f(\xi)}{c}, \max_{\xi\in\partial\Omega}\frac{g(\xi)}{\alpha}\Big\}.$$

Similarly, if $F_i \geq 0$ for all $i = 1, \ldots, n+n^\partial$, then the lower estimate in (11) holds. Let now $F_{i_0} < 0$ for some index $i_0 \in \{1, \ldots, n\}$. Then, as in the previous case,

$$\frac{F_{i_0}}{\delta_{i_0}(\mathbf{A})} = \frac{\int_\Omega f\phi_{i_0}dx}{\int_\Omega c\phi_{i_0}dx} \geq \frac{\int_\Omega \min\{0,\ \min_{\xi\in\overline{\Omega}} f(\xi)\}\,\phi_{i_0}dx}{c\int_\Omega\phi_{i_0}dx} = \min\Big\{0, \min_{\xi\in\overline{\Omega}}\frac{f(\xi)}{c}\Big\}.$$

Let now $F_{i_0} < 0$ for some index $i_0 \in \{n+1, \ldots, n+n^\partial\}$. Then, in view of (17), (14), (12), and (2), we observe that

$$\frac{F_{i_0}}{\delta_{i_0}(\mathbf{A})} = \frac{\int_\Omega f\phi_{i_0}dx + \int_{\partial\Omega} g\phi_{i_0}ds}{\int_\Omega c\phi_{i_0}dx + \int_{\partial\Omega} \alpha\phi_{i_0}ds} \geq$$

$$\geq \frac{\int_\Omega \min\{0,\ \min\limits_{\xi\in\overline{\Omega}} f(\xi)\}\,\phi_{i_0}dx + \int_\Omega \min\{0,\ \min\limits_{\xi\in\partial\Omega} g(\xi)\}\,\phi_{i_0}ds}{c\int_\Omega \phi_{i_0}dx + \alpha\int_{\partial\Omega}\phi_{i_0}ds} \geq$$

$$\geq \min\Big\{0, \min_{\xi\in\overline{\Omega}} \frac{f(\xi)}{c}, \min_{\xi\in\partial\Omega}\frac{g(\xi)}{\alpha}\Big\}.$$

□

Remark 9. If c and α are not constant, and f and g are general functions, then for computations of entries (which are sums of integrals over Ω and its boundary $\partial\Omega$) in system (5), we should, in practice, use certain quadrature rules, and, thus, each such a case may require a separate analysis. For example, we can obviously prove validity of estimates (11) in our case if a simple quadrature rule considered in [8] is used. For more complicated situations, the corresponding analysis can be done as in [13].

4 DMPs for some finite difference schemes

In this section, on the base of the two representative schemes, we shall demonstrate how our DMP can be proved for FDMs in principle.

Consider the following two-dimensional square domain $\Omega = (0,1)\times(0,1)$. Using the same step-size $h = 1/\hat{n}$ in both directions and the classical 5-point FD stencil, we arrive at the following equations inside the solution domain

$$\frac{-y_{i-1,j} - y_{i+1,j} - y_{i,j-1} - y_{i,j+1} + 4y_{i,j}}{h^2} + c_{i,j}y_{i,j} = f_{i,j}, \tag{18}$$

where $i,j = 1,...,\hat{n}-1$, i.e. we have n^* (interior) equations, where $n^* := (\hat{n}-1)^2$.

The well-known first order accurate discretization of the third boundary condition on $\partial\Omega$ (consisting, in our case, of four intervals) reads as follows:

$$\alpha_{i,0}y_{i,0} + \frac{y_{i,0} - y_{i,1}}{h} = g_{i,0} \quad \text{for all} \quad i = 1,2,...,\hat{n}-1, \tag{19}$$

$$\alpha_{i,\hat{n}}y_{i,\hat{n}} + \frac{y_{i,\hat{n}} - y_{i,\hat{n}-1}}{h} = g_{i,\hat{n}} \quad \text{for all} \quad i = 1,2,...,\hat{n}-1, \tag{20}$$

$$\alpha_{0,j}y_{0,j} + \frac{y_{0,j} - y_{1,j}}{h} = g_{0,j} \quad \text{for all} \quad j = 1,2,...,\hat{n}-1, \tag{21}$$

$$\alpha_{\hat{n},j}y_{\hat{n},j} + \frac{y_{\hat{n},j} - y_{\hat{n}-1,j}}{h} = g_{\hat{n},j} \quad \text{for all} \quad j = 1,2,...,\hat{n}-1, \tag{22}$$

i.e. we get, in addition, n^0 (boundary) equations, where $n^0 := 4(\hat{n}-1)$.

