Ivo Hrubec; Jiří Taufer

# Comparison of the factorization method and the method of combination of solutions

*Aplikace matematiky*, Vol. 17 (1972), No. 3, 209–224

Persistent URL: http://dml.cz/dmlcz/103410

# COMPARISON OF THE FACTORIZATION METHOD AND THE METHOD OF COMBINATION OF SOLUTIONS

IVO HRUBEC and JIŘÍ TAUFER

(Received March 19, 1971)

There is a number of methods of solution of the boundary value problem for a system of ordinary differential equations. One group of these methods consists in replacing the given problem by a sequence of problems with initial conditions. Neither the labouriousness nor the numerical stability of these methods are equivalent from the point of view of their numerical realization. A theoretical comparison of these methods is performed in [5]. The aim of the present paper is to show, on simple experimental examples, the unsatisfactory properties of the very simple and frequently used method of combination of solutions in comparison with the method of factorization. It was shown in [3] or [4] that the factorization method has certain advantages as regards its numerical realization. In [1] it was demonstrated that the combination method may happen to be numerically very unstable even for problems very stable from the physical view point. Experimental results corroborate fully this proposition. Even worse, it occurs that the less the solution of the original problem is sensitive to perturbations of coefficients, the less the combination method is stable. Moreover, we demostrate in the present paper that the combination method cannot be "saved" by increasing the precision of computation, e.g. by the double arithmetics. The analysis of the studied numerical processes is not presented in the paper; we give only illustrative examples showing the behaviour of the methods under consideration. The theoretical analysis of a wide class of methods, including both the combination and the factorization methods, is given in [5]. We consider the publication of merely experimental results to be useful since they explain immediately the frequent failure of the combination method, hence being possibly a warning as well as a clue for the choice of methods.

## FORMULATION OF THE COMBINATION METHOD AND THE COMPOSED FACTORIZATION METHOD FOR THE DIFFERENTIAL EQUATION OF THE SECOND ORDER

First of all, let us formulate a problem which we are going to solve by both the combination and the factorization methods.

**Problem 1.** Let the functions $p(t)$, $q(t)$ and $f(t)$ defined on the interval $\langle a, b\rangle$ be such that $1/p(t)$, $q(t)$ and $f(t)$ are Lebesgue integrable on $\langle a, b\rangle$. Find a function $y(t)$ satisfying

1) functions $y(t)$ and $p(t)\, y'(t)$ are absolutely continuous functions on $\langle a, b\rangle$;

2) function $y(t)$ fulfils the equation

$$-(p(t)\, y'(t))' + q(t)\, y(t) = f(t) \tag{1}$$

a.e. (almost everywhere) on $\langle a, b\rangle$;

3) function $y(t)$ satisfies the boundary value conditions

$$\alpha_1\, y(a) - \beta_1\, p(a)\, y'(a) = \gamma_1\,, \tag{2}$$

$$\alpha_2\, y(b) + \beta_2\, p(b)\, y'(b) = \gamma_2\,, \tag{3}$$

where $\alpha_1$, $\beta_1$, $\alpha_2$ and $\beta_2$ are numbers such that $\alpha_1^2 + \beta_1^2 \neq 0$ and $\alpha_2^2 + \beta_2^2 \neq 0$.

The combination method for Problem 1 may be described by the following algorithm:

**Step 1.** We find a function $z(t)$ such that $z(t)$ and $p(t)\, z'(t)$ are absolutely continuous functions on the interval $\langle a, b\rangle$ and $z(t)$ satisfies

$$-(p(t)\, z'(t))' + q(t)\, z(t) = 0 \quad \text{a.e. on} \quad \langle a, b\rangle\,, \tag{4}$$

$$z(a) = -\beta_1\,, \tag{5}$$

$$p(a)\, z'(a) = -\alpha_1\,. \tag{6}$$

Function $z(t)$ satisfies the homogeneous equation corresponding to Equation (1) and the homogeneous condition corresponding to Condition (2).

**Step 2.** We find a function $c(t)$ such that $c(t)$ and $p(t)\, c'(t)$ are absolutely continuous functions on the interval $\langle a, b\rangle$ and $c(t)$ satisfies

$$-(p(t)\, c'(t))' + q(t)\, c(t) = f(t) \quad \text{a.e. on} \quad \langle a, b\rangle\,, \tag{7}$$

$$c(a) = \frac{\alpha_1 \gamma_1}{\alpha_1^2 + \beta_1^2}\,, \tag{8}$$

$$p(a)\, c'(a) = -\frac{\beta_1 \gamma_1}{\alpha_1^2 + \beta_1^2}\,. \tag{9}$$

**Step 3.** We find a number $k$ such that

$$\alpha_2(c(b) + k\, z(b)) + \beta_2\, p(b)\, (c'(b) + k\, z'(b)) = \gamma_2 \tag{10}$$

holds. (Equation (10) has a solution if and only if Problem 1 has a solution.) Then the function

$$y(t) = c(t) + k\,z(t) \tag{11}$$

is a solution of Problem 1.

The methods of both composed and simple factorization for a very general boundary value problem for the system of linear differential equations with inner as well as transient conditions are described and discussed in [3]. However, the method simplifies essentially for Problem 1. Therefore we give the composed factorization method for Problem 1 without transforming it to a system of two equations of the first order.

