Helsinki University of Technology Institute of Mathematics Research Reports
Espoo 2008 A549

ON WEAKENING CONDITIONS FOR DISCRETE MAXIMUM PRINCIPLES FOR LINEAR FINITE ELEMENT SCHEMES

Antti Hannukainen Sergey Korotov Tomáš Vejchodský

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Institute of Mathematics Research Reports
Espoo 2008 A549

ON WEAKENING CONDITIONS FOR DISCRETE MAXIMUM PRINCIPLES FOR LINEAR FINITE ELEMENT SCHEMES

Antti Hannukainen **Sergey Korotov** **Tomáš Vejchodský**

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Mathematics and Systems Analysis

Antti Hannukainen, Sergey Korotov, Tomáš Vejchodský: *On weakening conditions for discrete maximum principles for linear finite element schemes*; Helsinki University of Technology Institute of Mathematics Research Reports A549 (2008).

Abstract: *In this work we discuss weakening requirements on the set of sufficient conditions due to Ph. Ciarlet [4, 5] for matrices associated to linear finite element schemes, which is commonly used for proving validity of discrete maximum principles (DMPs) for the second order elliptic problems.*

AMS subject classifications: 65N30, 65N50

Keywords: elliptic equations, maximum principles, finite element method, discrete maximum principle, monotone matrix, matrix irreducibility

Correspondence

Antti Hannukainen and Sergey Korotov
Institute of Mathematics, Helsinki University of Technology
P.O. Box 1100, FIN-02015 TKK, Finland

Tomáš Vejchodský
Institute of Mathematics, Czech Academy of Sciences
Žitná 25, CZ–115 67 Prague 1, Czech Republic

antti.hannukainen@hut.fi, sergey.korotov@hut.fi, vejchod@math.cas.cz

ISBN 978-951-22-9508-1 (print)
ISBN 978-951-22-9509-8 (PDF)
ISSN 0784-3143 (print)
ISSN 1797-5867 (PDF)

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Mathematics and Systems Analysis
P.O. Box 1100, FI-02015 TKK, Finland
email: math@tkk.fi http://math.tkk.fi/

1 Model problem and maximum principle

We consider the following test problem: Find a function u such that

$$-\operatorname{div}(\mathcal{A}\nabla u) + cu = f \quad \text{in} \quad \Omega, \tag{1}$$

$$u = g \quad \text{on} \quad \partial\Omega, \tag{2}$$

where $\Omega \subset \mathbf{R}^d$ is a bounded polytopic domain with Lipschitz boundary $\partial\Omega$. The diffusive tensor $\mathcal{A}$ is assumed to be a symmetric and uniformly positive definite matrix. The reactive coefficient c is assumed to be nonnegative in Ω.

The classical solutions of elliptic problems of the second order are known to satisfy the so-called maximum principles (MPs), see e.g. [11, 7]. For our test problem the corresponding MP is the following implication:

$$f \le 0 \qquad \Longrightarrow \qquad \max_{x\in\overline{\Omega}} u(x) \le \max\{0\,, \max_{s\in\partial\Omega} g(s)\}. \tag{3}$$

To the authors' knowledge the first reasonable DMP and conditions providing its validity were formulated in 1966 by R. Varga [14] for the finite difference method. Later, in 1970 in [4] (and [5]), Ph. Ciarlet presented a more general form of DMP suitable for finite element (FE) and finite difference types of discretizations. He also proposed a practical set of (sufficient) conditions on matrices involved, providing a validity of his DMP. Since that time these conditions became popular in numerical community, see e.g. [7, 8, 9, 10] and references therein, for proving various DMPs for problems of elliptic type. In this work we consider the issue of weakening the conditions proposed by Ciarlet.

2 Finite element discretization

Standard (linear) schemes for construction of FE approximations for the (unknown) solution u of (1)–(2) are based on the so-called weak formulation: Find $u \in g + H_0^1(\Omega)$ such that

$$a(u,v) = \mathcal{F}(v) \quad \forall v \in H_0^1(\Omega),$$

where

$$a(u,v) = \int_\Omega \mathcal{A}\nabla u \cdot \nabla v \,\mathrm{d}x + \int_\Omega cuv \,\mathrm{d}x \quad \text{and} \quad \mathcal{F}(v) = \int_\Omega fv \,\mathrm{d}x.$$

Here, the matrix $\mathcal{A}$ is assumed to be in $[L^\infty(\Omega)]^{d\times d}$, $c \in L^\infty(\Omega)$, $g \in H^1(\Omega)$, and $f \in L^2(\Omega)$. The existence and uniqueness of the weak solution u is provided by the standard Lax-Milgram lemma.

Let $\mathcal{T}_h$ be a FE partition (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 $\partial\Omega$. Further, let V_h be a finite-dimensional subspace of $H^1(\Omega)$, associated with $\mathcal{T}_h$ and its nodes, being spanned by the basis functions $\phi_1, \phi_2, \ldots, \phi_{N+N^\partial}$ with the following properties: $\phi_i \ge 0$ in $\overline{\Omega}$, $i = 1, \ldots, N + N^\partial$, and $\sum_{i=1}^{N+N^\partial} \phi_i \equiv 1$ in $\overline{\Omega}$.

We also assume that the basis functions $\phi_1, \phi_2, \ldots, \phi_N$ vanish on the boundary $\partial\Omega$. Thus, they span a finite-dimensional subspace V_h^0 of $H_0^1(\Omega)$. Let, in addition, $g_h = \sum_{i=N+1}^{N+N^\partial} g_i\phi_i \in V_h$ be a suitable approximation of the function g, for example its nodal interpolant.

The FE approximation is defined as a function $u_h \in g_h + V_h^0$ such that

$$a(u_h, v_h) = \mathcal{F}(v_h) \quad \forall v_h \in V_h^0,$$

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

Algorithmically, $u_h = \sum_{i=1}^{N+N^\partial} y_i\phi_i$, where the coefficients y_i are the entries of the solution $\bar{\mathbf{y}} = [y_1, \ldots, y_{N+N^\partial}]^\top$ of the following square system of $N+N^\partial$ linear algebraic equations

$$\bar{\mathbf{A}}\bar{\mathbf{y}} = \bar{\mathbf{F}}, \tag{4}$$

where

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{A}^\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \bar{\mathbf{F}} = \begin{bmatrix} \mathbf{F} \\ \mathbf{F}^\partial \end{bmatrix}, \quad \text{and} \quad \bar{\mathbf{A}}^{-\mathbf{1}} = \begin{bmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{A}^\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{5}$$

In the above, blocks $\mathbf{A}$ and $\mathbf{A}^\partial$ are matrices of size $N \times N$ and $N \times N^\partial$, respectively, $\mathbf{I}$ stands for the unit matrix, and $\mathbf{0}$ – for the zero matrix. The entries of $\bar{\mathbf{A}}$ are denoted by $a_{ij} = a(\phi_j, \phi_i),\ i = 1, \ldots, N,\ j = 1, \ldots, N+N^\partial$. The block-vector $\mathbf{F}$ consists of entries $f_i = \mathcal{F}(\phi_i),\ i = 1, \ldots, N$, and the block-vector $\mathbf{F}^\partial$ has entries $f_i = g_i,\ i = N+1, \ldots, N+N^\partial$, given by the boundary data. For the later reference we also include the formula for $\bar{\mathbf{A}}^{-1}$ in (5). Notice that $\bar{\mathbf{A}}$ is nonsingular if and only if $\mathbf{A}$ is nonsingular.

3 Sufficient algebraic conditions of Ph. Ciarlet

We will distinguish two essentially different types of DMPs.

Algebraic DMP: A natural algebraic analogue of (3) is as follows (cf. (4)):

$$\mathbf{F} \leq \mathbf{0} \qquad \Longrightarrow \qquad \max_{i=1,\ldots,N+N^\partial} y_i \leq \max\{0\,, \max_{j=N+1,\ldots,N+N^\partial} y_j\}.$$

Functional DMP: A natural functional imitation of (3) is as follows:

$$f \leq 0 \qquad \Longrightarrow \qquad \max_{\overline{\Omega}} u_h \leq \max\{0\,, \max_{\partial\Omega} u_h\}.$$

Remark 1. It is easy to see that the above types of DMPs are equivalent in the case of linear and multilinear finite elements. However, these DMPs are not equivalent, in general, for higher-order finite elements.

In [4], Ciarlet formulated and proved the following theorem:

Theorem 1. *The algebraic DMP is satisfied if and only if*

(A) $\bar{\mathbf{A}}$ *is monotone (i.e.,* $\bar{\mathbf{A}}$ *nonsingular and* $\bar{\mathbf{A}}^{-1} \geq 0$*)*

(B) $\xi + \mathbf{A}^{-1}\mathbf{A}^\partial\xi^\partial \geq 0$*, where* ξ *and* ξ^∂ *are vectors of all ones of sizes* N *and* N^∂*, respectively.*

Since conditions (A) and (B) are difficult to verify, Ciarlet proposed in [4] the following *standard set of sufficient conditions* which is more practical.

Theorem 2. *The algebraic DMP is valid provided the matrix* $\bar{\mathbf{A}}$ *satisfies*

(a) $a_{ii} > 0,\ i = 1, \ldots, N,$

(b) $a_{ij} \le 0,\ i \ne j,\ i = 1, \ldots, N,\ j = 1, \ldots, N + N^\partial,$

(c) $\sum\limits_{j=1}^{N+N^\partial} a_{ij} \ge 0,\ i = 1, \ldots, N,$

(d) $\mathbf{A}$ *is irreducibly diagonally dominant.*

Ciarlet essentially proposed the above conditions in order to utilize the following result of Varga [13, p. 85]:

Lemma 3. *If* $\mathbf{A} \in \mathbf{R}^{N\times N}$ *is an irreducibly diagonally dominant matrix with strictly positive diagonal and nonpositive off-diagonal entries then* $\mathbf{A}^{-1} > 0$.

Now, we can easily demonstrate the proof of Theorem 2. We follow the steps of Ciarlet [4]. Conditions (a), (b), and (d) together with Lemma 3 imply $\mathbf{A}^{-1} \ge 0$ and, hence, condition (A). Further, condition (c) is equivalent to $\mathbf{A}\xi + \mathbf{A}^\partial \xi^\partial \ge 0$ and since $\mathbf{A}^{-1} \ge 0$ we conclude that (B) is valid as well. Theorem 1 then guarantees the algebraic DMP.

Remark 2. In the case of homogeneous Dirichlet boundary conditions system (4) reduces to a simpler form $\mathbf{A}\mathbf{y} = \mathbf{F}$ (cf. [10]). Then the algebraic DMP holds if and only if $\mathbf{A}^{-1} \ge \mathbf{0}$, i.e., if and only if $\mathbf{A}$ is monotone.

4 Associated geometrical conditions on FE meshes

For some types of finite elements, the entries of $\bar{\mathbf{A}}$ can be computed explicitly, therefore condition (b) can often be guaranteed a priori by imposing suitable geometrical requirements on the shape (and size) of FE meshes employed.

For example, if $\mathcal{A}$ is a diagonal matrix then there exist the following popular geometrical conditions providing (b):

(i) for simplicial finite elements ($d \ge 2$) – all dihedral angles between facets of simplices have to be nonobtuse or acute [5, 2, 7, 8, 10];

(ii) for bilinear elements – all rectangular elements have to be nonnarrow ($\sqrt{2}/2 \le b_1/b_2 \le \sqrt{2}$, where b_1, b_2 are the edges of the rectangle), trilinear elements have to be cubes, see [8];

(iii) for 3D meshes consiting of right triangular prisms the altitudes of prisms are limited from both sides by certain quantities dependent on the area and angles of the triangular base (and the magnitude of the reactive coefficient c) [6].

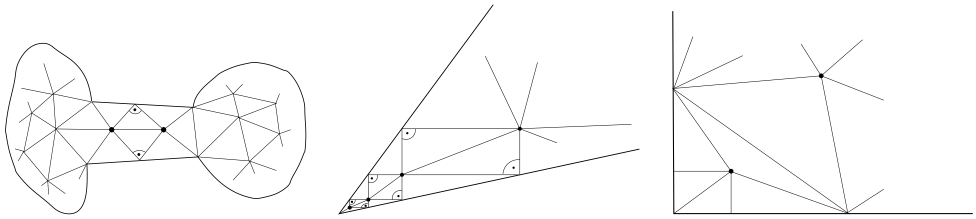

Figure 1: Examples of meshes leading to reducible matrix $\mathbf{A}$ for the Poisson problem with Dirichlet boundary conditions. The angles with dots are right.

5 Typical problems with standard conditions

Not only condition (b) but also the other conditions (a), (c), and (d) have to be addressed. The positivity of diagonal entries (a) is trivially satisfied for elliptic problems. Also the row sums (c) are nonnegative automatically for problem (1)–(2), because the basis functions form a partition of unity:

$$\sum_{j=1}^{N+N^\partial} a_{ij} = a\left(\sum_{j=1}^{N+N^\partial} \phi_j, \phi_i\right) = a(1,\phi_i) = \int_\Omega c\phi_i \geq 0, \quad i = 1,\dots,N. \tag{6}$$

On the other hand, the irreducibility of $\mathbf{A}$ required in (d) is not always obvious. For illustration, we present three examples of triangulations which lead to reducible matrices in Figure 1.

It might be a difficult task to satisfy all conditions (a)–(d) practically, especially in 3D. For example, the existence of a face-to-face partition of a cube into acute tetrahedra is still an open problem, see (i). Another practical problem in 3D is to keep the desired geometrical limitations on the elements during global and local refinements of meshes. Condition (b) leads to severe limitation in the case of 3D rectangular blocks (only cubes are allowed), see the point (ii) above. Moreover, if the diffusive tensor $\mathcal{A}$ is not diagonal then proving the irreducibility (d) could be a nontrivial task.

6 Less severe conditions: Stieltjes matrices

Conditions (a)–(d) can be weakened using the concept of M-matrices and Stieltjes matrices [13]. A real square matrix $\mathbf{A}$ is an *M-matrix* if all its off-diagonal entries are nonpositive and if it is nonsingular and $\mathbf{A}^{-1} \geq 0$. A real square matrix $\mathbf{A}$ is a *Stieltjes matrix* if all its off-diagonal entries are nonpositive and if it is symmetric and positive definite. The following lemma [13, p. 85] enables to eliminate conditions (a), (c), and (d) from the standard set as we state in Theorem 5 below.

Lemma 4. *If* $\mathbf{A}$ *is a Stieltjes matrix then it is also an M-matrix.*

Theorem 5. *If the finite element matrix* $\bar{\mathbf{A}}$*, associated to* (1)–(2)*, satisfies condition* (b) *from Theorem 2 then the algebraic DMP is valid.*

Proof. We verify conditions (A) and (B) of Theorem 1. (A) For problem (1)–(2), the FE matrix $\mathbf{A}$ is always symmetric and positive definite and hence if condition (b) is satisfied then $\mathbf{A}^{-1} \geq 0$ by Lemma 4. Further, since $\mathbf{A}^{\partial} \leq 0$ by (b), we obtain $-\mathbf{A}^{-1}\mathbf{A}^{\partial} \geq 0$ and therefore $\bar{\mathbf{A}}^{-1} \geq 0$, see (5). (B) Condition (c) is satisfied for problem (1)–(2) due to (6) and, as we already mentioned, (c) implies (B). $\square$

7 Testing the sharpness of theoretical conditions

The standard DMP results [2, 5, 7, 8] provide conditions which guarantee condition (b) and consequently that $\mathbf{A}$ is a Stieltjes matrix. However, the Stieltjes matrices form only a certain subclass of monotone matrices. In this section we test how sharp the conditions based on the Stieltjes matrix concept are, i.e., we try to test how wide is the class of meshes which lead to $\mathbf{A}$ being monotone but not Stieltjes. For this purpose we solve the 2D Poisson problem with homogenous Dirichlet boundary conditions on various domains using various triangulations.

We present three tests. In each test we construct a simple triangulation which is characterized by two parameters (angles) $0 < \alpha < \pi$ and $0 < \gamma < \pi$, see Figure 2. We prepare N sampling points and we go through all values $\alpha_i = i\pi/(N+1)$, $\gamma_j = j\pi/(N+1)$, $i, j = 1, 2, \dots, N$. For each pair α_i, γ_j, we construct the basic mesh as indicated in Figure 2, provided it is possible. Then we refine each triangle in the basic mesh into 100 similar subtriangles (each edge in the original mesh is divided into 10 segments). Then we assemble the stiffness matrix $\mathbf{A}$ on the refined mesh and we compute the inverse $\mathbf{A}^{-1}$. Finally, we investigate the entries of $\mathbf{A}$ and $\mathbf{A}^{-1}$. If all off-diagonal entries of $\mathbf{A}$ are nonpositive we mark the pair α_i, γ_j by 1. Otherwise, we check the nonnegativity of $\mathbf{A}^{-1}$. We mark the pair α_i, γ_j by 2 if $\mathbf{A}^{-1} \geq 0$ and by 3 if it is not.

The results of the computations are visualized in Figure 3, where we used $N = 200$ sampling points for each of the angles α and γ. The white areas correspond to the angles α_i, γ_j for which the indicated triangulation does not exist.

The stiffness matrix $\mathbf{A}$ for the Poisson equation in 2D is well-known to be Stieltjes matrix if and only if the sum of the two angles opposite to each interior edge in the mesh is at most π. For the investigated meshes this sufficient and necessary condition reduces to the requirement of non-obtuseness of the greatest angle in the triangulation, see the point (i) above. If we compare the sizes of domains 1 with domains 2 in Figure 3 we may conclude that the standard sufficient condition (b) is not very sharp. There is a wide space for its generalization. However, any generalization have to utilize more general criteria for the monotonicity of a matrix. These criteria, see e.g. [1], are more complicated and their application for the DMP is not straightforward. Certain success in this respect was reported in [3, 9, 12].

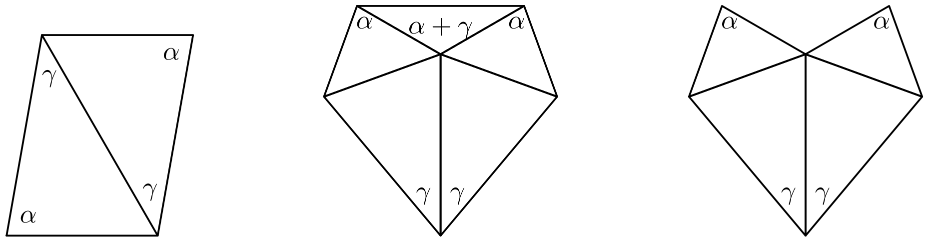

Figure 2: The basic meshes for the three tests. The meshes used for the actual computations are 10-fold refined basic meshes.

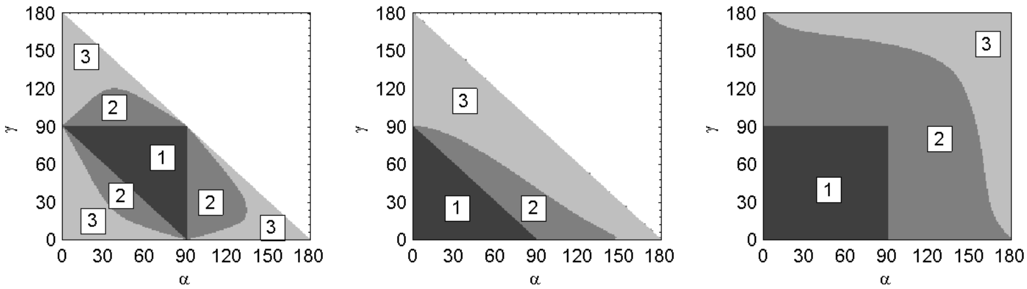

Figure 3: Results of three DMP tests. Domain 1: triangulations with non-obtuse maximal angle (matrix $\mathbf{A}$ is a Stieltjes matrix). Domain 2: triangulations with obtuse maximal angle providing the DMP (matrix $\mathbf{A}$ is monotone but not Stieltjes). Domain 3: triangulations with obtuse maximal angle, DMP is not valid (matrix $\mathbf{A}$ is not monotone).

8 Conclusions

We have analyzed the standard approach for proving the DMP for elliptic problems and showed that the positivity of the diagonal entries (a), the non-negativity of the row sums (c), and the irreducibility and diagonal dominance (d) are, in fact, not needed as sufficient conditions. Moreover, the presented numerical experiments indicate that the known geometric conditions guaranteeing $\mathbf{A}^{-1} \geq 0$ are not very sharp and that there is a space for possible generalization and further research.

References

[1] Bouchon, F., Monotonicity of some perturbations of irreducibly diagonally dominant M-matrices, *Numer. Math.* 105 (2007), 591–601.

[2] Brandts, J., Korotov, S., Křížek, M., The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, *Linear Algebra Appl.* (to appear).

[3] 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.

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

[7] 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.

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

[9] Korotov, S., Křížek, M., Neittaanmäki, P., Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, *Math. Comp.* 70 (2001), 107–119.

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

[11] Ladyzhenskaya, O.A., Ural'tseva, N.N., *Linear and quasilinear elliptic equations*, Leon Ehrenpreis Academic Press, New York-London, 1968.

[12] Ruas Santos, V., On the strong maximum principle for some piecewise linear finite element approximate problems of non-positive type, *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* 29 (1982), 473–491.

[13] Varga, R., *Matrix Iterative Analysis*, Prentice Hall, New Jersey, 1962.

[14] Varga, R., On discrete maximum principle, *J. SIAM Numer. Anal.* 3 (1966), 355–359.

(continued from the back cover)

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

The reports are available at *http://math.tkk.fi/reports/* .
The list of reports is continued inside the back cover.

A548 Kalle Mikkola
Weakly coprime factorization, continuous-time systems, and strong-H^p and Nevanlinna fractions
August 2008

A547 Wolfgang Desch, Stig-Olof Londen
A generalization of an inequality by N. V. Krylov
June 2008

A546 Olavi Nevanlinna
Resolvent and polynomial numerical hull
May 2008

A545 Ruth Kaila
The integrated volatility implied by option prices, a Bayesian approach
April 2008

A544 Stig-Olof Londen, Hana Petzeltová
Convergence of solutions of a non-local phase-field system
March 2008

ISBN 978-951-22-9508-1 (print)
ISBN 978-951-22-9509-8 (PDF)
ISSN 0784-3143 (print)
ISSN 1797-5867 (PDF)