# Computational Investigations of the Discrete Maximum Principles

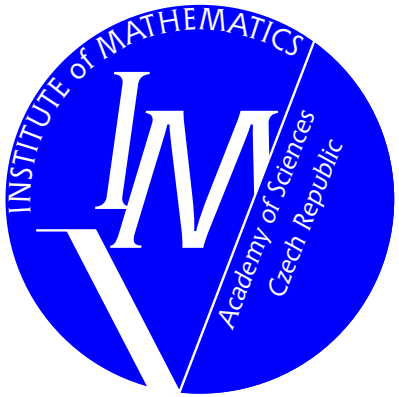
Tomáš Vejchodský[1] (vejchod@math.cas.cz)

Antti Hannukainen[2] (antti.hannukainen@hut.fi)

Sergey Korotov[2] (sergey.korotov@hut.fi)

[1] Institute of Mathematics, Academy of Sciences
Žitná 25, 115 67 Prague 1
Czech Republic

[2] Institute of Mathematics
Helsinki University of Technology
P.O. Box 1100, FIN–02015 Espoo
Finland




# (Continuous) maximum principle

$$-\operatorname{div}(\mathcal{A}\nabla u) + cu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega$$

$$\text{MaxP}: \qquad f \le 0 \quad \Rightarrow \quad \max_{\overline{\Omega}} u \le \max\{0, \max_{\partial\Omega} u\}$$

$$\text{MinP}: \qquad f \ge 0 \quad \Rightarrow \quad \min_{\overline{\Omega}} u \ge \min\{0, \min_{\partial\Omega} u\}$$

$$\text{ComP}: \quad f \ge 0 \ \&\ g \ge 0 \quad \Rightarrow \quad u \ge 0$$

$$\text{MaxP} \quad \Leftrightarrow \quad \text{MinP} \quad \Leftrightarrow \quad \text{ComP}$$

# Weak formulation and FEM

$$-\operatorname{div}(\mathcal{A}\nabla u) + cu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega$$

- Weak formulation: $\bar{u} = u + g$

$$u \in V: \quad a(u, v) = (f, v) - a(g, v) \quad \forall v \in V,$$

where $V = H_0^1(\Omega), \quad g \in H^1(\Omega)$

$$a(u, v) = \int_\Omega \left(\nabla u \cdot \nabla v + cuv\right) \mathrm{d}x, \qquad (f, v) = \int_\Omega fv \,\mathrm{d}x$$

- FEM: $\bar{u}_h = u_h + g_h, \quad g_h \approx g, \quad g \geq 0 \;\Leftrightarrow\; g_h \geq 0$

$$u_h \in V_h: \quad a(u_h, v_h) = (f, v_h) - a(g_h, v_h) \quad \forall v_h \in V_h,$$

where $V_h \subset V$ (continuous, piecewise linear on a mesh $\mathcal{T}_h$)

# Discrete maximum principle

$$\bar{u}_h = u_h + g_h,\ u_h \in V_h: \quad a(u_h, v_h) = (f, v_h) - a(g_h, v_h) \quad \forall v_h \in V_h$$

$$\text{DMaxP}: \qquad f \le 0 \quad \Rightarrow \quad \max_{\overline{\Omega}} \bar{u}_h \le \max\{0, \max_{\partial\Omega} \bar{u}_h\}$$

$$\text{DMinP}: \qquad f \ge 0 \quad \Rightarrow \quad \min_{\overline{\Omega}} \bar{u}_h \ge \min\{0, \min_{\partial\Omega} \bar{u}_h\}$$

$$\text{DComP}: \quad f \ge 0 \ \&\ g \ge 0 \quad \Rightarrow \quad \bar{u}_h \ge 0$$

$$\text{DMaxP} \quad \Leftrightarrow \quad \text{DMinP} \quad \Leftrightarrow \quad \text{DComP}$$

