Saul'yev and Group Explicit Methods

Ole Østerby

Department of Computer Science

Aarhus University, Denmark.

oleby@cs.au.dk

Abstract

The Saul'yev methods for parabolic equations are implicit in form, but can be solved explicitly and are therefore interesting in connection with non-linear problems. Abdullah's Group Explicit methods are parallel in nature and therefore interesting when using parallel computers. The main objective of this paper is to study the accuracy of these methods. Using global error estimation we show that for all these methods the time step must be bounded by the square of the space step size to ensure a global error which can be estimated. As a curiosity we show that the two original Saul'yev methods in fact solve two different differential equations.

MSC 65M06, 65M12, 65M15

1 Introduction

Explicit methods for parabolic equations are interesting because they are easy to program and especially so in connection with non-linear problems. But explicit methods must usually obey strict limitations on the time step size because of stability. Saul'yev methods ([7], [8]) are interesting because they are unconditionally stable. The time step is now restricted by consistency and it has been unclear to what extent averaging or alternation could compensate. We investigate this question using a global error estimation technique and show that the time step must in all cases be limited by the square of the space step size for reasons of accuracy.

Group Explicit methods ([1], [5]) are parallel in nature and therefore interesting in connection with parallel computers. The **GE** methods are only conditionally stable, subject to the usual restriction on the time step. Used in an alternating fashion this stability restriction is lifted. Again it is not easy to assess the global error because we are using a combination of different formulae. Using the global error estimation we conclude that the usual time-step restriction must be observed to ensure the accuracy of the computations.

It follows as a natural consequence that when computational economy is taken into account, the classical explicit method is a viable alternative.

2 Two Saul'yev Methods

In 1957 V. K. Saul'yev proposed two so-called asymmetric methods ([7], [8]) for the solution of the equation

$$u_t = bu_{xx} \tag{1}$$

The first method which we shall call **LR** can be written

$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^n - v_m^n - v_m^{n+1} + v_{m-1}^{n+1}}{h^2} \tag{2}$$

where h and k are the step sizes in the x- and t-direction, respectively, m and n are the corresponding step numbers, and v_m^n is an approximation to the true solution value $u(nk, mh)$. Here and in the following we shall use the notation of [9] (see pp. 7ff).

Equation (2) can be rewritten as

$$(1 + b\mu)v_m^{n+1} = b\mu v_{m-1}^{n+1} + (1 - b\mu)v_m^n + b\mu v_{m+1}^n \tag{3}$$

where $\mu = k/h^2$. The **LR-**formula is implicit in nature but can be solved in an explicit fashion from left to right using the (Dirichlet) boundary condition on the left boundary to get started.

The second Saul'yev method, called **RL**, for the same equation can be written

$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m-1}^n - v_m^n - v_m^{n+1} + v_{m+1}^{n+1}}{h^2} \tag{4}$$

or

$$(1 + b\mu)v_m^{n+1} = b\mu v_{m+1}^{n+1} + (1 - b\mu)v_m^n + b\mu v_{m-1}^n \tag{5}$$

This formula can also be solved in an explicit fashion, now from right to left using the (Dirichlet) boundary condition on the right boundary for the first step.

3 Group Explicit Methods

In 1983 A. R. B. Abdullah proposed a new way of applying the Saul'yev formulae in the so-called Group Explicit (**GE**) methods [1],[5]. If we apply the **LR-**formula

(3) to the point m and the **RL-**formula (5) to point $m-1$ then we have (with $\alpha = b\mu$)

$$\begin{aligned} -\alpha\ v_{m-1}^{n+1} + (1+\alpha)v_m^{n+1} &= (1-\alpha)v_m^n + \alpha\ v_{m+1}^n & (6)\\ (1+\alpha)v_{m-1}^{n+1} - \alpha\ v_m^{n+1} &= \alpha\ v_{m-2}^n + (1-\alpha)v_{m-1}^n & (7) \end{aligned}$$

The solution to these two equations in the two unknowns v_{m-1}^{n+1} and v_m^{n+1} can be written

$$\begin{aligned} v_{m-1}^{n+1} &= (a_1 v_{m-2}^n + a_2 v_{m-1}^n + a_3 v_m^n + a_4 v_{m+1}^n)/(1+2\alpha) & (8)\\ v_m^{n+1} &= (a_4 v_{m-2}^n + a_3 v_{m-1}^n + a_2 v_m^n + a_1 v_{m+1}^n)/(1+2\alpha) & (9) \end{aligned}$$

with $a_1 = \alpha(1+\alpha)$, $a_2 = 1-\alpha^2$, $a_3 = \alpha(1-\alpha)$, and $a_4 = \alpha^2$.

We therefore have formulae for v_{m-1}^{n+1} and v_m^{n+1} which can be solved independently of other pairs and therefore easily can be parallelized.

Since we often have an even number of subintervals and therefore an odd number of internal points, there will be one ungrouped point. If we start grouping points together from the left, the ungrouped point will be the last internal point to the right. We can use the **RL-**formula (5) here. Abdullah calls the resulting scheme **GER**. If we instead group points together from the right, the ungrouped point will be the first internal point (to the left) for which we use the **LR-**formula (3). This scheme is called **GEL**.

