Explicit stable methods for second order parabolic systems*

NEDŽAD LIMIĆ† AND MLADEN ROGINA‡

Abstract. *We show that it is possible to construct stable, explicit finite difference approximations for the classical solution of the initial value problem for the parabolic systems of the form* $\partial_t u = A(t,\mathbf{x})u + f$ *on* $\mathbb{R}^d$*, where* $A(t,\mathbf{x}) = \sum_{ij} a_{ij}(t,\mathbf{x})\partial_i\partial_j + \sum_i b_i(t,\mathbf{x})\partial_i + c(t,\mathbf{x})$*. The numerical scheme relies on an approximation of the elliptic operator* $A(t,\mathbf{x})$ *on an equidistant mesh by matrices that possess structure of a generator of Markov jump process. In the case of* $\mathbb{R}^2$ *scaling of second difference operators can be applied to get the necessary structure of approximations, while in the case of* $\mathbb{R}^d$*,* $d > 2$*, rotations at grid-knots are performed in order to get the mentioned structure. Numerical experiments illustrate the theory.*

Key words: *parabolic systems, finite difference schemes, Markov chains*

AMS subject classifications: 35J15, 35K45, 65L12, 60J25

Received February 22, 2000 Accepted July 17, 2000

1. Introduction

We derive and analyze a class of finite-difference methods for the L_∞-theory of 2^{nd} order parabolic system. The crucial step of construction is the space discretization of an elliptic operator, $A(\mathbf{x})$, on an equidistant grid so that its matrix approximations, A_n, have the structure of a generator of Markov jump process (MJP). Then we approximate the parabolic system by systems of ODE, for which explicit and stable finite-difference methods are constructed and analyzed.

Throughout our exposition the case of parabolic systems on $\mathbb{R}^d$ is considered although methods are applicable to problems on bounded domains as well.

In the course of discretization of $A(\mathbf{x})$ we encounter difficulties caused by too large values of $|a_{ij}|, i \neq j$, compared to the values of $|a_{ii}|, |a_{jj}|$. For bounded domains this problem is recognized and solved in [MW], where the corresponding

*This work was partially supported by grant 037011 from the Ministry of Science and Technology, Republic of Croatia

†Department of Mathematics, University of Zagreb, Bijenička c. 30, HR-10 000 Zagreb, Croatia, e-mail: `nlimic@math.hr`

‡Department of Mathematics, University of Zagreb, Bijenička c. 30, HR-10 000 Zagreb, Croatia, e-mail: `rogina@math.hr`

approximations are called positive. We prefer the terminology MJP in order to distinguish a similar structure of approximations for L_1-theory of elliptic operators in a divergence form. Results of *Section 2.* are generalizations of results in [MW]. The uniform continuity of coefficients enables the existence of a finite number of coordinate transformations in $\mathbb{R}^d$ ensuring the MJP-structure of approximations. Approximations A_n of *Section 2.* are consistent with the differential operator. The main result about approximations for general $\mathbb{R}^d$ is formulated in *Theorem 1.* Our analysis in *Section 2.* finishes with the problem of reducibility of approximations. Another construction of approximations for $\mathbb{R}^2$ is presented in *Section 3.* possessing a new feature. Contrary to approximating operators of *Section 2.*, the approximations A_n of *Section 3.* are not consistent with the differential operator, yet the convergence is of the same order as for approximations of *Section 2.* There is an essential difference between the two constructed approximations. The numerical neighbourhood of a grid-knot $(x, y) \in \mathbb{R}^2$ contains only surrounding grid knots $(x \pm h, y \pm h)$. This is an attractive feature which is not shared by other approaches in $\mathbb{R}^2$ (see for instance [BH]). The initial value problem for parabolic system is considered in *Section 4.* A class of explicit and stable methods, of any convergence order, for ODE with MJP-structure is constructed and analyzed. Methods are not linear [AS]. The efficiency of these methods is demonstrated by an example.

2. Numerical scheme for elliptic operators possessing MJP-structure

The linear space of continuous and uniformly bounded functions on $\mathbb{R}^d$ is denoted by $C(\mathbb{R}^d)$. The linear space of functions having partial derivatives up to the order m, and each partial derivative continuous and uniformly bounded, is denoted by $C^m(\mathbb{R}^d)$. By convention $C^0(\mathbb{R}^d) = C(\mathbb{R}^d)$. The linear subspace of $C^m(\mathbb{R}^d)$, for which the m^{th} partial derivatives are uniformly α-Hölder continuous, $\alpha \in (0, 1)$, is denoted by $C^{m+\alpha}(\mathbb{R}^d)$. All defined spaces are Banach spaces if supplied with the norm $\| \cdot \|_{k+\alpha}$ (see [LU1]). The Banach space $\dot{C}^{m+\alpha}(\mathbb{R}^d)$ is the closed subspace of $C^{m+\alpha}(\mathbb{R}^d)$, spanned by elements of $C^{m+\alpha}(\mathbb{R}^d)$ with compact supports.

Let the functions $a_{ij} = a_{ji}$, b_i $(i, j = 1, 2, \ldots, d)$ and c satisfy the following conditions:

$$|b_i|, |c| \quad \leq M, \quad c \leq 0, \quad a_{ij}, b_i, c \quad \text{uniformly continuous on } \mathbb{R}^d, \tag{1a}$$

$$M\,|z|_2^2 \geq \sum_{i,j=1}^{d} a_{ij}(\mathbf{x}) z_i \bar{z}_j \geq \mu |z|_2^2, \quad \mu > 0, \quad \mathbf{x} \in \mathbb{R}^d, \tag{1b}$$

where z_i are complex numbers, and $|z|_2$ the corresponding l_2-norm. The associated elliptic differential operator

$$A(\mathbf{x}) = \sum_{i,j=1}^{d} a_{ij}(\mathbf{x}) \partial_i \partial_j + \sum_{i=1}^{d} b_i(\mathbf{x}) \partial_i + c(\mathbf{x}), \tag{2}$$

maps $\dot{C}^{2+\alpha}(\mathbb{R}^d)$ to $\dot{C}(\mathbb{R}^d)$.

The orthogonal coordinate system in $\mathbb{R}^d$ is determined by unit vectors $\mathbf{e}_i$ of the canonical basis. For each $n \in \mathbb{N}$, points $\mathbf{x} = h \sum_{l=1}^{d} k_l \mathbf{e}_l, h = 2^{-n}$, $k_l \in \mathbb{Z}$, define a numerical grid G_n on $\mathbb{R}^d$. Let $E_2 = \dot{C}^{2+\alpha}(\mathbb{R}^d)$, and $E_2(n)$ be the closed subspace of all elements in E_2 vanishing at G_n. If the space $F_2 = E_2/E_2(n)$ is endowed with a quotient norm, it becomes isomorphic to a subspace of l_∞. Analogously, the spaces $E_0 = \dot{C}(\mathbb{R}^d)$, $E_0(n)$ and F_0 are defined. Because of the mentioned isomorphism, we identify F_2 with a subspace of F_0. The natural embedding from E_k into $F_k, k = 2, 0$, is denoted by $\phi_k(n)$. The image F_k of $\phi(n)$ is sometimes denoted by $F_k(G_n)$ to avoid ambiguity.

We shall say that an element $u \in E_k$ is approximated *with the order* $\beta > 0$ by elements $u_n \in F_k$ if there exists an n-independent positive number κ such that $\| \phi_k(n)u - u_n \|_k < \kappa h^\beta$. In this case, u_n *factor converges* to u with the order β. An operator $A \in \mathbf{L}(E_2, E_0)$ is *approximated by the operators* $A_n \in \mathbf{L}(F_2, F_0)$ with the order of approximation β, if there exists an n-independent positive number κ such that

$$\| \phi_0(n)A - A_n\phi_2(n) \|_{\mathbf{L}(E_2,F_0)} \leq \kappa h^\beta. \tag{3}$$

The approximation is called *stable* if there exists an n-independent $\rho > 0$ such that $\| (A_n)^{-1} \|_{\mathbf{L}(F_0,F_2)} \leq \rho$. Because of $F_2 \subset F_0 \subset l_\infty$, the operator A_n is considered as an infinite order matrix with matrix elements $(A_n)_{ij}$. Each row and/or column of this matrix is associated with a pair of grid-knots of G_n, and consequently, each matrix element $(A_n)_{ij}$ is associated with certain grid-knots $\mathbf{x}_i, \mathbf{x}_j \in G_n$.

For $\operatorname{Re}\lambda > 0$, solutions u and u_n of respective equations,

$$\begin{array}{rcl} (\lambda I - A)u &=& f, \\ (\lambda I_n - A_n)u_n &=& f_n, \quad n \in \mathbb{N}, \end{array} \tag{4}$$

will be compared.