## Definition (DMP)

$\mathcal{T}_h$ fixed, $\quad f \ge 0 \ \&\ g \ge 0 \quad \Rightarrow \quad \bar{u}_h \ge 0$ (everywhere in $\Omega$)

# Proof

$$\text{DMaxP} \quad \Leftrightarrow \quad \text{DMinP} \quad \Leftrightarrow \quad \text{DComP}$$

DMaxP $\Leftrightarrow$ DMinP:

$f \mapsto \bar{u}_h, \qquad -f \mapsto -\bar{u}_h, \qquad \min\limits_{\overline{\Omega}} \bar{u}_h = -\max\limits_{\overline{\Omega}} -\bar{u}_h$

# Proof

$$\text{DMaxP} \quad \Leftrightarrow \quad \text{DMinP} \quad \Leftrightarrow \quad \text{DComP}$$

DMinP $\Rightarrow$ DComP:

$f \geq 0,\ g \geq 0, \qquad f \mapsto \bar{u}_h$

$$\min_{\overline{\Omega}} \bar{u}_h \geq \min\{0, \min_{\partial\Omega} \bar{u}_h\} = \min\{0, \min_{\partial\Omega} g_h\} = 0$$

# Proof

$$\text{DMaxP} \quad \Leftrightarrow \quad \text{DMinP} \quad \Leftrightarrow \quad \text{DComP}$$

$\text{DComP} \quad \Rightarrow \quad \text{DMinP}$:

$f \geq 0, \quad f \mapsto \bar{u}_h, \quad g_h = u_h|_{\partial\Omega}, \quad \bar{g}_h = \min\limits_{\partial\Omega} \bar{u}_h$

a) $\bar{g}_h \geq 0 \quad \Rightarrow \quad \bar{u}_h \geq 0$

$$\min_{\overline{\Omega}} \bar{u}_h \geq 0 = \min\{0, \min_{\partial\Omega} \bar{u}_h\}$$

b) $\bar{g}_h < 0,$

$\widetilde{f} = c\bar{g}_h \leq 0, \quad \widetilde{g} = \bar{g}_h = \text{const.} \quad \mapsto \quad \widetilde{u}_h = \bar{g}_h$

$f \geq 0 \geq \widetilde{f}, \quad g_h \geq \widetilde{g} \quad \Rightarrow \quad \bar{u}_h \geq \widetilde{u}_h$

$$\min_{\overline{\Omega}} \bar{u}_h \geq \min_{\overline{\Omega}} \widetilde{u}_h = \bar{g}_h = \min_{\partial\Omega} \bar{u}_h = \min\{0, \min_{\partial\Omega} \bar{u}_h\}$$

# Algebraic system

$\mathcal{T}_h$ ... $N$ interior nodes, $N^\partial$ boundary nodes, $\bar{N} = N + N^\partial$

$$\bar{u}_h(x) = \sum_{j=1}^{\bar{N}} \bar{y}_j \phi_j(x) = \sum_{j=1}^{N} y_j \phi_j(x) + \sum_{k=1}^{N^\partial} g_k \phi_{N+k}(x)$$

$$\bar{\mathbf{A}}\bar{\mathbf{y}} = \bar{\mathbf{F}} \quad \Leftrightarrow \quad \begin{bmatrix} \mathbf{A} & \mathbf{A}^\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{g} \end{bmatrix}$$

- $A_{ij} = a(\phi_j, \phi_i) \quad F_i = (f, \phi_i) \quad i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, \bar{N}.$
- $\phi_j \geq 0 \qquad \sum_{j=1}^{\bar{N}} \phi_j \equiv 1 \qquad \bar{\mathbf{A}}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{A}^\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$

# Discrete maximum principle

## Definition (DMP)

$\mathcal{T}_h$ fixed, $\quad f \geq 0 \;\&\; g \geq 0 \quad \Rightarrow \quad \bar{u}_h \geq 0 \quad$ (everywhere in $\Omega$)

