

ON A NUMERICAL SOLUTION OF THE BOUNDARY
VALUE PROBLEM USING AN OPTIMAL NUMERICAL DIFFERENTIATION

Ljiljana Cvetković and Dragoslav Herceg

Prirodno-matematički fakultet. Institut za matematiku
21000 Novi Sad, ul.dr Ilije Djuričića br.4, Jugoslavija

ABSTRACT

In this paper we consider a numerical solution of two boundary value problems: (BVP1) $-u''(x) = f(x,u)$, $u(0) = \gamma_0$, $u(1) = \gamma_1$, and (BVP2) $-u''(x) - q(x)u(x) = f(x)$, $u(0) = \gamma_0$, $u(1) = \gamma_1$, using the authors' four-point rule of degree 3 for the second derivative. The discretisation meshes depend upon this rule. The discrete problem to (BVP1) has the usual form, but the discrete problem to (BVP2) has an unusual form. Under certain assumptions, our schemes have a third order of convergence.

1. INTRODUCTION

We consider two boundary value problems (BVP1) and (BVP2).

$$(BVP1) \quad -u''(x) = f(x,u) \quad \text{on} \quad [0,1], \quad u(0) = \gamma_0, \quad u(1) = \gamma_1,$$

where $f \in C^3([0,1] \times \mathbb{R})$ and $\gamma_0, \gamma_1 \in \mathbb{R}$. The nonlinearity $f(t,v)$ is assumed to satisfy the Lipschitz condition

$$(L) \quad q(v-w) \leq f(x,v) - f(x,w) \leq \mu(v-w), \quad v, w \in \mathbb{R}, \quad v \geq w, \quad x \in [0,1]$$

for some reals $q, \mu \in \mathbb{R}$, where

$$(1) \quad -h^{-2}q_0 \leq \mu < \lambda_0, \quad q_0 = 3(\sqrt{5} - 1), \quad 0 < \lambda_0 < 8,$$

and we let $h = (1 + n \frac{3+\sqrt{5}}{2})^{-1}$ be so small, i.e. $n \in \mathbb{N}$ so great, that

$$(2) \quad -h^{-2}q_0 < \frac{1}{2}(\lambda_0 + q)$$

(BVP2) $-u'' - q(x)u = f(x)$ on $[0, 1]$, $u(0) = \beta_0$, $u(1) = \beta_1$,
 where $q, f \in C^3[0, 1]$, $\beta_0, \beta_1 \in \mathbb{R}$. We assume

$$(3) \quad 0 \leq q(x) \leq \mu < 8 \quad \text{on } [0, 1].$$

The assumptions for (BVP1) are similar to the assumptions from [1], but here q_0 and λ_0 are different. The parameter h depends on the rule we used for the discretisation of (BVP1). For $q(x)$ from (BVP2) a usual assumption is $q(x) \leq 0$ in $[0, 1]$. But, our scheme for (BVP2) has a special form and we have the assumption $0 \leq q(x) < 8$.

The proofs in section 3 are based upon the results of this type in [1], [2], [7], [8].

2. FINITE DIFFERENCE SCHEMES

In this section we shall describe the schemes which we are going to discuss. It is convenient to distinguish (BVP1) and (BVP2). First we shall consider (BVP1).

For $u \in C^5[0, 1]$, the authors' four-point rule of degree 3 for the second derivative, see [4], is

$$(4) \quad -u''(x) = h^{-2} \left[af(x+h) + bf(x) + cf\left(x \mp \frac{1+\sqrt{5}}{2}h\right) + \right. \\ \left. + df\left(x \mp \frac{3+\sqrt{5}}{2}h\right) \right] + O(h^3),$$

where

$$a = -\frac{2\sqrt{5}}{5}, \quad b = 6 - 2\sqrt{5}, \quad c = -(3 - \sqrt{5}), \quad d = \frac{7\sqrt{5}}{5} - 3.$$