4 Stability

To study the stability of the **LR-**method we use the von Neumann approach ([3], [9], p. 23) and compute the growth factor

$$g_{LR}(\varphi) - 1 = b\mu(e^{i\varphi} - 1 - g_{LR}(1 - e^{-i\varphi})) \quad (10)$$

or

$$g_{LR} = \frac{1 + b\mu(e^{i\varphi}-1)}{1 + b\mu(1-e^{-i\varphi})} = \frac{1 - b\mu(1-\cos\varphi) + ib\mu\sin\varphi}{1 + b\mu(1-\cos\varphi) + ib\mu\sin\varphi} \quad (11)$$

The condition $|g_{LR}| \le 1$ is equivalent to

$$(1 - b\mu(1-\cos\varphi))^2 + b^2\mu^2\sin^2\varphi \le (1 + b\mu(1-\cos\varphi))^2 + b^2\mu^2\sin^2\varphi$$

or

$$-2b\mu(1-\cos\varphi) \quad \leq \quad 2b\mu(1-\cos\varphi)$$

which is always satisfied for $b > 0$, and the Saul'yev **LR-**method is therefore unconditionally stable.

A similar calculation reveals the same to be true for the **RL-**method.

The **GE-**formulae (8-9) can be written in matrix form

$$\left\{ \begin{array}{c} v_{m-1}^{n+1} \\ v_m^{n+1} \end{array} \right\} \quad = \quad \frac{1}{1+2\alpha} \left\{ \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \\ a_4 & a_3 & a_2 & a_1 \end{array} \right\} \left\{ \begin{array}{c} v_{m-2}^n \\ v_{m-1}^n \\ v_m^n \\ v_{m+1}^n \end{array} \right\} \tag{12}$$

and a full **GER-**step can be written

$$v^{n+1} \quad = \quad A_{GER} v^n + q_{GER}^n \tag{13}$$

with $v^n = \{v_1^n, v_2^n, \ldots, v_{M-1}^n\}^T$, $q_{GER}^n = \{-b_1 v_0^n, -b_4 v_0^n, 0, 0, \ldots, 0, -b_5 v_M^n\}^T$,

$$A_{GER} \quad = \quad \left\{ \begin{array}{cccccccc} b_2 & b_3 & b_4 & & & & & \\ b_3 & b_2 & b_1 & & & & & \\ & b_1 & b_2 & b_3 & b_4 & & & \\ & b_4 & b_3 & b_2 & b_1 & & & \\ & & . & . & . & . & & \\ & & & b_4 & b_3 & b_2 & b_1 & \\ & & & & & c & d \end{array} \right\} \tag{14}$$

with $b_i = a_i/(1+2\alpha)$, $i = 1, 2, 3, 4$, $c = \alpha/(1+\alpha)$, and $d = (1-\alpha)/(1+\alpha)$.

Similarly a **GEL-**step can be written

$$v^{n+1} \quad = \quad A_{GEL} v^n + q_{GEL}^n \tag{15}$$

with $q_{GEL}^n = \{-c v_0^n, 0, 0, \ldots, 0, -b_4 v_M^n, -b_1 v_M^n\}^T$ and

$$A_{GEL} \quad = \quad \left\{ \begin{array}{cccccccc} d & c & & & & & & \\ b_1 & b_2 & b_3 & b_4 & & & & \\ b_4 & b_3 & b_2 & b_1 & & & & \\ & & b_1 & b_2 & b_3 & b_4 & & \\ & & b_4 & b_3 & b_2 & b_1 & & \\ & & & . & . & . & . & \\ & & & & b_4 & b_3 & b_2 & \end{array} \right\} \tag{16}$$

When $0 < \alpha \leq 1$ then all elements in the A-matrices are non-negative and each row sum is ≤ 1 showing that **GER** and **GEL** are stable provided $0 < b\mu \leq 1$.

Using the von Neumann technique on (8) we get

$$\begin{aligned}(1+2\alpha)g &= \alpha^2 e^{-2i\varphi} + \alpha(1-\alpha)e^{-i\varphi} + 1 - \alpha^2 + \alpha(1+\alpha)e^{i\varphi} \\ &= 1 + 2\alpha\cos\varphi - 2\alpha^2\sin^2\varphi + 2i\alpha^2\sin\varphi(1-\cos\varphi) \qquad (17)\end{aligned}$$

and on (9) we get the complex conjugate. The extremal values of the imaginary part occur for $\varphi = 0$ and $\varphi = 2\pi/3$. In the latter case we get

$$(1+2\alpha)g = 1 - \alpha - \frac{3}{2}\alpha^2 + \frac{3\sqrt{3}}{2}i\alpha^2. \qquad (18)$$

When $\alpha = 1$ we have

$$3g = -\frac{3}{2} + \frac{3\sqrt{3}}{2}i \qquad (19)$$

such that $|g| = 1$. When $\alpha = 1 + \varepsilon$ a short calculation shows that $|g| > 1$ when $\varepsilon > 0$, so we can conclude that **GER** and **GEL** are unstable when $b\mu > 1$.