## Theorem (Ciarlet 1970)

$$DMP \quad \Leftrightarrow \quad (A1)\ \bar{\mathbf{A}}^{-1} \geq 0 \ (i.e.\ \bar{\mathbf{A}}\ monotone)$$
$$(G1)\ \xi + \mathbf{A}^{-1}\mathbf{A}^{\partial}\xi^{\partial} \geq 0,$$
$$where\ \xi = [\underbrace{1, \ldots, 1}_{N\ times}]^{\top}\ and\ \xi^{\partial} = [\underbrace{1, \ldots, 1}_{N^{\partial}\ times}]^{\top}$$

## Remark ($g = 0$)

$$f \geq 0 \ \Rightarrow\ \mathbf{F} \geq 0 \ \Leftrightarrow\ \underbrace{\mathbf{A}\mathbf{y} \geq 0 \ \Rightarrow\ \mathbf{y} \geq 0}_{\Leftrightarrow\ \mathbf{A}\ \text{monotone}} \ \Leftrightarrow\ u_h \geq 0 \text{ in } \Omega$$
$$\Leftrightarrow\ \mathbf{A}^{-1} \geq 0$$

# Monotone matrices

Let $\mathbf{A} \in \mathbb{R}^{N \times N}$.

- $\mathbf{A}$ monotone if $\mathbf{Ay} \geq 0 \quad \Rightarrow \quad \mathbf{y} \geq 0$.
- $\mathbf{A}$ M-matrix if off-diag$(\mathbf{A}) \leq 0$, $\mathbf{A}$ nonsingular, and $\mathbf{A}^{-1} \geq 0$.
- $\mathbf{A}$ Stieltjes if off-diag$(\mathbf{A}) \leq 0$ and $\mathbf{A}$ s.p.d.

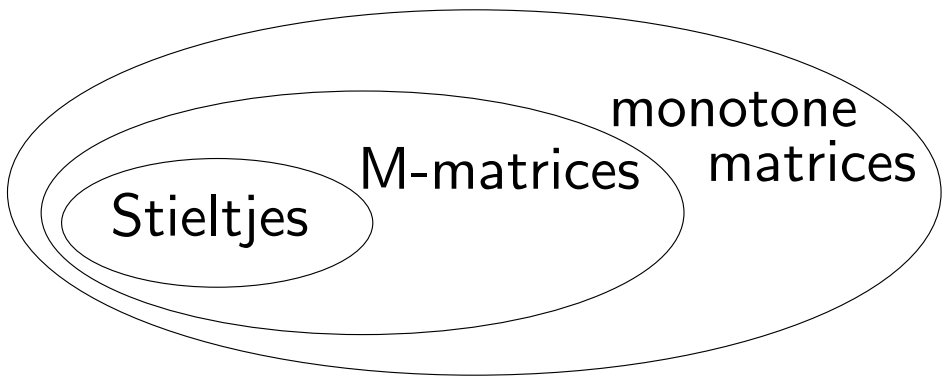
## Theorem

*Let* $\mathbf{A}$ *be nonsingular.*

$\mathbf{A}$ *monotone* $\quad \Leftrightarrow \quad \mathbf{A}^{-1} \geq 0$.

## Theorem (Varga 1962)

$\mathbf{A}$ *Stieltjes* $\quad \Rightarrow \quad \mathbf{A}$ *M-matrix.*



# Ph. Ciarlet conditions

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

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

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

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

## Theorem (Ciarlet 1970)

*(a)–(d)* $\Rightarrow$ *DMP*

## Proof.

(A1): Thm. (Varga): (a),(b),(d) $\Rightarrow$ $\mathbf{A}^{-1} > 0$

(b) $\Rightarrow$ $-\mathbf{A}^{-1}\mathbf{A}^{\partial} = \mathbf{A}^{-1}(-\mathbf{A}^{\partial}) \geq 0$ $\Rightarrow$ $\bar{\mathbf{A}}^{-1} \geq 0$

