Angle Conditions for Discrete Maximum Principles in Higher-Order FEM

Tomáš Vejchodský

Abstract This contribution reviews the general theory of the discrete Green's function and presents a numerical experiment indicating that the discrete maximum principle (DMP) fails to hold in the case of Poisson problem on any uniform triangulation of a triangular domain for orders of approximation three and higher. This extends the result [8] that the Laplace equation discretized by the higher-order FEM satisfies the DMP on a patch of triangular elements in exceptional cases only.

1 Introduction

The discrete maximum principle (DMP) is important in practice, because it guarantees nonnegativity of approximations of naturally nonnegative quantities like temperature, concentration, density, etc. Its theoretical significance lies in its connection with the uniform convergence of the finite element approximations [4]. In contrast to the lowest-order finite element method (FEM), the DMP for the higher-order FEM in dimension two and higher is not well understood, yet.

A stronger version of the DMP for the Laplace equation discretized by higher-order finite elements was studied by Höhn and Mittelman in [8]. This stronger version requires the validity of the DMP on all vertex patches (union of elements sharing a vertex) in the triangulation. They find that the quadratic elements do not satisfy the stronger DMP unless the triangulation is very special (e.g. all equilateral triangles) and that the restrictions for cubic elements are even more severe.

In the present contribution we briefly review the general theory about the discrete Green's function (DGF) and the standard DMP for the Poisson problem. Then we present a numerical experiment indicating that the standard DMP is not satisfied on any uniform triangulation for the finite elements of order three and higher.

Tomáš Vejchodský
Institute of Mathematics, Academy of Sciences, Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: vejchod@math.cas.cz

2 Model problem and its FEM discretization

First, we briefly introduce the Poisson problem and its discretization by the FEM. The main purpose of this section is to settle down the notation.

Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. The classical and the weak formulations of the Poisson problem reads as follows:

$$\text{Find } u \in C^2(\Omega) \cap C(\overline{\Omega}) \text{ such that} \quad -\Delta u = f \quad \text{in } \Omega, \text{ and} \quad u = 0 \text{ on } \partial\Omega. \tag{1}$$

$$\text{Find } u \in H^1_0(\Omega) \text{ such that} \quad a(u,v) = \mathscr{F}(v) \quad \forall v \in H^1_0(\Omega), \tag{2}$$

where $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x$ and $\mathscr{F}(v) = \int_\Omega fv \,\mathrm{d}x$. We require $f \in C(\Omega)$ for the classical formulation and $f \in L^2(\Omega)$ for the weak one.

In order to discretize problem (2) by the Galerkin method, we introduce a finite dimensional subspace V_h of $H^1_0(\Omega)$. We assume that $V_h \subset C(\overline{\Omega})$. The Galerkin solution $u_h \in V_h$ is given by the requirement

$$a(u_h, v_h) = \mathscr{F}(v_h) \quad \forall v_h \in V_h. \tag{3}$$

Considering a basis $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h, we can express $u_h = \sum_{i=1}^N z_i \varphi_i$ and verify that problem (3) is equivalent to the system $Az = F$ of linear algebraic equations, where the stiffness matrix $A \in \mathbb{R}^{N\times N}$ has entries $a_{ij} = a(\varphi_j, \varphi_i)$, the load vector $F \in \mathbb{R}^N$ has entries $F_i = \mathscr{F}(\varphi_i)$, and $z = (z_1, z_2, \ldots, z_N)^\top$.

The FEM can be seen as a special case of the Galerkin method, where the space V_h is chosen in a special way such that the stiffness matrix A is sparse. The particular choice of V_h is not important at this point and it will be specified later on.

3 Discrete maximum principle

Theorem 1 below is an equivalent formulation of the standard maximum principle due to E. Hopf [9] applied to problem (1). Similarly, Theorem 2 presents the same principle for the weak solution.

Theorem 1. *Let u be a classical solution to (1). If $f \geq 0$ in Ω then $u \geq 0$ in Ω.*

Theorem 2. *Let u be a weak solution to (2). If $f \geq 0$ a.e. in Ω then $u \geq 0$ a.e. in Ω.*

The same result for the the Galerkin solution $u_h \in V_h$ is known as the DMP. Unfortunately, it is not valid in general and various conditions for its validity are studied.