Let us first describe the factorization method for a special case of Problem 1. Namely, let us assume for the present in addition that $p(t) > 0$, $q(t) \geqq 0$ a.e. on $\langle a, b\rangle$ and that $\alpha_i \geqq 0$, $\beta_i \geqq 0$ for $i = 1, 2$. In this case the composed factorization method is described by the following algorithm:

**Step A.** If $\alpha_1 \neq 0$ and $\beta_1/\alpha_1 \leqq 1$ then proceed to Step B. Otherwise, proceed to Step C.

**Step B.** We find absolutely continuous functions $\eta_1(t)$ and $\zeta_1(t)$ on the interval $\langle a, b\rangle$ satisfying the differential equations

$$\eta_1'(t) = q(t)\,\eta_1^2(t) - \frac{1}{p(t)} \qquad \text{a.e. on} \quad \langle a, b\rangle\,, \tag{12}$$

$$\zeta_1'(t) = q(t)\,\eta_1(t)\,\zeta_1(t) - \eta_1(t) f(t) \quad \text{a.e. on} \quad \langle a, b\rangle \tag{13}$$

with the initial conditions

$$\eta_1(a) = -\frac{\beta_1}{\alpha_1}\,,$$

$$\zeta_1(a) = \frac{\gamma_1}{\alpha_1}\,.$$

Then we proceed to Step D.

**Step C.** We find absolutely continuous functions $\eta_2(t)$ and $\zeta_2(t)$ on the interval $\langle a, b\rangle$ satisfying the differential equations

$$\eta_2'(t) = \frac{1}{p(t)}\,\eta_2^2(t) - q(t) \qquad \text{a.e. on} \quad \langle a, b\rangle\,, \tag{14}$$

$$\zeta_2'(t) = \frac{1}{p(t)}\,\eta_2(t)\,\zeta_2(t) - f(t) \quad \text{a.e. on} \quad \langle a, b\rangle \tag{15}$$

with the initial conditions

$$\eta_2(a) = -\frac{\alpha_1}{\beta_1},$$

$$\zeta_2(a) = -\frac{\gamma_1}{\beta_1}.$$

Then we proceed to Step D.

**Step D.** If $\alpha_2 \neq 0$ and $\beta_2/\alpha_2 \leqq 1$ then proceed to Step E. Otherwise, proceed to Step F.

**Step E.** We find absolutely continuous functions $\hat{\eta}_1(t)$ and $\hat{\zeta}_1(t)$ on the interval $\langle a, b\rangle$ satisfying the differential equations

$$(16) \qquad \hat{\eta}_1'(t) = q(t)\,\hat{\eta}_1^2(t) - \frac{1}{p(t)} \qquad \text{a.e. on} \quad \langle a, b\rangle\,,$$

$$(17) \qquad \hat{\zeta}_1'(t) = q(t)\,\hat{\eta}_1(t)\,\hat{\zeta}_1(t) - \hat{\eta}_1(t)\,f(t) \quad \text{a.e. on} \quad \langle a, b\rangle$$

with the initial conditions

$$\hat{\eta}_1(b) = \frac{\beta_2}{\alpha_2},$$

$$\hat{\zeta}_1(b) = \frac{\gamma_2}{\alpha_2}.$$

(This means that the differential equations (16) and (17) are solved from the right to the left.) Then we proceed to Step G.

**Step F.** We find absolutely continuous functions $\hat{\eta}_2(t)$ and $\hat{\zeta}_2(t)$ on the interval $\langle a, b\rangle$ satisfying the differential equations

$$(18) \qquad \hat{\eta}_2'(t) = \frac{1}{p(t)}\,\hat{\eta}_2^2(t) - q(t) \qquad \text{a.e. on} \quad \langle a, b\rangle\,,$$

$$(19) \qquad \hat{\zeta}_2'(t) = \frac{1}{p(t)}\,\hat{\eta}_2(t)\,\hat{\zeta}_2(t) - f(t) \quad \text{a.e. on} \quad \langle a, b\rangle$$

with the initial conditions

$$\hat{\eta}_2(b) = \frac{\alpha_2}{\beta_2},$$

$$\hat{\zeta}_2(b) = \frac{\gamma_2}{\beta_2}.$$

Then we proceed to Step G.

**Step G.** We find the solution of Problem 1 by solving one of the following four systems:

1) $$\begin{pmatrix} 1, & \eta_1(t) \\ 1, & \hat{\eta}_1(t) \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_1(t) \\ \hat{\zeta}_1(t) \end{pmatrix}$$

in case we employed Steps B and E;

2) $$\begin{pmatrix} 1, & \eta_1(t) \\ \hat{\eta}_2(t), & 1 \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_1(t) \\ \hat{\zeta}_2(t) \end{pmatrix}$$

in case we employed Steps B and F;

3) $$\begin{pmatrix} \eta_2(t), & 1 \\ 1, & \hat{\eta}_1(t) \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_2(t) \\ \hat{\zeta}_1(t) \end{pmatrix}$$

in case we employed Steps C and E;

4) $$\begin{pmatrix} \eta_2(t), & 1 \\ \hat{\eta}_2(t), & 1 \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_2(t) \\ \hat{\zeta}_2(t) \end{pmatrix}$$

in case we employed Steps C and F.

(These systems have the unique solution if and only if Problem 1 has the unique solution.) Under our additional assumptions on Problem 1, the Ricatti equations (12) or (14) or (16) or (18) have a solution on the whole interval $\langle a, b\rangle$. If some of the additional assumptions concerning the sign of functions $p(t)$, $q(t)$ and numbers $\alpha_i$ and $\beta_i$ $(i = 1, 2)$ is not fulfilled, then these Ricatti equations need not have a solution on the whole interval. Let us now describe the factorization method for this more general case. For this purpose, it is necessary to introduce a parameter $\mu$ controlling the course of the method. Let $\mu$ be a real number greater than one. The following algorithm is called the composed $\mu$-factorization.

**Step A.** If $\alpha_1 \neq 0$ and $|\beta_1/\alpha_2| \leqq 1$, put $\tau_1 = a$ and $i = 1$ and proceed to Step B. Otherwise, put $\tau_2 = a$ and $i = 2$ and proceed to Step C.

