Some recent results on algebraic flux correction schemes

Gabriel R. Barrenechea[1], Volker John[2] & Petr Knobloch[3]

[1] Department of Mathematics and Statistics, University of Strathclyde, Scotland
[2] WIAS Institute, Berlin, Germany
[3] Charles University, Prague, Czech Republic

INRIA Projet NACHOS,
Nice, Sofia Antipolis, April 2, 2014

The continuous maximum principle :

Theorem

Let u be the solution of the problem

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

and $u = 0$ on $\partial\Omega$. Then, if $f \geq 0$ in Ω, then $u \geq 0$ in Ω, and attains its minimum at the boundary.

The discrete version :

Theorem

Let $u_h \in \mathbb{P}_1(\Omega)$ be the solution of the problem

$$(\nabla u_h, \nabla v_h)_\Omega = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Then, if $f \geq 0$ in Ω and the mesh is acute, then $u_h \geq 0$ in Ω, and attains its minimum at the boundary.

Remark : Under these hypothesis, the matrix $[(\nabla\lambda_j, \nabla\lambda_i)_\Omega]$ is an M-matrix. This is, it is invertible, all the diagonal elements are positive, and the off-diagonal ones are non-positive.

The discrete version :

Theorem

Let $u_h \in \mathbb{P}_1(\Omega)$ be the solution of the problem

$$(\nabla u_h, \nabla v_h)_\Omega = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Then, if $f \geq 0$ in Ω and the mesh is acute, then $u_h \geq 0$ in Ω, and attains its minimum at the boundary.

Remark : Under these hypothesis, the matrix $[(\nabla\lambda_j, \nabla\lambda_i)_\Omega]$ is an M-matrix. This is, it is invertible, all the diagonal elements are positive, and the off-diagonal ones are non-positive.

The convection-diffusion equation

The DMP :

Theorem

Let $u_h \in \mathbb{P}_1(\Omega)$ be the solution of the problem

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot\nabla u_h, v_h)_\Omega = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Then, if $f \geq 0$ in Ω, the mesh is acute, and $\frac{|\boldsymbol{b}|h}{2\varepsilon} < 1$, then $u_h \geq 0$ in Ω, and attains its minimum at the boundary.

Some early solutions

Artificial diffusion :

Find $u_h \in \mathbb{P}_1(\Omega)$ such that

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot\nabla u_h, v_h)_\Omega + s(u_h, v_h) = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Bad news : The linear schemes, such as the artificial diffusion, have two main drawbacks:

- their consistency error leads to a convergence of $O(\sqrt{h})$;
- they produce results which are extremely diffusive.

Some early solutions

Artificial diffusion :

Find $u_h \in \mathbb{P}_1(\Omega)$ such that

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot \nabla u_h, v_h)_\Omega + \alpha\, h\,(\nabla u_h, \nabla v_h)_\Omega = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Bad news : The linear schemes, such as the artificial diffusion, have two main drawbacks:

- their consistency error leads to a convergence of $O(\sqrt{h})$;
- they produce results which are extremely diffusive.

University of Strathclyde Glasgow

Some early solutions

Artificial diffusion :

Find $u_h \in \mathbb{P}_1(\Omega)$ such that

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot \nabla u_h, v_h)_\Omega + \alpha\, h\,(\nabla u_h, \nabla v_h)_\Omega = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Bad news : The linear schemes, such as the artificial diffusion, have two main drawbacks:

- their consistency error leads to a convergence of $O(\sqrt{h})$;
- they produce results which are extremely diffusive.

Some early solutions

Artificial diffusion :

Find $u_h \in \mathbb{P}_1(\Omega)$ such that

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot \nabla u_h, v_h)_\Omega + \alpha\, h\,(\nabla u_h, \nabla v_h)_\Omega = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Bad news : The linear schemes, such as the artificial diffusion, have two main drawbacks:

- their consistency error leads to a convergence of $O(\sqrt{h})$;
- they produce results which are extremely diffusive.

