

## ON A SECOND ORDER NON-NEGATIVITY CONSERVING METHOD

Z. Horváth (Győr, Hungary)

### 1. Introduction

Consider the following parabolic differential equation along with first kind boundary conditions

$$(1) \quad \begin{cases} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = F(x, t), & 0 < x < 1, t > 0, \\ u(x, 0) = u_0(x), & 0 < x < 1, \\ u(0, t) = u(1, t) = 0, & t \geq 0. \end{cases}$$

As it is well-known, this model problem arises by appropriate simplification of many physical problems, for example the problem of one-dimensional heat conduction. The exact solution of (1) is known to be non-negative if  $F(x, t) \geq 0$  and  $u_0(x) \geq 0$  ( $\forall x \in [0, 1], \forall t \geq 0$ ).

Using the standard difference approximation for the spatial derivative we get a semi-discrete approximation of (1)

$$(2) \quad \begin{cases} y' - Ay = f \\ y(0) = y_0 \end{cases}$$

where  $y_0 \in R^N$ ,  $A \in R^{N \times N}$ ,  $f, y : [0, \infty) \rightarrow R^N$ . This approximation preserves the non-negativity property of (1): if  $y_0 \geq 0$  and  $f(t) \geq 0$  then  $y(t) \geq 0$  since  $-a_{i,j} \geq 0$  for all  $i \neq j$ , see [1].

It is a natural requirement that a numerical method solving (2) should have this so-called non-negativity preserving property, too.

For solving (3) numerically we want to choose the time-stepsize  $\tau$  independently from the given spatial stepsize  $h = 1/(N + 1)$ . Then, in general, the requirement of preservation of non-negativity calls forth a barrier of the order of the numerical method: such a method is of order 0 or 1, see [2].

In this paper we give a method of order 2 that preserves the non-negativity unconditionally, i.e. without any conditions on stepsizes, such as e.g.  $\tau/h^2 \leq \text{const.}$

Concerning  $A$  we require only that  $A$  preserves non-negativity. We achieve this using a suitable approximation of the matrix exponential.

For the sake of brevity we introduce the following notations:

- for all  $v = (v_1, \dots, v_N)^T \in \mathbb{R}^N$ ,  $v \geq 0$  means  $v_i \geq 0$ ,  $1 \leq i \leq N$ ;
- if  $g$  is a real function defined on a set  $S$ ,  $g \geq 0$  means  $g(t) \geq 0$  for all  $t \in S$ .

## 2. Approximation of the matrix exponential

We start from the identity  $e^a = e^{\frac{1}{2}b}e^{a-b}e^{\frac{1}{2}b}$  for real numbers  $a$  and  $b$ . If we replace  $a$  and  $b$  by  $N \times N$ -matrices  $A$  and  $B$  then, in general, the equation will not be true. Namely, it is well-known (see e.g. [1]) that for all  $C, D \in \mathbb{R}^{N \times N}$  if  $CD = DC$  then  $e^C e^D = e^{C+D}$  but if  $CD \neq DC$  then generally  $e^C e^D \neq e^{C+D}$ .

So, in general,  $e^A \neq e^{\frac{1}{2}B}e^{A-B}e^{\frac{1}{2}B} =: E(A; B)$  as an approximation of the matrix exponential  $e^A$ .

Let us consider the following initial value problem:

$$(3) \quad \begin{cases} y' - Ay & = 0 \\ y(0) & = y_0 \end{cases}$$

where  $A$  is an arbitrary  $N \times N$ -matrix and  $y_0$  an arbitrary  $N$ -vector. We know that the exact solution of (3) is

$$(4) \quad y(t) = e^{At}y_0, \quad t \geq 0.$$

Let us replace in (4)  $e^{At}$  by  $E(At; Bt)$ , where  $B$  is an arbitrary real  $N \times N$ -matrix. Then

$$(5) \quad \bar{y}(t) = E(At; Bt)y_0$$

is an approximate solution of (4). In fact, let  $z = z(t)$  denote the error of  $\bar{y}$ , i.e.  $z(t) := \bar{y}(t) - y(t)$ .