**Step B.** We find absolutely continuous functions $\eta_i(t)$ and $\zeta_i(t)$ on the interval $\langle \tau_i, \tau_{i+1}\rangle$ satisfying the differential equations

(20) $$\eta_i'(t) = q(t)\, \eta_i^2(t) - \frac{1}{p(t)} \qquad \text{a.e. on} \quad \langle \tau_i, \tau_{i+1}\rangle\,,$$

(21) $$\zeta_i'(t) = q(t)\, \eta_i(t)\, \zeta_i(t) - \eta_i(t)\, f(t) \quad \text{a.e. on} \quad \langle \tau_i, \tau_{i+1}\rangle$$

with the initial conditions

$$\eta_i(\tau_i) = \frac{1}{\eta_{i-1}(\tau_i)} \quad \text{for} \quad i > 1\,, \quad \eta_1(a) = -\frac{\beta_1}{\alpha_1} \quad \text{for} \quad i = 1\,,$$

$$\zeta_i(\tau_i) = \frac{\zeta_{i-1}(\tau_i)}{\eta_{i-1}(\tau_i)} \quad \text{for} \quad i > 1\,, \quad \zeta_1(a) = \frac{\gamma_1}{\alpha_1} \quad \text{for} \quad i = 1\,.$$

The number $\tau_{i+1}$ satisfies the following four conditions:

1) $\tau_{i+1} \leqq b$;

2) there exists a solution of Equation (20) on the whole interval $\langle \tau_i, \tau_{i+1} \rangle$;

3) $|\eta_i(t)| < \mu$ for $t \in \langle \tau_i, \tau_{i+1} \rangle$;

4) if $\tau_{i+1} < b$ then $|\eta_i(\tau_{i+1})| > 1$.

(The point $\tau_{i+1}$ is constructed in the course of solution of Equation (20): we solve it either till we reach the point $b$ or till $\eta_i(t)$ in absolute value is not less than $\mu$.) If $\tau_{i+1} = b$, then proceed to Step D. If $\tau_{i+1} < b$, then increase $i$ by one and proceed to Step C.

**Step C.** We find absolutely continuous functions $\eta_i(t)$ and $\zeta_i(t)$ on the interval $\langle \tau_i, \tau_{i+1} \rangle$ satisfying the differential equations

$$\eta_i'(t) = \frac{1}{p(t)}\,\eta_i^2(t) - q(t) \qquad \text{a.e. on} \quad \langle \tau_i, \tau_{i+1} \rangle\,, \tag{22}$$

$$\zeta_i'(t) = \frac{1}{p(t)}\,\eta_i(t)\,\zeta_i(t) - f(t) \quad \text{a.e. on} \quad \langle \tau_i, \tau_{i+1} \rangle \tag{23}$$

with the initial conditions

$$\eta_i(\tau_i) = \frac{1}{\eta_{i-1}(\tau_i)} \quad \text{for} \quad i > 2\,, \qquad \eta_2(a) = -\frac{\alpha_1}{\beta_1} \quad \text{for} \quad i = 2\,,$$

$$\zeta_i(\tau_i) = \frac{\zeta_{i-1}(\tau_i)}{\eta_{i-1}(\tau_i)} \quad \text{for} \quad i > 2\,, \qquad \zeta_2(a) = -\frac{\gamma_1}{\beta_1} \quad \text{for} \quad i = 2\,.$$

The number $\tau_{i+1}$ satisfies the following four conditions:

1) $\tau_{i+1} \leqq b$;

2) there exists a solution of Equation (22) on the whole interval $\langle \tau_i, \tau_{i+1} \rangle$;

3) $|\eta_i(t)| < \mu$ for $t \in \langle \tau_i, \tau_{i+1} \rangle$;

4) if $\tau_{i+1} < b$ then $|\eta_i(\tau_{i+1})| > 1$.

(The point $\tau_{i+1}$ is constructed similarly as in Step B in the course of solution of Equation (22).) If $\tau_{i+1} = b$, then proceed to Step D. If $\tau_{i+1} < b$, then increase $i$ by one and proceed to Step B.

**Step D.** If $\alpha_2 \neq 0$ and $|\beta_2/\alpha_2| \leqq 1$, put $\hat{\tau}_1 = b$ and $i = 1$ and proceed to Step E. Otherwise, put $\hat{\tau}_2 = b$ and $i = 2$ and proceed to Step F.

**Step E.** We find absolutely continuous functions $\hat{\eta}_i(t)$ and $\hat{\zeta}_i(t)$ on the interval $\langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle$ satisfying the differential equations

$$\hat{\eta}_i'(t) = q(t)\,\hat{\eta}_i^2(t) - \frac{1}{p(t)} \qquad \text{a.e. on} \quad \langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle\,, \tag{24}$$

$$\hat{\zeta}_i'(t) = q(t)\,\hat{\eta}_i(t)\,\hat{\zeta}_i(t) - \hat{\eta}_i(t)\,f(t) \quad \text{a.e. on} \quad \langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle \tag{25}$$

with the initial conditions

$$\hat{\eta}_i(\hat{\tau}_i) = \frac{1}{\hat{\eta}_{i-1}(\hat{\tau}_i)} \quad \text{for} \quad i > 1\,, \qquad \hat{\eta}_1(b) = \frac{\beta_2}{\alpha_2} \quad \text{for} \quad i = 1\,,$$

$$\hat{\zeta}_i(\hat{\tau}_i) = \frac{\hat{\zeta}_{i-1}(\hat{\tau}_i)}{\hat{\eta}_{i-1}(\hat{\tau}_i)} \quad \text{for} \quad i > 1\,, \qquad \hat{\zeta}_1(b) = \frac{\gamma_2}{\alpha_2} \quad \text{for} \quad i = 1\,.$$

The number $\hat{\tau}_{i+1}$ satisfies the following four conditions:

1) $\hat{\tau}_{i+1} \geqq a$;

2) there exists a solution of Equation (24) on the whole interval $\langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle$;

3) $|\hat{\eta}_i(t)| < \mu$ for $t \in \langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle$;

4) if $\hat{\tau}_{i+1} > a$, then $|\hat{\eta}_i(\hat{\tau}_{i+1})| > 1$.