We let for $n \in \mathbb{N}$,

$$I_h = \{x_{2i} = i \frac{3+\sqrt{5}}{2} h, x_{2i+1} = h + x_{2i}, i=0, 1, \dots, n\},$$

where $h = (1 + n \frac{3+\sqrt{5}}{2})^{-1}$. Now we can form the discrete analogue to (BVP1) by using (4):

$$(DBVP 1) \quad (L_h y)_i = \begin{cases} \gamma_0, & i=0, \\ 0, & i=1, 2, \dots, 2n, \\ \gamma_1, & i=2n+1, \end{cases}$$

where

$$(9) \quad f_k = [\beta_0, f_2, f_1, f_4, f_3, \dots, f_{2n-3}, f_{2n}, f_{2n-1}, \beta_1]^T.$$

3. PROPERTIES OF THE SCHEMES

In this section we shall consider the discrete analogues (DBVP1) and (DBVP2) to (BVP1) and (BVP2). We shall prove that (DBVP1) and (DBVP2) have unique solutions, say y and z , for which we have

$$\|u^h - y\|_\infty \leq M_1 h^3, \quad \|u^k - z\|_\infty \leq M_2 k^3,$$

where u^h and u^k denote the restriction of the exact solution of (BVP1) and (BVP2) to the mesh I_h and J_k , and M_1, M_2 are the constants independent of h and k .

We shall begin with some notations (see [1], [2], [7], [8]). For $x, y \in \mathbb{R}^m$, we write

$$x \leq (<) y \quad \text{iff} \quad x_i \leq (<) y_i, \quad i=1, 2, \dots, m,$$

$$|x| = [|x_1|, |x_2|, \dots, |x_m|]^T.$$

Any $e \in \mathbb{R}^m$, $e > 0$, defines the norm $\|x\|_e = \max_{1 \leq i \leq m} \frac{|x_i|}{e_i}$ on \mathbb{R}^m . In particular $e = [1, 1, \dots, 1]^T$ yields the maximum norm $\|\cdot\|_\infty$.

A mapping F of \mathbb{R}^m into itself is called monotone if $x \leq y \Rightarrow Fx \leq Fy$ for any $x, y \in \mathbb{R}^m$.

For mappings F, G of \mathbb{R}^m into itself we write $F \leq G$ iff $G - F$ is monotone.

The set of matrices of the format $m \times m$ is denoted by $\mathbb{R}^{m,m}$.

For any $A = [a_{ij}] \in \mathbb{R}^{m,m}$ the matrices $A_d, A_a, A_a^- \in \mathbb{R}^{m,m}$ are defined via $A_d = \text{diag}(a_{11}, a_{22}, \dots, a_{mm})$, $A_a = A - A_d$, $A_a^- = [a_{ij}^-]$,

$$a_{ij}^- = \begin{cases} a_{ij} & \text{if } a_{ij} < 0, \\ 0 & \text{if } a_{ij} \geq 0. \end{cases}$$

Let $\tau^0(x) = \{i: i=1, 2, \dots, m, x_i = 0\}$, $\tau^+(x) = \{i: i=1, 2, \dots, m, x_i > 0\}$ for $x \in \mathbb{R}^m$.

If τ^1 and τ^2 are disjoint subsets of $\{1, 2, \dots, m\}$ we

say that $A = [a_{ij}] \in \mathbb{R}^{m,m}$ connects τ^1 with τ^2 if for all $i \in \tau^1$ there are point $i_0 = i, i_1, i_2, \dots, i_r \in \{1, 2, \dots, m\}$ such that $a_{i_{j-1} i_j} \neq 0$, $j=1, 2, \dots, r$ and $i_r \in \tau^2$.

The matrix A is called an L-matrix if $A_a \leq 0$.

$A \in \mathbb{R}^{m,m}$ is called an inverse-monotone matrix if A has an inverse $A^{-1} \geq 0$. Throughout this paper we shall use the abbreviation i.m. for inverse-monotone. The inverse-monotone L-matrix is called a M-matrix.

The following 5 theorems and their proofs can be found in [2], [3], [6], [7].