(G1): (c) $\Leftrightarrow$ $\mathbf{A}\xi + \mathbf{A}^{\partial}\xi^{\partial} \geq 0$ $\Rightarrow$ $\xi + \mathbf{A}^{-1}\mathbf{A}^{\partial}\xi^{\partial} \geq 0$ $\square$

# Reduced conditions

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

**Theorem**

*(b) $\Rightarrow$ DMP*

**Proof.**

(A1): **A** s.p.d. and (b) $\overset{def}{\Leftrightarrow}$ Stieltjes matrix $\Rightarrow$ M-matrix $\Rightarrow \quad \mathbf{A}^{-1} \ge 0$

$$\text{(b)} \quad \Rightarrow \quad -\mathbf{A}^{-1}\mathbf{A}^{\partial} = \mathbf{A}^{-1}(-\mathbf{A}^{\partial}) \ge 0 \quad \Rightarrow \quad \bar{\mathbf{A}}^{-1} \ge 0$$

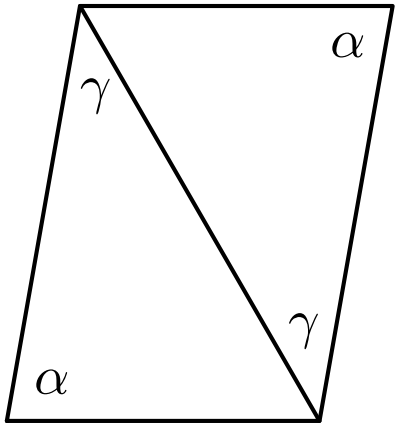
$$\text{(G1):} \quad \sum_{j=1}^{N+N^{\partial}} a_{ij} = a\left(\sum_{j=1}^{N+N^{\partial}} \phi_j, \phi_i\right) = a(1, \phi_i) = \int_{\Omega} c\phi_i \ge 0$$

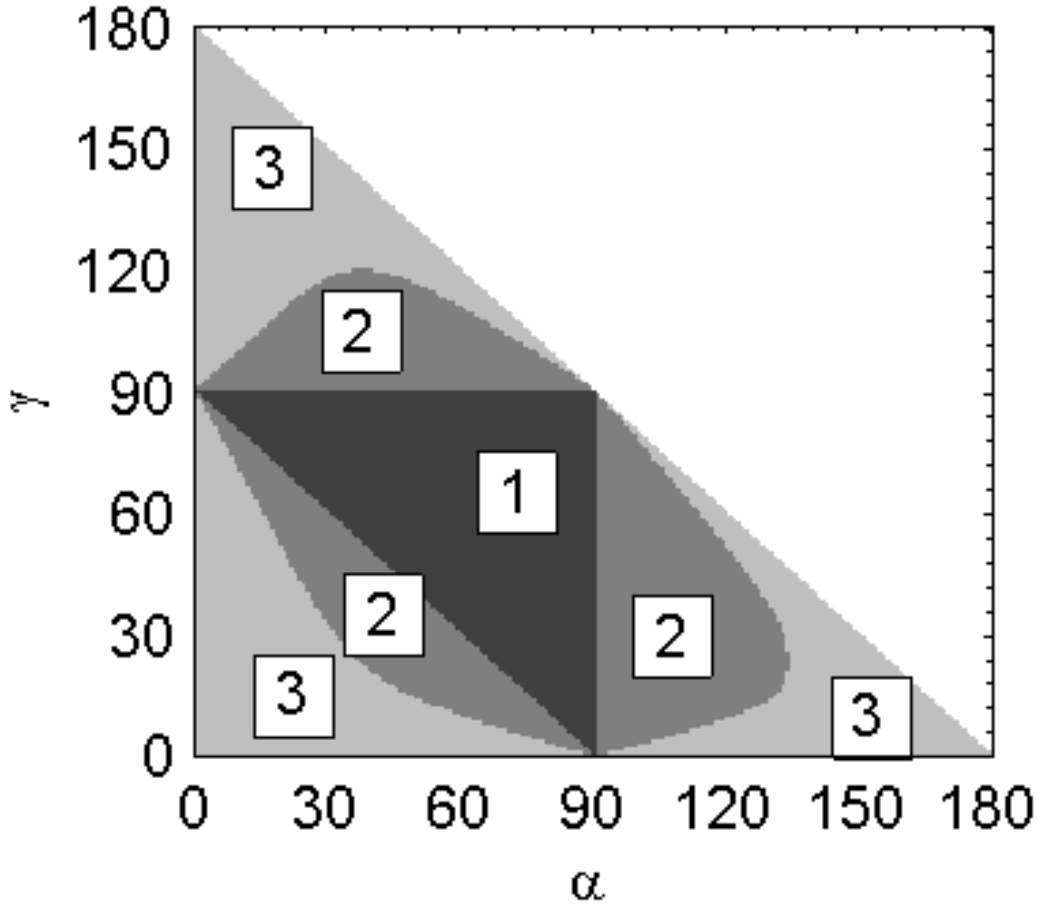
$$\Rightarrow \quad \text{(c)} \quad \Leftrightarrow \quad \mathbf{A}\xi + \mathbf{A}^{\partial}\xi^{\partial} \ge 0 \quad \Rightarrow \quad \xi + \mathbf{A}^{-1}\mathbf{A}^{\partial}\xi^{\partial} \ge 0$$