Finally, we compose a square FD system of linear equations with the matrix denoted by $\mathbf{A}$ and $n^* + n^0$ unknowns.

Theorem 5. *The finite difference discretization* (18)–(22) *has the following properties:*

a) for sufficiently smooth solution u, it approximates the exact solution with the first order,

b) the resulting FD matrix $\mathbf{A}$ is strictly diagonally dominant and monotone,

c) the DMP estimates (11) *are valid.*

Proof. The first statement is obvious, due to the first order accuracy of the approximation of the boundary condition.

Further, an easy computation shows that (we shall use a single numeration for all the nodes now, it should not lead to any misunderstanding, also we

use denotation like f_i, c_i, etc to denote of values of f, c, etc at the node with the index i)

$$\delta_i(\mathbf{A}) = c_i > 0 \quad \text{for} \quad i = 1, ..., n^* \quad (for\ interior\ nodes), \tag{23}$$

$$\delta_i(\mathbf{A}) = \alpha_i > 0 \quad \text{for} \quad i = n^* + 1, ..., n^* + n^0 \quad (for\ boundary\ nodes). \tag{24}$$

Since α and c are positive functions, the matrix $\mathbf{A}$ is thus SDD and monotone.

Further, for the right-hand side of the system we observe that

$$F_i = f_i \quad \text{for} \quad i = 1, ..., n^*, \tag{25}$$

$$F_i = g_i \quad \text{for} \quad i = n^* + 1, ..., n^* + n^0. \tag{26}$$

Now, for $i = 1, ..., n^*$, the following inequalities hold

$$\frac{F_i}{\delta_i(\mathbf{A})} = \frac{f_i}{c_i} \le \max\left\{0, \max_{i=1,...,n^*} \frac{f_i}{c_i}\right\} \le \max\left\{0, \max_{x\in\bar{\Omega}} \frac{f(x)}{c(x)}\right\}, \tag{27}$$

and for $i = n^* + 1, ..., n^* + n^0$, we observe that

$$\frac{F_i}{\delta_i(\mathbf{A})} = \frac{g_i}{\alpha_i} \le \max\left\{0, \max_{i=n^*+1,...,n^*+n^0} \frac{g_i}{\alpha_i}\right\} \le \max\left\{0, \max_{s\in\partial\Omega} \frac{g(s)}{\alpha(s)}\right\}. \tag{28}$$

Similarly we proceed with a proof of the lower estimate, therefore (11) holds. □

The approximation of the third boundary condition considered just above has only the first order of accuracy, which is not consistent with the second order of accuracy of FD approximation for the differential equation. Therefore, we shall present now an another FD scheme with an increased accuracy of the approximation of the third boundary condition. Namely, let us approximate the third boundary condition on the boundary of $\Omega = (0,1) \times (0,1)$ in the following manner:

- On the part of the boundary with $x = 0$:

$$\begin{aligned} &\frac{y_{0,j} - y_{1,j}}{h} - \frac{h}{2}\left(\frac{y_{0,j+1} - 2y_{0,j} + y_{0,j-1}}{h^2}\right) + \frac{h}{2}c_{0,j}y_{0,j} + \alpha_{0,j}y_{0,j} = \\ &= g_{0,j} + \frac{h}{2}f_{0,j}, \quad j = 1, 2, \ldots, \hat{n} - 1. \end{aligned} \tag{29}$$

- On the part of the boundary with $x = 1$:

$$\begin{aligned} &\frac{y_{\hat{n},j} - y_{\hat{n}-1,j}}{h} - \frac{h}{2}\left(\frac{y_{\hat{n},j+1} - 2y_{\hat{n},j} + y_{\hat{n},j-1}}{h^2}\right) + \frac{h}{2}c_{\hat{n},j}y_{\hat{n},j} + \alpha_{\hat{n},j}y_{\hat{n},j} = \\ &= g_{\hat{n},j} + \frac{h}{2}f_{\hat{n},j}, \quad j = 1, 2, \ldots, \hat{n} - 1. \end{aligned} \tag{30}$$