THEOREM 1. Let $A \in \mathbb{R}^{m,m}$, $\delta \in \mathbb{R}^m$. If $A_a \leq 0$, $\delta \geq 0$, $A\delta \geq 0$ and if A connects $\tau^0(A\delta)$ with $\tau^+(A\delta)$ then A is an M-matrix.

THEOREM 2. Let $A = [a_{ij}] \in \mathbb{R}^{m,m}$, $A_a^- = A^z + A^s$, $A^z = [a_{ij}^{(z)}] \leq 0$, $A^s = [a_{ij}^{(s)}] \leq 0$. The matrix A is i.m. if the following conditions are satisfied:

1. $A_d + A^z$ is an M-matrix,
2. $a_{ij} \leq \sum_{k=1}^m a_{ik}^{-1} a_{kj}^{(s)}$ for all $a_{ij} > 0$, $i \neq j$,
3. there exists $e \geq 0$ ($e \in \mathbb{R}^m$) such that $Ae \geq 0$ and A^z or A^s connects $\tau^0(Ae)$ with $\tau^+(Ae)$.

THEOREM 3. Let $A \leq B$ ($A, B \in \mathbb{R}^{m,m}$) and assume that B is i.m. Then A is i.m. iff there exists $e > 0$ ($e \in \mathbb{R}^m$) such that $B^{-1}Ae > 0$.

THEOREM 4. Let $C \leq A \leq B$ ($A, B, C \in \mathbb{R}^{m,m}$). If B and C are i.m. then A is i.m.

THEOREM 5. Let F be a nonlinear mapping \mathbb{R}^m into itself and let $P, Q \in \mathbb{R}^{m,m}$ be such that $Q \leq P$, $Q \leq F \leq P$. Let $A, S \in \mathbb{R}^{m,m}$ be such that $A-P$, $A-S$ are i.m., $2S \leq P+Q$. Then $(A-F)^{-1}$ exists and $|(A-F)^{-1} - (A-F)^{-1}w| \leq (A-P)^{-1}|v-w|$, for any $v, w \in \mathbb{R}^m$. Furthermore, the parallel chord iteration

$$x^0 \in \mathbb{R}^m, (A-S)x^n = (F-S)x^{n-1}, n \in \mathbb{N}$$

converges for any $x^0 \in \mathbb{R}^m$ to the unique solution of $Ax = Fx$.

Now we shall consider (DBVP1).

$$\lambda_i = \frac{2}{x_1(1-x_1)+\xi} \geq \frac{2}{0.25+\xi} = \lambda_0, \quad i=1,2,\dots,2n.$$

Let $D_1 = \text{diag}(0, d_1, d_2, \dots, d_{2n}, 0) \geq 0$ satisfy condition (10). Then $A_h - D_1 \leq A_h$, $(A_h - D_1)e > 0$. Since $A_h^{-1} \geq 0$ it follows $A_h^{-1}(A_h - D_1)e > 0$ and Theorem 3 yields that $A_h - D_1$ is i.m. Let $D_2 = \text{diag}(0, d'_1, d'_2, \dots, d'_{2n}, 0) \leq 0$ satisfy condition (10). Let $(A_h - D_2)_a = A^Z + A^S$, where $A^Z = [z_{ij}]$, $A^S = [s_{ij}]$ is defined via

$$z_{ij} = h^{-2} \begin{cases} -(3-\sqrt{5}) & \text{if } i=2m, j=2m-1, m=1,2,\dots,n, \\ & i=2m-1, j=2m, m=1,2,\dots,n, \\ 0 & \text{otherwise,} \end{cases}$$

$$s_{ij} = h^{-2} \begin{cases} -\frac{2\sqrt{5}}{5} & \text{if } i=2m, j=2m+1, m=1,2,\dots,n \\ & i=2m+1, j=2m, m=0,1,\dots,n-1, \\ 0 & \text{otherwise.} \end{cases}$$

Then applying Theorem 1 to $\delta = [1, 1, \dots, 1]^T$ we see that $(A_h - D_2)d + A^Z$ is i.m. Since

$$d(b - h^2 d'_1) \leq a \iff d'_1 \geq -h^{-2} q_0, \text{ and } (A_h - D_2)e \geq A_h e > 0,$$

Theorem 2 yields that $A_h - D_2$ is i.m.. Now let the matrix $D = \text{diag}(0, \mu_1, \mu_2, \dots, \mu_{2n}, 0)$ satisfy condition (10). Then there exist matrices $D_1 \geq 0$ and $D_2 \geq 0$ such that $D_2 \leq D \leq D_1$. Since $A_h - D_1 \leq A_h - D \leq A_h - D_2$, Theorem 4 yields that $A_h - D$ is i.m.