For an elliptic operator A satisfying (1), an approximation $A_n \in \mathbf{L}(F_2, F_0)$ is said to possess an MJP-structure if A_n are generators of (regular or irregular) Markov jump processes, *i.e.* matrix elements a_{ij} of A_n indexed by an index set J satisfy:

(a) $a_{ii}(t) \leq 0$, for all $i \in J, t \geq 0$,

(b) $a_{ij}(t) \geq 0$, for all $i, j \in J$, $i \neq j$, $t \geq 0$,

(c) $a_i(t) = \sum_j h_{ij}(t) \leq 0$, for all $i \in J, t \geq 0$.

The property is assumed to be valid for sufficiently large natural n, for which the basic results of the following theorem hold:

Theorem 1. *Let A be defined by (1) and (2). Then there exist approximations $A_n \in \mathbf{L}(F_2, F_0)$ such that:*

(i) *each A_n possesses an MJP-structure and approximates A with the order α;*

(ii) *there exists an n-independent $\theta_A > 0$ such that $(A_n)_{ij} = 0$ if the associated pair of grid-knots, $\mathbf{x}_i, \mathbf{x}_j$, fulfills the condition $|\mathbf{x}_i - \mathbf{x}_j| > \theta_A h$;*

(iii) *equations $(\lambda I - A_n)u_n = f_n$, $\operatorname{Re}\lambda > 0$, have unique solutions in F_0 satisfying $\| u_n \|_\infty < \| f_n \|_\infty / \operatorname{Re}\lambda$;*

(iv) *if f_n factor converges to f with the order α, then the sequence of solutions u_n of (4) also factor converges to u with the order α;*

(v) *if locally at $\mathbf{x} \in G_n$ p is a second degree polynomial, then $(\phi_0(n)A - A_n\phi_2(n))p(\mathbf{x}) = 0$.*

Proof. The statement (iii) is a consequence of the MJP-structure of A_n, thus ensuring the stability of operators $\lambda I - A_n$. The statement (iv) follows from (i), (iii) and the following result [KR1]: Let $A \in \mathbf{L}(E_2, E_0)$, and $A_n \in \mathbf{L}(F_2, F_0)$ be approximations of A with the order α. Let $f \in E_0$, and let f_n be approximations of f with the order α. Then, if u and u_n are unique solutions of (4) and $\lambda I - A_n$ are stable, it follows that u_n factor converges to u with the order α.

To prove (i), let the approximation A_n of A have the property described in (v). Then, by using standard methods of numerical analysis for PDE [RM], expression (3) can easily be derived with $\beta = \alpha$. To finish the proof we have therefore to construct a sequence A_n possessing the mentioned property and the MJP-structure. Our construction of the numerical approximations of operator A at a grid-knot $\mathbf{x}$ depends on the value $\omega(\mathbf{e}, \mathbf{x})$:

$$\omega(\mathbf{e}, \mathbf{x}) = \max_{i \neq j} \quad \frac{\max\left\{ \sum_{r \neq i} |a_{ir}(\mathbf{x})|, \sum_{r \neq j} |a_{jr}(\mathbf{x})| \right\}}{\min\{a_{ii}(\mathbf{x}),\, a_{jj}(\mathbf{x})\}}, \tag{5}$$

where the symbol $\mathbf{e}$ reminds of the fact that $\omega(\mathbf{e}, \mathbf{x})$ is calculated with respect to the original coordinate system $\{\mathbf{e_i}\}_1^n$. The construction is simple in case where $\rho \in (0,1)$ exists, such that

$$\omega(\mathbf{e}, \mathbf{x}) \leq 1 - \rho. \tag{6}$$

The partial differential operators of the first and second orders are approximated in the usual way:

$$\partial_i f(\mathbf{x}) \to \Box_i f(\mathbf{x}) \quad = \quad \frac{1}{2h}\left[f(\mathbf{x} + \mathbf{e}_i h) - f(\mathbf{x} - \mathbf{e}_i h)\right], \tag{7a}$$

$$\partial_i^2 f(\mathbf{x}) \to \Box_{ii} f(\mathbf{x}) \quad = \quad \frac{1}{h^2}\left[f(\mathbf{x} + \mathbf{e}_i h) - 2f(\mathbf{x}) + f(\mathbf{x} - \mathbf{e}_i h)\right]. \tag{7b}$$

The mixed partial differential operator can be approximated by one of the following four possibilities:

$$\partial_i \partial_j f(\mathbf{x}) \to \Box_{ij} f(\mathbf{x}) \ = \ \frac{1}{h^2} \begin{cases} f(\mathbf{x} \pm \mathbf{e}_i h \pm \mathbf{e}_j h) - f(\mathbf{x} \pm \mathbf{e}_i h) - f(\mathbf{x} \pm \mathbf{e}_j h) + f(\mathbf{x}), \\ -f(\mathbf{x} \pm \mathbf{e}_i h \mp \mathbf{x}_j h) + f(\mathbf{x} \pm \mathbf{e}_i h) + f(\mathbf{x} \mp \mathbf{e}_j h) - f(\mathbf{x}). \end{cases} \tag{8}$$

In this way, the quadratic operator $a_{ii}(\partial_i)^2 + a_{jj}(\partial_j)^2$ is approximated by the matrix $a_{ii}\Box_{ii} + a_{jj}\Box_{jj}$ with negative diagonal elements and non-negative off-diagonal elements. If $a_{ij} \geq 0$, then $a_{ij}\partial_i\partial_j$ is approximated by the half sum of the first two possibilities, otherwise by the half sum of the second two possibilities. Because of (6), the obtained approximation $\sum_{ij} a_{ij}\Box_{ij}$, is the matrix with the MJP-structure. The approximation of $\sum b_i\partial_i$ by the difference operator $\sum b_i\Box_i$ can violate the MJP-structure only for larger values of h. Hence, if (6) is valid for all $\mathbf{x} \in \mathbb{R}^d$, the approximations A_n the have MJP-structure for large values of n. The statement (ii) is obviously fulfilled with $\theta_A = \sqrt{d}$. The statement (v) follows directly from the constructed approximations $\Box_i f$ and $\Box_{ij} f$. In case of $d = 3$, and nonnegative mixed coefficients, $a_{ij} \geq 0, i \neq j$, nontrivial matrix elements corresponding to the grid knot (k, l, m) have the well known form:

$$\begin{array}{lcl}
A_{klm\ klm} & = & -2h^{-2}\big[a_{11} + a_{22} + a_{33} - |a_{12}| - |a_{23}| - |a_{31}|\big] \\
A_{klm\ k\pm 1lm} & = & h^{-2}[a_{11} - |a_{12}| - |a_{31}|] \\
A_{klm\ kl\pm 1m} & = & h^{-2}[a_{22} - |a_{12}| - |a_{23}|] \\
A_{klm\ klm\pm 1} & = & h^{-2}[a_{33} - |a_{23}| - |a_{31}|] \\
A_{klm\ k\pm 1l\pm 1m} & = & h^{-2}|a_{12}| \quad (= A_{klm\ k\pm 1l\mp 1m}) \\
A_{klm\ kl\pm 1m\pm 1} & = & h^{-2}|a_{23}| \quad (= A_{klm\ kl\pm 1m\mp 1}) \\
A_{klm\ k\pm 1lm\pm 1} & = & h^{-2}|a_{31}| \quad (= A_{klm\ k\pm 1lm\mp 1})
\end{array}$$

For negative mixed coefficients at the considered grid knot nontrivial matrix elements are given in brackets.

In the case that (6) is not valid at a grid-knot $\mathbf{x}$, the previous construction of numerical scheme does not yield an A_n with the MJP-structure. Still, the previous construction can be utilized upon applying rotations to grid-knots in which (6) is violated. Auxiliary results are given in *Lemma 1.–Lemma 4.* from which the statement (i) of *Theorem 1.* obviously follows. $\Box$

The original orthogonal coordinate system in $\mathbb{R}^d$ defined by the unit vectors $\mathbf{e}_i$ of the canonical basis shall be henceforth referred to as an $\mathbf{e}$-system; two such systems, defined by two sets of unit vectors, shall be assumed to have the same origin at $\mathbf{x} = \mathbf{0}$. We shall translate the origin to the considered grid-knot $\mathbf{x}$ and perform the transformation in translated systems, thus simplifying the expressions calculated at the considered grid-knot. The $\mathbf{e}$-system is used to define grids G_n. To each $\mathbf{x} \in \mathbb{R}^d$ there corresponds an orthogonal transformation $O(\mathbf{x})$ transforming $\mathbf{e}$-system into $\mathbf{g}$-system in which the diffusion tensor at $\mathbf{x}$ has a diagonal form. Diagonal elements have values in the interval $[\mu, M]$. The new axis are defined by elements $s(\mathbf{x}) = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_d\} \in (\partial B)^d$, where ∂B is the unit spherical surface in $\mathbb{R}^d$ centered at $\mathbf{x}$. If $\mathbf{f}_i$, i = 1,2, ... ,d, are linearly independent, and $|\mathbf{f}_i|_2 = 1$, where $|\cdot|_2$ is the l_2-norm in $\mathbb{R}^d$, then the elements $\mathbf{f}_i$ define a coordinate system called an $\mathbf{f}$-system.