5 Consistency

In order to check for consistency we apply the difference operator for the **LR**-method (cf. [9], p. 30) on a smooth function ψ and expand around the point $((n+\frac{1}{2})k, mh)$:

$$\begin{aligned}P^{LR}_{k,h}\psi &= \frac{\psi^{n+1}_m - \psi^n_m}{k} - b\frac{\psi^n_{m+1} - \psi^n_m - \psi^{n+1}_m + \psi^{n+1}_{m-1}}{h^2} \\ &= \psi_t + \frac{1}{24}k^2\psi_{ttt} + \cdots \\ &\quad - \frac{b}{h}(\psi_x + \frac{1}{2}h\psi_{xx} + \frac{1}{6}h^2\psi_{xxx} + \frac{1}{24}h^3\psi_{xxxx} \\ &\quad\quad - \frac{1}{2}k\psi_{xt} - \frac{1}{4}kh\psi_{xxt} - \frac{1}{12}kh^2\psi_{xxxt} \\ &\quad\quad + \frac{1}{8}k^2\psi_{xtt} + \frac{1}{16}k^2h\psi_{xxtt} - \frac{1}{48}k^3\psi_{xttt} + \cdots) \\ &\quad + \frac{b}{h}(\psi_x - \frac{1}{2}h\psi_{xx} + \frac{1}{6}h^2\psi_{xxx} - \frac{1}{24}h^3\psi_{xxxx} \\ &\quad\quad + \frac{1}{2}k\psi_{xt} - \frac{1}{4}kh\psi_{xxt} + \frac{1}{12}kh^2\psi_{xxxt} \\ &\quad\quad + \frac{1}{8}k^2\psi_{xtt} - \frac{1}{16}k^2h\psi_{xxtt} + \frac{1}{48}k^3\psi_{xttt} + \cdots)\end{aligned}$$

$$= \psi_t - b\psi_{xx} + b\frac{k}{h}\psi_{xt} + \frac{1}{24}k^2\psi_{ttt} \qquad (20)$$
$$- \frac{1}{12}bh^2\psi_{xxxx} + \frac{1}{6}bkh\psi_{xxxt} - \frac{1}{8}bk^2\psi_{xxtt} + \frac{1}{24}b\frac{k^3}{h}\psi_{xttt} + \cdots$$

We recognize the differential operator for (1) in the first two terms, and the remaining terms (which constitute what we call the *local truncation error*) must tend to 0 as h and k tend to 0 for the **LR-**method to be consistent. We must therefore require that k tends to 0 faster than h. For the method to be first order (in h) we must require k to be $O(h^2)$. This is a requirement much like the stability condition for the classical explicit method (cf. [9], p. 25), although we are no longer bound by the proportionality constant 0.5. On the other hand the **LR-**method is then only of order 1 in h.

A similar calculation for the **RL-**method gives

$$P^{RL}_{k,h}\psi = \psi_t - b\psi_{xx} - b\frac{k}{h}\psi_{xt} + \frac{1}{24}k^2\psi_{ttt} \qquad (21)$$
$$- \frac{1}{12}bh^2\psi_{xxxx} - \frac{1}{6}bkh\psi_{xxxt} - \frac{1}{8}bk^2\psi_{xxtt} - \frac{1}{24}b\frac{k^3}{h}\psi_{xttt} + \cdots$$

and similar comments on consistency and order apply for the **RL-**method. We note that the annoying $\frac{k}{h}$-term appears with opposite sign in the two expressions. Saul'yev himself did not advise to use these methods by themselves ([8], p. 29) but instead suggested to use **LR** and **RL** alternately, e.g. **LR** in the odd steps and **RL** in the even steps ([7], [8], p. 43), in order that the $\frac{k}{h}$-terms might partially compensate each other. Another suggestion ([2], [6]) with the same intention is to compute with both **LR** and **RL** in each step and take the average. This, however, means doubling the computational work. We shall refer to these methods by the names **ALT** and **AV**, respectively.
In [2] the average of **LR** and **RL** is defined by computing with **LR** and **RL** separately and taking the average at the end. We call this version **AVB**.
It is obvious that either approach is unconditionally stable.
It is less obvious what the consistency requirements are.
Practical experiments indicate that it is advisable to keep $b\mu \leq 1$.

6 A Word of Caution

The consistency requirement $\frac{k}{h} \to 0$ is concerned with the situation where the step sizes tend to 0 and we wish the numerical solution to converge towards the true solution. But in practice we compute with fixed, finite step sizes and wonder what the error might be.

Equation (2) can be rewritten as (cf. [8], pp. 30f)

$$\begin{aligned}\frac{v_m^{n+1}-v_m^n}{k} &= b\frac{v_{m+1}^n-2v_m^n+v_{m-1}^n}{2h^2}+b\frac{v_{m+1}^{n+1}-2v_m^{n+1}+v_{m-1}^{n+1}}{2h^2}- \\ & \quad b\frac{k}{h}\frac{v_{m+1}^{n+1}-v_{m-1}^{n+1}-v_{m+1}^n+v_{m-1}^n}{2hk}. \end{aligned} \tag{22}$$

We recognize the first two terms on the right-hand side as the (second order) Crank-Nicolson approximation to $u_{xx}((n+\frac{1}{2})k, mh)$ and the last term as an approximation to u_{xt} at the same point. So the **LR-**method is actually computing an approximate solution to

$$u_t = bu_{xx} - b\frac{k}{h}u_{xt} \tag{23}$$

a result which is actually apparent from formula (20).

Similarly it can be shown that the **RL-**method produces an approximate solution to

$$u_t = bu_{xx} + b\frac{k}{h}u_{xt} \tag{24}$$

When $\frac{k}{h} \to 0$ both these equations tend to the desired $u_t = bu_{xx}$, so everything works fine in the limit, but for finite step sizes there is a difference.

7 Consistency of the GE Methods

We begin by rewriting (9):

$$\begin{aligned}(1+2\alpha)(v_m^{n+1}-v_m^n) &= \alpha(v_{m+1}^n-2v_m^n+v_{m-1}^n)+\alpha^2(v_{m+1}^n-v_{m-1}^n+v_{m-2}^n-v_m^n) \\ &= \alpha(h^2v_{xx}+\frac{1}{12}h^4v_{xxxx}+\alpha^2(2hv_x+\frac{1}{3}h^3v_{xxx}) \\ &\quad +\alpha^2(-2hv_x+2h^2v_{xx}-\frac{4}{3}h^3v_{xxx}+\frac{2}{3}h^4v_{xxxx})+\cdots \\ &= \alpha(1+2\alpha)h^2v_{xx}-\alpha^2h^3v_{xxx}+\frac{1}{12}\alpha(1+8\alpha)h^4v_{xxxx}+\cdots\end{aligned}$$

such that

$$\begin{aligned}P_{k,h}^{(9)} &= \frac{\psi_m^{n+1}-\psi_m^n}{k}-\{\text{r.h.s.}\} \\ &= \psi_t+\frac{1}{2}k\psi_{tt}-b\psi_{xx}+\frac{\alpha}{1+2\alpha}bh\psi_{xxx}+\cdots\end{aligned} \tag{25}$$

showing that (9) is consistent with $u_t = bu_{xx}$ and of first order in h and k. For formula (8) we have in a similar fashion:

$$P_{k,h}^{(8)} = \psi_t + \frac{1}{2}k\psi_{tt} - b\psi_{xx} - \frac{\alpha}{1+2\alpha}bh\psi_{xxx} + \cdots \quad (26)$$

We note that the first order (in h) term appears with different sign in (25) and (26). Since we use the two formulas (9) and (8) alternately at even and odd points we may wonder if there is a compensating effect. We could also encourage this by using **GER** and **GEL** alternately, **AGE** (**(S)AGE** in [1] and [5]), or by taking the average (**GE-AV**). **AGE** is unconditionally stable ([1], [5]) but practical experiments indicate that the use of values of $b\mu$ larger than 1.0 is inadvisable

8 Computational Economy

The computational work involved in using Saul'yev or **GE** methods is proportional to the number of grid points as it is with the classical explicit (**EX**), implicit (**IM**), or Crank-Nicolson (**CN**, [4]) methods. The difference is the number of simple arithmetic operations (**SAO**) such as additions (**A**), multiplications (**M**), and divisions (**D**) per grid point. These are listed in Table 1.

Table 1: Simple arithmetic operations for the methods

Method	**A**	**M**	**D**	**SAO**
LR,RL,ALT	2	2		4
AV	5	5		10
AVB	4	4		8
GEL,GER,AGE	3	4		7
GE-AV	7	9		16
EX	2	2		4
IM	3	3	2	8
CN	5	5	2	12

We shall later see that the extra work needed for **AV** and **GE-AV** is not compensated by better results. On the other hand the extra work needed by **CN** (and **IM**) is compensated (somewhat) by the lesser restrictions on the time step size. Note that we have restrictions on $\mu = k/h^2$ for Saul'yev methods because of consistency, and for **GE** methods because of stability.

9 Estimating the Global Error

When assessing the global error of a numerical method we often compare the numerical solution with the true solution for a test problem where such is known. But it is more useful in practice to be able to estimate the global error and for this we use the technique of [9], ch. 10.

Here we assume that the numerical solution for small h can be written as a series:

$$v = u - hc - h^2 d - h^3 f - \cdots \tag{27}$$

where v is the numerical solution, computed with step size h, u is the true solution, and c, d, and f are (unknown) auxiliary functions.

Remark. If $c \neq 0$ then the method is called first order, if $c \equiv 0$ and $d \neq 0$ the method is second order etc. □

To gain information on the order, p, and the error, $u - v$, we perform calculations with h, $2h$, and $4h$:

$$v_1 = u - hc - h^2 d - h^3 f - \cdots \tag{28}$$
$$v_2 = u - 2hc - 4h^2 d - 8h^3 f - \cdots \tag{29}$$
$$v_3 = u - 4hc - 16h^2 d - 64h^3 f - \cdots \tag{30}$$

We now calculate two differences:

$$v_1 - v_2 = hc + 3h^2 d + 7h^3 f + \cdots \tag{31}$$
$$v_2 - v_3 = 2hc + 12h^2 d + 56h^3 f + \cdots \tag{32}$$

and the so-called *order ratio*

$$q = \frac{v_2 - v_3}{v_1 - v_2} = 2\frac{hc + 6h^2 d + 28h^3 f}{hc + 3h^2 d + 7h^3 f}. \tag{33}$$

The order ratio can be calculated for every fourth point and will often reveal the order of the method when h is sufficiently small such that hc dominates the following terms. If q assumes values near 2 for many points then the method is most likely of order $p = 1$. In this case the difference $v_1 - v_2$ is an estimate of the global error, $u - v_1$:

$$v_1 - v_2 = u - v_1 + 2h^2 d + 6h^3 f + \cdots \tag{34}$$

If the order ratio takes on values near 4 then the auxiliary function, c, is probably identically zero, the order of the method is 2, and the error estimate is $(v_1 - v_2)/3$:

$$\frac{v_1 - v_2}{3} = u - v_1 + \frac{4}{3}h^3 f + \cdots \tag{35}$$

Isolated deviations of the order ratio from 2 (or 4) may occur for points in the x-t-plane close to a line where the auxiliary function c (or d) is 0. The function c (d) will often change sign across such a line which is reflected in small values and a sign change in the difference $v_1 - v_2$.

Remark. Other values of p – even non-integral – may occur, but a closer analysis is required to decide whether to put any trust into such values. □

It is important to check the order ratio to correctly determine the order and decide which error estimate to apply. If examination of the order ratio is inconclusive, the reason might be interference from the next term(s) in the series. If we e.g. encounter a $q \approx 2.8$, how do we decide if we have a first order method with a relatively large second order term ($h^2 d \approx 0.22hc$) or a second order method with a relatively large third order term ($h^3 f \approx -0.1h^2 d$) or possibly a single term with $p = 1.5$. In this case a reduction of the step size might give a clearer picture, since a smaller step size will reduce the relative importance of the succeeding terms. In the three cases above a reduction of h by a factor 2 would most probably lead to order ratios of 2.5, 3.5 or 2.8, respectively. A further reduction by a factor 2 would show values 2.3, 3.76 or 2.8, respectively, thus making a decision easier. If a reduction of h does not help then the reason might be that the basic assumption (27) does not hold for the numerical method on this problem and a reliable order and error estimation can not be obtained in this way.

10 Independent Step Sizes

In fact we do have two different step sizes, h and k, and they can be varied independently within certain bounds. We may therefore take as a basic assumption that for small h and k

$$v_1 = u - hc - kd - hke - h^2 f - k^2 g - \cdots \tag{36}$$

where c, d, e, f, and g, are auxiliary functions of t and x. In order to gain information on these we perform extra calculations where we first double the step size h (twice) and then k

$$v_2 = u - 2hc - kd - 2hke - 4h^2 f - k^2 g - \cdots \tag{37}$$
$$v_3 = u - 4hc - kd - 4hke - 16h^2 f - k^2 g - \cdots \tag{38}$$
$$v_4 = u - hc - 2kd - 2hke - h^2 f - 4k^2 g - \cdots \tag{39}$$
$$v_5 = u - hc - 4kd - 4hke - h^2 f - 16k^2 g - \cdots \tag{40}$$

On a grid with step sizes $4h$ and $4k$ we can now compute

$$v_1 - v_2 = hc + hke + 3h^2 f + \cdots \tag{41}$$

$$v_2 - v_3 = 2hc + 2hke + 12h^2 f + \cdots \quad (42)$$
$$v_1 - v_4 = kd + hke + 3k^2 g + \cdots \quad (43)$$
$$v_4 - v_5 = 2kd + 2hke + 12k^2 g + \cdots \quad (44)$$

and

$$q_h = \frac{v_2 - v_3}{v_1 - v_2} \quad \text{and} \quad q_k = \frac{v_4 - v_5}{v_1 - v_4}. \quad (45)$$

An hk-term seldom arises and cannot be detected by the previous calculations. In order to be able to isolate a possible hk-term we double both the step sizes:

$$v_6 = u - 2hc - 2kd - 4hke - 4h^2 f - 4k^2 g - \cdots \quad (46)$$

and compute

$$(v_1 - v_6) - (v_1 - v_2) - (v_1 - v_4) = hke + \cdots \quad (47)$$

If the order ratio q_h (q_k) is close to 2.0 then the order in h (k) is probably 1 and the difference $v_1 - v_2$ ($v_1 - v_4$) is a reasonable estimate of the contribution to the error due to the finite step size h (k). If the order ratio is close to 4.0 then the order is probably 2, the first order term is 0, and the error estimate is one third of the relevant difference.

If one or both the order ratios are not close to 2.0 (or 4.0 or 8.0 or ...) then the interference from succeeding terms may be too big and smaller step size(s) are called for. If the information from the order ratios is still inconclusive then possibly the basic assumption (36) does not hold.

11 Computational Examples

We have chosen three test examples, all with the region $0 \le x \le 1$, $0 \le t \le 1$:

1.

$$u_t = u_{xx}; \qquad u(0,x) = sin(\pi x); \qquad u(t,0) = u(t,1) = 0$$

with true solution

$$u(t,x) = sin(\pi x) e^{-\pi^2 t}.$$

2. [2] formula (23)

$$u_t = u_{xx}; \qquad u(0,x) = 0; \qquad u(t,0) = 1; \qquad u(t,1) = 1$$

with true solution

$$u(t,x) = 1 - \frac{4}{\pi} \sum_{n=1,3,\ldots} \frac{sin(n\pi x)e^{-n^2\pi^2 t}}{n} \qquad (t > 0).$$

3. [6] formulas (5.1–4)

$$u_t = u_{xx}; \qquad u(0,x) = 1; \qquad u(t,0) = 0; \qquad u(t,1) = 1$$

with true solution

$$u(t,x) = x - \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{sin(n\pi x)e^{-n^2\pi^2 t}}{n} \qquad (t > 0).$$

Test problems **2** and **3** have a discontinuity at $t = 0$. The methods we compare are **LR, RL, AV, AVB, ALT, GER, GEL, GE-AV, AGE, IM, CN**.

Table 2: Order ratio q_h for **GER** with $h = 1/40$, $k = 1/6400$ on problem **1.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	3.49	4.00	3.88	4.00	4.00	4.00	4.12	4.01	4.51
0.2	3.76	4.02	3.96	4.02	4.02	4.02	4.08	4.02	4.27
0.3	3.86	4.03	3.99	4.03	4.03	4.03	4.07	4.03	4.20
0.4	3.92	4.05	4.02	4.05	4.05	4.05	4.08	4.05	4.18
0.5	3.96	4.06	4.04	4.06	4.06	4.06	4.09	4.06	4.17

Table 3: Order ratio q_k for **GER** with $h = 1/40$, $k = 1/6400$ on problem **1.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	1.47	1.32	0.86	-3.29	4.00	2.89	2.61	2.50	2.45
0.2	3.63	3.68	3.75	3.86	4.00	4.16	4.33	4.49	4.60
0.3	3.99	3.99	4.00	4.00	4.00	4.00	4.00	4.01	4.01
0.4	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
0.5	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00

In Tables 2 and 3 we show the order ratios q_h and q_k for $x = 0.1$ (0.1) 0.9 and $t = 0.1$ (0.1) 0.5 for **GER** on Problem **1** with $h = 1/40$, $k = 1/6400$. The order ratios are all close to 4.0 indicating that **GER** is second order in both h and k and that our assumption (36) is valid and the auxiliary functions c and d are 0.

Table 4: h-component of the error estimate ($\times 10^6$) for **GER** with $h = 1/40$, $k = 1/6400$ on problem **1.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	-59	-111	-153	-180	-189	-180	-153	-111	-58
0.2	-44	-83	-114	-134	-141	-134	-114	-83	-44
0.3	-24	-46	-64	-75	-79	-75	-64	-46	-24
0.4	-12	-23	-32	-37	-39	-37	-32	-23	-12
0.5	-6	-11	-15	-17	-18	-17	-15	-11	-6

Table 5: Order ratio q_h for **IM** with $h = 1/40$, $k = 1/6400$ on problem **3.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.10	4.07	4.04	4.00	3.85	5.61	4.26	4.16	4.11	4.09
0.20	4.00	3.99	3.98	3.98	3.97	3.96	3.95	3.94	3.93
0.30	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
0.40	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02
0.50	4.04	4.04	4.04	4.04	4.04	4.04	4.04	4.04	4.04

Table 4 show the estimates ($\times 10^6$) of the h-component of the error as calculated by $(v_1 - v_2)/3$. The k-component of the error is negligible due to the small time step which is necessary for reasons of stability.

Tables 5 and 6 show the order ratios for **IM** on Problem 3 demonstrating that **IM** is second order in h and first order in k and that (36) is valid. The two components of the error estimate are shown in Tables 7 and 8 calculated as $(v_1 - v_2)/3$ and $v_1 - v_4$, respectively.

Table 6: Order ratio q_k for **IM** with $h = 1/40$, $k = 1/6400$ on problem **3.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.10	2.00	2.00	2.00	2.00	1.99	1.99	1.98	1.97	1.96
0.20	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
0.30	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
0.40	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
0.50	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01

Table 7: h-component of the error estimate ($\times 10^6$) for **IM** with $h = 1/40$, $k = 1/6400$ on problem **3.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.10	-22	-36	-34	-19	3	25	37	36	22
0.20	-15	-28	-37	-43	-44	-41	-34	-24	-13
0.30	-10	-20	-27	-32	-33	-32	-27	-19	-10
0.40	-6	-11	-15	-18	-19	-18	-15	-11	-6
0.50	-3	-5	-8	-9	-9	-9	-8	-5	-3

Table 8: k-component of the error estimate ($\times 10^6$) for **IM** with $h = 1/40$, $k = 1/6400$ on problem **3.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.10	-102	-179	-218	-214	-178	-126	-75	-36	-13
0.20	-43	-82	-112	-130	-134	-126	-106	-76	-40
0.30	-23	-44	-61	-72	-75	-72	-61	-44	-23
0.40	-12	-22	-30	-36	-37	-36	-30	-22	-12
0.50	-5	-10	-14	-17	-17	-17	-14	-10	-5

The calculation of $v_2, \ldots, v_5$ represent a 150 % increase in computer time relative to the original v_1. What we hope to get in return is information on the size and shape of the error. If the values of the order ratios q_h and q_k are close to 2.0 or 4.0 (cf. Tables 2, 3, 5, 6 then we can get reliable error estimates and information on how to best adjust the step sizes in order to obtain a desired accuracy. Another possibility is to use Richardson extrapolation in one or both directions (i.e. to add the error estimates to the computed solution) to achieve better results. In this case extra calculations are needed to provide error estimates for the extrapolated results.

The assumption (36) works fine for **GER**, **GEL**, **IM**, and **CN**, but not so well for the other methods. As an example we show in Table 9 the computed order ratios q_h for **LR** on Problem 1. Reduction of the step sizes gives no improvement, and it seems reasonable to conclude that the assumption (36) does not apply in this case.

Table 9: Order ratio q_h for **LR** with $h = 1/40$, $k = 1/6400$ on problem **1.**

$t \setminus x$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.10	-0.71	-1.33	-3.26	-	4.03	2.30	1.71	1.41	1.23
0.20	-2.96	-6.49	-	7.33	3.95	2.81	2.24	1.89	1.66
0.30	-10.26	-	10.49	5.60	3.93	3.09	2.58	2.24	1.99
0.40	-	13.47	7.21	5.04	3.93	3.26	2.81	2.49	2.24
0.50	16.29	8.79	6.13	4.76	3.93	3.38	2.98	2.67	2.44

12 An extended assumption

For the Saul'yev methods the local truncation error involves terms with k/h, and an analysis along the lines of [9], ch. 9 indicate that we might expect terms with k/h and $(k/h)^2$ etc. in the expansion for the global error. We therefore introduce two new auxiliary functions a and b and take as our basic assumption that for small h and k and k/h

$$v_1 \;=\; u - \frac{k}{h}a - \frac{k^2}{h^2}b - hc - kd - hke - h^2 f - k^2 g - \cdots \tag{48}$$

and we also add two more calculations (v_7 and v_8) with $(4h, 2k)$ and $(2h, 4k)$, respectively.

At each point in the coarse $(4h, 4k)$ grid we now have 8 equations $(v_1, \ldots, v_8)$ in the 8 unknowns $u, \frac{k}{h}a, \frac{k^2}{h^2}b, hc, kd, hke, h^2 f, k^2 g$. One problem with this set-up is that we always get a solution, but with no indication whether the original assumption bears any relation to reality. To check this we must repeat the calculations with e.g. $h/2$ or $k/2$ or both and see if the solutions show the expected dependence of h and k. Small solution values will be contaminated by the succeeding terms in the expansion, so we can only hope to find approximate values for the leading two or three terms.

We show the results of the calculations at $t = 0.3$ where we are relatively far away from the discontinuities at $t = 0$ and at $x = 0.3$ which seems like a 'typical' point. We have not chosen the midpoint $x = 0.5$ because the symmetry in problems **1** and **2** gives a very small and atypical error for **LR** and **RL**.

In Table 10 we have given the leading terms of the error expansion (48) as calculated from the 8 equations for Problem **1** at $t = 0.3$, $x = 0.3$ using $h = 1/40$ and $k = 1/6400$ using the various Saul'yev and **GE** methods and (for comparison) **IM** and **CN**. With this choice of step sizes it is possible to quadruple k without crossing the stability limit for the **GE-**methods and to quadruple h and still have

Table 10: Error terms and errors times 10^6 for Problem **1** at $t = 0.3$ and $x = 0.3$ with $h = 1/40$ and $k = 1/6400$. Also shown are the order ratios q_h and q_k.

method	k/h	k	h^2	k^2/h^2	error	acc.	q_h	q_k
LR	-259		-65	7	-316	1		
RL	259		-65	6	201	1		
AV			-64	-43	-109	-2		3.93
AVB			-65	6	-57	2	3.94	4.01
ALT			-65	-46	-109	2		4.01
GER			-65		-64	1	3.99	4.00
GEL			-65		-64	1	4.07	4.00
GE-AV			-68		-33	35		
AGE			-65	47	-18	0		4.00
IM		-95	-65		-159	1	4.03	2.00
CN			-65		-64	1	4.03	4.00

values with $\Delta x = 0.1$. The column *acc* is the accuracy of the error estimate defined as the actual error minus the error contributions listed. The errors and error contributions are given in units of 10^{-6}. The last two columns give the order ratios as computed by (45) when these are close to 4.00 (or 2.00)

In the theoretical error expansion we expect a k^2-term but since we use a very small time step this term will usually not be among the leading ones.

The order ratios behave nicely for **GER**, **GEL**, **IM**, and **CN** indicating second order in both h and k, except for **IM** which is first order in k. The components of the error estimate as computed by (35) or (34) agree within one unit (times 10^{-6}) with the values listed in the table. For **AV**, **AVB**, **ALT**, and **AGE** q_k is close to 4.00, because for fixed h a k^2/h^2-term behaves like k^2. The difference is revealed in q_h which does not attain values near 4.00 except for **AVB** where the k^2/h^2-term is relatively small compared to the h^2-term. In all cases the k-component of the error estimate computed as $(v_1 - v_4)/3$ agrees within one unit (times 10^{-6}) with the values listed in the table.

We notice a dominant k/h-term for **LR** and with opposite sign for **RL**.
We notice also that this k/h-term is eliminated (as claimed in [2], [6], [8]) by using **LR** and **RL** alternately or by taking average. We notice, however, that a k^2/h^2-term appears and is amplified in **AV** and **ALT**.

Remark. Because of symmetry in Problems **1** and **2** the auxiliary function $a(t, x)$ is equal to 0 for $x = 0.5$ and therefore the k/h-term vanishes at $x = 0.5$ for **LR** and **RL** thus reducing the error considerably. □

The next term for **LR** and **RL** (and **IM**) and the leading term for all other methods is the h^2-term which has the same value for all methods (-65, $+54$, and -27 for Problems **1**, **2**, and **3**, respectively) indicating that the auxiliary function $f(t,x)$ is the same for all methods.

Comparing all methods the original **LR** and **RL** give the largest errors (away from $x = 0.5$). Since the h^2-term is the same for all methods the only difference appears when there is a second term, and depending on the signs, this extra term might increase or decrease the error. In the cases considered this gives a slight advantage to **AVB** and **AGE**.

