An efficient approach for solving stiff nonlinear boundary value problems

V. L. Makarov*  D. V. Dragunov†

November 15, 2017

Abstract

A new method for solving stiff boundary value problems is described and compared to other known approaches using the Troesch’s problem as a test example. The method is based on the general idea of alternate approximation of either the unknown function or its inverse and has a genuine “immunity” towards numerical difficulties invoked by the rapid variation (stiffness) of the unknown solution. A c++ implementation of the proposed method is available at https://github.com/imathsoft/MathSoftDevelopment.

AMS 2010 Subject Classification: 65L04, 65L05, 65L10, 65L20, 65L50, 65Y15

Key Words: stiff boundary value problem, simple shooting method, multiple shooting method, the Troesch’s problem

1 Introduction

In the present paper we consider a nonlinear boundary value problem

\[
\frac{d^2 u(x)}{dx^2} = N(u(x), x) u(x), \quad x \in [a, b], \quad N(u, x) \in C^2(\mathbb{R} \times [a, b]),
\]

(1)

\[
u(a) = u_l \in \mathbb{R}, \quad u(b) = u_r \in \mathbb{R},
\]

(2)

which arises in many areas of physics and mathematics. Although, there is a huge variety of known methods for solving problems of type (1), (2) (see, for example [12], [10], [2] and the references therein), almost none of them fill comfort when the problem turns out to be stiff. As it was pointed out in [4], a good mathematical definition of the concept of stiffness does not exist. The famous definition given in [11] says that ”stiff equations are problems for which explicit methods don’t work”, which, unfortunately, is not very constructive. According to [3], there is at least 6 different definitions of stiff problems which possess different levels of formality and are accepted by different schools of mathematics.

*Institute of Mathematics of NAS of Ukraine, e-mail: makarov@imath.kiev.ua
†Institute of Mathematics of NAS of Ukraine, e-mail: dragunovdenis@gmail.com
The authors of [3] came up with their own definition of "stiffness", based on the concept of stiffness ratio, which, according to them, encompasses all the known definitions.

In the present paper we confine ourselves to consider only a subclass of stiff boundary value problems (1), (2) whose stiffness is originated from the fact that the exact solution $u(x)$ possesses narrow intervals of rapid variation, known as the boundary layers. Such a behavior is typical for singularly perturbed problems, which are an important subclass of stiff problems (see, [7], [3]). The rapid variation is equivalent to having $|u'(x)| \gg 1$ on some subset of $[a, b]$. And it is the need to approximate the solution on this subset that makes the problem numerically difficult and unstable, i.e. stiff. Now to approximate the solution on the subset of $[a, b]$ where $|u'(x)|$ is comparatively small is much easier from the numerical point of view. To be more specific, let us consider a set $\chi_u \in [a, b]$ defined in the following way:

$$\chi_u = \{ x \in [a, b] : |u'(x)| \geq 1 \}. \quad (3)$$

It is easy to see that, defined in such a way, set $\chi_u$ consists of a finite or infinite number of distinctive closed intervals $\bar{\iota}_i$. Some of the intervals $\bar{\iota}_i$ might be those of rapid variation for the solution $u(x)$. At the same time, by the definition of $\chi_u$ (3), solution $u(x)$ is strictly monotonic on each interval $\bar{\iota}_i$, which means that we can consider the inverse function $x_{\bar{\iota}_i}(\cdot) = u^{-1}(\cdot)$ defined on the closed interval $u(\bar{\iota}_i) \in u([a, b])$. There are two remarkable things about the function $x_{\bar{\iota}_i}(\cdot)$:

1. $|x'_{\bar{\iota}_i}(u)| \leq 1, \forall u \in u(\bar{\iota}_i)$, which means that the initial BVP stated in terms of ”inverse solution” $x'_{\bar{\iota}_i}(u)$ is not stiff on $u(\bar{\iota}_i)$;

2. having function $x_{\bar{\iota}_i}(u)$ approximated on a discrete set of points from $u(\bar{\iota}_i)$ we automatically get function $u(x)$ approximated on some discrete set of points from $\bar{\iota}_i$.

The two observations give us the key insight on how to solve the subclass of stiff problems defined above. It is the divide and conquer principle: on the subintervals where solution $u(x)$ is well behaved (showing rather moderate variation) we solve the given problem (1), (2), whereas on the subintervals $\bar{\iota}_i$, where $u(x)$ varies rapidly (and the initial problem is stiff), we solve the corresponding problem for the inverse solution $x'_{\bar{\iota}_i}(u)$. Of course, this becomes feasible from the practical point of view only if there is a finite number of subintervals $\bar{\iota}_i$, which becomes our assumption from now on.

The main purpose of the present paper is not only to give a theoretical idea about how to treat some subclass of stiff boundary value problems in an efficient way, but also to describe and examine one of the possible practical implementations of the proposed theoretical approach for the class of boundary value problems (1), (2). That is why throughout the paper we will stay in touch with one of the most famous examples of stiff BVPs, known as the
Troesch’s problem:

\[
\frac{d^2 u(x)}{dx^2} = \lambda \sinh (\lambda u(x)), \quad x \in [0, 1]
\]

\[ u(0) = 0, \quad u(1) = 1, \]  

which is a partial case of problem (1), (2) with \( N(u(x), x) \equiv \lambda \sinh (\lambda u(x)) / u(x), \ a = u_l = 0, \ b = u_r = 1. \) It is well known, that the problem (4), (5) is inherently unstable and difficult (see [1], [3], [6], [13], [18], [19] and the references therein).

The Troesch’s problem, in addition to its application in physics of plasma, has drawn a lot of interest to itself as a test case for methods of solving unstable two-point boundary value problems because of its difficulties [1]. It is worth mentioning, that the approach described in the present paper was initially developed for the particular purpose of solving the Troesch’s problem. This can, in part, explain why despite the fact that eventually the approach was generalized to be applicable to the class of problems (1), (2) (and can potentially be generalized even further), the main theoretical result of the paper, Theorem 5, deals with a more narrow set of problems, which, however, contains the Troesch’s one.

Talking about the known methods for solving BVPs, it is impossible not to mention the simple shooting method (SSM) and the multiple shooting method (MSM) [16, Section 7.3] which are two the most simple and reliable techniques to deal with boundary value problems of type (1), (2). By calling them techniques and not just methods we would like to emphasize a fact that the basic idea behind them is very broad and can be used in many different modifications, which, in turn, might be called the methods. Since definitions of both SSM and MSM essentially relay on using methods for solving initial value problems (IVP), one of the ways to come up with a new modification of the methods consists in using a different IVP solver. Below we are going to adapt (or modify if you wish) the SSM and MSM for using a specific approach for numerical solution of IVP’s called Straight-Inverse method (or, simply SI-method) which is based on the general idea of alternate approximation of either straight \( u(x) \) or inverse \( x(u) \) solutions of the problem (1), (2) and has a genuine "immunity" towards numerical difficulties invoked by the rapid variation (stiffness) of the solution in question.

The paper is organized as follows. Section 2 introduces a pair of, so called, step functions, which are the simplest logical building blocks of the proposed implementation of the SI-method. The section pays special attention to the computational aspect of the step functions. In particular, it describes quite original and easy-to-implement approach for computing partial derivatives of the functions. The approach is based on the theory of matrix functions and is general enough to be mentioned on its own. In Section 3 we describe the SI-method for solving initial value problems associated with equation (1) and thoroughly investigate
approximation properties of the method, stating and proving the main theoretical result of the paper, Theorem 5. The SI-method for solving boundary value problems \((1) \quad (2)\) is given in Section 4 followed by the section of numerical examples, where we apply the proposed implementation of the SI-method to the Troesch’s problem and discuss the results by comparing them to the corresponding results from other papers. Section 6 contains conclusions.

2 Step functions

In this section we introduce a pair of, so called, step functions, which play a crucial role throughout the rest of the paper.

Let us define the step function \(U(h)\) to be the solution for the following initial value problem:

\[
\frac{d^2 U(s)}{ds^2} = (As + B)U(s), \quad U(0) = D; \quad U'(0) = C, \quad s \in \mathbb{R}.
\]  

Throughout the present section we assume that \(A, B, C, D\) are elements of some Banach space \(\mathfrak{X}\) over the field of real numbers \(\mathbb{R}\) unless otherwise stated.

It is easy to check that if \(A, B, C, D \in \mathbb{R}\), the function \(U(s)\) can be expressed explicitly through the Airy functions (see \([14, 283]\)):

\[
U(s) = C_1 \text{Ai} \left( \frac{As + B}{(-A)^{2/3}} \right) + C_2 \text{Bi} \left( \frac{As + B}{(-A)^{2/3}} \right),
\]  

where

\[
C_1 = \left( -\text{Bi} \left( \frac{B}{(-A)^{2/3}} \right) C (-A)^{2/3} + D A \text{Bi} \left( 1, \frac{B}{(-A)^{2/3}} \right) \right) E,
\]

\[
C_2 = \left( -C (-A)^{2/3} \text{Ai} \left( \frac{B}{(-A)^{2/3}} \right) + D A \text{Ai} \left( 1, \frac{B}{(-A)^{2/3}} \right) \right) E,
\]

\[
E = A^{-1} \left( \text{Bi} \left( \frac{B}{(-A)^{2/3}} \right) \text{Ai} \left( 1, \frac{B}{(-A)^{2/3}} \right) - \text{Bi} \left( 1, \frac{B}{(-A)^{2/3}} \right) \text{Ai} \left( \frac{B}{(-A)^{2/3}} \right) \right)^{-1}.
\]

Similarly to this, we define step function \(V(s)\) to be the solution for the problem

\[
\frac{d^2 V(s)}{ds^2} = (As + B)\frac{dV(s)}{ds}, \quad V(0) = D; \quad V'(0) = C, \quad s \in \mathbb{R},
\]  

\[
(8)
\]
possessing explicit representation in the form of

\[ V(s) = D + C \int_0^s \exp \left( \frac{A}{2} \xi^2 + B \xi \right) d\xi, \]  

(9)

provided that \( A, B, C, D \in \mathbb{R} \).