The following statement is implied by the fact that the (central projection) image of G_n onto ∂B is a dense set in ∂B.

Lemma 1. *Let the diffusion tensor at* $\mathbf{x}$ *be diagonal in the orthogonal* $\mathbf{g}$*-system, and let there exist at least one axis which does not cross grid-knots of* G_n*. Then for each* $\varepsilon_T > 0$ *there exists an* $m(\varepsilon_T) \in \mathbb{N}$*, and*

(i) *a coordinate system (not necessarily orthogonal) with unit vectors* $\mathbf{f}_i$*;*

(ii) *a cube* C *of the edge* $2m(\varepsilon_T)$*, centered at* $\mathbf{x}$*, in the* $\mathbf{e}$*-system,*

such that the axes of the $\mathbf{f}$*-system intersect grid-knots in* C*, and satisfy the following inequalities:*

$$\mathbf{f}_i \cdot \mathbf{g}_i \geq 1 - \varepsilon_T, \quad i = 1, 2, \ldots, d. \tag{9}$$

The coefficients $a_{ij}(\cdot), b_i(\cdot)$ are defined in the **e**-system. In the **g**-system they have different expressions and, consequently, they are denoted by $a_{ij}(\mathbf{g},\cdot)$ and $b_i(\mathbf{g},\cdot)$, respectively. Thus, by definition, $a_{ij}(\mathbf{x}) = a_{ij}(\mathbf{e},\mathbf{x})$. The matrix $a(\mathbf{x}) = \{a_{ij}(\mathbf{x})\}_{11}^{dd}$ is considered as a linear operator on a d-dimensional unitary space of columns. Its l_2-norm is denoted by $|a(\mathbf{x})|_2$. We have $|a(\mathbf{x})|_2 = |a(\mathbf{g},\mathbf{x})|_2 \leq Md$. Let $\varepsilon_a > 0$ be given and $\delta > 0$ such that $|a(\mathbf{x}) - a(\mathbf{y})|_2 < \varepsilon_a$ if $|\mathbf{x}-\mathbf{y}| < \delta$. Then

$$|\omega(\mathbf{e},\mathbf{x}) - \omega(\mathbf{e},\mathbf{y})| < \frac{2M}{\mu^2}\varepsilon_a,$$

holds whenever $|\mathbf{x}-\mathbf{y}| < \delta$. The index a in ε_a means diffusion tensor $a(\mathbf{x})$.

The value $\omega(\mathbf{e},\mathbf{x})$ refers to the original coordinate system, i.e. the **e**-system; in the **g**-system it is denoted by $\omega(\mathbf{g},\mathbf{x})$.

Lemma 2. (i) *Let an orthogonal system be defined by the unit vectors* $\mathbf{g}_i$*,* $i = 1,2,\ldots,d$*; we can then find for each* ε_a *a* **g***-independent* $\delta > 0$*, such that the inequality*

$$|\omega(\mathbf{g},\mathbf{x}) - \omega(\mathbf{g},\mathbf{y})| \quad < \quad \frac{2M}{\mu^2}\varepsilon_a, \tag{10}$$

holds for $|\mathbf{x}-\mathbf{y}| < \delta$*.*

(ii) *Let* $\varepsilon_T > 0$ *and* $\mathbf{f}_i$ *be unit vectors of another orthogonal coordinate system so that (9) is valid. Then*

$$|\omega(\mathbf{g},\mathbf{x}) - \omega(\mathbf{f},\mathbf{x})| \quad \leq \quad \frac{4M^2d^2}{\mu^2}\varepsilon_T. \tag{11}$$

(iii) *Let* $\varepsilon_T > 0$ *and* $\mathbf{f}_i$ *be unit vectors of another (not necessarily orthogonal) coordinate system, such that (9) is valid. Then, for sufficiently small* ε_T*, the following inequality is valid:*

$$\begin{aligned} |\omega(\mathbf{g},\mathbf{x}) - \omega(\mathbf{f},\mathbf{x})| \quad &\leq \quad \frac{4M_f^2d^2}{\mu_f^2}\varepsilon_T, \\ M_f := M(1+3\sqrt{2d\varepsilon_T}), \quad &\mu_f := \mu(1-3\sqrt{2d\varepsilon_T}). \end{aligned} \tag{12}$$

Proof. (Part (iii).) The transformations matrices from the original system to the **g**-system and **f**-system are denoted by O and T respectively; O is orthogonal. Matrix T has elements $T_{kl} = O_{kl} + \tau_{kl}$, where $\tau_{kl} = \mathbf{e}_k \cdot (\mathbf{f}_l - \mathbf{g}_l)$. It follows that $\sum_{kl} \tau_{kl}^2 \leq 2d\varepsilon_T$. An estimate of the matrix τ in l_2-norm is simple, $|\tau|_2 \leq (2d\varepsilon_T)^{1/2}$. The matrix $a(\mathbf{g},\mathbf{x}) = Oa(\mathbf{x})O^{-1}$ satisfies the ellipticity condition with the same constants as in (1). For sufficiently small ε_T, the matrix $a(\mathbf{f},\mathbf{x}) = Ta(\mathbf{x})T^{-1}$ satisfies another ellipticity inequality,

$$M_f|z|_2^2 \quad \geq \quad \sum_{i,j=1}^{d} a_{ij}(\mathbf{f},\mathbf{x})z_i\bar{z}_j \quad \geq \quad \mu_f|z|_2^2.$$

Property (12) then follows from this inequality in the same way as (11) follows from (1). □

It is convenient to denote $\omega(\mathbf{g}(m),\mathbf{x})$ by $\omega_m(\mathbf{x})$, and use the symbol T_m for the orthogonal system defined by unit vectors $\mathbf{g}_i(m)$.

Lemma 3. *Let $\rho \in (0, 1/4)$ be given. There exist $n \in \mathbb{N}$, $h = 2^{-n}$, and a finite number of disjoint sets $D_m, m = 0, 1, \ldots, M_a$, $\mathbb{R}^d = \cup_m D_m$, each supplied with an orthogonal coordinate system T_m, such that the function $\omega_m(\mathbf{x})$ on D_m in the T_m-system satisfies the inequality*

$$\sup\big\{|\omega_m(\mathbf{x})| \,:\, \mathbf{x} \in G_n \cap D_m\big\} \;\le\; 1 - 2\rho, \tag{13}$$

for any $m = 0, 1, \ldots, M_a$.

Proof. (A part of proof based on the uniform continuity of coefficients). To each $\mathbf{v} \in G_n$ there corresponds a grid cube $C(h, \mathbf{v}) = \prod_1^d \times [v_j, v_j + h)$, where v_j are coordinates of $\mathbf{v}$ in the original $\mathbf{e}$-system. Cubes $C(h, \mathbf{v})$ define a decomposition of $\mathbb{R}^d$ into disjoint sets. For the chosen ρ there is an $n \in \mathbb{N}$ and a ε_a of *Lemma 2.*, such that the difference (10) is less than ρ for all $\mathbf{x}, \mathbf{y} \in C(h, \mathbf{v})$, and for all $\mathbf{v} \in G_n$. Let us point out that (10) is valid for all orthogonal systems, i.e. $|\omega(\mathbf{g}, \mathbf{x}) - \omega(\mathbf{g}, \mathbf{y})| < \rho$. Let $D_0 \subset \mathbb{R}^d$ be the union of all those cubes $C(h, \mathbf{v})$ for which $|\omega(\mathbf{e}, \mathbf{v})| < 1 - 4\rho$. Let $C(h, \mathbf{v})$ be a cube outside of D_0. There are orthogonal unit vectors $\mathbf{g}_i(1)$, i = 1,2, … ,d, and the corresponding orthogonal transformation of coordinates, so that the diffusion tensor at $\mathbf{v}$ is diagonal in the $\mathbf{g}(1)$-system. Let $D_1 \subset \mathbb{R}^d$ be the union of all the cubes, for which $|\omega(\mathbf{g}, \mathbf{v})| < 1 - 4\rho$ is valid in the $\mathbf{g}(1)$-system. By induction, there is a decomposition $\mathbb{R}^d = \cup_0^\infty D_m$, where to each D_m there corresponds an orthogonal coordinate system, T_m, called a $\mathbf{g}$(m)-system, so that $|\omega_m(\mathbf{x})| < 1 - 4\rho$, $\mathbf{v} \in G_n \cap D_m$. Let $\mathbf{x} \in C(h, \mathbf{v}) \subset D_m$ for some m. Then $|\omega_m(\mathbf{x})| \le |\omega_m(\mathbf{x}) - \omega_m(\mathbf{v})| + |\omega_m(\mathbf{v})| < 1 - 3\rho$ for the chosen m.