**Definition 1.** We say that  $E(A; B)$  is an approximation for  $e^A$  of order  $p$  ( $p \in \mathbb{N}$ ), if  $z(t) = O(t^{p+1})$  ( $t \rightarrow 0$ ).

**Theorem 1.** For arbitrary  $A, B \in \mathbb{R}^{N \times N}$  the matrix  $E(A; B)$  is a second order approximation for  $e^A$ .

**Proof.** For all  $A, B$  the first three Taylor-coefficients at 0 of  $z$  are zero:

$$(6) \quad z(0) = \bar{y}(0) - y(0) = y_0 - y_0 = 0,$$

$$(7) \quad z'(t) = \bar{y}'(t) - Ay(t) = \left( \frac{1}{2}BE(At; Bt) + e^{\frac{1}{2}Bt}(A - B)e^{(A-B)t}e^{\frac{1}{2}Bt} + \frac{1}{2}E(At; Bt)B \right) y_0 - Ae^{At}y_0,$$

thus  $z'(0) = Ay_0 - Ay_0 = 0$ .

$$(8) \quad z''(0) = \left( \frac{1}{2}BA + \frac{1}{2}B(A - B) + (A - B)^2 + \frac{1}{2}(A - B)B + \frac{1}{2}AB \right) y_0 - A^2y_0 = 0.$$

Thus  $z(t) = O(t^3)$  ( $t \rightarrow 0$ ), which was to be proved.

**Remark 1.** By a similar calculation one can get

$$(9) \quad z'''(0) = \left( \frac{1}{4}B^2A + \frac{1}{4}AB^2 - \frac{1}{2}BAB + \frac{1}{2}A^2B + \frac{1}{2}BA^2 - ABA \right) y_0,$$

so in general  $z(t) \neq O(t^4)$  ( $t \rightarrow 0$ ).

### 3. On the preservation of non-negativity

Denoting the exact solution of (2) by  $y = y(t)$  the preservation of non-negativity can be formulated as follows (c.f. [2]).

**Definition 2.** We say that  $A \in R^{N \times N}$  preserves the non-negativity (or  $A$  is a non-negativity preserving matrix), if  $y_0 \geq 0$  and  $f \geq 0$  imply  $y \geq 0$ .

**Remark 2.** Since  $y(t) = e^{At}y_0 + \int_0^t e^{A(t-s)}f(s)ds$  ( $t \geq 0$ ), so

$$(10) \quad y \geq 0 \text{ for all } y_0 \geq 0, f \geq 0 \text{ iff } e^{At} \geq 0 \text{ } (\forall t \geq 0).$$

Moreover  $e^{At} \geq 0$  ( $\forall t \geq 0$ ) iff the elements of  $A$  that are not in the diagonal are non-negative (see e.g. [1]). Thus  $A$  preserves the non-negativity iff  $A - \text{diag } a_{i,i} \geq 0$ .

We consider now the following approximate solution of (2): let  $\tau > 0$  be an arbitrary stepsize,  $t_n := n\tau$  ( $\forall n \in N$ ) and

$$(11) \quad y_{n+1} := r(\tau A)y_n + \tau \sum_{i=1}^q r_i(\tau A)f(t_n + c_i\tau) \quad (n \in N)$$

where  $q, c_i$  are given constants and  $r, r_i$  are given functions ( $1 \leq i \leq q$ ).  $y_n$  will be regarded as an approximation of  $y(t_n)$ .

**Definition 3.** We say that a method of type (11) preserves non-negativity unconditionally, if for all  $\tau > 0, y_0 \geq 0, f \geq 0$  and for all non-negativity preserving matrices  $A$  there holds  $y_n \geq 0$  for every  $n \in N$ .

Bolley and Crouzeix proved the following important theorem ([2]):

**Theorem.** *If a method of type (11) preserves non-negativity unconditionally and  $r, r_i$  are rational functions, then the order of this method is at most one.*

Therefore, we will admit other than rational functions  $r$  and  $r_i$ 's in (11) and create (approximating  $e^{At}$  by  $E(At; Bt)$ ) an unconditionally non-negativity preserving method with a (global) error of order 2. In order to formulate such a method in a conveniently realizable way, we use the following modification of  $E(A; B)$ .

