Lagged diffusivity method for the solution of nonlinear diffusion convection problems with finite differences

Emanuele Galligani

August 2011

Technical Report of Numerical Analysis

Università degli Studi di Modena e Reggio Emilia

Department of Pure and Applied Mathematics ‘‘G.Vitali’’

TR NA-UniMoRE-3-2011
Abstract. This paper concerns with the analysis of an iterative procedure for the solution of a nonlinear nonstationary diffusion convection equation in a two dimensional bounded domain supplemented by Dirichlet boundary conditions. This procedure, denoted Lagged Diffusivity method, computes the solution by lagging the diffusion term.

A model problem is considered and a finite difference discretization for that model is described. Furthermore, properties of the finite difference operator are proved. Then, a sufficient condition for the convergence of the Lagged Diffusivity method is given.

At each stage of the iterative procedure a linear system has to be solved and the Arithmetic Mean method is used.

Numerical experiments show the efficiency, for different test functions, of the Lagged Diffusivity method combined with the Arithmetic Mean method as inner solver. Better results are obtained when the convection term increases.

Key Words: lagged diffusivity, nonlinear diffusion convection problem, finite differences, Arithmetic Mean method.

AMS Classification: 65H10, 65M06.

C.R. Categories: G.1.8, G.1.5.

1 Statement of the problem

Consider a nonlinear two dimensional diffusion convection equation of the form

$$\frac{\partial u}{\partial t} = \text{div}(\sigma \nabla u) - \mathbf{v} \cdot \nabla u - \alpha u + s$$  \hspace{1cm} (1)

where $u = u(x, y, t)$ is the density function at the point $(x, y)$ at the time $t$ of a diffusion medium $R$, $\sigma = \sigma(x, y, u) > 0$ is the diffusion coefficient or diffusivity and is dependent on the solution $u$, $\alpha = \alpha(x, y) \geq 0$ is the absorption term, $\mathbf{v} = \mathbf{v}(x, y, t, u)$ is the velocity vector and the source term $s(x, y, t)$ is a real valued sufficiently smooth function.

Equation (1) is supplemented by the initial condition ($t = 0$)

$$u(x, y, 0) = U_0(x, y)$$  \hspace{1cm} (2)

in the closure $\bar{R}$ of $R$ and by a mixed boundary condition on the contour $\partial R$ of $R$ of the form

$$\alpha_0 u(x, y, t) + \alpha_1 \frac{\partial u}{\partial \nu}(x, y, t) = U_1(x, y, t)$$  \hspace{1cm} (3)

where $\frac{\partial u}{\partial \nu}$ denotes the outward normal derivative of the function $u$ in the points of $\partial R$.

When the vector $\mathbf{v}$ dominates $\sigma$, the parabolic equation (1), for computational purpose, must be treated as being of hyperbolic type. The nonlinearity introduced by the $u$–dependence of the coefficient $\sigma(u)$ requires that, in general, the solution of equation (1) be approximated by numerical methods.

The method of lines ([3], [7]) is a popular numerical method for determining the solution to a problem of the form (1)–(2)–(3).

There exist various techniques for discretizing the problem (1)–(2)–(3) with respect to the space variable. The use of truncated Taylor series to represent the derivatives in (1) provides more insight into the nature of the truncation error that arises when the continuous model equation (1) is replaced by a discrete set of finite difference equations.

The use of a material balance argument to obtain a discrete version of the problem (1)–(2)–(3) often yields a more physically based system of difference equations that, when solved, ensure that material is conserved.

The box integration method ([13, §6.3]) and the method based on the variational formulation of the problem (1)–(2)–(3) are very useful for the treatment of problems with internal interfaces.

Using the Taylor series approach we obtain the finite difference equations as described below.

In the following we consider $R$ a rectangular domain with boundary $\partial \bar{R}$ and we assume that the functions $\sigma$, $\alpha$ and $s$ satisfy the “smoothness” conditions:
(i) the function $\sigma = \sigma(u)$ is continuous in $u$; the functions $\alpha(x, y)$ and $s(x, y, t)$ are continuous in $x, y$ and in $x, y, t$ respectively;

(ii) there exist two positive constants $\sigma_{\min}$ and $\sigma_{\max}$ such that

$$0 < \sigma_{\min} \leq \sigma(u) \leq \sigma_{\max}$$

uniformly in $u$; in addition, $\alpha(x, y) \geq \alpha_{\min} \geq 0$;

(iii) for fixed $(x, y) \in R$, the function $\sigma(u)$ satisfies Lipschitz condition in $u$ with constant $\Gamma$ (uniformly in $x, y$), $\Gamma > 0$;

Here, the vector $\mathbf{v} = (\tilde{v}_1, \tilde{v}_2)^T$ is assumed to be constant.

We superimpose on $R \cup \partial R$ a grid of points $R_h \cup \partial R_h$; the set of the internal points $R_h$ of the grid are the mesh points $(x_i, y_j)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, M$, with uniform mesh size $h$ along $x$ and $y$ directions respectively, i.e. $x_{i+1} = x_i + h$ and $y_{j+1} = y_j + h$ for $i = 0, \ldots, N$, $j = 0, \ldots, M$.

Furthermore, at the mesh points of $R \cup \partial R$, for $i = 0, \ldots, N + 1$ and $j = 0, \ldots, M + 1$, the solution $u(x_i, y_j, t)$ is approximated by a grid function $u_{ij}(t)$ defined on $R_h \cup \partial R_h$ and satisfying the boundary condition (3) on $\partial R_h$ for $t > 0$ and the initial condition (2) on $R_h \cup \partial R_h$ for $t = 0$.

In order to approximate partial derivatives in space of the equation (1) we shall make use of the centered difference quotients with respect to $x$ and $y$ of a grid function $v_{ij}(t)$ at the mesh point $(x_i, y_j)$:

$$\delta_x v_{ij}(t) = \frac{v_{i+1,j}(t) - v_{i-1,j}(t)}{2h} \quad \delta_y v_{ij}(t) = \frac{v_{i,j+1}(t) - v_{i,j-1}(t)}{2h}$$

Providing a discretization error $O(h^2)$, the finite difference approximation of the right hand side of (1) in $(x_i, y_j)$ is given by

$$\delta_x (\sigma(u_{ij}(t))\delta_x u_{ij}(t)) + \delta_y (\sigma(u_{ij}(t))\delta_y u_{ij}(t)) -$$

$$-\tilde{v}_1 \delta_x u_{ij}(t) - \tilde{v}_2 \delta_y u_{ij}(t) -$$

$$-\alpha(x_i, y_j) u_{ij}(t) + s(x_i, y_j, t)$$

that yields to

$$(B_{ij} + \tilde{B}_{ij})u_{ij-1}(t) + (L_{ij} + \tilde{L}_{ij})u_{i-1,j}(t) - (D_{ij} + \tilde{D}_{ij})u_{ij}(t) +$$

$$+(R_{ij} + \tilde{R}_{ij})u_{i+1,j}(t) + (T_{ij} + \tilde{T}_{ij})u_{ij+1}(t) + s(x_i, y_j, t)$$

where the coefficients are: for $i = 1, \ldots, N$ and $j = 1, \ldots, M$

$$L_{ij} \equiv L_{ij}(u(t)) = \frac{1}{h^2 \sigma} \left( \frac{u_{ij}(t) + u_{i-1,j}(t)}{2} \right)$$

$$B_{ij} \equiv B_{ij}(u(t)) = \frac{1}{h^2 \sigma} \left( \frac{u_{ij}(t) + u_{ij-1}(t)}{2} \right)$$

$$R_{ij} \equiv R_{ij}(u(t)) = \frac{1}{h^2 \sigma} \left( \frac{u_{i+1,j}(t) + u_{ij}(t)}{2} \right)$$

$$T_{ij} \equiv T_{ij}(u(t)) = \frac{1}{h^2 \sigma} \left( \frac{u_{ij+1}(t) + u_{ij}(t)}{2} \right)$$

$$\tilde{L}_{ij} = \frac{\tilde{v}_1}{2h} \quad \tilde{B}_{ij} = \frac{\tilde{v}_2}{2h} \quad \tilde{R}_{ij} = -\frac{\tilde{v}_1}{2h} \quad \tilde{T}_{ij} = -\frac{\tilde{v}_2}{2h} \quad \tilde{D}_{ij} = \alpha(x_i, y_j)$$

$^1$In the approximation of $\text{div}(\sigma \nabla u)$ we consider the centered difference quotients with spacing $h/2$ and then we replace the values of $u_{i+1/2,j}$, $u_{i,j+1/2}$ with $(u_{i,j+1})/2$ and $(u_{i,j+1})/2$ respectively, where we indicate $x_{i+1/2} = x_i \pm h/2$ and $y_{j+1/2} = y_j \pm h/2$.
We denote the mesh points $P_l = (x_i, y_j)$, $(i = 1, ..., N, j = 1, ..., M)$ and we order the points $P_l$ in row lexicographic order: $l = (j - 1) \cdot N + i$. We set $N = M$, and we denote the vector solution $u(t)$ whose components are the values of the grid function at the internal mesh points

$$u = (u_1(t), ..., u_N(t))^T \equiv (u_{11}(t), ..., u_{N1}(t), u_{12}(t), ..., u_{N2}(t), ..., u_{1M}(t), ..., u_{NM}(t))^T$$

Thus, in the case of Dirichlet boundary conditions, i.e., $a_0 = 1$ and $a_1 = 0$ in (3), the right hand side of (1) is approximated by

$$A(u)u + b(u) + s$$

(6)

where the matrix $A(u)$ of order $\mu$ has a block tridiagonal form.

The $M$ diagonal blocks of $A(u)$ are tridiagonal matrices of order $N$ with diagonal elements $a_{ii}(u) = -(D_{ij} + D_{ji})$, the sub– and super–diagonal elements are $a_{ij}(u) = (L_{ij} + L_{ji})$ and $a_{ii+1}(u) = (R_{ij} + R_{ji})$ respectively, $i = 1, ..., N, j = 1, ..., M$ and $l = (j - 1) \cdot N + i$.

The sub– and super–diagonal blocks are diagonal matrices of order $N$ with elements $a_{l-N1}(u) = (B_{ij} + \bar{B}_{ij})$ and $a_{l+N}(u) = (T_{ij} + \bar{T}_{ij})$ respectively, $i = 1, ..., N, j = 1, ..., M$ and $l = (j - 1) \cdot N + i$.

The matrix $A(u)$ is an irreducible matrix ([13, p. 18]).

Providing that the mesh spacing $h$ is sufficiently small, i.e.

$$h < \min \left\{ \frac{2\sigma_{\min}}{|e_1|}, \frac{2\sigma_{\min}}{|e_2|} \right\}$$

(7)

the matrix $A(u)$ is strictly ($\alpha(x, y) > 0$) or irreducibly ($\alpha(x, y) = 0$) diagonally dominant ([13, p. 23]) and has negative diagonal elements, $a_{ii}(u) < 0$ ($i = 1, ..., \mu$) and nonnegative off diagonal elements $a_{ij}(u) \geq 0, i \neq j$, with $i, j = 1, ..., \mu$; therefore $-A(u)$ is an $M$–matrix ([13, p. 91]).

In the case of diffusion equation ($\vec{\partial} = 0$), the matrix $A(u)$ is also symmetric; then $A(u)$ is symmetric negative definite ([13, p. 91]).