(A part of proof based on the compactness of $(\partial B)^d$). Each T_k is defined by an element $s_k \in (\partial B)^d$, where s_k is the central $\mathbf{v}$-projection of $\mathbf{g}_i(k), i = 1, 2, \ldots, d$, onto $(\partial B)^d$. Elements s_k and s_l are defined by the unit vectors $\mathbf{g}_i(k)$ and $\mathbf{g}_i(l)$, i = 1,2, … ,d, respectively. Let $\varepsilon_T > 0$ and $U_k \subset (\partial B)^d$ be a neighbourhood of s_k containing all the elements $s = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_d\} \in (\partial B)^d$, defined by unit vectors $\mathbf{f}_i$ so that (9) is valid. Due to the compactness of $(\partial B)^d$, there is a finite sub-cover $\{V_0, V_1, \ldots, V_M\} \subset \{U_j : j \in \mathbb{N}_0\}$ of the set $\{s_j : j \in \mathbb{N}_0\}$. Let r be chosen. The chosen element s_r, associated with D_r, is contained in some of the sets V_m, m = 0,1, … ,M, and

$$\mathbf{g}_i(r) \cdot \mathbf{g}_i(m) > 1 - \varepsilon_T.$$

Let ω_r and ω_m be functions defined by (5) in the $\mathbf{g}$(r)-system and $\mathbf{g}$(m)-system, respectively. Inequality (11) can be used to get:

$$\begin{aligned} |\omega_m(\mathbf{x})| \quad &\le \quad |\omega_m(\mathbf{x}) - \omega_m(\mathbf{v})| \;+\; |\omega_m(\mathbf{v}) - \omega_r(\mathbf{v})| \;+\; |\omega_r(\mathbf{v})| \\ &< \quad \rho \;+\; \frac{4M^2d^2}{\mu^2}\varepsilon_T \;+\; 1 - 4\rho. \end{aligned}$$

By choosing ε_T sufficiently small we have $|\omega_m(\mathbf{x})| \le 1 - 2\rho$. □

Lemma 4. *There exists a decomposition of $\mathbb{R}^d$ into $1 + M_a$ disjoint sets D_m, and there exist $1 + M_a$ associated coordinate systems T_m (not necessary orthogonal) with the following properties:*

(i) *the axis of T_m intersect grid-knots of G_n;*

(ii) *the function* ω_m *on* D_m *in the coordinate system* T_m *satisfies the inequality* $|\omega_m(\mathbf{x})| \leq 1-\rho$.

Proof. Let $\mathbf{x} \in G_n$. By *Lemma 3.* there exists $m \in \{0, 1, \cdots, M_a\}$ such that $|\omega_m(\mathbf{x})| < 1-2\rho$. A system T_m has at least one axis which does not intersect grid-knots of $\cup_m G_m$. Therefore, we have to rotate axis $\mathbf{g}_i(m)$, $i \in \{1, 2, \cdots, M_a\}$ slightly, and obtain a new axis $\mathbf{f}_i(m)$, satisfying (9). Then

$$\begin{aligned} |\omega(\mathbf{f}(m), \mathbf{x})| &\leq |\omega(\mathbf{f}(m), \mathbf{x}) - \omega(\mathbf{g}(m), \mathbf{x})| \;+\; |\omega(\mathbf{g}(m), \mathbf{x})| \\ &\leq \frac{4M_f^2 d^2}{\mu_f^2}\varepsilon_T \;+\; 1-2\rho. \end{aligned}$$

The statement (ii) is attained by choosing ε_T sufficiently small. □

The proof of *Theorem 1.* can now be finished:

Proof. For each $\mathbf{x} \in G_n$, the collection of all the grid-knots connected by the scheme defines the numerical neighbourhood $N(\mathbf{x})$, and there are at most $1+M_a$ different types of neighbourhoods. All the neighbourhoods at the same $\mathbf{x} \in \cup_n G_n$ are alike with respect to the scaling by positive constants 2^m, $m \in \mathbb{Z}$. All the neighbourhoods $N(\mathbf{x})$, for $\mathbf{x} \in D_m \cap G_n, m, n$, fixed, are alike with respect to translations. Each $N(\mathbf{x})$ contains at most $1 + d(1+d)$ grid-knots. This proves (ii) of *Theorem 1.*

To the differential operator $A_0(\mathbf{x}) = \sum_{ij} a_{ij}(\mathbf{x})\partial_i\partial_j$ we apply the coordinate transformation T_m. In the new coordinates we obtain $A_0(\mathbf{x}) = \sum_{ij} a_{ij}(\mathbf{f}, \mathbf{x})\hat{\partial}_i\hat{\partial}_j$, the diffusion tensor $a_{ij}(\mathbf{f}(m), \cdot)$ satisfying (6) on D_m. For each $\mathbf{x} \in D_m \cap G_n$, the corresponding numerical scheme is:

$$A_n u(\mathbf{x}) = \sum_{\mathbf{z} \in N(\mathbf{x})} a(\mathbf{z}, h) u(\mathbf{z}),$$

where $a(\mathbf{x}, h) < 0$ and $a(\mathbf{z}, h) > 0$ for all $\mathbf{z} \in N(\mathbf{x}) \setminus \{\mathbf{x}\}$. This completes the proof of *Theorem 1.* □

Since the set D_0 of the previous construction may be empty, it is assumed that an orthogonal transformation is performed at the beginning of the construction making D_0 nonempty. The cubes $C(h, \mathbf{v})$ entering the construction of the sets $D_m, m = 0, 1, \ldots, M$, are called basic cubes, i.e. the sets D_m are unions of basic cubes. Basic cubes are denoted by $C(\mathbf{v})$ rather than $C(h, \mathbf{v})$.

An infinite-order matrix $A = \{a_{ij}\}_{\mathcal{I}\mathcal{I}}$, with the index set $\mathcal{I}$, is called *reducible* if there exists a decomposition $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$, into disjoint nonempty subsets $\mathcal{I}_1, \mathcal{I}_2$, such that the sub-matrix X which corresponding to indices $i \in \mathcal{I}_1, j \in \mathcal{I}_2$, is zero. An infinite-order matrix A is called *irreducible* if such a decomposition is not possible. The following criterion of irreducibility will be used: Let $\mathcal{I}_1$ be any finite index set, and $A_{\mathcal{I}_1\mathcal{I}_1}$ be the corresponding submatrix of A. If there exists a finite index set $\mathcal{I}_2$, $\mathcal{I}_1 \subset \mathcal{I}_2$, such that the submatrix $A_{\mathcal{I}_2\mathcal{I}_2}$ is irreducible, then the matrix A is irreducible, too. The matrix A_n of *Theorem 1.* can be reducible. As an example, let us consider $A = d^2/dx^2$, $\mathbb{R} = D_0 \cup D_1$, $D_1 = [0, \infty)$. Let the grid steps in D_0 and D_1 be equal to h and $3h$ respectively; it follows that A_n are reducible. Reducibility can cause undesirable features of solutions of ODE which have a reducible matrix

A_n as the system matrix; a demonstration of such an undesired effect is illustrated in *Example 2*.

The well known notions of recurrent classes of the theory of Markov chains can be defined and utilized here. A grid-knot $\mathbf{y}$ can be *reached* from a grid-knot $\mathbf{x}$ if there exists a sequence of grid-knots, $\mathbf{x} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m = \mathbf{y}$, such that $\mathbf{x}_{k+1} \in N(\mathbf{x}_k)$. The sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ is called a *path*. The grid-knots $\mathbf{x}$ and $\mathbf{y}$ *communicate* if there exist two paths such that $\mathbf{y}$ can be reached from $\mathbf{x}$ by one path, and $\mathbf{x}$ can be reached from $\mathbf{y}$ by another one. The communication of any two grid-knots of G_n is equivalent to the irreducibility of A_n.