□

# Experiment I

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$

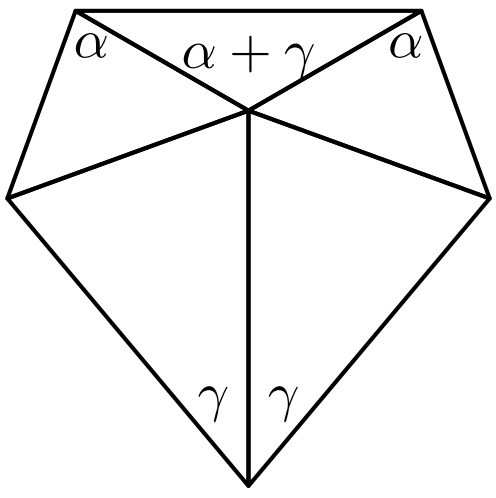


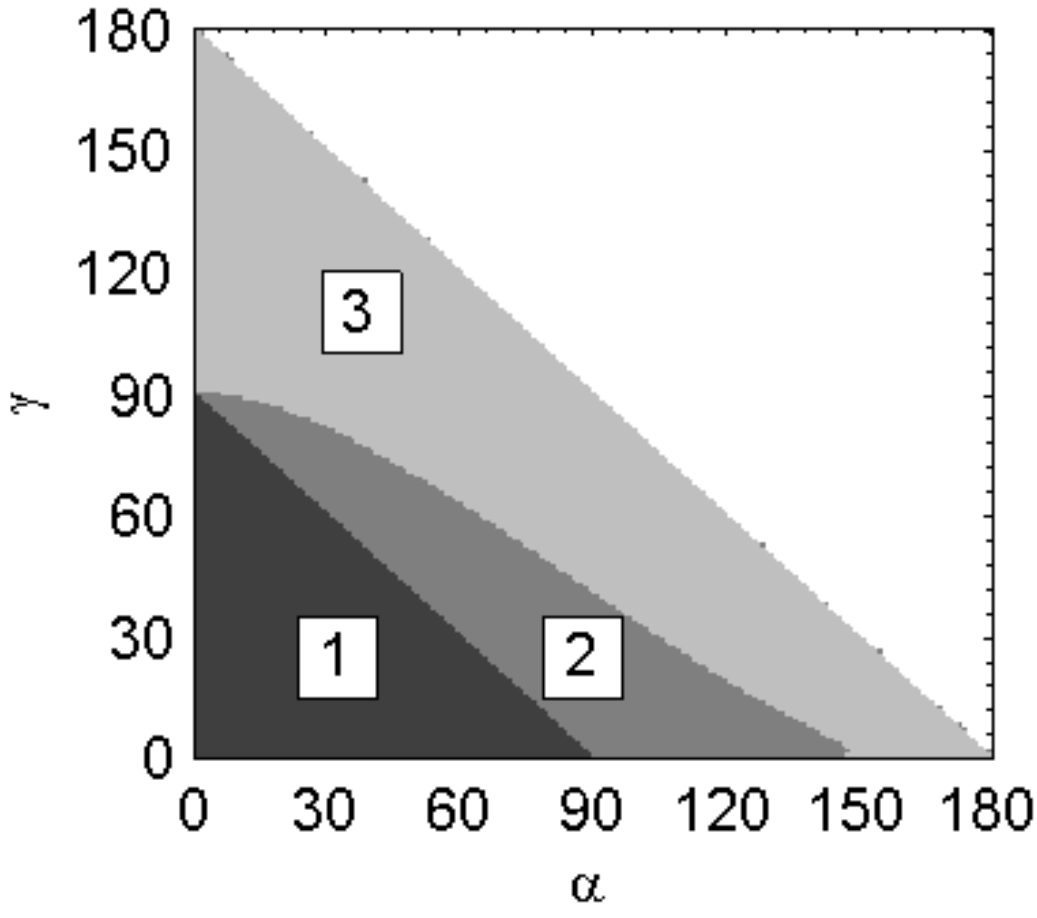




1 off-diag$(A) \leq 0$ 2 off-diag$(A) \not\leq 0$, $\mathbf{A}^{-1} \geq 0$ 3 $\mathbf{A}^{-1} \not\geq 0$