**Definition 4.** Let  $E_k(A; B) := e^{\frac{1}{2}B} \sum_{j=0}^k \frac{1}{j!} (A - B)^j e^{\frac{1}{2}B}$ , where  $k \in N$  and  $A, B \in R^{N \times N}$  are arbitrary.

**Remark 3.**

a) If  $B$  is a diagonal matrix, then  $e^{\frac{1}{2}B}$  (and thus  $E(A; B)$ ) can be easily computed:  $e^{\frac{1}{2}B} = \text{diag } e^{\frac{1}{2}b_{i,i}}$  and obviously  $e^{\frac{1}{2}B} \geq 0$ .

b) It is clear from Section 2, that  $k \geq 2$  implies  $E_k(At; Bt) = O(t^3)$  ( $t \rightarrow 0$ ).

Before formulating our main result it will be advantageous to give the following definition.

**Definition 5.** We say that  $q \in N^+, b_i, c_i$  ( $1 \leq i \leq q$ ) are the parameters of a positive quadrature of order  $p$ , if  $b_i \geq 0, c_i \in [0, 1]$  ( $1 \leq i \leq q$ ) and for any sufficiently smooth function  $g$  there holds  $\int_0^\tau g(s) ds = \tau \sum_{i=1}^q b_i g(c_i \tau) + O(\tau^{p+1})$  ( $t \rightarrow 0$ ).

**Remark 4.** For example,  $q = 2, c_1 = 0, c_2 = 1, b_1 = b_2 = \frac{1}{2}$  (i.e. the parameters of the trapezoidal rule) are parameters of a positive quadrature.

We come now to our main result.

**Theorem 2.** *Let  $k \geq 2$  and let  $q, b_i, c_i$  ( $1 \leq i \leq q$ ) be the parameters of a positive quadrature of order 2. Define the functions  $r, r_i$  as follows: for all  $H = (h_{i,j} \in R^{N \times N})$  let  $r(H) := E_k(H; \text{diag } h_{j,j})$  and  $r_i(H) := b_i r((1 - c_i)H)$ . Thus the method defined in (11) is of order 2 and preserves non-negativity unconditionally.*

**Proof.** One can easily see that the exact solution  $y = y(t)$  of (3) can be written in the form

$$(12) \quad y((n+1)\tau) = e^{A\tau}y(n\tau) + \int_0^\tau e^{A(\tau-s)}f(n\tau+s)ds, \quad n \geq 0$$

where  $\tau > 0$  is an arbitrary stepsize.

The previous remarks (3a, b), the conditions of the theorem and (12) imply that the approximation  $y_n \approx y(t_n)$  is of order two.

Let now  $A$  be an arbitrary non-negativity preserving matrix, then for every  $\tau > 0$  we have  $r(\tau A) \geq 0$  and  $r_i(\tau A) \geq 0$ . Indeed, e.g.

$$(13) \quad \begin{aligned} r(\tau A) &= E_k(\tau A; \tau \text{diag } a_{j,j}) = \\ &= e^{\frac{1}{2}\tau \text{diag}(a_{i,i})} \sum_{j=0}^k \frac{\tau^j}{j!} (A - \text{diag } a_{i,i})^j e^{\frac{1}{2}\tau \text{diag}(a_{i,i})} \end{aligned}$$

and all matrices and all coefficients on the right-hand side are non-negative, thus their product and sum is non-negative, too.

Hence the theorem is proved.

**Acknowledgement.** The author wishes to thank G. Stoyan for posing the problem of the construction of non-negativity preserving method using the matrix exponential of a diagonal matrix.

## References

- [1] **Bellman R.**, *Introduction to matrix analysis*, McGraw-Hill, N.Y., 1953.
- [2] **Bolley C., Crouzeix M.**, Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques, *R.A.I.R.O. Analyse numérique*, **12**(3) (1978), 237-245.

(Received June 30, 1992)

**Z. Horváth**

Department of Mathematics and Natural Sciences  
Széchenyi István Technical Highschool  
Győr, Hungary