Lemma 5. *Let $C(\mathbf{v}_1)$ and $C(\mathbf{v}_2)$ be any two basic cubes, and $\mathbf{x} \in C(\mathbf{v}_1) \cap G_n$ for some n. Then, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ there is a grid-knot $\mathbf{y} \in C(\mathbf{v}_2) \cap G_n$ which can be reached from $\mathbf{x}$.*

Proof. Let $C(\mathbf{v}) \subset D_k$ be a basic cube defined by the vertex $\mathbf{v} \in G_n$ and let $C(\mathbf{v}_r)$ be those 2^d surrounding basic cubes facing $C(\mathbf{v})$ by means of sides. Let $\mathbf{x} \in C(\mathbf{v})$. For sufficiently large n there exist grid-knots $\mathbf{x}_r \in C(\mathbf{v}_r) \cap G_n$ which can be reached from $\mathbf{x}$. By induction, all $3^d - 1$ basic cubes which surround $C(\mathbf{v})$ have this property, i.e. they have grid-knots in G_n reachable from $\mathbf{x}$. Since the number of types of numerical neighbourhoods is finite, there exists a sufficiently large n_0 such that the established property holds for any basic cube and $\mathbf{x} \in G_n, n \geq n_0$. □

Proposition 1. *Let A_n be approximations on G_n of the differential operator A defined by (1), (2), and let A_n have the properties (i) - (v) of* Theorem 1. *For large values of n, the set $G_n = G_n(A)$ is either one class of grid-knots with respect to A_n, or $G_n = G_n(A) \cup H_n$, where*

(i) *$G_n(A)$ are recurrent classes and $\cup_n G_n(A)$ is a dense set in $\mathbb{R}^d$;*

(ii) *the sets H_n contain no recurrent class;*

(iii) *if $H_n \neq \emptyset$, the matrix A_n on G_n has the following block structure:*

$$A_n = \begin{pmatrix} A_n(A) & 0 \\ R & Q \end{pmatrix},$$

where $A_n(A)$ is the irreducible part of A_n.

(iv) *If $H_n \neq \emptyset$, then $A_n(A)$ is an approximation on subgrids $G_n(A)$ of the operator A, and $A_n(A)$ has the properties (i) - (v) of* Theorem 1.

Proof. Let $\mathbf{x}_1, \mathbf{x}_2 \in G_n$ and $C(\mathbf{v}_0) \subset D_0$ be a basic cube. There are two grid-knots $\mathbf{y}_1, \mathbf{y}_2 \in C(\mathbf{v}_0)$ which can be reached from $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Because all the internal grid-knots in basic cubes of D_0 communicate among themselves, we conclude that all the grid-knots in D_0 communicate among themselves. Next we define the subgrid $G_n(A) \subset G_n$ containing D_0, and all of the grid-knots in G_n that communicate among themselves. This set is the only recurrent class of G_n. Since it contains grid-knots in each basic cube, there must be an n-independent $\kappa > 0$, such that any ball of the radius κh contains at least two grid-knots of $G_n(A)$. □

Generally, two grids $G_n(A)$ for different n cannot be ordered by inclusion, that is, $G_n(A) \subset G_m(A)$, or vice-versa. One would like at least a subsequence of such grids, $G_{n(r)}(A), r \in \mathbb{N}$, to exist, for which the inclusion $G_{n(r)}(A) \subset G_{n(r+1)}(A)$ is valid, and in addition, $\cup_r G_{n(r)}(A)$ to be a dense set in $\mathbb{R}^d$. The mentioned ordering is not always possible, as can be shown if we take $A = d^2/dx^2, G_n = \mathbb{N}, G_n(A) = \{k \in \mathbb{Z}\} : k \leq 0\} \cup \{1 + rk : k \in N\}$, where r is a fixed natural. The mentioned ordering is possible for $r = 2^m - 1$ with some natural m, and impossible for any other choice of r.

3. A weighted polynomial scheme in $\mathbb{R}^2$ possessing MJP-structure

It is possible to construct a numerical scheme possessing the MJP-structure without using rotations at the knots. The approximating operator then reproduces piecewise parabolæ, i.e. (v) of *Theorem 1.* holds for piecewise quadratic p, which does not influence local convergence properties; we will show how such a scheme can be constructed in the important case of an elliptic operator $\dot{C}^{2+\alpha}(\mathbb{R}^2) \rightarrow \dot{C}^{\alpha}(\mathbb{R}^d)$.

Let G_{n+1} be the numerical grid in $\mathbb{R}^1$ and let us define $w^{n+1} : \mathbb{R} \rightarrow \mathbb{R}^+$ to be a piecewise constant and right-continuous function having breakpoints at $G_{n+1} = \{x_i^{n+1} : i \in \mathbb{Z}\}$. Let

$$w^{n+1} \mid_{[x_i^n, x_{i+1}^n)} := \begin{cases} w_{2i}^{n+1} & \text{for} \quad x_{2i}^{n+1} < x_{2i+1}^{n+1} \\ w_{2i+1}^{n+1} & \text{for} \quad x_{2i+1}^{n+1} \leq x_{2i+2}^{n+1}. \end{cases}$$

The associated measures $d\sigma_w^{n+1}$ are defined by $d\sigma_w^{n+1}(\delta) = \displaystyle\int_\delta \frac{dx}{w^{n+1}(x)}$ for all Lebesgue measurable sets $\delta \subset \mathbb{R}$. For the sake of symmetry, let $w_{2i}^{n+1} = w_{2i-1}^{n+1}$, $w_{2i+1}^{n+1} = w_{2i-2}^{n+1}$ for all n. To simplify the notation, we keep h as the distance between knots in G_n, that is $h/2$ is the distance in G_{n+1}. If f is sufficiently smooth, we may use a sort of generalized Taylor expansion in the neighbourhood of a point in G_n, that is the following identity:

$$f(x_{i+1}^n) \quad = \quad f(x_i^n) + h\partial_x f(x_i^n) + \int_{x_i^n}^{x_{i+1}^n} d\eta \int_{x_i^n}^{\eta} D_w^2 f(\xi) d\sigma_w^{n+1}(\xi), \tag{14}$$

where products of measure derivatives are denoted as:

$$D_w^2 f \quad := \quad w^{n+1} \cdot \partial_{xx} f. \tag{15}$$

Upon integration of (14) we have

$$\begin{aligned} f(x_{i+1}^n) \quad = \quad & f(x_i^n) + h\partial_x f(x_i^n) + w_{2i}^{n+1} f''(x_i^n) \left(\int_{x_{2i}^{n+1}}^{x_{2i+1}^{n+1}} d\eta \int_{x_{2i}^{n+1}}^{\eta} \frac{d\xi}{w_{2i}^{n+1}} \right. \\ & + \int_{x_{2i+1}^{n+1}}^{x_{2i+2}^{n+1}} d\eta \int_{x_{2i+1}^{n+1}}^{\eta} \frac{d\xi}{w_{2i+1}^{n+1}} + \left. \int_{x_{2i+1}^{n+1}}^{x_{2i+2}^{n+1}} d\eta \int_{x_{2i+1}^{n+1}}^{x_{2i+2}^{n+1}} \frac{d\xi}{w_{2i+1}^{n+1}} \right) \\ & + \int_{x_i^n}^{x_{i+1}^n} d\eta \int_{x_i^n}^{\eta} (D_w^2 f(\xi) - D_w^2 f(x_i^n) d\sigma_w(\xi). \end{aligned}$$

Thus, for the piecewise constant w, we have, up to the terms of the 3rd order:

$$f(x_{i+1}^n) \quad \approx \quad f(x_i^n) + h\partial_x f(x_i) + D_w^2 f(x_i^n)\frac{h^2}{8}\left(\frac{1}{w_{2i}^{n+1}} + \frac{3}{w_{2i+1}^{n+1}}\right). \tag{16}$$

In the same way, by expanding f to the left of x_i^n, we obtain:

$$f(x_{i-1}^n) \quad \approx \quad f(x_i) - h\partial_x f(x_i^n) + D_w^2 f(x_i^n)\frac{h^2}{8}\left(\frac{1}{w_{2i-1}^{n+1}} + \frac{3}{w_{2i-2}^{n+1}}\right). \tag{17}$$

The "weighted central difference" approximation of the second derivative follows from (16), (17), and the symmetry properties of w:

$$\left(\frac{1}{w_{2i}^{n+1}} + \frac{3}{w_{2i+1}^{n+1}}\right) D_w^2 f(x_i) \quad \approx \quad \frac{4}{h^2}(f(x_{i+1}^n) + f(x_{i-1}^n) - 2f(x_i^n)).$$