(The point $\hat{\tau}_{i+1}$ is constructed in the course of solution of Equation (24): we solve it from the right to the left either till we reach the point $a$ or till the function $\hat{\eta}_i(t)$ in absolute value is not less than $\mu$.) If $\hat{\tau}_{i+1} = a$, then proceed to Step G. If $\hat{\tau}_{i+1} > a$, then increase $i$ by one and proceed to Step F.

**Step F.** We find absolutely continuous functions $\hat{\eta}_i(t)$ and $\hat{\zeta}_i(t)$ on the interval $\langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle$ satisfying the differential equations

$$\hat{\eta}_i'(t) = \frac{1}{p(t)}\,\hat{\eta}_i^2(t) - q(t) \qquad \text{a.e. on} \quad \langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle\,, \tag{26}$$

$$\hat{\zeta}_i'(t) = \frac{1}{p(t)}\,\hat{\eta}_i(t)\,\hat{\zeta}_i(t) - f(t) \quad \text{a.e. on} \quad \langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle \tag{27}$$

with the initial conditions

$$\hat{\eta}_i(\hat{\tau}_i) = \frac{1}{\hat{\eta}_{i-1}(\hat{\tau}_i)} \quad \text{for} \quad i > 2\,, \qquad \hat{\eta}_2(b) = \frac{\alpha_2}{\beta_2} \quad \text{for} \quad i = 2\,,$$

$$\hat{\zeta}_i(\hat{\tau}_i) = \frac{\hat{\zeta}_{i-1}(\hat{\tau}_i)}{\hat{\eta}_{i-1}(\hat{\tau}_i)} \quad \text{for} \quad i > 2\,, \qquad \hat{\zeta}_2(b) = \frac{\gamma_2}{\beta_2} \quad \text{for} \quad i = 2\,.$$

The number $\hat{\tau}_{i+1}$ satisfies the following four conditions:

1) $\hat{\tau}_{i+1} \geqq a$;

2) there exists a solution of Equation (26) on the whole interval $\langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle$;

3) $|\hat{\eta}_i(t)| < \mu$ for $t \in \langle \hat{\tau}_{i+1}, \hat{\tau}_i \rangle$;

4) if $\hat{\tau}_{i+1} > a$ then $|\hat{\eta}_i(\hat{\tau}_{i+1})| > 1$.

(The point $\hat{\tau}_{i+1}$ is constructed similarly as in Step E in the course of solution of Equation (26).) If $\hat{\tau}_{i+1} = a$, then proceed to Step G. If $\hat{\tau}_{i+1} > a$, then increase $i$ by one and proceed to Step F.

**Step G.** Let $i(t)$ be an integer valued function such that $t \in \langle \tau_{i(t)}, \tau_{i(t)+1} \rangle$ and let $j(t)$ be an integer valued function such that $t \in \langle \hat{\tau}_{j(t)+1}, \hat{\tau}_{j(t)} \rangle$. We find the solution of Problem 1 by solving the following systems:

1)
$$\begin{pmatrix} 1, & \eta_{i(t)}(t) \\ 1, & \hat{\eta}_{j(t)}(t) \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_{i(t)}(t) \\ \hat{\zeta}_{j(t)}(t) \end{pmatrix}$$

if both numbers $i(t)$ and $j(t)$ are odd;

2)
$$\begin{pmatrix} 1, & \eta_{i(t)}(t) \\ \hat{\eta}_{j(t)}(t), & 1 \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_{i(t)}(t) \\ \hat{\zeta}_{j(t)}(t) \end{pmatrix}$$

if $i(t)$ is odd and $j(t)$ even;

3)
$$\begin{pmatrix} \eta_{i(t)}(t), & 1 \\ 1, & \hat{\eta}_{j(t)}(t) \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_{i(t)}(t) \\ \hat{\zeta}_{j(t)}(t) \end{pmatrix}$$

if $i(t)$ is even and $j(t)$ odd;

4)
$$\begin{pmatrix} \eta_{i(t)}(t), & 1 \\ \hat{\eta}_{j(t)}(t), & 1 \end{pmatrix} \begin{pmatrix} y(t) \\ p(t)\, y'(t) \end{pmatrix} = \begin{pmatrix} \zeta_{i(t)}(t) \\ \hat{\zeta}_{j(t)}(t) \end{pmatrix}$$

if both $i(t)$ and $j(t)$ are even.

Tab. 1a. $y'' - y = 1$; $h = 0{\cdot}01$

| $x$ | $y$ | $y'$ | Factorization × $10^{12}$ | | Combination × $10^{12}$ | | Increased precision × $10^{12}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −0·46211 | 0 | 167 | 0 | −29 | 0 | −29 |
| 0·1 | −0·04128 | −0·36426 | 31 | 189 | − 4 | −14 | − 4 | −21 |
| 0·2 | −0·07297 | −0·27005 | 52 | 163 | − 1 | −14 | − 1 | −21 |
| 0·3 | −0·09538 | −0·17854 | 53 | 178 | 0 | −14 | 1 | −14 |
| 0·4 | −0·10874 | −0·08883 | 54 | 166 | −12 | − 7 | − 9 | −14 |
| 0·5 | −0·11318 | 0 | 75 | 152 | −10 | 25 | −10 | 10 |
| 0·6 | −0·10874 | 0·08883 | 90 | 93 | − 7 | 3 | − 7 | − 3 |
| 0·7 | −0·09538 | 0·17854 | 85 | 23 | − 3 | 36 | − 3 | 7 |
| 0·8 | −0·07297 | 0·27005 | 83 | − 32 | 7 | 43 | 0 | 14 |
| 0·9 | −0·04128 | 0·36426 | 61 | − 83 | 14 | 36 | 7 | 21 |
| 1·0 | 0 | 0·46211 | 14 | −156 | 21 | 29 | 7 | 43 |

Tab. 1b. $y'' - y = 1$; $h = 0{\cdot}001$

| $x$ | $y$ | $y'$ | Factorization × $10^{12}$ | | Combination × $10^{12}$ | | Increased precision × $10^{12}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −0·46211 | 0 | − 310 | 0 | − 29 | 0 | 43 |
| 0·1 | −0·04128 | −0·36426 | − 39 | − 462 | − 5 | − 43 | 2 | 50 |
| 0·2 | −0·07297 | −0·27005 | − 87 | − 454 | 0 | − 58 | 11 | 36 |
| 0·3 | −0·09538 | −0·17854 | −114 | − 603 | −18 | − 36 | 3 | 50 |
| 0·4 | −0·10874 | −0·08883 | −126 | − 903 | − 7 | − 29 | 14 | 65 |
| 0·5 | −0·11318 | 0 | −217 | −1219 | −29 | − 87 | 14 | 43 |
| 0·6 | −0·10874 | 0·08883 | −301 | − 916 | −25 | − 87 | 21 | 43 |
| 0·7 | −0·09538 | 0·17854 | −331 | − 627 | −32 | −116 | 18 | 43 |
| 0·8 | −0·07297 | 0·27005 | −287 | − 312 | −36 | −203 | 14 | 43 |
| 0·9 | −0·04128 | 0·36426 | −167 | 18 | −43 | −116 | 0 | 58 |
| 1·0 | 0 | 0·46211 | 14 | 305 | −14 | − 72 | 7 | 58 |

## INCREASING PRECISION BY THE METHOD OF COMBINATION OF SOLUTIONS

When solving Problem 1 by the combination method we actually solve two initial value problems (4) and (7) with initial conditions (5), (6) and (8), (9) by the Runge-Kutta-Gill method of the fourth order. When using methods of this type, the interval $\langle a, b\rangle$ is divided into $n$ parts of the length $h$ and the solution $y_i$ of the problem is evaluated successively at the points $a + ih$, $i = 1, 2, \ldots, n$:

$$(28) \qquad y_i = y_{i-1} + h\Phi(h, a, y_{i-1})\,.$$

For small values of $h$, the right hand side of (28) is the sum of two quantities of

Tab. 2a. $y'' - 100y = 1$; $h = 0{\cdot}01$

| $x$ | $y$ | $y'$ | Factorization $\times\ 10^{12}$ | | Combination $\times\ 10^{12}$ | | Increased precision $\times\ 10^{12}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −0·09999 | 0 | − 68 | 0 | − 72 | 0 | − 72 |
| 0·1 | −0·00632 | −0·03677 | −450 | 4338 | −3339 | 33224 | −3339 | 33221 |
| 0·2 | −0·00864 | −0·01349 | −520 | 4829 | −2472 | 24307 | −2472 | 24296 |
| 0·3 | −0·00949 | −0·00488 | −613 | 5272 | −1403 | 13046 | −1403 | 13032 |
| 0·4 | −0·00972 | −0·00158 | −467 | 2772 | − 775 | 5513 | − 775 | 5455 |
| 0·5 | −0·00986 | 0 | −395 | 3 | − 584 | 506 | − 562 | 389 |
| 0·6 | −0·00972 | 0·00158 | −466 | −2763 | − 678 | − 4253 | − 678 | − 4486 |
| 0·7 | −0·00949 | 0·00488 | −612 | −5261 | −1070 | −10472 | −1187 | −10006 |
| 0·8 | −0·00864 | 0·01349 | −519 | −4810 | −1857 | −16606 | −1625 | −20331 |
| 0·9 | −0·00632 | 0·03677 | −446 | −4304 | −1269 | −17773 | −1735 | −17773 |
| 1·0 | 0 | 0·09999 | 0 | 68 | 3395 | 7208 | 3395 | 7208 |

Tab. 2b. $y'' - 100y = 1$; $h = 0{\cdot}001$

| $x$ | $y$ | $y'$ | Factorization $\times\ 10^{12}$ | | Combination $\times\ 10^{12}$ | | Increased precision $\times\ 10^{12}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −0·09999 | 7 | 0 | 0 | 0 | 0 | 0 |
| 0·1 | −0·00632 | −0·03677 | 2 | 0 | 0 | 1 | 0 | 2 |
| 0·2 | −0·00864 | −0·01349 | 2 | − 1 | 0 | − 2 | − 1 | − 2 |
| 0·3 | −0·00949 | −0·00488 | − 1 | − 4 | − 4 | − 21 | − 3 | − 21 |
| 0·4 | −0·00972 | −0·00158 | − 3 | − 8 | − 7 | − 45 | − 7 | −103 |
| 0·5 | −0·00986 | 0 | − 16 | 1 | 5 | − 76 | − 2 | −192 |
| 0·6 | −0·00972 | 0·00158 | − 27 | 9 | 20 | − 529 | − 8 | −296 |
| 0·7 | −0·00949 | 0·00488 | − 60 | 5 | 34 | 701 | −23 | 701 |
| 0·8 | −0·00864 | 0·01349 | − 99 | 156 | 237 | −7297 | 237 | 2015 |
| 0·9 | −0·00632 | 0·03677 | −142 | 1302 | 1522 | −6611 | 1056 | 835 |
| 1·0 | 0 | 0·09999 | − 6 | −931 | −335 | 22055 | 2458 | 7150 |

different magnitude and it may happen that the greater part of the increment $h\Phi$ vanishes in the round-off error. For this reason it is recommended to carry out the evaluation of (28) with increased precision (e.g. in double arithmetics). Since some programming languages (ALGOL) have no possibility to use the double length of words, we used a method which simulates the double arithmetics. We show an example of a computer which guarantees the maximal accuracy of numbers in double length.

A number $x$ in the floating point is represented in the form

$$fl(x) = m\beta^e\,, \quad |m| < \beta^t\,, \quad |e| < k$$

Tab. 3a. $y'' - 100y = 100$; $h = 0{\cdot}01$

| $x$ | $y$ | $y'$ | Factorization $\times 10^{10}$ | | Combination $\times 10^{10}$ | | Increased precision $\times 10^{10}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −9·99909 | 0 | −6402 | 0 | −122 | 0 | −98 |
| 0·1 | −0·63201 | −3·67739 | −44987 | 434000 | −3344 | 33153 | −3342 | 33190 |
| 0·2 | −0·86433 | −1·34993 | −52081 | 483010 | −2491 | 24121 | −2477 | 24195 |
| 0·3 | −0·94930 | −0·48873 | −61365 | 527240 | −1443 | 12458 | −1425 | 12756 |
| 0·4 | −0·97920 | −0·15836 | −46719 | 277160 | −842 | 4704 | −842 | 4704 |
| 0·5 | −0·98652 | 0 | −39516 | 191 | −694 | −60 | −768 | −1252 |
| 0·6 | −0·97920 | 0·15836 | −46668 | −276750 | −1117 | −7276 | −521 | −2507 |
| 0·7 | −0·94930 | 0·48873 | −61271 | −526310 | −2056 | −22303 | −1460 | −12766 |
| 0·8 | −0·86433 | 1·34993 | −51899 | −481230 | −6038 | −25494 | −1269 | −6420 |
| 0·9 | −0·63201 | 3·67739 | −44645 | −430620 | −15214 | −132824 | −909 | −56530 |
| 1·0 | 0 | 9·99909 | 0 | 6402 | −900 | −238486 | −900 | 219277 |

Tab. 3b. $y'' - 100y = 100$; $h = 0{\cdot}001$

| $x$ | $y$ | $y'$ | Factorization $\times 10^{12}$ | | Combination $\times 10^{10}$ | | Increased precision $\times 10^{10}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −9·99909 | 0 | 116 | 0 | 37 | 0 | 73 |
| 0·1 | −0·63201 | −3·67739 | −65 | 902 | 4 | 52 | 5 | 107 |
| 0·2 | −0·86433 | −1·34993 | −36 | 480 | 25 | 278 | 23 | 278 |
| 0·3 | −0·94930 | −0·48873 | 50 | −1120 | 83 | 387 | 74 | 983 |
| 0·4 | −0·97920 | −0·15836 | 58 | −1189 | 200 | 829 | 200 | 2916 |
| 0·5 | −0·98652 | 0 | −138 | −1833 | 348 | 1130 | 572 | 5899 |
| 0·6 | −0·97920 | 0·15836 | −312 | −2968 | 372 | 7028 | 1266 | 18949 |
| 0·7 | −0·94930 | 0·48873 | −633 | −5824 | −1460 | −17536 | 1520 | 20610 |
| 0·8 | −0·86433 | 1·34993 | −1084 | −10827 | −961 | −6425 | 2306 | 31721 |
| 0·9 | −0·63201 | 3·67739 | −1491 | −14901 | −911 | −94695 | 8625 | 96039 |
| 1·0 | 0 | 9·99909 | 0 | −116 | −906 | −391132 | −906 | 371807 |

$m$ denoting the mantissa and $e$ the exponent of the number; $\beta$ is the base of the number system, $m$, $e$, $k$ are integers, $t$ the number of digits of the mantissa. If $x$ is non zero, we require in addition that $|x| \geqq \beta^{t-1}$ (normalization). Neither overflow nor underflow is considered. The result of the machine operation of addition must be known with at least one digit more and be normalized before the round-off. When evaluating the sum $x + y$ we make the error

$$\eta = x + y - z$$

where $z = fl(x + y)$ is the result of the machine operation of addition. It is shown in [6] that

$$\eta = fl(y - fl(z - x)) \quad \text{for} \quad |x| \geqq |y| \, .$$

Tab. 4a. $y'' - 1000y = 1$; $h =$ 0·01

| $x$ | $y$ | $y'$ | Factorization $\times 10^{12}$ | | Combination $\times 10^8$ | | Increased precision $\times 10^8$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −0·03162 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0·1 | −0·00096 | −0·00133 | −5325 | 168400 | −1 | 45 | −1 | 45 |
| 0·2 | −0·00100 | −0·00005 | − 837 | 26469 | −0·1 | 4 | −0·1 | 4 |
| 0·3 | −0·00100 | 0 | − 61 | 1944 | 0 | 0·6 | 0 | 0·4 |
| 0·4 | −0·00100 | 0 | − 4 | 116 | 0·1 | 8 | 0·3 | 8 |
| 0·5 | −0·00100 | 0 | 0 | 0 | 7 | 285 | 7 | 285 |
| 0·6 | −0·00100 | 0 | − 4 | − 116 | 231 | 9155 | 231 | 9155 |
| 0·7 | −0·00100 | 0 | − 61 | − 1944 | 9863 | ? | 6811 | ? |
| 0·8 | −0·00100 | 0·00005 | − 837 | − 26469 | ? | ? | ? | ? |
| 0·9 | −0·00096 | 0·00133 | −5325 | −168400 | ? | ? | ? | ? |
| 1·0 | 0 | 0·03162 | 0 | 0 | ? | ? | ? | ? |

Tab. 4b. $y'' - 1000y = 1$; $h =$ 0·001

| $x$ | $y$ | $y'$ | Factorization $\times 10^{14}$ | | Combination $\times 10^8$ | | Increased precision $\times 10^8$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −0·03162 | 0 | − 682 | 0 | 0 | 0 | 0 |
| 0·1 | −0·00096 | −0·00133 | − 77 | 1222 | 0 | 0·006 | 0 | 0·005 |
| 0·2 | −0·00100 | −0·00005 | − 27 | − 336 | 0·001 | 0·06 | 0·001 | 0·06 |
| 0·3 | −0·00100 | 0 | 0 | 26 | 0·01 | 0·8 | 0·03 | 0·4 |
| 0·4 | −0·00100 | 0 | − 1 | − 23 | 1 | 37 | 1 | 31 |
| 0·5 | −0·00100 | 0 | 1 | 0 | 25 | 953 | 25 | 667 |
| 0·6 | −0·00100 | 0 | − 1 | 23 | 994 | 21362 | 898 | 24414 |
| 0·7 | −0·00100 | 0 | 0 | − 26 | 22070 | ? | 22070 | ? |
| 0·8 | −0·00100 | 0·00005 | − 31 | 219 | ? | ? | ? | ? |
| 0·9 | −0·00096 | 0·00133 | −129 | −2879 | ? | ? | ? | ? |
| 1·0 | 0 | 0·03162 | 0 | 682 | ? | ? | ? | ? |

During the evaluation of (28) $y_{i-1}$ is stored in two cells $y$ and $yy$ while $y_i$ is stored again in two cells, $z$ and $zz$. The computation proceeds according to the formulas

$$z = fl(y + h)\,,$$

$$zz = yy + fl(h\Phi - fl(z - y))$$

(provided $|y| \geqq |h\Phi|$).

After completing the evaluation of initial value problem, the sum $z + zz$ yields a more accurate value of the solution or, if need be, its derivative, for the evaluation of the constant $k$ in Equation (10) as well as for the evaluation of $y(t)$.

Tab. 5a. $y'' - 1000y = 1000$; $h = 0{\cdot}01$

| $x$ | $y$ | $y'$ | Factorization × $10^{12}$ | | Combination × $10^8$ | | Increased precision × $10^8$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −31·62277 | 0 | 931 | 0 | −1 | 0 | −1 |
| 0·1 | −0·95767 | −1·33856 | −5325200 | 168400000 | −1453 | 45959 | −1453 | 45959 |
| 0·2 | −0·99820 | −0·05666 | −837050 | 26470000 | −126 | 4114 | −126 | 4114 |
| 0·3 | −0·99992 | −0·00239 | −61496 | 1944400 | −44 | −7905 | −44 | −7905 |
| 0·4 | −1 | −0·00010 | −3718 | 116800 | −5781 | ? | −5781 | ? |
| 0·5 | −1 | 0 | −422 | 0 | −97629 | ? | −97629 | ? |
| 0·6 | −1 | 0·00010 | −3718 | −116800 | −3124700 | ? | −6249700 | ? |
| 0·7 | −0·99992 | 0·00239 | −61496 | −1944400 | ? | ? | ? | ? |
| 0·8 | −0·99820 | 0·05666 | −837030 | −26470000 | ? | ? | ? | ? |
| 0·9 | −0·95767 | 1·33856 | −5325100 | −168390000 | ? | ? | ? | ? |
| 1·0 | 0 | 31·62277 | 0 | −931 | ? | ? | ? | ? |

Tab. 5b. $y'' - 1000y = 1000$; $h = 0{\cdot}001$

| $x$ | $y$ | $y'$ | Factorization × $10^{12}$ | | Combination × $10^8$ | | Increased precision × $10^8$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −31·62277 | 0 | −232 | 0 | −12 | 0 | −13 |
| 0·1 | −0·95767 | −1·33856 | −676 | 15425 | −5 | −162 | −5 | −162 |
| 0·2 | −0·99820 | −0·05666 | −138 | −1209 | −114 | −3896 | −126 | −3133 |
| 0·3 | −0·99992 | −0·00239 | −21 | 43 | −1761 | −75044 | −2333 | −68940 |
| 0·4 | −1 | −0·00010 | −29 | 262 | −24092 | ? | −36299 | ? |
| 0·5 | −1 | 0 | −7 | 0 | 488310 | ? | 27 | ? |
| 0·6 | −1 | 0·00010 | −36 | −378 | 6250300 | ? | 3125300 | ? |
| 0·7 | −0·99992 | 0·00239 | −21 | −45 | ? | ? | ? | ? |
| 0·8 | −0·99820 | 0·05666 | −196 | −671 | ? | ? | ? | ? |
| 0·9 | −0·95767 | 1·33856 | −1207 | −32131 | ? | ? | ? | ? |
| 1·0 | 0 | 31·62277 | 0 | 232 | 0 | ? | ? | ? |

## NUMERICAL EXPERIMENTS

We solved the boundary value problem

$$y'' - ay = b\,, \quad y(0) = y(1) = 0 \tag{29}$$

by the above methods on the computer ICL 1905 for several values $a$, $b$. The computation was carried out with the fixed integration steps 0·01 and 0·001.

Tables 1 to 5 provide the errors of computation of $y$ and $y'$ satisfying Problem (29) by the methods of factorization, of combination of solutions and of combination of solutions with increased precision, for $x = 0, 0{\cdot}1, 0{\cdot}2, \ldots, 1$. To give a possibility

Tab. 6. $y'' + 100y = 1$; $h = 0·001$

| $x$ | $y$ | $y'$ | Factorization × $10^{12}$ | | Combination × $10^{12}$ | | Increased precision × $10^{12}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | 0·33805 | 1 | −910010 | 1 | −596 | 1 | −567 |
| 0·1 | 0·03304 | 0·26680 | − 76585 | −491900 | −45 | −341 | −44 | −334 |
| 0·2 | 0·04490 | −0·04975 | −107950 | 312970 | −56 | 215 | −53 | 193 |
| 0·3 | 0·02467 | −0·32056 | − 28533 | 834210 | −17 | 563 | −16 | 520 |
| 0·4 | −0·00905 | −0·29664 | 54882 | 690140 | 36 | 422 | 35 | 403 |
| 0·5 | −0·02525 | 0 | 115510 | − 4859 | 55 | − 38 | 54 | − 31 |
| 0·6 | −0·00905 | 0·29664 | 54270 | −693840 | 29 | −476 | 29 | −447 |
| 0·7 | 0·02467 | 0·32056 | − 28933 | −832050 | −21 | −527 | −18 | −509 |
| 0·8 | 0·04490 | 0·04975 | −107960 | −310250 | −55 | − 93 | −53 | − 99 |
| 0·9 | 0·03304 | −0·26680 | − 76468 | 492630 | −46 | 371 | −44 | 341 |
| 1·0 | 0 | −0·33805 | − 1 | 909800 | − 2 | 600 | − 1 | 596 |