In all cases but one there is very good agreement between the error estimates and the actual errors. The one exception is **GE-AV** where the basic assumption (48) does not seem to fit very well.

Remark. The classical explicit method, **EX**, is not included in the comparison because this would necessitate yet another halving of the time step, k, but we can mention that whenever both **EX** and **IM** can be applied they show the same h^2-component of the global error and a k-component with opposite sign and the same magnitude. □

Series expansions such as (36) and (48) tend to work best when the function is differentiable. When we have jump discontinuities such as in Problems **2** and **3** we can expect more interference from higher order terms. The results for these two problems are not as clear as for Problem **1**. First of all the results for **GER** and **GEL** do not fit in with the theory when $\mu = 1$ which is the stability limit for these methods. (A similar effect is seen for the classical explicit method with $\mu = 0.5$). A reduction of the step sizes is necessary to clarify the situation and then the pattern from Table 10 is repeated with the exception that a k-term appears for **GER**, **GEL**, **LR**, and **RL** in Problem **3**.

Table 11: Order ratios and error terms times 10^6 at $t = 0.3$ and $x = 0.3$ with $h = 1/40$ and $k = 1/12800$ using **GER** on the three problems.

Problem	q_h	q_k	$(v_1 - v_2)/3$	$v_1 - v_4$	estimate	error
1	4.01	3.99	-64	0	-64	-64
2	3.97	4.84	54	0	54	54
3	3.97	2.00	-27	10	-17	-17

In Table 11 we give the order ratios and error components (times 10^6) at $t = 0.3$, $x = 0.3$ computed with $h = 1/40$, $k = 1/12800$ using **GER** on Problems **1** - **3**. We note that the first order k-term in Problem **3** is clearly detected by q_k and correctly estimated by $v_1 - v_4$. For Problem **2** the order ratio q_k is typically close

to 4.0 for $t \geq 0.3$ but since the chosen point $(t, x) = (0.3, 0.3)$ happens to lie close to a line in (t, x)-space where the auxiliary function g is zero the stated value is atypical. Since the value of $v_1 - v_4$ is very small, the uncertainty in q_k is of minor importance.

13 Conclusions

We have studied and compared various versions of the Saul'yev methods ([2], [6], [8]) and Abdullah's Group Explicit methods ([1], [5]) for the parabolic equation $u_t = bu_{xx}$. The time step k must be restricted to $k \leq h^2/b$ for the Saul'yev methods to enable us to provide a reliable global error estimate. A similar restriction applies to the **GE**-methods for reasons of stability, with a strict inequality when a jump discontinuity is present at $t = 0$.

LR and **RL** cannot be recommended because of a rather large k/h-component of the error. **AV**, **ALT**, and **AGE** have a large k^2/h^2-component which is difficult to estimate. **AVB** is a borderline case and **GE-AV** fails to comply with the theory. Abdullah's **GER** and **GEL** methods show nice behaviour with a single h^2-term in the error, and sometimes a k-term. Both can be readily detected through a study of the order ratios, and the error components can be estimated effectively. Looking at the computational economy a viable alternative is **EX** with half the time step, being also explicit and highly parallelizable.

We do not recommend the calculation of $v_2, \ldots, v_8$ and the solution of the eight linear equations as a general practice. We do, however, recommend to compute $v_1, \ldots, v_5$ and the order ratios q_h and q_k. They give important and valuable information on the size and shape of the error. The order ratios should be computed for a larger set of points as seen in Tables 2, 3, 5, 6, and 9 such that we can get the broader picture.

References

[1] A. R. B. Abdullah, *The study of some numerical methods for solving parabolic partial differential equations*,
Ph.D. thesis, Loughborough Univ. of Tech., 1983

[2] H. Z. Barakat and J. A. Clark,
On the Solution of the Diffusion Equations by Numerical Methods,
Trans. ASME J. Heat Transfer, 88 (1966), pp. 421–427.

[3] G. G. O'Brien, M. A. Hyman, and S. Kaplan,
A Study of the Numerical Solution of Partial Differential Equations,
J. Math. Phys., 29 (1951), pp. 223–251. doi:10.1002/sapm1950291223

[4] J. Crank and P. Nicolson, *A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type*,
Proc. Cambridge Philos. Soc., 43 (1947), pp. 50–67.
Reprinted in Adv. Comput. Math., 6 (1996), pp. 207–226.
doi:10.1007/BF02127704

[5] D. J. Evans and A. R. B. Abdullah, *Group Explicit Methods for Parabolic Equations*, Int. J. Computer Math., 14 (1983), pp. 73-105.

[6] B. K. Larkin, *Some Stable Explicit Difference Approximations to the Diffusion Equation*, Math. Comp. 18 (1964), pp. 196-202.

[7] V. K. Saul'yev, *A method of numerical solution for the diffusion equation*, Dokl. Akad. Nauk SSSR 115 (1957) pp. 1077-1079 (in Russian).

[8] V. K. Saul'yev, *Integration of Equations of Parabolic Type by the Method of Nets*, Pergamon Press, Oxford, 1964
Translated from the russian edition (Fizmatgiz, Moscow, 1960).

[9] O. Østerby, *Numerical Solution of Parabolic Equations*,
Department of Computer Science, Aarhus University, 2015
doi:10.7146/aul.5.5.