Definition 15 then leads to an approximation of the second derivative in the form of

$$\partial_{xx} f \quad \to \Box_{11,w} f := \tfrac{4}{h^2}\,\frac{f(x_{i+1}^n)+f(x_{i-1}^n)-2f(x_i^n)}{1+3w_{2i}^{n+1}/w_{2i+1}^{n+1}}. \tag{18}$$

For equal weights, one obtains the common result (7): $\Box_{11} f := \frac{1}{h^2}(f(x_{i+1}) + f(x_{i-1}) - 2f(x_i))$. In the same manner, we may approximate ∂_{yy} by $\Box_{22,r}$, whereas the mixed derivatives ∂_{xy} are approximated by the usual non-weighted differences (8). We assume that r^{n+1} are a positive piecewise constant functions defining measures with density $1/r^{n+1}$, playing the same role as w in the y-direction, and $r_j^{n+1} := r^{n+1}\,|_{[\,y_j, y_{j+1})}$. Let

$$\alpha_i^n := 1 + 3w_{2i}^{n+1}/w_{2i+1}^{n+1}, \qquad \beta_j^n := r_{2j}^{n+1}/r_{2j+1}^{n+1}. \tag{19}$$

The normalization constants α_i^n, β_j^n can be determined in such a way that the difference scheme posseses the MJP-structure:

Theorem 2. *Let $a_{11}\partial_{xx} + 2a_{12}\partial_{xy} + a_{22}\partial_{yy} + b_1\partial_x + b_2\partial_y + c$ be an elliptic operator $\dot{C}^{2+\alpha}(\mathbb{R}^2) \to \dot{C}^{\alpha}(\mathbb{R}^d)$. Then there exists a choice of weights (19) such that the difference operator*

$$a_{11}\Box_{11,w} + 2a_{12}\Box_1\Box_2 + a_{22}\Box_{22,r} + b_1\Box_1 + b_2\Box_2$$

possesses the MJP-structure.

Proof. The weights enter approximations of second derivatives only, the first derivatives being approximated as in (7). We may assume that a_{11}, $a_{22} > 0$, $a_{12} \le 0$ (the case $a_{12} > 0$ may be treated in the same way). Then we may consider a numerical scheme obtained by approximating derivatives by finite differences which depend on the weights; the appropriate factors are listed in the following table (indices for α, β omitted, and a_{ij} are calculated at the knot (x_i^n, y_i^n)):

knot	factor $\times 4/h^2$
(x_i^n, y_i^n)	$-2a_{12} - \frac{2}{\alpha}a_{11} - \frac{2}{\beta}a_{22}$
(x_{i+1}^n, y_i^n)	$a_{12} + \frac{a_{11}}{\alpha}$
(x_{i+1}^n, y_{i+1}^n)	0
(x_i^n, y_{i+1}^n)	$a_{12} + \frac{a_{22}}{\beta}$
(x_{i-1}^n, y_{i+1}^n)	$-a_{12}$
(x_{i-1}^n, y_i^n)	$a_{12} + \frac{a_{11}}{\alpha}$
(x_{i-1}^n, y_{i-1}^n)	0
(x_i^n, y_{i-1}^n)	$a_{12} + \frac{a_{22}}{\beta}$
(x_{i+1}^n, y_{i-1}^n)	$-a_{12}$

The MJP property requires that

$$\begin{aligned} a_{12} + \frac{1}{\alpha}a_{11} + \frac{1}{\beta}a_{22} &\geq 0, \\ a_{12} + \frac{a_{11}}{\alpha} &\geq 0, \\ a_{12} + \frac{a_{22}}{\beta} &\geq 0. \end{aligned}$$

The first equation is implied by the last two, and this can be achieved if we let

$$0 < \alpha = \beta = \min\{-\frac{a_{11}}{a_{12}}, -\frac{a_{22}}{a_{12}}\}.$$

The approximations of first derivatives do not influence the MJP-structure, since the argument used in the proof of *Theorem 1.* applies. □

It is not difficult to prove that the scheme reproduces constants and linear functions exactly. It is not consistent, since it does not reproduce quadratic polynomials but piecewise parabolæ, that is, it reproduces parabolic splines [R] having jumps in second derivatives at the interior knots (property (v) of *Theorem 1.*). This is enough to establish convergence of the 2nd order, since the proof relies on 'weighted' Taylor expansion.

Theorem 3. *Let the difference operator A_n be determined by the scheme as in* Theorem 2. *Then*

$$\| \phi_0(n)A - A_n\phi_2(n) \|_{\mathbf{L}(E_2,F_0)} \leq \kappa h^{\alpha}, \tag{20}$$

that is, (3) holds, and the order of approximation is $\beta = \alpha$.

Proof. We first establish a Taylor expansion, that is, we prove that for $w > 0$ and u such that wu'' is integrable with respect to $d\sigma_w$, the error term in the expansion

$$u(x+h) \quad = \quad u(x) + hu'(x) + D_w^2 u(x) \int_0^h (h-t)\,d\sigma_w(t) + R(w;x) \tag{21}$$

is of the form

$$R(w;x) \quad = \quad \int_0^h [w(t+x)u''(t+x) - w(x)u''(x)](h-t)\frac{dt}{w(t+x)}. \qquad (22)$$

In the identity

$$u(x+h) \quad = \quad u(x) + hu'(x) + \int_x^{x+h} d\eta \int_x^{\eta} D_w^2 u(\xi) d\sigma_w(\xi)$$

we can rearrange the order of integration in the last term to obtain

$$\int_x^{x+h} d\eta \int_x^{\eta} D_w^2 u(\xi) d\sigma_w(\xi) \quad = \quad \int_x^{x+h} D_w^2 u(\xi) d\sigma_w(\xi) \int_{\xi}^{x+h} d\eta =$$

$$\int_x^{x+h} (x+h-\xi) D_w^2 u(\xi) d\sigma_w(\xi) \quad = \quad \int_0^h w(t+x)\partial_{tt} u(t+x) \cdot (h-t)\frac{1}{w(t+x)} dt.$$

Let $d\sigma_{w,x} := \dfrac{dt}{w(t+x)}$. By adding and subtracting the "second order term"

$$(wu''(x)\int_x^{x+h} (x+h-\xi) d\sigma_w(\xi),$$

we obtain the Taylor expansion:

$$u(x+h) = u(x) + hu'(x) + D_w^2 u(x) \int_0^h (h-t) d\sigma_{w,x} + R(w;x).$$

The error term may be written as $\int_x^{x+h}(D_w^2 u(\xi) - D_w^2 u(x))(x+h-\xi)d\sigma_w(\xi)$, which implies (22) upon substitution $\xi = t+x$. One verifies easily that the second term in (21) is in fact the second term in (16),(17).

Let $d(x) := (\partial_1^2 - \Box_{1,w})u(x)$. If we multiply and divide Taylor error (22) by t^α, we a obtain a simple estimate $|d(x)| \leq \mu h^\alpha \|u\|_{2+\alpha}$, where μ is independent of h and α; other differentials can be estimated in more or less the same way. We have immediately that

$$\|a_{11}\, d\|_{F_0} \quad \leq \quad \mathrm{c}h^\alpha \|u\|_{E_2}, \qquad \mathrm{c} \leq 2\mu \|a_{11}\|_{C(\mathbb{R}^2)},$$

whence (3) follows. □

4. An application to parabolic systems

We now consider the 2nd-order elliptic operator

$$A(t,\mathbf{x}) \;=\; \sum_{ij} a_{ij}(t,\mathbf{x})\partial_i\partial_j \;+\; \sum_i b_i(t,\mathbf{x})\partial_i \;+\; c(t,\mathbf{x}) \qquad (23)$$

on $\mathbb{R}^d$. In order to bound the error of the numerical approximations in terms of the grid step, we impose the following continuity conditions on the coefficients:

$$\begin{gathered} |a_{ij}(t,\mathbf{x}) - a_{ij}(s,\mathbf{y})| \,+\, |b_i(t,\mathbf{x}) - b_i(s,\mathbf{y})| \,+\, |c(t,\mathbf{x}) - c(s,\mathbf{y})| \;\leq \\ M_H\left(|\mathbf{x}-\mathbf{y}|^\alpha \;+ |t-s|^{\alpha/2}\right), \\ |b_i|,\; |c| \,\leq\, M_H, \quad i,j \,=\, 1,2,...,d, \end{gathered} \qquad (24)$$

for some $M_H > 0$, and moreover, we require the strict ellipticity to be t-uniform, that is, there exist $M > 0$ and $\mu > 0$ such that for all $t \geq 0$ and $\mathbf{x} \in \mathbb{R}^d$:

$$M|z|_2^2 \geq \sum_{i,j=1}^{d} a_{ij}(t, \mathbf{x}) z_i \bar{z}_j \geq \mu |z|_2^2, \quad \mu > 0. \tag{25}$$