- On the part of the boundary with $y = 0$:

$$\begin{aligned}\frac{y_{i,0} - y_{i,1}}{h} - \frac{h}{2}\left(\frac{y_{i+1,0} - 2y_{i,0} + y_{i-1,0}}{h^2}\right) + \frac{h}{2}c_{i,0}y_{i,0} + \alpha_{i,0}y_{i,0} = \\ = g_{i,0} + \frac{h}{2}f_{i,0}, \quad i = 1, 2, \ldots, \hat{n} - 1.\end{aligned} \tag{31}$$

- On the part of the boundary with $y = 1$:

$$\begin{aligned}\frac{y_{i,\hat{n}} - y_{i,\hat{n}-1}}{h} - \frac{h}{2}\left(\frac{y_{i+1,\hat{n}} - 2y_{i,\hat{n}} + y_{i-1,\hat{n}}}{h^2}\right) + \frac{h}{2}c_{i,0}y_{i,\hat{n}} + \alpha_{i,\hat{n}}y_{i,\hat{n}} = \\ = g_{i,\hat{n}} + \frac{h}{2}f_{i,\hat{n}}, \quad i = 1, 2, \ldots, \hat{n} - 1.\end{aligned} \tag{32}$$

Theorem 6. *The finite difference discretization* (18),(29)–(32) *has the following properties:*

a) for the solution $u \in C^4(\overline{\Omega})$ it approximates the exact solution with the second order,

b) the resulting FD matrix $\mathbf{A}$ *is strictly diagonally dominant and monotone,*

c) the DMP estimates (11) *are valid.*

Proof. We shall prove the statement a) only for the case of the part of the boundary with $x = 1$. (The proofs of the other cases are quite similar.) Clearly, it is enough to show the second order of approximation at the boundary nodes only. Let us define

$$\begin{aligned}\Psi_j = &\frac{u(1, jh) - u(1 - h, jh)}{h} - \\ &- \frac{h}{2}\left(\frac{u(1, (j+1)h) - 2u(1, jh) + u(1, (j-1)h)}{h^2}\right) + \\ &+ \frac{h}{2}c(1, jh)u(1, jh) + \alpha(1, jh)u(1, jh) - g(1, jh) - \frac{h}{2}f(1, jh).\end{aligned} \tag{33}$$

Using the Taylor expansion, we get

$$\frac{u(1, jh) - u(1 - h, jh)}{h} = \left(\partial_1 u - \frac{h}{2}\partial_{11}^2 u\right)_{(1,jh)} + \mathcal{O}(h^2), \tag{34}$$

$$\frac{u(1, (j+1)h) - 2u(1, jh) + u(1, (j-1)h)}{h^2} = \left(\partial_{22}^2 u\right)_{(1,jh)} + \mathcal{O}(h^2). \tag{35}$$

Hence, putting (34) and (35) into (33), we obtain

$$\Psi_j = (\partial_1 u - \alpha u - g)_{(1,jh)} - \frac{h}{2}\left(\partial^2_{11}u + \partial^2_{22}u + cu + f\right)_{(1,jh)} + \mathcal{O}(h^2). \quad (36)$$

Since $\frac{\partial u}{\partial n}(1,y) = \partial_1 u(1,y)$, the first term in the right-hand side of (36) vanishes due to the boundary condition in (1). The second term is also equal to zero. This shows the validity of a).

Obviously, (23) holds, therefore to prove the statement b), it is enough to show the diagonally dominance at the boundary nodes only. At these nodes we have (we use again, for simplicity, a single indexing)

$$\delta_i(\mathbf{A}) = \frac{h}{2}c_i + \alpha_i > 0 \quad \text{for} \quad i = n^* + 1, ..., n^* + n^0. \quad (37)$$

Therefore, under our assumptions $\mathbf{A}$ is SDD matrix, and, due its sign-structure, it is also M-matrix and, therefore, monotone.