Despite being explicit, formulas (7) and (9) are quite difficult to evaluate, especially when \( \|A\| \) is quite small. However, in what follows, we are going to use functions \( U(s) \) and \( V(s) \) with \( |s| \) quite small, which allows us to use more convenient approach for evaluating them instead of formulas (7), (9). The approach naturally follows from the theorems below.

**Theorem 1.** Let \( A, B, C, D \in \mathfrak{B}, s \in \mathbb{R} \) then

\[ U(s) = U(A, B, C, D, s) = \lim_{n \to +\infty} U_n(A, B, C, D, s), \]  

(10)

where

\[ U_n(A, B, C, D, s) = \int_0^s \left( \int_0^\xi (A\eta + B)U_{n-1}(A, B, C, D, \eta)d\eta \right) d\xi + Cs + D, \]  

(11)

\[ U_0(A, B, C, D, s) = 0 \]

and the following estimation holds true:

\[ \|U_{n+1}(A, B, C, D, s) - U_n(A, B, C, D, s)\| \leq \frac{(\|A\|s + \|B\|)^n(\|C\|s + \|D\|)|s|^{2n}}{2n!}. \]  

(12)

**Theorem 2.** Let \( A, B, C, D \in \mathfrak{B}, s \in \mathbb{R} \) then

\[ V(s) = V(A, B, C, D, s) = \lim_{n \to +\infty} V_n(A, B, C, D, s), \]  

(13)

where the successive approximations \( V_n(A, B, C, D, s) \) can be found recursively

\[ V'_n(A, B, C, D, s) = \int_0^s (A\eta + B)V'_{n-1}(A, B, C, D, \eta)d\eta + C, \]  

\[ V_n(A, B, C, D, s) = \int_0^s V'_n(A, B, C, D, \xi)d\xi + D, \]  

(14)

\[ V'_0(A, B, C, D, s) = 0, \]

and the following estimation holds true:

\[ \|V_{n+1}(A, B, C, D, s) - V_n(A, B, C, D, s)\| \leq \frac{(\|A\|s + \|B\|)^n\|C\|s^{n+1}}{n!}. \]  

(15)

Formulas (10), (11) and (13), (14) provide straightforward algorithms for computing func-
tions $U(s)$ and $V(s)$ respectively. One might argue that, since the target problem (1) (2) is one dimensional, we should be interested in the case $\mathfrak{B} \equiv \mathbb{R}$ only. The argument is sound, however, having the formulas work for $\mathfrak{B} = M(n)$ allows us to approximate partial derivatives of $U(s)$ and $V(s)$ with respect to parameters $A, B, C, D$ in a quite original, efficient and easy-to-implement way. The key idea of what is meant comes from the theory of matrix functions and, in particular, from the following formula (see, for example [8, p. 98]):

$$f(J_2(a)) = f \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix} = \begin{bmatrix} f(a) & f'(a) \\ 0 & f(a) \end{bmatrix},$$

which holds true for every function $f(x)$ defined in $x = a$ together with its first derivative. Matrix $J_2(a)$ is also known as a Jordan block associated with the eigenvalue $a$. From (16) it follows that to calculate $U'_a(a_1, a_2, a_3, a_4, s)$, for $i = 1, 2, \ldots, 4$, we can use the linear matrix equation

$$U(a_1 E_2 + \delta_{i1} J_2(0), a_2 E_2 + \delta_{i2} J_2(0), \ldots, a_4 E_2 + \delta_{i4} J_2(0), s) =$$

$$= U(a_1, \ldots, a_4, s) E_2 + U'_{a_i}(a_1, \ldots, a_4, s) J_2(0),$$

Of course, the same remains true as related to the function $V(a_1, a_2, a_3, a_4, s)$. Here we are not going to discuss the efficiency of the proposed approach for finding partial derivatives of the step functions (and this might be a good topic for further investigations). At the same time, we would like to mention that on practice the approach proved to be efficient enough and is used in the implementation available at [https://github.com/imathsoft/MathSoftDevelopment](https://github.com/imathsoft/MathSoftDevelopment).

In principle, equation (17) can be used for calculating partial derivatives of $U(A, B, C, D, s)$ and $V(A, B, C, D, s)$ with respect to $s$. However, it is better not to. The most natural way to calculate $V'_s(s)$ follows directly from the Theorem 2 and to calculate $U'_s(s)$ it is easier to use the formula

$$U'_s(s) = \int_0^s (A\eta + B) U(\eta) d\eta + C.$$  

The importance of the step functions introduced above can be explained by the approximation properties which they possess. Some of the properties are described in the theorems below.

**Theorem 3.** Let $u(x)$ be the solution to equation (1) supplemented by the following initial conditions:

$$u(a) = u_l \in \mathbb{R}, \quad u'(a) = u'_l \in \mathbb{R}.$$  

(1) Linear space of $n \times n$ matrices over the field of real numbers
Then, for the sufficiently small \( h > 0 \), the inequalities

\[
\|u(a + x) - U(A, B, C, D, x)\|_{x \in [-h,h]} \leq h^4 K \|u(a + x)\|_{x \in [-h,h]} \exp(hM),
\]

\[
\|u'(a + x) - U'_x(A, B, C, D, x)\|_{x \in [-h,h]} \leq h^3 K \|u(a + x)\|_{x \in [-h,h]} \exp(hM)
\]

hold true, where

\[
A = N'_u(u_t, a) u'_t + N'_x(u_t, a),
B = N(u_t, a),
C = u'_t,
D = u_t,
\]

\[
K = \frac{1}{2} \left\| (N(u(x), x))'' \right\|_{x \in [a-h,a+h]} = \frac{1}{2} \left\| N''_u(u(x), x) (u'(x))^2 + 2 N''_x(u(x), x) u'(x) + N''_x(u(x), x) + N'_u(u(x), x) N(u(x), x) \right\|_{x \in [a-h,a+h]},
\]

\[
M = \max\{1, |B| + |A|h\}
\]

and

\[
\|f(x)\|_{x \in [a,b]} \overset{\text{def}}{=} \max_{x \in [a,b]} |f(x)|, \ f(x) \in C([a,b]).
\]

Proof. From the assumption (II) about smoothness of the function \( N(u,x) \) and the Piccard-Lindelöf theorem (see, for example [17, p. 38]) it follows that the solution \( u(x) \) of the IVP (II), (19) exists at least in some closed neighborhood \( B_\delta(a) = [a-\delta, a+\delta], \delta > 0 \) of the point \( x = a \). From now on we assume that \( h = \delta \) and the solution \( u(x) \in C^2(B_\delta(a)) \) is known.

Combining equations (II) and (6) we come to a linear system of the first order ordinary differential equations with respect to unknown vector-function \( Z(x) = [z(x), z'(x)]^T, z(x) = u(a + x) - U(A, B, C, D, x) \):

\[
\dot{Z}(x) = \begin{bmatrix} 0 & 1 \\ B + Ax & 0 \end{bmatrix} Z(x) + \begin{bmatrix} 0 \\ F(x)u(a + x) \end{bmatrix},
\]

supplemented with zero initial condition

\[
Z(0) = 0,
\]

where

\[
F(x) = N(u(a + x), a + x) - B - Ax.
\]
From (24), (25) it follows that
\[
\|Z(x)\| \overset{\text{def}}{=} \max \{ |z(x)|, |z'(x)| \} \leq M \int_0^x \|Z(\xi)\| d\xi, \quad x \in [0, h]
\]
\[
+ h^3 K \|u(x)\|_{x \in [a-h, a+h]}, \quad x \in [-h, h].
\]

The Gronwall’s inequality (see, for example, [17, 42]), being applied to (26), leads us to the inequality
\[
\|Z(x)\|_{x \in [-h, h]} \leq h^3 K \|u(x)\|_{x \in [a-h, a+h]} \exp(hM)
\]
which immediately implies estimation (21). Estimation (20) follows from (21) and the fact that
\[
\|z(x)\|_{x \in [-h, h]} = \left\| \int_0^x z'(\xi) d\xi \right\|_{x \in [-h, h]} \leq h \|z'(x)\|_{x \in [-h, h]}.
\]

\[\square\]

**Lemma 1.** Let \(u'_l \neq 0\). Then in some neighborhood \(B(u_l, h) = (u_l - h; u_l + h)\) of the point \(u_l\) there exists a unique solution \(x(u)\) to the IVP
\[
\frac{d^2 x(u)}{du^2} = -N(u, x(u)) u \left( \frac{dx(u)}{du} \right)^3,
\]
\[
x(u_l) = a, \quad x'(u_l) = x'_l = 1/u'_l.
\]
and \(x(u)\) is the inverse function to \(u(x)\) satisfying (11), (19), i.e. \(\forall u_1 \in B(u_l, h)\) we have that \(x(u_1) \in \bar{B}(a, h_1)\) for some \(h_1 > 0\) and \(u_1 \equiv u(x(u_1))\).

**Theorem 4.** Let \(x(u)\) be a function satisfying IVP (21), (28) in some neighborhood \(B(u_l, h), h > 0\). Then
\[
\|x(u_l + u) - V(A, B, C, D, u)\|_{u \in [-h, h]} = h^4 K \|x(u_l + u)\|_{u \in [-h, h]} \exp(hM),
\]
\[
\|x'(u_l + u) - V'_u(A, B, C, D, u)\|_{u \in [-h, h]} = h^3 K \|x(u_l + u)\|_{u \in [-h, h]} \exp(hM),
\]
where
\[ A = - \left( (N'_u(u, t, a) + N'_x(u, t, a)x'_t)u_t + N(u, t, a) \right) (x'_t)^2 + 2 (N(u, t, a)u_t)^2 (x'_t)^4, \]
\[ B = - N(u, t, a)u_t (x'_t)^2, \]
\[ C = x'_t, \]
\[ D = a, \]
\[ K = \frac{1}{2} \left\| \left( N(u, x(u)) u (x'(u))^2 \right)^{''} \right\|_{u \in [u_t - h, u_t + h]}, \]
\[ M = |B| + |A|h. \]

**Proof.** The proof is similar to that of Theorem 3. So we skip it. \(\Box\)

## 3 Straight-Inverse method for solving IVPs for the second order differential equations

### 3.1 Preliminary comments

In this section we describe the SI-method for solving equation (1) subjected to the initial conditions (19). However, before doing that, let us first introduce a list of requirements which we expect the SI-method to fulfill, and which, historically, led us to the SI-method as such:

1. the method should approximate the exact solution \(u(x)\) for IVP (1), (19) on a discrete mesh \(\emptyset \neq \omega \in [a, b]\), which should depend on the problem itself and on the desired accuracy of approximation;

2. for a given positive \(h \in \mathbb{R}\), the method should provide an algorithm to construct a mesh \(\omega(h) \in [a, b]\), containing \(N_{\omega} = \mathcal{O} \left( h^{-1} \int_a^b \sqrt{1 + (u'(x))^2} dx \right)\) points, and the SI-method’s approximation \(u_{\omega}(x)\) of the solution \(u(x)\) on the mesh \(\omega(h)\) should satisfy the following asymptotic equalities:

\[
\|u(x) - u_{\omega}(x)\|_{\omega} = \mathcal{O}(h^2), \tag{32}
\]
\[
\|u'(x) - u'_{\omega}(x)\|_{\omega} = \mathcal{O}(h^2) \tag{33};
\]

3. from the method’s point of view, there should be no essential difference between solving IVP (1), (19) with respect to \(u(x)\) or with respect to its inverse \(x(u)\), i.e. the method being applied to the IVP (27), (28), which is the ”inverse” equivalent of IVP (1), (19) (in the sense that the graph of \(x(u)\) (where it exists) coincide with that of \(u(x)\)), gives the same result (or almost the same result) as when it is applied to the ”straight” IVP(2).