Tab. 7. $y'' + 1000y = 1$; $h = 0·001$

| $x$ | $y$ | $y'$ | Factorization × $10^{12}$ | | Combination × $10^{12}$ | | Increased precision × $10^{12}$ | |
|---|---|---|---|---|---|---|---|---|
| 0·0 | 0 | −0·00328 | 0 | −26177 | 0 | −3133 | 0 | −3118 |
| 0·1 | 0·00200 | 0·00263 | 20 | 21624 | 4 | 2302 | 4 | 2287 |
| 0·2 | 0 | −0·00197 | −64 | −14966 | − 8 | −1463 | − 8 | −1448 |
| 0·3 | 0·00200 | 0·00131 | 78 | 11477 | 11 | 624 | 11 | 609 |
| 0·4 | −0·00001 | −0·00066 | −71 | − 4571 | −13 | 223 | −13 | 236 |
| 0·5 | 0·00200 | 0 | 81 | 1 | 13 | −1065 | 13 | −1081 |
| 0·6 | −0·00001 | 0·00066 | −71 | 4571 | −13 | 1910 | −13 | 1923 |
| 0·7 | 0·00200 | −0·00131 | 78 | −11477 | 11 | −2757 | 11 | −2773 |
| 0·8 | 0 | 0·00197 | −64 | 14966 | − 8 | 3602 | − 8 | 3615 |
| 0·9 | 0·00200 | −0·00263 | 20 | −21624 | 5 | −4456 | 4 | −4461 |
| 1·0 | 0 | 0·00328 | 0 | 26177 | 0 | 5299 | 0 | 5310 |

to compare the relative errors, exact values of $y$ and $y'$ are introduced in the second and third columns. The next columns show the errors multiplied by the scale given in the headings of the tables. Interrogation marks denote worthless results when the error of solution is greater than the solution itself (in absolute value). The increase of precision in the combination method reveals itself to a small extent in the tables, since the errors of the method are considerably greater than the round-off errors. While in Tab. 1 the combination method yields better results than the factorization method, in the next tables the accuracy of the combination method decreases rapidly and in Tab. 4 and 5 it yields absurd results for $x > 0·5$, in the factorization method this tendency is negligible.

Experiments for $a < 0$ were carried out as well. In these cases, not physically motivated, the combination method yields a little better results, quite in accordance with the theory [5]. These results are presented in Tab. 6 and 7.

The presented results demonstrate the advantages of the factorization method. The good behaviour of the factorization method has been verified on many other problems, both experimental and practical. The method was used also to solve big boundary value problems for a system of twelve to twenty differential equations. The experience acquired is favourable from the point of view of both numerical stability and machine time.

## *References*

[1] *Ivo Babuška, Milan Práger, Emil Vitásek:* Numerical Processes in Differential Equations. John Wiley & Sons, London—New York—Sydney; SNTL Praha (1966).

[2] *Ivo Babuška:* Numerical Stability in Mathematical Analysis. IFIP Congress 68, Invited papers, 1—13 (1968).

[3] *Jiří Taufer:* On Factorization Method. Aplikace matem. 11, 6, 427—451 (1966).

[4] *Jiří Taufer:* Faktorisierungsmethode für ein Randwertproblem eines linearen Systems von Differential Gleichungen. Aplikace matem. 13, 2, 191—198 (1968).

[5] *Jiří Taufer:* Lösung der Randwertprobleme für Systeme von linearen Differentialgleichungen. In print. (1968).

[6] *T. J. Dekker:* A floating-point technique for extending the available precision. Mathematisch Centrum MR 118/70 (1970).

## Souhrn

## SROVNÁNÍ FAKTORIZACE S METODOU STŘELBY

Ivo Hrubec a Jiří Taufer

Existuje řada metod řešení okrajového problému pro soustavu obyčejných diferenciálních lineárních rovnic. Jedna skupina těchto metod spočívá v tom, že se příslušná úloha nahrazuje posloupností úloh s počátečními podmínkami. Tyto metody z hlediska jejich numerické realizace nejsou si ekvivalentní ani co do pracnosti, ani co do numerické stability. Teoretické srovnávání těchto metod bylo provedeno v práci [5]. Cílem této práce je ukázat na jednoduchých experimentálních příkladech nedobré vlastnosti velmi jednoduché a často používané metody střelby ve srovnání s metodou faktorizace. Získané experimentální výsledky jsou ve shodě s teoretickou analýzou zkoumaných numerických procesů, která byla provedena v jiných pracích.

V práci je formulována metoda střelby a metody složené faktorizace pro diferenciální rovnici druhého řádu. V této práci není prováděna teoretická analýza zkouma-

ných numerických procesů, ukazuje se zde pouze na názorných příkladech chování vyšetřovaných metod.

Teoretická analýza velké třídy metod, která zahrnuje jak metodu střelby tak i faktorizaci, byla provedena v práci [5]. Uveřejnění pouze experimentálních výsledků považujeme za užitečné, neboť tyto výsledky bezprostředně vysvětlují častý nezdar metody střelby a mohou tedy být varováním i vodítkem při výběru metod.

*Authors' addresses:* Ing. *Ivo Hrubec,* Geodetický ústav, Arbesovo nám. 4, Praha 5, ***Jiří** Taufer,* CSc., Matematický ústav ČSAV v Praze, Žitná 25, Praha 1.