The vector $b(u)$ in (6) is obtained by formulae (4)–(5) imposing Dirichlet boundary conditions in (3) and it depends on the function $U_1(x_i, y_j, t)$ at points $(x_i, y_j)$ of $\partial R$ and on the solution $u(t)$ as in the following

$$b(u) = \begin{cases} b_i(u_i) & \text{for } i = 1, ..., N; \\ b_i(u_i) & \text{for } i = \mu - N + 1, ..., \mu; \\ i = l \cdot N + 1 & \text{for } l = 1, ..., M - 2 \\ i = l \cdot N & \text{for } l = 2, ..., M - 1 \\ 0 & \text{otherwise} \end{cases}$$

The vector $s$ has components $s_i(t) = s(x_i, y_j, t)$, $i = 1, ..., N, j = 1, ..., M$ and $l = (j - 1)N + i$.

In the case of homogeneous Neumann boundary conditions, i.e., $a_0 = 0, a_1 = 1$ and $U_1(x, y, t) = 0$ in (3), we consider the finite difference discretization above where the mesh points $P_l = (x_i, y_j)$, $(i = 0, ..., N + 1, j = 0, ..., M + 1)$ are ordered in column lexicographic order ($l = (i - 1) \cdot (M + 2) + j$; here $\mu = (N + 2) \cdot (M + 2)$). Thus the matrix $A(u)$ of order $\mu$ in (6) has a block tridiagonal form as in the previous case, but the diagonal element $a_{ll}(u)$ is the sum, with opposite sign, of the off diagonal elements of the $l$–column.

The matrix $A(u)$ is still irreducible and is singular when $\alpha(x, y) = 0$; in the case of $\alpha(x, y) > 0$ and condition (7), the matrix $-A(u)$ is a nonsingular $M$–matrix.

In this case, the vector $b(u)$ is the null vector.

Then, problem (1)–(2)–(3) is transformed into an initial value problem of $\mu$ equations

$$\frac{du(t)}{dt} = A(u(t))u(t) + b(u(t)) + s(t) \quad t > 0$$

(8)

$$u(0) = u^0$$

(9)

where the components of $u^0$ are the values of the function $U_0$ at the mesh points, i.e., the components $u_i^0$ of $u^0$ are $u_i^0 = U_0(x_i, y_j), i = 1, ..., N; j = 1, ..., M$ and $l = (j - 1)N + i$.

The numerical solution of the initial value problem (8)–(9) is obtained by a step–by–step method.

4
The stiffness phenomenon and the discrete conservation principle (i.e., the question that the analytic properties of the equation (8) reflecting important physical laws must be maintained in the numerical solution) requires to use implicit step-by-step methods.

Widely used is the θ-method (see e.g. [8], [10]), which computes the approximation $u^{n+1}$ to the solution $u(t_{n+1})$ of (8)–(9) using the approximate solution $u^n$ at the time level $t_n$, with $t_n = n\Delta t$ and $\Delta t$ the time step. We indicate $s^n = s(t_n)$. For $n = 0, 1, ...,$ we have

$$\frac{u^{n+1} - u^n}{\Delta t} = \theta(A(u^{n+1})u^{n+1} + b(u^{n+1}) + s^{n+1}) + (1 - \theta)(A(u^n)u^n + b(u^n) + s^n)$$

That is, at each time level $n = 0, 1, ...,$, the vector $u^{n+1} \in \mathbb{R}^\mu$ is the solution of the nonlinear system,

$$F(u) \equiv (I - \Delta t\theta A(u))u - \Delta t\theta b(u) - w = 0 \quad (10)$$

where $\theta$ is a real parameter such that $0 \leq \theta \leq 1$; for any $\theta \neq 0$, the method (10) is implicit. $I$ is the $\mu \times \mu$ identity matrix. The vector $w \in \mathbb{R}^\mu$ is given by

$$w \equiv w^n = (I + \Delta t(1 - \theta)A(u^n))u^n + \Delta t(1 - \theta)b(u^n) + \Delta t(\theta s^{n+1} + (1 - \theta)s^n)$$

We set $\tau = \Delta t\theta$.

We can introduce an iterative method of lagged diffusivity, computing the new iterate $u^{(k+1)}$ keeping the diffusivity term at the previous iteration $k$. That is, since the matrix $I - \tau A(u)$ is nonsingular for all $u \in \mathbb{R}^\mu$ then $u^{(k+1)}$ is the solution of the linear system

$$(I - \tau A(u^{(k)}))u = w + \tau b(u^{(k)})$$

such that the residual

$$r^{(k+1)} = (I - \tau A(u^{(k)}))u^{(k+1)} - w - \tau b(u^{(k)})$$

satisfies the stopping condition

$$||r^{(k+1)}|| \leq \varepsilon_{k+1}$$

where $\varepsilon_k$ is a given tolerance such that $\varepsilon_k \to 0$ when $k \to \infty$.

Here $||\cdot||$ indicates the Euclidean norm.

The initial iterate $u^{(0)}$ of this lagged diffusivity procedure can be set equal to $u^n$.

2 Uniform monotonicity and a convergence result

In this section we consider the nonlinear system $F(u) = 0$ in (10) obtained from the problem (1), with initial condition (2) and Dirichlet boundary conditions in (3), by the finite difference discretization described in the previous section.

We prove that $F(u)$ is continuously and uniformly monotone and then the system $F(u) = 0$ has a unique solution. Moreover we prove that the sequence $\{u^{(k)}\}$ generated by the lagged diffusivity procedure is convergent to the solution. Before this, we have to prove three lemmas on some properties of finite difference operators.

In the following we may consider the matrix $A(u)$ as

$$A(u) = \hat{A}(u) + \hat{D}$$

where $\hat{A}(u)$ and $\hat{D}$ are the block tridiagonal matrices whose row elements are $\{B_{ij}, L_{ij}, -D_{ij}, R_{ij}, T_{ij}\}$ and $\{\hat{B}_{ij}, \hat{L}_{ij}, \hat{R}_{ij}, \hat{T}_{ij}\}$ respectively, while the matrix $\hat{D}$ is a diagonal matrix whose diagonal entries are $\{-D_{ij}\}$.

Furthermore we denote

$$\hat{A}(u) = \hat{A}^x(u) + \hat{A}^y(u)$$

where $\hat{A}^x(u)$ is the block diagonal matrix whose row elements are $\{L_{ij}, -D_{ij}^x, R_{ij}\}$ with $D_{ij}^x = L_{ij} + R_{ij}$, and $\hat{A}^y(u)$ is the block tridiagonal matrix whose row elements are $\{B_{ij}, -D_{ij}^y, T_{ij}\}$ with $D_{ij}^y = B_{ij} + T_{ij}$.
Analogously, we can define $b(u) = b^x(u) + b^y(u)$ where $b^x(u)$ contains the contributions of $U_1(x_0, y_j, t)$ and $U_1(x_{N+1}, y_j, t)$ for $j = 1, ..., M$ and $b^y(u)$ contains the contributions of $U_1(x_i, y_0, t)$ and $U_1(x_i, y_{M+1}, t)$ for $i = 1, ..., N$.

For sake of clarity we set $u_{ij} = u_{ij}(t)$ and $v_{ij} = v_{ij}(t)$, $(i = 0, ..., N + 1, j = 0, ..., M + 1)$, the grid functions defined on $R_h \cup \partial R_h$ and satisfying the Dirichlet boundary condition on $\partial R_h$ for $t > 0$.

Before to prove the main result we summarize in three lemmas some properties of finite difference operators.

**Lemma 1.** Let $\{u_{ij}\}$, $\{v_{ij}\}$, $\{z_{ij}\}$ be three grid functions defined at the mesh points $(x_i, y_j)$ of a grid $R_h \cup \partial R_h$, $i = 0, ..., N + 1$, $j = 0, ..., M + 1$ which are equal to the prescribed function $U_1(x_i, y_j, t)$ at the point $(x_i, y_j)$ of $\partial R_h$ and $t > 0$; then

$$< -\hat{A}(z)u - b(z), v > = < -\hat{A}^x(z)u - b^x(z), v > + < -\hat{A}^y(z)u - b^y(z), v >$$

where

$$< -\hat{A}^x(z)u - b^x(z), v > = \sum_{j=1}^{M} \left[ \sigma(z_{i+1/2}) (u_{ij} - 0_0) v_{1j} + \sum_{i=0}^{N} \sigma(z_{i+1/2}) (u_{ij} - u_{i-1j}) (v_{ij} - v_{i-1j}) + \sigma(z_{N+1/2}) (u_{Nj} - u_{N+1j}) v_{Nj} \right]$$

$$< -\hat{A}^y(z)u - b^y(z), v > = \sum_{i=1}^{M} \left[ \sigma(z_{i+1/2}) (u_{ij} - 0_0) v_{i1} + \sum_{j=2}^{N} \sigma(z_{i+1/2}) (u_{ij} - u_{i-1j}) (v_{ij} - v_{i-1j}) + \sigma(z_{M+1/2}) (u_{Mj} - u_{M+1j}) v_{Mj} \right]$$
Proof. We prove formula (13).

From (5) the expression of $< -\hat{A}^x(z)u - b^x(z), v >$ is given by

$$\sum_{j=1}^{M} \left( - (\sigma(z_{1/j})u_{ij} + (\sigma(z_{3/2}) + \sigma(z_{3/2}))u_{ij} - \sigma(z_{3/2})u_{ij} \right) v_{ij} +$$

$$+ [ - (\sigma(z_{1/2})u_{ij} + (\sigma(z_{3/2}) + \sigma(z_{3/2}))u_{ij} - \sigma(z_{3/2})u_{ij} \right] v_{ij} +$$

$$+ [ - (\sigma(z_{3/2})u_{ij} + (\sigma(z_{3/2}) + \sigma(z_{3/2}))u_{ij} - \sigma(z_{3/2})u_{ij} \right] v_{ij} +$$

$$\vdots$$

$$+ [ - (\sigma(z_{N-3/2})u_{ij} + (\sigma(z_{N-3/2}) + \sigma(z_{N-3/2}))u_{ij} \sigma(z_{N-3/2})u_{ij} \right] v_{ij} +$$

$$+ [ - (\sigma(z_{N-1/2})u_{ij} + (\sigma(z_{N-1/2}) + \sigma(z_{N-1/2}))u_{ij} \sigma(z_{N-1/2})u_{ij} \right] v_{ij} +$$

Collecting the terms we obtain (13).

Lemma 2. Let $\{u_{ij}\}$ and $\{v_{ij}\}$ be two grid functions defined at the mesh points $(x_i, y_j)$ of the grid $R_h \cup \partial R_h$, $i = 0, \ldots, N + 1$, $j = 0, \ldots, M + 1$ such that, at the point $(x_i, y_j)$ of $\partial R_h$ and $t > 0$, the grid function $\{u_{ij}\}$ is equal to the prescribed function $U_1(x_i, y_j, t)$ and the grid function $\{v_{ij}\}$ is equal to the null function, respectively.