Definition 1. Let the finite dimensional space V_h be fixed. We say that discretization (3) satisfies the discrete maximum principle (DMP) if the solution $u_h \in V_h$ is nonnegative in Ω for any $f \in L^2(\Omega)$, $f \geq 0$ a.e. in Ω.

A usefull tool for investigation of the DMP especially for the higher-order FEM is the so-called discrete Green's function (DGF) which was already introduced in [2, 5]. For any $y \in \Omega$ let us define the DGF $G_{h,y} \in V_h$ as the unique function satisfying

$$a(v_h, G_{h,y}) = v_h(y) \quad \forall v_h \in V_h. \tag{4}$$

This definition together with (3) implies the representation formula

$$u_h(y) = \mathscr{F}(G_{h,y}) = \int_\Omega f(x) G_h(x,y)\,\mathrm{d}x \quad \forall y \in \Omega,$$

where we use the usual notation $G_h(x,y) = G_{h,y}(x)$. This representation formula immediately proves the following theorem.

Theorem 3. *The discretization* (3) *satisfies the DMP if and only if* $G_h(x,y) \geq 0$ *for all* $(x,y) \in \Omega^2$.

Interestingly, the DGF G_h can be expressed in terms of a basis of V_h [12]:

$$G_h(x,y) = \sum_{i=1}^{N} \sum_{j=1}^{N} (A^{-1})_{ij} \varphi_i(x) \varphi_j(y) \quad \forall (x,y) \in \Omega^2, \tag{5}$$

where $(A^{-1})_{ij}$ stand for entries of the inverse of the stiffness matrix A. Let us remark that a special case of this formula, where the basis is formed by the eigenvectors of the discrete Laplacian was already presented in [2]. Further, we remark that the concept of the DGF is relevant even for more general problems. However, in the case of nonhomogeneous Dirichlet boundary conditions the boundary Green's function has to be introduced [3]. General formula (5) is used below to analyze the nonnegativity of the DGF and consequently the validity of the DMP.

4 Nonnegativity of the DGF for the lowest-order FEM

The analysis of nonnegativity of expression (5) simplifies if the basis functions $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h have the following property

$$\sum_{i=1}^{N} z_i \varphi_i \geq 0 \quad \text{in } \Omega \quad \Leftrightarrow \quad z_i \geq 0 \quad \forall i = 1,2,\ldots,N. \tag{6}$$

This property is typically satisfied for the lowest-order finite elements such as linear functions on simplices and multilinear functions on blocks (Cartesian products of intervals). Before we state the following well-known theorem, we recall that a square matrix A is monotone if it is nonsingular and $A^{-1} \geq 0$ (i.e. all entries of A^{-1} are nonnegative).

Theorem 4. *Let the basis functions* $\varphi_1, \varphi_2, \ldots, \varphi_N$ *of* V_h *have property* (6). *Then the discretization* (3) *satisfies the DMP if and only if the stiffness matrix* A *is monotone.*

Proof. It follows immediately from assumption (6), formula (5), and Theorem 3.

If the off-diagonal entries of the stiffness matrix A are nonpositive then A is M-matrix and, hence, monotone. The nonpositivity of the off-diagonal entries can

be guaranteed by various geometric conditions on finite element meshes like the nonobtuseness condition for simplicial meshes [1] or the nonnarrowness condition for rectangular finite elements [6]. However, these conditions could be too restrictive, because it suffices to have the stiffness matrix monotone and not M-matrix. An experiment indicating how much the nonobtuseness condition for triangles can be weaken is described in Section 6 and its results are presented in Fig. 2 (top-left).

5 Nonnegativity of the DGF for the higher-order FEM

Let us investigate the case of the higher-order FEM in more details. For simplicity let us consider two dimensional Poisson problem (1) in a polygonal domain Ω. We define the finite element space as $V_h = \{v \in H_0^1(\Omega) : v|_K \in \mathbb{P}^p(K) \quad \forall K \in \mathscr{T}_h\}$, where $\mathscr{T}_h$ is a face-to-face triangulation of Ω and $\mathbb{P}^p(K)$ stands for the space of polynomials of degree at most p on the triangle K.