# Experiment II

$-\Delta u = f$ in $\Omega$, $\quad u = 0$ on $\partial\Omega$

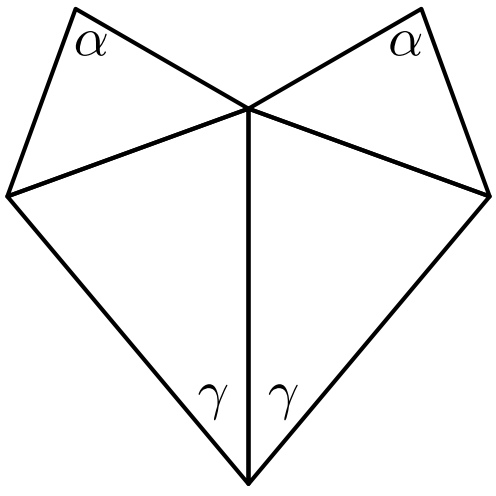


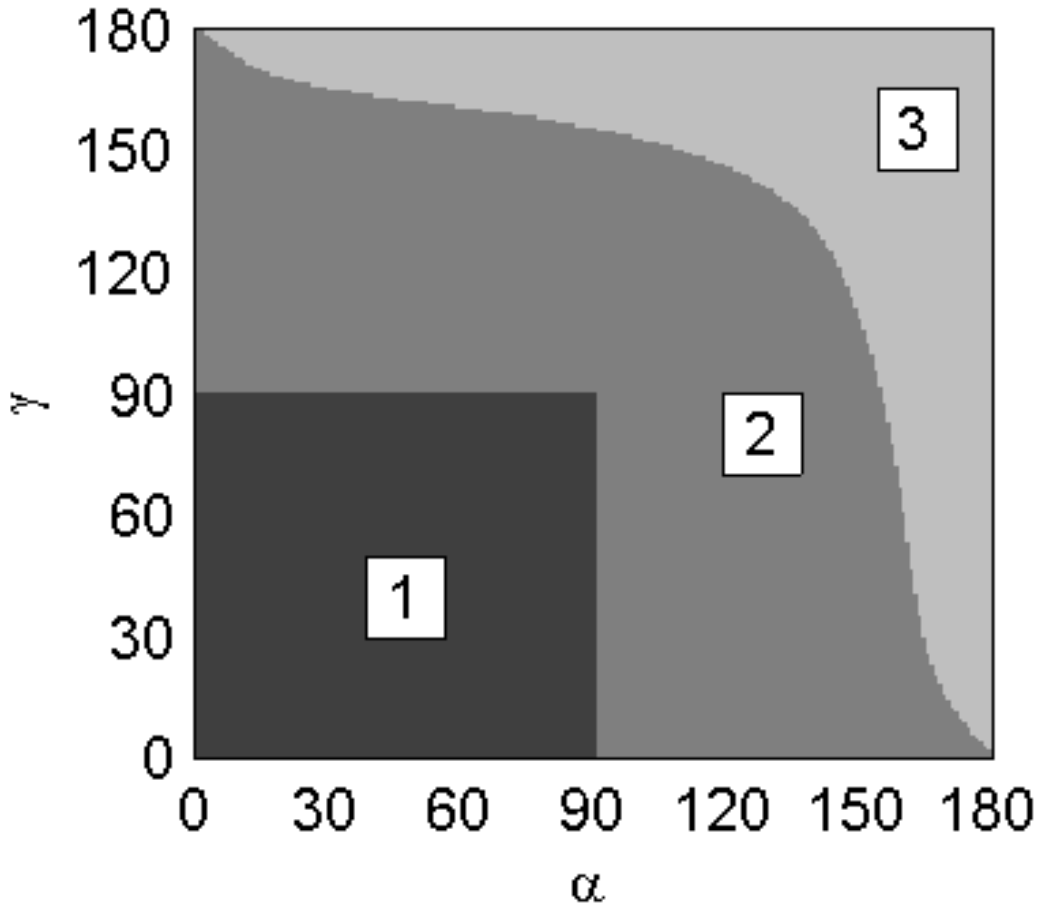




1 off-diag$(A) \leq 0$ 2 off-diag$(A) \not\leq 0$, $\mathbf{A}^{-1} \geq 0$ 3 $\mathbf{A}^{-1} \not\geq 0$