Then, we have the following expression for the discrete $l^2(R_h)$ inner product of the vectors $-\hat{A}(u)v$ and $v$

$$< -\hat{A}(u)v, v > = h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} \sigma(u_{i-1/2}) (\nabla u_{ij})^2 + h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} \sigma(u_{i-1/2}) (\nabla v_{ij})^2 +$$

$$\sum_{i=1}^{N} \sigma(u_{i+1/2}) v_{ij} v_{Nj} + \sum_{i=1}^{N} \sigma(u_{i+1/2}) v_{ij} v_{Mj}$$

(15)

Proof. From (5) the expression of $< -\hat{A}(u)v, v >$ is equal to

$$= \sum_{j=1}^{M} \left( - (\sigma(u_{i-1/2}) + \sigma(u_{i+1/2}))v_{ij} - \sigma(u_{i-1/2}) v_{ij} +$$

$$+ [ - (\sigma(u_{i-1/2})v_{ij} + (\sigma(u_{i+1/2}) + \sigma(u_{i+1/2}))v_{ij} - \sigma(u_{i+1/2})v_{ij} \right] v_{ij} +$$

$$+ [ - (\sigma(u_{i-1/2})v_{ij} + (\sigma(u_{i+1/2}) + \sigma(u_{i+1/2}))v_{ij} - \sigma(u_{i+1/2})v_{ij} \right] v_{ij} +$$

$$\vdots$$

$$+ [ - (\sigma(u_{i+1/2})v_{ij} + (\sigma(u_{i+1/2}) + \sigma(u_{i+1/2}))v_{ij} - \sigma(u_{i+1/2})v_{ij} \right] v_{ij} +$$

$$\vdots$$

$$+ [ - (\sigma(u_{i+1/2})v_{ij} + (\sigma(u_{i+1/2}) + \sigma(u_{i+1/2}))v_{ij} - \sigma(u_{i+1/2})v_{ij} \right] v_{ij} +$$

$$= \sum_{j=1}^{M} \left[ \sigma(u_{i+1/2})v_{ij} v_{ij} + \sum_{i=1}^{N} \sigma(u_{i+1/2}) v_{ij} v_{Nj} v_{ij} \right]$$

$$+ \sum_{i=1}^{N} \sigma(u_{i+1/2}) v_{ij} v_{Mj}$$
Using formulae (13) and (14) we have that the term
\[ > \sum_{j=1}^{M} \left[ \sigma(u_{1/2j})(v_{1j} - v_{0j})(v_{1j} - v_{0j}) + \sum_{i=2}^{N} \sigma(u_{i-1/2j})(v_{ij} - v_{i-1j})(v_{ij} - v_{i-1j}) + \sigma(u_{N+1/2j})v_{Nj}v_{Nj} \right] + \sum_{i=1}^{N} \left[ \sigma(v_{1/2i})(v_{1i} - v_{0i})(v_{1i} - v_{0i}) + \sum_{j=2}^{M} \sigma(v_{j-1/2i})(v_{ij} - v_{i-1j})(v_{ij} - v_{i-1j}) + \sigma(v_{M+1/2i})v_{Mi}v_{Mi} \right] \]

Then, we have the following expression for the discrete boundary
\[ \{ \sigma(v_{ij}) \}
\]
is equal to \( v_{\beta} \), and \( \{ \sigma(u_{ij}) \} \)
is equal to zero for all the points of \( \partial R_h \), the last expression can be written
\[ \sum_{j=1}^{M} \left[ \sigma(u_{1/2j})(v_{1j} - v_{0j})(v_{1j} - v_{0j}) + \sum_{i=2}^{N} \sigma(u_{i-1/2j})(v_{ij} - v_{i-1j})(v_{ij} - v_{i-1j}) + \right. \]
\[ \left. \sigma(u_{N+1/2j})v_{Nj}v_{Nj} \right] + \sum_{i=1}^{N} \left[ \sigma(v_{1/2i})(v_{1i} - v_{0i})(v_{1i} - v_{0i}) + \sum_{j=2}^{M} \sigma(v_{j-1/2i})(v_{ij} - v_{i-1j})(v_{ij} - v_{i-1j}) + \right. \]
\[ \left. \sigma(v_{M+1/2i})v_{Mi}v_{Mi} \right] \]

From the definition of backward difference quotients, we have formula (15).

**Lemma 3.** Let \( \{ u_{ij} \} \), \( \{ v_{ij} \} \) and \( \{ \tilde{v}_{ij} \} \) be three grid functions of \( B \) such that, at the points \( (x_i, y_j) \) of the boundary \( \partial R_h \) they are equal to the prescribed function \( U_1(x_i, y_j, t) \) \((t > 0)\).

Then, we have the following expression for the discrete \( I^2(R_h) \) inner product
\[ \left< -\hat{A}(u) + \hat{A}(v) \right> \hat{v} - b(u) + b(v), u - \hat{v} > \leq \frac{\beta \Gamma}{\phi} \| u - v \|_2^2 + \frac{\beta \Gamma}{2h^2} \| u - \hat{v} \|_2^2 + \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} (|\nabla_x(u_{ij} - \tilde{v}_{ij})|^2 + |\nabla_y(u_{ij} - \tilde{v}_{ij})|^2) \]

where \( \Gamma > 0 \) is the Lipschitz constant of condition (iii), \( \beta > 0 \) is a constant for which \( |\nabla_x v_{ij}| \leq \beta \) and \( |\nabla_y v_{ij}| \leq \beta \), and \( \phi \) is an arbitrary positive number.

**Proof.** Using formulae (13) and (14) we have that the term \( < -\hat{A}(u) + \hat{A}(v) \hat{v} - b(u) + b(v), u - \hat{v} > \)
is equal to
\[ \sum_{j=1}^{M} \left[ \sigma(u_{1/2j}) - \sigma(v_{1/2j}) \right] (\tilde{v}_{1j} - \tilde{v}_{0j})(u_{1j} - \tilde{v}_{1j}) + \sigma(u_{N+1/2j}) - \sigma(v_{N+1/2j}) \right] (\tilde{v}_{Nj} - \tilde{v}_{Nj})(u_{Nj} - \tilde{v}_{Nj}) \]
\[ + \sum_{i=1}^{N} \left[ \sigma(u_{i-1/2j}) - \sigma(v_{i-1/2j}) \right] (\tilde{v}_{ij} - \tilde{v}_{i-1j})(u_{ij} - u_{i-1j}) - (\tilde{v}_{ij} - \tilde{v}_{i-1j})) \]
\[ + \sum_{j=2}^{M} \left[ \sigma(u_{j-1/2i}) - \sigma(v_{j-1/2i}) \right] (\tilde{v}_{i} - \tilde{v}_{i-1})(u_{ij} - u_{i-1j}) - (\tilde{v}_{ij} - \tilde{v}_{i-1j})) \]

Since \( u_{ij} - \tilde{v}_{ij} \) is equal to zero for all the points of \( \partial R_h \), we have
\[ \sum_{j=1}^{M} \left[ \sigma(u_{1/2j}) - \sigma(v_{1/2j}) \right] (\tilde{v}_{1j} - \tilde{v}_{0j})(u_{1j} - \tilde{v}_{1j}) - (u_{0j} - \tilde{v}_{0j}) \]
\[ + \sigma(u_{N+1/2j}) - \sigma(v_{N+1/2j}) \right] (\tilde{v}_{Nj} - \tilde{v}_{Nj})(u_{Nj} - \tilde{v}_{Nj}) - (u_{Nj+1} - \tilde{v}_{Nj+1})) \]
\[ + \sum_{i=2}^{N} \left[ \sigma(u_{i-1/2j}) - \sigma(v_{i-1/2j}) \right] (\tilde{v}_{ij} - \tilde{v}_{i-1j})(u_{ij} - \tilde{v}_{ij}) - (u_{i-1j} - \tilde{v}_{i-1j})) \]
\[ + \sum_{j=2}^{M} \left[ \sigma(u_{j-1/2i}) - \sigma(v_{j-1/2i}) \right] (\tilde{v}_{i} - \tilde{v}_{i-1})(u_{ij} - u_{i-1j}) - (\tilde{v}_{ij} - \tilde{v}_{i-1j})) \]
\[ + \sigma(u_{M+1/2i}) - \sigma(v_{M+1/2i}) \right] (\tilde{v}_{Mi} - \tilde{v}_{Mi})(u_{Mi} - \tilde{v}_{Mi}) - (u_{Mi+1} - \tilde{v}_{Mi+1})) \]
\[ + \sum_{j=2}^{M} \left[ \sigma(u_{j-1/2i}) - \sigma(v_{j-1/2i}) \right] (\tilde{v}_{i} - \tilde{v}_{i-1})(u_{ij} - u_{i-1j}) - (\tilde{v}_{ij} - \tilde{v}_{i-1j})) \]
\[ + \sigma(u_{M+1/2i}) - \sigma(v_{M+1/2i}) \right] (\tilde{v}_{Mi} - \tilde{v}_{Mi})(u_{Mi} - \tilde{v}_{Mi}) - (u_{Mi+1} - \tilde{v}_{Mi+1})) \]
Then the term $| < -A(u) + A(v))\tilde{v} - b(u) + b(v), u - \tilde{v} > |$ is less than

$$
\leq h^2 \sum_{j=1}^{M} |(\sigma(u_{1/2}) - \sigma(v_{1/2}))||\nabla_x v_{1/2}||\nabla_x (u_{1/2} - v_{1/2})| + \\
+ |\sigma(u_{N+1/2}) - \sigma(v_{N+1/2})||\nabla_x v_{N+1/2}||\nabla_x (u_{N+1/2} - v_{N+1/2})| + \\
+ \sum_{i=2}^{N} |(\sigma(u_{i-1/2}) - \sigma(v_{i-1/2}))||\nabla_x v_{i-1/2}||\nabla_x (u_{i-1/2} - v_{i-1/2})| + \\
+ \sum_{j=1}^{N} |(\sigma(u_{1/2}) - \sigma(v_{1/2}))||\nabla_y v_{1/2}||\nabla_y (u_{1/2} - v_{1/2})| + \\
+ \sum_{j=2}^{N} |(\sigma(u_{1/2}) - \sigma(v_{1/2}))||\nabla_y v_{1/2}||\nabla_y (u_{1/2} - v_{1/2})|
$$

We used formula (12) for $\nabla_x \tilde{v}_{ij}$ and $\nabla_y \tilde{v}_{ij}$. Since the property of Lipschitz continuity on the function $\sigma$ with Lipschitz constant $\Gamma$ (property (iii)), we have

$$
\leq \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N+1} |(\sigma(u_{ij}) + u_{i-1/2} - \sigma(v_{ij}) + v_{i-1/2})||\nabla_x (u_{ij} - v_{ij})| + \\
+ \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N+1} |(\sigma(u_{i-1/2}) - \sigma(v_{i-1/2}))||\nabla_y (u_{i-1/2} - v_{i-1/2})|
$$

$$
\leq \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N+1} |(\sigma(u_{ij}) - \sigma(v_{ij}))||\nabla_x (u_{ij} - v_{ij})| + \\
+ \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N+1} |(\sigma(u_{i-1/2}) - \sigma(v_{i-1/2}))||\nabla_y (u_{i-1/2} - v_{i-1/2})|
$$

$$
\leq \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} |(\sigma(u_{i-j}) - \sigma(v_{i-j}))||\nabla_x (u_{i-j} - v_{i-j})| + \beta h^2 \sum_{j=1}^{M} |(u_{N+1-r} - v_{N+1-r})||\nabla_x (u_{N+1-r} - v_{N+1-r})| + \\
+ \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} |(u_{i-j} - v_{i-j})||\nabla_x (u_{i-j} - v_{i-j})| + \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} |(u_{i-j} - v_{i-j})||\nabla_x (u_{i-j} - v_{i-j})| + \\
+ \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} |(u_{i-j} - v_{i-j})||\nabla_y (u_{i-j} - v_{i-j})| + \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} |(u_{i-j} - v_{i-j})||\nabla_y (u_{i-j} - v_{i-j})|
$$

$$
\leq \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} |(u_{i-j} - v_{i-j})||\nabla_x (u_{i-j} - v_{i-j})| + \beta h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} |(u_{i-j} - v_{i-j})||\nabla_y (u_{i-j} - v_{i-j})|
$$
By the well known technical trick\(^2\) we can write

\[
\frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{|u_{ij} - v_{ij}|^2}{2\phi} + \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2}{2} \phi + \\
+ \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{|u_{ij} - v_{ij}|^2}{2\phi} + \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2}{2} \phi + \\
+ \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{|u_{ij} - v_{ij}|^2}{2\phi} + \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\left|\nabla_x(u_{i+1,j} - \tilde{v}_{i+1,j})\right|^2}{2} \phi + \\
+ \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{|u_{ij} - v_{ij}|^2}{2\phi} + \frac{\beta h^2 \Gamma}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\left|\nabla_y(u_{i+1,j} - \tilde{v}_{i+1,j})\right|^2}{2} \phi 
\leq \frac{\beta \Gamma}{\phi} \|u - v\|^2_h + \frac{\beta h^2 \Gamma \phi}{4} \sum_{j=1}^{M} \sum_{i=1}^{N} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) + \\
+ \frac{\beta h^2 \Gamma \phi}{4} \sum_{j=1}^{M} \sum_{i=1}^{N} \left(\left|\nabla_x(u_{i+1,j} - \tilde{v}_{i+1,j})\right|^2 + \left|\nabla_y(u_{i+1,j} - \tilde{v}_{i+1,j})\right|^2\right)
\]