To prove the statement c), we observe that for the right-hand side of the resulting FD system we have (25) for the interior nodes, and

$$F_i = g_i + \frac{h}{2}f_i \quad \text{for} \quad i = n^* + 1, ..., n^* + n^0 \quad (38)$$

at the boundary points, respectively. Due to the property b) Theorem 3 can now be used. Obviously, at the interior nodes we have again estimates (27). Using Lemma 1, at the boundary nodes (i.e., for indices $i = n^*+1, ..., n^*+n^0$) we obtain

$$\begin{aligned} \frac{F_i}{\delta_i(\mathbf{A})} = \frac{g_i + f_i h/2}{\alpha_i + c_i h/2} &\le \max\left\{0, \max_{i=n^*+1,...,n^*+n^0}\frac{g_i}{\alpha_i}, \max_{i=n^*+1,...,n^*+n^0}\frac{f_i}{c_i}\right\} \\ &\le \max\left\{0, \max_{s\in\partial\Omega}\frac{g(s)}{\alpha(s)}, \max_{x\in\Omega}\frac{f(x)}{c(x)}\right\}. \end{aligned} \quad (39)$$

The inequalities (27) and (39) altogether prove the upper estimate in (11). Similarly we can prove the associated lower estimate in (11). □

5 Numerical tests

In this section the two-sided estimation for the solution of the following problem is checked for the cases of linear FEM and two FDM schemes analysed in the previous section:

$$-\frac{d^2u}{dx^2} + u = 4xe^x, \ x \in \Omega = (0,1), \quad (40)$$

$$\frac{\partial u}{\partial n} + u(s) = g, \quad s \in \partial\Omega\ . \quad (41)$$

Here g is the function defined at the boundary points $x = 0$ and $x = 1$ as $g(0) = -1$ and $g(1) = -e$. We devide $(0, 1)$ into N equal segments of the length denoted by h, i.e. $h = 1/N$. For the first order accurate approximation the FD scheme is straightforward, for the second order accurate approximation we use the scheme

$$u_0 \left(\frac{1}{h} + \alpha(0) + \frac{h}{2} c(0) \right) - u_1 \frac{1}{h} = g(0) + \frac{h}{2} f(0), \tag{42}$$

$$u_N \left(\frac{1}{h} + \alpha(1) + \frac{h}{2} c(1) \right) - u_{N-1} \frac{1}{h} = g(1) + \frac{h}{2} f(1), \tag{43}$$

i.e.,

$$u_0 \left(\frac{1}{h} + 1 + \frac{h}{2} \right) - u_1 \frac{1}{h} = -1, \tag{44}$$

$$u_N \left(\frac{1}{h} + 1 + \frac{h}{2} \right) - u_{N-1} \frac{1}{h} = e\,(2h - 1)\,. \tag{45}$$

The exact solution of problem (40)–(41) is

$$u(x) = x(1 - x)e^x. \tag{46}$$

Therefore, we get

$$\min_{\bar{\Omega}}(u(x)) = 0,\ \max_{\bar{\Omega}}(u(x)) = 2 \left(\sqrt{\frac{5}{4}} - 1 \right) \exp \left(\sqrt{\frac{5}{4} - \frac{1}{2}} \right) \approx 0.43797. \tag{47}$$

It is clear that the two-sided estimation in (3) is valid, and after a short calculation we obtain the following estimates

$$u(x) \geq \min \Big\{ 0, \min_{x \in \overline{\Omega}} \frac{f(x)}{c(x)}, \min_{s \in \partial\Omega} \frac{g(s)}{\alpha(s)} \Big\} = -e \approx -2.71828 \tag{48}$$

and

$$u(x) \leq \max \Big\{ 0, \max_{x \in \overline{\Omega}} \frac{f(x)}{c(x)}, \max_{s \in \partial\Omega} \frac{g(s)}{\alpha(s)} \Big\} = 4e \approx 10.8731\ . \tag{49}$$

Table 1 and Table 2 contain the results of two-sided estimation for the numerical solutions for our test problem.

Table 1

	Estimated discrete				Real discrete			
	FDM 1st		FDM 2nd		FDM 1st		FDM 2nd	
N	**min**	**max**	**min**	**max**	**min**	**max**	**min**	**max**
10	-2.7182	8.8546	-2.0711	8.8546	-0.2663	0.2544	7.37e-5	0.4424
100	-2.7182	10.6573	-2.6507	10.6573	-0.0271	0.4194	6.94e-7	0.438
1000	-2.7182	10.8514	-2.7115	10.8514	-0.0027	0.4361	6.94e-9	0.438

Table 2

	linear FEM			
	Estimated discrete		Real discrete	
N	**min**	**max**	**min**	**max**
10	-0.0834	7.1633	-0.0695	0.3777
100	-0.1261	10.4452	-0.0088	0.4318
1000	-0.1627	10.8297	-9.03e-4	0.4374

The numbers in Table 3 support the theoretical analysis, namely, that the second order method converges much faster (in the maximum norm) than the numerical solution based on the first order approximation.