The standard basis of V_h consists of N^{V} vertex (piecewise linear) functions $\varphi_1, \varphi_2, \ldots, \varphi_{N^{\mathrm{V}}}$ and of $N - N^{\mathrm{V}}$ higher-order basis functions $\varphi_{N^{\mathrm{V}}+1}, \varphi_{N^{\mathrm{V}}+2}, \ldots, \varphi_N$, see e.g. [11]. The vertex functions are the usual piecewise linear "hat" functions. Thus, if B_j, $j = 1,2,\ldots,N^{\mathrm{V}}$, denote the interior vertices of the triangulation $\mathscr{T}_h$ then the vertex functions satisfy $\varphi_i(B_j) = \delta_{ij}$, $i, j = 1,2,\ldots,N^{\mathrm{V}}$.

The vertex and the higher-order (non-vertex) basis functions yield a natural 2×2 block structure of the stiffness matrix and its inverse

$$A = \begin{pmatrix} A^{\mathrm{VV}} & A^{\mathrm{VN}} \\ A^{\mathrm{NV}} & A^{\mathrm{NN}} \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} S^{-1} & -(A^{\mathrm{VV}})^{-1}A^{\mathrm{VN}}R^{-1} \\ -(A^{\mathrm{NN}})^{-1}A^{\mathrm{NV}}S^{-1} & R^{-1} \end{pmatrix},$$

where $A^{\mathrm{VV}} \in \mathbb{R}^{N^{\mathrm{V}} \times N^{\mathrm{V}}}$, $A^{\mathrm{NN}} \in \mathbb{R}^{(N-N^{\mathrm{V}}) \times (N-N^{\mathrm{V}})}$, etc., $S = A^{\mathrm{VV}} - A^{\mathrm{VN}}(A^{\mathrm{NN}})^{-1}A^{\mathrm{NV}}$, and $R = A^{\mathrm{NN}} - A^{\mathrm{NV}}(A^{\mathrm{VV}})^{-1}A^{\mathrm{VN}}$.

The Schur complement S has the following interesting property. Let B_i and B_j, $i, j = 1,2,\ldots,N^{\mathrm{V}}$, be two interior vertices of the triangulation $\mathscr{T}_h$. Since $\varphi_i(B_j) = \delta_{ij}$ and due to (5) we obtain

$$G_h(B_i, B_j) = (A^{-1})_{ij}\varphi_i(B_i)\varphi_j(B_j) = (A^{-1})_{ij} = (S^{-1})_{ij}. \tag{7}$$

Hence, the values of the DGF at the vertices of $\mathscr{T}_h$ coincide with the entries of S^{-1}. Furthermore, the DGF has a natural structure given by the Cartesian product of the mesh $\mathscr{T}_h$ with itself. In particular, if K and L are two elements from $\mathscr{T}_h$ and ι_K and ι_L denote the sets of indices of basis functions supported in K and L, respectively, i.e., $\iota_K = \{i : \mathrm{meas}(K \cap \mathrm{supp}\,\varphi_i) > 0\}$, then the DGF restricted to $K \times L$ is given by

$$G_h|_{K\times L}(x,y) = \sum_{i\in\iota_K}\sum_{j\in\iota_L}(A^{-1})_{ij}\varphi_i|_K(x)\varphi_j|_L(y), \quad (x,y) \in K \times L. \tag{8}$$

This formula contains a small number of basis functions and we use it for fast evaluation of the DGF at a given point.

6 Numerical experiment

In this experiment we test nonnegativity of the DGF on uniform meshes. We consider Poisson problem (1) on a triangle Ω. The finite element mesh is constructed by three successive uniform (red) refinements of Ω, see Fig. 1 (left).

To speed up the test of the nonnegativity of the DGF, we first check the values at vertices, using the Schur complement S, see (7). If S is monotone, it remains to verify the nonnegativity at the other points. We proceed by inspection of all pairs of elements $K,L \in \mathscr{T}_h$ using formula (8). Function $G_h|_{K\times L}$ is a polynomial. The test of nonnegativity of a multivariate polynomial is a complicated task (connected with the 17th Hilbert's problem [10]). Therefore, we sample the values of $G_h|_{K\times L}$ in a number of points $(x^K_{k\ell}, x^L_{mn}) \in K \times L$, where the sample point $x^K_{k\ell}$ has barycentric coordinates $(k,\ell,M-k-\ell)/M$, $0 \le k+\ell \le M$, see Fig. 1 (right). The total number of sample points in an element is $(M+1)(M+2)/2$. To ensure that the number of sample points is sufficient, we always perform a series of computations starting with $M = 8$ and doubling M until the results do not change.

Fig. 2 presents the results. Each point in a panel corresponds to a pair of angles α and β, which represent the vertex angles of the triangle Ω. The color of this point is given by the properties of the DGF. If the DGF is nonnegative at all vertices and at all sample points then the color is black. This is the only case when the DMP is hopefully satisfied. If the DGF is not nonnegative then we distinguish three more cases. (i) The DGF is negative in a sample point and S is M-matrix (dark gray). (ii) The DGF is negative in a sample point and S is monotone but not M-matrix (lighter gray). (iii) The DGF is negative in a vertex, i.e., S is nonmonotone (lightest gray).

The above description, however, applies for higher-order elements only ($p \ge 2$). The case of linear elements ($p = 1$) is exceptional, because just the vertex values of the DGF are relevant for its nonnegativity. Due to Theorem 4, we distinguish in the top-left panel of Fig. 2 the cases (a) A is nonmonotone, (b) A is monotone but not M-matrix, (c) A is M-matrix. Notice that the DMP is satisfied in cases (b) and (c).

Clear conclusion from Fig. 2 is that the DGF has negative values for all tested pairs of angles for orders $p \ge 3$. However, if we look on vertex values of the DGF only, we observe that the area of this region increases with p. The increase is not monotone but in principle the higher polynomial degree p we use the wider range of angles can be used in order to keep the vertex values of the DGF nonnegative.

The only polynomial degrees allowing the DMP on uniform meshes are $p = 1$ and $p = 2$. For the case $p = 1$ (see Section 4 above) the black area in the top-left panel of Fig. 2 clearly shows that the stiffness matrix A is M-matrix provided the maximal angle is at most 90°. In addition, we observe that the stiffness matrix can be monotone even if the maximal angle is about 117°. In the case $p = 2$ the DMP is satisfied only if all the angles are close to 60°. We also check the nonnegativity of the DGF for meshes finer than the mesh sketched in Fig. 1 (left). The results on meshes one and two times refined are exactly the same as those presented in Fig. 2.

It might be of further interest to see how the DGF really looks like. For illustration we choose $p = 3$ and $\alpha = \beta = 60^\circ$. For these values the DGF is nonnegative in the vertices and negative somewhere in between. The graph of the function $G_h(x,y)$,

$(x,y) \in \Omega^2$, is difficult to visualize, because it is a five dimensional object. However, each pair of elements $K_i \in \mathscr{T}_h$ and $K_j \in \mathscr{T}_h$ corresponds to a point in a plane and the color of this point can be chosen according to some characteristic of the DGF restricted to the polytope $K_i \times K_j$. The left panel of Fig. 3 presents the mean values of G_h over $K_i \times K_j$. The right panel illustrates the negative part of the minimum of G_h in $K_i \times K_j$, i.e., $(\min_{K_i \times K_j} G_h)^-$, where $\chi^- = (|\chi| - \chi)/2$. Both these quantities are approximated using the sample points as described above. The used triangulation together with indices of elements is shown in Fig. 1 (left). Notice that the elements with indices 1–39 are adjacent to the boundary of Ω while the elements 40–64 are interior. The right panel of Fig. 3 clearly shows that the DGF is negative in polytopes $K_i \times K_j$, where K_i and K_j are both adjacent to the boundary and they are neighbors to each other including the case $K_i = K_j$. Another choice of angles α and β leads, however, to the negativity of the DGF for more pairs K_i, K_j.

7 Conclusions

We discussed the nonnegativity of the DGF and equivalently the validity of the DMP for Galerkin solutions of Poisson problem (1) with homogeneous Dirichlet boundary conditions. Results of the performed experiment indicate that the DGF is not nonnegative on uniform meshes for all shapes of triangular elements for the order three and higher. The quadratic elements yield nonnegative DGF for triangles close to equilateral ones.

The results also indicate that the DGF is negative in the areas close to the boundary. In accordance with [7] we could speculate that the nonnegativity of the DGF is not primarily determined by the angles in the triangulation but by the way how the boundary is resolved. In addition, the domain, where the DGF is negative, is relatively small with respect to the entire Ω^2 and it lies close to the boundary. This means that a nonnegative f corrupting the DMP (Definition 1) must have great values in an element close to the boundary and small values in the interior of Ω (like an approximation of the Dirac delta function). Such data are rare in practice, however. This leads us to another generalization of the (continuous) maximum principle from Theorem 2. If $f \geq 0$ is given, we may ask how must the mesh look like in order to obtain the nonnegative finite element solution. Up to the author's knowledge, this question was not considered in the literature, yet.

A possible remedy of the failure of the DMP for higher-order elements could be a modification of the higher-order basis functions based on the exact eigenfunctions of the Laplacian. This approach was successfully applied in [5] for 1D elliptic problems. A generalization to higher dimension is still an unsolved problem.

Acknowledgements The author acknowledges the support of the Czech Science Foundation, Grant no. 102/07/0496, and of the Czech Academy of Sciences, Grant no. IAA100760702, and Institutional Research Plan no. AV0Z10190503.

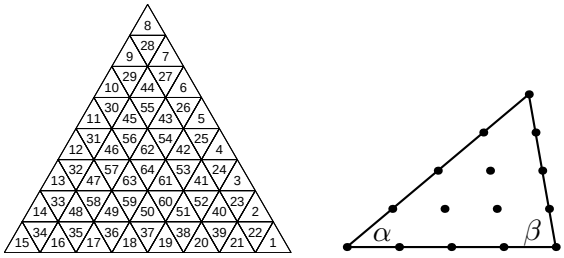

Fig. 1 A uniform mesh with 64 triangles enumerated in a spiral way (left). A triangular element characterized by a pair of of angles α and β with sample points for $M = 4$ (right).

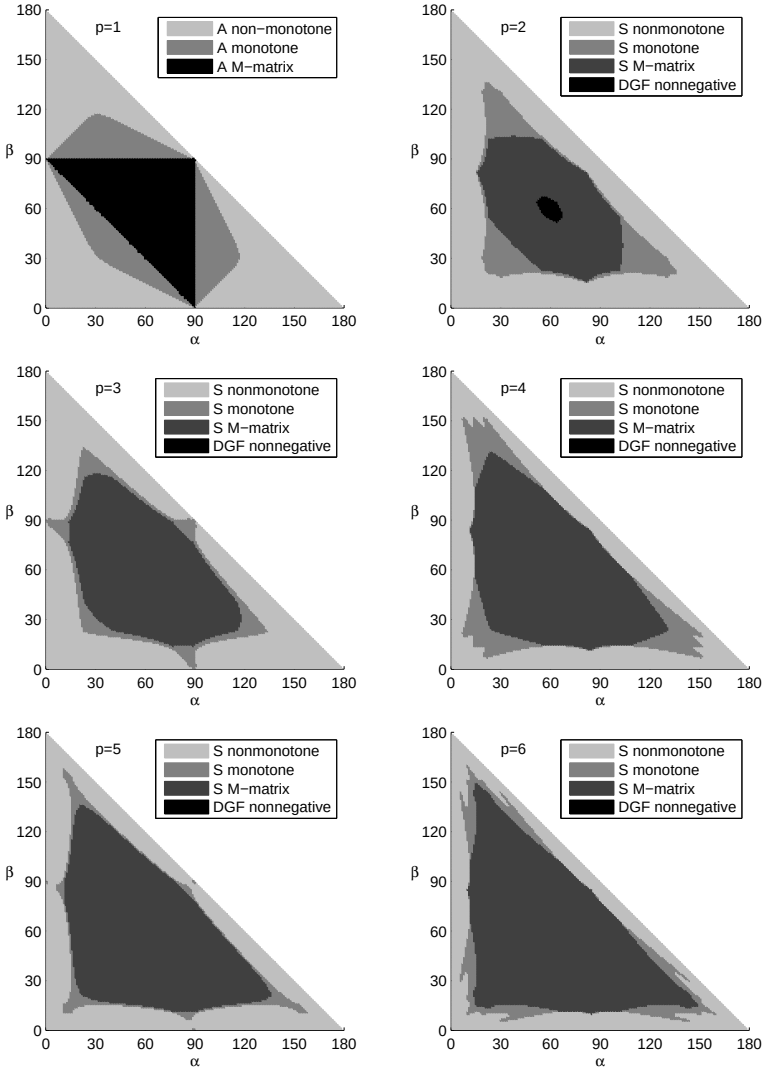

Fig. 2 The nonnegativity of the DGF and its dependence on the angles in the triangulation for orders $p = 1, 2, \ldots, 6$.

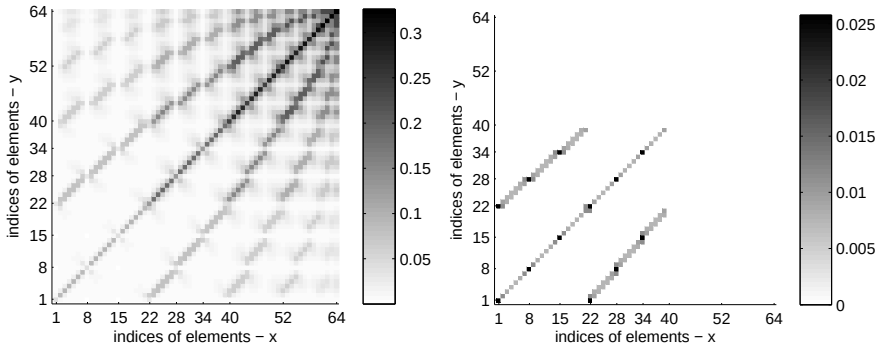

Fig. 3 A visualization of the entire DGF. A point with coordinates i, j corresponds to a pair of elements K_i, K_j. The color of this point represents the mean value (left) and the negative part of the minimum (right) of G_h in $K_i \times K_j$.

References

1. Brandts, J., Korotov, S., Křížek, M.: Dissection of the path-simplex in $\mathbf{R}^n$ into n path-subsimplices. Linear Algebra Appl. **421**, 382–393 (2007)
2. Ciarlet, P.G.: Discrete variational Green's function. I. Aequationes Math. **4**, 74–82 (1970)
3. Ciarlet, P.G.: Discrete maximum principle for finite-difference operators. Aequationes Math. **4**, 338–352 (1970)
4. Ciarlet, P.G., Raviart, P.A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Engrg. **2**, 17–31 (1973)
5. Ciarlet, P.G., Varga, R.S.: Discrete variational Green's function. II. One dimensional problem. Numer. Math. **16**, 115–128 (1970)
6. Christie, I., Hall, C.: The maximum principle for bilinear elements. Internat. J. Numer. Methods Engrg. **20**, 549–553 (1984)
7. Drăgănescu, A., Dupont, T.F., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comp. **74**, 1–23 (2005)
8. Höhn, W., Mittelmann, H.-D.: Some remarks on the discrete maximum-principle for finite elements of higher order. Computing **27**, 145–154 (1981)
9. Hopf, E.: Elementäre Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus. Sitzungsberichte Preussische Akademie der Wissenschaften, Berlin, 147–152 (1927)
10. Prestel, A., Delzell, C. N.: Positive polynomials: From Hilbert's 17th problem to real algebra. Springer-Verlag, Berlin (2001)
11. Šolín, P., Segeth, K., Doležel, I.: Higher-order finite element methods. Chapman & Hall/CRC, Boca Raton, FL (2004)
12. Vejchodský, T., Šolín, P.: Discrete maximum principle for higher-order finite elements in 1D. Math. Comp. **76**, 1833–1846 (2007)
13. Vejchodský, T., Šolín, P.: Discrete maximum principle for a 1D problem with piecewise-constant coefficients solved by *hp*-FEM. J. Numer. Math. **15**, 233–243 (2007)
14. Vejchodský, T., Šolín, P.: Discrete maximum principle for Poisson equation with mixed boundary conditions solved by *hp*-FEM. Adv. Appl. Math. Mech. **1**, 201–214 (2009)