Now we analyse the term \(\sum_{j=1}^{M} \sum_{i=1}^{N} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right)\). We have

\[
\sum_{j=1}^{M} \sum_{i=1}^{N} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) \\
= \sum_{j=1}^{M} \sum_{i=1}^{N+1} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) \\
\leq \sum_{j=1}^{M} \sum_{i=1}^{N+1} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) \\
\leq \sum_{j=1}^{M} \sum_{i=1}^{N+1} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) + \\
+ \sum_{j=1}^{M} \left|\nabla_x(u_{N+1,j} - \tilde{v}_{N+1,j})\right|^2 + \sum_{i=1}^{N} \left|\nabla_y(u_{i,M+1} - \tilde{v}_{i,M+1})\right|^2 \\
\leq \sum_{j=1}^{M} \sum_{i=1}^{N+1} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) + \\
+ \sum_{j=1}^{M} \frac{|u_{N+1,j} - \tilde{v}_{N+1,j}|^2}{h^2} + \sum_{i=1}^{N} \frac{|u_{i,M} - \tilde{v}_{i,M}|^2}{h^2} \\
\leq \sum_{j=1}^{M} \sum_{i=1}^{N+1} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) + \\
+ \frac{1}{h^2} \sum_{j=1}^{M} \sum_{i=1}^{N} |u_{ij} - \tilde{v}_{ij}|^2 + \frac{1}{h^2} \sum_{j=1}^{M} \sum_{i=1}^{N} |u_{ij} - \tilde{v}_{ij}|^2 \\
\leq \sum_{j=1}^{M} \sum_{i=1}^{N} \left(\left|\nabla_x(u_{ij} - \tilde{v}_{ij})\right|^2 + \left|\nabla_y(u_{ij} - \tilde{v}_{ij})\right|^2\right) + \frac{2}{h^2} \|u - \tilde{v}\|_h^2
\]

\(^2\)For any positive numbers \(a, b\) we have \(ab < \frac{1}{2}a^2 + \frac{1}{2}b^2\).
Thus, the term $|<(-\hat{A}(u) + A(u))\tilde{v} - b(u) + b(v), u - \tilde{v}>|$ is less than

$$\leq \frac{\beta \Gamma}{\phi} ||u - v||^2_h + \frac{\beta h^2 \Gamma \phi}{4} \sum_{j=1}^{M} \sum_{i=1}^{N} (||\nabla_x(u_{ij} - \tilde{v}_{ij})||^2 + ||\nabla_y(u_{ij} - \tilde{v}_{ij})||^2) +$$

$$+ \frac{\beta \Gamma \phi}{2h^2} ||u - \tilde{v}||^2_h + \frac{\beta h^2 \Gamma \phi}{4} \sum_{j=1}^{M} \sum_{i=1}^{N} (||\nabla_x(u_{ij} - \tilde{v}_{ij})||^2 + ||\nabla_y(u_{ij} - \tilde{v}_{ij})||^2)$$

Then, we have the result (16).

As consequence of lemmas 1, 2 and 3 we prove the result of the uniform monotonicity of the mapping $F(u)$; thus the nonlinear system (10) has a unique solution in $\mathcal{B}$.

**Theorem 3.** If

$$\alpha_{\text{min}} + \frac{1}{\tau} > \frac{\sigma_{\text{min}}}{h^2} + \frac{\beta^2 \Gamma^2}{2\sigma_{\text{min}}}$$

(17)

then the nonlinear system $F(u) = 0$, with $F(u)$ as in (10), has a unique solution in $\mathcal{B}$.

**Proof.** From (10) we can write

$$\frac{1}{\tau} F(u) = H(u) - \frac{1}{\tau} w$$

with

$$H(u) = \frac{1}{\tau} u - A(u)u - b(u)$$

We next show that the mapping $F(u)$ (or $H(u)$) is continuous and uniformly monotone, i.e. there exists a positive scalar $\gamma$ such that

$$< F(u) - F(v), u - v > \geq \gamma < u - v, u - v >$$

for all $u$ and $v$ in $\mathcal{B}$. Then, $H(u) = 1/\tau w$ has a unique solution for any vector $w$ ([9, p. 143, 167]).

From

$$\frac{1}{\tau} (F(u) - F(v)) = H(u) - H(v) = (-A(u) + \frac{1}{\tau} I)u - b(u) - (-A(v) + \frac{1}{\tau} I)v + b(v)$$

$$= (-A(u) + \frac{1}{\tau} I)(u - v) + (-A(u) + A(v))v - b(u) + b(v)$$

we have

$$\frac{1}{\tau} < F(u) - F(v), u - v > = < (-A(u) + \frac{1}{\tau} I)(u - v), u - v > +$$

$$+ < (-A(u) + A(v))v - b(u) + b(v), u - v >$$

$$= < (-\hat{A}(u) - \hat{A} - \hat{D} + \frac{1}{\tau} I)(u - v), u - v > +$$

$$+ < (-\hat{A}(u) + \hat{A}(v))v - b(u) + b(v), u - v >$$

$$= < -\hat{A}(u)(u - v), u - v > + < -\hat{A}(u - v), u - v > +$$

$$+ < (-\hat{D} + \frac{1}{\tau} I)(u - v), u - v > +$$

$$+ < (-\hat{A}(u) + \hat{A}(v))v - b(u) + b(v), u - v >$$

Since $\hat{A}$ is the skew–symmetric part of $A(u)$,

3We define $A_s(u)$ and $A_{ss}(u)$ the symmetric and the skew–symmetric part of $A(u)$ respectively:

$$A_s(u) = \frac{A(u) + A(u)^T}{2} \quad A_{ss}(u) = \frac{A(u) - A(u)^T}{2}$$

and we have $A_{ss}(u) = \hat{A}$.}

11
Then
\[ \frac{1}{\tau} < F(u) - F(v), u - v > = < -A(u) - (u - v), u - v > + (-D + \frac{1}{\tau}I)(u - v), u - v > + < -\hat{A}(u) + \hat{A}(v))v - b(u) + b(v), u - v > \]
We separately examine the terms < -\hat{A}(u)(u - v), u - v >, < (-\hat{D} + \frac{1}{\tau}I)(u - v), u - v > and < (-\hat{A}(u) + \hat{A}(v))v - b(u) + b(v), u - v > of the right hand side of the last expression.

By using Lemma 2 we have

\[ < -\hat{A}(u)(u - v), u - v > = h^2 \sum_{j=1}^{M} \sum_{i=1}^{N} \sigma(u_{i-1/2})((\nabla_x (u_{ij} - v_{ij}))^2 + \sum_{i=1}^{N} \sigma(u_{i-1/2})(\nabla_y (u_{ij} - v_{ij}))^2 + \sum_{j=1}^{M} \sigma(u_{N+j-1/2})(u_{Nj} - v_{Nj})^2 + \sum_{i=1}^{N} \sigma(u_{iM+1/2})(u_{iM} - v_{iM})^2 \]

The assumption (ii) on the uniform lower boundedness of \( \sigma \) respect to the variable \( u \) permits to write

\[ < -\hat{A}(u)(u - v), u - v > = h^2 \tau \sigma_{\min} \sum_{j=1}^{M} \sum_{i=1}^{N} (|\nabla_x (u_{ij} - v_{ij})|^2 + |\nabla_y (u_{ij} - v_{ij})|^2) + \tau \sigma_{\min} \sum_{j=1}^{M} |u_{Nj} - v_{Nj}|^2 + \tau \sigma_{\min} \sum_{i=1}^{N} |u_{iM} - v_{iM}|^2 \]

Since \( \alpha(x, y) \geq \alpha_{\min} \geq 0 \) we have

\[ < (-\hat{D} + \frac{1}{\tau}I)(u - v), u - v > \geq h^2 (\alpha_{\min} + \frac{1}{\tau}) \sum_{j=1}^{M} \sum_{i=1}^{N} (u_{ij} - v_{ij})^2 \]

\[ \geq (\alpha_{\min} + \frac{1}{\tau}) \|u - v\|_h^2 \]

By using Lemma 3 we have

\[ < (-\hat{A}(u) + \hat{A}(v))v - b(u) + b(v), u - v > \geq -\beta \Gamma (\frac{1}{\phi} + \frac{\phi}{2h^2}) \|u - v\|_h^2 - \frac{\beta h^2 \Gamma \phi}{2} \sum_{j=1}^{M} \sum_{i=1}^{N} (|\nabla_x (u_{ij} - v_{ij})|^2 + |\nabla_y (u_{ij} - v_{ij})|^2) \]

Then, collecting the last three inequalities we obtain

\[ \frac{1}{\tau} < F(u) - F(v), u - v > \geq (\alpha_{\min} + \frac{1}{\tau} - \beta \Gamma (\frac{1}{\phi} + \frac{\phi}{2h^2})) \|u - v\|_h^2 + h^2 (\tau \sigma_{\min} - \frac{\beta \Gamma \phi}{2}) \sum_{j=1}^{M} \sum_{i=1}^{N} (|\nabla_x (u_{ij} - v_{ij})|^2 + |\nabla_y (u_{ij} - v_{ij})|^2) + \sum_{j=1}^{M} \sum_{i=1}^{N} |u_{Nj} - v_{Nj}|^2 + \tau \sigma_{\min} \sum_{i=1}^{N} |u_{iM} - v_{iM}|^2 \]

\[ \geq (\alpha_{\min} + \frac{1}{\tau} - \beta \Gamma (\frac{1}{\phi} + \frac{\phi}{2h^2})) \|u - v\|_h^2 + h^2 (\tau \sigma_{\min} - \frac{\beta \Gamma \phi}{2}) \sum_{j=1}^{M} \sum_{i=1}^{N} (|\nabla_x (u_{ij} - v_{ij})|^2 + |\nabla_y (u_{ij} - v_{ij})|^2) \]

12
If we set 
\[ \phi = \frac{2\sigma_{\min}}{\beta \Gamma} \]
we obtain
\[ \frac{1}{\tau} < F(u) - F(v), u - v > \geq (\alpha_{\min} + \frac{1}{\tau} - \frac{\beta^2 \Gamma^2}{\sigma_{\min}} - \frac{\sigma_{\min}}{h^2})||u - v||_h^2 \]

When condition (17) holds, then the mapping \( F(u) \) is uniformly monotone on \( \mathcal{B} \) with constant
\[ \gamma = \tau(\alpha_{\min} - \frac{\sigma_{\min}}{h^2} - \frac{\beta^2 \Gamma^2}{2\sigma_{\min}}) + 1 \]
in (18).