A representative numerical result

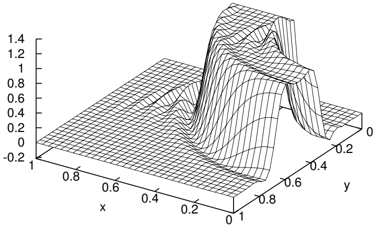


Figure 1 : Solution using a standard LPS method

A representative numerical result - II

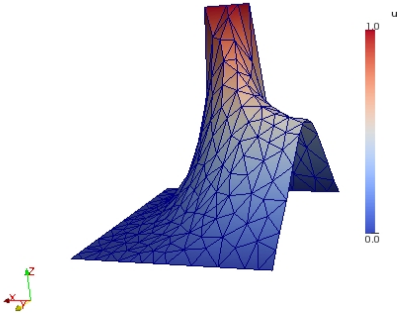


Figure 2 : Solution using the first order artificial diffusion method

Solution: nonlinear schemes

Idea :

Find $u_h \in \mathbb{P}_1(\Omega)$ such that

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot \nabla u_h, v_h)_\Omega + N(u_h; u_h, v_h) = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Main features :

- N is a continuous form, may depend on the residual, or not.
- In some cases (not that many!), the maximum principle can be proved (cf. Burman & Ern).
- Optimal convergence can be proved in most cases.

A more recent alternative (D. Kuzmin) : Algebraic Flux Correction schemes. These work at the matrix level, and have provided very convincing numerical results.

Solution: nonlinear schemes

Idea :

Find $u_h \in \mathbb{P}_1(\Omega)$ such that

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot\nabla u_h, v_h)_\Omega + N(u_h; u_h, v_h) = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Main features :

- N is a continuous form, may depend on the residual, or not.
- In some cases (not that many!), the maximum principle can be proved (cf. Burman & Ern).
- Optimal convergence can be proved in most cases.

A more recent alternative (D. Kuzmin) : Algebraic Flux Correction schemes. These work at the matrix level, and have provided very convincing numerical results.

Solution: nonlinear schemes

Idea :

Find $u_h \in \mathbb{P}_1(\Omega)$ such that

$$\varepsilon\,(\nabla u_h, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot\nabla u_h, v_h)_\Omega + N(u_h; u_h, v_h) = (f, v_h)_\Omega \quad \forall\, v_h \in \mathbb{P}_1(\Omega)\,.$$

Main features :

- N is a continuous form, may depend on the residual, or not.
- In some cases (not that many!), the maximum principle can be proved (cf. Burman & Ern).
- Optimal convergence can be proved in most cases.

A more recent alternative (D. Kuzmin) : Algebraic Flux Correction schemes. These work at the matrix level, and have provided very convincing numerical results.

University of Strathclyde Glasgow

Goals and Outline

1. Goals:
   - Understand the method, and its main features.
   - Give the first steps towards a numerical analysis of it.
   - Study its numerical behaviour.
2. The method for the 1D problem.
3. The discrete maximum principle.
4. Solvability of the linear problems, and the nonlinear one.
5. Concluding remarks.

Goals and Outline

1. Goals:
   - Understand the method, and its main features.
   - Give the first steps towards a numerical analysis of it.
   - Study its numerical behaviour.
2. The method for the 1D problem.
3. The discrete maximum principle.
4. Solvability of the linear problems, and the nonlinear one.
5. Concluding remarks.

Algebraic flux correction schemes

Starting point : A finite element discretisation of our problem of the form:

$$\mathbb{A}\mathrm{U} = \mathrm{G}\,.$$

Define:

$$\mathbb{D} := (d_{ij}) \quad \text{where} \quad d_{ij} := -\max\{a_{ij}, 0, a_{ji}\} \text{ for } i \neq j\,, \quad d_{ii} = -\sum_{j \neq i} d_{ij}\,.$$

Remark: The matrix $\tilde{\mathbb{A}}$ is an M-matrix. Then, it preserves positivity.

Algebraic flux correction schemes

Starting point : A finite element discretisation of our problem of the form:

$$\mathbb{A}\mathrm{U} = \mathrm{G}\,.$$

Define:

$$\mathbb{D} := (d_{ij}) \quad \text{where} \quad d_{ij} := -\max\{a_{ij}, 0, a_{ji}\} \text{ for } i \neq j\,, \quad d_{ii} = -\sum_{j\neq i} d_{ij}\,.$$

Remark: The matrix $\tilde{\mathbb{A}}$ is an M-matrix. Then, it preserves positivity.

Algebraic flux correction schemes

Starting point : A finite element discretisation of our problem of the form:

$$\big(\mathbb{A} + \mathbb{D}\big)\mathrm{U} = \mathrm{G} + \mathbb{D}\mathrm{U}\,.$$

Define:

$$\mathbb{D} := (d_{ij}) \quad \text{where} \quad d_{ij} := -\max\{a_{ij}, 0, a_{ji}\} \text{ for } i \neq j\,, \quad d_{ii} = -\sum_{j \neq i} d_{ij}\,.$$

Remark: The matrix $\tilde{\mathbb{A}}$ is an M-matrix. Then, it preserves positivity.

Algebraic flux correction schemes

Starting point : A finite element discretisation of our problem of the form:

$$\big(\underbrace{\mathbb{A} + \mathbb{D}}_{=:\tilde{\mathbb{A}}}\big)\mathrm{U} = \mathrm{G} + \mathbb{D}\mathrm{U}\,.$$

Define:

$$\mathbb{D} := (d_{ij}) \quad \text{where} \quad d_{ij} := -\max\{a_{ij}, 0, a_{ji}\} \text{ for } i \neq j\,, \quad d_{ii} = -\sum_{j\neq i} d_{ij}\,.$$

Remark: The matrix $\tilde{\mathbb{A}}$ is an M-matrix. Then, it preserves positivity.

Algebraic flux correction schemes

Starting point : A finite element discretisation of our problem of the form:

$$\big(\underbrace{\mathbb{A}+\mathbb{D}}_{=:\tilde{\mathbb{A}}}\big)\mathrm{U} = \mathrm{G} + \mathbb{D}\mathrm{U}\,.$$

Define:

$$\mathbb{D} := (d_{ij}) \quad \text{where} \quad d_{ij} := -\max\{a_{ij}, 0, a_{ji}\} \text{ for } i \neq j\,, \quad d_{ii} = -\sum_{j\neq i} d_{ij}\,.$$

Remark: The matrix $\tilde{\mathbb{A}}$ is an M-matrix. Then, it preserves positivity.

Algebraic flux correction schemes

Equivalent system :

$$\tilde{\mathbb{A}}\,\mathrm{U} = \mathrm{G} + \mathbb{D}\mathrm{U}\,.$$

From the properties of $\mathbb{D}$ it follows that

$$(\mathbb{D}\mathrm{U})_i = \sum_{j \neq i} f_{ij} \quad \text{where } f_{ij} = d_{ij}(u_j - u_i) \text{ are the fluxes}\,.$$

Goal : To limit the fluxes f_{ij} which are responsible for spurious oscillations.

Algebraic flux correction schemes

Equivalent system :

$$\tilde{\mathbb{A}}\,\mathrm{U} = \mathrm{G} + \mathbb{D}\mathrm{U}\,.$$

From the properties of $\mathbb{D}$ it follows that

$$\big(\mathbb{D}\mathrm{U}\big)_i = \sum_{j\neq i} f_{ij} \quad \text{where } f_{ij} = d_{ij}(u_j - u_i) \text{ are the fluxes}\,.$$

Goal : To limit the fluxes f_{ij} which are responsible for spurious oscillations.

Algebraic flux correction schemes

Equivalent system :

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j \neq i} f_{ij}$$

From the properties of $\mathbb{D}$ it follows that

$$\big(\mathbb{D}\mathrm{U}\big)_i = \sum_{j \neq i} f_{ij} \quad \text{where } f_{ij} = d_{ij}(u_j - u_i) \text{ are the fluxes}\,.$$

Goal : To limit the fluxes f_{ij} which are responsible for spurious oscillations.
The limiters α_{ij} should satisfy the following:

- $\alpha_{ij} \in [0, 1]$;
- α_{ij} should be as close to 1 as possible;
- $\alpha_{ij} \approx 1$ where the Galerkin solution is smooth.

Algebraic flux correction schemes

Equivalent system :

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j\neq i} f_{ij}$$

From the properties of $\mathbb{D}$ it follows that

$$\big(\mathbb{D}\mathrm{U}\big)_i = \sum_{j\neq i} f_{ij} \quad \text{where } f_{ij} = d_{ij}(u_j - u_i) \text{ are the fluxes}\,.$$

Goal : To limit the fluxes f_{ij} which are responsible for spurious oscillations.
The limiters α_{ij} should satisfy the following:

- $\alpha_{ij} \in [0, 1]$;
- α_{ij} should be as close to 1 as possible;
- $\alpha_{ij} \approx 1$ where the Galerkin solution is smooth.

Algebraic flux correction schemes

Equivalent system :

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j\neq i} \alpha_{ij}(U) f_{ij}$$

From the properties of $\mathbb{D}$ it follows that

$$\big(\mathbb{D}\mathrm{U}\big)_i = \sum_{j\neq i} f_{ij} \quad \text{where } f_{ij} = d_{ij}(u_j - u_i) \text{ are the fluxes} \,.$$

Goal : To limit the fluxes f_{ij} which are responsible for spurious oscillations.
The limiters α_{ij} should satisfy the following:

- $\alpha_{ij} \in [0,1]$;
- α_{ij} should be as close to 1 as possible;
- $\alpha_{ij} \approx 1$ where the Galerkin solution is smooth.

University of Strathclyde Glasgow

Algebraic flux correction schemes

Equivalent system :

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j \neq i} \alpha_{ij}(U) f_{ij}$$

From the properties of $\mathbb{D}$ it follows that

$$\big(\mathbb{D}\mathrm{U}\big)_i = \sum_{j \neq i} f_{ij} \quad \text{where } f_{ij} = d_{ij}(u_j - u_i) \text{ are the fluxes} .$$

Goal : To limit the fluxes f_{ij} which are responsible for spurious oscillations.
The limiters α_{ij} should satisfy the following:

- $\alpha_{ij} \in [0, 1]$;
- α_{ij} should be as close to 1 as possible;
- $\alpha_{ij} \approx 1$ where the Galerkin solution is smooth.

University of Strathclyde Glasgow

Definition of the limiters

1. Compute $P_i^+, P_i^-, Q_i^+, Q_i^-$ in such a way that, for each pair of neighbouring nodes x_i, x_j with indices such that $a_{ji} \leq a_{ij}$ one performs the updates

$$\begin{aligned} P_i^+ &:= P_i^+ + \max\{0, f_{ij}\}\,, & P_i^- &:= P_i^- - \max\{0, f_{ji}\}\,,\\ Q_i^+ &:= Q_i^+ + \max\{0, f_{ji}\}\,, & Q_i^- &:= Q_i^- - \max\{0, f_{ij}\}\,,\\ Q_j^+ &:= Q_j^+ + \max\{0, f_{ij}\}\,, & Q_j^- &:= Q_j^- - \max\{0, f_{ji}\}\,, \end{aligned}$$

2. Set

$$R_i^+ := \min\left\{1, \frac{Q_i^+}{P_i^+}\right\} \quad , \quad R_i^- := \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}.$$

3. Finally,

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,,\\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad\qquad i,j = 1,\ldots,N\,.$$

The 1D convection-diffusion equation

Model problem :

$$-\varepsilon\, u'' + bu' = g \quad \text{in } (0,1) \quad u(0) = u(1) = 0\,,$$

with positive constants ε and b.

Galerkin FEM : Equidistant nodes $x_i = ih$, with $h = 1/N$. Find $u_h \in \mathbb{P}_1(0,1)$ such that $u_h(0) = u_h(1) = 0$ and

$$\varepsilon(u_h', v_h') + (bu_h', v_h) = (g, v_h) \qquad \forall\, v_h \in \mathbb{P}_1(0,1)\,.$$

Difference equation form : Setting $u_i = u_h(x_i)$, this problem is rewritten as

$$-\varepsilon\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i \quad i = 1, \ldots, N-1.$$

The 1D convection-diffusion equation

Model problem :

$$-\varepsilon\, u'' + bu' = g \quad \text{in } (0,1) \quad u(0) = u(1) = 0\,,$$

with positive constants ε and b.
Galerkin FEM : Equidistant nodes $x_i = ih$, with $h = 1/N$. Find $u_h \in \mathbb{P}_1(0,1)$ such that $u_h(0) = u_h(1) = 0$ and

$$\varepsilon(u_h', v_h') + (bu_h', v_h) = (g, v_h) \qquad \forall\, v_h \in \mathbb{P}_1(0,1)\,.$$

Difference equation form : Setting $u_i = u_h(x_i)$, this problem is rewritten as

$$-\varepsilon\, \frac{u_{i-1} - 2\, u_i + u_{i+1}}{h^2} + b\, \frac{u_{i+1} - u_{i-1}}{2\, h} = g_i \quad i = 1, \ldots, N-1.$$

Assume: $Pe := \frac{bh}{2\varepsilon} > 1$.

The 1D convection-diffusion equation

Model problem :

$$-\varepsilon\, u'' + bu' = g \quad \text{in } (0,1) \quad u(0) = u(1) = 0\,,$$

with positive constants ε and b.
Galerkin FEM : Equidistant nodes $x_i = ih$, with $h = 1/N$. Find $u_h \in \mathbb{P}_1(0,1)$ such that $u_h(0) = u_h(1) = 0$ and

$$\varepsilon(u_h', v_h') + (bu_h', v_h) = (g, v_h) \qquad \forall\, v_h \in \mathbb{P}_1(0,1)\,.$$

Difference equation form : Setting $u_i = u_h(x_i)$, this problem is rewritten as

$$-\varepsilon\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i \quad i = 1, \ldots, N-1.$$

Assume: $Pe := \frac{bh}{2\varepsilon} > 1.$

University of Strathclyde Glasgow

The 1D convection-diffusion equation

Model problem :

$$-\varepsilon\, u'' + bu' = g \quad \text{in } (0,1) \quad u(0) = u(1) = 0\,,$$

with positive constants ε and b.
Galerkin FEM : Equidistant nodes $x_i = ih$, with $h = 1/N$. Find $u_h \in \mathbb{P}_1(0,1)$ such that $u_h(0) = u_h(1) = 0$ and

$$\varepsilon(u_h', v_h') + (bu_h', v_h) = (g, v_h) \qquad \forall\, v_h \in \mathbb{P}_1(0,1)\,.$$

Difference equation form : Setting $u_i = u_h(x_i)$, this problem is rewritten as

$$-\varepsilon\, \frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\, \frac{u_{i+1} - u_{i-1}}{2\,h} = g_i \quad i = 1, \ldots, N-1.$$

Assume: $Pe := \frac{bh}{2\varepsilon} > 1$.

The 1D convection-diffusion equation

Algebraic problem with limited fluxes:

$$\big(\mathbb{A}\mathrm{U}\big)_i + \sum_{j\neq i}(1-\alpha_{ij})f_{ij} = g_i \quad \text{with} \quad f_{ij} = d_{ij}(u_j - u_i)\,.$$

For the 1D problem: the system reduces to $u_0 = u_N = 0$, and

$$-(\varepsilon + \beta_i\,\tilde{\varepsilon})\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i\,, \qquad i = 1,\ldots,N-1\,,$$

where

$$\beta_i = \begin{cases} 1 & \text{if } u_{i+1} \neq u_i \quad \text{and} \quad \dfrac{u_i - u_{i-1}}{u_{i+1} - u_i} < 1\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

and $\tilde{\varepsilon} = \frac{b\,h}{2} - \varepsilon = \varepsilon\,(Pe - 1)$.

University of Strathclyde Glasgow

The Discrete Maximum Principle

Theorem

Consider any $\tilde{\varepsilon} \geq b\,h/2 - \varepsilon$. Then any solution of the nonlinear problem satisfies the discrete maximum principle, i.e., for any $i \in \{1, \ldots, N\}$, one has

$$g_i \geq 0 \qquad \Rightarrow \qquad u_i \geq \min\{u_{i-1}, u_{i+1}\}\,.$$

Moreover, for any $k, l \in \{0, 1, \ldots, N+1\}$ with $k + 1 < l$, one has

$$g_i \geq 0\,, \quad i = k+1, \ldots, l-1 \qquad \Rightarrow \qquad u_i \geq \min\{u_k, u_l\}\,, \quad i = k, \ldots, l\,.$$

Some numerics and the choice of $\tilde{\varepsilon}$

Other possible choices: The artificial diffusion matrix $\mathbb{D}$ can be defined using different combinations of the diffusion and convection matrices. For example:

(F) $\tilde{\varepsilon} = \frac{b\,h}{2} - \varepsilon = \varepsilon(Pe - 1)$.

(C) $\tilde{\varepsilon} = \frac{b\,h}{2}$.

(P) $\tilde{\varepsilon} = \frac{b\,h}{2}\left(\coth Pe - \frac{1}{Pe}\right)$.

Data: $b = f = 1$, $N = 16, \varepsilon = 0.03$, i.e., we solve

$$-0.03u'' + u' = 1 \quad \text{in } (0,1)\,,$$

and $u(0) = u(1) = 0$.

University of Strathclyde Glasgow

Some numerics and the choice of $\tilde{\varepsilon}$

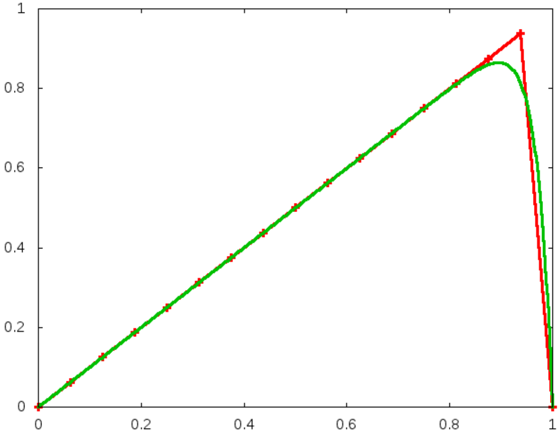


Figure 3 : Comparison of the exact solution (green) and discrete solution with $\tilde{\varepsilon}$ from (F).