We start an analysis of the numerical schemes by considering the initial value problem on $\mathbb{R}^d$:

$$\begin{aligned} \frac{d}{dt} u(t) &= A(t)u(t) + f(t), \\ u(0) &= u_0. \end{aligned} \tag{26}$$

Our objective is to analyse numerical approximations of (26) defined in terms of ODE:

$$\begin{aligned} \frac{d}{dt} u_n(t) &= A_n(t)u_n(t) + f_n(t), \\ u_n(0) &= u_{n,0}, \end{aligned} \tag{27}$$

where $A_n(\cdot)$ are the approximations of (23) on G_n, defined and described in *Sections 2.* and *3.*

If u_0 belongs to the class $\dot{C}^{2+\alpha}(\mathbb{R}^d)$, the parabolic system (26) has a classical solution. Since error estimates are derived from the properties of classical theory, it is necessary to define basic objects used in the course of the proofs. For an analysis of parabolic systems it is convenient to define the Banach space $\mathcal{H}(m, \alpha)$, spanned by functions u on $Q^{T,d} = [0, T] \times \mathbb{R}^d$, vanishing for large values of $\mathbf{x}$, the norm being defined as:

$$\|u\|_{\mathcal{H}(m,\alpha)} = \sup_{0 \leq t \leq T} \|u(t)\|_{m+\alpha} + \sum_{0 \leq l \leq m/2} \sup_{|t-s|<1} \left\| \frac{\partial_t^l u(t) - \partial_s^l u(s)}{|t-s|^{\alpha/2}} \right\|_{m-2l}.$$

Due to (24), (25), a solution v of the initial value problem (26) may be estimated in the following way [LSU]: For each $T > 0$ there exists a $\kappa(T)$ such that

$$\|u\|_{\mathcal{H}(2,\alpha)} \leq \kappa(T) \Big(\|u_0\|_{2+\alpha} + \|f\|_{\mathcal{H}(0,\alpha)} \Big). \tag{28}$$

Let $\triangle = \{(s, t) \in \mathbb{R}^2 : s \leq t\}$. An elliptic differential operator $A(\cdot)$ on $\mathbb{R}^d$, satisfying (24) - (25), generates an evolution family of positive operators $U(t, s) \in \mathbf{L}(\dot{C}(\mathbb{R}^d)), s, t \in \triangle$ (see [KR1, T1]). If $c = 0$ on $\mathbb{R}^d$, the operators are conservative, while for $c \leq 0$ they are contractions. Similarly, the operators $A_n(\cdot) \in \mathbf{L}_\infty(G_n)$ generate the corresponding evolution families $U_n(t, s)$. By (28), one obtains $u(t) = U(t, s)v_s$ and $|U(t, s)|_{2+\alpha} \leq \kappa(T)$, uniformly with respect to $0 \leq s \leq t \leq T$.

The space $\dot{C}(\mathbb{R}^d)$ is a Banach algebra and U(t,s) are its continuous mappings. Therefore, products of $u(t)U(t, s), u \in \dot{C}(\mathbb{R}^d)$, are also continuous mappings of this algebra. Similarly, the space $F_0(G_n) \subset l_\infty$ is an algebra, and $U_n(t, s)$ are its continuous mappings. The natural imbedding $\phi_0(n)$ is an isomorhism of the algebra $\dot{C}(\mathbb{R}^d)$ into F_0, i.e. $\phi_0(n)(uv) = (\phi_0(n)u)(\phi_0(n)v)$. It happens that the products are in $\mathcal{H}(m, \alpha)$ for $u, v \in \mathcal{H}(m, \alpha)$, so that the following result holds true:

Lemma 6. *Let $A(\cdot)$ satisfy (24)–(25). Then*

(i) *for each $T > 0$ there exists a $\rho(T)$ such that*

$$\Big\| \Big(\phi_0(n)U(t,s) \;-\; U_n(t,s)\phi_0(n)\Big)u \Big\|_\infty \;\le\; \rho(T)\|u\|_{2+\alpha}\, h^\alpha,$$

whenever $0 \le s < t \le T$, and $u \in E = \dot{C}^{2+\alpha}(\mathbb{R}^d)$,

(ii) *for each $m \in \mathbb{N}$ and $T > 0$, there exists a $\rho(m,T) > 0$ such that*

$$\Big\| \,\phi_0(n) \prod_{k=1}^m U(t_{k+1},t_k)u_k(t_k) \;-\; \prod_{k=1}^m U_n(t_{k+1},t_k)\Big(\phi_0(n)u_k(t_k)\Big)\Big\|_\infty \;\le\; \rho(m,T)\prod_{k=1}^m \|u_k\|_{2+\alpha} h^\alpha,$$

whenever $0 \le t_1 < t_2 \cdots < t_{m+1} \le T$.

Proof. The following identity [KR1, T1] is used in the proof of (i):

$$\phi_0(n)U(t,s) \;-\; U_n(t,s)\,\phi_0(n) \;=\; \int_s^t U_n(t,\tau)\,\big(\,\phi_0(n)\,A(\tau) \;-\; A_n(\tau)\,\phi_0(n)\,\big)\,U(\tau,s)\,d\tau.$$

Here, the operator $A_n(t)$ is bounded on both, F_0 and F_2 of *Section 2.*, so that $A_n(t)\phi_0(n) = A_n(t)\phi_2(n)$. By (28):

$$\Big\|\,\Big(\phi_0(n)U(t,s) \;-\; U_n(t,s)\,\phi_0(n)\,\Big)u\,\Big\|_\infty \;\le\; T\,\sup_{t\in[0,T]} \|A_n(t)\,\phi_2(n) \;-\; \phi_0(n)\,A(t)\|_{\mathbf{L}(E,F_0)}\;\|\,u\,\|_{2+\alpha}\,\kappa(T),$$

and the statement (i) follows from *Theorem 1.*

The statement (ii) can be derived by induction from (i). □

Our intention is to describe a class of schemes for solving (27). They are explicit, of any order of convergence, and absolutely stable. The initial value problems (27), with zero nonhomogeneous terms, have the general form:

$$\begin{aligned} \frac{d}{dt}v(t) &= H(t)v(t), \\ v(0) &= v_0. \end{aligned} \tag{29}$$

with the matrix $H(t)$ possessing the MJP-structure. The corresponding evolution family is denoted by $Q(t,s)$.

The matrix $H(t)$ can be represented as $H(t) = D(t) + B(t)$, where $D(t)$ is a diagonal matrix with diagonal elements $h_{ii}(t)$, and $B(t)$ is a matrix with nonnegative elements and zero diagonal elements. It is easy to describe the schemes for matrices H of (29) which are t-independent, and have all diagonal elements equal, $h_{ii} = -p$. Let τ be increment in time, and $t_m = m\tau$, $m = 0,1,\ldots.$ Let v_m be the approximations of $v(t_m)$. For t-independent H, the approximations of order L are defined by:

$$v_m \;=\; V(\tau)v_{m-1}, \quad m \in \mathbb{N}, \quad V(\tau) \;=\; P_L(\tau,p)^{-1}\,P_L(\tau,B), \tag{30}$$

where the polynomial $P_L(\tau, x)$ of order L is equal to the first $L+1$ terms of expansion of $x \to exp(\tau x)$, i.e. $P_L(\tau, x) = \sum_{r=0}^{L} \tau^r x^r / r!$. Matrix norm of the numerator, $P_L(\tau, B)$, is not larger than the denominator, so that $\|V(\tau)\|_\infty \leq exp(-\kappa\tau)$, where $\kappa = inf_i |h_i|$. The order of convergence is L, as follows from the corresponding error bounds:

$$\|v_m - v(t_m)\|_\infty \;\leq\; \|V(\tau)\|_\infty \|v_{m-1} - v(t_{m-1})\|_\infty + \frac{2(p\tau)^{L+1}}{(L+1)!}\|v(t_{m-1})\|_\infty. \quad (31)$$