Now we can state a result for the convergence of the lagged diffusivity method where the vector \( u^{(k+1)} \)
is the approximate solution of the linear system
\[ (I - \tau A(u^{(k)}))u = w + \tau b(u^{(k)}) \]
(19)
such that the residual
\[ r^{(k+1)} = (I - \tau A(u^{(k)}))u^{(k+1)} - w - \tau b(u^{(k)}) \]
satisfies the stopping condition
\[ ||r^{(k+1)}|| \leq \varepsilon_{k+1} \]
(20)
where \( \varepsilon_k \) is a given tolerance such that \( \varepsilon_k \to 0 \) when \( k \to \infty \).
Thus, the iterate \( u^{(k+1)} \) is the solution of the system (10) whose diffusivity term \( \sigma \) in \( A(u) \) and \( b(u) \)
depends on the iterate \( u^{(k)} \) and the inhomogeneous term now depends by \( u^{(k+1)} \).
We suppose that the grid functions \( \{u_{ij}^{(k)}\} \), \( k = 0, 1, ... \), are belonging to the set \( \mathcal{B} \). In particular, the backward difference quotients of each grid function \( \{u_{ij}^{(k)}\} \) are bounded. Since this bound depends on the inhomogeneous term, we have that there exist two constants \( \beta > 0 \) and \( \beta_0 > 0 \) such that
\[ |\nabla_x u_{ij}^{(k)}| \leq \tilde{\beta}_k \leq |\nabla_y u_{ij}^{(k)}| \leq \tilde{\beta}_k \]
(21)
with \( \tilde{\beta}_k = \beta + \varepsilon_k \beta_0 \). (Formula (21) replaces formula (12)).

**Theorem 4.** Let \( u^* \in \mathcal{B} \) be the solution of the nonlinear system \( F(u) = 0 \) in (10); let \( u^{(k+1)} \) be the solution of the linear system in (19) with condition (20). Then, if
\[ \alpha_{\min} + \frac{1}{\tau} > \frac{\sigma_{\min}}{h^2} \]
(22)
then, the sequence \( \{u^{(k)}\} \) converges to \( u^* \).

**Proof.** The solution \( u^* \) in \( \mathcal{B} \) of (10) satisfies the equation
\[ u^* - \tau A(u^*)u^* - w - \tau b(u^*) = 0 \]
(23)
and the iterate \( u^{(k+1)} \) satisfies the equation
\[ u^{(k+1)} - \tau A(u^{(k)})u^{(k+1)} - w - \tau b(u^{(k)}) = r^{(k+1)} \]
(24)
Subtracting (24) from (23), we obtain
\[ -A(u^*)u^* + A(u^{(k)})u^{(k+1)} + \frac{1}{\tau} u^* - \frac{1}{\tau} u^{(k+1)} - b(u^*) + b(u^{(k)}) = -\frac{1}{\tau} r^{(k+1)} \]
Taking into account of the identity
\[ -A(u)u + A(u)v = -A(u)(u - v) + (-A(u) + A(u))v \]
for all $u$, $v$ and $w$ belonging to $B$, we can write

$$(-A(u^*) + \frac{1}{\tau} I)(u^* - u^{(k+1)}) + (-A(u^*) + A(u^{(k)}))u^{(k+1)} - b(u^*) + b(u^{(k)}) = -\frac{1}{\tau}r^{(k+1)}$$

Thus, we have

$$< (-A(u^*) + \frac{1}{\tau} I)(u^* - u^{(k+1)}), u^* - u^{(k+1)} > +$$
$$+ < (-A(u^*) + A(u^{(k)}))u^{(k+1)} - b(u^*) + b(u^{(k)}), u^* - u^{(k+1)} >$$
$$= < -\frac{1}{\tau}r^{(k+1)}, u^* - u^{(k+1)} >$$

Since $\hat{A}$ is the skew-symmetric part of $A(u)$, for any vector $z \in \mathbb{R}^d$ we have

$$< (\hat{A}(u) + \hat{\Delta} + \hat{D})z, z > = < (\hat{A}(u) + \hat{\Delta})z, z >$$

we can write

$$< -\hat{A}(u^*)(u^* - u^{(k+1)}), u^* - u^{(k+1)} > + < (-\hat{\Delta} + \frac{1}{\tau} I)(u^* - u^{(k+1)}), u^* - u^{(k+1)} > +$$
$$+ < (-\hat{A}(u^*) + \hat{A}(u^{(k)}))u^{(k+1)} - b(u^*) + b(u^{(k)}), u^* - u^{(k+1)} >$$
$$= < -\frac{1}{\tau}r^{(k+1)}, u^* - u^{(k+1)} >$$

Then

$$< -\frac{1}{\tau}r^{(k+1)}, u^* - u^{(k+1)} > \geq$$
$$< -\hat{A}(u^*)(u^* - u^{(k+1)}), u^* - u^{(k+1)} > + < (-\hat{\Delta} + \frac{1}{\tau} I)(u^* - u^{(k+1)}), u^* - u^{(k+1)} > -$$
$$- | < (-\hat{A}(u^*) + \hat{A}(u^{(k)}))u^{(k+1)} - b(u^*) + b(u^{(k)}), u^* - u^{(k+1)} > |$$

From lemmas 2 and 3 we obtain

$$< -\frac{1}{\tau}r^{(k+1)}, u^* - u^{(k+1)} > \geq$$
$$\geq h^2\sigma_{\min}(N) \sum_{i=1}^{M} \sum_{j=1}^{N} |(\nabla_x (u_{ij}^* - u_{ij}^{(k+1)}))^2 + |\nabla_y (u_{ij}^* - u_{ij}^{(k+1)})|^2| +$$
$$+ \sigma_{\min} \sum_{j=1}^{N} |u_{Nj}^* - u_{Nj}^{(k+1)}|^2 + \sigma_{\min} \sum_{i=1}^{N} |u_{iM}^* - u_{iM}^{(k+1)}|^2 +$$
$$+ (\sigma_{\min} + \frac{1}{\tau}) \|u^* - u^{(k+1)}\|^2 -$$
$$- \frac{\gamma_{k+1}}{\phi} \|u^* - u^{(k)}\|^2 - \frac{\gamma_{k+1}}{2h^2} \|u^* - u^{(k+1)}\|^2 -$$
$$- \frac{\gamma_{k+1}}{2h^2} \frac{\int_{\Omega} \int \sum_{i=1}^{N} |(\nabla_x (u_{ij}^* - u_{ij}^{(k+1)}))^2 + |\nabla_y (u_{ij}^* - u_{ij}^{(k+1)})|^2|}{2} \geq$$
$$\geq h^2\sigma_{\min}(N) \sum_{i=1}^{M} \sum_{j=1}^{N} |(\nabla_x (u_{ij}^* - u_{ij}^{(k+1)}))^2 + |\nabla_y (u_{ij}^* - u_{ij}^{(k+1)})|^2| +$$
$$+ (\sigma_{\min} + \frac{1}{\tau}) \|u^* - u^{(k+1)}\|^2 -$$
$$- \frac{\gamma_{k+1}}{\phi} \|u^* - u^{(k)}\|^2 - \frac{\gamma_{k+1}}{2h^2} \|u^* - u^{(k+1)}\|^2 -$$
$$- \frac{\gamma_{k+1}}{2h^2} \frac{\int_{\Omega} \int \sum_{i=1}^{N} |(\nabla_x (u_{ij}^* - u_{ij}^{(k+1)}))^2 + |\nabla_y (u_{ij}^* - u_{ij}^{(k+1)})|^2|}{2} \geq$$

14
Therefore the sequence 
\[ \frac{1}{\tau} \| r^{(k+1)} \|_h \leq | \frac{1}{\tau} r^{(k+1)}, u^* - u^{(k+1)} | \geq | \frac{1}{\tau} r^{(k+1)}, u^* - u^{(k+1)} | > 0 \]
and keeping into account of the stopping condition (20) we can write
\[
\frac{1}{\tau} \varepsilon_{k+1} \| u^* - u^{(k+1)} \|_h \geq h^2 (\sigma_{\min} - \frac{\beta_{k+1} \Gamma \phi}{\phi}) \sum_{i=1}^{N} \sum_{j=1}^{M} |(\nabla u(u_{ij} - u_{ij}^{(k+1)}))^2 + |\nabla u(u_{ij}^* - u_{ij}^{(k+1)})|^2 | + \\
+ (\alpha_{\min} + \frac{1}{\tau}) \| u^* - u^{(k+1)} \|_h^2 - \frac{\beta_{k+1} \Gamma \phi}{\phi} \| u^* - u^{(k+1)} \|_h^2
\]
Since \( \phi \) is an arbitrary positive number, we can choose \( \phi \) such that
\[
\sigma_{\min} = \frac{\beta_{k+1} \Gamma \phi}{2} = 0
\]
that is
\[
\phi = \frac{2\sigma_{\min}}{\beta_{k+1} \Gamma}
\]
Thus,
\[
\frac{1}{\tau} \varepsilon_{k+1} \| u^* - u^{(k+1)} \|_h \geq (\alpha_{\min} + \frac{1}{\tau} - \frac{\sigma_{\min}}{h^2}) \| u^* - u^{(k+1)} \|_h^2 - \frac{\beta_{k+1} \Gamma \phi}{2\sigma_{\min}} \| u^* - u^{(k)} \|_h^2
\]
Since the grid function \( \{ u_{ij}^{(k+1)} \} \) belongs to \( B \), and then it satisfies inequality (11), we have
\[
\| u^* - u^{(k+1)} \|_h \leq 2\rho
\]
then
\[
\frac{2\rho}{\tau} \varepsilon_{k+1} \geq (\alpha_{\min} + \frac{1}{\tau} - \frac{\sigma_{\min}}{h^2}) \| u^* - u^{(k+1)} \|_h^2 - \frac{\beta_{k+1} \Gamma \phi}{2\sigma_{\min}} \| u^* - u^{(k)} \|_h^2
\]
We assume that condition (22) holds.\(^4\) Then, keeping into account of the expression of \( \tilde{\beta}_k \), we have the inequality
\[
\| u^* - u^{(k+1)} \|_h^2 \leq \tilde{\gamma} \| u^* - u^{(k)} \|_h^2 + \tilde{\alpha} \varepsilon_{k+1}
\]
(25)
where
\[
a = \alpha_{\min} + \frac{1}{\tau} - \frac{\sigma_{\min}}{h^2}, \quad \tilde{\gamma} = \frac{(\beta + \varepsilon_{k+1} \beta_0)^2 \Gamma^2}{2\sigma_{\min} a}, \quad \tilde{\alpha} = \frac{2\rho}{\tau a}
\]
Now, if there exists an integer \( k_0 \) such that \( \tilde{\gamma} < 1 \) for all \( k \geq k_0 \),\(^5\) we can write formula (25) as
\[
\| u^* - u^{(k_0 + r)} \|_h^2 \leq \tilde{\gamma}^r \| u^* - u^{(k_0)} \|_h^2 + \tilde{\alpha} \sum_{j=1}^{r} \tilde{\gamma}^{r-j} \varepsilon_{k_0 + j}
\]
r = 1, 2, ..., and since \( \varepsilon_k \to 0 \) for \( k \to \infty \), it follows from the general Toeplitz Lemma (e.g., see [9, p. 399]) that
\[
\lim_{k \to \infty} \| u^* - u^{(k)} \|_h^2 = 0
\]
Therefore the sequence \( \{ u^{(k)} \} \) of approximate solutions of the lagged diffusivity method converges to the solution \( u^* \) of the system (10).

\(^4\) We observe that condition (22) is satisfied when condition (17) holds.