THEOREM 7. Suppose that conditions (L), (1), (2) are satisfied. Then for any $v, w \in \mathbb{R}^{2n+2}$

$$|v-w| \leq (A_h - M)^{-1} |L_h v - L_h w|,$$

where $M = \text{diag}(0, \mu, \mu, \dots, \mu, 0) \in \mathbb{R}^{2n+2, 2n+2}$

THEOREM 8. Let $S = \text{diag}(0, s_1, s_2, \dots, s_{2n}, 0)$, $s_i \geq -h^{-2} q_0$, $i=1, 2, \dots, 2n$; and $s = \max_{1 \leq i \leq 2n} s_i \leq \frac{1}{2}(\mu + q)$. Then the parallel chord iteration

$$y^0 \in \mathbb{R}^{2n+2}, \quad (A_h - S)y^m = (F_h - S)y^{m-1}, \quad m \in \mathbb{N}$$

converges for any y^0 to the unique solution of $A_h y = F_h y$.

The proofs of Theorem 7 and Theorem 8 follow directly from Theorem 5.

THEOREM 9. *Suppose that conditions (L), (1), (2) are satisfied. Then*

$$\|v-w\|_e \leq \frac{1}{K} \|L_h v - L_h w\|_e, \quad K = \min(\xi, \lambda_0^{-1}), \quad \text{for any } v, w \in \mathbb{R}^{2n+2},$$

$$\|v-w\|_\infty \leq \frac{2}{\lambda_0 \xi K} \|L_h v - L_h w\|_\infty, \quad \text{for any } v, w \in \mathbb{R}^{2n+2}.$$

P r o o f. Since $(A_h - M)e \geq Ke$ and $(A_h - M)^{-1} \geq 0$ it follows that

$$(A_h - M)^{-1} e \leq \frac{1}{K} e, \quad \|(A_h - M)^{-1}\|_e = \|(A_h - M)^{-1} e\|_e \leq \frac{1}{K}.$$

From Theorem 7 it follows that

$$\|v-w\|_e \leq \|(A_h - M)^{-1}\|_e \|L_h v - L_h w\|_e \leq \frac{1}{K} \|L_h v - L_h w\|_e \quad \text{for any } v, w \in \mathbb{R}^{2n+2}.$$

Since

$$\frac{1}{0.25 + \xi} \|z\|_\infty \leq \|z\|_e \leq \frac{1}{\xi} \|z\|_\infty \quad \text{for any } z \in \mathbb{R}^{2n+2},$$

we have

$$\|v-w\|_\infty \leq \frac{0.25 + \xi}{\xi K} \|L_h v - L_h w\|_\infty, \quad \text{for any } v, w \in \mathbb{R}^{2n+2},$$

which completes the proof.

COROLLARY 1. *Suppose that conditions (L), (1), (2) are satisfied. Let y be the solution of (DBVP1), u the solution of (BVP1) and the vector u^h be the restriction of u to the mesh I_h . Then*

$$u \in C^5[0, 1] \Rightarrow \|y - u^h\|_\infty \leq M_1 h^3$$

where M_1 is a constant independent of h .

P r o o f. In [5] it is proved that $u \in C^5[0, 1]$ implies that

$$L_h u^h - L_h y = O(h^3).$$

Assuming $v = u^h$ and $w = y$, from Theorem 9, there, follows the proof.

We shall consider now the boundary value problem (BVP2) and its discrete analogue $B_k z = f_k$.

THEOREM 10. *Let condition (3) be satisfied. Then the matrix B_k is i.m.*

P r o o f. Let $(B_k)_{\bar{a}} = B^Z + B^S$, where $B^Z = [b_{ij}^{(z)}]$,

$$b_{ij}^{(z)} = \begin{cases} \frac{C}{2} & \text{if } i=2m, j=2m+2, m=1, 2, \dots, n-2, \\ & i=2m+1, j=2m-1, m=2, 3, \dots, n-1, \\ 0 & \text{otherwise.} \end{cases}$$

Since $(B_k)_d + B^Z$ is an M-matrix, $2DB_2 \leq A_2 C$, $4BD \leq C^2$ and

$$B_k e = [\xi, 2-q_2 e_2, 2-q_1 e_1, 2-q_4 e_4, 2-q_3 e_3, \dots$$

$$\dots, 2-q_{2n} e_{2n}, 2-q_{2n-1} e_{2n-1}, \xi]^T > 0, \text{ for}$$

$$e = [e_0, e_1, \dots, e_{2n+1}]^T > 0, e_i = x_i(1-x_i) + \xi, x_i \in J_k,$$

where $\xi > 0$ is determined so that $\mu < \frac{2}{0.25+\xi} < 8$, we conclude applying Theorem 2, that B_k is i.m.

THEOREM 11. *Suppose that condition (3) is satisfied. Let z be the solution of (DBVP2), u the solution of (BVP2) and the vector u^k the restriction of u to the mesh J_k . Then*

$$u \in C^5[0, 1] \Rightarrow \|u^k - z\|_{\infty} \leq M_2 k^3,$$

where M_2 is a constant independent of k .

P r o o f. Let $\varepsilon \in (0, \min(-A_1, \frac{A_1 A_2}{2B_1 - A_2}))$. Now multiply the zeroth row of the matrix B_k by εk^{-2} and add the result to the first and third row. Also, multiply the $(2n+1)$ th row of matrix B_k by εk^{-2} and add the result to the $(2n)$ th and $(2n-2)$ th row. Now multiply the first, third, $(2n)$ th and $(2n-2)$ th row of matrix B_k by k^2 . As a result one obtains a new matrix \tilde{B}_k . The equations

$$B_k z = f_k \text{ and } \tilde{B}_k z = \tilde{f}_k, \text{ with}$$

$$\tilde{f}_k = [\beta_0, k^2 f_2 + \varepsilon \beta_0, f_1, k^2 f_4 + \varepsilon \beta_0, f_3, f_5, \dots, f_{2n-2}, k^2 f_{2n-3} + \varepsilon \beta_1, f_{2n}, k^2 f_{2n-1} + \varepsilon \beta_1, \beta_1]^T$$

are equivalent. Using Theorem 2, with vector e from Theorem 10, we obtain that \tilde{B}_k is i.m.

Let $\lambda = \min(1, \frac{\epsilon\xi}{0.25+\xi}, \frac{2}{0.25+\xi} - \mu) > 0$. Then $\tilde{B}_k e \geq \lambda e$ and $\|\tilde{B}_k^{-1}\|_e \leq \frac{1}{\lambda}$.

Since $\tau(u) = \tilde{B}_k u^k - \tilde{B}_k z = \tilde{B}_k u^k - \tilde{f}_k = O(k^3)$ and $u^k - z = \tilde{B}_k^{-1} \tau(u)$, we have

$$\begin{aligned} \|u^k - z\|_\infty &\leq (0.25+\xi) \|u^k - z\|_e \leq (0.25+\xi) \|\tilde{B}_k^{-1}\|_e \|\tau(u)\|_e \leq \\ &\leq (0.25+\xi) \frac{1}{\lambda} \cdot \frac{1}{\xi} \|\tau(u)\|_\infty \leq M_2 k^3. \end{aligned}$$

4. NUMERICAL RESULT

In this section we shall present a numerical example :

$$-u'' = -e^u, \quad x \in [0, 1], \quad u(0) = u(1) = 0.$$

This problem is considered in [3] and [6], and its exact solution is

$$u(x) = -\ln 2 + 2 \ln \left(c \sec \frac{c(x-0.5)}{2} \right), \quad c = 1.3360557 \dots$$

For $x \in [0, 1]$ is $-1 \leq u(x) \leq 0$.