If all $h_i = 0$, then $\|v_m\|_\infty = \|v_0\|_\infty, \quad m \in \mathbb{N}$. An extension to t-dependent matrices $H(\cdot)$ is simple: Let

$$H(t) \;=\; -p(t)I \;+\; B(t),$$

where both, p and matrix elements of B, are smooth enough, so that integration formulæ, to be used in proceeding construction, have desired error bounds. Let $W(t, s)$ be the evolution family generated by $B(\cdot)$. In the proposed generalization of (30) the numerator must contain the first $L+1$ terms of expansion of $W((m-1)\tau, m\tau)$. The denominator of (30) must contain the corresponding $L+1$ terms of expansion of exponential function having the integral of p in the exponent. Each term of expansion must be approximated by an integration formula, with an error bound of the order $L+1$. The sequence

$$v_m = V(\tau, m)v_{m-1}, \quad V(\tau, m) = X_L(\tau, m, p)^{-1}\, X_L(\tau, m, B)$$

obtained in such a way, converges to the solution of (27) with the order L. For lower order convergence integration formulæ are simple [LR]. For the third order scheme, we define the third degree polynomial by:

$$\begin{aligned} X_3(\tau, m, g) \quad &= \quad 1 + \frac{\tau}{6}\Big[g(t_b) + 4g(t_m) + g(t_e)\Big] \\ &+\frac{\tau^2}{12}\Big[g(t_e)g(t_b) + g(t_e)^2 + 2g(t_m)^2 + 2g(t_m)g(t_b)\Big] \;+\; \frac{\tau^3}{6}g(t_e)g(t_m)g(t_b), \end{aligned}$$

where t_b, t_m and t_e are the beginning, midpoint and endpoint of the interval $[(m-1)\tau, m\tau]$. In order to apply the scheme to non-commuting matrices $B(t), t \in [0, \infty)$, we must respect the order of factors $B(t)$.

If the diagonal elements (functions) h_{ii} of H are not mutually equal, the proposed schemes can be used after slight adjustment. Let $[0, T]$ be the interval of interest. It is assumed that there exists a function p on the considered interval such that $p(t) \geq -h_{ii}(t)$ for all i, and $t \in [0, T]$. After splitting $H(t) = -p(t)I + B(t)$ the proposed schemes can be used. To the smaller difference $\sup_{i,t}(p(t) + h_{ii}(t))$ there corresponds smaller error of schemes.

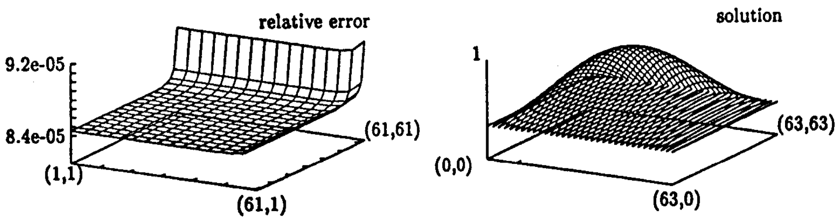

Figure 1. *Test example*

K	100	300	500
$\varepsilon(3,\tau)$	0.2208E+00	0.4670E-02	0.9019E-03
$\frac{tim(S)}{tim(E)}$	1.29E-03	1.33E-03	.34E-03

Table 1. *Error bounds*

Example 1. *The function*

$$u(t,x_1,x_2) \;=\; exp(-\pi^2\sigma^2 t)\, sin(\pi x_1) sin(\pi x_2)$$

is a solution of the initial value problem with the elliptic operator $A = (\sigma^2/2)\Delta$ *on* $D = (0,1)\times(0,1)$*, and the homogeneous Dirichlet boundary conditions. Numerical approximations are considered for* $\sigma^2 = 0.1/\pi^2, t = 1$*, with the grid step* $h = 2^{-6}$*. The resulting ODE (29) is solved by the method (30) with* $L = 3$ *and Euler implicit method. For various time increments,* $\tau = 1/K$*, the corresponding numerical approximations of* $u(1,\frac{1}{2},\frac{1}{2})$ *are presented in* Table 1*. The error bounds can be obtained from (31):*

$$\varepsilon(L,\tau) \;=\; \|v_k \,-\, v(t_k)\|_\infty \;\le\; \frac{2(1+\tau)}{(1+L)!}\, exp\,(\,p\tau\,)\, p\,(p\tau)^L, \quad k = 1,2,...,K.$$

The last column of table contains the ratio of computing times of the two methods. All calculations were carried out in double precision, with the same level of optimization, on SUN-SPARC (Ultra-Enterprise).

In this section irreducibility versus reducibility of approximations A_n was not discussed. The following example demonstrates that reducibility must be additionally interpreted.

Example 2. *Let* $D = (0,1)$*, and* $A = d^2/dx^2$ *with the homogeneous Neumann boundary conditions at the endpoints of* D*. We consider grids* $G_n = \{hk/n : k = 0,1,\ldots,n\}$ *of [0,1], and the corresponding approximations* A_n *on* G_n *defined as follows. Let* x_i *be in* $G_n \cap (0,1/3)$*, or in* $G_n \cap (2/3,1)$*. At these grid-knots the value* $u"(x_i)$ *is approximated in a standard way by* $h^{-2}(u(x_i-h) - 2u(x_i) + u(x_i+h))$*. At grid-knots of the middle part, [1/3,2/3], the approximation is constructed with* *the step equal to* $3h$*,* $u"(x_i) \to (3h)^{-2}(u(x_i - 3h) - 2u(x_i) + u(x_i + 3h))$*. Boundary conditions are approximated in the standard way. The matrices* A_n *are reducible. Solutions of the initial value problems (29) have a peculiar behaviour as* $t \to \infty$*. Depending on the positions of grid-knots, the vector valued function* $u_n(t)$ *tends either towards zero or towards two real numbers. In* Figure 2 *the solution* $u_n(t), n = 90$*, for various times* t *is graphically presented.*

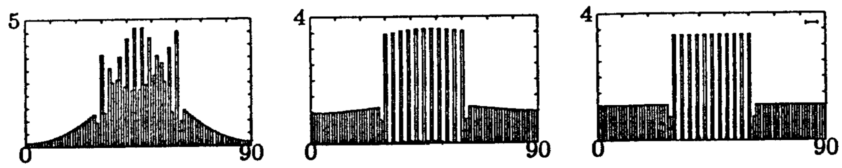

Figure 2. Example showing reducibility

5. Conclusion

Numerical method proposed in *Section 4.* for the initial value problem for the 2^{nd}-order parabolic systems is very useful due to the small requirements on memory, and possible implementation to parallel architectures. The only needed routine is the multiplication of the system matrix A_n with column u_n. The method is explicit and stable for matrices A_n possessing the MJP or compartmental structure [LR]. The second order elliptic operator on $\mathbb{R}^d$ can be approximated by matrices of such structure. It may be shown that higher order elliptic operators cannot possess this property, which restricts the described class of methods to the 2^{nd}-order parabolic systems.

Coefficients of the elliptic operator are assumed to be smooth enough, so that the classical theory of 2^{nd}-order differential equations can be applied. This ensures the factor convergence of solution in the spaces $\dot{C}(\mathbb{R}^d)$, the errors behaving like h^α for small grid step h. We did not consider elliptic operators in divergence form for which coefficients lack necessary smoothness. In such cases we must use the theory of weak solutions in Sobolev spaces, the approach that is under development.

References

[AS] A. Ashyralyev, P. E. Sobolevskiĭ, *Well-posedness of parabolic difference equations*, Birkhuser Verlag, Basel, 1994.

[BH] J. Bramble, B. Hubbard, *A theorem on error estimation for finite difference analogues of the Dirichlet problem for elliptic equations*, Contributions to Differential Equations **2**(1963), 319–340.

[LU1] O. A. Ladyzhenskaya, N. N. Ural'tseva, *Linear and quasilinear elliptic partial differential equations*, Academic Press, New York, 1968.

[LSU] O. A. Ladyzhenskaya, V. A. Solonikov, N. N. Ural'tseva, *Linear and quasi-linear equations of parabolic type*, Nauka, Moskow, 1967.

[KR1] S. G. Krein, *Linear differential equations in Banach spaces*, Nauka, Moscow, 1967.

[LR] N. Limić, M. Rogina, *Explicit Stable Methods for the Initial Value Problem for 2^{nd}-Order Parabolic Systems*, preprint, 2000.

[MW] T. S. Motzkyn, W. Wasov, *On the approximation of linear elliptic differential equations by difference equations with positive coefficients*, J. Mathematical Physics **31**(1953), 253–259.

[RM] R. D. Richtmyer, K. W. Morton, *Difference Methods for Initial-Value Problems, 2nd ed.*, Interscience, New York, 1967.

[R] M. Rogina, *Knot Insertion Algorithm for Weighted Cubic Splines*, in: *Curves and Surfaces with Applications in CAGD* (A. Le Méhauté, C. Rabut and L. L. Schumaker, Eds.), Vanderbilt Univ. Press, 1997, 387–395.

[T1] H. Tanabe, *Equations of evolution*, Pitman, 1975.