Some numerics and the choice of $\tilde{\varepsilon}$

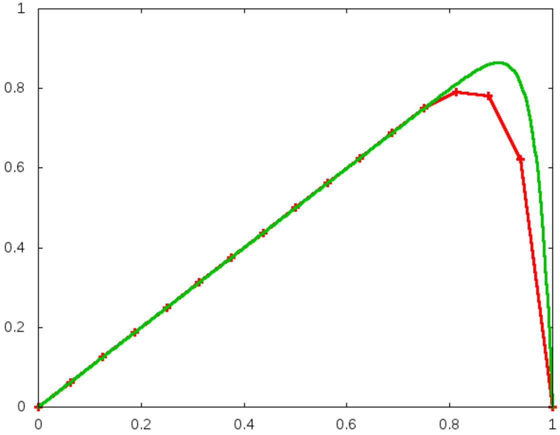


Figure 4 : Comparison of the exact solution (green) and discrete solution with $\tilde{\varepsilon}$ from (C).

Some numerics and the choice of $\tilde{\varepsilon}$

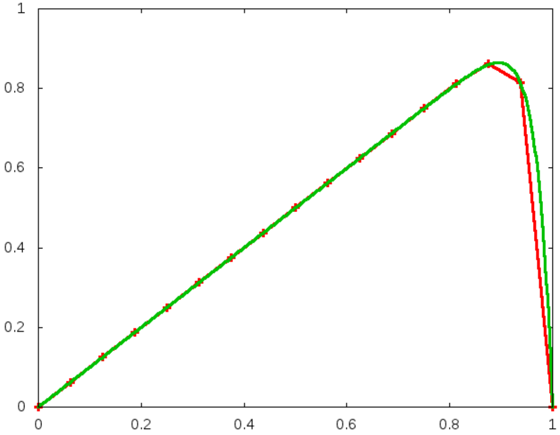


Figure 5 : Comparison of the exact solution (green) and discrete solution with $\tilde{\varepsilon}$ from (P).

University of Strathclyde Glasgow

Bad news from the numerics

- Computations very sensitive to rounding errors.

Idea : replace the condition $u_i < \min\{u_{i-1}, u_{i+1}\}$ by $u_i < \min\{u_{i-1}, u_{i+1}\} - \tau$.

- Not a remedy!

Bad news from the numerics

- Computations very sensitive to rounding errors.

Idea : replace the condition $u_i < \min\{u_{i-1}, u_{i+1}\}$ by $u_i < \min\{u_{i-1}, u_{i+1}\} - \tau$.

- Not a remedy!

Conclusion: The nonlinear problem is not solvable in general!

Bad news from the numerics

- Computations very sensitive to rounding errors.

Idea : replace the condition $u_i < \min\{u_{i-1}, u_{i+1}\}$ by $u_i < \min\{u_{i-1}, u_{i+1}\} - \tau$.

- Not a remedy!

Conclusion: The nonlinear problem is not solvable in general!

Example: $N = 4$, $\varepsilon = 0.03$, $b = 1$, $f_1 = 6$, $f_2 = -6$, $f_3 = 3$, $f_4 = -2$, and $\tilde{\varepsilon}$ from (F).

Bad news from the numerics

- Computations very sensitive to rounding errors.

Idea : replace the condition $u_i < \min\{u_{i-1}, u_{i+1}\}$ by $u_i < \min\{u_{i-1}, u_{i+1}\} - \tau$.

- Not a remedy!

Conclusion: The nonlinear problem is not solvable in general!

Example: $N = 4$, $\varepsilon = 0.03, b = 1, f_1 = 6, f_2 = -6, f_3 = 3, f_4 = -2$, and $\tilde{\varepsilon}$ from (F).

Reminder of the problem:

$$-(\varepsilon + \beta_i(u)\,\tilde{\varepsilon})\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i\,.$$

Bad news from the numerics

- Computations very sensitive to rounding errors.

Idea : replace the condition $u_i < \min\{u_{i-1}, u_{i+1}\}$ by $u_i < \min\{u_{i-1}, u_{i+1}\} - \tau$.

- Not a remedy!

Conclusion: The nonlinear problem is not solvable in general!

Example: $N = 4$, $\varepsilon = 0.03, b = 1, f_1 = 6, f_2 = -6, f_3 = 3, f_4 = -2$, and $\tilde{\varepsilon}$ from (F).

Reminder of the problem:

$$-(\varepsilon + \beta_i(\boldsymbol{u})\,\tilde{\varepsilon})\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i\,.$$

Bad news from the numerics

- Computations very sensitive to rounding errors.

Idea : replace the condition $u_i < \min\{u_{i-1}, u_{i+1}\}$ by $u_i < \min\{u_{i-1}, u_{i+1}\} - \tau$.

- Not a remedy!

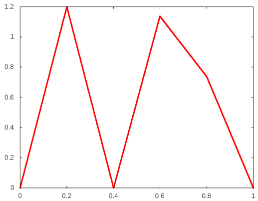
Conclusion: The nonlinear problem is not solvable in general!

Example: $N = 4$, $\varepsilon = 0.03, b = 1, f_1 = 6, f_2 = -6, f_3 = 3, f_4 = -2$, and $\tilde{\varepsilon}$ from (F).

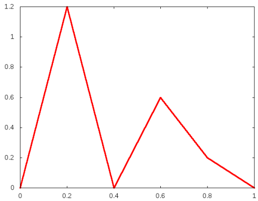
Reminder of the problem:

$$-(\varepsilon + \beta_i(\boldsymbol{u})\,\tilde{\varepsilon})\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i\,.$$

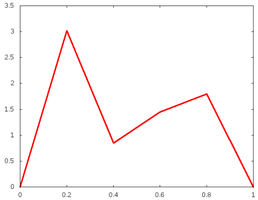
Bad news from the numerics



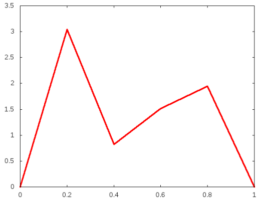
$1\ 1\ 0\ 1 \rightarrow 1\ 1\ 1\ 1$



$1\ 1\ 1\ 1 \rightarrow 1\ 1\ 1\ 0$



$0\ 0\ 1\ 0 \rightarrow 1\ 1\ 0\ 1$



$0\ 0\ 0\ 0 \rightarrow 1\ 1\ 0\ 1$

University of Strathclyde Glasgow

Solvability of the linear subproblems

Theorem

For every choice of $\tilde{\varepsilon} \in \left[\frac{b\,h}{2} - \varepsilon, \frac{b\,h}{2}\right]$ *and every possible* $\beta_i \in [0,1]$, *the problem*

$$-(\varepsilon + \beta_i\,\tilde{\varepsilon})\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i\,,$$

has a unique solution.

Solvability of the nonlinear problem

Main remark : The lack of solvability is due to the discontinuity of the coefficients β_i

Theorem

Let us suppose that the functions $\beta_i : \mathbb{R}^{N+1} \to [0,1]$, $i = 1, ..., N-1$, are continuous, and let $\tilde{\varepsilon}$ be any of the previous choices. Then, the nonlinear FCT scheme has a solution.

Proof: Write the method as the fixed point equation

$$\mathbb{M}(\boldsymbol{\beta}(\boldsymbol{u}))\,\boldsymbol{u} = \boldsymbol{g}\,,$$

apply the fact that the determinant is a continuous function of the entries of a matrix, and Brouwer's fixed point Theorem. □

Solvability of the nonlinear problem

Main remark : The lack of solvability is due to the discontinuity of the coefficients β_i

Theorem

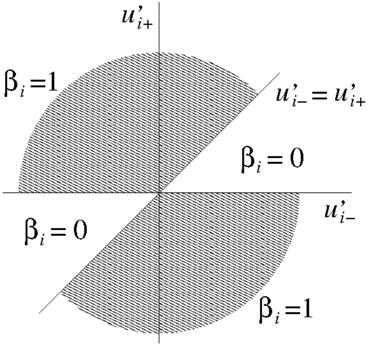
Let us suppose that the functions $\beta_i : \mathbb{R}^{N+1} \to [0,1]$, $i = 1, ..., N-1$, are continuous, and let $\tilde{\varepsilon}$ be any of the previous choices. Then, the nonlinear FCT scheme has a solution.

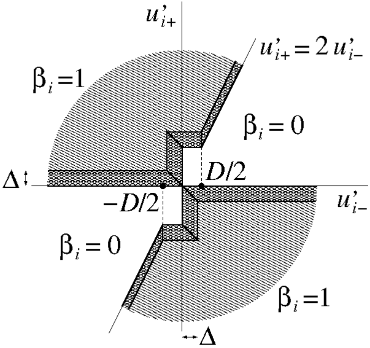
Proof: Write the method as the fixed point equation

$$\mathbb{M}(\boldsymbol{\beta}(\boldsymbol{u}))\,\boldsymbol{u} = \boldsymbol{g}\,,$$

apply the fact that the determinant is a continuous function of the entries of a matrix, and Brouwer's fixed point Theorem. □

Graphical representation of the regularisation





The price to pay: A weak version of the DMP

Theorem

Let $u_0, \ldots, u_{N+1}$ *be a solution of the modified FCT scheme with any functions* $\beta_1, \ldots, \beta_N \in [0,1]$ *as described before. Then*

$$g_i \geq 0 \qquad \Rightarrow \qquad u_i \geq \min\{u_{i-1}, u_{i+1}\} \qquad \text{or} \qquad u_i \geq \max\{u_{i-1}, u_{i+1}\} - \delta\, h\,,$$

for $i = 1, \ldots, N$.

University of Strathclyde Glasgow

Numerical evidence on the violation of the DMP

The problem : $-\varepsilon u'' + u' = 0$ subject to $u(0) = 1$ and $u(1) = 0$. We measured

- $MAX := u_h^{\max} - 1$;
- $RMAX := \max\{(u_h^{\max} - 1)/h\}$;
- Pe_{RMAX} the value of Pe for which the maximum $RMAX$ is attained.

Table 1 : Violation of the discrete maximum principle for the continuous β_i.

	$Pe \in [1, 20)$			$Pe \in [20, \infty)$		
ε	MAX	$RMAX$	Pe_{RMAX}	MAX	$RMAX$	Pe_{RMAX}
10^{-1}	6.62−3	2.65−2	1.25	no $Pe \geq 20$		
10^{-2}	3.55−3	9.27−2	1.85	no $Pe \geq 20$		
10^{-3}	7.14−4	1.28−1	2.79	4.88−15	4.88−14	25.0
10^{-4}	1.06−4	1.40−1	3.77	5.60−14	9.23−13	21.6
10^{-5}	1.41−5	1.47−1	4.80	4.81−13	5.59−10	21.6
10^{-6}	1.77−6	1.51−1	5.84	6.06−12	6.92−8	22.9

University of Strathclyde Glasgow

Some preliminary numerics in 2D: The Hemker problem

Data: $\varepsilon = 10^{-4}$, $\approx 12,000$ $\mathbb{Q}_1$ elements, discontinuous α_{ij} as before, continuous as follows

$$R_i^+ = R_i^- = \min\left\{1, \frac{\min\{Q_i^+, -Q_i^-\}}{\max\{P_i^+, -P_i^-, \tau\}}\right\}.$$

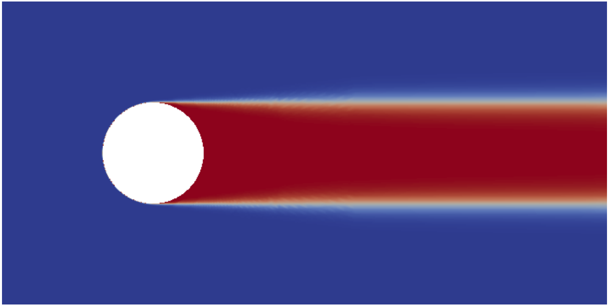


Figure 6 : Discontinuous α_{ij}, non-symmetric

University of Strathclyde Glasgow

Some preliminary numerics in 2D: The Hemker problem

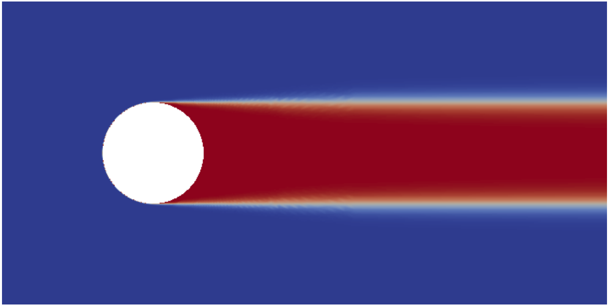


Figure 7 : Continuous α_{ij}, non-symmetric

Conclusions and perspectives

1. Some further insight on FCT schemes.
2. Analysis of a wider class of schemes.
3. Counter-examples of existence of solutions for the original method.
4. A modification that is proved to possess solutions, but satisfies only a weak version of the DMP.

Future extensions:

- Deeper study of the symmetric version in higher dimensions.
- Maximum principle on general meshes.
- (Order of) convergence.
- Time-dependent problems.
- Coupled nonlinear problems in chemical reactions.

University of Strathclyde Glasgow

Conclusions and perspectives

1. Some further insight on FCT schemes.
2. Analysis of a wider class of schemes.
3. Counter-examples of existence of solutions for the original method.
4. A modification that is proved to possess solutions, but satisfies only a weak version of the DMP.

Future extensions:

- Deeper study of the symmetric version in higher dimensions.
- Maximum principle on general meshes.
- (Order of) convergence.
- Time-dependent problems.
- Coupled nonlinear problems in chemical reactions.