---

(2) Apparently, the statement, as it is now, has a lack of rigor and it might even seem contradictory. However, its meaning will
3.2 Definition

Let \( \omega_{IVP}(h) \) denotes an ordered set of quadruples of the form

\[
[u'_i, x'_i, u_i, x_i], \quad u'_i = \frac{1}{x'_i}, \quad i = 0, 1, 2, \ldots
\]

which are defined by means of the following chain of recurrence equalities:

\[
x_0 = a, \quad u_0 = u_l, \quad u'_0 = u'_l;
\]

if \(|u'_{i-1}| \leq 1\) then

\[
x_i = x_{i-1} + h, \\
u_i = U(A_{i-1}, B_{i-1}, C_{i-1}, D_{i-1}, h), \\
u'_i = U'(A_{i-1}, B_{i-1}, C_{i-1}, D_{i-1}, h), \\
A_{i-1} = N'(u_{i-1}, x_{i-1})u'_{i-1} + N'(u_{i-1}, x_{i-1}), \\
B_{i-1} = N(x_{i-1}, x_{i-1}), \\
C_{i-1} = u'_1, \\
D_{i-1} = u_{i-1},
\]

otherwise, if \(|u'_{i-1}| > 1\)

\[
x_i = V(A_{i-1}, B_{i-1}, C_{i-1}, D_{i-1}, h^*_i), \\
u_i = u_{i-1} + h^*_i, \\
x'_i = V'(A_{i-1}, B_{i-1}, C_{i-1}, D_{i-1}, h^*_i), \\
A_{i-1} = - \left( (N'_u(u_{i-1}, x_{i-1}) + N'_x(u_{i-1}, x_{i-1})x'_{i-1})u_{i-1} + \\
+ N(u_{i-1}, x_{i-1}) \right) (x'_{i-1})^2 + \\
+ 2 (N(u_{i-1}, x_{i-1})u_{i-1})^2 (x'_{i-1})^4, \\
B_{i-1} = - N(u_{i-1}, x_{i-1})u_{i-1} (x'_{i-1})^2, \\
C_{i-1} = x'_1, \\
D_{i-1} = x_{i-1}, \\
h^*_i = \text{sign}(x'_{i-1})h,
\]

where \( h \) — some fixed positive real number hereinafter referenced to as a step size of the SI-method.

The ordered set \( \omega_{IVP}(h) \) will be referenced to as a mesh of the SI-method that corresponds become more clear as we get more familiar with the SI-method and its properties.
to IVP (1), (19). From the recurrence formulas (34), (35), (36) it follows that if function \( N(u, x) \) belongs to \( C^1(\mathbb{R} \times [0, +\infty)) \) then the mesh \( \omega_{IVP}(h) \) contains infinite number of elements, i.e. the recurrence process of calculating quadruples \([u'_i, x'_i, u_i, x_i]\) can be continued infinitely long. In the light of this, a reasonable question arises: whether the mesh \( \omega_{IVP}(h) \) (which is infinite) have something to do with the exact solution \( u(x) \) of the Cauchy problem (1), (19) (which might exist only on some finite subinterval of \([a, +\infty)\)) and, if yes, what approximation properties does the mesh possess with respect to the exact solution? To some extent the question is addressed in the paragraph below.

### 3.3 Error analysis.

In the present section we investigate approximation properties of the SI-method introduced above. The main result can be stated as the following theorem.

**Theorem 5.** Let the nonlinear function \( N(u, x) \) be independent on \( x \), i.e.

\[
\frac{\partial N(u, x)}{\partial x} \equiv 0, \tag{37}
\]

and

\[
N(u) \equiv N(u, x) \in C^3(\mathbb{R}), \tag{38}
\]

\[
N(u) > 0, \ \forall u \in \mathbb{R}, \tag{39}
\]

\[
\lim_{u \to +\infty} \frac{1}{u^{2+\lambda}} \int_{u_i}^{u} N(\xi)\xi d\xi > 0, \tag{40}
\]

for some \( \lambda > 0 \).

If

\[
0 < h < \min \left\{ 1, \sqrt{\varepsilon \frac{\varepsilon}{P^*}}, \frac{\varepsilon}{L_0^* M^*} \right\}, \tag{41}
\]

then there exists an integer \( i^* \geq 0 \), such that

\[
u'_i < 1, \ \forall i \in (0, \ldots, i^* - 1), \ u'_{i^*} \geq 1 \tag{42}
\]

and the estimation holds true

\[
\max \{|u(x_i) - u_i|, |u'(x_i) - u'_i|\} \leq h^2 P^*, \ \forall i \in (0, i^*), \tag{43}
\]

where

\[
P^* = \frac{M^*}{2} (L_0^* + L_1^* M^*) \exp ((S^* + 1) \max\{1, L_0^* + L_1^*\} + S^* M^*(2L_1^* + L_2^*)) \frac{L_1^* M^* + \max\{1, L_0^*\}}{L_1^* M^* + \max\{1, L_0^*\}}, \tag{44}
\]
\[ L_i^* = \max_{|u| < M^* + \varepsilon} |N^{(i)}(u)|, \ i = 0, 1, 2, \quad (45) \]

\[ M^* = \frac{1}{2} S^*(1 + 3\varepsilon - u'_i), \quad (46) \]

\[ S^* = \lim_{u \to +\infty} \int_{u_i}^{u} \frac{d\eta}{\sqrt{(u'_i)^2 + 2 \int_{u_i}^{\eta} N(\xi) d\xi}}, \quad (47) \]

and \( \varepsilon \) denotes an arbitrary parameter from \((0, 1/6)\).

If, additionally,

\[ N^{(i)}(u) \geq 0, \ \forall u \in \mathbb{R}, \ i = 1, 2, 3, \quad (48) \]

\[ \lim_{u \to +\infty} \frac{N(u)}{\int_{0}^{u} N(\xi) d\xi} < +\infty, \quad (49) \]

\[ h < \min \left\{ \frac{1 - 2\varepsilon}{P^*}, \frac{1}{3\mu} \right\}, \quad (50) \]

where

\[ \mu = \sup_{u \in [u_i^*, +\infty)} \frac{N(u)}{1 + \int_{u_i^*}^{u} N(\xi) d\xi}, \quad N(u) \overset{\text{def}}{=} N(u)u, \quad (51) \]

then \( u'_i \geq 1, \ \forall i > i^* \) and the estimations hold true

\[ |x_i - x(x_i)| \leq \frac{P^* h^2}{1 - P^* h^2 - L_0^* M^* h} + \]

\[ + \mathcal{E}_i \left( \frac{L_0^* M^*}{(1 - 2\varepsilon)^2} + 1 \right) \frac{P^*}{1 - 2\varepsilon} (u_i - u_i^*) + \frac{1}{2} \int_{u_i^*}^{u_i} \int_{u_i^*}^{\eta} T(\zeta) d\zeta d\eta \right) h^2, \quad (52) \]

\[ |x'_i - x'(x_i)| \leq \mathcal{E}_i \left( \frac{L_0^* M^*}{(1 - 2\varepsilon)^2} + 1 \right) \frac{P^*}{1 - 2\varepsilon} + \frac{1}{2} \int_{u_i^*}^{u_i} T(\zeta) d\zeta \right) h^2, \quad (53) \]

where \( x(\cdot) \overset{\text{def}}{=} u^{-1}(\cdot), \)

\[ \mathcal{E}_i = \exp \left( \int_{u_i^*}^{u_i} \frac{N(\zeta) + hN'(\zeta)}{1 - 6\varepsilon + \int_{u_i^*}^{\zeta - 2h} N(\xi) d\xi} \right) \times \left( \int_{1 - 6\varepsilon + \int_{u_i^* - 2h}^{\zeta - 2h} N(\xi) d\xi}^{\zeta - 2h} \frac{2h N(\zeta)^2}{2h N(\zeta)^2} \right), \quad (54) \]

\(^{(3)}\) The existence of the limit follows from condition \((50)\).
\[
\times \left( (1 - 2\epsilon)^2 + 2 \int_{u_i}^{\zeta} N(\xi) d\xi \right)^{-\frac{1}{2}}, \quad i = i^* + 1, i^* + 2, \ldots,
\]

\[
\mathcal{T}(\zeta) = \left\{ \begin{array}{c}
\frac{N''(\zeta)}{1 - 6\epsilon + 2 \int_{u_i}^{\zeta - h} N(\xi) d\xi} + \frac{6N''(\zeta)N(\zeta)}{\left(1 - 6\epsilon + 2 \int_{u_i}^{\zeta - h} N(\xi) d\xi\right)^2} + \\
\frac{8(N(\zeta))^3}{\left(1 - 6\epsilon + 2 \int_{u_i}^{\zeta - h} N(\xi) d\xi\right)^3} \left( (1 - 2\epsilon)^2 + 2 \int_{u_i}^{\zeta} N(\xi) d\xi \right)^{-\frac{1}{2}}.
\end{array} \right.
\]

**Proof.** As it was pointed out above, the function \(x(u)\), which is (by definition) inverse of the exact solution \(u(x)\) should be solution to the IVP (27), (28). Under the assumptions of the theorem, equation (27) becomes a partial case of the well known Bernoulli equation, which allows us to express the function \(x(u)\) in the closed form (see, for example, [20]):

\[
x(u) = a + \int_{u}^{u} \frac{d\eta}{\sqrt{(u')^2 + 2 \int_{\eta}^{\xi} N(\xi) d\xi}},
\]

(56)

From (37), (38), (39) and the Picard-Lindelöf theorem (see [17, p. 38]) it follows that function \(x(u)\) belongs to \(C^3([0, +\infty))\) and is the unique solution to the IVP (27), (28) on \([0, +\infty)\).

Using inequalities (39), (40), assumption \(u'_1 > 0\) and the Limit Comparison Theorem for Improper Integrals, from (56) we can easily derive that \(x(u)\) is a monotonically increasing function on \([0, +\infty)\) with bounded range:

\([0, +\infty) \xrightarrow{x(\cdot)} [a, S], \quad S = \lim_{u \to +\infty} x(u) < +\infty.\]

The latter fact means that its inverse \(u(x)\), exists on \([a, S]\) and is the unique solution to (11), (19) on the segment. As it follows from equation (56), conditions (39), (40) also mean that \(x'(u)\) is a monotonically decreasing towards zero function

\[
\lim_{u \to +\infty} x'(u) = 0,
\]

(57)

and, consequently, \(u'(x)\) is a function which monotonically increase towards infinity as \(x\)
tends to $S$:

$$u'(\xi_1) < u'(\xi_2), \ \forall \xi_1, \xi_2 \in [a, S] : \xi_1 < \xi_2, \ \lim_{x \to +S} u'(x) = +\infty. \ (58)$$

From (58) it follows that for each $\delta \geq u'_l$, there exists a unique $x_\delta \in [a, S)$ such that

$$u'(x_\delta) = \delta.$$ 

This in conjunction with the fact that function $u'(x)$ is convex on $[a, S)$, allows us to establish the following inequality (see Figure 1)

$$\max_{x \in [a, S]: u'(x) \leq \delta} |u(x)| = \int_a^{x_\delta} u'(\xi) d\xi \leq \frac{1}{2} (x_\delta - a)(\delta - u'_l) < \frac{1}{2} (S - a)(\delta - u'_l), \ (59)$$

which is of crucial importance for the rest of the proof.

Without loss of generality, we confine ourselves to consider a case when $u'_l < 1$.

Using notation

$$e_i = \max\{|u(x_i) - u_i|, |u'(x_i) - u'_i|\},$$

we can easily find

$$e_0 = 0.$$ 

From Theorem 3 it follows that

$$e_1 \leq h^3 K \|u(x)\|_{x \in [a, a + h]} \exp(hM),$$

Figure 1: Graph of $u = u'(x)$ (solid line). Area of the shaded region is equal to $\int_a^{x_\delta} u'(\xi) d\xi$, which is, apparently, less or equal to the area of $\triangle ABC$, which, in turn, is equal to $\frac{1}{2} (x_\delta - a)(\delta - u'_l)$.
where
\[ K = \frac{1}{2} \| N''(u(x))(u'(x))^2 + N'(u(x))u(x)u(x) \|_{[a,a+h]}, \]
\[ M = \max \{1, |N(u_0)| + |N'(u_0)u'_0|h\}. \]

In general case, \( e_i \) can be estimated from the system of differential equation
\[
\dot{Z}_i(x) = \begin{bmatrix} 0 & 1 \\ N(u_{i-1}) + N'(u_{i-1})u'_{i-1}(x-x_{i-1}) & 0 \end{bmatrix} Z_i(x) + \begin{bmatrix} 0 \\ F_i(x)u(x) \end{bmatrix},
\]
\[ x \in [x_{i-1}, x_i], \ Z_i(x_i) = Z_{i-1}(x_i), \]
where
\[ F_i(x) = N(u(x)) - N(u_{i-1}) - N'(u_{i-1})u'_{i-1}(x-x_{i-1}), \]
\[ Z_i(x) = \begin{bmatrix} z_i(x) \\ z'_i(x) \end{bmatrix}, \ z_i(x) = u(x) - U(N'(u_{i-1})u'_{i-1}, N(u_{i-1}), u'_{i-1}, x-x_{i-1}), \ i = 1, 2, \ldots, \]
\[ Z_0(x) \equiv 0. \] (61)

We are not going to estimate \( e_i \) for all integer \( i \), but only for those satisfying inequality \( i \leq i^* \), where \( i^* \) is defined in (42). However, at this point, the very existence of such an integer value \( i^* \) is yet to be proved. Let us fix some arbitrary \( \varepsilon \in (0, 1/6) \) and assume that
\[ e_j \leq \| Z_i(x) \|_{x \in [x_{j-1}, x_j]} < \varepsilon, \ \forall j : u'(x_j) \leq 1 + 3\varepsilon, \ u_j < 1, \] (62)
for sufficiently small values of \( h \).

We are also free to consider the constants \( L_i^*, i = 0, 1, 2 \) defined in (45), where
\[ M^* = \frac{1}{2} (S - a)(1 + 3\varepsilon - u'_0) > \max_{x \in [a;S]; u'(x) \leq 1+3\varepsilon} |u(x)|. \] (63)
Here, in the latter inequality of (63), we have used (59).

Now, requiring that
\[ h < \frac{\varepsilon}{L_0^*M^*}, \] (64)
we can easily prove that \( i^* \) exists and \( x_{i^*} \) belongs to \((a, u^{-1}(1 + 2\varepsilon)) \). Indeed, if there exists \( x_j \in (a, u^{-1}(1 + 2\varepsilon)) \), such that \( u'_j \geq 1 \), then we can put
\[ i^* = \min_{u'_j \geq 1} \{ j \}. \]
If this is not the case at least for a single \( h \) satisfying (64), then, from (64) it follows that

\[
h < u'^{-1}(1 + 2\varepsilon) - u'^{-1}(1 + \varepsilon) \geq \frac{\varepsilon}{\max_{x \in (a, u'^{-1}(1 + 2\varepsilon)) |u''(x)|} \geq \frac{\varepsilon}{L^*M^*},
\]

which, in turn, means that there exists at least one \( x_j \) belonging to the interval

\[
(u'^{-1}(1 + \varepsilon), u'^{-1}(1 + 2\varepsilon)].
\]

Taking into account (62), the latter fact yields us \( u'_j \geq 1 \) and, consequently, we get a contradiction.

Note, that constants defined in (45), (63), do not depend on \( h \) and, technically, have nothing to do with the assumption (62). Using the constants and assumption (62), from (60) we can derive estimation for \( \|Z_i(x)\|_{x \in [x_{i-1}, x_i]} \), for \( i = 1, 2, \ldots, i^* \) in the following way:

\[
\|Z_i(x)\|_{x \in [x_{i-1}, x_i]} \leq (1 + hM^* (L_1^* + h(L_1^* + L_2^*))) \|Z_{i-1}(x)\|_{x \in [x_{i-2}, x_{i-1}]} + \]

\[
+ \max \{1, L_0^* + L_1^* h\} \int_{x_{i-1}}^{x} \|Z_i(\xi)\|d\xi + h^3 \frac{M^*}{2} (L_2^* + L_1^* L_0^* M^*),
\]

\[
\forall x \in [x_{i-1}, x_i], \ i = 1, 2, \ldots, i^*, \ \|Z_0(x)\|_{x \in [x_{i-1}, x_0]} \equiv 0.
\]

Applying the Gronwall’s inequality (see, for example, [17, 42]) to (65) we get

\[
\|Z_i(x)\|_{x \in [x_{i-1}, x_i]} \leq \left( (1 + hM^* (L_1^* + h(L_1^* + L_2^*))) \|Z_{i-1}(x)\|_{x \in [x_{i-2}, x_{i-1}]} + h^3 \frac{M^*}{2} (L_2^* + L_1^* L_0^* M^*) \right) \times \]

\[
\times \exp \left( h \left( \max \{1, L_0^* + L_1^* h\} \right) \right), \ i = 1, 2, \ldots, i^*.
\]

From (66), taking into account (61), we can get the estimation

\[
\|Z_i(x)\|_{x \in [x_{i-1}, x_i]} \leq h^3 \frac{M^*}{2} (L_2^* + L_1^* L_0^* M^*) \sum_{j=1}^{j=i} (1 + hM^* (L_1^* + h(L_1^* + L_2^*)))^{j-1} \exp (jhE^*) =
\]

\[
= h^3 \frac{M^*}{2} (L_2^* + L_1^* L_0^* M^*) \exp (hE^*) \left( \frac{1 + hM^* (L_1^* + h(L_1^* + L_2^*)))^i \exp (ihE^*) - 1}{(1 + hM^* (L_1^* + h(L_1^* + L_2^*))) \exp (hE^*) - 1} <
\]

\[
< h^3 \frac{M^*}{2} (L_2^* + L_1^* L_0^* M^*) \exp (hE^*) \frac{\exp ((S - a) (E^* + M^* (L_1^* + h(L_1^* + L_2^*)))) - 1}{(1 + hM^* (L_1^* + h(L_1^* + L_2^*))) \exp (hE^*) - 1} \leq
\]

\[
\leq h^2 \frac{M^*}{2} (L_2^* + L_1^* L_0^* M^*) \frac{\exp ((S + 1)E^* + S^* M^* (2L_1^* + L_2^*))}{L_1^* M^* + E^*} = h^2 P^*,
\]

16
where
\[ E^* = \max \{1, L_0^* + L_1^* h\}, \quad \overline{E}^* = \max \{1, L_0^* + L_1^*\}, \]
\[ \tilde{E}^* = \max \{1, L_0^*\}, \quad S^* = S - a. \]

The last inequality in (67) holds true under assumption that
\[ 0 \leq h \leq 1, \quad (68) \]
which we accept from now on.

Going back to inequality (65), it is important to mention that to derive it for each particular \( i = 1, 2, \ldots, i^* \), we have used assumption (62) for \( j = i - 1 \) only. Besides that, the inequality (65) for \( i = 1 \) does not rely upon (62) at all and follows from Theorem 3. With this in mind and taking into account (67), we can easily prove (by means of mathematical induction) that for \( h \) satisfying
\[ h < \sqrt{\frac{\varepsilon}{P^*}}, \quad (69) \]
assumption (62) holds true. This concludes proof of the first part of the theorem, which states the existence of \( i^* \) (42) and the fulfillment of approximation estimates (43), provided that \( h \) satisfies (41).

Above we showed that \( x_{i^*} \in [\alpha_0, \alpha_1] \), where \( \alpha_0 = u^{r-1}(1 - P^* h^2), \alpha_1 = u^{r-1}(1 + 2\varepsilon) \). Using similar reasoning and restrictions (64), (69), it is easy to verify that
\[ u^{r-1}(1 - 2\varepsilon) \leq x_{i^*} - h \leq x_{i^*} + h \leq u^{r-1}(1 + 3\varepsilon). \quad (70) \]

In what follows we also will need the estimate
\[ \min_{x \in [x_{i^*} - h, x_{i^*} + h]} u'(x) = u'(x_{i^*} - h) = \quad (71) \]
\[ = u'(x_{i^*} - h) - u'(x_{i^*}) + u'(x_{i^*}) \geq u'(x_{i^*} - h) - u'(x_{i^*}) + u'(\alpha_0) \geq \]
\[ \geq -u''(u^{r-1}(1 + 3\varepsilon))h + 1 - P^* h^2 \geq 1 - P^* h^2 - L_0^* M^* h = \tau(h). \]

Using (71) we can easily find that
\[ \max_{u \in [u(x_{i^*} - h), u(x_{i^*} + h)]} x'(u) \leq (\tau(h))^{-1}. \quad (72) \]

Now if we require that
\[ h \leq \frac{\tau(h)}{P^*}, \quad (73) \]
then

\[ u_{i^*} \in [u(x_{i^*} - h), u(x_{i^*} + h)], \quad (74) \]

indeed

\[ u_{i^*} - u(x_{i^*} - h) = u_{i^*} - u(x_{i^*}) + u(x_{i^*}) - u(x_{i^*} - h) \geq h\tau(h) - P^*h^2 \geq 0, \]

\[ u(x_{i^*} + h) - u_{i^*} = u(x_{i^*} + h) - u(x_{i^*}) + u(x_{i^*}) - u_{i^*} \geq h\tau(h) - P^*h^2 \geq 0. \]

From (74) it follows that \( u(x_{i^*}) + \theta(u_{i^*} - u(x_{i^*})) \in [u(x_{i^*} - h), u(x_{i^*} + h)], \forall \theta \in [0, 1]. \)

With this in mind, and taking into account (72), we can easily derive the inequality

\[ |x(u_{i^*}) - x| = |x(u_{i^*}) - x(u(x_{i^*}))| = x'(u(x_{i^*}) + \theta(u_{i^*} - u(x_{i^*})))|u_{i^*} - u(x_{i^*})| \leq \]

\[ \leq (\tau(h))^{-1} P^*h^2. \]

where \( 0 \leq \theta \leq 1. \)

From (70) it follows that

\[ \max_{u \in [u(x_{i^*} - h), u(x_{i^*} + h)]} |x''(u)| = \max_{u \in [u(x_{i^*} - h), u(x_{i^*} + h)]} N(u)u(x'(u))^3 \leq \frac{L_0^*M^*}{(1 - 2\varepsilon)^3}. \quad (76) \]

Now using (74) and inequality (76) we can get the estimate

\[ |x'(u_{i^*}) - x'\bar{u}| \leq |x'(u_{i^*}) - x'(u(x_{i^*}))| + |x'(u(x_{i^*})) - x'\bar{u}| \leq \]

\[ \leq \frac{L_0^*M^*P^*}{(1 - 2\varepsilon)^3} h^2 + \left| \frac{1}{u'(x_{i^*})} - \frac{1}{u'\bar{u}} \right| \leq \left( \frac{L_0^*M^*}{(1 - 2\varepsilon)^2} + 1 \right) \frac{P^*}{1 - 2\varepsilon} h^2. \]

At this point we have proved that after the first \( i^* \) iterations the algorithm should switch from formulas (35) to formulas (36). The lemma below, among other things, states that starting from \( i^* \) the algorithm will never switch back to formulas (35), i.e. \( u' \) remains greater or equal to 1, for all \( i > i^* \), provided \( h \) is small enough. For the sake of convenience, hereinafter we will use notation \( N(u) = N(u)u. \)

**Lemma 2.** Let \( h > 0 \) satisfy condition

\[ h < \frac{1}{3\mu}, \quad (78) \]

where constant \( \mu \) is defined in (51). Then

\[ A_i \frac{u^2}{2} + B_i u \leq 0 \quad \forall u \in [0, h], \quad (79) \]
\[ 0 \leq x_i' \leq \left( \frac{1}{(x_i')^2} + \int_{u_{i-1}}^{u_i} \mathcal{N}(u) \, du \right)^{-\frac{1}{2}}, \quad i = i^*, i^* + 1, \ldots \quad (80) \]

**Proof.** Let us consider an auxiliary sequence

\[ x_i' = x_{i-1}' \exp \left( s_{1,i} \mathcal{N}(u_{i-1}) \left( x_{i-1}' \right)^2 h + s_{2,i} \right), \quad s_{1,i} \leq \frac{-1}{2}, \; s_{2,i} \leq 0, \quad i = i^* + 1, i^* + 2, \ldots, \quad x_{i^*}' = x_{i^*}' = \frac{1}{u_{i^*}'}. \]

It is easy to see that \( x_{i^*}' > 0 \) and from

\[ (x_i')^2 - \left( \frac{1}{(x_i')^2} + \mathcal{N}(u_{i-1}) h \right)^{-1} = \]

\[ = (x_{i-1}')^2 \exp \left( 2s_{1,i} \mathcal{N}(u_{i-1}) (x_{i-1}')^2 h + 2s_{2,i} \right) - \frac{(x_{i-1}')^2}{1 + (x_{i-1}')^2 \mathcal{N}(u_i) h} = \]

\[ = (x_{i-1}')^2 \left( \exp \left( 2s_{1,i} \mathcal{N}(u_{i-1}) (x_{i-1}')^2 h + 2s_{2,i} \right) \left( 1 + (x_{i-1}')^2 \mathcal{N}(u_{i-1}) h \right) - 1 \right) \leq \]

\[ \leq \frac{(x_{i-1}')^2}{1 + (x_{i-1}')^2 \mathcal{N}(u_i) h} \]

it follows that

\[ x_i' \leq \left( \frac{1}{(x_{i-1}')^2} + \mathcal{N}(u_{i-1}) h \right)^{-\frac{1}{2}}. \quad (82) \]

Applying inequality (82) recursively we get the estimate

\[ x_i' \leq \left( \frac{1}{(x_{i-1}')^2} + \sum_{j=i^*}^{i-1} \mathcal{N}(u_j) h \right)^{-\frac{1}{2}} \leq \left( 1 + \int_{u_{i^*} - h}^{u_i} \mathcal{N}(u) du - \mathcal{N}(u_i) h \right)^{-\frac{1}{2}}. \quad (83) \]

To derive the last inequality in (83) we exploit the fact that function \( \mathcal{N}(u) \) is non-decreasing (see (48)). From (83), using (78) we can easily get

\[ 0 \leq \mathcal{N}(u_i) (x_i')^2 h \leq \mathcal{N}(u_i) h \left( 1 + \int_{u_{i^*} - h}^{u_i} \mathcal{N}(u) du - \mathcal{N}(u_i) h \right)^{-1} \leq \frac{\mu h}{1 - \mu h} \leq \frac{1}{2}, \quad (84) \]

\[ i = i^*, i^* + 1, \ldots \]

19
Inequality \((81)\) together with \((39)\) and \((48)\) imply that estimate \((83)\) remains valid if
\[
s_{1,i} = \mathcal{N}(u_{i-1}) \left( x'_{i-1} \right)^2 h - 1 \leq -\frac{1}{2}, \quad s_{2,i} = -\frac{h^2}{2} \mathcal{N}'(u) \left( x'_{i-1} \right)^2.
\] (85)

On the other hand, sequence \(\{x'_i\}\) \((81)\) together with substitution \((85)\) totally coincide with sequence \(\{x'_i\}\) \((36)\):
\[
x'_i = x'_{i-1} \exp \left( -h \mathcal{N}(u_{i-1}) \left( x'_{i-1} \right)^2 + \frac{h^2}{2} \left( -\mathcal{N}'(u_{i-1}) \left( x'_{i-1} \right)^2 + 2 \left( \mathcal{N}(u_{i-1}) \left( x'_{i-1} \right)^2 \right)^2 \right) \right) =
\]
\[
x'_i = x'_{i-1} \exp \left( A_i \frac{h^2}{2} + Bh \right), \quad i = i^* + 1, i^* + 2, \ldots.
\] (86)

In the light of the latter observation, estimates \((80)\), immediately follow from \((83)\), whereas inequalities \((79)\) follow from \((84)\), \((85)\). \(\square\)

From now on we assume that \(h\) satisfies requirements \((78)\), \((73)\), which is equivalent to \((50)\).

Let us now consider a sequence of functions \(\{y_i(u)\}\), defined as follows
\[
y_i(u) = x(u) - V_i(u),
\] (87)
\[
V_i(u) = V(A_i, B_i, x'_{i-1}, x_{i-1} - u, u_{i-1}),
\]
\[
A_i = -\mathcal{N}'(u_{i-1}) \left( x'_{i-1} \right)^2 + 2 \left( \mathcal{N}(u_{i-1}) \right)^2 \left( x'_{i-1} \right)^4,
\]
\[
B_i = -\mathcal{N}(u_{i-1}) \left( x'_{i-1} \right)^2, \quad i = i^* + 1, i^* + 2, \ldots.
\]

It is easy to see that \(y_i(u)\) should satisfy the recurrence system of Cauchy problems
\[
y_i''(u) = G_i(u)y'_i(u) + F_i(u)x'(u),
\] (88)
\[
F_i(u) = -\mathcal{N}(u) \left( x'(u) \right)^2 - A_i(u - u_{i-1}) - B_i, \quad u \in [u_{i-1}, u_i],
\]
\[
G_i(u) = B_i + A_i(u - u_{i-1}),
\]
\[
y_i(u_i) = y_{i+1}(u_i), \quad y'_i(u_i) = y'_{i+1}(u_i), \quad u_i = u_{i-1} + h, \quad i = i^* + 1, i^* + 2, \ldots.
\] (89)

Inequalities \((75)\), \((77)\) allow us to estimate \(|y_{i^*}(u_{i^*})|\), \(k = 0, 1\) in the following way:
\[
|y_{i^*}(u_{i^*})| \leq \frac{P^*h^2}{1 - P^*h^2 - L^*_0M^*h}, \quad |y'_{i^*}(u_{i^*})| \leq \left( \frac{L^*_0M^*}{(1 - 2\varepsilon)^2} + 1 \right) \frac{P^*h^2}{1 - 2\varepsilon}.
\] (90)

Using mean value theorem, we can easily find that
\[
F_i(u) = \frac{(u - u_{i-1})^2}{2} \left( -\mathcal{N}''(u_{i-1} + \theta(u - u_{i-1})) \right) \left( x'(u_{i-1} + \theta(u - u_{i-1})) \right)^2 + \quad (91)
\]
+6N(u_{i-1} + \theta(u - u_{i-1}))N'(u_{i-1} + \theta(u - u_{i-1}))(x'(u_{i-1} + \theta(u - u_{i-1})))^4 - \\
-8(N(u_{i-1} + \theta(u - u_{i-1})))^3(x'(u_{i-1} + \theta(u - u_{i-1})))^3 - \\
-N(u_{i-1})\left((x'(u_{i-1}))^2 - (x'_{i-1})^2\right) + \\
(u - u_{i-1})\left(-N'(u_{i-1})\left((x'(u_{i-1}))^2 - (x'_{i-1})^2\right) + 2(N(u_{i-1}))^2\left((x'(u_{i-1}))^4 - (x'_{i-1})^4\right)\right), \\
\forall u \in [u_{i-1}, u_i], \theta = \theta(u) \in (0, 1), \\
which, together with (80) and (84), yields us an estimate \(^{4}\)
\[
\begin{aligned}
&= \left( (1 - 2\varepsilon)^2 + 2 \int_{u_i^*}^u \mathcal{N}(\xi) d\xi + 2\mathcal{N}(u_i^*)h - 2\mathcal{N}(u_i^*)h \right)^{\frac{1}{2}} \\
&\leq \left( (1 - 2\varepsilon)^2 + 2 \int_{u_i^*}^u \mathcal{N}(\xi) d\xi + 2 \int_{u_i^* - h}^u \mathcal{N}(\xi) d\xi - 2\varepsilon \right)^{\frac{1}{2}} \\
&= \left( (1 - 2\varepsilon)^2 - 2\varepsilon + 2 \int_{u_i^* - h}^u \mathcal{N}(\xi) d\xi \right)^{\frac{1}{2}} \\
&\leq \left( 1 - 6\varepsilon + 2 \int_{u_i^* - h}^u \mathcal{N}(\xi) d\xi \right)^{\frac{1}{2}}.
\end{aligned}
\]

In a similar way, from (80) we get

\[
x_i \leq \left( 1 + \int_{u_i^* - h}^{u_i - h} \mathcal{N}(\xi) d\xi \right)^{\frac{1}{2}} = \left( 1 + \int_{u_i^* - h}^{u_i - h} \mathcal{N}(\xi) d\xi + \mathcal{N}(u_i^* - h)h - \mathcal{N}(u_i^* - h)h \right)^{\frac{1}{2}} \\
\leq \left( 1 - \varepsilon + \int_{u_i^* - 2h}^{u_i - h} \mathcal{N}(\xi) d\xi \right)^{\frac{1}{2}}, \quad i = i^*, i^* + 1, \ldots.
\]

Solution to (88) can be expressed in the form

\[
y'_i(u) = \int_{u_{i-1}}^u \exp \left( \int_{\xi}^{u_{i-1}} G_i(\zeta) d\zeta \right) F_i(\xi) x'(\xi) d\xi + y'_{i-1}(u_{i-1}) \exp \left( \int_{u_{i-1}}^{u_{i-1}} G_i(\zeta) d\zeta \right), \quad (93)
\]

\[
y_i(u) = \int_{u_{i-1}}^u y'_i(\xi) d\xi + y_{i-1}(u_{i-1}), \quad (94)
\]

\[
u \in [u_{i-1}, u_i], \quad i = i^*, i^* + 1, i^* + 2, \ldots.
\]

Using initial estimates (90), from (93) we can easily get inequalities

\[
\| y'_i(u) \|_{[u_{i-1}, u_i]} \leq \mathcal{E}_i \left( |y'_{i^*}(u_i^*)| + \frac{h^2}{2} \int_{u_i^*}^{u_i} \mathcal{T}(\zeta) d\zeta \right), \quad i = i^* + 1, i^* + 2, \ldots, \quad (95)
\]

where \( \mathcal{E}_i \) and \( \mathcal{T}(\zeta) \) are defined in (54) and (53) respectively. Combining (95) with (94), we get

\[
\| y_i(u) \|_{[u_{i-1}, u_i]} \leq |y'_{i^*}(u_i^*)| +
\]

22
Estimates (52), (53) follows immediately from (90) and estimates (95), (96) respectively. This completes the proof.

As it can be easily verified, the estimations given in Theorem 5 are quite rough and become almost useless when, for instance, applied to the Troesch’s equation (4) with \( \lambda \) sufficiently large. The main reason for this is the roughness of estimate (63). The latter can be improved, which is addressed in the remark below.

Remark 1. The estimates of Theorem 5 remain valid and can be improved if the constant \( M^* \) defined by formula (46) is substituted by the one defined as

\[
M^* = \Phi^{-1}\left(\frac{1}{2} \left(1 + 3\varepsilon \right)^2 - \left(u_l^0\right)^2 \right), \quad \Phi(u) = \int_{u_l}^{u} N(\xi)\xi d\xi \tag{97}
\]

and the constant \( P^* \) defined at (44) is treated as a function of \( h \) defined as

\[
P^*(h) = \frac{h M^*}{2} (L_2^* + L_1^* L_0^* M^*) \exp \left( h E^* \right) \frac{\exp \left((S - a) \left( E^* + M^* (L_1^* + h(L_1^* + L_2^*))\right)\right) - 1}{(1 + h M^* (L_1^* + h(L_1^* + L_2^*))) \exp \left( h E^* \right) - 1}, \tag{98}
\]

where

\[
E^* = \max \{1, L_0^* + L_1^* h\}.
\]

Remark 2. If the IVP (1), (19) is considered on some finite interval, i.e. \( b < +\infty \) then the constant \( S^* \) defined in (47) can be substituted with

\[
S^* = \min \left\{ b, \lim_{u \to +\infty} \int_{u_l}^{u} \frac{d\eta}{\sqrt{(u_l')^2 + 2 \int_{u_l}^{\eta} N(\xi)\xi d\xi}} \right\} \tag{99}
\]

in order to make error estimates of Theorem 5 more precise.

4 Straight-Inverse method for solving BVPs for second order differential equations.

In this section we introduce the SI-method for solving boundary value problems. As a matter of the fact we are going to consider the simple and multiple shooting techniques supplemented by the SI-method for solving IVPs described above.
Since the *simple shooting technique* is nothing but a bisection algorithm supplemented by an IVP solver, the meaning of the *SI simple shooting* method is evident and self-explanatory. At the same time the *SI multiple shooting method* requires considerably deeper introduction and the rest of the current section is devoted strictly to it.

### 4.1 SI multiple shooting method.

Talking about the *multiple shooting technique* for solving boundary value problems we (as a rule) mean a way how the given BVP can be transformed to a system of nonlinear algebraic equations together with an algorithm for solving the system.

Assume that we fixed some positive parameter $h$, hereinafter referenced to as a *step size*. In addition to that we have at our disposal some initial guess $\Omega_k$ which is a discrete approximation (5 of the exact solution $u(x)$ of the BVP (1), (2):

$$\Omega_k = \{\omega_{k,i} = (u'_{k,i}, u_{k,i}, x_{k,i}) | x_{k,i} \in [a, b], u(x_{k,i}) \approx u_{k,i}, u'(x_{k,i}) \approx u'_{k,i}, i \in 0, N_k, x_{k,i} < x_{k,j} \Leftrightarrow i < j < N_k \} \quad (100)$$

Here we avoid discussing on how close the approximation should be, however, in practice, if the approximation is too rough the method described below can do not work at all. At the same time, using the simple shooting approach described above, it is always possible to get the desired approximation.

Given that, we can transform the set $\Omega_k$ into an ordered set of nonlinear equations

$$\Sigma_k = \{\sigma_{k,i,j}, i \in 0, N_k - 1, j = 0, 1\}$$

using the rule described below.

The first two equations can be represented in the form of

$$\sigma_{k,0,0}(\omega_{k,0}, \omega_{k,1}) \overset{\text{def}}{=}$$

$$\left\{ \begin{array}{l}
U(A_U(x_{k,0}, u_{k,0}, u'_0), B_U(x_{k,0}, u_{k,0}, u'_0), u'_0, u_{k,0}, h_{k,0}) = u_1, |u'_{k,0}| \leq 1, \\
V(A_V(u_{k,0}, x_{k,0}, x'_0), B_V(u_{k,0}, x_{k,0}, x'_0), x'_0, x_{k,0}, \bar{h}_{k,0}) = x_1, |u'_{k,0}| > 1,
\end{array} \right. \quad (102)$$

(5) Here we avoid discussing on how close the approximation should be, however, in practice, if the approximation is too rough the method described below can do not work at all. At the same time, using the simple shooting approach described above, it is always possible to get the desired approximation.
\[
\sigma_{k,0,1}(\omega_{k,0}, \omega_{k,1}) \overset{def}{=} \begin{cases}
U'_h(A_U(x_{k,0}, u_{k,0}, u'_0), B_U(x_{k,0}, u_{k,0}, u'_0), u'_0, u_{k,0}, h)|_{h=b_{k,0}}, & |u'_0| \leq 1, \\
V'_h(A_V(u_{k,0}, x_{k,0}, x'_0), B_V(u_{k,0}, x_{k,0}, x'_0), x'_0, x_{k,0}, h)|_{h=b_{k,0}}, & |u'_k| > 1,
\end{cases}
\]

(103)

For \(i \in \overline{1, N_k-1}\) we have:

\[
\sigma_{k,i,0}(\omega_{k,i-1}, \omega_{k,i}, \omega_{k,i+1}) \overset{def}{=} \begin{cases}
U(A_U(x_{k,i}, u_i, u'_i), B_U(x_{k,i}, u_i, u'_i), u'_i, u_i, h_{k,i}), & |u'_{k,i}| \leq 1, |u'_{k,i-1}| \leq 1, \\
U(A_U(x_i, u_{k,i}, u'_i), B_U(x_i, u_{k,i}, u'_i), u'_i, u_{k,i}, x_{k,i+1} - x_i), & |u'_{k,i}| \leq 1, |u'_{k,i-1}| > 1, \\
V(A_V(u_{k,i}, x_i, x'_i), B_V(u_{k,i}, x_i, x'_i), x'_i, x_i, h_{k,i}), & |u'_{k,i}| > 1, |u'_{k,i-1}| > 1, \\
V(A_V(u_i, x_{k,i}, x'_i), B_V(u_i, x_{k,i}, x'_i), x'_i, x_{k,i}, u_{k,i+1} - u_i), & |u'_{k,i}| > 1, |u'_{k,i-1}| \leq 1,
\end{cases}
\]

(104)

\[
\sigma_{k,i,1}(\omega_{k,i-1}, \omega_{k,i}, \omega_{k,i+1}) \overset{def}{=} \begin{cases}
U'_h(A_U(x_{k,i}, u_i, u'_i), B_U(x_{k,i}, u_i, u'_i), u'_i, u_i, h)|_{h=b_{k,i}}, & |u'_{k,i-1}| \leq 1, |u'_{k,i}| \leq 1, \\
U'_h(A_U(x_i, u_{k,i}, u'_i), B_U(x_i, u_{k,i}, u'_i), u'_i, u_{k,i}, h)|_{h=x_{k,i+1} - x_i}, & |u'_{k,i-1}| > 1, |u'_{k,i}| \leq 1, \\
V'_h(A_V(u_{k,i}, x_i, x'_i), B_V(u_{k,i}, x_i, x'_i), x'_i, x_i, h)|_{h=b_{k,i}}, & |u'_{k,i-1}| > 1, |u'_{k,i}| > 1, \\
V'_h(A_V(u_i, x_{k,i}, x'_i), B_V(u_i, x_{k,i}, x'_i), x'_i, x_{k,i}, h)|_{h=u_{k,i+1} - u_i}, & |u'_{k,i-1}| \leq 1, |u'_{k,i}| > 1,
\end{cases}
\]

(105)
where

\[
A_U(x, u, u') = N'_u(u, x)u' + N'_x(u, x),
\]
\[
B_U(x, u, u') = N(u, x),
\]
\[
A_V(u, x, x') = -((N'_u(u, x) + N'_x(u, x)x')u + N(u, x)) (x')^2 + 2 (N(u, x)u)^2 (x')^4,
\]
\[
B_V(u, x, x') = -N(u, x)u (x')^2,
\]

(106)

bold variables describe unknowns and

\[
u'_i \overset{\text{def}}{=} 1/x'_i, \forall i \in 0, N_k.
\]

As you can see, the equations are dependent on the absolute values of \(u'_{k,i}\), which, according to our assumptions about \(\Omega_k\), characterize rapidity of variation of the unknown solution \(u(x)\) at different points of segment \([a, b]\). This follows the general idea of the straight-inverse approach consisting in switching between the straight (i.e. \(u(x)\)) and inverse (i.e. \(x(u)\)) solutions depending on which of the two behaves better (that is, possesses lower variation in a vicinity of a given point).

Using \(\Omega_k\) as an initial guess and applying a single iteration of the generalized Newton’s method (see, for example, [16, p. 293]) to the system \(\Sigma_k\) (102), (103), (104), (105) we will get a new set \(\Omega_{k+1}\) as a combination of \(\Omega_k\) and the results brought by the Newton’s method iteration, assuming that

\[
u_{k+1,i} \overset{\text{def}}{=} u_i, \quad u'_{k+1,i} \overset{\text{def}}{=} u'_i,
\]
\[
x_{k+1,i} \overset{\text{def}}{=} x_i, \quad x'_{k+1,i} \overset{\text{def}}{=} x'_i
\]

wherever it makes sense. In practice it happens that the set \(\Omega_{k+1}\) obtained as described above needs to be sorted out (to fulfill the requirement \(x_{k+1,i} < x_{k+1,j} \Leftrightarrow i < j < N_{k+1}\)) and then refined (by linear interpolation to satisfy the inequality \(\max\{h_{k+1,i}, h_{k+1,i}'\} \leq h, \forall i \in 0, N_{k+1} - 1\)). Once it is done, we can use \(\Omega_{k+1}\) to construct a new system \(\Sigma_{k+1}\) and, applying another iteration of the Newton’s method to it, get \(\Omega_{k+2}\) and so on and so forth, until the difference between the two subsequent \(\Omega\)s is not small enough.\(^6\)

5 Numerical examples

In this section we present numerical results of the SI-method applied to the Troesch’s problem \(\Box\). At the same time, in what follows, we evaluate and examine estimates of Theorem \(\Box\) for the Troesch’s equation with some fixed value of the parameter \(\lambda\).

Using the algorithm described in the previous section, the Troesch’s problem was solved

\(^6\)Of course, it might not always be the case and requires the iteration process to be convergent.
for multiple values of $\lambda$ and the initial slopes $u'(0)$ of the solutions can be found in Tab. 1. For the purpose of comparison, the table also contains the corresponding values calculated by other methods and algorithms. The two rightmost columns of Tab. 1 contain slopes calculated by the SI-method with different values of step size $h$. Comparing these values to those calculated by other methods, we can get a general impression about the order of approximation of the SI-method with respect to $h$. With a great deal of evidence, we can conclude that the order is 2, which is perfectly coherent with the results of Theorem 5.

| $\lambda$ | 10 | 9 | Maple 2016 (7) | SI-method, $h = 10^{-4}$ | SI-method, $h = 10^{-5}$ |
|---------|----|---|----------------|--------------------------|--------------------------|
| 2       | 0.5186322404 | – | 0.518621219269 | 0.518621219577035 | 0.51862121922419 |
| 3       | 0.255607567 | – | 0.255604215562 | 0.255604216455332 | 0.25560421571849 |
| 5       | 4.57504633e-02 | – | 4.575046140632e-02 | 4.575046196263e-02 | 4.575046141188e-02 |
| 8       | 2.587169418e-03 | – | 2.587169418963e-3 | 2.587169500425e-03 | 2.587169419777e-03 |
| 20      | 1.648773182e-08 | 1.6487734e-8 | – | 1.648773647e-08 | 1.648773188e-00 |
| 30      | 7.486093793e-13 | 7.4861194e-13 | – | 7.486093841e-13 | 7.486093844e-13 |
| 50      | 1.542999878e-21 | 1.5430022e-21 | – | 1.543002448e-21 | 1.542999906e-21 |
| 61      | – | 2.5770722e-26 | – | 2.577078525e-26 | 2.577072999e-26 |
| 100     | 2.976060781e-43 | – | – | 2.976075557e-043 | 2.976060927e-043 |

Table 1: Values of $u'(0)$ for the Troesch’s problem calculated by different approaches.

As soon as the initial slope $u'(0)$ is known, we can use Theorem 5 to calculate error estimates. Let us do that for $\lambda = 2$, using value $u'(0)$ from Tab. 1 calculated by Maple 2016. As it was pointed out in [15], the initial value problem associated with (4) has a pole approximately in

$$x_\infty = \frac{1}{\lambda} \ln \left( \frac{8}{u'(0)} \right).$$

This allows us to get an approximation for $S^*$

$$S^* \approx x_\infty = \frac{1}{2} \ln \left( \frac{8}{0.518621219269} \right) \approx 1.368011516.$$

At the same time, Remark 2 allows us to lower the value of $S^*$, taking into account that in the case of Troesch’s problem $b = 1$:

$$S^* = \min\{1.0, 1.368011516\} = 1.0.$$

Now using Remark 3 and taking into account that for the Troesch’s problem

$$\Phi(u) = \cosh(\lambda u) - \cosh(\lambda u_t),$$

(7) Using numeric "dsolve" procedure with "abserr = 1e-12"
we can calculate $M^*$ in the following way

$$
M^* = \frac{1}{\lambda} \cosh^{-1}
\left( \frac{1}{2} \left( (1 + 3\varepsilon)^2 - (u_l')^2 \right) + \cosh(\lambda u_l) \right)
$$

(110)

Assuming that

$$
\varepsilon = 0.1,
$$

we get

$$
M^* \approx \frac{1}{2} \cosh^{-1}
\left( \frac{1}{2} \left( 1.3^2 - 0.518621219269^2 \right) + 1 \right) \approx 0.5654221730.
$$

(111)

With the value of $M^*$ available, we are in the position to evaluate $L_i^*$ via formulas (115):

$$
L_0^* = \max_{|u| < M^* + \varepsilon} \left\{ \frac{\lambda \sinh(\lambda u)}{u} \right\} \approx 5.289849576,
$$

(112)

$$
L_1^* = \max_{|u| < M^* + \varepsilon} \left\{ \left( \frac{\lambda \sinh(\lambda u)}{u} \right)' \right\} \approx 4.218574488,
$$

(113)

$$
L_2^* = \max_{|u| < M^* + \varepsilon} \left\{ \left( \frac{\lambda \sinh(\lambda u)}{u} \right)'' \right\} \approx 8.48000570,
$$

(114)

Using formula (98) we can calculate

$$
P^*(h = 1e - 4) \approx 1675.2, \quad P^*(h = 1e - 5) \approx 1673.5.
$$

(115)

According to the Theorem 5 values (115) give us an error estimate of the SI method on $[0, x_i^*]$, see (43). At the same time, the corresponding data from Table (1) suggests that the estimates (43) are quite overpriced.

| $\lambda$ | Other | Maple 2016 | SI-method, $h = 10^{-4}$ | SI-method, $h = 10^{-5}$ |
|-----------|-------|-------------|--------------------------|--------------------------|
| 2         | 2.406790318 | 2.406939711 | 15 | 2.4069393831274 | 2.40693982969129 | 2.4069398312315 |
| 3         | 4.266151411 | 4.266222862 | 15 | 4.26622285457896 | 4.2662228617306 |
| 5         | 12.10049478 | 12.10049546 | 15 | 12.1004954501778 | 12.1004954506293 |
| 8         | 54.57983465 | 54.57983447 | 15 | 54.5798344412402 | 54.5798344554302 |
| 10        | 148.4064126 | 148.4064212 | 15 | 148.406421145524 | 148.406421155906 |
| 20        | 22026.29966 | 22026.4657 | 15 | 22026.4657404062 | 22026.4657494068 |
| 30        | – | – | – | 3269017.37247181 | 3269017.3724718 |
| 50        | – | – | – | 72004899337.3858 | 72004899337.386 |

Table 2: Values of $u'(1)$ for the Troesch's problem calculated by different approaches.

The order of the SI-method's error on $[u_i^*, 1]$ with respect to $h$ can be estimated empirically from Tab. 2 which contains values of $u'(1)$ calculated by different methods for different values of $\lambda$. The two rightmost columns of the table contain values of $u'(1)$ calculated by

*(Using numeric “dsolve” procedure with “abser = 1e-12”*)
the SI-method with different values of step size $h$. Examining the table, we should keep in mind that the values calculated by other (than SI) methods are actually inverse to those approximated by the SI-method, i.e., roughly speaking, on the segment where derivative of the unknown function $u(x)$ gets bigger than 1 the method approximates values of $x'(u) = \frac{1}{u'(x)}$. Estimates of Theorem 5 does not deal with the values presented in the Tab. 2 but with their inverse. Nevertheless, we still can see that the method’s error is of order 2 with respect to $h$, just as it is predicted by the theorem. At the same time, it is worth mentioning that the scalar multiplier in front of $h^2$ (see (53)), on practice, can be of significantly lower magnitude.

| Value | [6] | [19] | SI-method, $h = 10^{-4}$ | SI-method, $h = 10^{-5}$ |
|-------|-----|-----|-------------------------|-------------------------|
| $u(0.1)$ | 4.211183679705e-05 | 4.211189927237e-05 | 4.21119023173e-05 | 4.21118993037e-05 |
| $u(0.2)$ | 1.299639238293e-04 | 3.58978401386e-04 | 3.58978427345e-04 | 3.58978410657e-04 |
| $u(0.3)$ | 9.779014227050e-04 | 9.779027718029e-04 | 9.77902842508e-04 | 9.77902772532e-04 |
| $u(0.4)$ | 2.659017178062e-03 | 2.659020490351e-03 | 2.659020682593e-03 | 2.65902049234e-03 |
| $u(0.999)$ | 8.889931171768e-01 | 8.889931181558e-01 | 8.89035025083e-01 | 8.88994611223e-01 |

Table 3: Solution to the Troesch’s problem with $\lambda = 10$ obtained by different approaches.

Table 3 presents comparison between approximations of the Troesch’s problem solution $u(x)$ calculated by different methods at points other than the ends of segment $[0, 1]$. The approximations obtained by the SI-method for different values of $h$, confirm that the order of the method’s error with respect to $h$ is 2. It is worth mentioning, that because of specifics of the SI-method, one cannot have a control over the points $x_i$ belonging to the rightmost part of the segment $[0, 1]$, where the absolute value of the derivative $u'(x)$ exceeds 1. In the latter case, the method “works” with the inverse function $x(u)$ and it is rather possible to choose points $u_i$ where to calculate the approximation of $x(u) = u^{-1}(u)$. This explains why the bottom row in the Table 3 contains approximations by the SI-method for value $x$ “close” but not equal to 0.999.

| Source | $u'(0)$ | $\|\Omega(h)\|_{11}$ | CPU time, sec. | Rel. diff. to [19] |
|--------|--------|----------------|--------------|-----------------|
| SIM, $h = 10^{-2}$ | 3.141990565e-43 | 240 | 0.022 | 5.6e-2 |
| SIM, $h = 10^{-3}$ | 2.977378936e-43 | 2208 | 0.054 | 4.4e-4 |
| SIM, $h = 10^{-4}$ | 2.976075557e-43 | 21753 | 0.275 | 5.0e-6 |
| SIM, $h = 10^{-5}$ | 2.97600782e-43 | 203143 | 2.135 | 4.9e-8 |
| SIM, $h = 10^{-6}$ | 2.97600782e-43 | 2081478 | 16.05 | 3.4e-10 |
| [19] | 2.97600781e-43 | – | – | 0.0 |

Table 4: Solution to the Troesch’s problem with $\lambda = 100$.

Tab. 4 allows us to get an insight about the performance of the SI-method and the complexity of the algorithm with respect to the number of knots in the method’s mesh.

---

(9) For $x = 0.9990000491899$
(10) For $x = 0.999000017539$
(11) Number of knots in the final mesh.
(12) Relative difference as compared to $u'(0)$ calculated in [19].
The absolute values of execution time listed in the table are obtained on a laptop with CPU Intel(R) Core(TM) i3-3120M, 2.5 GHz and 8 Gb of RAM, using the single thread implementation available at https://github.com/imathsoft/MathSoftDevelopment. It is easy to notice that the dependency between the number of knots and the execution time is quite close to a linear one, which gives us an evidence that the complexity of the algorithm is close to $O(\|\Omega(h)\|)$, although the question of complexity was not investigated thoroughly and remains beyond the scope of the present paper. Potentially the implementation can be speeded up by parallelization of some subroutines.

6 Conclusions

The SI-method presented above can be considered as a particular implementation of a quite general idea about switching between ”straight” and ”inverse” problems when one of them becomes essentially more difficult in terms of numerical calculations than the other one. In the other words, the approach presented here can be quite easily modified and applied to ordinary differential equations of different types, by choosing different step functions $U(s)$ and $V(s)$. Moreover, the authors of the paper suggest that the idea of the method can be successfully applied to problems with partial differential equations and, potentially, to operator equations of general type.

The particular version of the SI-method presented above, is quite straightforward and efficient in terms of programming. One of its possible c++ implementations is available at GitHub[13] and can be used for solving problems other than the Troesch’s problem exploited in the present paper.

The results of numerical example based on the Troesch’s problem clearly show that the proposed implementation of the SI-method behaves very well as compared to the other approaches for solving the problem in terms of both accuracy and efficiency. It is worth mentioning that this is despite the fact that the SI-method is general and does not have anything in it which is designed specifically for the purpose of solving the Troesch’s problem (as it is in many other papers referenced in Section 5).

References

[1] Snyman J. A. Continuous and discontinuous numerical solutions to the troesch problem. *Journal of Computational and Applied Mathematics*, 5(3):171–175, 1979.

[13]https://github.com/imathsoft/MathSoftDevelopment
[2] E. L. Allgower and S. F. McCormick. Newton’s method with mesh refinements for numerical solution of nonlinear two-point boundary value problems. *Numerische Mathematik*, 29(3):237–260, 1978.

[3] Luigi Brugnano, Francesca Mazzia, and Donato Trigiante. Fifty years of stiffness. In *Recent advances in computational and applied mathematics*, pages 1–21. Springer, Dordrecht, 2011.

[4] J. R. Cash. Efficient numerical methods for the solution of stiff initial-value problems and differential algebraic equations. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 459(2032):797–815, 2003.

[5] Shih-Hsiang Chang. Numerical solution of troesch’s problem by simple shooting method. *Applied Mathematics and Computation*, 216(11):3303 – 3306, 2010.

[6] Shih-Hsiang Chang. A variational iteration method for solving troesch’s problem. *Journal of Computational and Applied Mathematics*, 234(10):3043 – 3047, 2010.

[7] Joseph E. Flaherty and R. E. O’Malley, Jr. The numerical solution of boundary value problems for stiff differential equations. *Math. Comput.*, 31(137):66–93, 1977.

[8] F. R. Gantmacher. *The theory of matrices. Vol. 1*. AMS Chelsea Publishing, Providence, RI, 1998. Translated from the Russian by K. A. Hirsch, Reprint of the 1959 translation.

[9] I. P. Gavrilyuk, M. Hermann, M. V. Kutniv, and V. L. Makarov. Difference schemes for nonlinear bvp’s using runge-kutta ivp-solvers. *Advances in Difference Equations*, 2006:Article ID 12167, 29 pages, 2006.

[10] Sung N. Ha. A nonlinear shooting method for two-point boundary value problems. *Computers and Mathematics with Applications*, 42(1011):1411 – 1420, 2001.

[11] E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. Stiff and differential-algebraic problems, Second revised edition, paperback.

[12] M. Hermann and D. Kaiser. Rwpm: a software package of shooting methods for nonlinear two-point boundary value problems. *Applied Numerical Mathematics*, 13(1):103 – 108, 1993.

[13] D.J Jones. Solution of troesch’s, and other, two point boundary value problems by shooting techniques. *Journal of Computational Physics*, 12(3):429 – 434, 1973.

[14] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.
[15] S.M Roberts and J.S Shipman. On the closed form solution of troesch’s problem. *Journal of Computational Physics*, 21(3):291 – 304, 1976.

[16] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*. Springer-Verlag, 2002.

[17] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[18] B.A Troesch. A simple approach to a sensitive two-point boundary value problem. *Journal of Computational Physics*, 21(3):279 – 290, 1976.

[19] Hector Vazquez-Leal, Yasir Khan, Guillermo Fernández-Anaya, Agustín Herrera-May, Arturo Sarmiento-Reyes, Uriel Filobello-Nino, Víctor-M. Jimenez-Fernández, and Domítilo Pereyra-Díaz. A general solution for Troesch’s problem. *Mathematical Problems in Engineering*, 2012.

[20] V.F. Zaitsev and A.D. Polyanin. *Handbook of Exact Solutions for Ordinary Differential Equations*. CRC Press, 2002.