Table 3

N	**FDM 1st**	**FDM 2st**	**linear FEM**
10	2.66e-1	1.303e-2	8.24e-2
100	2.71e-2	1.305e-4	8.97e-3
1000	2.71e-3	1.305e-6	9.05e-4

6 Final remarks

It would be interesting to obtain suitable practical conditions guaranteeing the validity of our variant of DMP also for various hp-versions of FEM (see [21]), and analyse the case of elliptic problems with full diffusive tensors (cf. [18]).

References

[1] Axelsson, O., Kolotilina, L., Monotonicity and discretization error estimates, *SIAM J. Numer. Anal.* 27 (1990), 1591–1611.

[2] Bramble, J. H., Hubbard, B. E., On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type, *J. Math. and Phys.* 43 (1964), 117–132.

[3] Brandts, J., Korotov, S., Křížek, M., Šolc, J., On nonobtuse simplicial partitions, *SIAM Rev.* 51 (2009), 317–335.

[4] Ciarlet, P.G., Discrete maximum principle for finite-difference operators, *Aequationes Math.* 4 (1970), 338–352.

[5] Ciarlet, P.G., Raviart, P.-A., Maximum principle and uniform convergence for the finite element method, *Comput. Methods Appl. Mech. Engrg.* 2 (1973), 17–31.

[6] Faragó, I., Discrete maximum principle for finite element parabolic models in higher dimensions, *Math. Comp. Sim.* 80 (2010) 1601–1611.

[7] Faragó, I., Horváth, R., Continuous and discrete parabolic operators and their qualitative properties, *IMA Numer. Anal.*, 29 (2009), 606–631.

[8] Faragó, I., Korotov, S., Szabó, T., On modifications of continuous and discrete maximum principles for reaction-diffusion problems *Adv. Appl. Math. Mech.* 3 (2011), 109–120.

[9] Forsythe, G. E., Wasow, W. R., *Finite-Difference Methods for Partial Differential Equations*. Reprint of the 1960 original. Dover Phoenix Editions. Dover Publications, Inc., Mineola, NY, 2004.

[10] Gustafson, K., Abe, T., The third boundary condition - was it Robin's? *The Mathematical Intelligencer* 20(1) (1998), 63–71.

[11] Hannukainen, A., Korotov, S., Vejchodský, T., Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes, *J. Comput. Appl. Math.* 226 (2009), 275–287.

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

[13] Karátson, J., Korotov, S., Discrete maximum principles for finite element solutions of some mixed nonlinear elliptic problems using quadratures, *J. Comput. Appl. Math.* 192 (2006), 75–88.

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

[15] Křížek, M., Qun Lin, On diagonal dominance of stiffness matrices in 3D, *East-West J. Numer. Math.*, 3 (1995), 59–69.

[16] Křížek, M., Neittaanmäki, P., *Finite Element Approximation of Variational Problems and Applications*, Longman Scientific & Technical, Harlow, 1990.

[17] Křížek, M., Neittaanmäki, P., *Mathematical and Numerical Modelling in Electrical Engineering. Theory and Applications*, Kluwer Academic Publishers, Dordrecht, 1996.

[18] Kuzmin, D., Shashkov, M. J., Svyatskiy, D., A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems, *J. Comput. Phys.* 228 (2009), 3448–3463.

[19] Preiningerová, V., Křížek, M., Kahoun, V., Temperature distribution in large transformer cores. In: Proc. of CANZA Conf. (M. Franyó, ed.), Budapest, 1985, pp. 254–261.

[20] Smelov, V. V., Extension of the algebraic aspect of the discrete maximum principle, *Russian J. Numer. Anal. Math. Modelling* 22 (2007), 601–614.

[21] Vejchodský, T., Šolín, P., Discrete maximum principle for higher-order finite elements in 1D, *Math. Comp.* 76 (2007), 1833–1846.

[22] Windisch, G., A maximum principle for systems with diagonally dominant M-matrices. In: Discretization in Differential Equations and Enclosures (ed. E. Adams et al.), Math. Res., 36, Akademie-Verlag, Berlin, 1987, 243–250.