\(^5\) We remark that the condition \( \tilde{\gamma} < 1 \) is satisfied for \( k \) sufficiently large when condition (17) holds.
3 Numerical experiments

In this section we consider a numerical experimentation of the lagged diffusivity method for the solution of the nonlinear system generated by the \( \theta \)-method applied on the model problem (1) in a rectangular domain with Dirichlet boundary conditions. Indeed we solve with the lagged diffusivity procedure the nonlinear system

\[
(I - \tau A(u))u - w - b(u) = 0
\]

In the experiments, the vector solution \( u^* \) is prefixed and is composed by the values of prescribed functions \( u(x, y, t) \) defined on \( [a, b] \times [a, b] \times [0, \infty) \). In all the experiments we have \( a = 0 \) and \( b = 1 \). We choose different solution functions where the time value \( t \) is set equal to 1:

\[
\begin{align*}
\ u_1 & : u(x, y, t) = xyt \\
\ u_2 & : u(x, y, t) = ce^{-8((x-\frac{1}{2})^2+(y-\frac{1}{2})^2)t} \\
\ u_3 & : u(x, y, t) = \sum_{i=1}^{3} 15e^{-8((x-\xi_i)^2+(y-\eta_i)^2)t} \\
\ u_4 & : u(x, y, t) = 1 + \sin(3\pi x) \cos(3\pi y)t^2 \\
\ u_5 & : u(x, y, t) = (x + y)t \\
\ u_6 & : u(x, y, t) = (1 + x - y)^3t
\end{align*}
\]

For the function \( u_2 \) we define \( u_2a \) and \( u_2b \) for \( c = 1 \) and \( c = 15 \) respectively.

For the function \( u_3 \) we have \( \xi_1 = \eta_1 = 0.2, \xi_2 = 0.2, \eta_2 = 0.8 \) and \( \xi_3 = \eta_3 = 0.8 \).

The chosen functions \( \sigma(u) \) are:

\[
\begin{align*}
\sigma_1 & : \sigma(u) = 0.01 + 0.5u \\
\sigma_2 & : \sigma(u) = 0.01 + 0.5u^2 \\
\sigma_3 & : \sigma(u) = \frac{100}{1 + 500u}
\end{align*}
\]

Table 1 gathers the values of the minimum and the maximum values of \( \sigma(u^*) \) denoted with \( \sigma^* \) and \( \sigma^\star \) respectively. Here, the function \( \sigma(u^*) \) is computed on the values of \( u_{ij}^* \) as in (5) and the writing \( 1.00 \cdot 10^{-2} \) means \( 1.00 \cdot 10^{-2} \).

The vector \( w \) is computed as

\[
w = (I - \tau A(u^*))u^* - b(u^*)
\]

where the matrix \( A(u^*) \) and the vector \( b(u^*) \) have elements as in formula (5) with \( N = M \) and \( \alpha(x, y) = 0 \). We set \( \theta = 1 \) and \( \Delta t = 10^{-3} \) or \( \Delta t = 10^{-4} \). \( \tau = \theta\Delta t \). Here (see Table 1), the condition (22) holds.

At each iteration \( k \) of the diffusivity lagging procedure, we have to solve the linear system of order \( \mu = N \times N \):

\[
(I - \tau A(u^{(k)}))u = w + b(u^{(k)})
\]

We consider the case that the coefficient matrix is an M-matrix (\( \tilde{\nu} \neq 0 \)). In these experiments the iterative method of the Arithmetic Mean (AM) is used as linear solver in the form introduced in [12]. This method is convergent when the coefficient matrix is a nonsingular M-matrix. This form of the Arithmetic Mean method is a variant of the Alterating Group Explicit (AGE) decomposition introduced by Evans (e.g., see [1], [2]). The effectiveness of the Arithmetic Mean method, even on parallel architectures, is highlighted in the papers [4], [5], [6] and [12].

We call \( u^{(k+1)} \) the solution of the linear system above computed with \( j_k \) iterations of the Arithmetic Mean solver where the inner residual

\[
r^{(k+1)} = (I - \tau A(u^{(k)}))u^{(k+1)} - w - b(u^{(k)})
\]

satisfies the condition

\[
\|r^{(k+1)}\| \leq \varepsilon_{k+1}
\]
with \( \varepsilon_1 = 0.1 \| F(u(0)) \| \) and
\[
\varepsilon_{k+1} = 0.5 \varepsilon_k
\]  \hfill (26)
The vector \( F(u(0)) = (I - \tau A(u(0)))u(0) - w - b(u(0)) \) is the initial outer residual and it is called \( \text{res0} \).

The initial vector \( u(0) \) is taken as the null vector \( (u(0) = 0) \) or as the vector \( e \) which is the vector with all the components equal to 1 \( (u(0) = e) \).

The diffusivity lagging procedure has been implemented in a Fortran code with machine precision \( 2 \times 10^{-16} \) and stops when
\[
\varepsilon_{k+1} \leq \varepsilon
\]  \hfill (27)
with \( \varepsilon = 10^{-4} \).

An experiment evaluates the effectiveness of the lagged diffusivity method for different values of \( \varepsilon \); another experiment shows the behaviour of the method for different choices of \( \varepsilon_{k+1} \) respect to the one in (26) with \( \varepsilon = 10^{-4} \).

We call \( k^* \) the iteration of the diffusivity lagging procedure for which condition (27) is satisfied.

In the tables we report the number of iterations \( k \), the total number of iterations of the Arithmetic Mean method \( j_T \), the discrete \( l_2(R_h) \) norm of the error, \( \text{err} = \| u^* - u^{(k^*)} \|_h \), the Euclidean norm of the outer residual
\[
\text{res} = \| F(u^{(k^*)}) \| = \| (I - \tau A(u^{(k^*)}))u^{(k^*)} - w - b(u^{(k^*)}) \|
\]
and the Euclidean norm of the initial outer residual \( \text{res0} \).

The symbol \( \ast \) close to the value of \( \text{res} \) indicates that the behaviour of the norm of the outer residual \( \| F(u^{(k^*)}) \| \) is not monotone decreasing.

Furthermore, in the tables, \( n.c.(1) \) or \( n.c.(2) \) indicate that condition (7) is not satisfied at the first or at the second iteration of the diffusivity lagging procedure, respectively; analogously, \( \max.i.i.(1) \) or \( \max.i.i.(2) \) indicate that the maximum number of iterations of the inner iterative solver (8000) is reached at the first or at the second iteration of the diffusivity lagging procedure, respectively.

\[
\begin{array}{cccc}
  u(x,y,t) & \sigma_1 & \sigma_2 & \sigma_3 \\
  u1 & 1.00(-2)-5.10(-1) & 1.00(-2)-5.10(-1) & 1.99(-1)-100 \\
  u2a & 1.00(-2)-5.10(-1) & 1.00(-2)-5.10(-1) & 1.99(-1)-9.84 \\
  u2b & 1.00(-2)-7.51 & 1.00(-2)-112.51 & 1.33(-2)-7.22(-1) \\
  u5 & 1.00(-2)-8.52 & 1.00(-2)-145.03 & 1.17(-2)-1.50 \\
  u4 & 1.00(-2)-1.01 & 1.00(-2)-2.01 & 9.99(-2)-100 \\
  u5 & 1.00(-2)-1.01 & 1.00(-2)-2.01 & 9.99(-2)-100 \\
  u6 & 1.00(-2)-4.01 & 1.00(-2)-32.01 & 2.49(-2)-100 \\
\end{array}
\]

Table 1: Values of \( \sigma^* \) and \( \pi^* \).

From the numerical experiments we can draw the conclusions below.

- We observe that, since \( \varepsilon_{k+1} \) decreases, for \( k \) increasing, as (26) and the lagged diffusivity method stops at the iteration \( k^* \) when the criterium for \( \varepsilon_{k^*+1} \) in (27) is satisfied, we have
\[
\varepsilon_{k^*+1} = \frac{1}{2} \varepsilon_{k^*} = \frac{1}{2^2} \varepsilon_{k^*-1} = \ldots = \frac{1}{2^k} \varepsilon_1 \leq \varepsilon
\]
where we set \( \varepsilon_1 = 0.1 \| F(u(0)) \| \). Then
\[
k^* > \log_2 \left( \frac{\varepsilon_1}{\varepsilon} \right)
\]
In the experiments we obtain
\[
k^* = \lceil \log_2 \left( \frac{\varepsilon_1}{\varepsilon} \right) \rceil
\]
- We observe that in all the experiments with the rule (26) the outer residual \( \| F(u^{(k^*)}) \| \) has the same order of \( \varepsilon \) with an error in the discrete \( l_2(R_h) \) norm of order of \( h_{\varepsilon} \).
About the initial vectors, we can say that, generally, the null vector is a good choice, in terms of iterations on the total number 1534. A detection of that the null vector is not a good initial vector for the problems with $\sigma(u) = \sigma 3$, is the number of inner iterations at the first outer iteration. Indeed, for all the functions $u(x, y, t)$ and $\sigma(u) = \sigma 3$ we have better results when $u^{(0)} = e$.

A detection of that the null vector is not a good initial vector for the problems with $\sigma(u) = \sigma 3$, is the number of inner iterations at the first outer iteration. Indeed, for all the functions $u_{1},...u_{6}$, most of the inner iterations happen at the first iteration, $k = 1$; i.e., Table 3, $\sigma(u) = \sigma 3$, for $u_{2b}$ we have 1396 inner iterations at $k = 1$ on the total number 1429, or for $u_{3}$ we have 1484 inner iterations on the total number 1534.

The same phenomenon happens when $N$ is smaller; i.e., Table 4, $\sigma(u) = \sigma 3$, for $u_{2b}$ we have 354 (on 373) inner iterations at $k = 1$. In this last case, if we start from $u^{(0)} = e$ we have

$$
\begin{array}{|c|c|c|}
\hline
k^{*} & j_T & \text{err} \\
\hline
20 & 21 & 2.12(-11) \\
\hline
\end{array}
$$

Furthermore, we also observe that a good choice of the initial vector for a problem with time step $\tau = 10^{-3}$ does not imply that is still a good choice when the time step is reduced (see the problem with $\sigma = \sigma 1$ and $u(x, y, t) = u_{2b}$).

When the behaviour of the norm of the outer residual $F(u^{(k)})$ is not monotone decreasing (i.e., $\sigma(u) = \sigma 2$, $u(x, y, t) = u_{2b}, u_{3}, u_{6}, \tau = 10^{-3}$ and $N = 256$ or $N = 128$) we can have a large total number of inner iterations or the reaching of the maximum number of iterations (fixed at 8000) at an outer iteration. We suggest to change the initial vector to obtain a monotone decreasing of the norm of the outer residual that implies a reduction of the total number of inner iterations and avoids to reach the maximum number of inner iterations. Changing the initial vector, we have the following results for $N = 256$ (see Table 2):

<table>
<thead>
<tr>
<th>$u(x, y, t)$</th>
<th>$u^{(0)}$</th>
<th>$k^{*}$</th>
<th>$j_T$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$\text{res0}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{2b}$</td>
<td>10e</td>
<td>29</td>
<td>12706</td>
<td>2.70(-7)</td>
<td>9.50(-5)</td>
<td>292440.08</td>
</tr>
<tr>
<td>$u_{3}$</td>
<td>10e</td>
<td>28</td>
<td>26839</td>
<td>3.01(-7)</td>
<td>1.33(-4)</td>
<td>189726.62</td>
</tr>
<tr>
<td>$u_{6}$</td>
<td>4e</td>
<td>26</td>
<td>3669</td>
<td>2.67(-7)</td>
<td>1.05(-4)</td>
<td>35807.59</td>
</tr>
</tbody>
</table>

Table 2: Results with $\tilde{v}_1 = \tilde{v}_2 = 1$, $\tau = 10^{-3}$ and $u^{(0)} = 0$.

- From the experiments in Table 7, the choice for $\varepsilon_{k+1}$ as in (26) gives satisfactory results in terms of iterations of the inner solver, in terms of outer residual and of the error.