# Experiment III

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$

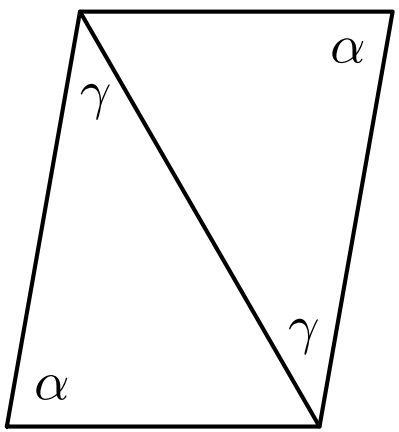


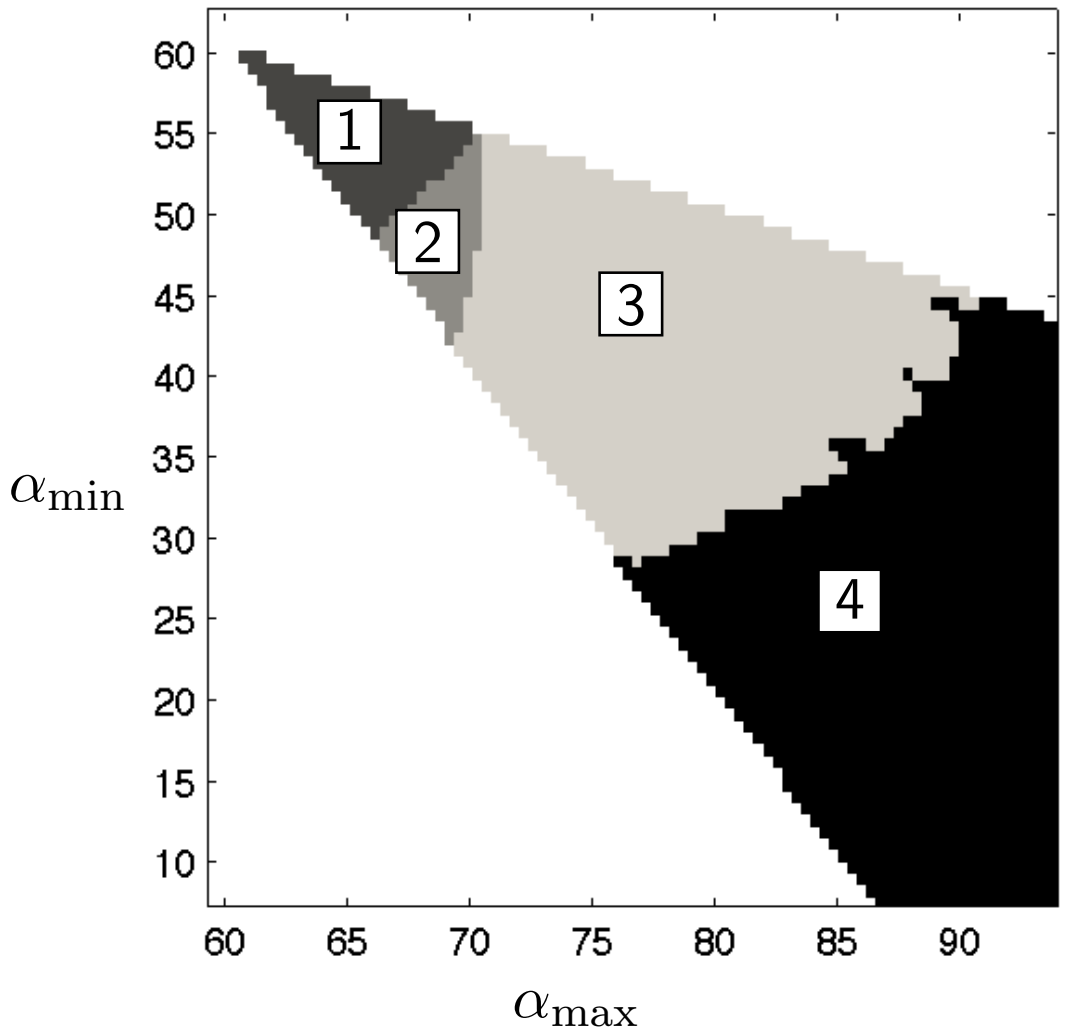




1 off-diag$(A) \leq 0$ 2 off-diag$(A) \not\leq 0$, $\mathbf{A}^{-1} \geq 0$ 3 $\mathbf{A}^{-1} \not\geq 0$

# Experiment prisms

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$



# Experiment prisms

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$

1. $\frac{1}{2}|T_{\max}| \tan \alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} \leq d_i^2 \leq |T_{\min}| \tan \alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \quad \Rightarrow \quad$ DMP

2. $d_L^{(P)} \leq d^{(P)} \leq d_U^{(P)}$ for all $P \in \mathcal{T}_{h,\tau} \quad \Rightarrow \quad$ DMP

$$d_L^{(P)} = \left( \frac{2 \cot \alpha_{\max}^{(T)}}{|T|} \right)^{-\frac{1}{2}}, \quad d_U^{(P)} = \left( \frac{\cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\min}^{(T)}}{2|T|} \right)^{-\frac{1}{2}}$$

3. $\mathbf{A}^{-1} \geq 0 \quad \Leftrightarrow \quad$ DMP

4. no DMP

A. Hannukainen, S. Korotov, T. Vejchodský: Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes, accepted by J. Comput. Appl. Math., 2008.

# Thank you for your attention

Tomáš Vejchodský[1] (vejchod@math.cas.cz)

Antti Hannukainen[2] (antti.hannukainen@hut.fi)

Sergey Korotov[2] (sergey.korotov@hut.fi)

[1] Institute of Mathematics, Academy of Sciences
Žitná 25, 115 67 Prague 1
Czech Republic

[2] Institute of Mathematics
Helsinki University of Technology
P.O. Box 1100, FIN–02015 Espoo
Finland