To compute the approximation of $u(x)$ we define (see [2])

$$f(x, v) = \begin{cases} -e^{-1}, & \text{for } v \leq -1, \\ -e^v, & \text{for } -1 \leq v \leq 0, \\ -1, & \text{for } v \geq 0, \end{cases}$$

satisfying (L) with $q = e^{-1} - 1$, $\mu = 0$. Using our first scheme, with $n = 20$, we iterate according to

$$y^0 = [1, 1, \dots, 1]^T \in \mathbb{R}^{42}, \quad A_h y^m = F_h y^{m-1}, \quad m \in \mathbb{N}.$$

From Theorem 7 (see [2]) it follows that

$$|y^m - y| \leq A_h^{-1} |A_h y^m - F_h y^m|, \quad m \in \mathbb{N},$$

where y is the solution of (DBVP1) for our example. Also it holds that

$$|y^m - y| \leq w^m, \quad m \in \mathbb{N},$$

where w^m is a solution of $A_h w^m \geq |A_h y^m - F_h y^m|$, $m \in \mathbb{N}$.

We have calculated in double precision arithmetic on PDP 11/340.

The numerical results are

$$h = 0.18740391 \cdot 10^{-1}, \quad m = 16, \quad \|y^m - y\|_{\infty} < 10^{-16},$$

$$\|y^m - u^h\|_{\infty} \leq 10^{-7},$$

where u^h denotes the restriction of u to the mesh I_h .

REFERENCES

- [1] Bohl, E., Lorenz, J., *Inverse monotonicity and difference schemes of higher order. A summary for two-point boundary value problems.* *Aeq. Math.* 19(1979), 1-36.
- [2] Bohl, E., *Finite Modelle gewöhnlicher Randwertaufgaben.* Teubner, Stuttgart, 1981.
- [3] Ciarlet, P.G., Schultz, M.H., Varga, R.S., *Numerical methods of higher order accuracy for nonlinear boundary value problems, I. One dimensional problem,* *Numer. Math.* 9(1967), 394-430.
- [4] Herceg, D., Aleksić, Lj., *An optimal numerical approximation of second derivative. III Conference on Applied Mathematics, D. Herceg, ed., Institute of Mathematics, Novi Sad, 1982, 41-47.*
- [5] Herceg, D., Cvetković, Lj., *On a numerical differentiation (to appear).*
- [6] Jerome, J.W., Varga, R.S., *Generalizations of spline functions and applications to nonlinear boundary value and eigenvalue problems. Theory and Applications of Spline Functions, T.N.E. Greville, ed., Academic Press, New York-London, 1969, 103-155.*
- [7] Lorenz, J., *Die Inversmonotonie von Matrizen und ihre Anwendung beim Stabilitätsnachweis von Differenzenverfahren. Dissertation, Universität Münster, 1975.*
- [8] Lorenz, J., *Zur Inversmonotonie diskreter Probleme.* *Numer. Math.* 27 (1977), 227-238.

Received by the editors April 7, 1983.

REZIME

O NUMERIČKOM REŠAVANJU KONTURNOG PROBLEMA
POMOĆU OPTIMALNOG NUMERIČKOG DIFERENCIRANJA

U radu se posmatra numeričko rešavanje konturnih problema (BVP1) i (BVP2) pod pretpostavkama (L), (1), (2) odnosno (3). Za formiranje diskretnih analogona (DBVP1) i (DBVP2) koriste se optimalne četvorotačkaste formule reda 3 za numeričku aproksimaciju drugog izvoda iz [4] i [5]. Pri tom se koriste specijalne mreže diskretizacije I_h i J_k . Dokazano je da pod navedenim pretpostavkama opisani diferencni postupci imaju red konvergencije 3 i da postupak paralelne sečice za rešavanje (DBVP1) konvergira. Numerički je rešen primer iz [3] i [6].