- About the initial vectors, we can say that, generally, the null vector is a good choice, in terms of total number of inner iterations, for $\sigma(u) = \sigma 1$ and $\sigma(u) = \sigma 2$; while for all the functions $u(x, y, t)$ and $\sigma(u) = \sigma 3$ we have better results when $u^{(0)} = e$.

<table>
<thead>
<tr>
<th>$\sigma(u) = \sigma 1$;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
</tr>
<tr>
<td>$u_{2a}$</td>
</tr>
<tr>
<td>$u_{2b}$</td>
</tr>
<tr>
<td>$u_{3}$</td>
</tr>
<tr>
<td>$u_{4}$</td>
</tr>
<tr>
<td>$u_{5}$</td>
</tr>
<tr>
<td>$u_{6}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\sigma(u) = \sigma 2$;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{2b}$</td>
</tr>
<tr>
<td>$u_{3}$</td>
</tr>
<tr>
<td>$\sigma(u) = \sigma 3$;</td>
</tr>
</tbody>
</table>

1. From the experiments in Table 7, the choice for $\varepsilon_{k+1}$ as in (26) gives satisfactory results in terms of iterations of the inner solver, in terms of outer residual and of the error.

2. About the initial vectors, we can say that, generally, the null vector is a good choice, in terms of total number of inner iterations, for $\sigma(u) = \sigma 1$ and $\sigma(u) = \sigma 2$; while for all the functions $u(x, y, t)$ and $\sigma(u) = \sigma 3$ we have better results when $u^{(0)} = e$.

Table 2: Results with $\tilde{v}_1 = \tilde{v}_2 = 1$, $\tau = 10^{-3}$ and $u^{(0)} = 0$. 
\( N = 256; \tilde{v}_1 = \tilde{v}_2 = 1; \tau = 10^{-4}; \)

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
u(x, y, t) & k^* & j_T & err & res & res0 \\
\hline
\hline
u1 & 17 & 33 & 4.22(-7) & 1.12(-4) & 89.44 \\
u2a & 17 & 37 & 6.66(-7) & 1.72(-4) & 113.39 \\
u2b & 21 & 545 & 5.83(-7) & 1.60(-4) & 202.62 \\
u3 & 23 & 676 & 5.40(-7) & 1.45(-4) & 620.36 \\
u4 & 19 & 74 & 3.89(-7) & 1.03(-4) & 310.31 \\
u5 & 19 & 70 & 3.75(-7) & 9.94(-5) & 304.43 \\
u6 & 21 & 216 & 3.80(-7) & 1.10(-4) & 1170.79 \\
\hline
\end{array}
\]

Table 3: Results with \( \tilde{v}_1 = \tilde{v}_2 = 1, \tau = 10^{-4} \) and \( u^{(0)} = 0. \)

and for \( N = 128 \) (see Table 4):

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
u(x, y, t) & u^{(0)} k^* & j_T & err & res & res0 \\
\hline
u2b & 10e & 26 & 3074 & 8.63(-7) & 1.36(-4) & 52462.80 \\
u3 & 10e & 26 & 6619 & 4.31(-7) & 9.58(-5) & 34067.73 \\
u4 & 4e & 23 & 857 & 7.64(-7) & 1.51(-4) & 6471.45 \\
u6 & 19 & 1515 & 3.16(-7) & 9.58(-5) & 297.96 \\
\hline
\end{array}
\]

- In the case of \( \sigma(u) = \sigma_3, \) \( u(x, y, t) = u4 \) and \( \tau = 10^{-4} \) (Table3), at the iteration \( k = 2 \) the minimum component of the solution \( u^{(2)} \) is equal to a small negative number that implies a large negative value for \( \sigma \) and the condition (7) can not be satisfied. This also happens for \( N = 128 \) (Table 4).

For example, in the case \( N = 256, \) changing the initial value \( u^{(0)} = 2 \cdot 10^{-2}, \) with \( \varepsilon_{k+1} = 0.8 \varepsilon_k, \) we have the result

\[
\begin{array}{|c|c|c|c|c|}
\hline
k^* & j_T & err & res & res0 \\
\hline
57 & 288 & 3.45(-7) & 1.09(-4) & 301.32 \\
\hline
\end{array}
\]

- From Table 2, we observe that the Arithmetic Mean method gives better performances when the ratio \( \Omega^*/\tilde{v}_1 (\tilde{v}_1 = \tilde{v}_2) \) is small, that is the coefficient matrix of the inner linear system is strongly asymmetric (see ([11])).

In Table 2, in the cases of \( \sigma(u) = \sigma_2 \) with \( u2b \) and \( \sigma(u) = \sigma_3 \) with \( u2b \) and \( u3 \) (also the cases in tables 4 and 5 with \( \sigma(u) = \sigma_2 \) and \( u2b \)) we have that the maximum number of inner iterations is reached.

In the case of \( \sigma(u) = \sigma_2 \) we have \( \sigma_{\text{max}}(u^{(1)}) \gg \sigma_{\text{max}}(u^*) \) while for \( \sigma(u) = \sigma_3 \) we have \( \sigma_{\text{max}}(u^{(0)}) \gg \sigma_{\text{max}}(u^*) \); for very large values of the diffusivity function, the Arithmetic Mean method is less efficient since the coefficient matrix is nearly symmetric.
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
$u(x,y,t)$ & $k^*$ & $j\tau$ & $\text{err}$ & $\text{res}$ & $\text{res}0$ \\
\hline
$N = 128; \bar{v}_1 = \bar{v}_2 = 1; \tau = 10^{-3}$; & & & & & \\
$\sigma(u) = \sigma 2$; & & & & & \\
$u1$ & 16 & 46 & 9.21(-7) & 1.38(-4) & 47.17 \\
u2a & 16 & 59 & 1.15(-6) & 1.58(-4) & 57.01 \\
u2b & $\max_{i,i}(2)$ & * & & & 1740.31 \\
u3 & 20 & 11202 & 6.27(-7) & 1.39(-4) & 49473.64 \\
u4 & 18 & 206 & 9.53(-7) & 1.48(-4) & 215.00 \\
u5 & 18 & 217 & 1.04(-6) & 1.59(-4) & 217.75 \\
u6 & 23 & 967 & 6.77(-7) & 1.33(-4) & 5694.88 \\
$\sigma(u) = \sigma 3$; & & & & & \\
u2b & 20 & 3010 & 8.27(-7) & 1.38(-4) & 872.17 \\
u3 & 21 & 3552 & 7.91(-7) & 1.38(-4) & 1555.36 \\
$N = 128; \bar{v}_1 = \bar{v}_2 = 1; \tau = 10^{-3}$; & & & & & \\
$\sigma(u) = \sigma 3$; & & & & & \\
u2b$ & 20 & 373 & 2.34(-10) & 3.44(-8) & 853.85 \\
u3 & 21 & 394 & 2.48(-8) & 3.81(-6) & 1530.99 \\
u4 & $\text{n.c.}(2)$ & & & & 146.38 \\
u5 & 18 & 395 & 5.02(-7) & 8.65(-5) & 141.33 \\
u6 & 19 & 495 & 5.07(-7) & 1.00(-4) & 273.30 \\
\hline
\end{tabular}
\caption{Results with $N = 128, \bar{v}_1 = \bar{v}_2 = 1$ and $u^{(0)} = 0$.}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{sigma1.png}
\caption{1}
\end{figure}

\section*{References}


\( N = 256; \tilde{v}_1 = \tilde{v}_2 = 1; \tau = 10^{-3}; \)

<table>
<thead>
<tr>
<th>( u(x, y, t) )</th>
<th>( k^2 )</th>
<th>( j_T )</th>
<th>( \text{err} )</th>
<th>( \text{res} )</th>
<th>( \text{res0} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma(u) = \sigma_1 )</td>
<td>| |</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 ( u_1 )</td>
<td>20</td>
<td>302</td>
<td>3.69(-7)</td>
<td>1.03(-4)</td>
<td>546.04</td>
</tr>
<tr>
<td>2 ( u_{2a} )</td>
<td>20</td>
<td>347</td>
<td>4.13(-7)</td>
<td>1.10(-4)</td>
<td>600.82</td>
</tr>
<tr>
<td>2 ( u_{2b} )</td>
<td>21</td>
<td>3287</td>
<td>5.37(-7)</td>
<td>1.63(-4)</td>
<td>1846.07</td>
</tr>
<tr>
<td>3 ( u_3 )</td>
<td>26</td>
<td>6066</td>
<td>4.76(-7)</td>
<td>1.47(-4)</td>
<td>50940.66</td>
</tr>
<tr>
<td>4 ( u_4 )</td>
<td>20</td>
<td>651</td>
<td>3.87(-7)</td>
<td>1.12(-4)</td>
<td>619.17</td>
</tr>
<tr>
<td>5 ( u_5 )</td>
<td>20</td>
<td>633</td>
<td>4.63(-7)</td>
<td>1.31(-4)</td>
<td>692.97</td>
</tr>
<tr>
<td>6 ( u_6 )</td>
<td>24</td>
<td>1631</td>
<td>3.37(-7)</td>
<td>1.10(-4)</td>
<td>9323.23</td>
</tr>
<tr>
<td>( \sigma(u) = \sigma_2 )</td>
<td>| |</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 ( u_2 )</td>
<td>19</td>
<td>249</td>
<td>4.51(-7)</td>
<td>1.32(-4)</td>
<td>362.86</td>
</tr>
<tr>
<td>2 ( u_{2a} )</td>
<td>19</td>
<td>311</td>
<td>5.12(-7)</td>
<td>1.40(-4)</td>
<td>375.94</td>
</tr>
<tr>
<td>2 ( u_{2b} )</td>
<td>21</td>
<td>3567</td>
<td>* *</td>
<td>3567.73</td>
<td></td>
</tr>
<tr>
<td>3 ( u_3 )</td>
<td>29</td>
<td>38837</td>
<td>2.36(-7)</td>
<td>1.06(-4)*</td>
<td>298830.38</td>
</tr>
<tr>
<td>4 ( u_4 )</td>
<td>20</td>
<td>950</td>
<td>4.52(-7)</td>
<td>1.40(-4)</td>
<td>816.18</td>
</tr>
<tr>
<td>5 ( u_5 )</td>
<td>20</td>
<td>932</td>
<td>5.45(-7)</td>
<td>1.67(-4)</td>
<td>881.87</td>
</tr>
<tr>
<td>6 ( u_6 )</td>
<td>26</td>
<td>4122</td>
<td>2.63(-7)</td>
<td>1.03(-4)*</td>
<td>35275.82</td>
</tr>
<tr>
<td>( \sigma(u) = \sigma_3 )</td>
<td>| |</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 ( u_{2b} )</td>
<td>21</td>
<td>176</td>
<td>4.00(-7)</td>
<td>1.36(-4)</td>
<td>1522.73</td>
</tr>
<tr>
<td>3 ( u_3 )</td>
<td>22</td>
<td>267</td>
<td>3.81(-7)</td>
<td>1.35(-4)</td>
<td>2946.38</td>
</tr>
</tbody>
</table>

Table 5: Results with \( \tilde{v}_1 = \tilde{v}_2 = 1, \tau = 10^{-3} \) and \( u^{(0)} = e \).

Figure 2:


\( N = 256; \bar{v}_1 = \bar{v}_2 = 1; \tau = 10^{-4}; \)

<table>
<thead>
<tr>
<th>( u(x, y, t) )</th>
<th>( k^2 )</th>
<th>( jr )</th>
<th>( err )</th>
<th>( res )</th>
<th>( res\theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma(u) = \sigma 1; )</td>
<td>( \sigma(u) = \sigma 2; )</td>
<td>( \sigma(u) = \sigma 3; )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( u_1 )</td>
<td>18</td>
<td>37</td>
<td>5.17(-7)</td>
<td>1.34(-4)</td>
<td>211.75</td>
</tr>
<tr>
<td>( u_{2a} )</td>
<td>18</td>
<td>39</td>
<td>5.64(-7)</td>
<td>1.45(-4)</td>
<td>193.32</td>
</tr>
<tr>
<td>( u_{2b} )</td>
<td>21</td>
<td>578</td>
<td>5.11(-7)</td>
<td>1.40(-4)</td>
<td>1513.47</td>
</tr>
<tr>
<td>( u_3 )</td>
<td>23</td>
<td>737</td>
<td>5.23(-7)</td>
<td>1.40(-4)</td>
<td>6026.72</td>
</tr>
<tr>
<td>( u_4 )</td>
<td>18</td>
<td>80</td>
<td>3.89(-7)</td>
<td>1.03(-4)</td>
<td>147.91</td>
</tr>
<tr>
<td>( u_5 )</td>
<td>18</td>
<td>76</td>
<td>3.73(-7)</td>
<td>9.89(-5)</td>
<td>132.60</td>
</tr>
<tr>
<td>( u_6 )</td>
<td>21</td>
<td>240</td>
<td>3.46(-7)</td>
<td>1.00(-4)</td>
<td>1074.69</td>
</tr>
<tr>
<td>( u_1 )</td>
<td>18</td>
<td>33</td>
<td>4.50(-7)</td>
<td>1.19(-4)</td>
<td>205.39</td>
</tr>
<tr>
<td>( u_{2a} )</td>
<td>18</td>
<td>36</td>
<td>5.18(-7)</td>
<td>1.34(-4)</td>
<td>184.12</td>
</tr>
<tr>
<td>( u_{2b} )</td>
<td>21</td>
<td>5017</td>
<td>4.12(-7)</td>
<td>1.30(-4)</td>
<td>1610.41</td>
</tr>
<tr>
<td>( u_3 )</td>
<td>25</td>
<td>7615</td>
<td>5.02(-7)</td>
<td>1.68(-4)*</td>
<td>30266.34</td>
</tr>
<tr>
<td>( u_4 )</td>
<td>18</td>
<td>132</td>
<td>4.23(-7)</td>
<td>1.17(-4)</td>
<td>158.24</td>
</tr>
<tr>
<td>( u_5 )</td>
<td>18</td>
<td>121</td>
<td>3.82(-7)</td>
<td>1.05(-4)</td>
<td>144.41</td>
</tr>
<tr>
<td>( u_6 )</td>
<td>22</td>
<td>910</td>
<td>4.89(-7)</td>
<td>1.69(-4)</td>
<td>3611.96</td>
</tr>
<tr>
<td>( u_{2b} )</td>
<td>21</td>
<td>31</td>
<td>4.25(-7)</td>
<td>1.27(-4)</td>
<td>1499.84</td>
</tr>
<tr>
<td>( u_3 )</td>
<td>22</td>
<td>48</td>
<td>3.98(-7)</td>
<td>1.24(-4)</td>
<td>2825.73</td>
</tr>
<tr>
<td>( u_4 )</td>
<td>n.c.</td>
<td>135.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( u_5 )</td>
<td>17</td>
<td>102</td>
<td>4.76(-7)</td>
<td>1.66(-4)</td>
<td>113.34</td>
</tr>
<tr>
<td>( u_6 )</td>
<td>19</td>
<td>604</td>
<td>3.99(-7)</td>
<td>1.58(-4)</td>
<td>416.74</td>
</tr>
</tbody>
</table>

Table 6: Results with \( \bar{v}_1 = \bar{v}_2 = 1, \tau = 10^{-4} \) and \( u^{(0)} = e \).

Figure 3:

Figure 4:

Figure 5:
\[ N = 256; \ u = u_1; \ \sigma(u) = \sigma_1; \ \tilde{v}_1 = \tilde{v}_2 = 1; \ \tau = 10^{-3}; \ \text{res}0 = 205.61; \]

<table>
<thead>
<tr>
<th>( \varepsilon )</th>
<th>( k^* )</th>
<th>( j_T )</th>
<th>( \text{err} )</th>
<th>( \text{res} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^{-3}</td>
<td>15</td>
<td>210</td>
<td>4.21(-6)</td>
<td>1.20(-3)</td>
</tr>
<tr>
<td>10^{-4}</td>
<td>18</td>
<td>260</td>
<td>5.26(-7)</td>
<td>1.51(-4)</td>
</tr>
<tr>
<td>10^{-5}</td>
<td>21</td>
<td>310</td>
<td>6.59(-8)</td>
<td>1.90(-5)</td>
</tr>
<tr>
<td>10^{-6}</td>
<td>25</td>
<td>377</td>
<td>4.09(-9)</td>
<td>1.18(-6)</td>
</tr>
</tbody>
</table>

\[ N = 256; \ u = u_4; \ \sigma(u) = \sigma_2; \ \tilde{v}_1 = \tilde{v}_2 = 1; \ \tau = 10^{-3}; \ \text{res}0 = 842.77; \]

<table>
<thead>
<tr>
<th>( \varepsilon )</th>
<th>( k^* )</th>
<th>( j_T )</th>
<th>( \text{err} )</th>
<th>( \text{res} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^{-3}</td>
<td>17</td>
<td>717</td>
<td>3.76(-6)</td>
<td>1.16(-3)</td>
</tr>
<tr>
<td>10^{-4}</td>
<td>20</td>
<td>875</td>
<td>4.66(-7)</td>
<td>1.44(-4)</td>
</tr>
<tr>
<td>10^{-5}</td>
<td>24</td>
<td>1085</td>
<td>2.91(-8)</td>
<td>9.03(-6)</td>
</tr>
<tr>
<td>10^{-6}</td>
<td>27</td>
<td>1242</td>
<td>3.66(-9)</td>
<td>1.13(-6)</td>
</tr>
</tbody>
</table>

\[ N = 256; \ u = u_1; \ \sigma(u) = \sigma_1; \ \tilde{v}_1 = \tilde{v}_2 = 1; \ \tau = 10^{-3}; \ \text{res}0 = 205.61; \]

<table>
<thead>
<tr>
<th>( \varepsilon_{k+1} )</th>
<th>( k^* )</th>
<th>( j_T )</th>
<th>( \text{err} )</th>
<th>( \text{res} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.7 \varepsilon_k )</td>
<td>35</td>
<td>268</td>
<td>3.85(-7)</td>
<td>1.10(-4)</td>
</tr>
<tr>
<td>( 0.5 \varepsilon_k )</td>
<td>18</td>
<td>260</td>
<td>5.26(-7)</td>
<td>1.51(-4)</td>
</tr>
<tr>
<td>( 0.1 \varepsilon_k )</td>
<td>6</td>
<td>250</td>
<td>7.08(-7)</td>
<td>2.01(-4)</td>
</tr>
<tr>
<td>( 0.05 \varepsilon_k )</td>
<td>5</td>
<td>260</td>
<td>4.36(-7)</td>
<td>1.23(-4)</td>
</tr>
<tr>
<td>( 0.01 \varepsilon_k )</td>
<td>3</td>
<td>189</td>
<td>7.67(-6)</td>
<td>2.31(-3)</td>
</tr>
</tbody>
</table>

\[ N = 256; \ u = u_6; \ \sigma(u) = \sigma_1; \ \tilde{v}_1 = \tilde{v}_2 = 1; \ \tau = 10^{-3}; \ \text{res}0 = 9617.82; \]

<table>
<thead>
<tr>
<th>( \varepsilon_{k+1} )</th>
<th>( k^* )</th>
<th>( j_T )</th>
<th>( \text{err} )</th>
<th>( \text{res} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.7 \varepsilon_k )</td>
<td>46</td>
<td>1625</td>
<td>3.10(-7)</td>
<td>1.02(-4)</td>
</tr>
<tr>
<td>( 0.5 \varepsilon_k )</td>
<td>24</td>
<td>1097</td>
<td>3.46(-7)</td>
<td>1.33(-4)</td>
</tr>
<tr>
<td>( 0.1 \varepsilon_k )</td>
<td>7</td>
<td>1327</td>
<td>3.06(-6)</td>
<td>9.34(-4)</td>
</tr>
<tr>
<td>( 0.05 \varepsilon_k )</td>
<td>6</td>
<td>1408</td>
<td>1.02(-6)</td>
<td>3.09(-4)</td>
</tr>
<tr>
<td>( 0.01 \varepsilon_k )</td>
<td>4</td>
<td>1384</td>
<td>4.12(-5)</td>
<td>2.46(-2)</td>
</tr>
</tbody>
</table>

\[ N = 256; \ u = u_4; \ \sigma(u) = \sigma_2; \ \tilde{v}_1 = \tilde{v}_2 = 1; \ \tau = 10^{-3}; \ \text{res}0 = 842.77; \]

<table>
<thead>
<tr>
<th>( \varepsilon_{k+1} )</th>
<th>( k^* )</th>
<th>( j_T )</th>
<th>( \text{err} )</th>
<th>( \text{res} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.7 \varepsilon_k )</td>
<td>39</td>
<td>925</td>
<td>3.30(-7)</td>
<td>1.04(-4)</td>
</tr>
<tr>
<td>( 0.5 \varepsilon_k )</td>
<td>20</td>
<td>875</td>
<td>4.66(-7)</td>
<td>1.44(-4)</td>
</tr>
<tr>
<td>( 0.1 \varepsilon_k )</td>
<td>6</td>
<td>627</td>
<td>2.58(-6)</td>
<td>1.14(-3)</td>
</tr>
<tr>
<td>( 0.05 \varepsilon_k )</td>
<td>5</td>
<td>670</td>
<td>2.18(-5)</td>
<td>1.26(-2)</td>
</tr>
<tr>
<td>( 0.01 \varepsilon_k )</td>
<td>3</td>
<td>772</td>
<td>1.83(-3)</td>
<td>1.01</td>
</tr>
</tbody>
</table>

\[ N = 256; \ u = u_6; \ \sigma(u) = \sigma_2; \ \tilde{v}_1 = \tilde{v}_2 = 1; \ \tau = 10^{-3}; \ \text{res}0 = 31481.01; \]

<table>
<thead>
<tr>
<th>( \varepsilon_{k+1} )</th>
<th>( k^* )</th>
<th>( j_T )</th>
<th>( \text{err} )</th>
<th>( \text{res} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.7 \varepsilon_k )</td>
<td>49</td>
<td>4146</td>
<td>2.78(-7)</td>
<td>1.14(-4)*</td>
</tr>
<tr>
<td>( 0.5 \varepsilon_k )</td>
<td>25</td>
<td>3960</td>
<td>4.70(-7)</td>
<td>1.85(-4)*</td>
</tr>
<tr>
<td>( 0.1 \varepsilon_k )</td>
<td>8</td>
<td>4930</td>
<td>1.75(-5)</td>
<td>1.42(-2)*</td>
</tr>
<tr>
<td>( 0.05 \varepsilon_k )</td>
<td>6</td>
<td>4827</td>
<td>3.16(-4)</td>
<td>0.27*</td>
</tr>
<tr>
<td>( 0.01 \varepsilon_k )</td>
<td>4</td>
<td>3800</td>
<td>4.70(-3)</td>
<td>4.63*</td>
</tr>
</tbody>
</table>

Table 7: Results for different \( \varepsilon \) and \( \varepsilon_{k+1} \) (\( u(0) = 0 \)).
Figure 6:

\[ \sigma_2(u5) \quad \sigma_2(u6) \]

Figure 7:

\[ \sigma_3(u1) \quad \sigma_3(u2a) \quad \sigma_3(u2b) \]

Figure 8:

\[ \sigma_3(u3) \quad \sigma_3(u4) \]
Figure 9: