NUMERICAL SOLUTIONS TO SOME ILL-POSED PROBLEMS

by

NGUYEN SI HOANG

B. A., Vietnam National University, Vietnam, 2002

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Mathematics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2011

Abstract

Several methods for a stable solution to the equation $F(u) = f$ have been developed. Here $F : H \to H$ is an operator in a Hilbert space H, and we assume that noisy data f_δ, $\|f_\delta - f\| \le \delta$, are given in place of the exact data f.

When F is a linear bounded operator, two versions of the Dynamical Systems Method (DSM) with stopping rules of Discrepancy Principle type are proposed and justified mathematically.

When F is a non-linear monotone operator, various versions of the DSM are studied. A Discrepancy Principle for solving the equation is formulated and justified. Several versions of the DSM for solving the equation are formulated. These methods consist of a Newton-type method, a gradient-type method, and a simple iteration method. *A priori* and *a posteriori* choices of stopping rules for these methods are proposed and justified. Convergence of the solutions, obtained by these methods, to the minimal norm solution to the equation $F(u) = f$ is proved. Iterative schemes with a posteriori choices of stopping rule corresponding to the proposed DSM are formulated. Convergence of these iterative schemes to a solution to the equation $F(u) = f$ is proved.

This dissertation consists of six chapters which are based on joint papers by the author and his advisor Prof. Alexander G. Ramm. These papers are published in different journals. The first two chapters deal with equations with linear and bounded operators and the last four chapters deal with non-linear equations with monotone operators.

NUMERICAL SOLUTIONS TO SOME ILL-POSED PROBLEMS

by

NGUYEN SI HOANG

B. A., Vietnam National University, Vietnam, 2002

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Mathematics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2011

Approved by:

Major Professor
Alexander G. Ramm

Copyright

Abstract

Several methods for a stable solution to the equation $F(u) = f$ have been developed. Here $F : H \to H$ is an operator in a Hilbert space H, and we assume that noisy data f_δ, $\|f_\delta - f\| \le \delta$, are given in place of the exact data f.

When F is a linear bounded operator, two versions of the Dynamical Systems Method (DSM) with stopping rules of Discrepancy Principle type are proposed and justified mathematically.

When F is a non-linear monotone operator, various versions of the DSM are studied. A Discrepancy Principle for solving the equation is formulated and justified. Several versions of the DSM for solving the equation are formulated. These methods consist of a Newton-type method, a gradient-type method, and a simple iteration method. *A priori* and *a posteriori* choices of stopping rules for these methods are proposed and justified. Convergence of the solutions, obtained by these methods, to the minimal norm solution to the equation $F(u) = f$ is proved. Iterative schemes with a posteriori choices of stopping rule corresponding to the proposed DSM are formulated. Convergence of these iterative schemes to a solution to the equation $F(u) = f$ is proved.

This dissertation consists of six chapters which are based on joint papers by the author and his advisor Prof. Alexander G. Ramm. These papers are published in different journals. The first two chapters deal with equations with linear and bounded operators and the last four chapters deal with non-linear equations with monotone operators.

Table of Contents

Acknowledgments

I would like to express my most sincere thanks to my advisor Prof. Alexander G. Ramm whose inspiring guidance and critical help brought me to the successful completion of this dissertation.

Dedication

to my family

Chapter 1

Dynamical systems gradient method for solving ill-conditioned linear algebraic systems.

Published in Acta Appl. Math., 111, N2, (2010), 189-204.

Dynamical systems gradient method for solving ill-conditioned linear algebraic systems

N. S. Hoang†* A. G. Ramm†‡

†Mathematics Department, Kansas State University,
Manhattan, KS 66506-2602, USA

Abstract

A version of the Dynamical Systems Method (DSM) for solving ill-conditioned linear algebraic systems is studied in this paper. An *a priori* and *a posteriori* stopping rules are justified. An algorithm for computing the solution using a spectral decomposition of the left-hand side matrix is proposed. Numerical results show that when a spectral decompositon of the left-hand side matrix is available or not computationally expensive to obtain the new method can be considered as an alternative to the Variational Regularization.

Keywords. Ill-conditioned linear algebraic systems , Dynamical Systems Method (DSM), Variational Regularization

MSC: 65F10; 65F22

1 Introduction

The Dynamical Systems Method (DSM) was systematically introduced and investigated in [19] as a general method for solving operator equations, linear and nonlinear, especially ill-posed operator equations (see also [20]-[23]). In several recent publications various versions of the DSM, proposed in [19], were shown to be as efficient and economical as variational regularization methods (see [4]-[10], [15]). This was demonstrated, for example, for the problems of solving ill-conditioned linear algebraic systems (cf. [2]), and stable numerical differentiation of noisy data (see [16], [17], [3]).

The aim of this paper is to formulate a version of the DSM gradient method for solving ill-posed linear equations and to demonstrate numerical efficiency of this method. There is a large literature

[*] Email: nguyenhs@math.ksu.edu

[‡] Corresponding author. Email: ramm@math.ksu.edu

on iterative regularization methods. These methods can be derived from a suitable version of the DSM by a discretization (see [19]). In the Gauss-Newton-type version of the DSM one has to invert some linear operator, which is an expensive procedure. The same is true for regularized Newton-type versions of the DSM and of their iterative counterparts. In contrast, the DSM gradient method we study in this paper *does not* require inversion of operators.

We want to solve equation

$$Au = f, \tag{1}$$

where A is a linear bounded operator in a Hilbert space H. We assume that (1) has a solution, possibly nonunique, and denote by y the unique minimal-norm solution to (1), $y \perp \mathcal{N} := \mathcal{N}(A) := \{u : Au = 0\}$, $Ay = f$. We assume that the range of A, $R(A)$, is not closed, so problem (1) is ill-posed. Let f_δ, $\|f - f_\delta\| \leq \delta$, be the noisy data. We want to construct a stable approximation of y, given $\{\delta, f_\delta, A\}$. There are many methods for doing this, see, e.g., [11], [12], [13], [19], [25], to mention a few books, where variational regularization, quasisolutions, quasiinversion, iterative regularization, and the DSM are studied.

The DSM version we study in this paper consists of solving the Cauchy problem

$$\dot{u}(t) = -A^*(Au(t) - f), \quad u(0) = u_0, \quad u_0 \perp N, \quad \dot{u} := \frac{du}{dt}, \tag{2}$$

where A^* is the adjoint to operator A, and proving the existence of the limit $\lim_{t\to\infty} u(t) = u(\infty)$, and the relation $u(\infty) = y$, i.e.,

$$\lim_{t\to\infty} \|u(t) - y\| = 0. \tag{3}$$

If the noisy data f_δ are given, then we solve the problem

$$\dot{u}_\delta(t) = -A^*(Au_\delta(t) - f_\delta), \quad u_\delta(0) = u_0, \tag{4}$$

and prove that, for a suitable stopping time t_δ, and $u_\delta := u_\delta(t_\delta)$, one has

$$\lim_{\delta\to 0} \|u_\delta - y\| = 0. \tag{5}$$

In Section 2 these results are formulated precisely and recipes for choosing t_δ are proposed.

The novel results in this paper include the proof of the discrepancy principle (Theorem 3), an efficient method for computing $u_\delta(t_\delta)$ (Section 3), and an a priori stopping rule (Theorem 2).

Our presentation is essentially self-contained.

Our results show that the DSM provides a method for solving a wide range of ill-posed problems, which is quite competitive with other methods, currently used. The DSM yields sometimes better accuracy and stability than variational regularization, and is simple in computational implementation.

2 Results

Suppose $A : H \to H$ is a linear bounded operator in a Hilbert space H. Assume that equation

$$Au = f \tag{6}$$

has a solution not necessarily unique. Denote by y the unique minimal-norm solution i.e., $y \perp \mathcal{N} := \mathcal{N}(A)$. Consider the following Dynamical Systems Method (DSM)

$$\begin{aligned} \dot{u} &= -A^*(Au - f), \\ u(0) &= u_0, \end{aligned} \tag{7}$$

where $u_0 \perp \mathcal{N}$ is arbitrary. Denote $T := A^*A$, $Q := AA^*$. The unique solution to (7) is

$$u(t) = e^{-tT}u_0 + e^{-tT}\int_0^t e^{sT}dsA^*f.$$

Let us show that any ill-posed linear equation (6) with exact data can be solved by the DSM.

2.1 Exact data

Theorem 1 *Suppose $u_0 \perp \mathcal{N}$. Then problem (7) has a unique solution defined on $[0, \infty)$, and $u(\infty) = y$, where $u(\infty) = \lim_{t\to\infty} u(t)$.*

Proof. Denote $w := u(t) - y$, $w_0 = w(0)$. Note that $w_0 \perp \mathcal{N}$. One has

$$\dot{w} = -Tw, \quad T = A^*A. \tag{8}$$

The unique solution to (8) is $w = e^{-tT}w_0$. Thus,

$$\|w\|^2 = \int_0^{\|T\|} e^{-2t\lambda}d\langle E_\lambda w_0, w_0\rangle.$$

where $\langle u, v\rangle$ is the inner product in H, and E_λ is the resolution of the identity of the selfadjoint operator T. Thus,

$$\|w(\infty)\|^2 = \lim_{t\to\infty}\int_0^{\|T\|} e^{-2t\lambda}d\langle E_\lambda w_0, w_0\rangle = \|P_{\mathcal{N}}w_0\|^2 = 0,$$

where $P_{\mathcal{N}} = E_0 - E_{-0}$ is the orthogonal projector onto $\mathcal{N}$. Theorem 1 is proved. □

2.2 Noisy data f_δ

Let us solve stably equation (6) assuming that f is not known, but f_δ, the noisy data, are known, where $\|f_\delta - f\| \le \delta$. Consider the following DSM

$$\dot{u}_\delta = -A^*(Au_\delta - f_\delta), \quad u_\delta(0) = u_0.$$

Denote

$$w_\delta := u_\delta - y, \quad T := A^*A, \quad w_\delta(0) = w_0 := u_0 - y \in \mathcal{N}^\perp.$$

Let us prove the following result:

Theorem 2 *If* $\lim_{\delta\to 0} t_\delta = \infty$, $\lim_{\delta\to 0} t_\delta\delta = 0$, *and* $w_0 \perp \mathcal{N}$, *then*

$$\lim_{\delta\to 0} \|w_\delta(t_\delta)\| = 0.$$

Proof. One has

$$\dot{w}_\delta = -Tw_\delta + \eta_\delta, \quad \eta_\delta = A^*(f_\delta - f), \quad \|\eta_\delta\| \le \|A\|\delta. \tag{9}$$

The unique solution of equation (9) is

$$w_\delta(t) = e^{-tT}w_\delta(0) + \int_0^t e^{-(t-s)T}\eta_\delta ds.$$

Let us show that $\lim_{t\to\infty} \|w_\delta(t)\| = 0$. One has

$$\lim_{t\to\infty} \|w_\delta(t)\| \le \lim_{t\to\infty} \|e^{-tT}w_\delta(0)\| + \lim_{t\to\infty} \left\| \int_0^t e^{-(t-s)T}\eta_\delta ds \right\|. \tag{10}$$

One uses the spectral theorem and gets:

$$\begin{aligned} \int_0^t e^{-(t-s)T} ds\eta_\delta &= \int_0^t \int_0^{\|T\|} dE_\lambda \eta_\delta e^{-(t-s)\lambda} ds \\ &= \int_0^{\|T\|} e^{-t\lambda} \frac{e^{t\lambda} - 1}{\lambda} dE_\lambda \eta_\delta = \int_0^{\|T\|} \frac{1 - e^{-t\lambda}}{\lambda} dE_\lambda \eta_\delta. \end{aligned} \tag{11}$$

Note that

$$0 \le \frac{1 - e^{-t\lambda}}{\lambda} \le t, \quad \forall \lambda > 0, t \ge 0, \tag{12}$$

since $1 - x \le e^{-x}$ for $x \ge 0$. From (11) and (12), one obtains

$$\begin{aligned} \left\| \int_0^t e^{-(t-s)T} ds\eta_\delta \right\|^2 &= \int_0^{\|T\|} \left|\frac{1 - e^{-t\lambda}}{\lambda}\right|^2 d\langle E_\lambda \eta_\delta, \eta_\delta\rangle \\ &\le t^2 \int^{\|T\|} d\langle E_\lambda \eta_\delta, \eta_\delta\rangle \\ &= t^2 \|\eta_\delta\|^2. \end{aligned} \tag{13}$$

Since $\|\eta_\delta\| \le \|A\|\delta$, from (10) and (13), one gets

$$\lim_{\delta\to 0}\|w_\delta(t_\delta)\| \le \lim_{\delta\to 0}\Big(\|e^{-t_\delta T}w_\delta(0)\| + t_\delta\delta\|A\|\Big) = 0.$$

Here we have used the relation:

$$\lim_{\delta\to 0}\|e^{-t_\delta T}w_\delta(0)\| = \|P_{\mathcal{N}}w_0\| = 0,$$

and the last equality holds because $w_0 \in \mathcal{N}^\perp$. Theorem 2 is proved. □

From Theorem 2, it follows that the relation $t_\delta = \frac{C}{\delta^\gamma}$, $\gamma = \text{const}$, $\gamma \in (0,1)$ and $C > 0$ is a constant, can be used as an *a priori* stopping rule, i.e., for such t_δ one has

$$\lim_{\delta\to 0}\|u_\delta(t_\delta) - y\| = 0. \tag{14}$$

2.3 Discrepancy principle

Let us consider equation (6) with noisy data f_δ, and a DSM of the form

$$\dot{u}_\delta = -A^*Au_\delta + A^*f_\delta, \quad u_\delta(0) = u_0. \tag{15}$$

for solving this equation. Equation (15) has been used in Section 2.2. Recall that y denotes the minimal-norm solution of equation (6).

Theorem 3 *Assume that $\|Au_0 - f_\delta\| > C\delta$. The solution t_δ to the equation*

$$h(t) := \|Au_\delta(t) - f_\delta\| = C\delta, \quad 1 < C = \mathit{const}, \tag{16}$$

does exist, is unique, and

$$\lim_{\delta\to 0}\|u_\delta(t_\delta) - y\| = 0. \tag{17}$$

Proof. Denote

$$v_\delta(t) := Au_\delta(t) - f_\delta, \qquad T := A^*A, \qquad Q = AA^*$$

and

$$w_\delta(t) := u_\delta(t) - y, \qquad w_0 := u_0 - y.$$

One has

$$\begin{aligned}\frac{d}{dt}\|v_\delta(t)\|^2 &= 2\,\mathrm{Re}\langle A\dot{u}_\delta(t), Au_\delta(t) - f_\delta\rangle \\ &= 2\,\mathrm{Re}\langle A[-A^*(Au_\delta(t) - f_\delta)], Au_\delta(t) - f_\delta\rangle \\ &= -2\|A^*v_\delta(t)\|^2 \le 0.\end{aligned} \tag{18}$$

Thus, $\|v_\delta(t)\|$ is a nonincreasing function. Let us prove that equation (16) has a solution for $C > 1$. Recall the known commutation formulas:

$$e^{-sT}A^* = A^*e^{-sQ},\ Ae^{-sT} = e^{-tQ}A.$$

Using these formulas and the representation

$$u_\delta(t) = e^{-tT}u_0 + \int_0^t e^{-(t-s)T}A^*f_\delta ds,$$

one gets:

$$\begin{aligned}
v_\delta(t) &= Au_\delta(t) - f_\delta \\
&= Ae^{-tT}u_0 + A\int_0^t e^{-(t-s)T}A^*f_\delta ds - f_\delta \\
&= e^{-tQ}Au_0 + e^{-tQ}\int_0^t e^{sQ}dsQf_\delta - f_\delta \\
&= e^{-tQ}A(u_0 - y) + e^{-tQ}f + e^{-tQ}(e^{tQ} - I)f_\delta - f_\delta \\
&= e^{-tQ}Aw_0 + e^{-tQ}f - e^{-tQ}f_\delta.
\end{aligned} \tag{19}$$

Note that

$$\lim_{t\to\infty} e^{-tQ}Aw_0 = \lim_{t\to\infty} Ae^{-tT}w_0 = AP_{\mathcal{N}}w_0 = 0.$$

Here the continuity of A, and the following relations

$$\lim_{t\to\infty} e^{-tT}w_0 = \lim_{t\to\infty}\int_0^{\|T\|} e^{-st}dE_s w_0 = (E_0 - E_{-0})w_0 = P_{\mathcal{N}}w_0,$$

were used. Therefore,

$$\lim_{t\to\infty}\|v_\delta(t)\| = \lim_{t\to\infty}\|e^{-tQ}(f - f_\delta)\| \le \|f - f_\delta\| \le \delta, \tag{20}$$

because $\|e^{-tQ}\| \le 1$. The function $h(t)$ is continuous on $[0,\infty)$, $h(0) = \|Au_0 - f_\delta\| > C\delta$, $h(\infty) \le \delta$. Thus, equation (16) must have a solution t_δ.

Let us prove the uniqueness of t_δ. Without loss of generality we can assume that there exists $t_1 > t_\delta$ such that $\|Au_\delta(t_1) - f_\delta\| = C\delta$. Since $\|v_\delta(t)\|$ is nonincreasing and $\|v_\delta(t_\delta)\| = \|v_\delta(t_1)\|$, one has

$$\|v_\delta(t)\| = \|v_\delta(t_\delta)\|, \quad \forall t \in [t_\delta, t_1].$$

Thus,

$$\frac{d}{dt}\|v_\delta(t)\|^2 = 0, \quad \forall t \in (t_\delta, t_1). \tag{21}$$

Using (18) and (21) one obtains

$$A^*v_\delta(t) = A^*(Au_\delta(t) - f_\delta) = 0, \quad \forall t \in [t_\delta, t_1].$$

This and (15) imply

$$\dot{u}_\delta(t) = 0, \quad \forall t \in (t_\delta, t_1). \tag{22}$$

One has

$$\begin{aligned} \dot{u}_\delta(t) &= -Tu_\delta(t) + A^*f_\delta \\ &= -T\bigg(e^{-tT}u_0 + \int_0^t e^{-(t-s)T}A^*f_\delta ds\bigg) + A^*f_\delta \\ &= -Te^{-tT}u_0 - (I - e^{-tT})A^*f_\delta + A^*f_\delta \\ &= -e^{-tT}(Tu_0 - A^*f_\delta). \end{aligned} \tag{23}$$

From (23) and (22), one gets $Tu_0 - A^*f = e^{tT}e^{-tT}(Tu_0 - A^*f) = 0$. Note that the operator e^{tT} is an isomorphism for any fixed t since T is selfadjoint and bounded. Since $Tu_0 - A^*f = 0$, by (23) one has $\dot{u}_\delta(t) = 0$, $u_\delta(t) = u_\delta(0)$, $\forall t \geq 0$. Consequently,

$$C\delta < \|Au_\delta(0) - f_\delta\| = \|Au_\delta(t_\delta) - f_\delta\| = C\delta.$$

This is a contradiction which proves the uniqueness of t_δ.

Let us prove (17). First, we have the following estimate:

$$\begin{aligned} \|Au(t_\delta) - f\| &\leq \|Au(t_\delta) - Au_\delta(t_\delta)\| + \|Au_\delta(t_\delta) - f_\delta\| + \|f_\delta - f\| \\ &\leq \bigg\|e^{-t_\delta Q}\int_0^{t_\delta} e^{sQ}Qds\bigg\|\|f_\delta - f\| + C\delta + \delta. \end{aligned} \tag{24}$$

Let us use the inequality:

$$\big\|e^{-t_\delta Q}\int_0^{t_\delta} e^{sQ}Qds\big\| = \|I - e^{-t_\delta Q}\| \leq 2,$$

and conclude from (24), that

$$\lim_{\delta\to 0} \|Au(t_\delta) - f\| = 0. \tag{25}$$

Secondly, we claim that

$$\lim_{\delta\to 0} t_\delta = \infty. \tag{26}$$

Assume the contrary. Then there exist $t_0 > 0$ and a sequence $(t_{\delta_n})_{n=1}^\infty$, $t_{\delta_n} < t_0$, such that

$$\lim_{n\to\infty} \|Au(t_{\delta_n}) - f\| = 0. \tag{27}$$

Analogously to (18), one proves that

$$\frac{d\|v\|^2}{dt} \leq 0,$$

where $v(t) := Au(t) - f$. Thus, $\|v(t)\|$ is nonincreasing. This and (27) imply the relation $\|v(t_0)\| = \|Au(t_0) - f\| = 0$. Thus,

$$0 = v(t_0) = e^{-t_0 Q} A(u_0 - y).$$

This implies $A(u_0 - y) = e^{t_0 Q} e^{-t_0 Q} A(u_0 - y) = 0$, so $u_0 - y \in \mathcal{N}$. Since $u_0 - y \in \mathcal{N}^\perp$, it follows that $u_0 = y$. This is a contradiction because

$$C\delta \leq \|Au_0 - f_\delta\| = \|f - f_\delta\| \leq \delta, \quad 1 < C.$$

Thus, $\lim_{\delta\to 0} t_\delta = \infty$.

Let us continue the proof of (17). Let $w_\delta(t) := u_\delta(t) - y$. We claim that $\|w_\delta(t)\|$ is nonincreasing on $[0, t_\delta]$. One has

$$\begin{aligned}
\frac{d}{dt}\|w_\delta(t)\|^2 &= 2\operatorname{Re}\langle \dot{u_\delta}(t), u_\delta(t) - y\rangle \\
&= 2\operatorname{Re}\langle -A^*(Au_\delta(t) - f_\delta), u_\delta(t) - y\rangle \\
&= -2\operatorname{Re}\langle Au_\delta(t) - f_\delta, Au_\delta(t) - f_\delta + f_\delta - Ay\rangle \\
&\leq -2\|Au_\delta(t) - f_\delta\|\Big(\|Au_\delta(t) - f_\delta\| - \|f_\delta - f\|\Big) \\
&\leq 0.
\end{aligned}$$

Here we have used the inequalities:

$$\|Au_\delta(t) - f_\delta\| \geq C\delta > \|f_\delta - Ay\| = \delta, \quad \forall t \in [0, t_\delta].$$

Let $\epsilon > 0$ be arbitrary small. Since $\lim_{t\to\infty} u(t) = y$, there exists $t_0 > 0$, independent of δ, such that

$$\|u(t_0) - y\| \leq \frac{\epsilon}{2}. \tag{28}$$

Since $\lim_{\delta\to 0} t_\delta = \infty$ (see (26)), there exists $\delta_0 > 0$ such that $t_\delta > t_0$, $\forall \delta \in (0, \delta_0)$. Since $\|w_\delta(t)\|$ is nonincreasing on $[0, t_\delta]$ one has

$$\|w_\delta(t_\delta)\| \leq \|w_\delta(t_0)\| \leq \|u_\delta(t_0) - u(t_0)\| + \|u(t_0) - y\|, \quad \forall \delta \in (0, \delta_0). \tag{29}$$

Note that

$$\|u_\delta(t_0) - u(t_0)\| = \|e^{-t_0 T}\int_0^{t_0} e^{sT} ds A^*(f_\delta - f)\| \leq \|e^{-t_0 T}\int_0^{t_0} e^{sT} ds A^*\|\delta. \tag{30}$$

Since $e^{-t_0 T}\int_0^{t_0} e^{sT}dsA^*$ is a bounded operator for any fixed t_0, one concludes from (30) that $\lim_{\delta\to 0}\|u_\delta(t_0)-u(t_0)\|=0$. Hence, there exists $\delta_1\in(0,\delta_0)$ such that

$$\|u_\delta(t_0)-u(t_0)\|\le\frac{\epsilon}{2},\quad \forall\delta\in(0,\delta_1). \tag{31}$$

From (28)–(31), one obtains

$$\|u_\delta(t_\delta)-y\|=\|w_\delta(t_\delta)\|\le\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon,\quad \forall\delta\in(0,\delta_1).$$

This means that $\lim_{\delta\to 0}u_\delta(t_\delta)=y$. Theorem 3 is proved. □

3 Computing $u_\delta(t_\delta)$

3.1 Systems with known spectral decomposition

One way to solve the Cauchy problem (15) is to use explicit Euler or Runge-Kutta methods with a constant or adaptive stepsize h. However, stepsize h for solving (15) by explicit numerical methods is often smaller than 1 and the stopping time $t_\delta=nh$ may be large. Therefore, the computation time, characterized by the number of iterations n, for this approach may be large. This fact is also reported in [2], where one of the most efficient numerical methods for solving ordinary differential equations (ODEs), the DOPRI45 (see [1]), is used for solving a Cauchy problem in a DSM. Indeed, the use of explicit Euler method leads to a Landweber iteration which is known for slow convergence. Thus, it may be computationally expensive to compute $u_\delta(t_\delta)$ by numerical methods for ODEs.

However, when A in (15) is a matrix and a decomposition $A=USV^*$, where U and V are unitary matrices and S is a diagonal matrix, is known, it is possible to compute $u_\delta(t_\delta)$ at a speed comparable to other methods such as the variational regularization (VR) as it will be shown below.

We have

$$u_\delta(t)=e^{-tT}u_0+e^{-tT}\int_0^t e^{sT}dsA^*f_\delta,\quad T:=A^*A. \tag{32}$$

Suppose that a decomposition

$$A=USV^*, \tag{33}$$

where U and V are unitary matrices and S is a diagonal matrix is known. These matrices possibly contain complex entries. Thus, $T=A^*A=V\bar{S}SV^*$ and $e^T=e^{V\bar{S}SV^*}$. Using the formula $e^{V\bar{S}SV^*}=Ve^{\bar{S}S}V^*$, which is valid if V is unitary and $\bar{S}S$ is diagonal, equation (32) can be rewritten as

$$u_\delta(t)=Ve^{-t\bar{S}S}V^*u_0+V\int_0^t e^{(s-t)\bar{S}S}ds\bar{S}U^*f_\delta. \tag{34}$$

Here, the overbar stands for complex conjugation. Choose $u_0 = 0$. Then

$$u_\delta(t) = V \int_0^t e^{(s-t)\bar{S}S} ds \bar{S} h_\delta, \quad h_\delta := U^* f_\delta. \tag{35}$$

Let us assume that

$$\delta < \|f\|. \tag{36}$$

This is a natural assumption. Let us check that

$$A^* f_\delta \neq 0. \tag{37}$$

Indeed, if $A^* f_\delta = 0$, then one gets

$$\langle f_\delta, f \rangle = \langle f_\delta, Ay \rangle = \langle A^* f_\delta, y \rangle = 0. \tag{38}$$

This implies

$$\delta^2 \geq \|f - f_\delta\|^2 = \|f\|^2 + \|f_\delta\|^2 > \delta^2. \tag{39}$$

This contradiction implies (37).

The stopping time t_δ we choose by the following discrepancy principle:

$$\|Au_\delta(t_\delta) - f_\delta\| = \left\| \int_0^{t_\delta} e^{(s-t_\delta)\bar{S}S} ds \bar{S} S h_\delta - h_\delta \right\| = \|e^{-t_\delta \bar{S}S} h_\delta\| = C\delta.$$

where $1 < C$.

Let us find t_δ from the equation

$$\phi(t) := \psi(t) - C\delta = 0, \qquad \psi(t) := \|e^{-t\bar{S}S} h_\delta\|. \tag{40}$$

The existence and uniqueness of the solution t_δ to equation (40) follow from Theorem 3.

We claim that *equation* (40) *can be solved by using Newton's iteration* (48) *for any initial value* t_0 *such that* $\phi(t_0) > 0$.

Let us prove this claim. It is sufficient to prove that $\phi(t)$ is a monotone strictly convex function. This is proved below.

Without loss of generality, we can assume that h_δ (see (40)) is a vector with real components. The proof remained essentially the same for h_δ with complex components.

First, we claim that

$$\sqrt{\bar{S}S} h_\delta \neq 0, \qquad \text{and} \qquad \|\sqrt{\bar{S}S} e^{-t\bar{S}S} h_\delta\| \neq 0, \tag{41}$$

so $\psi(t) > 0$.

Indeed, since $e^{-t\bar{S}S}$ is an isomorphism and $e^{-t\bar{S}S}$ commutes with $\sqrt{\bar{S}S}$ one concludes that $\|\sqrt{\bar{S}S}e^{-t\bar{S}S}h_\delta\| = 0$ iff $\sqrt{\bar{S}S}h_\delta = 0$. If $\sqrt{\bar{S}S}h_\delta = 0$ then $\bar{S}h_\delta = 0$, and, therefore,

$$0 = \bar{S}h_\delta = \bar{S}U^* f_\delta = V^*V\bar{S}U^*f_\delta = V^*A^*f_\delta. \tag{42}$$

Since V is a unitary matrix, it follows from (42) that $A^*f_\delta = 0$. This contradicts to relation (37).

Let us now prove that ϕ monotonically decays and is strictly convex. Then our claim will be proved.

One has

$$\frac{d}{dt}\langle e^{-t\bar{S}S}h_\delta, e^{-t\bar{S}S}h_\delta\rangle = -2\langle e^{-t\bar{S}S}h_\delta, \bar{S}Se^{-t\bar{S}S}h_\delta\rangle.$$

Thus,

$$\dot{\psi}(t) = \frac{d}{dt}\|e^{-t\bar{S}S}h_\delta\| = \frac{\frac{d}{dt}\|e^{-t\bar{S}S}h_\delta\|^2}{2\|e^{-t\bar{S}S}h_\delta\|} = -\frac{\langle e^{-t\bar{S}S}h_\delta, \bar{S}Se^{-t\bar{S}S}h_\delta\rangle}{\|e^{-t\bar{S}S}h_\delta\|}. \tag{43}$$

Equation (43), relation (41), and the fact that $\langle e^{-t\bar{S}S}h_\delta, \bar{S}Se^{-t\bar{S}S}h_\delta\rangle = \|\sqrt{\bar{S}S}e^{-t\bar{S}S}h_\delta\|^2$ imply

$$\dot{\psi}(t) < 0. \tag{44}$$

From equation (43) and the definition of ψ in (40), one gets

$$\psi(t)\dot{\psi}(t) = -\langle e^{-t\bar{S}S}h_\delta, \bar{S}Se^{-t\bar{S}S}h_\delta\rangle \tag{45}$$

Differentiating equation (45) with respect to t, one obtains

$$\begin{aligned}\psi(t)\ddot{\psi}(t) + \dot{\psi}^2(t) &= \langle \bar{S}Se^{-t\bar{S}S}h_\delta, \bar{S}Se^{-t\bar{S}S}h_\delta\rangle + \langle e^{-t\bar{S}S}h_\delta, \bar{S}S\bar{S}Se^{-t\bar{S}S}h_\delta\rangle \\ &= 2\|\bar{S}Se^{-t\bar{S}S}h_\delta\|^2.\end{aligned}$$

This equation and equation (43) imply

$$\psi(t)\ddot{\psi}(t) = 2\|\bar{S}Se^{-t\bar{S}S}h_\delta\|^2 - \frac{\langle e^{-t\bar{S}S}h_\delta, \bar{S}Se^{-t\bar{S}S}h_\delta\rangle^2}{\|e^{-t\bar{S}S}h_\delta\|^2} \geq \|\bar{S}Se^{-t\bar{S}S}h_\delta\|^2 > 0. \tag{46}$$

Here the inequality: $\langle e^{-t\bar{S}S}h_\delta, \bar{S}Se^{-t\bar{S}S}h_\delta\rangle \leq \|e^{-t\bar{S}S}h_\delta\|\|\bar{S}Se^{-t\bar{S}S}h_\delta\|$ was used. Since $\psi > 0$, inequality (46) implies

$$\ddot{\psi}(t) > 0. \tag{47}$$

It follows from inequalities (44) and (47) that $\phi(t)$ is a strictly convex and decreasing function on $(0, \infty)$. Therefore, t_δ can be found by Newton's iterations:

$$\begin{aligned}t_{n+1} &= t_n - \frac{\phi(t_n)}{\dot{\phi}(t_n)} \\ &= t_n + \frac{\|e^{-t_n\bar{S}S}h_\delta\| - C\delta}{\langle \bar{S}Se^{-t_n\bar{S}S}h_\delta, e^{-t_n\bar{S}S}h_\delta\rangle}\|e^{-t_n\bar{S}S}h_\delta\|, \quad n = 0, 1, ...,\end{aligned} \tag{48}$$

for any initial guess t_0 of t_δ such that $\phi(t_0) > 0$. Once t_δ is found, the solution $u_\delta(t_\delta)$ is computed by (35).

Remark 1 In the decomposition $A = VSU^*$ we do not assume that U, V and S are matrices with real entries. The singular value decomposition (SVD) is a particular case of this decomposition.

It is computationally expensive to get the SVD of a matrix in general. However, there are many problems in which the decomposition (33) can be computed fast using the fast Fourier transform (FFT). Examples include image restoration problems with circulant block matrices (see [14]) and deconvolution problems. (see Section 4.2).

3.2 On the choice of t_0

Let us discuss a strategy for choosing the initial value t_0 in Newton's iterations for finding t_δ. We choose t_0 satisfying condition:

$$0 < \phi(t_0) = \|e^{-t_0\bar{S}S}h_\delta\| - \delta \leq \delta \tag{49}$$

by the following strategy

1. Choose $t_0 := 10\frac{\|h_\delta\|}{\delta}$ as an initial guess for t_0.

2. Compute $\phi(t_0)$. If t_0 satisfying (49) we are done. Otherwise, we go to step 3.

3. If $\phi(t_0) < 0$ and the inequality $\phi(t_0) > \delta$ has not occurred in iteration, we replace t_0 by $\frac{t_0}{10}$ and go back to step 2. If $\phi(t_0) < 0$ and the inequality $\phi(t_0) > \delta$ has occurred in iteration, we replace t_0 by $\frac{t_0}{3}$ and go back to step 2. If $\phi(t_0) > \delta$, we go to step 4.

4. If $\phi(t_0) > \delta$ and the inequality $\phi(t_0) < 0$ has not occured in iterations, we replace t_0 by $3t_0$ and go back to step 2. If the inequality $\phi(t_0) < 0$ has occured in some iteration before, we stop the iteration and use t_0 as an initial guess in Newton's iterations for finding t_δ.

4 Numerical experiments

In this section results of some numerical experiments with ill-conditioned linear algebraic systems are reported. In all the experiments, by DSMG we denote the version of the DSM described in this paper, by VR we denote the Variational Regularization, implemented using the discrepancy principle, and by DSM-[2] we denote the method developed in [2].

4.1 A linear algebraic system for the computation of second derivatives

Let us do some numerical experiments with linear algebraic systems arising in a numerical experiment of computing the second derivative of a noisy function.

The problem is reduced to an integral equation of the first kind. A linear algebraic system is obtained by a discretization of the integral equation whose kernel K is Green's function

$$K(s,t) = \begin{cases} s(t-1), & \text{if} \quad s < t \\ t(s-1), & \text{if} \quad s \geq t \end{cases}.$$

Here $s, t \in [0,1]$. Using A_N from [2], we do some numerical experiments for solving u_N from the linear algebraic system $A_N u_N = b_{N,\delta}$. In the experiments the exact right-hand side is computed by the formula $b_N = A_N u_N$ when u_N is given. In this test, u_N is computed by

$$u_N := \big(u(t_{N,1}), u(t_{N,2}), ..., u(t_{N,N})\big)^T, \qquad t_{N,i} := \frac{i}{N}, \quad i = 1, ..., N,$$

where $u(t)$ is a given function. We use $N = 10, 20, ..., 100$ and $b_{N,\delta} = b_N + e_N$, where e_N is a random vector whose coordinates are independent, normally distributed, with mean 0 and variance 1, and scaled so that $\|e_N\| = \delta_{rel}\|b_N\|$. This linear algebraic system is mildly ill-posed: the condition number of A_{100} is 1.2158×10^4.

In Figure 1, the difference between the exaction and solution obtained by the DSMG, VR and DSM-[2] are plotted. In these experiments, we used $N = 100$ and $u(t) = \sin(\pi t)$ with $\delta_{rel} = 0.05$ and $\delta_{rel} = 0.01$. Figure 1 shows that the results obtained by the VR and the DSM-[2] are very close to each other. The results obtained by the DSMG are much better than those by the DSM-[2] and by the VR.

Table 1 presents numerical results when N varies from 10 to 100, $u(t) = \sin(2\pi t)$, and $t \in [0,1]$. In this experiment the DSMG yields more accurate solutions than the DSM-[2] and the VR. The DSMG in this experiment takes more iterations than the DSM-[2] and the VR to get a solution.

In this experiment the DSMG is implemented using the SVD of A obtained by the function *svd* in Matlab. As already mentioned, the SVD is a special case of the spectral decomposition (33). It is expensive to compute the SVD, in general. However, there are practically important problems where the spectral decomposition (33) can be computed fast (see Section 4.2 below). These problems consist of deconvolution problems using the Fast Fourier Transform (FFTs).

The conclusion from this experiment is: the DSMG may yield results with much better accuracy than the VR and DSM-[2]. Numerical experiments for various $u(t)$ show that the DSMG competes favorably with the VR and the DSM-[2].

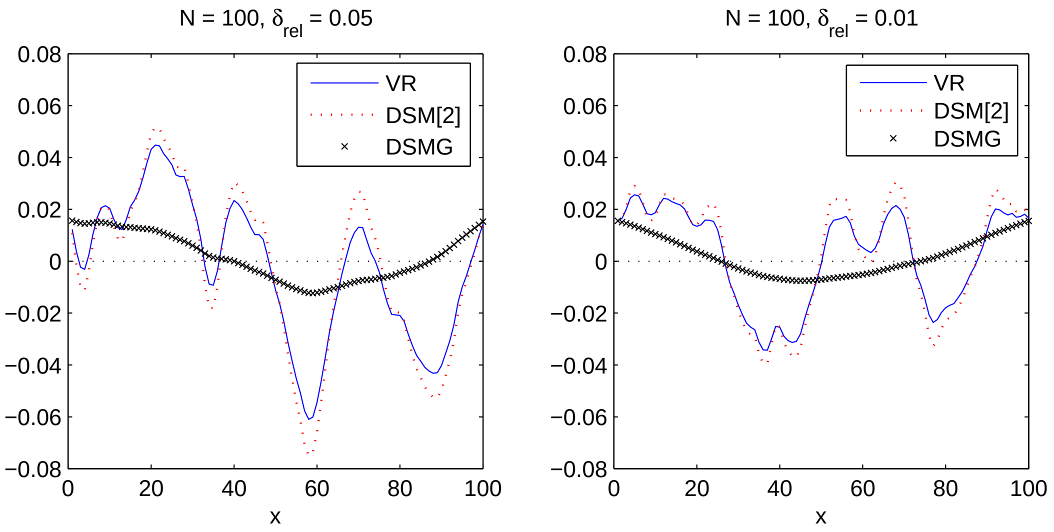

Figure 1: *Plots of differences between the exact solution and solutions obtained by the DSMG, VR and DSM-[2].*

4.2 An application to image restoration

The image degradation process can be modeled by the following equation:

$$g_\delta = g + w, \quad g = h * f, \quad \|w\| \le \delta, \tag{50}$$

where h represents a convolution function that models the blurring that many imaging systems introduce. For example, camera defocus, motion blur, imperfections of the lenses, all these phenomenon can be modeled by choosing a suitable h. The functions g_δ, f, and w are the observed image, the original signal, and the noise, respectively. The noise w can be due to the electronics used (thermal and shot noise), the recording medium (film grain), or the imaging process (photon noise).

In practice g, h and f in equation (50) are often given as functions of a discrete argument and equation (50) can be written in this case as

$$g_{\delta,i} = g_i + w_i = \sum_{j=-\infty}^{\infty} f_j h_{i-j} + w_i, \quad i \in \mathbb{Z}. \tag{51}$$

Note that one (or both) signals f_j and h_j have compact support (finite length). Suppose that signal f is periodic with period N, i.e., $f_{i+N} = f_i$, and $h_j = 0$ for $j < 0$ and $j \ge N$. Assume that f is represented by a sequence $f_0, ..., f_{N-1}$ and h is represented by $h_0, ..., h_{N-1}$. Then the convolution

Table 1: Numerical results for computing second derivatives with $\delta_{rel} = 0.01$.

	DSM		DSM-[2]		VR	
N	n_{iter}	$\frac{\lVert u_\delta - y\rVert_2}{\lVert y\rVert_2}$	n_{linsol}	$\frac{\lVert u_\delta - y\rVert_2}{\lVert y\rVert_2}$	n_{linsol}	$\frac{\lVert u_\delta - y\rVert_2}{\lVert y\rVert_2}$
20	9	0.0973	3	0.1130	6	0.1079
30	5	0.0831	4	0.1316	6	0.1160
40	7	0.0488	4	0.1150	6	0.1045
50	9	0.0614	4	0.1415	6	0.1063
60	6	0.0419	4	0.0919	6	0.0817
70	9	0.0513	4	0.0961	6	0.0842
80	6	0.0418	4	0.1225	6	0.0981
90	7	0.0287	4	0.0919	7	0.0840
100	7	0.0248	5	0.0778	7	0.0553

$h * f$ is periodic signal g with period N, and the elements of g are defined as

$$g_i = \sum_{j=0}^{N-1} h_j f_{(i-j)\, mod\, N}, \quad i = 0, 1, ..., N-1. \tag{52}$$

Here $(i-j)\, mod\, N$ is $i - j$ modulo N. The discrete Fourier transform (DFT) of g is defined as the sequence

$$\hat{g}_k := \sum_{j=0}^{N-1} g_j e^{-i2\pi jk/N}, \qquad k = 0, 1, ..., N-1.$$

Denote $\hat{g} = (\hat{g}_0, ..., \hat{g}_{N-1})^T$. Then equation (52) implies

$$\hat{g} = \hat{f}\hat{h}, \qquad \hat{f}\hat{h} := (\hat{f}_0\hat{h}_0, \hat{f}_1\hat{h}_1, ..., \hat{f}_{N-1}\hat{h}_{N-1})^T. \tag{53}$$

Let $a = (a_0, ..., a_{N-1})^T$ and $\text{diag}(a)$ denote a diagonal matrix whose diagonal is $(a_0, ..., a_{N-1})$ and other entries are zeros. Then equation (53) can be rewritten as

$$\hat{g} = A\hat{f}, \qquad A := \text{diag}(\hat{h}). \tag{54}$$

Since A is of the form (33) with $U = V = I$ and $S = \text{diag}(\hat{h})$, one can use the DSMG method to solve equation (54) stably for $\hat{f}$.

The image restoration test problem we use is taken from [14]. This test problem was developed at the US Air Force Phillips Laboratory, Lasers and Imaging Directorate, Kirtland Air Force Base, New Mexico. The original and blurred images have 256 × 256 pixels, and are shown in Figure 2. These data has been widely used in the literature for testing image restoration algorithms.

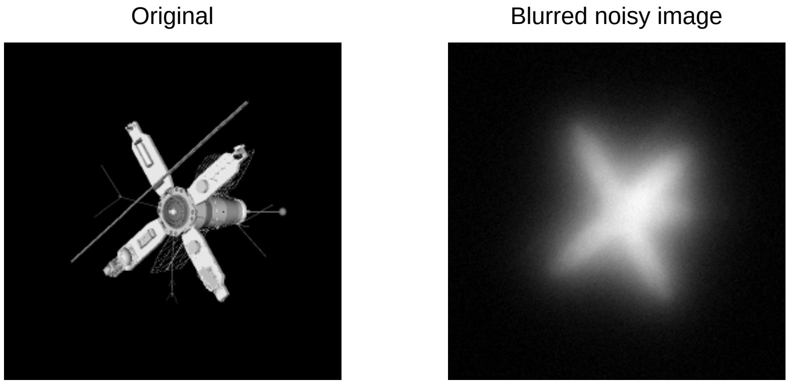

Figure 2: *Original and blurred noisy images.*

Figure 3 plots the regularized images by the VR and the DSMG when $\delta_{rel} = 0.01$. Again, with an input value for δ_{rel}, the observed blurred noisy images is computed by

$$g_\delta = g + \delta_{rel} \frac{\|g\|}{\|err\|} err,$$

where err is a vector with random entries normally distributed with mean 0 and variance 1. In this experiment, it took 5 and 8 iterations for the DSMG and the VR, respectively, to yield numerical results. From Figure 3 one concludes that the DSMG is comparable to the VR in terms of accuracy. The time of computation in this experiment is about the same for the VR and DSMG.

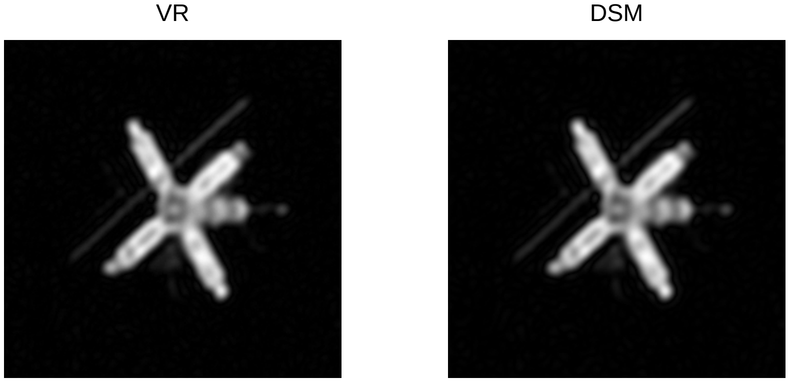

Figure 3: *Regularized images when noise level is 1%.*

Figure 4 plots the regularized images by the VR and the DSMG when $\delta_{rel} = 0.05$. It took 4 and 7 iterations for the DSMG and the VR, respectively, to yield numerical results. Figure 4 shows that the images obtained by the DSMG and the VR are about the same.

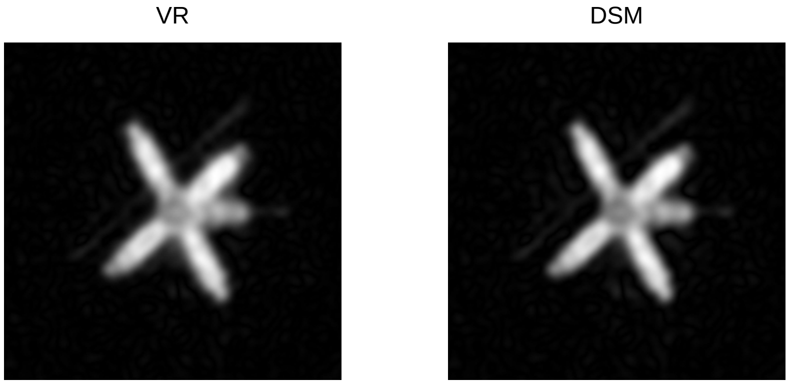

Figure 4: *Regularized images when noise level is 5%.*

The conclusions from this experiment are: the DSMG yields results with the same accuracy as the VR, and requires less iterations than the VR. The restored images by the DSM-[2] are about the same as those by the VR.

Remark 2 Equation (50) can be reduced to equation (53) whenever one of the two functions f and h has compact support and the other is periodic.

5 Concluding remarks

A version of the Dynamical Systems Method for solving ill-conditioned linear algebraic systems is studied in this paper. An *a priori* and *a posteriori* stopping rules are formulated and justified. An algorithm for computing the solution in the case when a spectral decomposition of the matrix A is available is presented. Numerical results show that the DSMG, i.e., the DSM version developed in this paper, yields results comparable to those obtained by the VR and the DSM-[2] developed in [2], and the DSMG method may yield much more accurate results than the VR method. It is demonstrated in [14] that the rate of convergence of the Landweber method can be increased by using preconditioning techniques. The rate of convergence of the DSM version, presented in this paper, might be improved by a similar technique. The advantage of our method over the steepest descent in [14] is the following: *the stopping time t_δ can be found from a discrepancy principle by Newton's iterations for a wide range of initial guess t_0; when t_δ is found one can compute the solution without any iterations.* Also, our method requires less iterations than the steepest descent in [14], which is an accelerated version of the Landweber method.

References

[1] Hairer, E., and Nørsett, S. P., and Wanner, G., Solving ordinary differential equation I, nonstiff problems, Springer, Berlin, 1987.

[2] Hoang, N. S. and Ramm, A. G., Solving ill-conditioned linear algebraic systems by the dynamical systems method (DSM), Inverse Problems in Sci. and Engineering, 16, N5, (2008), pp. 617 - 630.

[3] Hoang, N. S. and Ramm, A. G., On stable numerical differentiation, Australian J. Math. Anal. Appl., 5, N1, (2008), Article 5, pp.1-7.

[4] Hoang, N. S. and Ramm, A. G., An iterative scheme for solving nonlinear equations with monotone operators, BIT Numer. Math. 48, N4, (2008), 725-741.

[5] Hoang, N. S. and Ramm, A. G., Dynamical systems method for solving linear finite-rank operator equations, Ann. Polon. Math., 95, N1, (2009), 77-93.

[6] Hoang, N. S. and Ramm, A. G., Dynamical systems method for solving nonlinear equations with monotone operators, Math. of Comput., (to appear)

[7] Hoang, N. S. and Ramm, A. G., Dynamical Systems Gradient method for solving nonlinear equations with monotone operators, Acta Appl. Math., 106, (2009), 473-499.

[8] Hoang, N. S. and Ramm, A. G., A new version of the Dynamical Systems Method (DSM) for solving nonlinear equations with monotone operators, Diff. Eqns and Appl., 1, N1, (2009), 1-25.

[9] Hoang, N. S. and Ramm, A. G., A discrepancy principle for equations with monotone continuous operators, Nonlinear Analysis: Theory, Methods and Appl., 70, (2009), 4307-4315.

[10] Hoang, N. S. and Ramm, A. G., The Dynamical Systems Method for solving nonlinear equations with monotone operators, Asian Europ. Math. Journ., (to appear)

[11] Ivanov, V., Tanana, V., Vasin, V., Theory of ill-posed problems, VSP, Utrecht, 2002.

[12] Lattes, J., Lions, J., Mèthode de quasi-réversibilité et applications, Dunod, Paris, 1967.

[13] Morozov, Methods of solving incorrectly posed problems, Springer Verlag, New York, 1984.

[14] Nagy, J. G. and Palmer, K. M., Steepest descent, CG, and iterative regularization of ill-posed problems, BIT Numerical Mathematics, **43**(2003), 1003-1017.

[15] Ramm, A. G., Dynamical systems method for solving linear ill-posed problems, Ann. Polon. Math., 95, N3, (2009), 253-272.

[16] Ramm, A. G. and Smirnova, A. B., On stable numerical differentiation, Mathem. of Computation, 70, (2001), 1131-1153.

[17] Ramm, A. G. and Smirnova, A. B., Stable numerical differentiation: when is it possible? Jour. Korean SIAM, 7, N1, (2003), 47-61.

[18] Ramm, A. G. and Airapetyan, R., Dynamical systems and discrete methods for solving nonlinear ill-posed problems, Appl. Math. Reviews, vol. 1, Ed. G. Anastassiou, World Sci. Publishers, 2000, pp.491-536.

[19] Ramm, A. G., Dynamical systems method for solving operator equations, Elsevier, Amsterdam, 2007.

[20] Ramm, A. G., Dynamical systems method for solving operator equations, Communic. in Nonlinear Sci. and Numer. Simulation, 9, N2, (2004), 383-402.

[21] Ramm, A. G., Dynamical systems method (DSM) and nonlinear problems, in the book: Spectral Theory and Nonlinear Analysis, World Scientific Publishers, Singapore, 2005, 201-228. (ed J. Lopez-Gomez).

[22] Ramm, A. G., Dynamical systems method (DSM) for unbounded operators, Proc.Amer. Math. Soc., 134, N4, (2006), 1059-1063.

[23] Ramm, A. G., Dynamical systems method for nonlinear equations in Banach spaces, Communic. in Nonlinear Sci. and Numer. Simulation, 11, N3, (2006), 306-310.

[24] Tautenhahn, U., On the asymptotical regularization of nonlinear ill-posed problems, Inverse Problems, 10 (1994) 1405-1418.

[25] Vainikko, G., Veretennikov, A., Iterative processes in ill-posed problems, Nauka, Moscow, 1996.

Chapter 2

Dynamical systems method for solving finite-rank operator equations

Published in Ann. Polon. Math., 95, N1, (2009), 77-93.

Dynamical systems method for solving linear finite-rank operator equations

N. S. Hoang†[*]   A. G. Ramm†[‡]

†Mathematics Department, Kansas State University,
Manhattan, KS 66506-2602, USA

Abstract

A version of the Dynamical Systems Method (DSM) for solving ill-conditioned linear algebraic systems is studied in this paper. An *a priori* and *a posteriori* stopping rules are justified. An iterative scheme is constructed for solving ill-conditioned linear algebraic systems.

Keywords. Ill-posed problems, Dynamical Systems Method, Variational Regularization

1 Introduction

We want to solve stably the equation

$$Au = f, \tag{1}$$

where A is a linear bounded operator in a real Hilbert space H. We assume that (1) has a solution, possibly nonunique, and denote by y the unique minimal-norm solution to (1), $y \perp \mathcal{N} := \mathcal{N}(A) := \{u : Au = 0\}$, $Ay = f$. We assume that the range of A, $R(A)$, is not closed, so problem (1) is ill-posed. Let f_δ, $\|f - f_\delta\| \leq \delta$, be the noisy data. We want to construct a stable approximation of y, given $\{\delta, f_\delta, A\}$. There are many methods for doing this, see, e.g., [4]–[6], [7], [14], [15], to mention some (of the many) books, where variational regularization, quasisolutions, quasiinversion, and iterative regularization are studied, and [7]-[12], where the Dynamical Systems Method (DSM) is studied systematically (see also [1], [14], [13], and references therein for related results). The basic

[*] Email: nguyenhs@math.ksu.edu

[‡] Corresponding author. Email: ramm@math.ksu.edu

new results of this paper are: 1) a new version of the DSM for solving equation (1) is justified; 2) a stable method for solving equation (1) with noisy data by the DSM is given; a priori and a posteriori stopping rules are proposed and justified; 3) an iterative method for solving linear ill-conditioned algebraic systems, based on the proposed version of DSM, is formulated; its convergence is proved; 4) numerical results are given; these results show that the proposed method yields a good alternative to some of the standard methods (e.g., to variational regularization, Landweber iterations, and some other methods).

The DSM version we study in this paper consists of solving the Cauchy problem

$$\dot{u}(t) = -P(Au(t) - f), \quad u(0) = u_0, \quad u_0 \perp \mathcal{N}, \quad \dot{u} := \frac{du}{dt}, \tag{2}$$

and proving the existence of the limit $\lim_{t\to\infty} u(t) = u(\infty)$, and the relation $u(\infty) = y$, i.e.,

$$\lim_{t\to\infty} \|u(t) - y\| = 0. \tag{3}$$

Here P is a bounded operator such that $T := PA \geq 0$ is selfadjoint, $\mathcal{N}(T) = \mathcal{N}(A)$.

For any linear (not necessarily bounded) operator A there exists a bounded operator P such that $T = PA \geq 0$. For example, if $A = U|A|$ is the polar decomposition of A, then $|A| := (A^*A)^{\frac{1}{2}}$ is a selfadjoint operator, $T := |A| \geq 0$, U is a partial isometry, $\|U\| = 1$, and if $P := U^*$, then $\|P\| = 1$ and $PA = T$. Another choice of P, namely, $P = (A^*A + aI)^{-1}A^*$, $a = const > 0$, i s used in Section 3. For this choice $Q := AP \geq 0$.

If the noisy data f_δ are given, $\|f_\delta - f\| \leq \delta$, then we solve the problem

$$\dot{u}_\delta(t) = -P(Au_\delta(t) - f_\delta), \quad u_\delta(0) = u_0, \tag{4}$$

and prove that, for a suitable stopping time t_δ, and $u_\delta := u_\delta(t_\delta)$, one has

$$\lim_{\delta\to 0} \|u_\delta - y\| = 0. \tag{5}$$

An *a priori* and an *a posteriori* methods for choosing t_δ are given.

In Section 2 these results are formulated and recipes for choosing t_δ are proposed. In Section 3 a numerical example is presented.

2 Formulation and results

Suppose $A : H \to H$ is a linear bounded operator in a real Hilbert space H. Assume that equation (1) has a solution not necessarily unique. Denote by y the unique minimal-norm solution i.e.,

$y \perp \mathcal{N} := \mathcal{N}(A)$. Consider the DSM (2) where $u_0 \perp \mathcal{N}$ is arbitrary. Denote

$$T := PA, \quad Q := AP. \tag{6}$$

The unique solution to (2) is

$$u(t) = e^{-tT}u_0 + e^{-tT}\int_0^t e^{sT}dsPf. \tag{7}$$

Let us first show that any ill-posed linear equation (1) with exact data can be solved by the DSM. We assume below that $P = (A^*A + aI)^{-1}A^*$, where $a = const > 0$. With this choice of P one has $\mathcal{N}(T) = \mathcal{N}(A)$, $\|T\| \le 1$.

2.1 Exact data

The following result is known (see [7]) but a short proof is included for completeness.

Theorem 1 *Suppose $u_0 \perp \mathcal{N}$ and $T^* = T \ge 0$. Then problem (2) has a unique solution defined on $[0, \infty)$, and $u(\infty) = y$, where $u(\infty) = \lim_{t\to\infty} u(t)$.*

Proof. Denote $w := u(t) - y$, $w_0 := w(0) = u_0 - y$. Note that $w_0 \perp \mathcal{N}$. One has

$$\dot{w} = -Tw, \quad T := PA, \quad w(0) = u_0 - y. \tag{8}$$

The unique solution to (8) is $w = e^{-tT}w_0$. Thus,

$$\|w\|^2 = \int_0^{\|T\|} e^{-2t\lambda}d\langle E_\lambda w_0, w_0\rangle.$$

where $\langle u, v\rangle$ is the inner product in H, and E_λ is the resolution of the identity of T. Thus,

$$\|w(\infty)\|^2 = \lim_{t\to\infty}\int_0^{\|T\|} e^{-2t\lambda}d\langle E_\lambda w_0, w_0\rangle = \|P_{\mathcal{N}}w_0\|^2 = 0,$$

where $P_{\mathcal{N}} = E_0 - E_{-0}$ is the orthogonal projector onto $\mathcal{N}$. Theorem 1 is proved. □

2.2 Noisy data f_δ

Let us solve stably equation (1) assuming that f is not known, but f_δ, the noisy data, are known, where $\|f_\delta - f\| \le \delta$. Consider the following DSM

$$\dot{u}_\delta = -P(Au_\delta - f_\delta), \quad u_\delta(0) = u_0. \tag{9}$$

Denote

$$w_\delta := u_\delta - y, \quad T := PA, \quad w_\delta(0) = w_0 := u_0 - y \in \mathcal{N}^\perp.$$

Let us prove the following result:

Theorem 2 *If* $T = T^* \geq 0$, $\lim_{\delta\to 0} t_\delta = \infty$, $\lim_{\delta\to 0} t_\delta \delta = 0$, *and* $w_0 \in \mathcal{N}^\perp$, *then*

$$\lim_{\delta\to 0} \|w_\delta(t_\delta)\| = 0.$$

Proof. One has

$$\dot{w}_\delta = -Tw_\delta + \zeta_\delta, \quad \zeta_\delta = P(f_\delta - f), \quad \|\zeta_\delta\| \leq \|P\|\delta. \tag{10}$$

The unique solution of equation (10) is

$$w_\delta(t) = e^{-tT} w_\delta(0) + \int_0^t e^{-(t-s)T} \zeta_\delta ds.$$

Let us show that $\lim_{\delta\to 0} \|w_\delta(t_\delta)\| = 0$. One has

$$\lim_{t\to\infty} \|w_\delta(t)\| \leq \lim_{t\to\infty} \|e^{-tT} w_\delta(0)\| + \lim_{t\to\infty} \left\| \int_0^t e^{-(t-s)T} \zeta_\delta ds \right\|. \tag{11}$$

Let E_λ be the resolution of identity corresponding to T. One uses the spectral theorem and gets:

$$\begin{aligned}\int_0^t e^{-(t-s)T} ds \zeta_\delta &= \int_0^t \int_0^{\|T\|} dE_\lambda \zeta_\delta e^{-(t-s)\lambda} ds \\ &= \int_0^{\|T\|} e^{-t\lambda} \frac{e^{t\lambda} - 1}{\lambda} dE_\lambda \zeta_\delta = \int_0^{\|T\|} \frac{1 - e^{-t\lambda}}{\lambda} dE_\lambda \zeta_\delta.\end{aligned} \tag{12}$$

Note that

$$0 \leq \frac{1 - e^{-t\lambda}}{\lambda} \leq t, \quad \forall \lambda > 0, \quad t \geq 0, \tag{13}$$

since $1 - x \leq e^{-x}$ for $x \geq 0$. From (12) and (13), one obtains

$$\begin{aligned}\left\| \int_0^t e^{-(t-s)T} ds \zeta_\delta \right\|^2 &= \int_0^{\|T\|} \left|\frac{1 - e^{-t\lambda}}{\lambda}\right|^2 d\langle E_\lambda \zeta_\delta, \zeta_\delta\rangle \\ &\leq t^2 \int_0^{\|T\|} d\langle E_\lambda \zeta_\delta, \zeta_\delta\rangle \\ &= t^2 \|\zeta_\delta\|^2.\end{aligned} \tag{14}$$

This estimate follows also from the inequality: $\|e^{-(t-s)T}\| \leq 1$, which holds for $T^* = T \geq 0$ and $t \geq s$. Indeed, one has $\| \int_0^t e^{-(t-s)T} ds\| \leq t$, and estimate (14) follows.

Since $\|\zeta_\delta\| \leq \|P\|\delta$, from (11) and (14), one gets

$$\lim_{\delta\to 0} \|w_\delta(t_\delta)\| \leq \lim_{\delta\to 0} \left(\|e^{-t_\delta T} w_\delta(0)\| + t_\delta \delta \|P\| \right) = 0.$$

Here we have used the relation:

$$\lim_{\delta\to 0} \|e^{-t_\delta T} w_\delta(0)\| = \|P_{\mathcal{N}} w_0\| = 0,$$

and the last equality holds because $w_0 \in \mathcal{N}^\perp$. Theorem 2 is proved. □

From Theorem 2, it follows that the relation

$$t_\delta = \frac{C}{\delta^\gamma}, \quad \gamma = \text{const}, \quad \gamma \in (0,1)$$

where $C > 0$ is a constant, can be used as an *a priori* stopping rule, i.e., for such t_δ one has

$$\lim_{\delta\to 0} \|u_\delta(t_\delta) - y\| = 0. \tag{15}$$

2.3 Discrepancy principle

In this section we assume that A is a linear finite-rank operator. Thus, it is a linear bounded operator. Let us consider equation (1) with noisy data f_δ, and a DSM of the form

$$\dot{u}_\delta = -PAu_\delta + Pf_\delta, \quad u_\delta(0) = u_0, \tag{16}$$

for solving this equation. Equation (16) has been used in Section 2.2. Recall that y denotes the minimal-norm solution of equation (1), and that $\mathcal{N}(T) = \mathcal{N}(A)$ with our choice of P.

Theorem 3 *Let $T := PA$, $Q := AP$. Assume that $\|Au_0 - f_\delta\| > C\delta$, $Q = Q^* \geq 0$, $T^* = T \geq 0$, and T is a finite-rank operator. Then the solution t_δ to the equation*

$$h(t) := \|Au_\delta(t) - f_\delta\| = C\delta, \quad C = const, \quad C \in (1,2), \tag{17}$$

does exist, is unique, $\lim_{\delta\to 0} t_\delta = \infty$, and

$$\lim_{\delta\to 0} \|u_\delta(t_\delta) - y\| = 0, \tag{18}$$

where y is the unique minimal-norm solution to (1).

Proof. Denote

$$v_\delta(t) := Au_\delta(t) - f_\delta, \quad w(t) := u(t) - y, \quad w_0 := u_0 - y.$$

One has

$$\begin{aligned} \frac{d}{dt}\|v_\delta(t)\|^2 &= 2\langle A\dot{u}_\delta(t), Au_\delta(t) - f_\delta\rangle \\ &= 2\langle A[-P(Au_\delta(t) - f_\delta)], Au_\delta(t) - f_\delta\rangle \\ &= -2\langle AP(Au_\delta - f_\delta), Au_\delta - f_\delta\rangle \leq 0, \end{aligned} \tag{19}$$

where the last inequality holds because $AP = Q \geq 0$. Thus, $\|v_\delta(t)\|$ is a nonincreasing function.

Let us prove that equation (17) has a solution for $C \in (1, 2)$. One has the following commutation formulas:

$$e^{-sT}P = Pe^{-sQ}, \quad Ae^{-sT} = e^{-sQ}A.$$

Using these formulas and the representation

$$u_\delta(t) = e^{-tT}u_0 + \int_0^t e^{-(t-s)T}Pf_\delta ds,$$

one gets:

$$\begin{aligned} v_\delta(t) &= Au_\delta(t) - f_\delta \\ &= Ae^{-tT}u_0 + A\int_0^t e^{-(t-s)T}Pf_\delta ds - f_\delta \\ &= e^{-tQ}Au_0 + e^{-tQ}\int_0^t e^{sQ}dsQf_\delta - f_\delta \\ &= e^{-tQ}A(u_0 - y) + e^{-tQ}f + e^{-tQ}(e^{tQ} - I)f_\delta - f_\delta \\ &= e^{-tQ}Aw_0 - e^{-tQ}f_\delta + e^{-tQ}f = e^{-tQ}Au_0 - e^{-tQ}f_\delta. \end{aligned} \tag{20}$$

Note that

$$\lim_{t\to\infty} e^{-tQ}Aw_0 = \lim_{t\to\infty} Ae^{-tT}w_0 = AP_{\mathcal{N}}w_0 = 0.$$

Here the continuity of A and the following relation

$$\lim_{t\to\infty} e^{-tT}w_0 = \lim_{t\to\infty}\int_0^{\|T\|} e^{-st}dE_s w_0 = (E_0 - E_{-0})w_0 = P_{\mathcal{N}}w_0,$$

were used. Therefore,

$$\lim_{t\to\infty}\|v_\delta(t)\| = \lim_{t\to\infty}\|e^{-tQ}(f - f_\delta)\| \le \|f - f_\delta\| \le \delta, \tag{21}$$

where $\|e^{-tQ}\| \le 1$ because $Q \ge 0$. The function $h(t)$ is continuous on $[0, \infty)$, $h(0) = \|Au_0 - f_\delta\| > C\delta$, $h(\infty) \le \delta$. Thus, equation (17) must have a solution t_δ.

Let us prove the uniqueness of t_δ. If t_δ is non-unique, then without loss of generality we can assume that there exists $t_1 > t_\delta$ such that $\|Au_\delta(t_1) - f_\delta\| = C\delta$. Since $\|v_\delta(t)\|$ is nonincreasing and $\|v_\delta(t_\delta)\| = \|v_\delta(t_1)\|$, one has

$$\|v_\delta(t)\| = \|v_\delta(t_\delta)\|, \quad \forall t \in [t_\delta, t_1].$$

Thus,

$$\frac{d}{dt}\|v_\delta(t)\|^2 = 0, \quad \forall t \in (t_\delta, t_1). \tag{22}$$

Using (19) and (22) one obtains

$$\|\sqrt{AP}(Au_\delta(t)-f_\delta)\|^2 = \langle AP(Au_\delta(t)-f_\delta), Au_\delta(t)-f_\delta\rangle = 0, \quad \forall t\in[t_\delta,t_1],$$

where $\sqrt{AP}=Q^{\frac{1}{2}}\geq 0$ is well defined since $Q=Q^*\geq 0$. This implies $Q^{\frac{1}{2}}(Au_\delta - f_\delta)=0$. Thus

$$Q(Au_\delta(t)-f_\delta)=0, \quad \forall t\in[t_\delta,t_1]. \tag{23}$$

From (20) one gets:

$$v_\delta(t) = Au_\delta(t)-f_\delta = e^{-tQ}Au_0 - e^{-tQ}f_\delta. \tag{24}$$

Since $Qe^{-tQ}=e^{-tQ}Q$ and e^{-tQ} is an isomorphism, equalities (23) and (24) imply

$$Q(Au_0-f_\delta)=0.$$

This and (24) imply

$$AP(Au_\delta(t)-f_\delta) = e^{-tQ}(QAu_0 - Qf_\delta)=0, \quad t\geq 0.$$

This and (19) imply

$$\frac{d}{dt}\|v_\delta\|^2 = 0, \quad t\geq 0. \tag{25}$$

Consequently,

$$C\delta < \|Au_\delta(0)-f_\delta\| = \|v_\delta(0)\| = \|v_\delta(t_\delta)\| = \|Au_\delta(t_\delta)-f_\delta\| = C\delta.$$

This is a contradiction which proves the uniqueness of t_δ.

Let us prove (18). First, we have the following estimate:

$$\begin{aligned}\|Au(t_\delta)-f\| &\leq \|Au(t_\delta)-Au_\delta(t_\delta)\| + \|Au_\delta(t_\delta)-f_\delta\| + \|f_\delta - f\| \\ &\leq \left\|e^{-t_\delta Q}\int_0^{t_\delta} e^{sQ}Q ds\right\| \|f_\delta - f\| + C\delta + \delta,\end{aligned} \tag{26}$$

where $u(t)$ solves (2) and $u_\delta(t)$ solves (9). One uses the inequality:

$$\big\|e^{-t_\delta Q}\int_0^{t_\delta} e^{sQ}Q ds\big\| = \|I - e^{-t_\delta Q}\| \leq 2,$$

and concludes from (26), that

$$\lim_{\delta\to 0}\|Au(t_\delta)-f\| = 0. \tag{27}$$

Secondly, we claim that

$$\lim_{\delta\to 0} t_\delta = \infty.$$

Assume the contrary. Then there exist $t_0 > 0$ and a sequence $(t_{\delta_n})_{n=1}^{\infty}$, $t_{\delta_n} < t_0$, $\lim_{n\to\infty} \delta_n = 0$, such that

$$\lim_{n\to\infty} \|Au(t_{\delta_n}) - f\| = 0. \tag{28}$$

Analogously to (19), one proves that

$$\frac{d}{dt}\|v\|^2 \le 0,$$

where $v(t) := Au(t) - f$. Thus, $\|v(t)\|$ is nonincreasing. This and (28) imply the relation $\|v(t_0)\| = \|Au(t_0) - f\| = 0$. Thus,

$$0 = v(t_0) = e^{-t_0 Q} A(u_0 - y).$$

This implies $A(u_0 - y) = e^{t_0 Q} e^{-t_0 Q} A(u_0 - y) = 0$, so $u_0 - y \in \mathcal{N}$. Since $u_0 - y \in \mathcal{N}^{\perp}$, it follows that $u_0 = y$. This is a contradiction because

$$C\delta \le \|Au_0 - f_\delta\| = \|f - f_\delta\| \le \delta, \quad 1 < C < 2.$$

Thus,

$$\lim_{\delta\to 0} t_\delta = \infty. \tag{29}$$

Let us continue the proof of (18). From (20) and the relation $\|Au_\delta(t_\delta) - f_\delta\| = C\delta$, one has

$$\begin{aligned} C\delta t_\delta &= \|t_\delta e^{-t_\delta Q} A w_0 - t_\delta e^{-t_\delta Q}(f_\delta - f)\| \\ &\le \|t_\delta e^{-t_\delta Q} A w_0\| + \|t_\delta e^{-t_\delta Q}(f_\delta - f)\| \\ &\le \|t_\delta e^{-t_\delta Q} A w_0\| + t_\delta \delta. \end{aligned} \tag{30}$$

We claim that

$$\lim_{\delta\to 0} t_\delta e^{-t_\delta Q} A w_0 = \lim_{\delta\to 0} t_\delta A e^{-t_\delta T} w_0 = 0. \tag{31}$$

Note that (31) holds if $T \ge 0$ *has finite rank, and* $w_0 \in \mathcal{N}^{\perp}$. It also holds if $T \ge 0$ is compact and the Fourier coefficients $w_{0j} := \langle w_0, \phi_j\rangle$, $T\phi_j = \lambda_j \phi_j$, decay sufficiently fast. In this case

$$\|Ae^{-tT} w_0\|^2 \le \|T^{\frac{1}{2}} e^{-tT} w_0\|^2 = \sum_{j=1}^{\infty} \lambda_j e^{-2\lambda_j t} |w_{0j}|^2 := S = o(\frac{1}{t^2}), \quad t \to \infty,$$

provided that $\sum_{j=1}^{\infty} |w_{0j}|\lambda_j^{-2} < \infty$. Indeed,

$$S = \sum_{\lambda_j \le \frac{1}{t^{\frac{2}{3}}}} + \sum_{\lambda_j > \frac{1}{t^{\frac{2}{3}}}} := S_1 + S_2.$$

One has

$$S_1 \le \frac{1}{t^2} \sum_{\lambda_j \le t^{-\frac{2}{3}}} \frac{|w_{0j}|^2}{\lambda_j^2} = o(\frac{1}{t^2}), \quad S_2 \le ce^{-2t^{\frac{1}{3}}} = o(\frac{1}{t^2}), \quad t \to \infty,$$

where $c > 0$ is a constant.

From (31) and (30), one gets

$$0 \le \lim_{\delta\to 0}(C-1)\delta t_\delta \le \lim_{\delta\to 0}\|t_\delta e^{-t_\delta Q}Aw_0\| = 0.$$

Thus,

$$\lim_{\delta\to 0}\delta t_\delta = 0 \tag{32}$$

Now, the desired conclusion (18) follows from (29), (32) and Theorem 2. Theorem 3 is proved. □

2.4 An iterative scheme

Let us solve stably equation (1) assuming that f is not known, but f_δ, the noisy data, are known, where $\|f_\delta - f\| \le \delta$. Consider the following discrete version of the DSM:

$$u_{n+1,\delta} = u_{n,\delta} - hP(Au_{n,\delta} - f_\delta), \quad u_{\delta,0} = u_0. \tag{33}$$

Let us denote $u_n := u_{n,\delta}$ when $\delta \ne 0$, and set

$$w_n := u_n - y, \quad T := PA, \quad w_0 := u_0 - y \in \mathcal{N}^\perp.$$

Let $n = n_\delta$ be the stopping rule for iterations (33). Let us prove the following result:

Theorem 4 *Assume that $T = T^* \ge 0$, $h\|T\| < 2$, $\lim_{\delta\to 0} n_\delta h = \infty$, $\lim_{\delta\to 0} n_\delta h\delta = 0$, and $w_0 \in \mathcal{N}^\perp$. Then*

$$\lim_{\delta\to 0}\|w_{n_\delta}\| = \lim_{\delta\to 0}\|u_{n_\delta} - y\| = 0. \tag{34}$$

Proof. One has

$$w_{n+1} = w_n - hTw_n + h\zeta_\delta, \quad \zeta_\delta = P(f_\delta - f), \quad \|\zeta_\delta\| \le \|P\|\delta, \quad w_0 = u_0 - y. \tag{35}$$

The unique solution of equation (35) is

$$w_{n+1} = (I-hT)^{n+1}w_0 + h\sum_{i=0}^{n}(I-hT)^i\zeta_\delta.$$

Let us show that $\lim_{\delta\to 0}\|w_{n_\delta}\| = 0$. One has

$$\|w_n\| \le \|(I-hT)^n w_0\| + \left\| h\sum_{i=0}^{n-1}(I-hT)^i\zeta_\delta \right\|. \tag{36}$$

Let E_λ be the resolution of the identity corresponding to T. One uses the spectral theorem and gets:

$$\begin{aligned} h\sum_{i=0}^{n-1}(I-hT)^i &= h\sum_{i=0}^{n-1}\int_0^{\|T\|}(1-h\lambda)^i dE_\lambda \\ &= h\int_0^{\|T\|}\frac{1-(1-\lambda h)^n}{1-(1-h\lambda)}dE_\lambda = \int_0^{\|T\|}\frac{1-(1-\lambda h)^n}{\lambda}dE_\lambda. \end{aligned} \tag{37}$$

Note that

$$0 \le \frac{1-(1-h\lambda)^n}{\lambda} \le hn, \quad \forall \lambda > 0, \quad t \ge 0, \tag{38}$$

since $1-(1-\alpha)^n \le \alpha n$ for all $\alpha \in [0,2]$. From (37) and (38), one obtains

$$\begin{aligned} \Big\|h\sum_{i=0}^{n-1}(I-hT)^i\zeta_\delta\Big\|^2 &= \int_0^{\|T\|}\big|\frac{1-(1-\lambda h)^n}{\lambda}\big|^2 d\langle E_\lambda\zeta_\delta,\zeta_\delta\rangle \\ &\le (hn)^2\int_0^{\|T\|} d\langle E_\lambda\zeta_\delta,\zeta_\delta\rangle \\ &= (nh)^2\|\zeta_\delta\|^2. \end{aligned} \tag{39}$$

Alternatively, this estimate follows from the inequality $\|(I-hT)^i\| \le 1$, provided that $0 \le hT < 2$. Indeed, in this case one has $\|\sum_{i=0}^{n-1}(I-hT)^i\| \le n$, and this implies estimate (39).

Since $\|\zeta_\delta\| \le \|P\|\delta$, from (36) and (39), one gets

$$\lim_{\delta\to 0}\|w_{n_\delta}\| \le \lim_{\delta\to 0}\Big(\|(I-hT)^{n_\delta}w_\delta(0)\| + hn_\delta\delta\|P\|\Big) = 0.$$

Here we have used the relation:

$$\lim_{\delta\to 0}\|(I-hT)^{n_\delta}w_\delta(0)\| = \|P_{\mathcal{N}}w_0\| = 0,$$

and the last equality holds because $w_0 \in \mathcal{N}^\perp$. Theorem 4 is proved. □

From Theorem 4, it follows that the relation

$$n_\delta = \frac{C}{h\delta^\gamma}, \quad \gamma = \text{const}, \quad \gamma \in (0,1)$$

where $C > 0$ is a constant, can be used as an *a priori* stopping rule, i.e., for such n_δ one has

$$\lim_{\delta\to 0}\|u_{n_\delta} - y\| = 0. \tag{40}$$

2.5 An iterative scheme with a stopping rule based on a discrepancy principle

In this section we assume that A is a linear finite-rank operator. Thus, it is a linear bounded operator. Let us consider equation (1) with noisy data f_δ, and a DSM of the form

$$u_{n+1} = u_n - hP(Au_n - f_\delta), \quad u_0 = u_0, \tag{41}$$

for solving this equation. Equation (41) has been used in Section 2.4. Recall that y denotes the minimal-norm solution of equation (1). Example of a choice of P is given in Section 3.

Note that $\mathcal{N} := \mathcal{N}(T) = \mathcal{N}(A)$.

Theorem 5 *Let $T := PA$, $Q := AP$. Assume that $\|Au_0 - f_\delta\| > C\delta$, $Q = Q^* \geq 0$, $T^* = T \geq 0$, $h\|T\| < 2$, $h\|Q\| < 2$, and T is a finite-rank operator. Then there exists a unique n_δ such that*

$$\|Au_{n_\delta} - f_\delta\| \leq C\delta < \|Au_{n_\delta - 1} - f_\delta\|, \quad C = const, \quad C \in (1, 2). \tag{42}$$

For this n_δ one has:

$$\lim_{\delta \to 0} \|u_{n_\delta} - y\| = 0. \tag{43}$$

Proof. Denote

$$v_n := Au_n - f_\delta, \quad w_n := u_n - y, \quad w_0 := u_0 - y.$$

From (41), one gets

$$v_{n+1} = Au_{n+1} - f_\delta = Au_n - f_\delta - hAP(Au_n - f_\delta) = v_n - hQv_n.$$

This implies

$$\begin{aligned} \|v_{n+1}\|^2 - \|v_n\|^2 &= \langle v_{n+1} - v_n, v_{n+1} + v_n \rangle \\ &= \langle -hQv_n, v_n - hQv_n + v_n \rangle \\ &= -\langle v_n, hQ(2 - hQ)v_n \rangle \leq 0 \end{aligned} \tag{44}$$

where the last inequality holds because $AP = Q \geq 0$ and $\|hQ\| < 2$. Thus, $(\|v_n\|)_{n=1}^\infty$ is a nonincreasing sequence.

Let us prove that equation (42) has a solution for $C \in (1, 2)$. One has the following commutation formulas:

$$(I - hT)^n P = P(I - hQ)^n, \quad A(I - hT)^n = (I - hQ)^n A.$$

Using these formulas, the representation

$$u_n = (I - hT)^n u_0 + h\sum_{i=0}^{n-1}(I - hT)^i P f_\delta,$$

and the identity $(I - B)\sum_{i=0}^{n-1} B^i = I - B^n$, with $B = I - hQ$, $I - B = hQ$, one gets:

$$\begin{aligned}
v_n &= Au_n - f_\delta \\
&= A(I - hT)^n u_0 + Ah\sum_{i=0}^{n-1}(I - hT)^i P f_\delta - f_\delta \\
&= (I - hQ)^n A u_0 + \sum_{i=0}^{n-1}(I - hQ)^i hQ f_\delta - f_\delta \\
&= (I - hQ)^n A u_0 - (I - (I - hQ)^n) f_\delta - f_\delta \\
&= (I - hQ)^n (Au_0 - f) + (I - hQ)^n (f - f_\delta) \\
&= (I - hQ)^n A w_0 + (I - hQ)^n (f - f_\delta).
\end{aligned} \tag{45}$$

If $V = V^* \geq 0$ is an operator with $||V|| \leq 2$, then $||I - V|| = \sup_{0\leq s\leq 2}|1 - s| \leq 1$.

Note that

$$\lim_{n\to\infty}(I - hQ)^n A w_0 = \lim_{n\to\infty} A(I - hT)^n w_0 = AP_{\mathcal{N}} w_0 = 0,$$

where $P_{\mathcal{N}}$ is the orthoprojection onto the null-space $\mathcal{N}$ of the operator T, and the continuity of A and the following relation

$$\lim_{n\to\infty}(I - hT)^n w_0 = \lim_{n\to\infty}\int_0^{\|T\|}(1 - sh)^n dE_s w_0 = (E_0 - E_{-0})w_0 = P_{\mathcal{N}} w_0, \quad 0 \leq sh < 2,$$

were used. Therefore,

$$\lim_{n\to\infty}\|v_\delta(t)\| = \lim_{n\to\infty}\|(I - hQ)^n (f - f_\delta)\| \leq \|f - f_\delta\| \leq \delta, \tag{46}$$

where $\|I - hQ\| \leq 1$ because $Q \geq 0$ and $\|hQ\| < 2$. The sequence $\{\|v_n\|\}_{n=1}^\infty$ is nonincreasing with $\|v_0\| > C\delta$ and $\lim_{n\to\infty}\|v_n\| \leq \delta$. Thus, there exists $n_\delta > 0$ such that (42) holds.

Let us prove (43). Let $u_{n,0}$ be the sequence defined by the relations:

$$u_{n+1,0} = u_{n,0} - hP(Au_{n,0} - f), \quad u_{0,0} = u_0.$$

First, we have the following estimate:

$$\begin{aligned}
\|Au_{n_\delta,0} - f\| &\leq \|Au_{n_\delta} - Au_{n_\delta,0}\| + \|Au_{n_\delta} - f_\delta\| + \|f_\delta - f\| \\
&\leq \left\|\sum_{i=0}^{n_\delta - 1}(I - hQ)^i hQ\right\|\|f_\delta - f\| + C\delta + \delta.
\end{aligned} \tag{47}$$

Since $0 \le hQ < 2$, one has $||I - hQ|| \le 1$. This implies the following inequality:

$$\left\| \sum_{i=0}^{n_\delta - 1} (I - hQ)^i hQ \right\| = \|I - (I - hQ)^{n_\delta}\| \le 2,$$

and one concludes from (47) that

$$\lim_{\delta \to 0} \|Au_{n_\delta,0} - f\| = 0. \tag{48}$$

Secondly, we claim that

$$\lim_{\delta \to 0} hn_\delta = \infty.$$

Assume the contrary. Then there exist $n_0 > 0$ and a sequence $(n_{\delta_n})_{n=1}^{\infty}$, $n_{\delta_n} < n_0$, such that

$$\lim_{n \to \infty} \|Au_{n_\delta,0} - f\| = 0. \tag{49}$$

Analogously to (44), one proves that

$$\|v_{n,0}\| \le \|v_{n-1,0}\|,$$

where $v_{n,0} = Au_{n,0} - f$. Thus, the sequence $\|v_{n,0}\|$ is nonincreasing. This and (49) imply the relation $\|v_{n_0,0}\| = \|Au_{n_0,0} - f\| = 0$. Thus,

$$0 = v_{n_0,0} = (I - hQ)^{n_0} A(u_0 - y).$$

This implies $A(u_0 - y) = (I - hQ)^{-n_0}(I - hQ)^{n_0} A(u_0 - y) = 0$, so $u_0 - y \in \mathcal{N}$. Since, by the assumption, $u_0 - y \in \mathcal{N}^\perp$, it follows that $u_0 = y$. This is a contradiction because

$$C\delta \le \|Au_0 - f_\delta\| = \|f - f_\delta\| \le \delta, \quad 1 < C < 2.$$

Thus,

$$\lim_{\delta \to 0} hn_\delta = \infty. \tag{50}$$

Let us continue the proof of (43). From (45) and $\|Au_{n_\delta} - f_\delta\| = C\delta$, one has

$$\begin{aligned} C\delta n_\delta h &= \|n_\delta h(I - hQ)^{n_\delta} Aw_0 - n_\delta h(I - hQ)^{n_\delta}(f_\delta - f)\| \\ &\le \|n_\delta h(I - hQ)^{n_\delta} Aw_0\| + \|n_\delta h(I - hQ)^{n_\delta}(f_\delta - f)\| \\ &\le \|n_\delta h(I - hQ)^{n_\delta} Aw_0\| + n_\delta h\delta. \end{aligned} \tag{51}$$

We claim that if $w_0 \in \mathcal{N}^\perp$, $0 \le hT < 2$, and T is a finite-rank operator, then

$$\lim_{\delta \to 0} n_\delta h(I - hQ)^{n_\delta} Aw_0 = \lim_{\delta \to 0} n_\delta hA(I - hT)^{n_\delta} w_0 = 0. \tag{52}$$

From (51) and (52) one gets

$$0 \le \lim_{\delta\to 0}(C-1)\delta h n_\delta \le \lim_{\delta\to 0}\|n_\delta h(I-hQ)^{n_\delta}Aw_0\| = 0.$$

Thus,

$$\lim_{\delta\to 0}\delta n_\delta h = 0 \tag{53}$$

Now (43) follows from (50), (53) and Theorem 4. Theorem 5 is proved. □

3 Numerical experiments

3.1 Computing $u_\delta(t_\delta)$

In [3] a DSM (9) was investigated with $P = A^*$ and the singular value decomposition (SVD) of A was assumed known. In general, it is computationally expensive to get the SVD of large scale matrices. In this paper, we have derived an iterative scheme for solving ill-conditioned linear algebraic systems $Au = f_\delta$ without using SVD of A.

Choose $P = (A^*A + a)^{-1}A^*$ where a is a fixed positive constant. This choice of P satisfies all the conditions in Theorem 3. In particular, $Q = AP = A(A^*A + aI)^{-1}A^* = AA^*(AA^* + aI)^{-1} \ge 0$ is a selfadjoint operator, and $T = PA = (A^*A + aI)^{-1}A^*A \ge 0$ is a selfadjoint operator. Since

$$\|T\| = \left\| \int_0^{\|A^*A\|} \frac{\lambda}{\lambda + a} dE_\lambda \right\| = \sup_{0\le\lambda\le\|A^*A\|} \frac{\lambda}{\lambda+a} < 1,$$

where E_λ is the resolution of the identity of A^*A, the condition $h\|T\| < 2$ in Theorem 5 is satisfied for all $0 < h \le 1$. Set $h = 1$ and $P = (A^*A + a)^{-1}A^*$ in (41). Then one gets the following iterative scheme:

$$u_{n+1} = u_n - (A^*A + aI)^{-1}(A^*Au_n - A^*f_\delta), \quad u_0 = 0. \tag{54}$$

For simplicity we have chosen $u_0 = 0$. However, one may choose $u_0 = v_0$ if v_0 is known to be a better approximation to y than 0 and $v_0 \in \mathcal{N}^\perp$. In iterations (54) we use a stopping rule of discrepancy type. Indeed, we stop iterations if u_n satisfies the following condition

$$\|Au_n - f_\delta\| \le 1.01\delta. \tag{55}$$

The choice of a affects both the accuracy and the computation time of the method. If a is too large, one needs more iterations to approach the desired accuracy, so the computation time will be large. If a is too small, then the results become less accurate because for too small a the inversion of the operator A^*A+aI is an ill-posed problem since the operator A^*A is not boundedly invertible. Using the idea of the choice of the initial guess of the regularization parameter in [2], we choose a to satisfy the following condition:

$$\delta \leq \phi(a) := \|A(A^*A+a)^{-1}A^*f_\delta - f_\delta\| \leq 2\delta. \tag{56}$$

This can be done by using the following strategy:

1. Choose $a := \frac{\delta\|A\|^2}{3\|f_\delta\|}$ as an initial guess for a.

2. Compute $\phi(a)$. If a satisfies (56), then we are done. Otherwise, we go to step 3.

3. If $c = \frac{\phi(a)}{\delta} > 3$, we replace a by $\frac{a}{2(c-1)}$ and go back to step 2. If $2 < c \leq 3$, then we replace a by $\frac{a}{2(c-1)}$ and go back to step 2. Otherwise, we go to step 4.

4. If $c = \frac{\phi(a)}{\delta} < 1$, we replace a by $3a$. If the inequality $c < 1$ has occured in an earlier iteration, we stop the iterations and use $3a$ as our choice for a in iterations (54). Otherwise we go back to step 2.

In our experiments, we denote by DSM the iterative scheme (54), by VR$_i$ a Variational Regularization method (VR) with a as the regularization parameter, and by VR$_n$ the VR in which Newton's method is used for finding the regularization parameter from a discrepancy principle. We compare these methods in terms of relative error and number of iterations, denoted by n$_{iter}$.

All the experiments were carried in double arithmetics precision environment using MATLAB.

3.2 A linear algebraic system related to an inverse problem for the heat equation

In this section, we apply the DSM and the VR to solve a linear algebraic system used in [2]. This linear algebraic system is a part of numerical solutions to an inverse problem for the heat equation. This problem is reduced to a Volterra integral equation of the first kind with $[0,1]$ as the integration interval. The kernel is $K(s,t) = k(s-t)$ with

$$k(t) = \frac{t^{-3/2}}{2\kappa\sqrt{\pi}}\exp(-\frac{1}{4\kappa^2 t}).$$

Here, we use the value $\kappa = 1$. In this test in [2] the integral equation was discretized by means of simple collocation and the midpoint rule with n points. The unique exact solution u_n is constructed, and then the right-hand side b_n is produced as $b_n = A_n u_n$ (see [2]). In our test, we use $n =$

$10, 20, ..., 100$ and $b_{n,\delta} = b_n + e_n$, where e_n is a vector containing random entries, normally distributed with mean 0, variance 1, and scaled so that $\|e_n\| = \delta_{rel}\|b_n\|$. This linear system is ill-posed: the condition number of A_{100} obtained by using the function *cond* provided in MATLAB is 1.3717×10^{37}. This number shows that the corresponding linear algebraic system is severely ill-conditioned.

Table 2: Numerical results for the inverse heat equation with $\delta_{rel} = 0.05$, $n = 10i$, $i = \overline{1, 10}$.

	DSM		VR$_i$		VR$_n$	
n	n$_{\text{iter}}$	$\frac{\|u_\delta - y\|_2}{\|y\|_2}$	n$_{\text{iter}}$	$\frac{\|u_\delta - y\|_2}{\|y\|_2}$	n$_{\text{iter}}$	$\frac{\|u_\delta - y\|_2}{\|y\|_2}$
10	3	0.1971	1	0.2627	5	0.2117
20	4	0.3359	1	0.4589	5	0.3551
30	4	0.3729	1	0.4969	5	0.3843
40	4	0.3856	1	0.5071	5	0.3864
50	5	0.3158	1	0.4789	6	0.3141
60	6	0.2892	1	0.4909	6	0.3060
70	7	0.2262	1	0.4792	8	0.2156
80	6	0.2623	1	0.4809	7	0.2600
90	5	0.2856	1	0.4816	7	0.2715
100	7	0.2358	1	0.4826	7	0.3405

Table 2 shows that the results obtained by the DSM are comparable to those by the VR$_n$ in terms of accuracy. The time of computation of the DSM is comparable to that of the VR$_n$. In some situations, the results by VR$_n$ and the DSM are the same although the VR$_n$ uses 3 more iterations than does the DSM. The conclusion from this Table is that DSM competes favorably with the VR$_n$ in both accuracy and time of computation.

Figure 5 plots numerical solutions to the inverse heat equation for $\delta_{rel} = 0.05$ and $\delta_{rel} = 0.01$ when $n = 100$. From the figure one can see that the numerical solutions obtained by the DSM are about the same those by the VR$_n$. In these examples, the time of computation of the DSM is about the same as that of the VR$_n$.

The conclusion is that the DSM competes favorably with the VR$_n$ in this experiment.

4 Concluding remarks

Iterative scheme (54) can be considered as a modification the Landweber iterations. The difference between the two methods is the multiplication by $(A^*A + aI)^{-1}$. Our iterative method is much faster than the conventional Landweber iterations. Iterative method (54) is an analog of the

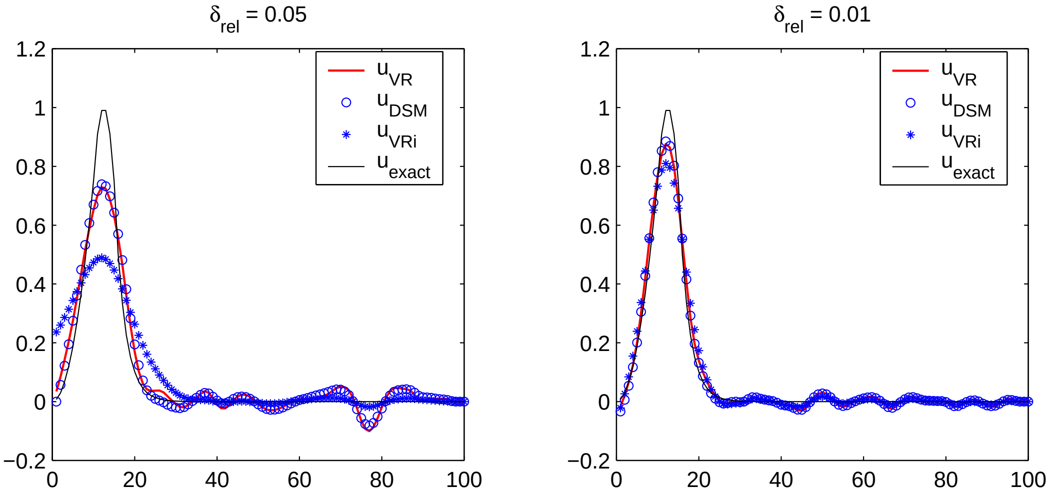

Figure 5: Plots of solutions obtained by DSM, VR for the inverse heat equation when $n = 100$, $\delta_{rel} = 0.05$ (left) and $\delta_{rel} = 0.01$ (right).

Gauss-Newton method. It can be considered as a regularized Gauss-Newton method for solving ill-conditioned linear algebraic systems. The advantage of using (54) instead of using (4.1.3) in [2] is that one only has to compute the lower upper (LU) decomposition of $A^*A + aI$ once while the algorithm in [2] requires computing LU at every step. Note that computing the LU is the main cost for solving a linear system. Numerical experiments show that the new method competes favorably with the VR in our experiments.

References

[1] Airapetyan, R. and Ramm, A. G., Dynamical systems and discrete methods for solving nonlinear ill-posed problems, *Appl. Math. Reviews*, vol. 1, Ed. G. Anastassiou, World Sci. Publishers, 2000, 491–536.

[2] Hoang, N. S. and Ramm, A. G., Solving ill-conditioned linear algebraic systems by the dynamical systems method (DSM), *Inverse Probl. Sci. Eng.*, 16 (2008), no. 5, 617–630.

[3] Hoang, N. S. and Ramm, A. G., Dynamical systems gradient method for solving ill-conditioned linear algebraic systems, *Acta Appl. Math., doi: 10.1007/s10440-009-9540-3.*

[4] Ivanov, V., Tanana, V., Vasin, V., *Theory of ill-posed problems*, VSP, Utrecht, 2002.

[5] Lattes, J., Lions, J., *Mèthode de quasi-réversibilité et applications*, Dunod, Paris, 1967.

[6] Morozov, V.A., *Methods for solving incorrectly posed problems*, Springer Verlag, New York, 1984.

[7] Ramm, A. G., *Dynamical systems method for solving operator equations*, Elsevier, Amsterdam, 2007.

[8] Ramm, A. G., Dynamical systems method for solving nonlinear operator equations, *International Jour. of Applied Math. Sci.*, 1, N1, (2004), 97-110.

[9] Ramm, A. G., Dynamical systems method for solving operator equations, *Communic. in Nonlinear Sci. and Numer. Simulation*, 9, N2, (2004), 383-402.

[10] Ramm, A. G., Discrepancy principle for the dynamical systems method, *Communic. in Nonlinear Sci. and Numer. Simulation*, 10, N1, (2005), 95-101

[11] Ramm, A. G., Dynamical systems method (DSM) and nonlinear problems, in the book: *Spectral Theory and Nonlinear Analysis*, World Scientific Publishers, Singapore, 2005, 201-228. (ed J. Lopez-Gomez).

[12] Ramm, A. G., Dynamical systems method (DSM) for unbounded operators, *Proc.Amer. Math. Soc.*, 134, N4, (2006), 1059-1063.

[13] Tautenhahn, U., On the asymptotical regularization of nonlinear ill-posed problems, *Inverse Problems*, 10 (1994) 1405-1418.

[14] Vainikko, G., Veretennikov, A., *Iterative processes in ill-posed problems*, Nauka, Moscow, 1996.

[15] Vasin, V., Ageev, A., *Ill-posed problems with a priori information*, Nauka, Ekaterinburg, 1993.

Chapter 3

A discrepancy principle for equations with monotone continuous operators

Published in Nonlinear Analysis, 70, (2009), 4307-4315.

A discrepancy principle for equations with monotone continuous operators

N. S. Hoang†* and A. G. Ramm†‡

†Mathematics Department, Kansas State University,
Manhattan, KS 66506-2602, USA

Abstract

A discrepancy principle for solving nonlinear equations with monotone operators given noisy data is formulated. The existence and uniqueness of the corresponding regularization parameter $a(\delta)$ is proved. Convergence of the solution obtained by the discrepancy principle is justified. The results are obtained under natural assumptions on the nonlinear operator.

MSC: 47J05, 47J06, 47J35, 65R30

Key words: Discrepancy principle, monotone operators, regularization, nonlinear operator equations, ill-posed problems.

1 Introduction

Consider the equation:

$$F(u) = f, \tag{1}$$

where F is a monotone operator in a real Hilbert space H. Monotonicity is understood in the following sense:

$$\langle F(u) - F(v), u - v\rangle \geq 0, \quad \forall u, v \in H. \tag{2}$$

Here $\langle \cdot, \cdot \rangle$ denotes the inner product in H. Assume that F is continuous.

Equations with monotone operators are important in many applications and were studied extensively, see, for example, [1]–[3], [9], [10], [12], and references therein. There are many technical

‡Corresponding author. Email: ramm@math.ksu.edu

*Email: nguyenhs@math.ksu.edu

and physical problems leading to equations with such operators in the cases when dissipation of energy occurs. For example, in [5] and [4], Chapter 3, pp.156-189, a wide class of nonlinear dissipative systems is studied, and the basic equations of such systems can be reduced to equation (1) with monotone operators. Many examples of equations with monotone operators can be found in [2] and in references mentioned above. In [6] and [7] it is proved that any solvable linear operator equation with a closed densely defined operator in a Hilbert space H can be reduced to an equation with a monotone operator and solved by a convergent iterative process.

In this paper, apparently for the first time, a discrepancy principle for solving equation (3) with noisy data (see Section 2) is proved under natural assumptions. No smallness assumptions on the nonlinearity, no global restrictions on its growth, or other special properties of the nonlinearity, except the monotonicity and continuity, are imposed. No source-type assumptions are used. Our result is widely applicable. It is well known that without extra assumptions, usually source-type assumption concerning the right-hand side, or some equivalent assumption concerning the smoothness of the solution, one cannot get a rate of convergence even for linear ill-posed equations (see, for example, [9]). On the other hand, such assumptions are usually not algorithmically verifiable and often they do not hold. By this reason we do not make such assumptions and do not give estimates of the rate of convergence.

In [11] a stationary equation $F(u) = f$ with a nonlinear monotone operator F was studied. The assumptions A1-A3 on p.197 in [11] are more restrictive than ours, and the Rule R2 on p.199, formula (4.1) in [11], for the choice of the regularization parameter is more difficult to use computationally: one has to solve nonlinear equation (4.1) in [11] for the regularization parameter. Moreover, to use this equation one has to invert an ill-conditioned linear operator $A + aI$ for small values of a. Assumption A1 in [11] is not verifiable, because the solution $x^\dagger$ is not known. Assumption A3 in [11] requires F to be constant in a ball $B_r(x^\dagger)$ if $F'(x^\dagger) = 0$. Our discrepancy principle does not require these assumptions, and, in contrast to equation (4.1) in [11], it does not require inversion of ill-conditioned linear operators.

The novel results in our paper include Theorem 5 in Section 3 and Theorem 7 in Section 4. In Theorem 5 a new discrepancy principle is proposed and justified assuming only the monotonicity and continuity of F. Implementing the discrepancy principle in Theorem 5 requires solving equation (3) and then solving nonlinear equation (15) for the regularization parameter $a(\delta)$. Theorem 7 allows one to solve equations (3) and (15) approximately. Thus, when δ is not too small one can save a large amount of computations in solving equations (3) and (15) by applying Theorem 7 and using our new stopping rule. Our results allow one to solve numerically stably equation (1) if F is locally Lipschitz and monotone. Based on Theorem 7, an algorithm for stable solution of equation (1) is formulated for locally Lipschitz monotone operators.

2 Auxiliary results

Let us consider the following equation

$$F(V_{\delta,a}) + aV_{\delta,a} - f_\delta = 0, \qquad a > 0, \tag{3}$$

where $a = const.$ It is known (see, e.g., [9, p.111]) that equation (3) with monotone continuous operator F has a unique solution for any $f_\delta \in H$.

Throughout the paper we assume that F is a monotone continuous operator and the inner product in H is denoted $\langle u, v\rangle$. Below the word decreasing means strictly decreasing and increasing means strictly increasing.

Recall the following result from [9, p.112]:

Lemma 1 *Assume that equation* (1) *is solvable, y is its minimal-norm solution, assumption* (2) *holds, and F is continuous. Then*

$$\lim_{a\to 0} \|V_a - y\| = 0, \tag{4}$$

where V_a solves equation (3) *with $\delta = 0$.*

Lemma 2 *Assume $\|F(0) - f_\delta\| > 0$. Let $a > 0$, and F be monotone. Denote*

$$\psi(a) := \|V_{\delta,a}\|, \qquad \phi(a) := a\psi(a) = \|F(V_{\delta,a}) - f_\delta\|,$$

where $V_{\delta,a}$ solves (3)*. Then $\psi(a)$ is decreasing, and $\phi(a)$ is increasing.*

Proof. Since $\|F(0) - f_\delta\| > 0$, one has $\psi(a) \neq 0$, $\forall a \geq 0$. Indeed, if $\psi(a)\big|_{a=\tau} = 0$, then $V_{\delta,a} = 0$, and equation (3) implies $\|F(0) - f_\delta\| = 0$, which is a contradiction. Note that $\phi(a) = a\|V_{\delta,a}\|$. One has

$$\begin{aligned} 0 &\leq \langle F(V_{\delta,a}) - F(V_{\delta,b}), V_{\delta,a} - V_{\delta,b}\rangle \\ &= \langle -aV_{\delta,a} + bV_{\delta,b}, V_{\delta,a} - V_{\delta,b}\rangle \\ &= (a+b)\langle V_{\delta,a}, V_{\delta,b}\rangle - a\|V_{\delta,a}\|^2 - b\|V_{\delta,b}\|^2. \end{aligned} \tag{5}$$

Thus,

$$\begin{aligned} 0 &\le (a+b)\langle V_{\delta,a}, V_{\delta,b}\rangle - a\|V_{\delta,a}\|^2 - b\|V_{\delta,b}\|^2 \\ &\le (a+b)\|V_{\delta,a}\|\|V_{\delta,b}\| - a\|V_{\delta,a}\|^2 - b\|V_{\delta,b}\|^2 \\ &= (a\|V_{\delta,a}\| - b\|V_{\delta,b}\|)(\|V_{\delta,b}\| - \|V_{\delta,a}\|) \\ &= (\phi(a) - \phi(b))(\psi(b) - \psi(a)). \end{aligned} \tag{6}$$

If $\psi(b) > \psi(a)$ then (6) implies $\phi(a) \ge \phi(b)$, so

$$a\psi(a) \ge b\psi(b) > b\psi(a).$$

Therefore, if $\psi(b) > \psi(a)$ then $b < a$.

Similarly, if $\psi(b) < \psi(a)$ then $\phi(a) \le \phi(b)$. This implies $b > a$.

Suppose $\psi(a) = \psi(b)$, i.e., $\|V_{\delta,a}\| = \|V_{\delta,b}\|$. From (5) one has

$$\|V_{\delta,a}\|^2 \le \langle V_{\delta,a}, V_{\delta,b}\rangle \le \|V_{\delta,a}\|\|V_{\delta,b}\| = \|V_{\delta,a}\|^2.$$

This implies $V_{\delta,a} = V_{\delta,b}$, and then equation (3) implies $a = b$.

Therefore ϕ is increasing and ψ is decreasing. □

Lemma 3 *If F is monotone and continuous, then $\|V_{\delta,a}\| = O(\frac{1}{a})$ as $a \to \infty$, and*

$$\lim_{a\to\infty} \|F(V_{\delta,a}) - f_\delta\| = \|F(0) - f_\delta\|. \tag{7}$$

Proof. Rewrite (3) as

$$F(V_{\delta,a}) - F(0) + aV_{\delta,a} + F(0) - f_\delta = 0.$$

Multiply this equation by $V_{\delta,a}$, use the monotonicity of F and get:

$$a\|V_{\delta,a}\|^2 \le \langle aV_{\delta,a} + F(V_{\delta,a}) - F(0), V_{\delta,a}\rangle = \langle f_\delta - F(0), V_{\delta,a}\rangle \le \|f_\delta - F(0)\|\|V_{\delta,a}\|.$$

Therefore, $\|V_{\delta,a}\| = O(\frac{1}{a})$. This and the continuity of F imply (7). □

Remark 1 If $\|F(0) - f_\delta\| > C\delta^\gamma, 0 < \gamma \le 1$ then relation (7) implies

$$\|F(V_{\delta,a}) - f_\delta\| \ge C\delta^\gamma, \qquad 0 < \gamma \le 1, \tag{8}$$

for sufficiently large $a > 0$.

Lemma 4 *Let $C > 0$ and $\gamma \in (0,1]$ be constants such that $C\delta^\gamma > \delta$. Suppose that $\|F(0) - f_\delta\| > C\delta^\gamma$. Then, there exists a unique $a(\delta) > 0$ such that $\|F(V_{\delta,a(\delta)}) - f_\delta\| = C\delta^\gamma$.*

Proof. We have $F(y) = f$, and

$$\begin{aligned}0 =&\langle F(V_{\delta,a}) + aV_{\delta,a} - f_\delta, F(V_{\delta,a}) - f_\delta\rangle\\ =&\|F(V_{\delta,a}) - f_\delta\|^2 + a\langle V_{\delta,a} - y, F(V_{\delta,a}) - f_\delta\rangle + a\langle y, F(V_{\delta,a}) - f_\delta\rangle\\ =&\|F(V_{\delta,a}) - f_\delta\|^2 + a\langle V_{\delta,a} - y, F(V_{\delta,a}) - F(y)\rangle + a\langle V_{\delta,a} - y, f - f_\delta\rangle\\ &+ a\langle y, F(V_{\delta,a}) - f_\delta\rangle\\ \ge&\|F(V_{\delta,a}) - f_\delta\|^2 + a\langle V_{\delta,a} - y, f - f_\delta\rangle + a\langle y, F(V_{\delta,a}) - f_\delta\rangle.\end{aligned}$$

Here the monotonicity of F was used. Therefore

$$\begin{aligned}\|F(V_{\delta,a}) - f_\delta\|^2 &\le -a\langle V_{\delta,a} - y, f - f_\delta\rangle - a\langle y, F(V_{\delta,a}) - f_\delta\rangle\\ &\le a\|V_{\delta,a} - y\|\|f - f_\delta\| + a\|y\|\|F(V_{\delta,a}) - f_\delta\|\\ &\le a\delta\|V_{\delta,a} - y\| + a\|y\|\|F(V_{\delta,a}) - f_\delta\|.\end{aligned} \tag{9}$$

Also,

$$\begin{aligned}0 &= \langle F(V_{\delta,a}) - F(y) + aV_{\delta,a} + f - f_\delta, V_{\delta,a} - y\rangle\\ &= \langle F(V_{\delta,a}) - F(y), V_{\delta,a} - y\rangle + a\|V_{\delta,a} - y\|^2 + a\langle y, V_{\delta,a} - y\rangle + \langle f - f_\delta, V_{\delta,a} - y\rangle\\ &\ge a\|V_{\delta,a} - y\|^2 + a\langle y, V_{\delta,a} - y\rangle + \langle f - f_\delta, V_{\delta,a} - y\rangle,\end{aligned}$$

where the monotonicity of F was used again. Therefore,

$$a\|V_{\delta,a} - y\|^2 \le a\|y\|\|V_{\delta,a} - y\| + \delta\|V_{\delta,a} - y\|.$$

This implies

$$a\|V_{\delta,a} - y\| \le a\|y\| + \delta. \tag{10}$$

From (9), (10), and an elementary inequality $ab \le \epsilon a^2 + \frac{b^2}{4\epsilon}$, $\forall\epsilon > 0$, one gets:

$$\begin{aligned}\|F(V_{\delta,a}) - f_\delta\|^2 &\le \delta^2 + a\|y\|\delta + a\|y\|\|F(V_{\delta,a}) - f_\delta\|\\ &\le \delta^2 + a\|y\|\delta + \epsilon\|F(V_{\delta,a}) - f_\delta\|^2 + \frac{1}{4\epsilon}a^2\|y\|^2,\end{aligned} \tag{11}$$

where $\epsilon > 0$ is arbitrary small, fixed, independent of a, and can be chosen arbitrary small. Let $a \searrow 0$. Then (11) implies $\lim_{a\to 0}(1-\epsilon)\|F(V_{\delta,a}) - f_\delta\|^2 \le \delta^2 < (C\delta^\gamma)^2$. Thus,

$$\lim_{a\to 0}\|F(V_{\delta,a}) - f_\delta\| < C\delta^\gamma, \qquad C > 0, \quad 0 < \gamma \le 1.$$

This, the continuity of F, the continuity of $V_{\delta,a}$ with respect to $a \in [0,\infty)$, and inequality (8), imply that equation $\|F(V_{\delta,a}) - f_\delta\| = C\delta^\gamma$ must have a solution $a(\delta) > 0$. □

Remark 2 Let $V_a := V_{\delta,a}|_{\delta=0}$, so $F(V_a) + aV - f = 0$. Let y be the minimal-norm solution to equation (1). We claim that

$$\|V_{\delta,a} - V_a\| \le \frac{\delta}{a}. \tag{12}$$

Indeed, from (3) one gets

$$F(V_{\delta,a}) - F(V_a) + a(V_{\delta,a} - V_a) = f - f_\delta.$$

Multiply this equality by $(V_{\delta,a} - V_a)$ and use (2) to obtain

$$\begin{aligned}\delta\|V_{\delta,a} - V_a\| &\ge \langle f - f_\delta, V_{\delta,a} - V_a\rangle \\ &= \langle F(V_{\delta,a}) - F(V_a) + a(V_{\delta,a} - V_a), V_{\delta,a} - V_a)\rangle \\ &\ge a\|V_{\delta,a} - V_a\|^2.\end{aligned}$$

This implies (12).

Let us derive a uniform with respect to a bound on $\|V_a\|$. From the equation

$$F(V_a) + aV_a - F(y) = 0,$$

and the monotonicity of F one gets

$$0 = \langle F(V_a) + aV_a - F(y), V_a - y\rangle \ge a\langle V_a, V_a - y\rangle.$$

This implies the desired bound:

$$\|V_a\| \le \|y\|, \qquad \forall a > 0. \tag{13}$$

Similar arguments one can find in [9, p. 113].

From (12) and (13), one gets the following estimate:

$$\|V_{\delta,a}\| \le \|V_a\| + \frac{\delta}{a} \le \|y\| + \frac{\delta}{a}. \tag{14}$$

3 A discrepancy principle

Our standing assumptions are the monotonicity and continuity of F and the solvability of equation (1). They are not repeated below. We assume without loss of generality that $\delta \in (0, 1)$.

Theorem 5 *Let $\gamma \in (0, 1]$ and $C > 0$ be some constants such that $C\delta^\gamma > \delta$. Assume that $\|F(0) - f_\delta\| > C\delta^\gamma$. Let y be its minimal-norm solution. Then there exists a unique $a(\delta) > 0$ such that*

$$\|F(V_{\delta,a(\delta)}) - f_\delta\| = C\delta^\gamma, \tag{15}$$

where $V_{\delta,a(\delta)}$ solves (3) *with $a = a(\delta)$.*

If $0 < \gamma < 1$ then

$$\lim_{\delta\to 0} \|V_{\delta,a(\delta)} - y\| = 0. \tag{16}$$

Proof. The existence and uniqueness of $a(\delta)$ follow from Lemma 4. Let us show that

$$\lim_{\delta\to 0} a(\delta) = 0. \tag{17}$$

The triangle inequality, inequality (12) and equality (15) imply

$$\begin{aligned} a(\delta)\|V_{a(\delta)}\| &\leq a(\delta)\big(\|V_{\delta,a(\delta)} - V_{a(\delta)}\| + \|V_{\delta,a(\delta)}\|\big) \\ &\leq \delta + a(\delta)\|V_{\delta,a(\delta)}\| = \delta + C\delta^\gamma. \end{aligned} \tag{18}$$

From inequality (18), one gets

$$\lim_{\delta\to 0} a(\delta)\|V_{a(\delta)}\| = 0. \tag{19}$$

It follows from Lemma 2 with $f_\delta = f$, i.e., $\delta = 0$, that the function $\phi_0(a) := a\|V_a\|$ is nonnegative and strictly increasing on $(0,\infty)$. This and relation (19) imply:

$$\lim_{\delta\to 0} a(\delta) = 0. \tag{20}$$

From (15) and (14), one gets

$$C\delta^\gamma = a\|V_{\delta,a}\| \leq a(\delta)\|y\| + \delta. \tag{21}$$

Thus, one gets:

$$C\delta^\gamma - \delta \leq a(\delta)\|y\|. \tag{22}$$

If $\gamma < 1$ then $C - \delta^{1-\gamma} > 0$ for sufficiently small δ. This implies:

$$0 \leq \lim_{\delta\to 0} \frac{\delta}{a(\delta)} \leq \lim_{\delta\to 0} \frac{\delta^{1-\gamma}\|y\|}{C - \delta^{1-\gamma}} = 0. \tag{23}$$

By the triangle inequality and inequality (12), one has

$$\|V_{\delta,a(\delta)} - y\| \leq \|V_{a(\delta)} - y\| + \|V_{a(\delta)} - V_{\delta,a(\delta)}\| \leq \|V_{a(\delta)} - y\| + \frac{\delta}{a(\delta)}. \tag{24}$$

Relation (16) follows from (23), (24) and Lemma 1. □

Instead of using (3), one may use the following equation:

$$F(V_{\delta,a}) + a(V_{\delta,a} - \bar{u}) - f_\delta = 0, \qquad a > 0, \tag{25}$$

where $\bar{u}$ is an element of H. Denote $F_1(u) := F(u+\bar{u})$. Then F_1 is monotone and continuous. Equation (3) can be written as:

$$F_1(U_{\delta,a}) + aU_{\delta,a} - f_\delta = 0, \qquad U_{\delta,a} := V_{\delta,a} - \bar{u}, \quad a > 0. \tag{26}$$

By applying Theorem 5 with $F = F_1$ one gets the following result:

Corollary 6 *Let $\gamma \in (0,1]$ and $C > 0$ be some constants such that $C\delta^\gamma > \delta$. Let $\bar{u} \in H$ and z be the solution to* (1) *with minimal distance to $\bar{u}$. Assume that $\|F(\bar{u}) - f_\delta\| > C\delta^\gamma$. Then there exists a unique $a(\delta) > 0$ such that*

$$\|F(\tilde{V}_{\delta,a(\delta)}) - f_\delta\| = C\delta^\gamma, \tag{27}$$

where $\tilde{V}_{\delta,a(\delta)}$ solves the following equation:

$$F(\tilde{V}_{\delta,a}) + a(\delta)(\tilde{V}_{\delta,a} - \bar{u}) - f_\delta = 0.$$

If $\gamma \in (0,1)$ then this $a(\delta)$ satisfies

$$\lim_{\delta\to 0} \|\tilde{V}_{\delta,a(\delta)} - z\| = 0. \tag{28}$$

Remark 3 It is an open problem to choose γ and C optimal in some sense.

Remark 4 Theorem 5 and Theorem 7 do not hold, in general, for $\gamma = 1$. Indeed, let $Fu = \langle u, p\rangle p$, $\|p = 1\|$, $p \perp \mathcal{N}(F) := \{u \in H : Fu = 0\}$, $f = p$, $f_\delta = p + q\delta$, where $\langle p, q\rangle = 0$, $\|q\| = 1$, $Fq = 0$, $\|q\delta\| = \delta$. One has $Fy = p$, where $y = p$, is the minimal-norm solution to the equation $Fu = p$. Equation $Fu + au = p + q\delta$, has the unique solution $V_{\delta,a} = q\delta/a + p/(1+a)$. Equation (15) is $C\delta = \|q\delta + (ap)/(1+a)\|$. This equation yields $a = a(\delta) = c\delta/(1-c\delta)$, where $c := (C^2-1)^{1/2}$, and we assume $c\delta < 1$. Thus, $\lim_{\delta\to 0} V_{\delta,a(\delta)} = p + c^{-1}q := v$, and $Fv = p$. Therefore $v = \lim_{\delta\to 0} V_{\delta,a(\delta)}$ is not p, i.e., is not the minimal-norm solution to the equation $Fu = p$. Similar arguments one can find in [8, p. 29].

4 Applications

In this section we discuss methods for solving equations (3) and (1) using the new discrepancy principle, i.e., Theorem 5. Implementing this principle, i.e., solving equation (15), requires solving equation (3). *If F is linear*, then equation (3) has the form:

$$(F + aI)u = f_\delta. \tag{29}$$

Since $F \geq 0$ the operator $F + aI$ is boundedly invertible, $\|(F + aI)^{-1}\| \leq \frac{1}{a}$, and equation (29) is well-posed if $a > 0$ is not too small. There are many methods for solving efficiently well-posed linear equations with positive-definite operators. For this reason we *mainly discuss some methods for stable solution of equation* (1) *with nonlinear operators.* In this section a method is developed for a stable solution of equation (1) with locally Lipschitz monotone operator F, so we assume that

$$\|F(u) - F(v)\| \leq L\|u - v\|, \quad u, v \in B(u_0, R) := \{u : \|u - u_0\| \leq R\}, \quad L = L(R). \tag{30}$$

Here $u_0 \in H$ is an arbitrary fixed element. Consider the operator

$$G(u) := u - \lambda[F(u) + au - f_\delta], \quad \lambda > 0.$$

We claim that G is a contraction mapping in H provided that λ is sufficiently small. Let $F_1 := F + aI$. Then (30) implies $\|F_1(u) - F_1(v)\| \leq (a + L)\|u - v\|$. Using the monotonicity of F, one gets

$$\begin{aligned} \|G(u) - G(v)\|^2 &= \|(u - v) - \lambda\big(F_1(u) - F_1(v)\big)\|^2 \\ &= \|u - v\|^2 - 2\lambda\langle u - v, F_1(u) - F_1(v)\rangle + \lambda^2\|F_1(u) - F_1(v)\|^2 \\ &\leq \|u - v\|^2[1 - 2\lambda a + \lambda^2(a + L)^2]. \end{aligned} \tag{31}$$

This implies that G is a contraction mapping if

$$0 < \lambda < \frac{2a}{(a + L)^2}.$$

For these λ the solution $V_{\delta,a}$ of equation (3) can be found by the following iterative process:

$$u_{n+1} = u_n - \lambda[F(u_n) + au_n - f_\delta], \quad u_0 := u_0. \tag{32}$$

After finding $V_{\delta,a}$, one finds $a(\delta)$ from the discrepancy principle (15), i.e., by solving the nonlinear equation:

$$\phi(a(\delta)) := \|F(V_{\delta,a(\delta)}) - f_\delta\| = C\delta^\gamma. \tag{33}$$

There are many methods for solving this equation. For example, one can use the bisection method or the golden section method. If $a(\delta)$ is found, one solves equation (3) with $a = a(\delta)$ for $V_{\delta,a(\delta)}$ and takes its solution as an approximate solution to (1).

Although the sequence u_n, defined by (32), converges to the solution of equation (3) at the rate of a geometrical series with a denominator $q \in (0, 1)$, it is very time consuming to try to solve equation (3) with high accuracy if q is close to 1. Theorem 7 (see below) allows one to stop iterations (32) at the first value of n which satisfies the following condition:

$$\|F(u_n) + au_n - f_\delta\| \le \theta\delta, \qquad \theta > 0, \tag{34}$$

where θ is a fixed constant. This saves the time of computation.

Theorem 7 *Let δ, F, f_δ, and y be as in Theorem 5 and $0 < \gamma < 1$. Assume that $v_\delta \in H$ and $\alpha(\delta) > 0$ satisfy the following conditions:*

$$\|F(v_\delta) + \alpha(\delta)v_\delta - f_\delta\| \le \theta\delta, \qquad \theta > 0, \tag{35}$$

and

$$C_1\delta^\gamma \le \|F(v_\delta) - f_\delta\| \le C_2\delta^\gamma, \qquad 0 < C_1 < C_2. \tag{36}$$

Then one has:

$$\lim_{\delta\to 0} \|v_\delta - y\| = 0. \tag{37}$$

Proof. Let u and v be arbitrary elements in H. By the monotonicity of F one gets

$$\begin{aligned} a\|u - v\|^2 &\le \langle u - v, F(u) - F(v) + au - av\rangle \\ &\le \|u - v\|\|F(u) - F(v) + au - av\|, \qquad \forall a > 0. \end{aligned} \tag{38}$$

This implies

$$a\|u - v\| \le \|F(u) - F(v) + au - av\|, \qquad \forall v, u \in H, \quad \forall a > 0. \tag{39}$$

Using inequality (39) with $v = v_\delta$ and $u = V_{\delta,\alpha(\delta)}$, equation (3) with $a = \alpha(\delta)$, and inequality (35), one gets

$$\begin{aligned} \alpha(\delta)\|v_\delta - V_{\delta,\alpha(\delta)}\| &\le \|F(v_\delta) - F(V_{\delta,\alpha(\delta)}) + \alpha(\delta)v_\delta - \alpha(\delta)V_{\delta,\alpha(\delta)}\| \\ &= \|F(v_\delta) + \alpha(\delta)v_\delta - f_\delta\| \le \theta\delta. \end{aligned} \tag{40}$$

Therefore,

$$\|v_\delta - V_{\delta,\alpha(\delta)}\| \le \frac{\theta\delta}{\alpha(\delta)}. \tag{41}$$

Using (14) and (41), one gets:

$$\alpha(\delta)\|v_\delta\| \le \alpha(\delta)\|V_{\delta,\alpha(\delta)}\| + \alpha(\delta)\|v_\delta - V_{\delta,\alpha(\delta)}\| \le \theta\delta + \alpha(\delta)\|y\| + \delta. \tag{42}$$

From the triangle inequality and inequalities (35) and (36) one obtains:

$$\alpha(\delta)\|v_\delta\| \ge \|F(v_\delta) - f_\delta\| - \|F(v_\delta) + \alpha(\delta)v_\delta - f_\delta\| \ge C_1\delta^\gamma - \theta\delta. \tag{43}$$

Inequalities (42) and (43) imply

$$C_1\delta^\gamma - \theta\delta \le \theta\delta + \alpha(\delta)\|y\| + \delta. \tag{44}$$

This inequality and the fact that $C_1 - \delta^{1-\gamma} - 2\theta\delta^{1-\gamma} > 0$ for sufficiently small δ and $0 < \gamma < 1$ imply

$$\frac{\delta}{\alpha(\delta)} \le \frac{\delta^{1-\gamma}\|y\|}{C_1 - \delta^{1-\gamma} - 2\theta\delta^{1-\gamma}}, \qquad 0 < \delta \ll 1. \tag{45}$$

Thus, one obtains

$$\lim_{\delta\to 0} \frac{\delta}{\alpha(\delta)} = 0. \tag{46}$$

From the triangle inequality and inequalities (35), (36) and (41), one gets

$$\begin{aligned} \alpha(\delta)\|V_{\delta,\alpha(\delta)}\| &\le \|F(v_\delta) - f_\delta\| + \|F(v_\delta) + \alpha(\delta)v_\delta - f_\delta\| + \alpha(\delta)\|v_\delta - V_{\delta,\alpha(\delta)}\| \\ &\le C_2\delta^\gamma + \theta\delta + \theta\delta. \end{aligned} \tag{47}$$

This inequality implies

$$\lim_{\delta\to 0} \alpha(\delta)\|V_{\delta,\alpha(\delta)}\| = 0. \tag{48}$$

The triangle inequality and inequality (12) imply

$$\begin{aligned} \alpha\|V_\alpha\| &\le \alpha\big(\|V_{\delta,\alpha} - V_\alpha\| + \|V_{\delta,\alpha}\|\big) \\ &\le \delta + \alpha\|V_{\delta,\alpha}\|. \end{aligned} \tag{49}$$

From formulas (49) and (48), one gets

$$\lim_{\delta\to 0} \alpha(\delta)\|V_{\alpha(\delta)}\| = 0. \tag{50}$$

It follows from Lemma 2 with $f_\delta = f$, i.e., $\delta = 0$, that the function $\phi_0(a) := a\|V_a\|$ is nonnegative and strictly increasing on $(0, \infty)$. This and relation (50) imply

$$\lim_{\delta\to 0} \alpha(\delta) = 0. \tag{51}$$

From the triangle inequality and inequalities (41) and (12) one obtains

$$\begin{aligned} \|v_\delta - y\| &\le \|v_\delta - V_{\delta,\alpha(\delta)}\| + \|V_{\delta,\alpha(\delta)} - V_{\alpha(\delta)}\| + \|V_{\alpha(\delta)} - y\| \\ &\le \frac{\theta\delta}{\alpha(\delta)} + \frac{\delta}{\alpha(\delta)} + \|V_{\alpha(\delta)} - y\|, \end{aligned} \tag{52}$$

where $V_{\alpha(\delta)}$ solves equation (3) with $a = \alpha(\delta)$ and $f_\delta = f$.

The conclusion (37) follows from inequalities (46), (51), (52) and Lemma 1. Theorem 7 is proved. □

Remark 5 Inequalities (35) and (36) are used as stopping rules for finding approximations:

$$\alpha(\delta) \approx a(\delta), \quad \text{and} \quad v(\delta) \approx V_{\delta,a(\delta)}.$$

Remark 6 By the monotonicity of F one gets

$$\begin{aligned}\|F(u) - F(v)\|^2 &\le \langle F(u) - F(v), F(u) - F(v) + a(u - v)\rangle \\ &\le \|F(u) - F(v)\|\|F(u) - F(v) + a(u - v)\|, \quad \forall u, v \in H, \quad \forall a > 0.\end{aligned}$$

This implies

$$\|F(u) - F(v)\| \le \|F(u) - F(v) + a(u - v)\|, \qquad \forall u, v \in H, \quad a > 0. \tag{53}$$

Fix $\delta > 0$ and $\theta > 0$. Let C be as in Theorem 5. Choose C_1 and C_2 such that

$$C_1\delta^\gamma + \theta\delta < C\delta^\gamma < C_2\delta^\gamma - \theta\delta. \tag{54}$$

Suppose α_i and v_i, $i = 1, 2$, satisfy condition (35) and

$$\|F(v_1) - f_\delta\| < C_1\delta^\gamma, \qquad C_2\delta^\gamma < \|F(v_2) - f_\delta\|. \tag{55}$$

Let us show that

$$\alpha_{low} := \alpha_1 < a(\delta) < \alpha_2 := \alpha_{up}, \tag{56}$$

where $a(\delta)$ satisfies conditions of Theorem 5. Using inequality (53) for v_i and V_{δ,α_i}, $i = 1, 2$, and inequality (35), one gets

$$\begin{aligned}\|F(v_i) - F(V_{\delta,\alpha_i})\| &\le \|F(v_i) - F(V_{\delta,\alpha_i}) + \alpha_i v_i - \alpha_i V_{\delta,\alpha_i}\| \\ &\le \|F(v_i) + \alpha_i v_i - f_\delta\| \le \theta\delta.\end{aligned} \tag{57}$$

From inequalities (55), (57) and the triangle inequality, one derives:

$$\|F(V_{\delta,\alpha_1}) - f_\delta\| < C_1\delta^\gamma + \theta\delta \quad \text{and} \quad C_2\delta^\gamma - \theta\delta < \|F(V_{\delta,\alpha_2}) - f_\delta\|. \tag{58}$$

Recall that $\|F(V_{\delta,a(\delta)}) - f_\delta\| = C\delta^\gamma$. Inequality (56) is obtained from inequalities (54), (58) and the fact that the function $\phi(\alpha) = \|F(V_{\delta,\alpha}) - f_\delta\|$ is strictly increasing (see Lemma 2).

Let $f_\delta, F, C, \theta, \gamma$, and δ be as in Theorem 5 and 7, and C_1 and C_2 satisfy inequality (54). Let us formulate an algorithm (see **Algorithm 1** below) for finding $\alpha(\delta) \approx a(\delta)$ and $v(\delta) \approx V_{\delta,a(\delta)}$, using the bisection method and assuming that F is a locally Lipschitz monotone operator and α_{low} and α_{up} are known. By Theorem 7, $v(\delta)$ can be considered as a stable solution to equation (1).

Algorithm 1: *Finding $\alpha(\delta) \approx a(\delta)$ and $v_\delta \approx V_{\delta,a(\delta)}$ given α_{low} and α_{up}.*

1. Let $a := \frac{\alpha_{up}+\alpha_{low}}{2}$ and u_0 be an initial guess for $V_{\delta,a}$. Compute u_n by formula (32) and stop at n_{stop}, where n_{stop} is the smallest $n > 0$ for which condition (35) is satisfied. Then go to step 2.

2. If $C_2\delta^\gamma < \|F(u_{n_{stop}}) - f_\delta\|$, then set $\alpha_{up} := a$ and go to step 4. Otherwise, go to step 3.

3. If $C_1\delta^\gamma \leq \|F(u_{n_{stop}}) - f_\delta\|$, then stop the process and take $v(\delta) := u_{n_{stop}}$ as a solution to (1). If $\|F(u_{n_{stop}}) - f_\delta\| < C_1\delta^\gamma$, then set $\alpha_{low} := a$ and go to step 4.

4. Check if $\|a - \alpha_{low}\|$ is less than a desirable small value $\epsilon > 0$. If it is, then take $v(\delta) := u_{n_{stop}}$ as a solution to (1). If is is not, then go back to step 1.

Let us formulate algorithms for finding α_{up} and α_{low}.

Algorithm 2: *Finding* α_{up}.

1. Let $a = \alpha$ be an initial guess for $\alpha(\delta)$ and u_0 be an initial guess for v_δ. Compute u_n by formula (32) with a and stop at n_{stop}, the smallest $n > 0$ for which condition (35) is satisfied. Then go to step 2.

2. If condition (36) holds for $v_\delta := u_{n_{stop}}$, then stop the process and take $u_{n_{stop}}$ as a solution to (1). Otherwise, go to step 3.

3. If $C_2\delta^\gamma < \|F(u_{n_{stop}}) - f_\delta\|$, then set $\alpha_{up} := a$. Otherwise, set $\alpha := 2a$ and go back step 1.

Algorithm 3: *Finding* α_{low}.

1. Let $a = \alpha$ be an initial guess for $\alpha(\delta)$ and u_0 be an initial guess for v_δ. Compute u_n by formula (32) with a and stop at n_{stop}, the smallest $n > 0$ for which condition (35) is satisfied. Then go to step 2.

2. If condition (36) holds for $v_\delta := u_{n_{stop}}$, then stop the process and take $u_{n_{stop}}$ as a solution to (1). Otherwise, go to step 3.

3. If $\|F(u_{n_{stop}}) - f_\delta\| < C_1\delta^\gamma$, then set $\alpha_{low} := a$. Otherwise, set $\alpha := \frac{a}{2}$ and go back step 1.

In practice these algorithms are often implemented at the same time to avoid repetition calculations.

Remark 7 The sequence $(\|u_n - V_{\delta,a(\delta)}\|)_{n=0}^\infty$, where u_n is computed by formula (32) and $V_{\delta,a(\delta)}$ is the solution to (3) with $a = a(\delta)$, is decreasing. Thus, the sequence u_n will stay inside a ball $B(0, R)$ assuming that $R > 0$ is chosen sufficiently large, so that $y, u_0 \in B(0, R)$.

Remark 8 Theorem 7 and the above algorithms are not only useful for solving nonlinear equations with monotone operators but also for solving linear equations with monotone operators. If one uses iterative methods to solve equation (29) then, by using Theorem 7, one can stop iterations whenever inequality (35) holds. By using stopping rule (35) one saves time of computations compared to solving (29) exactly. If F is a positive matrix then one can solve (29) by conjugate gradient, or Jacobi, or Gauss-Seidel, or successive over-relaxation methods, with stopping rule (35).

Remark 9 If F is twice Fréchet differentiable, there are more options for solving equations (3) and (33): they can be solved by gradient-type methods, Newton-type methods, or a combination of these methods.

References

[1] K. Deimling, *Nonlinear functional analysis*, Springer Verlag, Berlin, 1985.

[2] J. L. Lions, *Quelques methodes de resolution des problemes aux limites non lineaires*, Dunod, Gauthier-Villars, Paris, 1969.

[3] D. Pascali and S. Sburlan, *Nonlinear Mappings of Monotone Type*, Noordhoff, Leyden, 1978.

[4] A. G. Ramm, *Theory and applications of some new classes of integral equations*, Springer-Verlag, New York, 1980.

[5] A. G. Ramm, Stationary regimes in passive nonlinear networks, in the book *Nonlinear Electromagnetics*, Ed. P.Uslenghi, Acad. Press, New York, 1980, pp. 263-302.

[6] A. G. Ramm, Iterative solution of linear equations with unbounded operators, *J. Math. Anal. Appl.*, 1338-1346.

[7] A. G. Ramm, On unbounded operators and applications, *Appl. Math. Lett.*, 21, (2008), 377-382.

[8] A. G. Ramm, *Inverse problems*, Springer, New York, 2005.

[9] A. G. Ramm, *Dynamical systems method for solving operator equations*, Elsevier, Amsterdam, 2007.

[10] I. V. Skrypnik, *Methods for Analysis of Nonlinear Elliptic Boundary Value Problems*, American Mathematical Society, Providence, RI, 1994.

[11] U. Tautenhahn, On the method of Lavrentiev regularization for nonlinear ill-posed problems, *Inverse Probl.*, 18, (2002), 191-207.

[12] M. M. Vainberg, *Variational methods and method of monotone operators in the theory of nonlinear equations*, Wiley, London, 1973.

Chapter 4

Dynamical systems method for solving non-linear equations with monotone operators

Published in Math. Comp., 79, (2010), 239–258

MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

DYNAMICAL SYSTEMS METHOD FOR SOLVING NONLINEAR EQUATIONS WITH MONOTONE OPERATORS

N. S. HOANG AND A. G. RAMM

Abstract. A version of the Dynamical Systems Method (DSM) for solving ill-posed nonlinear equations with monotone operators in a Hilbert space is studied in this paper. An *a posteriori* stopping rule, based on a discrepancy-type principle is proposed and justified mathematically. The results of two numerical experiments are presented. They show that the proposed version of DSM is numerically efficient. The numerical experiments consist of solving nonlinear integral equations.

1. Introduction

In this paper we study a Dynamical Systems Method (DSM) for solving the equation

$$F(u) = f, \tag{1.1}$$

where F is a nonlinear twice Fréchet differentiable monotone operator in a real Hilbert space H, and equation (1.1) is assumed solvable. Monotonicity means that

$$\langle F(u) - F(v), u - v\rangle \geq 0, \quad \forall u, v \in H. \tag{1.2}$$

Here, $\langle \cdot, \cdot\rangle$ denotes the inner product in H. It is known (see, e.g., [8]), that the set $\mathcal{N} := \{u : F(u) = f\}$ is closed and convex if F is monotone and continuous. A closed and convex set in a Hilbert space has a unique minimal-norm element. This element in $\mathcal{N}$ we denote y, $F(y) = f$. We assume that

$$\sup_{\|u-u_0\|\leq R} \|F^{(j)}(u)\| \leq M_j(u_0, R), \quad 0 \leq j \leq 2, \tag{1.3}$$

where $u_0 \in H$ is an element of H, $R > 0$ is arbitrary, and $f = F(y)$ is not known; but f_δ, the noisy data, are known and $\|f_\delta - f\| \leq \delta$. If $F'(u)$ is not boundedly invertible, then solving for u given noisy data f_δ is often (but not always) an ill-posed problem. When F is a linear bounded operator

Received by the editor April 3, 2008, and in revised form, January 7, 2009.

2000 *Mathematics Subject Classification.* Primary: 65R30; Secondary: 47J05, 47J06, 47J35,

Key words and phrases. Dynamical systems method (DSM), nonlinear operator equations, monotone operators, discrepancy principle.

many methods for a stable solution of (1.1) were proposed (see [4]–[8] and the references therein). However, when F is nonlinear then the theory is less complete.

The DSM for solving equation (1.1) was extensively studied in [8]–[15]. In [8] the following version of the DSM for solving equation (1.1) was studied:

$$\dot{u}_\delta = -\big(F'(u_\delta) + a(t)I\big)^{-1}\big(F(u_\delta) + a(t)u_\delta - f_\delta\big), \quad u_\delta(0) = u_0. \tag{1.4}$$

Here F is a monotone operator, and $a(t) > 0$ is a continuous function, defined for all $t \geq 0$, strictly monotonically decaying, $\lim_{t\to\infty} a(t) = 0$. These assumptions on $a(t)$ hold throughout the paper and are not repeated. Additional assumptions on $a(t)$ will appear later. Convergence of the above DSM was proved in [8] for any initial value u_0 with an *a priori* choice of stopping time t_δ, provided that $a(t)$ is suitably chosen.

The theory of monotone operators is presented in many books, e.g., in [1], [7], [16]. Most of the results of the theory of monotone operators, used in this paper, can be found in [8]. In [6] methods for solving nonlinear equations in a finite-dimensional space are discussed.

In this paper we propose and justify a stopping rule based on a discrepancy principle (DP) for the DSM (1.4). The main result of this paper is Theorem 3.1 in which a DP is formulated, the existence of the stopping time t_δ is proved, and the convergence of the DSM with the proposed DP is justified under some natural assumptions apparently for the first time for a wide class of nonlinear equations with monotone operators.

These results are new from the theoretical point of view and very useful practically. The auxiliary results in our paper are also new and can be used in other problems of numerical analysis. These auxiliary results are formulated in Lemmas 2.2–2.4, 2.7, 2.10, 2.11, and in the remarks. In particular, in Remark 3.3 we emphasize that the trajectory of the solution stays in a ball of a fixed radius R for all $t \geq 0$.

In Section 4 the results of two numerical experiments are presented. In the second experiment we demonstrate numerically that our method for solving equation (1.1) can be used even for a wider class of equations than the basic Theorem 3.1 guarantees.

2. Auxiliary results

Let us consider the following equation:

$$F(V_{\delta,a}) + aV_{\delta,a} - f_\delta = 0, \qquad a > 0, \tag{2.1}$$

where $a = const$. It is known (see, e.g., [8]) that equation (2.1) with monotone continuous operator F has a unique solution for any $f_\delta \in H$.

Let us recall the following result from [8, p. #112]:

Lemma 2.1. *Assume that equation* (1.1) *is solvable, y is its minimal-norm solution, and assumptions* (1.2) *and* (1.3) *hold. Then*

$$\lim_{a\to 0}\|V_{0,a}-y\|=0,$$

where $V_{0,a}$ solves (2.1) *with $\delta=0$.*

Lemma 2.2. *If* (1.2) *holds and F is continuous, then $\|V_{\delta,a}\|=O(\frac{1}{a})$ as $a\to\infty$, and*

$$\lim_{a\to\infty}\|F(V_{\delta,a})-f_\delta\|=\|F(0)-f_\delta\|. \tag{2.2}$$

Proof. Rewrite (2.1) as

$$F(V_{\delta,a})-F(0)+aV_{\delta,a}+F(0)-f_\delta=0.$$

Multiply this equation by $V_{\delta,a}$, use inequality $\langle F(V_{\delta,a})-F(0),V_{\delta,a}-0\rangle\ge 0$ from (1.2) and get:

$$a\|V_{\delta,a}\|^2\le\langle aV_{\delta,a}+F(V_{\delta,a})-F(0),V_{\delta,a}\rangle=\langle f_\delta-F(0),V_{\delta,a}\rangle\le\|f_\delta-F(0)\|\|V_{\delta,a}\|.$$

Therefore, $\|V_{\delta,a}\|=O(\frac{1}{a})$. This and the continuity of F imply (2.2). □

Let $a=a(t)$, $0<a(t)\searrow 0$, and assume $a\in C^1[0,\infty)$. Then the solution $V_\delta(t):=V_{\delta,a(t)}$ of (2.1) is a function of t. From the triangle inequality one gets

$$\|F(V_\delta(0))-f_\delta\|\ge\|F(0)-f_\delta\|-\|F(V_\delta(0))-F(0)\|.$$

From Lemma 2.2 it follows that for large $a(0)$ one has

$$\|F(V_\delta(0))-F(0)\|\le M_1\|V_\delta(0)\|=O\Big(\frac{1}{a(0)}\Big).$$

Therefore, if $\|F(0)-f_\delta\|>C\delta$, then $\|F(V_\delta(0))-f_\delta\|\ge(C-\epsilon)\delta$, where $\epsilon>0$ is sufficiently small, for sufficiently large $a(0)>0$.

Below the words decreasing and increasing mean strictly decreasing and strictly increasing.

Lemma 2.3. *Assume $\|F(0)-f_\delta\|>0$. Let $0<a(t)\searrow 0$, and let F be monotone. Denote*

$$\phi(t):=\|F(V_\delta(t))-f_\delta\|,\quad \psi(t):=\|V_\delta(t)\|,$$

where $V_\delta(t)$ solves (2.1) *with $a=a(t)$. Then $\phi(t)$ is decreasing, and $\psi(t)$ is increasing.*

Proof. Since $\|F(0) - f_\delta\| > 0$, it follows that $\psi(t) \neq 0$, $\forall t \geq 0$. Note that $\phi(t) = a(t)\|V_\delta(t)\|$. One has

$$\begin{aligned} 0 &\leq \langle F(V_\delta(t_1)) - F(V_\delta(t_2)), V_\delta(t_1) - V_\delta(t_2)\rangle \\ &= \langle -a(t_1)V_\delta(t_1) + a(t_2)V_\delta(t_2), V_\delta(t_1) - V_\delta(t_2)\rangle \\ &= (a(t_1) + a(t_2))\langle V_\delta(t_1), V_\delta(t_2)\rangle - a(t_1)\|V_\delta(t_1)\|^2 - a(t_2)\|V_\delta(t_2)\|^2. \end{aligned} \tag{2.3}$$

Thus,

$$\begin{aligned} 0 &\leq (a(t_1) + a(t_2))\langle V_\delta(t_1), V_\delta(t_2)\rangle - a(t_1)\|V_\delta(t_1)\|^2 - a(t_2)\|V_\delta(t_2)\|^2 \\ &\leq (a(t_1) + a(t_2))\|V_\delta(t_1)\|\|V_\delta(t_2)\| - a(t_1)\|V_\delta(t_1)\|^2 - a(t_2)\|V_\delta(t_2)\|^2 \\ &= (a(t_1)\|V_\delta(t_1)\| - a(t_2)\|V_\delta(t_2)\|)(\|V_\delta(t_2)\| - \|V_\delta(t_1)\|) \\ &= (\phi(t_1) - \phi(t_2))(\psi(t_2) - \psi(t_1)). \end{aligned} \tag{2.4}$$

If $\psi(t_2) > \psi(t_1)$, then (2.4) implies $\phi(t_1) \geq \phi(t_2)$, so

$$a(t_1)\psi(t_1) \geq a(t_2)\psi(t_2) > a(t_2)\psi(t_1).$$

Thus, if $\psi(t_2) > \psi(t_1)$, then $a(t_2) < a(t_1)$ and, therefore, $t_2 > t_1$, because $a(t)$ is decreasing.

Similarly, if $\psi(t_2) < \psi(t_1)$, then $\phi(t_1) < \phi(t_2)$. This implies $a(t_2) > a(t_1)$, so $t_2 < t_1$.

If $\psi(t_2) = \psi(t_1)$, then (2.3) implies

$$\|V_\delta(t_1)\|^2 \leq \langle V_\delta(t_1), V_\delta(t_2)\rangle \leq \|V_\delta(t_1)\|\|V_\delta(t_2)\| = \|V_\delta(t_1)\|^2.$$

This implies $V_\delta(t_1) = V_\delta(t_2)$, and then $a(t_1) = a(t_2)$. Hence, $t_1 = t_2$, because $a(t)$ is decreasing.

Therefore, $\phi(t)$ is decreasing and $\psi(t)$ is increasing. □

Lemma 2.4. *Suppose that $\|F(0) - f_\delta\| > C\delta$, $C > 1$, and $a(0)$ is sufficiently large. Then, there exists a unique $t_1 > 0$ such that $\|F(V_\delta(t_1)) - f_\delta\| = C\delta$.*

Proof. The uniqueness of t_1 follows from Lemma 2.3. We have $F(y) = f$, and

$$\begin{aligned} 0 =& \langle F(V_\delta) + aV_\delta - f_\delta, F(V_\delta) - f_\delta\rangle \\ =& \|F(V_\delta) - f_\delta\|^2 + a\langle V_\delta - y, F(V_\delta) - f_\delta\rangle + a\langle y, F(V_\delta) - f_\delta\rangle \\ =& \|F(V_\delta) - f_\delta\|^2 + a\langle V_\delta - y, F(V_\delta) - F(y)\rangle + a\langle V_\delta - y, f - f_\delta\rangle \\ & + a\langle y, F(V_\delta) - f_\delta\rangle \\ \geq& \|F(V_\delta) - f_\delta\|^2 + a\langle V_\delta - y, f - f_\delta\rangle + a\langle y, F(V_\delta) - f_\delta\rangle. \end{aligned}$$

Here the inequality $\langle V_\delta - y, F(V_\delta) - F(y)\rangle \ge 0$ was used. Therefore,

$$\begin{aligned} \|F(V_\delta) - f_\delta\|^2 &\le -a\langle V_\delta - y, f - f_\delta\rangle - a\langle y, F(V_\delta) - f_\delta\rangle \\ &\le a\|V_\delta - y\|\|f - f_\delta\| + a\|y\|\|F(V_\delta) - f_\delta\| \\ &\le a\delta\|V_\delta - y\| + a\|y\|\|F(V_\delta) - f_\delta\|. \end{aligned} \tag{2.5}$$

On the other hand, we have

$$\begin{aligned} 0 &= \langle F(V_\delta) - F(y) + aV_\delta + f - f_\delta, V_\delta - y\rangle \\ &= \langle F(V_\delta) - F(y), V_\delta - y\rangle + a\|V_\delta - y\|^2 + a\langle y, V_\delta - y\rangle + \langle f - f_\delta, V_\delta - y\rangle \\ &\ge a\|V_\delta - y\|^2 + a\langle y, V_\delta - y\rangle + \langle f - f_\delta, V_\delta - y\rangle, \end{aligned}$$

where the inequality $\langle V_\delta - y, F(V_\delta) - F(y)\rangle \ge 0$ was used. Therefore,

$$a\|V_\delta - y\|^2 \le a\|y\|\|V_\delta - y\| + \delta\|V_\delta - y\|.$$

This implies

$$a\|V_\delta - y\| \le a\|y\| + \delta. \tag{2.6}$$

From (2.5) and (2.6), and an elementary inequality $ab \le \epsilon a^2 + \frac{b^2}{4\epsilon}$, $\forall\epsilon > 0$, one gets

$$\begin{aligned} \|F(V_\delta) - f_\delta\|^2 &\le \delta^2 + a\|y\|\delta + a\|y\|\|F(V_\delta) - f_\delta\| \\ &\le \delta^2 + a\|y\|\delta + \epsilon\|F(V_\delta) - f_\delta\|^2 + \frac{1}{4\epsilon}a^2\|y\|^2, \end{aligned} \tag{2.7}$$

where $\epsilon > 0$ is fixed, independent of t, and can be chosen arbitrarily small. Let $t \to \infty$ and $a = a(t) \searrow 0$. Then (2.7) implies $\limsup_{t\to\infty}(1-\epsilon)\|F(V_\delta) - f_\delta\|^2 \le \delta^2$. This, the continuity of F, the continuity of $V_\delta(t)$ on $[0,\infty)$, and the assumption $\|F(0) - f_\delta\| > C\delta$, where $C > 1$, imply that equation $\|F(V_\delta(t)) - f_\delta\| = C\delta$ must have a solution $t_1 > 0$. □

Remark 2.5. Let $V := V_\delta(t)|_{\delta=0}$, so $F(V) + a(t)V - f = 0$. Let y be the minimal-norm solution to $F(u) = f$. We claim that

$$\|V_\delta - V\| \le \frac{\delta}{a}. \tag{2.8}$$

Indeed, from (2.1) one gets

$$F(V_\delta) - F(V) + a(V_\delta - V) = f - f_\delta.$$

Multiply this equality by $(V_\delta - V)$ and use (1.2) to obtain

$$\begin{aligned}\delta\|V_\delta - V\| &\ge \langle f - f_\delta, V_\delta - V\rangle \\ &= \langle F(V_\delta) - F(V) + a(V_\delta - V), V_\delta - V\rangle \\ &\ge a\|V_\delta - V\|^2.\end{aligned}$$

This implies (2.8).

Similarly, from the equation

$$F(V) + aV - F(y) = 0,$$

one can derive that

$$\|V\| \le \|y\|. \tag{2.9}$$

From (2.8) and (2.9), one gets the following estimate:

$$\|V_\delta\| \le \|V\| + \frac{\delta}{a} \le \|y\| + \frac{\delta}{a}. \tag{2.10}$$

Let us recall the following lemma, which is basic in our proofs.

Lemma 2.6 ([8], p. 97)**.** *Let $\alpha(t)$, $\beta(t)$, $\gamma(t)$ be continuous nonnegative functions on $[\tau_0,\infty)$, $\tau_0 \ge 0$ is a fixed number. If there exists a function $\mu := \mu(t)$,*

$$\mu \in C^1[\tau_0,\infty), \quad \mu(t) > 0, \quad \lim_{t\to\infty}\mu(t) = \infty,$$

such that

$$0 \le \alpha(t) \le \frac{\mu(t)}{2}\left[\gamma - \frac{\dot\mu(t)}{\mu(t)}\right], \qquad \dot u := \frac{du}{dt}, \tag{2.11}$$

$$\beta(t) \le \frac{1}{2\mu(t)}\left[\gamma - \frac{\dot\mu(t)}{\mu(t)}\right], \tag{2.12}$$

$$\mu(\tau_0)g(\tau_0) < 1, \tag{2.13}$$

and $g(t) \ge 0$ satisfies the inequality

$$\dot g(t) \le -\gamma(t)g(t) + \alpha(t)g^2(t) + \beta(t), \quad t \ge \tau_0, \tag{2.14}$$

then

$$0 \le g(t) < \frac{1}{\mu(t)} \to 0, \quad as \quad t \to \infty. \tag{2.15}$$

If inequalities (2.11)–(2.13) hold on an interval $[\tau_0, T)$, then $g(t)$, the solution to inequality (2.14), exists on this interval and inequality (2.15) holds on $[\tau_0, T)$.

Lemma 2.7. *Suppose M_1, c_0, and c_1 are positive constants and $0 \neq y \in H$. Then there exist $\lambda > 0$ and a function $a(t) \in C^1[0,\infty)$, $0 < a(t) \searrow 0$, such that the following conditions hold:*

$$\frac{M_1}{\|y\|} \le \lambda, \tag{2.16}$$

$$\frac{c_0}{a(t)} \le \frac{\lambda}{2a(t)}\Big[1 - \frac{|\dot{a}(t)|}{a(t)}\Big], \tag{2.17}$$

$$c_1\frac{|\dot{a}(t)|}{a(t)} \le \frac{a(t)}{2\lambda}\Big[1 - \frac{|\dot{a}(t)|}{a(t)}\Big], \tag{2.18}$$

$$\|F(0) - f_\delta\| \le \frac{a^2(0)}{\lambda}. \tag{2.19}$$

Proof. Take

$$a(t) = \frac{d}{(c+t)^b}, \quad 0 < b \le 1, \quad c \ge \max\big(2b, 1\big). \tag{2.20}$$

Note that $|\dot{a}| = -\dot{a}$. We have

$$\frac{|\dot{a}(t)|}{a(t)} = \frac{b}{c+t} \le \frac{b}{c} \le \frac{1}{2}, \qquad \forall t \ge 0. \tag{2.21}$$

Hence,

$$\frac{1}{2} \le 1 - \frac{|\dot{a}(t)|}{a(t)}, \qquad \forall t \ge 0. \tag{2.22}$$

Take

$$\lambda \ge \frac{M_1}{\|y\|}. \tag{2.23}$$

Then (2.16) is satisfied.

Choose d such that

$$d \ge \max\Big(\sqrt{c^{2b}\lambda\|F(0) - f_\delta\|}, 4b\lambda c_1\Big). \tag{2.24}$$

From equality (2.20) and inequality (2.24) one gets

$$\frac{|\dot{a}(t)|}{a^2(t)} = \frac{b}{d(c+t)^{1-b}} \le \frac{b}{d} \le \frac{1}{4\lambda c_1}, \qquad \forall t \ge 0. \tag{2.25}$$

This and inequality (2.21) imply inequality (2.18). It follows from inequality (2.24) that

$$\|F(0) - f_\delta\| \le \frac{d^2}{c^{2b}\lambda} = \frac{a^2(0)}{\lambda}. \tag{2.26}$$

Thus, inequality (2.19) is satisfied.

Choose $\kappa \ge 1$ such that

$$\kappa > \max\Big(\frac{4c_0}{\lambda}, 1\Big). \tag{2.27}$$

Define

$$\nu(t) := \kappa a(t), \quad \lambda_\kappa := \kappa\lambda. \tag{2.28}$$

Note that inequalities (2.16), (2.18), (2.19) and (2.21) still hold for $a(t) = \nu(t)$ and $\lambda = \lambda_\kappa$.

Using the inequalities (2.27) and $c \geq 1$ and the definition (2.28), one obtains

$$\frac{c_0}{\nu(t)} \leq \frac{\lambda\kappa}{4\nu(t)} \leq \frac{\lambda_\kappa}{2\nu(t)}\left[1 - \frac{|\dot{\nu}|}{\nu}\right]. \tag{2.29}$$

Thus, one can replace the function $a(t)$ by $\nu(t) = \kappa a(t)$ and λ by $\lambda = \lambda_\kappa$ to satisfy inequalities (2.16)–(2.19). □

Remark 2.8. In the proof of Lemma 2.7 $a(0)$ and λ can be chosen so that $\frac{a(0)}{\lambda}$ is uniformly bounded as $\delta \to 0$ regardless of the rate of growth of the constant $M_1 = M_1(R)$ from formula (1.3) when $R \to \infty$, i.e., regardless of the strength of the nonlinearity $F(u)$.

Indeed, to satisfy (2.23) one can choose $\lambda = \frac{M_1}{\|y\|}$. To satisfy (2.24) one can choose

$$d = \max\left(\sqrt{c^{2b}\lambda\|f_\delta - F(0)\|}, 4b\lambda c_1\right) \leq \max\left(\sqrt{c^{2b}\lambda(\|f - F(0)\| + 1)}, 4b\lambda c_1\right),$$

where we have assumed, without loss of generality, that $0 < \delta < 1$. With this choice of d and λ, the ratio $\frac{a(0)}{\lambda}$ is bounded uniformly with respect to $\delta \in (0,1)$ and does not depend on R.

Indeed, with the above choice one has $\frac{a(0)}{\lambda} = \frac{d}{c^b\lambda} \leq \tilde{c}(1 + \sqrt{\lambda^{-1}}) \leq \tilde{c}$, where $\tilde{c} > 0$ is a constant independent of δ, and one can assume that $\lambda \geq 1$ without loss of generality.

This remark is used in Remark 3.3, where we prove that the trajectory of $u_\delta(t)$, defined by (3.1), stays in a ball $B(u_0, R)$ for all $0 \leq t \leq t_\delta$, where the number t_δ is defined by formula (3.3) (see below), and $R > 0$ is sufficiently large. An upper bound on R is given in Remark 3.3.

Remark 2.9. It is easy to choose $u_0 \in H$ such that

$$g_0 := \|u_0 - V_\delta(0)\| \leq \frac{\|F(0) - f_\delta\|}{a(0)}. \tag{2.30}$$

Indeed, if, for example, $u_0 = 0$, then by Lemmas 2.2 and 2.3 one gets

$$g_0 = \|V_\delta(0)\| = \frac{a(0)\|V_\delta(0)\|}{a(0)} \leq \frac{\|F(0) - f_\delta\|}{a(0)}.$$

If (2.19) and (2.30) hold, then $g_0 \leq \frac{a(0)}{\lambda}$. Inequality (2.30) also holds if $||u_0 - V_\delta(0)||$ is sufficiently small.

Lemma 2.10. *Let p, b and c be positive constants. Then*

$$\left(p - \frac{b}{c}\right)\int_0^t \frac{e^{ps}}{(s+c)^b}ds < \frac{e^{pt}}{(c+t)^b}, \qquad \forall c, b > 0, \quad t > 0. \tag{2.31}$$

Proof. One has

$$\begin{aligned}\frac{d}{dt}\Big(\frac{e^{pt}}{(c+t)^b}\Big) &= \frac{pe^{pt}}{(c+t)^b} - \frac{be^{pt}}{(c+t)^{b+1}}\\ &\ge \Big(p-\frac{b}{c}\Big)\frac{e^{pt}}{(c+t)^b}, \qquad t\ge 0.\end{aligned}$$

Therefore,

$$\begin{aligned}\Big(p-\frac{b}{c}\Big)\int_0^t \frac{e^{ps}}{(s+c)^b}ds &\le \int_0^t \frac{d}{ds}\frac{e^{ps}}{(c+s)^b}ds\\ &\le \frac{e^{pt}}{(c+t)^b} - \frac{1}{c^b} \le \frac{e^{pt}}{(c+t)^b}.\end{aligned}$$

Lemma 2.10 is proved. □

Lemma 2.11. *Let* $a(t)=\frac{d}{(c+t)^b}$ *where* $d,c,b>0,\ c\ge 6b$. *One has*

$$e^{-\frac{t}{2}}\int_0^t e^{\frac{s}{2}}|\dot a(s)|\|V_\delta(s)\|ds \le \frac{1}{2}a(t)\|V_\delta(t)\|, \qquad t\ge 0. \tag{2.32}$$

Proof. Let $p=\frac{1}{2}$ in Lemma 2.10. Then

$$\Big(\frac{1}{2}-\frac{b}{c}\Big)\int_0^t \frac{e^{\frac{s}{2}}}{(s+c)^b}ds < \frac{e^{\frac{t}{2}}}{(c+t)^b}, \qquad \forall c,b\ge 0. \tag{2.33}$$

Since $c\ge 6b$ or $\frac{3b}{c}\le\frac{1}{2}$, one has

$$\frac{1}{2}-\frac{b}{c}\ge\frac{2b}{c}\ge\frac{2b}{c+s}, \qquad s\ge 0.$$

This implies

$$a(s)\Big(\frac{1}{2}-\frac{b}{c}\Big) = \frac{d}{(c+s)^b}\Big(\frac{1}{2}-\frac{b}{c}\Big) \ge \frac{2db}{(c+s)^{b+1}} = 2|\dot a(s)|, \qquad s\ge 0. \tag{2.34}$$

Multiplying (2.34) by $e^{\frac{s}{2}}\|V_\delta(s)\|$, integrating from 0 to t, using inequality (2.33) and the fact that $\|V_\delta(s)\|$ is nondecreasing, one gets

$$e^{\frac{t}{2}}a(t)\|V_\delta(t)\| > \int_0^t e^{\frac{s}{2}}\|V_\delta(t)\|a(s)\Big(\frac{1}{2}-\frac{b}{c}\Big)ds \ge 2\int_0^t e^{\frac{s}{2}}|\dot a(s)|\|V_\delta(s)\|ds, \qquad t\ge 0.$$

This implies inequality (2.32). Lemma 2.11 is proved. □

3. Main result

Denote

$$A := F'(u_\delta(t)), \quad A_a := A + aI,$$

where I is the identity operator, and $u_\delta(t)$ solves the following Cauchy problem:

$$\dot{u}_\delta = -A_{a(t)}^{-1}[F(u_\delta) + a(t)u_\delta - f_\delta], \quad u_\delta(0) = u_0. \tag{3.1}$$

We assume below that $||F(u_0) - f_\delta|| > C_1\delta^\zeta$, where $C_1 > 1$ and $\zeta \in (0, 1]$ are some constants. We also assume, without loss of generality, that $\delta \in (0, 1)$.

Assume that equation $F(u) = f$ has a solution, possibly nonunique, and y is the minimal norm solution to this equation. Let f be unknown, but f_δ be given, $\|f_\delta - f\| \le \delta$.

Theorem 3.1. *Assume $a(t) = \frac{d}{(c+t)^b}$, where $b \in (0, 1]$, $c, d > 0$ are constants, $c > 6b$, and d is sufficiently large so that conditions (2.17)–(2.19) hold. Assume that $F : H \to H$ is a monotone operator, twice Fréchet differentiable, $\sup_{u \in B(u_0,R)} \|F^{(j)}(u)\| \le M_j(u_0, R)$, $0 \le j \le 2$, $B(u_0, R) := \{u : \|u - u_0\| \le R\}$, u_0 is an element of H, satisfying inequality (2.30) and*

$$\|F(u_0) + a(0)u_0 - f_\delta\| \le \frac{1}{4}a(0)\|V_\delta(0)\|, \tag{3.2}$$

where $V_\delta(t) := V_{\delta,a(t)}$ solves (2.1) with $a = a(t)$. Then the solution $u_\delta(t)$ to problem (3.1) exists on an interval $[0, T_\delta]$, $\lim_{\delta\to 0} T_\delta = \infty$, and there exists a unique t_δ, $t_\delta \in (0, T_\delta)$ such that $\lim_{\delta\to 0} t_\delta = \infty$ and

$$\|F(u_\delta(t_\delta)) - f_\delta\| = C_1\delta^\zeta, \quad \|F(u_\delta(t)) - f_\delta\| > C_1\delta^\zeta, \quad \forall t \in [0, t_\delta), \tag{3.3}$$

where $C_1 > 1$ and $0 < \zeta \le 1$. If $\zeta \in (0, 1)$ and t_δ satisfies (3.3), then

$$\lim_{\delta\to 0} \|u_\delta(t_\delta) - y\| = 0. \tag{3.4}$$

Remark 3.2. One can choose u_0 satisfying inequalities (2.30) and (3.2) (see also (3.34) below). Indeed, if u_0 is a sufficiently close approximation to $V_\delta(0)$, the solution to equation (2.1), then inequalities (2.30) and (3.2) are satisfied. Note that inequality (3.2) is a sufficient condition for (3.35) to hold. In our proof inequality (3.35) is used at $t = t_\delta$. The stopping time t_δ is often sufficiently large for the quantity $e^{-\frac{t_\delta}{2}}h_0$ to be small. In this case inequality (3.35) with $t = t_\delta$ is satisfied for a wide range of u_0. For example, in our numerical experiment in Section 4 the method converged rapidly when $u_0 = 0$. Condition $c > 6b$ is used in the proof of Lemma 2.11.

Proof of Theorem 3.1. Denote

$$C := \frac{C_1 + 1}{2}. \tag{3.5}$$

Let

$$w := u_\delta - V_\delta, \quad g(t) := \|w\|.$$

One has

$$\dot{w} = -\dot{V}_\delta - A^{-1}_{a(t)}\big[F(u_\delta) - F(V_\delta) + a(t)w\big]. \tag{3.6}$$

We use Taylor's formula and get

$$F(u_\delta) - F(V_\delta) + aw = A_a w + K, \quad \|K\| \le \frac{M_2}{2}\|w\|^2, \tag{3.7}$$

where $K := F(u_\delta) - F(V_\delta) - Aw$, and M_2 is the constant from the estimate (1.3). Multiplying (3.6) by w and using (3.7) one gets

$$g\dot{g} \le -g^2 + \frac{M_2}{2}\|A^{-1}_{a(t)}\|g^3 + \|\dot{V}_\delta\|g. \tag{3.8}$$

Let t_0 be such that

$$\frac{\delta}{a(t_0)} = \frac{1}{C-1}\|y\|, \qquad C > 1. \tag{3.9}$$

This t_0 exists and is unique since $a(t) > 0$ monotonically decays to 0 as $t \to \infty$.

Since $a(t) > 0$ monotonically decays, one has

$$\frac{\delta}{a(t)} \le \frac{1}{C-1}\|y\|, \qquad 0 \le t \le t_0. \tag{3.10}$$

By Lemma 2.4, there exists t_1 such that

$$\|F(V_\delta(t_1)) - f_\delta\| = C\delta, \quad F(V_\delta(t_1)) + a(t_1)V_\delta(t_1) - f_\delta = 0. \tag{3.11}$$

We claim that $t_1 \in [0, t_0]$.

Indeed, from (2.1) and (2.10) one gets

$$C\delta = a(t_1)\|V_\delta(t_1)\| \le a(t_1)\bigg(\|y\| + \frac{\delta}{a(t_1)}\bigg) = a(t_1)\|y\| + \delta, \quad C > 1,$$

so

$$\delta \le \frac{a(t_1)\|y\|}{C-1}.$$

Thus,

$$\frac{\delta}{a(t_1)} \le \frac{\|y\|}{C-1} = \frac{\delta}{a(t_0)}.$$

Since $a(t) \searrow 0$, one has $t_1 \le t_0$.

Differentiating both sides of (2.1) with respect to t, one obtains

$$A_{a(t)}\dot{V}_\delta = -\dot{a}V_\delta.$$

This implies

$$\begin{aligned}\|\dot{V}_\delta\| \le |\dot{a}|\|A^{-1}_{a(t)}V_\delta\| \le \frac{|\dot{a}|}{a}\|V_\delta\| &\le \frac{|\dot{a}|}{a}\Big(\|y\| + \frac{\delta}{a}\Big)\\ &\le \frac{|\dot{a}|}{a}\|y\|\Big(1+\frac{1}{C-1}\Big), \qquad \forall t \le t_0.\end{aligned} \tag{3.12}$$

Since $g \ge 0$, inequalities (3.8) and (3.12) imply

$$\dot{g} \le -g(t) + \frac{c_0}{a(t)}g^2 + \frac{|\dot{a}|}{a(t)}c_1, \quad c_0 = \frac{M_2}{2}, \quad c_1 = \|y\|\Big(1+\frac{1}{C-1}\Big). \tag{3.13}$$

Here we have used the estimate

$$\|A_a^{-1}\| \le \frac{1}{a}$$

and the relations

$$A_a := F'(u) + aI, \quad F'(u) := A \ge 0.$$

Inequality (3.13) is of the type (2.14) with

$$\gamma(t) = 1, \quad \alpha(t) = \frac{c_0}{a(t)}, \quad \beta(t) = c_1\frac{|\dot{a}|}{a(t)}.$$

Let us check assumptions (2.11)–(2.13). Take

$$\mu(t) = \frac{\lambda}{a(t)},$$

where $\lambda = const > 0$ and satisfies conditions (2.11)–(2.13) in Lemma 2.7. Since u_0 satisfies inequality (2.30), one gets $g(0) \le \frac{a(0)}{\lambda}$, by Remark 2.9. This, inequalities (2.11)–(2.13), and Lemma 2.6 yield

$$g(t) < \frac{a(t)}{\lambda}, \quad \forall t \le t_0, \qquad g(t) := \|u_\delta(t) - V_\delta(t)\|. \tag{3.14}$$

Therefore,

$$\begin{aligned}\|F(u_\delta(t)) - f_\delta\| \le &\|F(u_\delta(t)) - F(V_\delta(t))\| + \|F(V_\delta(t)) - f_\delta\|\\ \le &M_1 g(t) + \|F(V_\delta(t)) - f_\delta\|\\ \le &\frac{M_1 a(t)}{\lambda} + \|F(V_\delta(t)) - f_\delta\|, \qquad \forall t \le t_0.\end{aligned} \tag{3.15}$$

It is proved in Section 2, Lemma 2.3, that $\|F(V_\delta(t)) - f_\delta\|$ *is decreasing.* Since $t_1 \le t_0$, one gets

$$\|F(V_\delta(t_0)) - f_\delta\| \le \|F(V_\delta(t_1)) - f_\delta\| = C\delta. \tag{3.16}$$

This, inequality (3.15), the inequality $\frac{M_1}{\lambda} \le \|y\|$ (see (2.23)), the relation (3.9), and the definition $C_1 = 2C - 1$ (see (3.5)), imply

$$\begin{aligned} \|F(u_\delta(t_0)) - f_\delta\| \le & \frac{M_1 a(t_0)}{\lambda} + C\delta \\ \le & \frac{M_1\delta(C-1)}{\lambda\|y\|} + C\delta \le (2C-1)\delta = C_1\delta. \end{aligned} \tag{3.17}$$

Thus, if

$$\|F(u_\delta(0)) - f_\delta\| > C_1\delta^\gamma, \quad 0 < \gamma \le 1,$$

then, by the continuity of the function $t \to \|F(u_\delta(t)) - f_\delta\|$ on $[0,\infty)$, there exists $t_\delta \in (0, t_0)$ such that

$$\|F(u_\delta(t_\delta)) - f_\delta\| = C_1\delta^\gamma \tag{3.18}$$

for any given $\gamma \in (0,1]$, and any fixed $C_1 > 1$.

Let us prove (3.4).

From (3.15) with $t = t_\delta$, and from (2.10), one gets

$$\begin{aligned} C_1\delta^\zeta &\le M_1\frac{a(t_\delta)}{\lambda} + a(t_\delta)\|V_\delta(t_\delta)\| \\ &\le M_1\frac{a(t_\delta)}{\lambda} + \|y\|a(t_\delta) + \delta. \end{aligned}$$

Thus, for sufficiently small δ, one gets

$$\tilde{C}\delta^\zeta \le a(t_\delta)\Big(\frac{M_1}{\lambda} + \|y\|\Big), \quad \tilde{C} > 0,$$

where $\tilde{C} < C_1$ is a constant. Therefore,

$$\lim_{\delta\to 0}\frac{\delta}{a(t_\delta)} \le \lim_{\delta\to 0}\frac{\delta^{1-\zeta}}{\tilde{C}}\Big(\frac{M_1}{\lambda} + \|y\|\Big) = 0, \quad 0 < \zeta < 1. \tag{3.19}$$

We claim that

$$\lim_{\delta\to 0} t_\delta = \infty. \tag{3.20}$$

Let us prove (3.20). Using (3.1), one obtains

$$\frac{d}{dt}\big(F(u_\delta) + au_\delta - f_\delta\big) = A_a\dot{u}_\delta + \dot{a}u_\delta = -\big(F(u_\delta) + au_\delta - f_\delta\big) + \dot{a}u_\delta.$$

This and (2.1) imply

$$\frac{d}{dt}\big[F(u_\delta) - F(V_\delta) + a(u_\delta - V_\delta)\big] = -\big[F(u_\delta) - F(V_\delta) + a(u_\delta - V_\delta)\big] + \dot{a}u_\delta. \tag{3.21}$$

Denote

$$v := v(t) := F(u_\delta(t)) - F(V_\delta(t)) + a(t)(u_\delta(t) - V_\delta(t)), \qquad h := h(t) := \|v\|.$$

Multiplying (3.21) by v, one obtains

$$\begin{aligned} h\dot{h} &= -h^2 + \langle v, \dot{a}(u_\delta - V_\delta)\rangle + \dot{a}\langle v, V_\delta\rangle \\ &\le -h^2 + h|\dot{a}|\|u_\delta - V_\delta\| + |\dot{a}|h\|V_\delta\|, \qquad h \ge 0. \end{aligned} \tag{3.22}$$

Thus,

$$\dot{h} \le -h + |\dot{a}|\|u_\delta - V_\delta\| + |\dot{a}|\|V_\delta\|. \tag{3.23}$$

Since $\langle F(u_\delta) - F(V_\delta), u_\delta - V_\delta\rangle \ge 0$, one obtains from the two equations

$$\langle v, u_\delta - V_\delta\rangle = \langle F(u_\delta) - F(V_\delta) + a(t)(u_\delta - V_\delta), u_\delta - V_\delta\rangle$$

and

$$\langle v, F(u_\delta) - F(V_\delta)\rangle = \|F(u_\delta) - F(V_\delta)\|^2 + a(t)\langle u_\delta - V_\delta, F(u_\delta) - F(V_\delta)\rangle,$$

the following two inequalities:

$$a\|u_\delta - V_\delta\|^2 \le \langle v, u_\delta - V_\delta\rangle \le \|u_\delta - V_\delta\|h \tag{3.24}$$

and

$$\|F(u_\delta) - F(V_\delta)\|^2 \le \langle v, F(u_\delta) - F(V_\delta)\rangle \le h\|F(u_\delta) - F(V_\delta)\|. \tag{3.25}$$

Inequalities (3.24) and (3.25) imply

$$a\|u_\delta - V_\delta\| \le h, \quad \|F(u_\delta) - F(V_\delta)\| \le h. \tag{3.26}$$

Inequalities (3.23) and (3.26) imply

$$\dot{h} \le -h\Big(1 - \frac{|\dot{a}|}{a}\Big) + |\dot{a}|\|V_\delta\|. \tag{3.27}$$

Since $1 - \frac{|\dot{a}|}{a} \ge \frac{1}{2}$ because $c \ge 2b$, inequality (3.27) holds if

$$\dot{h} \le -\frac{1}{2}h + |\dot{a}|\|V_\delta\|. \tag{3.28}$$

Inequality (3.28) implies

$$h(t) \le h(0)e^{-\frac{t}{2}} + e^{-\frac{t}{2}}\int_0^t e^{\frac{s}{2}}|\dot{a}|\|V_\delta\|ds. \tag{3.29}$$

From (3.29) and (3.26), one gets

$$\|F(u_\delta(t)) - F(V_\delta(t))\| \le h(0)e^{-\frac{t}{2}} + e^{-\frac{t}{2}}\int_0^t e^{\frac{s}{2}}|\dot{a}|\|V_\delta\|ds. \tag{3.30}$$

Therefore,

$$\begin{aligned}\|F(u_\delta(t))-f_\delta\| &\geq \|F(V_\delta(t))-f_\delta\| - \|F(V_\delta(t))-F(u_\delta(t))\|\\ &\geq a(t)\|V_\delta(t)\| - h(0)e^{-\frac{t}{2}} - e^{-\frac{t}{2}}\int_0^t e^{\frac{s}{2}}|\dot{a}|\|V_\delta\|ds.\end{aligned} \tag{3.31}$$

From the results in Section 2 (see Lemma 2.11), it follows that there exists an $a(t)$ such that

$$\frac{1}{2}a(t)\|V_\delta(t)\| \geq e^{-\frac{t}{2}}\int_0^t e^{\frac{s}{2}}|\dot{a}|\|V_\delta(s)\|ds. \tag{3.32}$$

For example, one can choose

$$a(t) = \frac{d}{(c+t)^b}, \quad 6b < c, \tag{3.33}$$

where $d, c, b > 0$. Moreover, one can always choose u_0 such that

$$h(0) = \|F(u_0) + a(0)u_0 - f_\delta\| \leq \frac{1}{4}a(0)\|V_\delta(0)\|, \tag{3.34}$$

because the equation $F(u_0) + a(0)u_0 - f_\delta = 0$ is solvable. If (3.34) holds, then

$$h(0)e^{-\frac{t}{2}} \leq \frac{1}{4}a(0)\|V_\delta(0)\|e^{-\frac{t}{2}}, \qquad t \geq 0.$$

If $2b < c$, then (3.33) implies

$$e^{-\frac{t}{2}}a(0) \leq a(t).$$

Therefore,

$$e^{-\frac{t}{2}}h(0) \leq \frac{1}{4}a(t)\|V_\delta(0)\| \leq \frac{1}{4}a(t)\|V_\delta(t)\|, \quad t \geq 0, \tag{3.35}$$

where we have used the inequality $\|V_\delta(t)\| \leq \|V_\delta(t')\|$ for $t < t'$, established in Lemma 2.3 in Section 2. From (3.18) and (3.31)–(3.35), one gets

$$C_1\delta^\zeta = \|F(u_\delta(t_\delta)) - f_\delta\| \geq \frac{1}{4}a(t_\delta)\|V_\delta(t_\delta)\|.$$

Thus,

$$\lim_{\delta\to 0} a(t_\delta)\|V_\delta(t_\delta)\| \leq \lim_{\delta\to 0} 4C_1\delta^\zeta = 0.$$

Since $\|V_\delta(t)\|$ increases (see Lemma 2.3), the above formula implies $\lim_{\delta\to 0} a(t_\delta) = 0$. Since $0 < a(t) \searrow 0$, it follows that $\lim_{\delta\to 0} t_\delta = \infty$, i.e., (3.20) holds.

It is now easy to finish the proof of the Theorem 3.1.

From the triangle inequality and inequalities (3.14) and (2.8) one obtains

$$\begin{aligned}\|u_\delta(t_\delta) - y\| &\leq \|u_\delta(t_\delta) - V_\delta(t_\delta)\| + \|V(t_\delta) - V_\delta(t_\delta)\| + \|V(t_\delta) - y\|\\ &\leq \frac{a(t_\delta)}{\lambda} + \frac{\delta}{a(t_\delta)} + \|V(t_\delta) - y\|.\end{aligned} \tag{3.36}$$

Note that $V(t_\delta) = V_{0,a(t_\delta)}$ (see equation (2.1)). From (3.19), (3.20), inequality (3.36) and Lemma 2.1, one obtains (3.4). Theorem 3.1 is proved. □

Remark 3.3. The trajectory $u_\delta(t)$ remains in the ball $B(u_0, R) := \{u : \|u - u_0\| < R\}$ for all $t \le t_\delta$, where R does not depend on δ as $\delta \to 0$. Indeed, estimates (3.14), (2.10) and (3.10) imply

$$\begin{aligned} \|u_\delta(t) - u_0\| &\le \|u_\delta(t) - V_\delta(t)\| + \|V_\delta(t)\| + \|u_0\| \\ &\le \frac{a(0)}{\lambda} + \frac{C\|y\|}{C-1} + \|u_0\| := R, \qquad \forall t \le t_\delta. \end{aligned} \tag{3.37}$$

Here we have used the fact that $t_\delta < t_0$ (see the proof of Theorem 3.1). Since one can choose $a(t)$ and λ so that $\frac{a(0)}{\lambda}$ is uniformly bounded as $\delta \to 0$ and regardless of the growth of M_1 (see Remark 2.8) one concludes that R can be chosen independent of δ and M_1.

4. Numerical experiments

4.1. **An experiment with an operator defined on $H = L^2[0,1]$.** Let us do a numerical experiment solving nonlinear equation (1.1) with

$$F(u) := B(u) + \big(\arctan(u)\big)^3 := \int_0^1 e^{-|x-y|}u(y)dy + \big(\arctan(u)\big)^3. \tag{4.1}$$

Since the function $u \to \arctan^3 u$ is increasing on $\mathbb{R}$, one has

$$\langle \big(\arctan(u)\big)^3 - \big(\arctan(v)\big)^3, u - v\rangle \ge 0, \qquad \forall\, u, v \in H. \tag{4.2}$$

Moreover,

$$e^{-|x|} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{e^{i\lambda x}}{1+\lambda^2} d\lambda. \tag{4.3}$$

Therefore, $\langle B(u-v), u-v\rangle \ge 0$, so

$$\langle F(u) - F(v), u - v\rangle \ge 0, \qquad \forall\, u, v \in H. \tag{4.4}$$

Thus, F is a monotone operator. Note that

$$\langle \big(\arctan(u)\big)^3 - \big(\arctan(v)\big)^3, u - v\rangle = 0 \quad \text{iff} \quad u = v \quad a.e.$$

Therefore, the operator F, defined in (4.1), is injective and equation (1.1), with this F, has at most one solution.

The Fréchet derivative of F is

$$F'(u)w = \frac{3\big(\arctan(u)\big)^2}{1+u^2} w + \int_0^1 e^{-|x-y|}w(y)dy. \tag{4.5}$$

If $u(x)$ vanishes on a set of positive Lebesgue measure, then $F'(u)$ is not boundedly invertible. If $u \in C[0,1]$ vanishes even at one point x_0, then $F'(u)$ is not boundedly invertible in H.

In numerical implementation of the DSM, one often discretizes the Cauchy problem (3.1) and gets a system of ordinary differential equations (ODEs). Then, one can use numerical methods for solving ODEs to solve the system of ordinary differential equations obtained from discretization. There are many numerical methods for solving ODEs (see, e.g., [2]).

In practice one does not have to compute $u_\delta(t_\delta)$ exactly but can use an approximation to $u_\delta(t_\delta)$ as a stable solution to equation (1.1). To calculate such an approximation, one can use, for example, the iterative scheme

$$\begin{aligned} u_{n+1} &= u_n - (F'(u_n) + a_n I)^{-1}(F(u_n) + a_n u_n - f_\delta), \\ u_0 &= 0, \end{aligned} \tag{4.6}$$

and stop iterations at $n := n_\delta$ such that the following inequality holds:

$$\|F(u_{n_\delta}) - f_\delta\| < C\delta^\gamma, \quad \|F(u_n) - f_\delta\| \geq C\delta^\gamma, \quad n < n_\delta, \quad C > 1, \quad \gamma \in (0,1). \tag{4.7}$$

The existence of the stopping time n_δ is proved in [3, p. 733] and the choice $u_0 = 0$ is also justified in this paper. Iterative scheme (4.6) and stopping rule (4.7) are used in the numerical experiments. We proved in [3, p. 733] that u_{n_δ} converges to u^*, a solution of (1.1). Since F is injective as discussed above, we conclude that u_{n_δ} converges to the unique solution of equation (1.1) as δ tends to 0. The accuracy and stability are the key issues in solving the Cauchy problem. The iterative scheme (4.6) can be considered formally as the explicit Euler's method with the stepsize $h = 1$ (see, e.g., [2]). There might be other iterative schemes which are more efficient than scheme (4.6), but this scheme is simple and easy to implement.

Integrals of the form $\int_0^1 e^{-|x-y|}h(y)dy$ in (4.1) and (4.5) are computed by using the trapezoidal rule. The noisy function used in the test is

$$f_\delta(x) = f(x) + \kappa f_{noise}(x), \quad \kappa > 0.$$

The noise level δ and the relative noise level are defined by the formulas

$$\delta = \kappa \|f_{noise}\|, \quad \delta_{rel} := \frac{\delta}{\|f\|}.$$

In the test κ is computed in such a way that the relative noise level δ_{rel} equals some desired value, i.e.,

$$\kappa = \frac{\delta}{\|f_{noise}\|} = \frac{\delta_{rel}\|f\|}{\|f_{noise}\|}.$$

We have used the relative noise level as an input parameter in the test.

In all the figures the x-variable runs through the interval $[0,1]$, and the graphs represent the numerical solutions $u_{DSM}(x)$ and the exact solution $u_{exact}(x)$.

In the test we took $h = 1$, $C = 1.01$, and $\gamma = 0.99$. The exact solution in the test is

$$u_e(x) = \begin{cases} 0 & \text{if} \quad \frac{1}{3} \le x \le \frac{2}{3}, \\ 1 & \quad \text{otherwise}, \end{cases} \tag{4.8}$$

here $x \in [0,1]$, and the right-hand side is $f = F(u_e)$. As mentioned above, $F'(u)$ is not boundedly invertible in any neighborhood of u_e.

It is proved in [3] that one can take $a_n = \frac{d}{1+n}$, and d is sufficiently large. However, in practice, if we choose d too large, then the method will use too many iterations before reaching the stopping time n_δ in (4.7). This means that the computation time will be large in this case. Since

$$\|F(V_\delta) - f_\delta\| = a(t)\|V_\delta\|,$$

and $\|V_\delta(t_\delta) - u_\delta(t_\delta)\| = O(a(t_\delta))$, we have

$$C\delta^\gamma = \|F(u_\delta(t_\delta)) - f_\delta\| \le a(t_\delta)\|V_\delta\| + O(a(t_\delta)),$$

and we choose

$$d = C_0\delta^\gamma, \qquad C_0 > 0.$$

In the experiments our method works well with $C_0 \in [7, 10]$. In numerical experiments, we found out that the method diverged for smaller C_0. In the test we chose a_n by the formula $a_n := C_0\frac{\delta^{0.99}}{n+1}$. The number of nodal points, used in computing integrals in (4.1) and (4.5), was $N = 100$. The accuracy of the solutions obtained in the tests with $N = 30$ and $N = 50$ was slightly less accurate than the one for $N = 100$.

Numerical results for various values of δ_{rel} are presented in Table 3. In this experiment, the noise function f_{noise} is a vector with random entries normally distributed, with mean value 0 and variance 1. Table 3 shows that the iterative scheme yields good numerical results.

TABLE 3. Results when $C_0 = 7$, $N = 100$ and $u = u_e$.

δ_{rel}	0.02	0.01	0.005	0.003	0.001
Number of iterations	57	57	58	58	59
$\frac{\|u_{DSM}-u_{exact}\|}{\|u_{exact}\|}$	0.1437	0.1217	0.0829	0.0746	0.0544

Figure 6 presents the numerical results when $N = 100$ and $C_0 = 7$ with $\delta_{rel} = 0.01$ and $\delta_{rel} = 0.005$. The numbers of iterations for $\delta = 0.01$ and $\delta = 0.005$ were 57 and 58, respectively.

Figure 7 presents the numerical results when $N = 100$ and $C_0 = 7$ with $\delta = 0.003$ and $\delta = 0.001$. In these cases, it took 58 and 59 iterations to get the numerical solutions for $\delta_{rel} = 0.003$ and $\delta_{rel} = 0.001$, respectively.

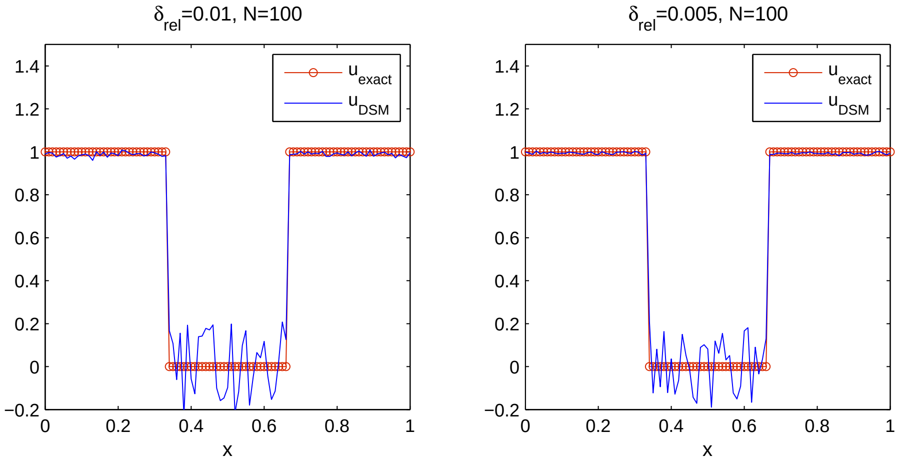

FIGURE 6. Plots of solutions obtained by the DSM when $N = 100$, $\delta_{rel} = 0.01$ (left) and $\delta_{rel} = 0.005$ (right).

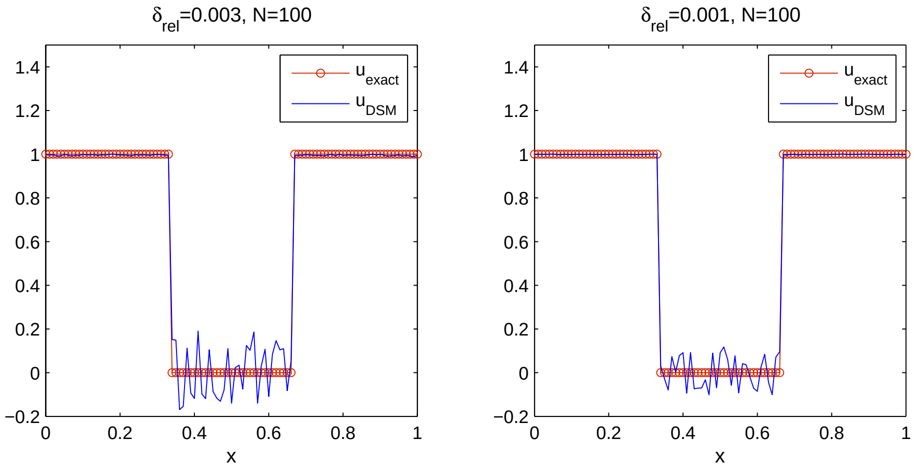

FIGURE 7. Plots of solutions obtained by the DSM when $N = 100$, $\delta_{rel} = 0.003$ (left) and $\delta_{rel} = 0.001$ (right).

We also carried out numerical experiments with $u(x) \equiv 1$, $x \in [0,1]$, as the exact solution. Note that $F'(u)$ is boundedly invertible at this exact solution. However, in any arbitrarily small (in L^2 norm) neighborhood of this solution, there are infinitely many elements u at which $F'(u)$ is not boundedly invertible, because, as we have pointed out earlier, $F'(u)$ is not boundedly invertible if $u(x)$ is continuous and vanishes at some point $x \in [0,1]$. In this case one cannot use the usual methods like Newton's method or the Newton-Kantorovich method. Numerical results for this experiment are presented in Table 4.

From Table 4 one concludes that the method works well in this experiment.

TABLE 4. Results when $C_0 = 4$, $N = 50$ and $u(x) \equiv 1$, $x \in [0,1]$.

δ_{rel}	0.05	0.03	0.02	0.01	0.003	0.001
Number of iterations	28	29	28	29	29	29
$\frac{\|u_{DSM}-u_{exact}\|}{\|u_{exact}\|}$	0.0770	0.0411	0.0314	0.0146	0.0046	0.0015

4.2. **An experiment with an operator defined on a dense subset of $H = L^2[0,1]$.** Our second numerical experiment with the equation $F(u) = f$ deals with the operator F which is not defined on all of $H = L^2[0,1]$, but on a dense subset $D = C[0,1]$ of H:

$$F(u) := B(u) + u^3 := \int_0^1 e^{-|x-y|}u(y)dy + u^3. \tag{4.9}$$

Therefore, the assumptions of Theorem 3.1 are not satisfied. Our goal is to show by this numerical example, that numerically our method may work for an even wider class of problems than that covered by Theorem 3.1.

The operator B is compact in $H = L^2[0,1]$. The operator $u \longmapsto u^3$ is defined on a dense subset D of $L^2[0,1]$, for example, on $D := C[0,1]$. If $u, v \in D$, then

$$\langle u^3 - v^3, u - v\rangle = \int_0^1 (u^3 - v^3)(u - v)dx \geq 0. \tag{4.10}$$

This and the inequality $\langle B(u - v), u - v\rangle \geq 0$, followed from equality (4.3), imply

$$\langle F(u) - F(v), u - v\rangle \geq 0, \qquad \forall u, v \in D.$$

Note that the equal sign of inequality (4.10) happens iff $u = v$ a.e. in Lebesgue measure. Thus, F is injective. Therefore, the element u_{n_δ} obtained from the iterative scheme (4.6) and the stopping rule (4.7) converges to the exact solution u_e as δ goes to 0.

Note that D does not contain subsets open in $H = L^2[0,1]$, i.e., it does not contain interior points of H. This is a reflection of the fact that the operator $G(u) = u^3$ is unbounded on any open subset of H. For example, in any ball $\|u\| \leq C$, $C = const > 0$, where $\|u\| := \|u\|_{L^2[0,1]}$, there is an element u such that $\|u^3\| = \infty$. As such an element one can take, for example, $u(x) = c_1x^{-b}$, $\frac{1}{3} < b < \frac{1}{2}$. Here $c_1 > 0$ is a constant chosen so that $\|u\| \leq C$. The operator $u \longmapsto F(u) = G(u) + B(u)$ is maximal monotone on $D_F := \{u : u \in H,\ F(u) \in H\}$ (see [1, p. #102]), so that equation (2.1) is uniquely solvable for any $f_\delta \in H$.

The Fréchet derivative of F is

$$F'(u)w = 3u^2w + \int_0^1 e^{-|x-y|}w(y)dy. \tag{4.11}$$

If $u(x)$ vanishes on a set of positive Lebesgue measure, then $F'(u)$ is obviously not boundedly invertible. If $u \in C[0,1]$ vanishes even at one point x_0, then $F'(u)$ is not boundedly invertible in H.

We also use the iterative scheme (4.6) with the stopping rule (4.7).

We use the same exact solution u_e as in (4.8). The right-hand side f is computed by $f = F(u_e)$. Note that F' is not boundedly invertible in any neighborhood of u_e.

In experiments we found that our method works well with $C_0 \in [1,4]$. Indeed, in the test we chose a_n by the formula $a_n := C_0\frac{\delta^{0.9}}{n+6}$. The number of node points used in computing integrals in (4.1) and (4.5) was $N = 30$. In the test, the accuracy of the solutions obtained when $N = 30$, $N = 50$ were slightly less accurate than the one when $N = 100$.

Numerical results for various values of δ_{rel} are presented in Table 5. In this experiment, the noise function f_{noise} is a vector with random entries normally distributed of mean 0 and variance 1. Table 5 shows that the iterative scheme yields good numerical results.

TABLE 5. Results when $C_0 = 2$ and $N = 100$.

δ_{rel}	0.02	0.01	0.005	0.003	0.001
Number of iterations	16	17	17	17	18
$\frac{\|u_{DSM}-u_{exact}\|}{\|u_{exact}\|}$	0.1387	0.1281	0.0966	0.0784	0.0626

Figure 8 presents the numerical results when $f_{noise}(x) = \sin(3\pi x)$ for $\delta_{rel} = 0.02$ and $\delta_{rel} = 0.01$. The number of iterations when $C_0 = 2$ for $\delta_{rel} = 0.02$ and $\delta_{rel} = 0.01$ were 16 and 17, respectively.

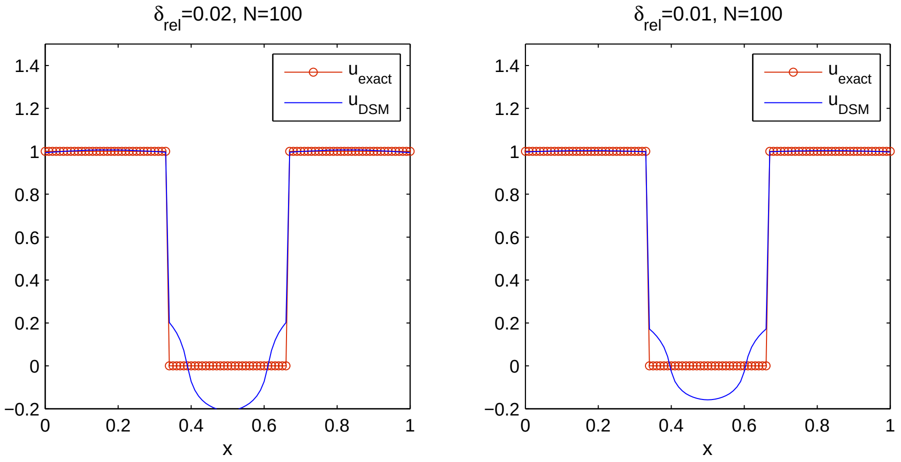

FIGURE 8. Plots of solutions obtained by the DSM with $f_{noise}(x) = \sin(3\pi x)$ when $N = 100$, $\delta_{rel} = 0.02$ (left) and $\delta_{rel} = 0.01$ (right).

Figure 9 presents the numerical results when $f_{noise}(x) = \sin(3\pi x)$ with $\delta_{rel} = 0.003$ and $\delta_{rel} = 0.001$. We also used $C_0 = 2$. In these cases, it took 17 and 18 iterations to give the numerical solutions for $\delta_{rel} = 0.003$ and $\delta_{rel} = 0.001$, respectively.

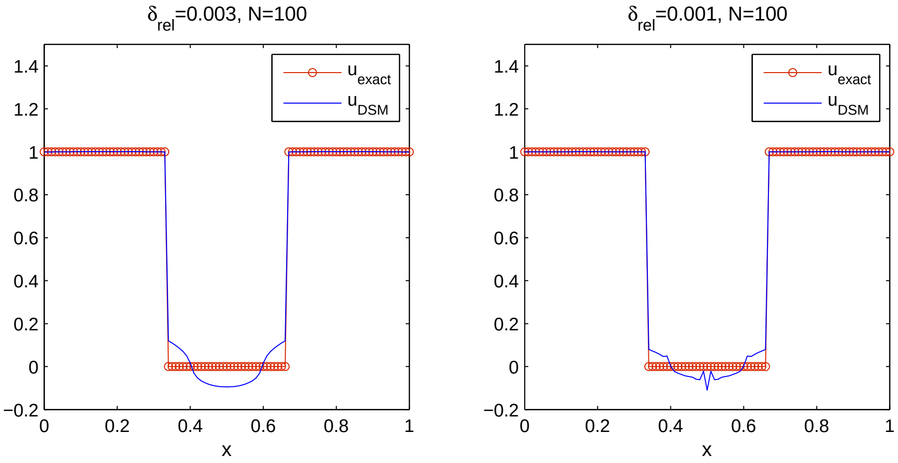

FIGURE 9. Plots of solutions obtained by the DSM with $f_{noise}(x) = \sin(3\pi x)$ when $N = 100$, $\delta_{rel} = 0.003$ (left) and $\delta_{rel} = 0.001$ (right).

We have included the results of the numerical experiments with $u(x) \equiv 1$, $x \in [0, 1]$, as the exact solution. The operator $F'(u)$ is boundedly invertible in $L^2([0, 1])$ at this exact solution. However, in any arbitrarily small L^2-neighborhood of this solution, there are infinitely many elements u at which $F'(u)$ is not boundedly invertible as was mentioned above. Therefore, even in this case one cannot use the usual methods such as Newton's method or the Newton-Kantorovich method. Numerical results for this experiment are presented in Table 6.

TABLE 6. Results when $C_0 = 1$, $N = 30$ and $u(x) = 1$, $x \in [0, 1]$.

δ_{rel}	0.05	0.03	0.02	0.01	0.003	0.001
Number of iterations	7	8	8	9	10	10
$\frac{\|u_{DSM}-u_{exact}\|}{\|u_{exact}\|}$	0.0436	0.0245	0.0172	0.0092	0.0026	0.0009

From the numerical experiments we can conclude that the method works well in this experiment. Note that the function F used in this experiment is not defined on the whole space $H = L^2[0, 1]$ but defined on a dense subset $D = C[0, 1]$ of H.

REFERENCES

[1] K. Deimling, *Nonlinear functional analysis*, Springer-Verlag, Berlin, 1985.

[2] E. Hairer, and S. P. Norsett, and G. Wanner, *Solving ordinary differential equations. I, Nonstiff problems*, Springer-Verlag, Berlin, 1993.

[3] N. S. Hoang and A. G. Ramm, An iterative scheme for solving equations with monotone operators, *BIT*, 48, N4, (2008), 725-741.

[4] V. Ivanov, V. Tanana and V. Vasin, *Theory of ill-posed problems*, VSP, Utrecht, 2002.

[5] V. A. Morozov, *Methods of solving incorrectly posed problems*, Springer-Verlag, New York, 1984.

[6] J. Ortega, W. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, SIAM, Philadelphia, 2000.

[7] D. Pascali and S. Sburlan, *Nonlinear mappings of monotone type*, Noordhoff, Leyden, 1978.

[8] A. G. Ramm, *Dynamical systems method for solving operator equations*, Elsevier, Amsterdam, 2007.

[9] A. G. Ramm, Global convergence for ill-posed equations with monotone operators: the dynamical systems method, *J. Phys A*, 36, (2003), L249-L254.

[10] A. G. Ramm, Dynamical systems method for solving nonlinear operator equations, *International Jour. of Applied Math. Sci.*, 1, N1, (2004), 97-110.

[11] A. G. Ramm, Dynamical systems method for solving operator equations, *Communic. in Nonlinear Sci. and Numer. Simulation*, 9, N2, (2004), 383-402.

[12] A. G. Ramm, DSM for ill-posed equations with monotone operators, *Comm. in Nonlinear Sci. and Numer. Simulation*, 10, N8, (2005), 935-940.

[13] A. G. Ramm, Discrepancy principle for the dynamical systems method, *Communic. in Nonlinear Sci. and Numer. Simulation*, 10, N1, (2005), 95-101

[14] A. G. Ramm, Dynamical systems method (DSM) and nonlinear problems, in the book: *Spectral Theory and Nonlinear Analysis*, World Scientific Publishers, Singapore, 2005, 201-228. (ed J. Lopez-Gomez).

[15] A. G. Ramm, Dynamical systems method (DSM) for unbounded operators, *Proc. Amer. Math. Soc.*, 134, N4, (2006), 1059-1063.

[16] M. M. Vainberg, *Variational methods and method of monotone operators in the theory of nonlinear equations*, Wiley, London, 1973.

Mathematics Department, Kansas State University, Manhattan, KS 66506-2602, USA
E-mail address: nguyenhs@math.ksu.edu

Mathematics Department, Kansas State University, Manhattan, KS 66506-2602, USA
E-mail address: ramm@math.ksu.edu

Chapter 5

An iterative scheme for solving equations with monotone operators

Published in BIT, 48, N4, (2008), 725-741.

AN ITERATIVE SCHEME FOR SOLVING NONLINEAR EQUATIONS WITH MONOTONE OPERATORS

N. S. HOANG[1] and A. G. RAMM[2]

[1] *Department of Mathematics, Kansas State University, Manhattan, KS 66502, USA. email: nguyenhs@math.ksu.edu*

[2] *Department of Mathematics, Kansas State University, Manhattan, KS 66502, USA. email: ramm@math.ksu.edu*

Abstract.

An iterative scheme for solving ill-posed nonlinear operator equations with monotone operators is introduced and studied in this paper. A discrete version of the Dynamical Systems Method (DSM) algorithm for stable solution of ill-posed operator equations with monotone operators is proposed and its convergence is proved. A discrepancy principle is proposed and justified. *A priori* and *a posteriori* stopping rules for the iterative scheme are formulated and justified.

AMS subject classification (2000): 47J05, 47J06, 47J35, 65R30.

Key words: Dynamical systems method (DSM), nonlinear operator equations, monotone operators, discrepancy principle..

1 Introduction

In this paper we study a discrete version of the Dynamical Systems Method (DSM) for solving the equation

$$F(u) = f, \tag{1.1}$$

where F is a nonlinear twice Fréchet differentiable monotone operator in a real Hilbert space H, and equation (1.1) is assumed solvable. Monotonicity is understood in the following sense:

$$\langle F(u) - F(v), u - v\rangle \geq 0, \quad \forall u, v \in H. \tag{1.2}$$

Here $\langle u, v\rangle$ denotes the inner product in H. It is known (see, e.g., [6]), that the set $\mathcal{N} := \{u : F(u) = f\}$ is closed and convex if F is monotone and continuous. A closed and convex set in a Hilbert space has a unique minimal-norm element. This element in $\mathcal{N}$ we denote by y, $F(y) = f$. We assume that

$$\sup_{\|u-u_0\|\leq R} \|F^{(j)}(u)\| \leq M_j = M_j(u_0, R), \quad 0 \leq j \leq 2, \tag{1.3}$$

where $F^{(j)}(u)$ is the $j-$th Fréchet derivative of F at the point $u \in H$, $u_0 \in H$ is an element of H, $R > 0$ is arbitrary, and $f = F(y)$ is not known but f_δ, the noisy data, are known and $\|f_\delta - f\| \le \delta$. If $F'(u)$ is not boundedly invertible then solving for u given noisy data f_δ is often (but not always) an ill-posed problem.

Our goal is to develop an iterative process discrepancy principle type for a stable solution of equation (1.1), given noisy data f_δ, $\|f - f_\delta\| \le \delta$. In [6] a general approach to construction of convergent iterative processes for solving (1.1) on the basis of the DSM is developed. Some results on the DSM and its applications one finds in [2], [6]–[13]. In [3]–[6] and references therein methods for solving ill-posed problems are discussed.

Although the DSM is presented in detail in the monograph [6], we briefly give its main idea for convenience of the reader. The idea of solving equation (1.1) by a version of the DSM consists of finding a nonlinear map $\Phi(t,u)$, such that:

a) The Cauchy problem:

$$\dot{u}(t) = \Phi(t,u), \quad u(0) = u_0,$$

has a unique global solution,

b) There exists the limit:

$$\lim_{t\to\infty} u(t) := u(\infty),$$

and this limit solves (1.1):

c)

$$F(u(\infty)) = f.$$

Several versions of DSM were proposed and justified mathematically in [6]–[12].

In this paper the following iterative scheme for stable solution to (1.1) is investigated:

$$u_{n+1} = u_n - A_n^{-1}[F(u_n) + a_n u_n - f_\delta], \quad A_n := F'(u_n) + a_n I, \quad u_0 = u_0.$$

For this iterative scheme we formulate and justify an *a posteriori* stopping rule based on a discrepancy principle:

$$\|F(u_{n_\delta}) - f_\delta\| \le C_1\delta^\gamma, \quad C_1\delta^\gamma < \|F(u_n) - f_\delta\|, \quad \forall n < n_\delta,$$

where $C_1 > 1$, $0 < \gamma \le 1$. The existence of n_δ and the convergence of u_{n_δ} to a solution of equation (1.1) are justified provided that u_0 and a_n are suitably chosen (see Theorem 2.6).

The novel points in this paper are formulated in Lemmas 2.1, 2.3, 2.4, 2.5, and in Theorem 2.6. The ideas of the proofs of these results are new and these results have no intersection with the results in the published literature and with the results in the papers, mentioned in the references. The new discrepancy principle, stated in Theorem 2.6 and justified in the proof of this main Theorem may look similar to the well-known Morozov's discrepancy principle (with $\gamma = 1$) for linear equations, but in fact it is a completely different principle both in its proof and in its numerical application. Its proof is completely different from the proof of Morozov's principle because we do not use variational regularization, and our problem is fully nonlinear in the sense that no restriction on the global growth of the nonlinearity are made. The essential practical difference of our discrepancy principle from Morozov's principle consists of the following: in Morozov's principle one has to solve a nonlinear equation for the regularization parameter, while in our principle the "stopping rule", that is, the choice of n_δ is made automatically. Our results are new not only for nonlinear equations but for linear equations as well. Note that solving the nonlinear equation for the regularization parameter in Morozov's principle is by itself a non-trivial and time consuming task.

If $\gamma = 1$, then, in general, one cannot prove convergence to the minimal-norm solution y even for linear equations $Au = f$ regularized by the method $(A + a)u = f$, where $A \geq 0$ is a linear operator in H and $a > 0$ is the regularization parameter (see [5, p. 29]).

2 Auxiliary and main results

2.1 Auxiliary results

Let us consider the following equation:

$$F(\tilde{V}_{a,\delta}) + a\tilde{V}_{a,\delta} - f_\delta = 0, \qquad a > 0. \tag{2.1}$$

It is known (see, e.g., [1] and [6]) that equation (2.1) with monotone continuous operator F has a unique solution for any fixed $a > 0$ and $f_\delta \in H$.

Lemma 2.1. *If* (1.2) *holds and* F *is continuous, then* $\|\tilde{V}_{a,\delta}\| = O(\frac{1}{a})$ *as* $a \to \infty$, *and*

$$\lim_{a\to\infty} \|F(\tilde{V}_{a,\delta}) - f_\delta\| = \|F(0) - f_\delta\|. \tag{2.2}$$

Proof. Rewrite (2.1) as

$$F(\tilde{V}_{a,\delta}) - F(0) + a\tilde{V}_{a,\delta} + F(0) - f_\delta = 0.$$

Multiply this equation by $\tilde{V}_{a,\delta}$, use the inequality $\langle F(\tilde{V}_{a,\delta}) - F(0), \tilde{V}_{a,\delta} - 0\rangle \geq 0$, which follows from (1.2), and get:

$$a\|\tilde{V}_{a,\delta}\|^2 \leq \langle a\tilde{V}_{a,\delta} + F(\tilde{V}_{a,\delta}) - F(0), \tilde{V}_{a,\delta}\rangle = \langle f_\delta - F(0), \tilde{V}_{a,\delta}\rangle \leq \|f_\delta - F(0)\|\|\tilde{V}_{a,\delta}\|.$$

Therefore, $\|\tilde{V}_{a,\delta}\| = O(\frac{1}{a})$. This and the continuity of F imply (2.2). □

Let us recall the following result (see Lemma 6.1.7 [6, p. 112]):

LEMMA 2.2. *Assume that equation* (1.1) *is solvable. Let* y *be its minimal-norm solution. Assume that conditions* (1.2) *and* (1.3) *hold. Then*

$$\lim_{a\to 0}\|\tilde{V}_a - y\| = 0,$$

where $\tilde{V}_a := \tilde{V}_{a,0}$ *which solves* (2.1) *with* $\delta = 0$.

Let us consider the following equation

$$F(V_{n,\delta}) + a_n V_{n,\delta} - f_\delta = 0, \qquad a_n > 0, \tag{2.3}$$

and denote $V_n := V_{n,\delta}$ when $\delta \neq 0$. From the triangle inequality one gets:

$$\|F(V_0) - f_\delta\| \geq \|F(0) - f_\delta\| - \|F(V_0) - F(0)\|.$$

From the inequality $\|F(V_0) - F(0)\| \leq M_1\|V_0\|$ and Lemma 2.1 it follows that for large a_0 one has:

$$\|F(V_0) - F(0)\| \leq M_1\|V_0\| = O\Big(\frac{1}{a_0}\Big),$$

where $V_0 = \tilde{V}_{a_0,\delta}$. Therefore, if $\|F(0) - f_\delta\| > C\delta$, then $\|F(V_0) - f_\delta\| \geq (C-\epsilon)\delta$, where $\epsilon > 0$ is arbitrarily small for sufficiently large $a_0 > 0$.

LEMMA 2.3. *Suppose that* $\|F(0) - f_\delta\| > C\delta,\ C > 1$. *Assume that* $0 < (a_n)_{n=0}^\infty \searrow 0$, *and* a_0 *is sufficiently large. Then, there exists a unique* $n_\delta > 0$, *such that*

$$\|F(V_{n_\delta}) - f_\delta\| \leq C\delta < \|F(V_n) - f_\delta\|, \quad \forall n < n_\delta. \tag{2.4}$$

PROOF. We have $F(y) = f$, and

$$\begin{aligned}
0 =&\langle F(V_n) + a_n V_n - f_\delta, F(V_n) - f_\delta\rangle\\
=&\|F(V_n) - f_\delta\|^2 + a_n\langle V_n - y, F(V_n) - f_\delta\rangle + a_n\langle y, F(V_n) - f_\delta\rangle\\
=&\|F(V_n) - f_\delta\|^2 + a_n\langle V_n - y, F(V_n) - F(y)\rangle + a_n\langle V_n - y, f - f_\delta\rangle\\
&+ a_n\langle y, F(V_n) - f_\delta\rangle\\
\geq&\|F(V_n) - f_\delta\|^2 + a_n\langle V_n - y, f - f_\delta\rangle + a_n\langle y, F(V_n) - f_\delta\rangle.
\end{aligned}$$

Here the inequality $\langle V_n - y, F(V_n) - F(y)\rangle \geq 0$ was used. Therefore

$$\begin{aligned}
\|F(V_n) - f_\delta\|^2 &\leq -a_n\langle V_n - y, f - f_\delta\rangle - a_n\langle y, F(V_n) - f_\delta\rangle\\
&\leq a_n\|V_n - y\|\|f - f_\delta\| + a_n\|y\|\|F(V_n) - f_\delta\| \qquad (2.5)\\
&\leq a_n\delta\|V_n - y\| + a_n\|y\|\|F(V_n) - f_\delta\|.
\end{aligned}$$

On the other hand, one has:

$$\begin{aligned}0 &= \langle F(V_n) - F(y) + a_n V_n + f - f_\delta, V_n - y\rangle \\ &= \langle F(V_n) - F(y), V_n - y\rangle + a_n \|V_n - y\|^2 + a_n\langle y, V_n - y\rangle + \langle f - f_\delta, V_n - y\rangle \\ &\geq a_n\|V_n - y\|^2 + a_n\langle y, V_n - y\rangle + \langle f - f_\delta, V_n - y\rangle,\end{aligned}$$

where the inequality $\langle V_n - y, F(V_n) - F(y)\rangle \geq 0$ was used. Therefore,

$$a_n\|V_n - y\|^2 \leq a_n\|y\|\|V_n - y\| + \delta\|V_n - y\|.$$

This implies

$$a_n\|V_n - y\| \leq a_n\|y\| + \delta. \tag{2.6}$$

From (2.5) and (2.6), and an elementary inequality $ab \leq \epsilon a^2 + \frac{b^2}{4\epsilon}$, $\forall \epsilon > 0$, one gets:

$$\begin{aligned}\|F(V_n) - f_\delta\|^2 &\leq \delta^2 + a_n\|y\|\delta + a_n\|y\|\|F(V_n) - f_\delta\| \\ &\leq \delta^2 + a_n\|y\|\delta + \epsilon\|F(V_n) - f_\delta\|^2 + \frac{1}{4\epsilon}a_n^2\|y\|^2,\end{aligned} \tag{2.7}$$

where $\epsilon > 0$ is fixed, independent of n, and can be chosen arbitrary small. Let $n \to \infty$ so $a_n \searrow 0$. Then (2.7) implies $\limsup_{n\to\infty}(1-\epsilon)\|F(V_n)-f_\delta\|^2 \leq \delta^2$, $\forall\, \epsilon > 0$. This implies $\limsup_{n\to\infty}\|F(V_n)-f_\delta\| \leq \delta$. This, the assumption $\|F(0) - f_\delta\| > C\delta$, and the fact that $\|F(V_n) - f_\delta\|$ is nonincreasing (see Lemma 2.4), imply that there exists a unique $n_\delta > 0$ such that (2.4) holds. Lemma 2.3 is proved. □

Remark 2.1. Let $V_{0,n} := V_{\delta,n}|_{\delta=0}$. Then $F(V_{0,n}) + a_n V_{0,n} - f = 0$. Note that we have

$$\|V_{\delta,n} - V_{0,n}\| \leq \frac{\delta}{a_n}. \tag{2.8}$$

Indeed, from (2.1) one gets

$$F(V_{\delta,n}) - F(V_{0,n}) + a_n(V_{\delta,n} - V_{0,n}) = f - f_\delta.$$

Multiply this equality with $(V_{\delta,n} - V_{0,n})$ and use (1.2) to get:

$$\begin{aligned}\delta\|V_{\delta,n} - V_{0,n}\| &\geq \langle f - f_\delta, V_{\delta,n} - V_{0,n}\rangle \\ &= \langle F(V_\delta, n) - F(V_{0,n}) + a_n(V_{\delta,n} - V_{0,n}), V_{\delta,n} - V_{0,n}\rangle \\ &\geq a_n\|V_{\delta,n} - V_{0,n}\|^2.\end{aligned}$$

This implies (2.8). Similarly, from the equation

$$F(V_{0,n}) + a_n V_{0,n} - F(y) = 0,$$

one can derive that

$$\|V_{0,n}\| \le \|y\|. \tag{2.9}$$

Similar arguments one can find in [6].

From (2.8) and (2.9), one gets the following estimate:

$$\|V_n\| \le \|V_{0,n}\| + \frac{\delta}{a_n} \le \|y\| + \frac{\delta}{a_n}, \quad V_n := V_{\delta,n}. \tag{2.10}$$

Lemma 2.4. *Assume $\|F(0) - f_\delta\| > 0$. Let $0 < a_n \searrow 0$, and F be monotone. Denote*

$$h_n := \|F(V_n) - f_\delta\|, \quad k_n := \|V_n\|, \qquad n = 0, 1, ...,$$

where V_n solves (2.3). Then h_n is decreasing, and k_n is increasing.

Proof. Since $\|F(0) - f_\delta\| > 0$, it follows that $k_n \ne 0$, $\forall n \ge 0$. Note that $h_n = a_n\|V_n\|$. One has

$$\begin{aligned} 0 &\le \langle F(V_n) - F(V_m), V_n - V_m\rangle \\ &= \langle -a_nV_n + a_mV_m, V_n - V_m\rangle \\ &= (a_n + a_m)\langle V_n, V_m\rangle - a_n\|V_n\|^2 - a_m\|V_m\|^2. \end{aligned} \tag{2.11}$$

Thus,

$$\begin{aligned} 0 &\le (a_n + a_m)\langle V_n, V_m\rangle - a_n\|V_n\|^2 - a_m\|V_m\|^2 \\ &\le (a_n + a_m)\|V_n\|\|V_m\| - a_n\|V_n\|^2 - a_m\|V_m\|^2 \\ &= (a_n\|V_n\| - a_m\|V_m\|)(\|V_m\| - \|V_n\|) \\ &= (h_n - h_m)(k_m - k_n). \end{aligned} \tag{2.12}$$

If $k_m > k_n$ then (2.12) implies $h_n \ge h_m$, so

$$a_nk_n \ge a_mk_m > a_mk_n.$$

Thus, if $k_m > k_n$ then $a_m < a_n$ and, therefore, $m > n$, because a_n is decreasing.

Similarly, if $k_m < k_n$ then $h_n \le h_m$. This implies $a_m > a_n$, so $m < n$.

If $k_m = k_n$ then (2.11) implies

$$\|V_m\|^2 \le \langle V_m, V_n\rangle \le \|V_m\|\|V_n\| = \|V_m\|^2.$$

This implies $V_m = V_n$, and then $a_n = a_m$. Hence, $m = n$, because a_n is decreasing.

Therefore h_n is decreasing and k_n is increasing. Lemma 2.4 is proved. □

REMARK 2.2. From Lemma 2.1 and Lemma 2.4 one concludes that

$$a_n\|V_n\| = \|F(V_n) - f_\delta\| \le \|F(0) - f_\delta\|, \qquad \forall n \ge 0.$$

LEMMA 2.5. *Suppose* $M_1, c_0,$ *and* c_1 *are positive constants and* $0 \ne y \in H$*. Then there exist* $\lambda > 0$ *and a sequence* $0 < (a_n)_{n=0}^\infty \searrow 0$ *such that the following conditions hold*

$$a_n \le 2a_{n+1}, \tag{2.13}$$

$$\|f_\delta - F(0)\| \le \frac{a_0^2}{\lambda}, \tag{2.14}$$

$$\frac{M_1}{\lambda} \le \|y\|, \tag{2.15}$$

$$\frac{a_n - a_{n+1}}{a_{n+1}^2} \le \frac{1}{2c_1\lambda}, \tag{2.16}$$

$$c_0\frac{a_n}{\lambda^2} + \frac{a_n - a_{n+1}}{a_{n+1}}c_1 \le \frac{a_{n+1}}{\lambda}. \tag{2.17}$$

PROOF. Let us show that if $0 < a_0$ is sufficiently large then the following sequence

$$a_n = \frac{a_0}{1+n}, \tag{2.18}$$

satisfy conditions (2.13)–(2.17). One has

$$\frac{a_n}{a_{n+1}} = \frac{n+2}{n+1} \le 2, \qquad \forall\, n \ge 0.$$

Thus, inequality (2.13) is obtained.

Choose

$$\lambda \ge \frac{M_1}{\|y\|}. \tag{2.19}$$

Then inequality (2.15) is satisfied.

Inequality (2.14) is obtained if a_0 is sufficiently large. Indeed, (2.14) holds if

$$a_0 \ge \sqrt{\lambda\|f_\delta - F(0)\|}. \tag{2.20}$$

Let us check inequality (2.16). One has

$$\frac{a_n - a_{n+1}}{a_{n+1}^2} = \left(\frac{a_0}{1+n} - \frac{a_0}{2+n}\right)\frac{(n+2)^2}{a_0^2} = \frac{n+2}{a_0(n+1)} \le \frac{2}{a_0}, \quad n \ge 0.$$

Thus, (2.16) holds if

$$\frac{2}{a_0} \le \frac{1}{2c_1\lambda}, \tag{2.21}$$

i.e., if a_0 is sufficiently large.

Let us verify inequality (2.17). Assume that $(a_n)_{n=0}^\infty$ and λ satisfy (2.13)–(2.16) and (2.18). Choose $\kappa \geq 1$ such that

$$\frac{2c_0}{\kappa\lambda} \leq \frac{1}{2}. \tag{2.22}$$

Consider the sequence $(b_n)_{n=0}^\infty := (\kappa a_n)_{n=0}^\infty$ and let $\lambda_\kappa := \kappa\lambda$. Using inequalities (2.13), (2.16) and (2.22), one gets

$$\begin{aligned} c_0\frac{b_n}{\lambda_\kappa^2} + \frac{b_n - b_{n+1}}{b_{n+1}}c_1 &= \frac{2c_0}{\kappa\lambda}\frac{a_n}{2\lambda} + \frac{a_n - a_{n+1}}{a_{n+1}}c_1 \\ &\leq \frac{1}{2}\frac{a_{n+1}}{\lambda} + \frac{a_{n+1}}{2\lambda} = \frac{a_{n+1}}{\lambda} = \frac{b_{n+1}}{\lambda_\kappa}. \end{aligned}$$

Thus, inequality (2.17) holds for a_n replaced by $b_n = \kappa a_n$ and λ replaced by $\lambda_\kappa = \kappa\lambda$, where $\kappa \geq \max(1, \frac{4c_0}{\lambda})$ (see (2.22)). Inequalities (2.13)–(2.16) hold as well under this transformation. Thus, the choices $a_n = \frac{a_0\kappa}{n+1}$ and $\lambda := \kappa\frac{M_1}{\|y\|}$, $\kappa \geq \max(1, \frac{4c_0\|y\|}{M_1})$, satisfy all the conditions of Lemma 2.5. $\Box$

Remark 2.3. Using similar arguments one can show that the choices $\lambda > 0$, $a_n = \frac{d_0}{(n+1)^b}$, $d_0 \geq 1$, $0 < b \leq 1$, satisfy all conditions of Lemma 2.5 provided that d_0 is sufficiently large and λ is chosen so that inequality (2.19) holds.

Remark 2.4. In the proof of Lemma 2.5 a_0 and λ can be chosen so that $\frac{a_0}{\lambda}$ is uniformly bounded as $\delta \to 0$ regardless of the rate of growth of the constant $M_1 = M_1(R)$ from formula (1.3) when $R \to \infty$, i.e., regardless of the strength of the nonlinearity $F(u)$.

Indeed, to satisfy (2.19) one can choose $\lambda = \frac{M_1}{\|y\|}$. To satisfy (2.20) and (2.21) one can choose

$$a_0 = \max\left(\sqrt{\lambda\|f_\delta - F(0)\|}, 4c_1\lambda\right) \leq \max\left(\sqrt{\lambda(\|f - F(0)\| + 1)}, 4c_1\lambda\right),$$

where we have assumed without loss of generality that $0 < \delta < 1$. With this choice of a_0 and λ, the ratio $\frac{a_0}{\lambda}$ is bounded uniformly with respect to $\delta \in (0,1)$ and does not depend on R.

Indeed, with the above choice one has $\frac{a_0}{\lambda} \leq c(1+\sqrt{\lambda^{-1}}) \leq c$, where $c > 0$ is a constant independent of δ, and one can assume that $\lambda \geq 1$ without loss of generality.

This Remark is used in the proof of main result in Section 2.2. Specifically, it will be used to prove that an iterative process (2.24) generates a sequence which stays in a ball $B(u_0, R)$ for all $n \leq n_0 + 1$, where the number n_0 is defined by formula (2.33) (see below), and $R > 0$ is sufficiently large. An upper bound on R is given in the proof of Theorem 2.6, below formula (2.46).

Remark 2.5. It is easy to choose $u_0 \in H$ such that

$$g_0 := \|u_0 - V_0\| \leq \frac{\|F(0) - f_\delta\|}{a_0}. \tag{2.23}$$

Indeed, if, for example, $u_0 = 0$, then by Remark 2.2 one gets

$$g_0 = \|V_0\| = \frac{a_0\|V_0\|}{a_0} \le \frac{\|F(0) - f_\delta\|}{a_0}.$$

If (2.14) and (2.23) hold then $g_0 \le \frac{a_0}{\lambda}$.

2.2 Main result

Recall that $V_n := V_{n,\delta}$, and

$$F(V_{n,\delta}) + a_n V_{n,\delta} - f_\delta = 0.$$

Consider the following iterative scheme:

$$u_{n+1} = u_n - A_n^{-1}[F(u_n) + a_n u_n - f_\delta], \quad A_n := F'(u_n) + a_n I, \quad u_0 = u_0, \tag{2.24}$$

where u_0 is chosen so that inequality (2.23) holds. Note that $F'(u_n) \ge 0$ since F is monotone. Thus, $\|A_n^{-1}\| \le \frac{1}{a_n}$.

Let a_n and λ satisfy conditions (2.13)–(2.17). Assume that equation $F(u) = f$ has a solution $y \in B(u_0, R)$, possibly nonunique, and y is the minimal-norm solution to this equation. Let f be unknown but f_δ be given, and $\|f_\delta - f\| \le \delta$. We have the following result:

THEOREM 2.6. *Assume $a_n = \frac{d_0}{(d+n)^b}$ where $d \ge 1$, $0 < b \le 1$, and d_0 is sufficiently large so that conditions* (2.13)–(2.17) *hold. Let u_n be defined by* (2.24)*. Assume that u_0 is chosen so that* (2.23) *holds and $\|F(u_0) - f_\delta\| > C_1\delta^\gamma > \delta$. Then there exists a unique n_δ, depending on C_1 and γ (see below), such that*

$$\|F(u_{n_\delta}) - f_\delta\| \le C_1\delta^\gamma, \quad C_1\delta^\gamma < \|F(u_n) - f_\delta\|, \quad \forall n < n_\delta, \tag{2.25}$$

where $C_1 > 1$, $0 < \gamma \le 1$.

Let $0 < (\delta_m)_{m=1}^\infty$ be a sequence such that $\delta_m \to 0$. If N is a cluster point of the sequence n_{δ_m} satisfying (2.25)*, then*

$$\lim_{m\to\infty} u_{n_{\delta_m}} = u^*, \tag{2.26}$$

where u^* is a solution to the equation $F(u) = f$. If

$$\lim_{m\to\infty} n_{\delta_m} = \infty, \tag{2.27}$$

where $\gamma \in (0, 1)$, then

$$\lim_{m\to\infty} \|u_{n_{\delta_m}} - y\| = 0. \tag{2.28}$$

Proof. Denote

$$C := \frac{C_1 + 1}{2}. \tag{2.29}$$

Let

$$z_n := u_n - V_n, \quad g_n := \|z_n\|.$$

We use Taylor's formula and get:

$$F(u_n) - F(V_n) + a_n z_n = A_{a_n} z_n + K_n, \quad \|K_n\| \le \frac{M_2}{2}\|z_n\|^2, \tag{2.30}$$

where $K_n := F(u_n) - F(V_n) - F'(u_n)z_n$ and M_2 is the constant from (1.3). From (2.24) and (2.30) one obtains

$$z_{n+1} = z_n - z_n - A_n^{-1}K(z_n) - (V_{n+1} - V_n). \tag{2.31}$$

From (2.31), (2.30), and the estimate $\|A_n^{-1}\| \le \frac{1}{a_n}$, one gets

$$g_{n+1} \le \frac{M_2 g_n^2}{2a_n} + \|V_{n+1} - V_n\|. \tag{2.32}$$

Since $0 < a_n \searrow 0$, for any fixed $\delta > 0$ there exists n_0 such that

$$\frac{\delta}{a_{n_0+1}} > \frac{1}{C-1}\|y\| \ge \frac{\delta}{a_{n_0}}, \qquad C > 1. \tag{2.33}$$

By (2.13), one has $\frac{a_n}{a_{n+1}} \le 2$, $\forall\, n \ge 0$. This and (2.33) imply

$$\frac{2}{C-1}\|y\| \ge \frac{2\delta}{a_{n_0}} > \frac{\delta}{a_{n_0+1}} > \frac{1}{C-1}\|y\| \ge \frac{\delta}{a_{n_0}}, \qquad C > 1. \tag{2.34}$$

Thus,

$$\frac{2}{C-1}\|y\| > \frac{\delta}{a_n}, \quad \forall n \le n_0 + 1. \tag{2.35}$$

The number n_0, satisfying (2.35), exists and is unique since $a_n > 0$ monotonically decays to 0 as $n \to \infty$. By Lemma 2.3, there exists a number n_1 such that

$$\|F(V_{n_1+1}) - f_\delta\| \le C\delta < \|F(V_{n_1}) - f_\delta\|, \tag{2.36}$$

where V_n solves the equation $F(V_n) + a_n V_n - f_\delta = 0$. *We claim that* $n_1 \in [0, n_0]$. Indeed, one has $\|F(V_{n_1}) - f_\delta\| = a_{n_1}\|V_{n_1}\|$, and $\|V_{n_1}\| \le \|y\| + \frac{\delta}{a_{n_1}}$ (cf. (2.10)), so

$$C\delta < a_{n_1}\|V_{n_1}\| \le a_{n_1}\left(\|y\| + \frac{\delta}{a_{n_1}}\right) = a_{n_1}\|y\| + \delta, \quad C > 1. \tag{2.37}$$

Therefore,

$$\delta < \frac{a_{n_1}\|y\|}{C-1}. \tag{2.38}$$

Thus, by (2.34),

$$\frac{\delta}{a_{n_1}} < \frac{\|y\|}{C-1} < \frac{\delta}{a_{n_0+1}}. \tag{2.39}$$

Here the last inequality is a consequence of (2.34). Since a_n decreases monotonically, inequality (2.39) implies $n_1 \le n_0$. One has

$$\begin{aligned} a_{n+1}\|V_n - V_{n+1}\|^2 &= \langle (a_{n+1} - a_n)V_n - F(V_n) + F(V_{n+1}), V_n - V_{n+1}\rangle \\ &\le \langle (a_{n+1} - a_n)V_n, V_n - V_{n+1}\rangle \\ &\le (a_n - a_{n+1})\|V_n\|\|V_n - V_{n+1}\|. \end{aligned} \tag{2.40}$$

By (2.10), $\|V_n\| \le \|y\| + \frac{\delta}{a_n}$, and, by (2.35), $\frac{\delta}{a_n} \le \frac{2\|y\|}{C-1}$ for all $n \le n_0 + 1$. Therefore,

$$\|V_n\| \le \|y\|\bigg(1 + \frac{2}{C-1}\bigg), \qquad \forall n \le n_0 + 1, \tag{2.41}$$

and, by (2.40),

$$\|V_n - V_{n+1}\| \le \frac{a_n - a_{n+1}}{a_{n+1}}\|V_n\| \le \frac{a_n - a_{n+1}}{a_{n+1}}\|y\|\bigg(1 + \frac{2}{C-1}\bigg), \quad \forall n \le n_0 + 1. \tag{2.42}$$

Inequalities (2.32) and (2.42) imply

$$g_{n+1} \le \frac{c_0}{a_n}g_n^2 + \frac{a_n - a_{n+1}}{a_{n+1}}c_1, \quad c_0 = \frac{M_2}{2}, \quad c_1 = \|y\|\bigg(1 + \frac{2}{C-1}\bigg), \tag{2.43}$$

for all $n \le n_0 + 1$.

By Lemma 2.5 and Remark 2.3, the sequence $(a_n)_{n=1}^\infty$, satisfies conditions (2.13)–(2.17), provided that d_0 is sufficiently large and $\lambda > 0$ is chosen so that (2.19) holds. Let us show by induction that

$$g_n < \frac{a_n}{\lambda}, \qquad 0 \le n \le n_0 + 1. \tag{2.44}$$

Inequality (2.44) holds for $n = 0$ by Remark 2.5. Suppose (2.44) holds for some $n \ge 0$. From (2.43), (2.44) and (2.17), one gets

$$\begin{aligned} g_{n+1} &\le \frac{c_0}{a_n}\bigg(\frac{a_n}{\lambda}\bigg)^2 + \frac{a_n - a_{n+1}}{a_{n+1}}c_1 \\ &= \frac{c_0 a_n}{\lambda^2} + \frac{a_n - a_{n+1}}{a_{n+1}}c_1 \\ &\le \frac{a_{n+1}}{\lambda}. \end{aligned} \tag{2.45}$$

Thus, by induction, inequality (2.44) holds for all n in the region $0 \le n \le n_0 + 1$.

From Remark 2.1 one has $\|V_n\| \le \|y\| + \frac{\delta}{a_n}$. This and the triangle inequality imply

$$\|u_0 - u_n\| \le \|u_0\| + \|z_n\| + \|V_n\| \le \|u_0\| + \|z_n\| + \|y\| + \frac{\delta}{a_n}. \tag{2.46}$$

Inequalities (2.41), (2.44), and (2.46) guarantee that the sequence u_n, generated by the iterative process (2.24), remains in the ball $B(u_0,R)$ for all $n \le n_0+1$, where $R \le \frac{a_0}{\lambda} + \|u_0\| + \|y\| + \frac{\delta}{a_n}$. This inequality and the estimate (2.35) imply that the sequence u_n, $n \le n_0+1$, stays in the ball $B(u_0,R)$, where

$$R \le \frac{a_0}{\lambda} + \|u_0\| + \|y\| + \|y\|\frac{C+1}{C-1}.$$

By Remark 2.4, one can choose a_0 and λ so that $\frac{a_0}{\lambda}$ is uniformly bounded as $\delta \to 0$ even if $M_1(R) \to \infty$ as $R \to \infty$ at an arbitrary fast rate. Thus, the sequence u_n stays in the ball $B(u_0,R)$ for $n \le n_0+1$ when $\delta \to 0$. An upper bound on R is given above. It does not depend on δ as $\delta \to 0$.

One has:

$$\begin{aligned}\|F(u_n)-f_\delta\| \le&\|F(u_n)-F(V_n)\| + \|F(V_n)-f_\delta\|\\ \le&M_1 g_n + \|F(V_n)-f_\delta\| \\ \le&\frac{M_1 a_n}{\lambda} + \|F(V_n)-f_\delta\|, \qquad \forall n \le n_0+1,\end{aligned} \tag{2.47}$$

where (2.44) was used and M_1 is the constant from (1.3). Since $\|F(V_n)-f_\delta\|$ is nonincreasing, by Lemma 2.4, and $n_1 \le n_0$, one gets

$$\|F(V_{n_0+1})-f_\delta\| \le \|F(V_{n_1+1})-f_\delta\| \le C\delta. \tag{2.48}$$

From (2.15), (2.47), (2.48), the relation (2.33), and the definition $C_1 = 2C-1$ (see (2.29)), one concludes that

$$\begin{aligned}\|F(u_{n_0+1})-f_\delta\| \le&\frac{M_1 a_{n_0+1}}{\lambda} + C\delta\\ \le&\frac{M_1\delta(C-1)}{\lambda\|y\|} + C\delta \le (2C-1)\delta = C_1\delta.\end{aligned} \tag{2.49}$$

Thus, if

$$\|F(u_0)-f_\delta\| > C_1\delta^\gamma, \quad 0<\gamma\le 1,$$

then one concludes from (2.49) *that there exists* n_δ, $0 < n_\delta \le n_0+1$, *such that*

$$\|F(u_{n_\delta})-f_\delta\| \le C_1\delta^\gamma < \|F(u_n)-f_\delta\|, \quad 0 \le n < n_\delta, \tag{2.50}$$

for any given $\gamma \in (0,1]$, *and any fixed* $C_1 > 1$.

Let us prove (2.26). If $n>0$ is fixed, then $u_{\delta,n}$ is a continuous function of f_δ. Denote

$$\tilde{u}_N = \lim_{\delta\to 0} u_{\delta,N}, \tag{2.51}$$

where $N < \infty$ is a cluster point of n_{δ_m}, so that there exists a subsequence of n_{δ_m}, which we denote by n_m, such that

$$\lim_{m\to\infty} n_m = N.$$

From (2.51) and the continuity of F, one obtains:

$$\|F(\tilde{u}_N) - f\| = \lim_{m\to\infty} \|F(u_{n_{\delta_m}}) - f_{\delta_m}\| \le \lim_{m\to\infty} C_1\delta_m^\gamma = 0.$$

Thus, $\tilde{u}_N$ is a solution to the equation $F(u) = f$, and (2.26) is proved.

Let us prove (2.28) *assuming that* (2.27) *holds.* From (2.25) and (2.47) with $n = n_\delta - 1$, and from (2.50), one gets

$$C_1\delta^\gamma \le M_1\frac{a_{n_\delta-1}}{\lambda} + a_{n_\delta-1}\|V_{n_\delta-1}\| \le M_1\frac{a_{n_\delta-1}}{\lambda} + \|y\|a_{n_\delta-1} + \delta.$$

If $0 < \delta < 1$ and δ is sufficiently small, then

$$\tilde{C}\delta^\gamma \le a_{n_\delta-1}\Big(\frac{M_1}{\lambda} + \|y\|\Big), \quad \tilde{C} > 0,$$

where $\tilde{C}$ is a constant. Therefore, by (2.13),

$$\lim_{\delta\to0}\frac{\delta}{2a_{n_\delta}} \le \lim_{\delta\to0}\frac{\delta}{a_{n_\delta-1}} \le \lim_{\delta\to0}\frac{\delta^{1-\gamma}}{\tilde{C}}\Big(\frac{M_1}{\lambda} + \|y\|\Big) = 0, \quad 0 < \gamma < 1. \tag{2.52}$$

In particular, for $\delta = \delta_m$, one gets

$$\lim_{\delta_m\to0}\frac{\delta_m}{a_{n_{\delta_m}}} = 0. \tag{2.53}$$

From the triangle inequality, inequalities (2.8) and (2.44), one obtains

$$\begin{aligned}\|u_{n_{\delta_m}} - y\| &\le \|u_{n_{\delta_m}} - V_{n_{\delta_m}}\| + \|V_{n_{\delta_m}} - V_{n_{\delta_m},0}\| + \|V_{n_{\delta_m},0} - y\|\\ &\le \frac{a_{n_{\delta_m}}}{\lambda} + \frac{\delta_m}{a_{n_{\delta_m}}} + \|V_{n_{\delta_m},0} - y\|.\end{aligned} \tag{2.54}$$

Recall that $V_{n,0} = \tilde{V}_{a_n}$. From (2.27), (2.53), inequality (2.54) and Lemma 2.2, one obtains (2.28). Theorem 2.6 is proved. □

Remark 2.6. It is practically convenient to choose $u_0 = 0$. In this case inequality (2.23) holds and we assume that $\|F(0) - f_\delta\| > C_1\delta^\gamma > \delta$.

Remark 2.7. It follows from inequality (2.54) that the following rule:

$$a_{n_\delta} = O(\delta^\eta), \qquad 0 < \eta < 1, \tag{2.55}$$

can be used as an *a priori* choice of stopping rule for n_δ. Indeed, if n_δ is chosen as in equation (2.55) then, by inequality (2.54) with $n_{\delta_m} = n_\delta$, one gets

$$\lim_{\delta\to0}\|u_{n_\delta} - y\| = 0. \tag{2.56}$$

3 Numerical experiments

Let us present a numerical experiment solving nonlinear integral equation (1.1) with

$$F(u) := B(u) + u^2 \arctan(u) := \int_0^1 e^{-|x-y|} u(y)dy + u^2 \arctan(u). \tag{3.1}$$

The operator B is compact in $H = L^2[0,1]$. The operator $u \longmapsto u^2 \arctan(u)$ is defined on a dense subset D of of $L^2[0,1]$, for example, on $D := C[0,1]$. If $u, v \in D$, then

$$\langle u^2 \arctan u - v^2 \arctan v, u - v\rangle = \int_0^1 (u^2 \arctan u - v^2 \arctan v)(u - v)dx \geq 0.$$

Here we have used the fact that the function: $x^2 \arctan(x)$ is increasing on $\mathbb{R}$. Moreover,

$$e^{-|x|} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{e^{i\lambda x}}{1+\lambda^2} d\lambda.$$

Therefore, $\langle B(u - v), u - v\rangle \geq 0$, so

$$\langle F(u) - F(v), u - v\rangle \geq 0, \qquad \forall u, v \in D.$$

The Fréchet derivative of F is:

$$F'(u)h = \left(\frac{u^2}{1+u^2} + 2u \arctan(u)\right)h + \int_0^1 e^{-|x-y|} h(y)dy. \tag{3.2}$$

If $u(x)$ vanishes on a set of positive Lebesgue's measure, then $F'(u)$ is not boundedly invertible. If $u \in C[0,1]$ vanishes even at one point x_0, then $F'(u)$ is not boundedly invertible in H.

Let us use the iterative process (2.24):

$$\begin{aligned} u_{n+1} &= u_n - (F'(u_n) + a_n I)^{-1}(F(u_n) + a_n u_n - f_\delta), \\ u_0 &= 0. \end{aligned} \tag{3.3}$$

We stop iterations at $n := n_\delta$ such that the following inequality holds

$$\|F(u_{n_\delta}) - f_\delta\| < C\delta^\gamma, \quad \|F(u_n) - f_\delta\| \geq C\delta^\gamma, \quad n < n_\delta, \quad C > 1, \quad \gamma \in (0,1). \tag{3.4}$$

Integrals of the form $\int_0^1 e^{-|x-y|} h(y)dy$ in (3.1) and (3.2) are computed by using the trapezoidal rule. The noisy function used in the test is

$$f_\delta(x) = f(x) + \kappa f_{noise}(x), \quad \kappa = \kappa(\delta) > 0.$$

The noise level δ and the relative noise level are defined by

$$\delta = \kappa\|f_{noise}(x)\|, \quad \delta_{rel} := \frac{\delta}{\|f\|}.$$

In the test κ is computed in such a way that the relative noise level δ_{rel} equals to some desired value, i.e.,

$$\kappa = \frac{\delta}{\|f_{noise}(x)\|} = \frac{\delta_{rel}\|f\|}{\|f_{noise}\|}.$$

We have used the relative noise level as an input parameter in the test.

In all figures the x-variable runs through the interval $[0,1]$, and the graphs represent the numerical solutions $u_{DSM}(x)$ and the exact solution $u_{exact}(x)$.

In the test we have used $C = 1.01$ and $\gamma = 0.9$. As we have proved, the iterative scheme converges when $a_n = \frac{d}{1+n}$, and d is sufficiently large. However, in practice, if we choose d too large, then the method will use too many iterations before reaching the stopping time n_δ in (3.4). This means that the computation time will be large in this case. Since

$$\|F(V_{n_\delta}) - f_\delta\| = a_{n_\delta}\|V_{n_\delta}\|,$$

and $\|V_{n_\delta} - u_{n_\delta}\| = O(a_{n_\delta})$, we have

$$C\delta^\gamma = \|F(u_{n_\delta}) - f_\delta\| \sim a_{n_\delta}.$$

Thus, we choose

$$d = C_0\delta^\gamma, \qquad C_0 > 0.$$

In experiments we found that our method works well with $C_0 \in [1,4]$. Indeed, in the test we chose a_n by the formula $a_n := C_0\frac{\delta^{0.9}}{n+6}$. The number of node points used in computing integrals in (3.1) and (3.2) was $N = 100$. In all experiments, the noise function f_{noise} is a vector with random entries normally distributed of mean 0 and variance 1.

Numerical results for various values of δ_{rel} are presented in Table 7. Table 7 shows that the iterative scheme yields good numerical results.

Table 7: Results when $C_0 = 1$.

δ_{rel}	0.05	0.03	0.02	0.01	0.003	0.001
Number of iterations	10	10	10	10	11	11
$\frac{\|u_{DSM}-u_{exact}\|}{\|u_{exact}\|}$	0.0458	0.0273	0.0189	0.0094	0.0027	0.0009

Figure 10 plots the numerical results when relative noise levels are $\delta_{rel} = 0.01$ and $\delta_{rel} = 0.003$.

Figure 11 plots the numerical results when the noise levels are $\delta_{rel} = 0.05$ and $\delta_{rel} = 0.02$.

In computations the functions u, f and f_δ are vectors in $\mathbb{R}^N$ where N is the number of nodal points. The norm used in computations is the 2-norm of $\mathbb{R}^N$.

From the numerical results we conclude that the proposed stopping rule yields good results in this problem.

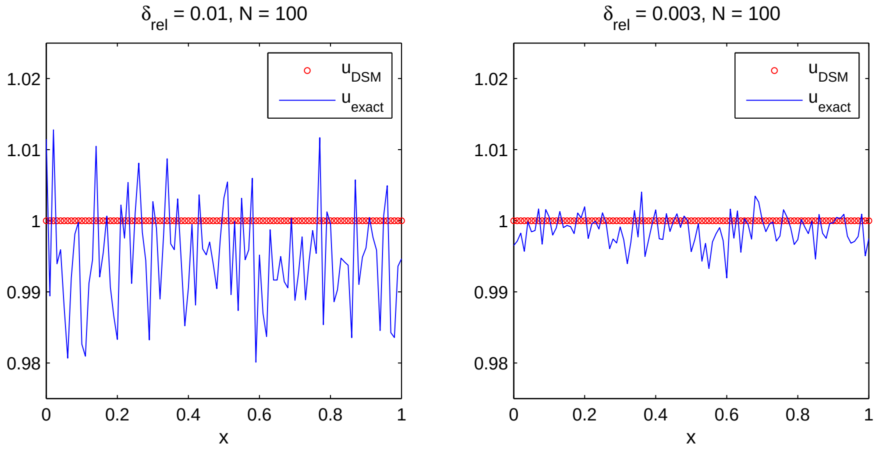

Figure 10: Plots of solutions obtained by the DSM when $N = 100$, $\delta_{rel} = 0.01$ (left) and $\delta_{rel} = 0.003$ (right).

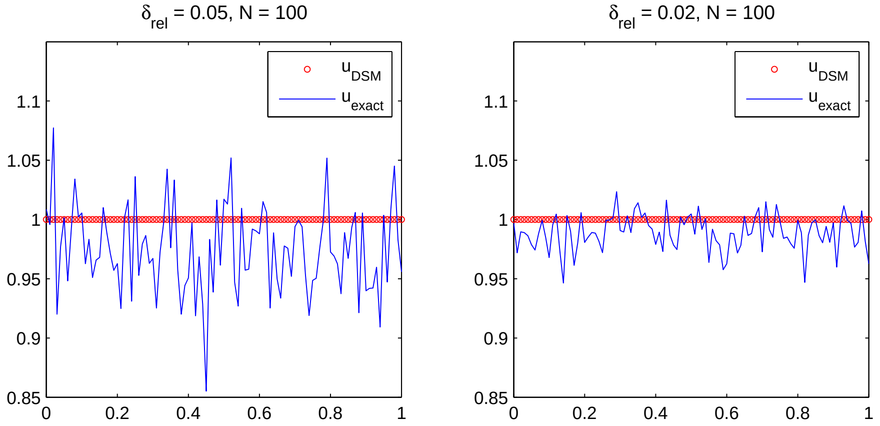

Figure 11: Plots of solutions obtained by the DSM when $N = 100$, $\delta_{rel} = 0.05$ (left) and $\delta_{rel} = 0.02$ (right).

REFERENCES

1. K. Deimling. *Nonlinear functional analysis.* Springer Verlag, Berlin, 1985.

2. N. S. Hoang and A. G. Ramm. Dynamical systems gradient method for solving ill-conditioned linear algebraic systems. (submitted).

3. V. Ivanov, V. Tanana and V. Vasin, *Theory of ill-posed problems.* VSP, Utrecht, 2002.

4. V. A. Morozov, *Methods of solving incorrectly posed problems.* Springer Verlag, New York, 1984.

5. A. G. Ramm, *Inverse problems.* Springer, New York, 2005.

6. A. G. Ramm, *Dynamical systems method for solving operator equations.* Elsevier, Amsterdam, 2007.

7. A. G. Ramm, Global convergence for ill-posed equations with monotone operators: the dynamical

systems method. *J. Phys A* 36, (2003), L249-L254.

8. A. G. Ramm, Dynamical systems method for solving nonlinear operator equations. *International Jour. of Applied Math. Sci.*, 1, N1, (2004), 97-110.

9. A. G. Ramm, Dynamical systems method for solving operator equations. *Communic. in Nonlinear Sci. and Numer. Simulation*, 9, N2, (2004), 383-402.

10. A. G. Ramm, DSM for ill-posed equations with monotone operators, *Comm. in Nonlinear Sci. and Numer. Simulation*, 10, N8, (2005),935-940.

11. A. G. Ramm, Discrepancy principle for the dynamical systems method, *Communic. in Nonlinear Sci. and Numer. Simulation*, 10, N1, (2005), 95-101

12. A. G. Ramm, Dynamical systems method (DSM) and nonlinear problems, in the book: *Spectral Theory and Nonlinear Analysis.* World Scientific Publishers, Singapore, 2005, 201-228. (ed J. Lopez-Gomez).

13. A. G. Ramm, Dynamical systems method (DSM) for unbounded operators, *Proc.Amer. Math. Soc.*, 134, N4, (2006), 1059-1063.

Chapter 6

A new version of the Dynamical systems method (DSM) for solving nonlinear quations with monotone operators

Published in Diff. Eq. Appl., 1, N1, (2009), 1-25.

A new version of the Dynamical Systems Method (DSM) for solving nonlinear equations with monotone operators

N. S. Hoang†* A. G. Ramm†‡

†Mathematics Department, Kansas State University,
Manhattan, KS 66506-2602, USA

Abstract

A version of the Dynamical Systems Method for solving ill-posed nonlinear monotone operator equations is studied in this paper. A discrepancy principle is proposed and justified. A numerical experiment was carried out with the new stopping rule. Numerical experiments show that the proposed stopping rule is efficient.

Mathematics Subject Classification. 47J05, 47J06, 47J35, 65R30

Keywords. Dynamical systems method (DSM), nonlinear operator equations, monotone operators, discrepancy principle.

1 Introduction

In this paper we study a version of the Dynamical Systems Method (DSM) for solving the equation

$$F(u) = f, \tag{1}$$

where F is a nonlinear, Fréchet differentiable, monotone operator in a real Hilbert space H, and equation (1) is assumed solvable, possibly nonuniquely. Monotonicity means that

$$\langle F(u) - F(v), u - v \rangle \geq 0, \quad \forall u, v \in H. \tag{2}$$

[*] Email: nguyenhs@math.ksu.edu

[‡] Corresponding author. Email: ramm@math.ksu.edu

It is known (see, e.g., [7]), that the set $\mathcal{N} := \{u : F(u) = f\}$ is closed and convex if F is monotone and continuous. A closed and convex set in a Hilbert space has a unique minimal-norm element. This element in $\mathcal{N}$ we denote by y, $F(y) = f$, and call it the minimal-norm solution to equation (1). We assume that

$$\sup_{\|u-u_0\|\le R} \|F'(u)\| \le M_1(R), \tag{3}$$

where $u_0 \in H$ is an element of H, $R > 0$ is arbitrary, and $f = F(y)$ is not known but f_δ, the noisy data, are known, and $\|f_\delta - f\| \le \delta$. If $F'(u)$ is not boundedly invertible then solving equation (1) for u given noisy data f_δ is often (but not always) an ill-posed problem. When F is a linear bounded operator many methods for stable solution of (1) were proposed (see [5]–[7] and references therein). However, when F is nonlinear then the theory is less complete.

DSM consists of finding a nonlinear map $\Phi(t, u)$ such that the Cauchy problem

$$\dot{u} = \Phi(t, u), \qquad u(0) = u_0,$$

has a unique solution for all $t \ge 0$, there exists $\lim_{t\to\infty} u(t) := u(\infty)$, and $F(u(\infty)) = f$,

$$\exists!\ u(t) \quad \forall t \ge 0; \qquad \exists u(\infty); \qquad F(u(\infty)) = f. \tag{4}$$

Various choices of Φ were proposed in [7] for (4) to hold. Each such choice yields a version of the DSM.

The DSM for solving equation (1) was extensively studied in [7]–[14]. In [7], the following version of the DSM was investigated for monotone operators F:

$$\dot{u}_\delta = -\big(F'(u_\delta) + a(t)I\big)^{-1}\big(F(u_\delta) + a(t)u_\delta - f_\delta\big), \quad u_\delta(0) = u_0. \tag{5}$$

The convergence of this method was justified with some *a apriori* choice of stopping rule. A DSM gradient method was formulated and justified in [4].

In this paper we consider a version of the DSM for solving equation (1):

$$\dot{u}_\delta = -\big(F(u_\delta) + a(t)u_\delta - f_\delta\big), \quad u_\delta(0) = u_0, \tag{6}$$

where F is a monotone operator.

The advantage of this version compared with (5) is the absence of the inverse operator in the algorithm, which makes the algorithm (6) less expensive than (5). On the other hand, algorithm (5) converges faster than (6) in many cases. The algorithm (6) is cheaper than the DSM gradient algorithm proposed in [4].

The convergence of the method (6) for any initial value u_0 is proved for a stopping rule based on a discrepancy principle. This *a posteriori* choice of stopping time t_δ is justified provided that $a(t)$ is suitably chosen.

The advantage of method (6), a modified version of the simple iteration method, over the Gauss-Newton method and the version (5) of the DSM is the following: neither inversion of matrices nor evaluation of F' is needed in a discretized version of (6). Although the convergence rate of the DSM (6) maybe slower than that of the DSM (5), the DSM (6) might be faster than the DSM (5) for large-scale systems due to its lower computation cost.

In this paper we investigate a stopping rule based on a discrepancy principle (DP) for the DSM (6). The main results of this paper are Theorem 17 and Theorem 19 in which a DP is formulated, the existence of a stopping time t_δ is proved, and the convergence of the DSM with the proposed DP is justified under some natural assumptions.

2 Auxiliary results

The inner product in H is denoted $\langle u, v\rangle$. Let us consider the following equation

$$F(V_\delta) + aV_\delta - f_\delta = 0, \qquad a > 0, \tag{7}$$

where $a = const$. It is known (see, e.g., [7], [15]) that equation (7) with monotone continuous operator F has a unique solution for any $f_\delta \in H$.

Let us recall the following result from [7]:

Lemma 1 *Assume that equation* (1) *is solvable, y is its minimal-norm solution, assumption* (2) *holds, and F is continuous. Then*

$$\lim_{a\to 0} \|V_a - y\| = 0,$$

where V_a solves (7) *with $\delta = 0$.*

Clearly, under our assumption (3), F is continuous.

Lemma 2 *If* (2) *holds and F is continuous, then $\|V_\delta\| = O(\frac{1}{a})$ as $a \to \infty$, and*

$$\lim_{a\to\infty} \|F(V_\delta) - f_\delta\| = \|F(0) - f_\delta\|. \tag{8}$$

Proof. Rewrite (7) as

$$F(V_\delta) - F(0) + aV_\delta + F(0) - f_\delta = 0.$$

Multiply this equation by V_δ, use inequality $\langle F(V_\delta) - F(0), V_\delta - 0\rangle \geq 0$ and get:

$$a\|V_\delta\|^2 \leq \|f_\delta - F(0)\|\|V_\delta\|.$$

Therefore,

$$\|V_\delta\| = O(\frac{1}{a}).$$

This and the continuity of F imply (8). □

Let $a = a(t)$ be a strictly monotonically decaying continuous positive function on $[0, \infty)$, $0 < a(t) \searrow 0$, and assume $a \in C^1[0, \infty)$. These assumptions hold throughout the paper and often are not repeated. Then the solution V_δ of (7) is a function of t, $V_\delta = V_\delta(t)$. From the triangle inequality one gets:

$$\|F(V_\delta(0)) - f_\delta\| \geq \|F(0) - f_\delta\| - \|F(V_\delta(0)) - F(0)\|.$$

From Lemma 2 it follows that for large $a(0)$ one has:

$$\|F(V_\delta(0)) - F(0)\| \leq M_1\|V_\delta(0)\| = O\biggl(\frac{1}{a(0)}\biggr).$$

Therefore, if $\|F(0) - f_\delta\| > C\delta$, then $\|F(V_\delta(0)) - f_\delta\| \geq (C - \epsilon)\delta$, where $\epsilon > 0$ is sufficiently small and $a(0) > 0$ is sufficiently large.

Below the words decreasing and increasing mean strictly decreasing and strictly increasing.

Lemma 3 *Assume $\|F(0) - f_\delta\| > 0$. Let $0 < a(t) \searrow 0$, and F be monotone. Denote*

$$\psi(t) := \|V_\delta(t)\|, \qquad \phi(t) := a(t)\psi(t) = \|F(V_\delta(t)) - f_\delta\|,$$

where $V_\delta(t)$ solves (7) with $a = a(t)$. Then $\phi(t)$ is decreasing, and $\psi(t)$ is increasing.

Proof. Since $\|F(0) - f_\delta\| > 0$, one has $\psi(t) \neq 0$, $\forall t \geq 0$. Indeed, if $\psi(t)\big|_{t=\tau} = 0$, then $V_\delta(\tau) = 0$, and equation (7) implies $\|F(0) - f_\delta\| = 0$, which is a contradiction. Note that $\phi(t) = a(t)\|V_\delta(t)\|$. One has

$$\begin{aligned} 0 &\leq \langle F(V_\delta(t_1)) - F(V_\delta(t_2)), V_\delta(t_1) - V_\delta(t_2)\rangle \\ &= \langle -a(t_1)V_\delta(t_1) + a(t_2)V_\delta(t_2), V_\delta(t_1) - V_\delta(t_2)\rangle \\ &= (a(t_1) + a(t_2))\langle V_\delta(t_1), V_\delta(t_2)\rangle - a(t_1)\|V_\delta(t_1)\|^2 - a(t_2)\|V_\delta(t_2)\|^2. \end{aligned} \tag{9}$$

Thus,

$$\begin{aligned}0 &\le (a(t_1)+a(t_2))\langle V_\delta(t_1), V_\delta(t_2)\rangle - a(t_1)\|V_\delta(t_1)\|^2 - a(t_2)\|V_\delta(t_2)\|^2 \\ &\le (a(t_1)+a(t_2))\|V_\delta(t_1)\|\|V_\delta(t_2)\| - a(t_1)\|V_\delta(t_1)\|^2 - a(t_2)\|V_\delta(t_2)\|^2 \\ &= (a(t_1)\|V_\delta(t_1)\| - a(t_2)\|V_\delta(t_2)\|)(\|V_\delta(t_2)\| - \|V_\delta(t_1)\|) \\ &= (\phi(t_1)-\phi(t_2))(\psi(t_2)-\psi(t_1)).\end{aligned} \tag{10}$$

If $\psi(t_2) > \psi(t_1)$ then (10) implies $\phi(t_1) \ge \phi(t_2)$, so

$$a(t_1)\psi(t_1) \ge a(t_2)\psi(t_2) > a(t_2)\psi(t_1).$$

Thus, if $\psi(t_2) > \psi(t_1)$ then $a(t_2) < a(t_1)$ and, therefore, $t_2 > t_1$, because $a(t)$ is strictly decreasing.

Similarly, if $\psi(t_2) < \psi(t_1)$ then $\phi(t_1) \le \phi(t_2)$. This implies $a(t_2) > a(t_1)$, so $t_2 < t_1$.

Suppose $\psi(t_1) = \psi(t_2)$, i.e., $\|V_\delta(t_1)\| = \|V_\delta(t_2)\|$. From (9), one has

$$\|V_\delta(t_1)\|^2 \le \langle V_\delta(t_1), V_\delta(t_2)\rangle \le \|V_\delta(t_1)\|\|V_\delta(t_2)\| = \|V_\delta(t_1)\|^2.$$

This implies $V_\delta(t_1) = V_\delta(t_2)$, and then equation (7) implies $a(t_1) = a(t_2)$. Hence, $t_1 = t_2$, because $a(t)$ is strictly decreasing.

Therefore $\phi(t)$ is decreasing and $\psi(t)$ is increasing. □

Lemma 4 *Suppose that $\|F(0) - f_\delta\| > C\delta$, $C > 1$, and $a(0)$ is sufficiently large. Then, there exists a unique $t_1 > 0$ such that $\|F(V_\delta(t_1)) - f_\delta\| = C\delta$.*

Proof. The uniqueness of t_1 follows from Lemma 3 because $\|F(V_\delta(t)) - f_\delta\| = \phi(t)$, and ϕ is decreasing. We have $F(y) = f$, and

$$\begin{aligned}0 &= \langle F(V_\delta) + aV_\delta - f_\delta, F(V_\delta) - f_\delta\rangle \\ &= \|F(V_\delta) - f_\delta\|^2 + a\langle V_\delta - y, F(V_\delta) - f_\delta\rangle + a\langle y, F(V_\delta) - f_\delta\rangle \\ &= \|F(V_\delta) - f_\delta\|^2 + a\langle V_\delta - y, F(V_\delta) - F(y)\rangle + a\langle V_\delta - y, f - f_\delta\rangle + a\langle y, F(V_\delta) - f_\delta\rangle \\ &\ge \|F(V_\delta) - f_\delta\|^2 + a\langle V_\delta - y, f - f_\delta\rangle + a\langle y, F(V_\delta) - f_\delta\rangle.\end{aligned}$$

Here the inequality $\langle V_\delta - y, F(V_\delta) - F(y)\rangle \ge 0$ was used. Therefore

$$\begin{aligned}\|F(V_\delta) - f_\delta\|^2 &\le -a\langle V_\delta - y, f - f_\delta\rangle - a\langle y, F(V_\delta) - f_\delta\rangle \\ &\le a\|V_\delta - y\|\|f - f_\delta\| + a\|y\|\|F(V_\delta) - f_\delta\| \\ &\le a\delta\|V_\delta - y\| + a\|y\|\|F(V_\delta) - f_\delta\|.\end{aligned} \tag{11}$$

On the other hand, we have

$$\begin{aligned}0 &= \langle F(V_\delta) - F(y) + aV_\delta + f - f_\delta, V_\delta - y\rangle \\ &= \langle F(V_\delta) - F(y), V_\delta - y\rangle + a\|V_\delta - y\|^2 + a\langle y, V_\delta - y\rangle + \langle f - f_\delta, V_\delta - y\rangle \\ &\geq a\|V_\delta - y\|^2 + a\langle y, V_\delta - y\rangle + \langle f - f_\delta, V_\delta - y\rangle,\end{aligned}$$

where the inequality $\langle V_\delta - y, F(V_\delta) - F(y)\rangle \geq 0$ was used. Therefore,

$$a\|V_\delta - y\|^2 \leq a\|y\|\|V_\delta - y\| + \delta\|V_\delta - y\|.$$

This implies

$$a\|V_\delta - y\| \leq a\|y\| + \delta. \tag{12}$$

From (11) and (12), and an elementary inequality $ab \leq \epsilon a^2 + \frac{b^2}{4\epsilon}$, $\forall \epsilon > 0$, one gets:

$$\begin{aligned}\|F(V_\delta) - f_\delta\|^2 &\leq \delta^2 + a\|y\|\delta + a\|y\|\|F(V_\delta) - f_\delta\| \\ &\leq \delta^2 + a\|y\|\delta + \epsilon\|F(V_\delta) - f_\delta\|^2 + \frac{1}{4\epsilon}a^2\|y\|^2,\end{aligned} \tag{13}$$

where $\epsilon > 0$ is fixed, independent of t, and can be chosen arbitrary small. Let $t \to \infty$ and $a = a(t) \searrow 0$. Then (13) implies

$$\overline{\lim}_{t\to\infty}(1 - \epsilon)\|F(V_\delta) - f_\delta\|^2 \leq \delta^2.$$

This, the continuity of F, the continuity of $V_\delta(t)$ on $[0, \infty)$, and the assumption $\|F(0) - f_\delta\| > C\delta$ imply that equation $\|F(V_\delta(t)) - f_\delta\| = C\delta$ must have a solution $t_1 > 0$. The uniqueness of this solution was already established. □

Remark 5 From the proof of Lemma 4 one obtains the following result:

If $t_n \nearrow \infty$ then there exists a unique $n_1 > 0$ such that

$$\|F(V_{n_1+1}) - f_\delta\| \leq C\delta < \|F(V_{n_1}) - f_\delta\|, \qquad V_n := V_\delta(t_n).$$

Remark 6 From Lemma 2 and Lemma 3 one concludes that

$$a_n\|V_n\| = \|F(V_n) - f_\delta\| \leq \|F(0) - f_\delta\|, \qquad a_n := a(t_n), \quad \forall n \geq 0.$$

Remark 7 Let $V := V_\delta(t)|_{\delta=0}$, so

$$F(V) + a(t)V - f = 0.$$

Let y be the minimal-norm solution to equation (1). We claim that

$$\|V_\delta - V\| \le \frac{\delta}{a}. \tag{14}$$

Indeed, from (7) one gets

$$F(V_\delta) - F(V) + a(V_\delta - V) = f - f_\delta.$$

Multiply this equality with $(V_\delta - V)$ and use the monotonicity of F to get

$$a\|V_\delta - V\|^2 \le \delta\|V_\delta - V\|.$$

This implies (14).

Similarly, multiplying the equation

$$F(V) + aV - F(y) = 0,$$

by $V - y$ one derives the inequality:

$$\|V\| \le \|y\|. \tag{15}$$

Similar arguments one can find in [7].

From (14) and (15), one gets the following estimate:

$$\|V_\delta\| \le \|V\| + \frac{\delta}{a} \le \|y\| + \frac{\delta}{a}. \tag{16}$$

Lemma 8 *Suppose* $a(t) = \frac{d}{(c+t)^b}$, $\varphi(t) = \int_0^t \frac{a(s)}{2} ds$ *where* $b \in (0, \frac{1}{2}]$, d *and* c *are positive constants. Then*

$$\frac{d}{2}\bigg(1 - \frac{2b}{c^\theta d}\bigg) \int_0^t \frac{e^{\varphi(s)}}{(s+c)^{2b}} ds < \frac{e^{\varphi(t)}}{(c+t)^b}, \qquad \forall t > 0, \quad \theta = 1 - b > 0. \tag{17}$$

Proof. We have

$$\varphi(t) = \int_0^t \frac{d}{2(c+s)^b} ds = \frac{d}{2(1-b)}\Big((c+t)^{1-b} - c^{1-b}\Big) = p(c+t)^\theta - C_3, \tag{18}$$

where $\theta := 1 - b$, $p := \frac{d}{2\theta}$, $C_3 := pc^\theta$. One has

$$\begin{aligned}
\frac{d}{dt}\frac{e^{p(c+t)^\theta}}{(c+t)^b} &= \frac{p\theta e^{p(c+t)^\theta}}{(c+t)^{b+1-\theta}} - \frac{be^{p(c+t)^\theta}}{(c+t)^{b+1}} \\
&= \frac{e^{p(c+t)^\theta}}{(c+t)^b}\bigg(\frac{d}{2(c+t)^b} - \frac{b}{c+t}\bigg) \\
&\ge \frac{e^{p(c+t)^\theta}}{(c+t)^b}\frac{d}{2(c+t)^b}\bigg(1 - \frac{2b}{c^\theta d}\bigg).
\end{aligned}$$

Therefore,

$$\frac{d}{2}\left(1-\frac{2b}{c^\theta d}\right)\int_0^t \frac{e^{p(c+s)^\theta}}{(s+c)^{2b}}ds \le \int_0^t \frac{d}{ds}\frac{e^{p(c+s)^\theta}}{(c+s)^b}ds$$
$$\le \frac{e^{p(c+t)^\theta}}{(c+t)^b}-\frac{e^{pc^\theta}}{c^b}\le \frac{e^{p(c+t)^\theta}}{(c+t)^b}.$$

Multiplying this inequality by e^{-C_3} and using (18), one obtains (17). Lemma 8 is proved. □

Lemma 9 *Let* $a(t)=\frac{d}{(c+t)^b}$ *and* $\varphi(t):=\int_0^t\frac{a(s)}{2}ds$ *where* $d,c>0$, $b\in(0,\frac{1}{2}]$ *and* $c^{1-b}d\ge 6b$. *One has*

$$e^{-\varphi(t)}\int_0^t e^{\varphi(s)}|\dot a(s)|\|V_\delta(s)\|ds\le \frac{1}{2}a(t)\|V_\delta(t)\|,\qquad t\ge 0. \tag{19}$$

Proof. From Lemma 8, one has

$$\frac{1}{2}\left(1-\frac{2b}{c^\theta d}\right)\int_0^t e^{\varphi(s)}\frac{d^2}{(s+c)^{2b}}ds< e^{\varphi(t)}\frac{d}{(c+t)^b},\qquad \forall c,b\ge 0,\quad \theta=1-b>0. \tag{20}$$

Since $c^{1-b}d\ge 6b$ or $\frac{6b}{c^\theta d}\le 1$, one has

$$1-\frac{2b}{c^\theta d}\ge\frac{4b}{c^\theta d}\ge\frac{4b}{(c+s)^{1-b}d},\qquad s\ge 0.$$

This implies

$$\frac{a^2(s)}{2}\left(1-\frac{2b}{c^\theta d}\right)=\frac{d^2}{2(c+s)^{2b}}\left(1-\frac{2b}{c^\theta d^2}\right)\ge\frac{4db}{2(c+s)^{b+1}}=2|\dot a(s)|,\qquad s\ge 0. \tag{21}$$

Multiplying (20) by $\|V_\delta(t)\|$, using inequality (21) and the fact that $\|V_\delta(t)\|$ is increasing, one gets, for all $t\ge 0$, the inequalities:

$$e^{\varphi(t)}a(t)\|V_\delta(t)\|>\int_0^t e^{\varphi(s)}\|V_\delta(t)\|\frac{a^2(s)}{2}\left(1-\frac{2b}{c^\theta d}\right)ds\ge 2\int_0^t e^{\varphi(s)}|\dot a(s)|\|V_\delta(s)\|ds.$$

This implies inequality (19). Lemma 9 is proved. □

Let us recall the following lemma, which is basic in our proofs.

Lemma 10 ([7], p. 97) *Let* $\alpha(t)$, $\beta(t)$, $\gamma(t)$ *be continuous nonnegative functions on* $[t_0,\infty)$, $t_0\ge 0$ *is a fixed number. If there exists a function*

$$\mu\in C^1[t_0,\infty),\quad \mu>0,\quad \lim_{t\to\infty}\mu(t)=\infty,$$

such that

$$0\le\alpha(t)\le\frac{\mu}{2}\left[\gamma-\frac{\dot\mu(t)}{\mu(t)}\right],\qquad \dot\mu:=\frac{d\mu}{dt}, \tag{22}$$

$$\beta(t)\le\frac{1}{2\mu}\left[\gamma-\frac{\dot\mu(t)}{\mu(t)}\right], \tag{23}$$

$$\mu(0)g(0)<1, \tag{24}$$

and $g(t) \geq 0$ satisfies the inequality

$$\dot{g}(t) \leq -\gamma(t)g(t) + \alpha(t)g^2(t) + \beta(t), \quad t \geq t_0, \tag{25}$$

then $g(t)$ exists on $[t_0, \infty)$ and

$$0 \leq g(t) < \frac{1}{\mu(t)} \to 0, \quad as \quad t \to \infty. \tag{26}$$

If inequalities (22)–(24) hold on an interval $[t_0, T)$, then $g(t)$ exists on this interval and inequality (26) holds on $[t_0, T)$.

Lemma 11 *Suppose M_1 and c_1 are positive constants and $0 \neq y \in H$. Then there exist a number $\lambda > 0$ and a function $a(t) \in C^1[0, \infty)$, $0 < a(t) \searrow 0$, such that*

$$|\dot{a}(t)| \leq \frac{a^2(t)}{2},$$

and the following conditions hold

$$\frac{M_1}{\|y\|} \leq \lambda, \tag{27}$$

$$0 \leq \frac{\lambda}{2a(t)}\left[a(t) - \frac{|\dot{a}(t)|}{a(t)}\right], \tag{28}$$

$$c_1\frac{|\dot{a}(t)|}{a(t)} \leq \frac{a(t)}{2\lambda}\left[a(t) - \frac{|\dot{a}(t)|}{a(t)}\right], \tag{29}$$

$$\frac{\lambda}{a(0)}g(0) < 1. \tag{30}$$

Proof. Take

$$a(t) = \frac{d}{(c+t)^b}, \quad 0 < b \leq \frac{1}{2}, \quad 2b \leq c^{1-b}d, \quad c \geq 1. \tag{31}$$

Note that $|\dot{a}| = -\dot{a}$. We have

$$\frac{|\dot{a}|}{a^2} = \frac{b}{d(c+t)^{1-b}} \leq \frac{b}{dc^{1-b}} \leq \frac{1}{2}.$$

Hence,

$$\frac{a(t)}{2} \leq a(t) - \frac{|\dot{a}(t)|}{a(t)}. \tag{32}$$

Thus, inequality (28) is satisfied. Take

$$\lambda \geq \frac{M_1}{\|y\|}, \tag{33}$$

then (27) is satisfied. For any given $g(0)$, choose $a(0)$ sufficiently large so that

$$\frac{\lambda}{a(0)}g(0) < 1.$$

Therefore, inequality (30) is satisfied.

Choose $\kappa \geq 1$ such that

$$\kappa > \max\left(\frac{4\lambda c_1 b}{d^2}, 1\right). \tag{34}$$

Define

$$\nu(t) := \kappa a(t) \qquad \lambda_\kappa := \kappa\lambda. \tag{35}$$

Note that (28) holds for $a(t) = \nu(t)$, $\lambda = \lambda_\kappa$ since (32) holds as well under this transformation, i.e.,

$$\frac{\nu(t)}{2} \leq \nu(t) - \frac{|\dot{\nu}(t)|}{\nu(t)}. \tag{36}$$

Using the inequalities (34) and $c \geq 1$ and the definition (35), one obtains

$$4\lambda_\kappa c_1 \frac{|\dot{\nu}(t)|}{\nu^3(t)} = 4\lambda c_1 \frac{b}{\kappa d^2 (c+t)^{1-2b}} \leq 4\lambda c_1 \frac{b}{\kappa d^2} \leq 1.$$

This implies

$$c_1 \frac{|\dot{\nu}|}{\nu(t)} \leq \frac{\nu^2(t)}{4\lambda_\kappa} \leq \frac{\nu(t)}{2\lambda_\kappa}\left[\nu - \frac{2|\dot{\nu}|}{\nu}\right].$$

Thus, one can replace the function $a(t)$ by $\nu(t) = \kappa a(t)$ and λ by $\lambda_\kappa = \kappa\lambda$ in the inequalities (27)–(30). □

Lemma 12 *Suppose M_1, c_1 and $\tilde{\alpha}$ are positive constants and $0 \neq y \in H$. Then there exist a number $\lambda > 0$ and a sequence $0 < (a_n)_{n=0}^\infty \searrow 0$ such that the following conditions hold*

$$\frac{a_n}{a_{n+1}} \leq 2, \tag{37}$$

$$\|f_\delta - F(0)\| \leq \frac{a_0^2}{\lambda}, \tag{38}$$

$$\frac{M_1}{\lambda} \leq \|y\|, \tag{39}$$

$$\frac{a_n}{\lambda} - \frac{\tilde{\alpha} a_n^2}{\lambda} + \frac{a_n - a_{n+1}}{a_{n+1}} c_1 \leq \frac{a_{n+1}}{\lambda}. \tag{40}$$

Proof. Let us show that if $a_0 > 0$ is sufficiently large, then the following sequence

$$a_n = \frac{a_0}{(1+n)^b}, \qquad b = \frac{1}{2}, \tag{41}$$

satisfies conditions (38)–(40) if

$$\lambda \geq \frac{M_1}{\|y\|}. \tag{42}$$

Condition (37) is satisfied by the sequence (41). Inequality (39) is satisfied since (42) holds. Choose $a(0)$ so that

$$a_0 \geq \sqrt{\|f_\delta - F(0)\|\lambda}, \tag{43}$$

then (38) is satisfied.

Assume that $(a_n)_{n=0}^\infty$ and λ satisfy (37), (38) and (39). Choose $\kappa \geq 1$ such that

$$\kappa \geq \max\left(\frac{1}{\tilde{\alpha} a_0 \sqrt{2}}, \frac{\lambda c_1}{\tilde{\alpha} a_0^2}\right). \tag{44}$$

It follows from (44) that

$$\frac{1}{\kappa a_0 \sqrt{2}} \leq \tilde{\alpha}, \qquad \frac{\lambda c_1}{\kappa a_0^2} \leq \tilde{\alpha}. \tag{45}$$

Define

$$(b_n)_{n=0}^\infty := (\kappa a_n)_{n=0}^\infty, \qquad \lambda_\kappa := \kappa\lambda. \tag{46}$$

For all $n \geq 0$ one has

$$\frac{a_n - a_{n+1}}{a_n^2} = \frac{a_n^2 - a_{n+1}^2}{a_n^2(a_n + a_{n+1})} \leq \frac{a_n^2 - a_{n+1}^2}{2a_{n+1}a_n^2} = \frac{\frac{a_0^2}{n+1} - \frac{a_0^2}{n+2}}{2\frac{a_0}{\sqrt{n+2}}\frac{a_0^2}{n+1}} = \frac{1}{a_0 2\sqrt{n+2}} \leq \frac{1}{a_0 2\sqrt{2}}. \tag{47}$$

Since a_n is decreasing, one has

$$\begin{aligned} \frac{a_n - a_{n+1}}{a_n^2 a_{n+1}} &= \frac{a_n^2 - a_{n+1}^2}{a_n^2 a_{n+1}(a_n + a_{n+1})} \\ &\leq \frac{a_n^2 - a_{n+1}^2}{2a_n^2 a_{n+1}^2} = \frac{\frac{a_0^2}{n+1} - \frac{a_0^2}{n+2}}{2\frac{a_0^2}{n+2}\frac{a_0^2}{n+1}} \leq \frac{1}{2a_0^2}, \qquad \forall n \geq 0. \end{aligned} \tag{48}$$

Using inequalities (47) and (45), one gets

$$\frac{2(a_n - a_{n+1})}{\kappa a_n^2} \leq \frac{1}{\kappa a_0 \sqrt{2}} \leq \tilde{\alpha}. \tag{49}$$

Similarly, using inequalities (48) and (45), one gets

$$\frac{2\lambda(a_n - a_{n+1})c_1}{\kappa a_n^2 a_{n+1}} \leq \frac{\lambda c_1}{\kappa a_0^2} \leq \tilde{\alpha}. \tag{50}$$

Inequalities (49) and (50) imply

$$\begin{aligned} \frac{b_n - b_{n+1}}{\lambda_\kappa} + \frac{b_n - b_{n+1}}{b_{n+1}}c_1 &= \frac{a_n - a_{n+1}}{\lambda} + \frac{a_n - a_{n+1}}{a_{n+1}}c_1 \\ &= \frac{\kappa a_n^2}{2\lambda}\frac{2(a_n - a_{n+1})}{\kappa a_n^2} + \frac{\kappa a_n^2}{2\lambda}\frac{2\lambda(a_n - a_{n+1})c_1}{\kappa a_n^2 a_{n+1}} \\ &\leq \frac{\kappa a_n^2}{2\lambda}\tilde{\alpha} + \frac{\kappa a_n^2}{2\lambda}\tilde{\alpha} = \frac{\kappa a_n^2 \tilde{\alpha}}{\lambda} = \frac{\tilde{\alpha} b_n^2}{\lambda_\kappa}. \end{aligned}$$

Thus, inequality (40) holds for a_n replaced by $b_n = \kappa a_n$ and λ replaced by $\lambda_\kappa = \kappa\lambda$, where κ satisfies (44). Inequalities (37)–(39) hold as well under this transformation. Thus, the choices $a_n = b_n$ and $\lambda := \kappa\frac{M_1}{\|y\|}$, where κ satisfies (44), satisfy all the conditions of Lemma 12. □

Remark 13 The constant c_1, used in Lemmas 11 and 12, will be used in Theorems 17 and 19. This constant is defined in equation (62). The constant $\tilde{\alpha}$, used in Lemma 12, is the one from Theorem 19. This constant is defined in equation (89).

Remark 14 Using similar arguments one can show that the sequence $a_n = \frac{d}{(c+n)^b}$, where $c \geq 1$, $0 < b \leq \frac{1}{2}$, satisfy all conditions of Lemma 4 provided that d is sufficiently large and λ is chosen so that inequality (42) holds.

Remark 15 In the proof of Lemmas 11 and 12 the numbers a_0 and λ can be chosen so that $\frac{a_0}{\lambda}$ is uniformly bounded as $\delta \to 0$ regardless of the rate of growth of the constant $M_1 = M_1(R)$ from formula (3) when $R \to \infty$, i.e., regardless of the strength of the nonlinearity $F(u)$.

To satisfy (42) one can choose $\lambda = M_1 \frac{1}{\|y\|}$. To satisfy (43) one can choose

$$a_0 = \sqrt{\lambda(\|f - F(0)\| + \|f\|)} \geq \sqrt{\lambda \|f_\delta - F(0)\|},$$

where we have assumed without loss of generality that $0 < \|f_\delta - f\| < \|f\|$. With this choice of a_0 and λ, the ratio $\frac{a_0}{\lambda}$ is bounded uniformly with respect to $\delta \in (0, 1)$ and does not depend on R. The dependence of a_0 on δ is seen from (43) since f_δ depends on δ. In practice one has $\|f_\delta - f\| < \|f\|$. Consequently,

$$\sqrt{\|f_\delta - F(0)\| \lambda} \leq \sqrt{(\|f - F(0)\| + \|f\|)\lambda}.$$

Thus, we can practically choose $a(0)$ independent of δ from the following inequality

$$a_0 \geq \sqrt{\lambda(\|f - F(0)\| + \|f\|)}.$$

Indeed, with the above choice one has $\frac{a_0}{\lambda} \leq c(1 + \sqrt{\lambda^{-1}}) \leq c$, where $c > 0$ is a constant independent of δ, and one can assume that $\lambda \geq 1$ without loss of generality.

This Remark is used in the proof of the main result in Section 3. Specifically, it is used to prove that an iterative process (88) generates a sequence which stays in the ball $B(u_0, R)$ for all $n \leq n_0 + 1$, where the number n_0 is defined by formula (99) (see below), and $R > 0$ is sufficiently large. An upper bound on R is given in the proof of Theorem 19, below formula (112).

Remark 16 One can choose $u_0 \in H$ such that

$$g_0 := \|u_0 - V_0\| \leq \frac{\|F(0) - f_\delta\|}{a_0}. \tag{51}$$

Indeed, if, for example, $u_0 = 0$, then by Remark 6 one gets

$$g_0 = \|V_0\| = \frac{a_0\|V_0\|}{a_0} \leq \frac{\|F(0) - f_\delta\|}{a_0}.$$

If (38) and (51) hold then $g_0 \leq \frac{a_0}{\lambda}$.

3 Main results

3.1 Dynamical systems method

Assume:

$$0 < a(t) \searrow 0, \quad \lim_{t\to\infty} \frac{\dot{a}(t)}{a(t)} = 0, \quad \frac{|\dot{a}(t)|}{a^2(t)} \le \frac{1}{2}. \tag{52}$$

Let $u_\delta(t)$ solve the following Cauchy problem:

$$\dot{u}_\delta = -[F(u_\delta) + a(t)u_\delta - f_\delta], \quad u_\delta(0) = u_0. \tag{53}$$

Theorem 17 *Assume that $F : H \to H$ is a monotone operator, condition (3) holds, and u_0 is an element of H, satisfying inequality* (83) *(see below). Let $a(t)$ satisfy conditions of Lemma 11. For example, one can choose $a(t) = \frac{d}{(c+t)^b}$, where $b \in (0, \frac{1}{2}]$, $c \ge 1$ and $d > 0$ are constants, and d is sufficiently large. Assume that equation $F(u) = f$ has a solution $y \in B(u_0, R)$, possibly nonunique, and y is the minimal-norm solution to this equation. Let f be unknown but f_δ be given, $\|f_\delta - f\| \le \delta$. Then the solution $u_\delta(t)$ to problem* (53) *exists on an interval* $[0, T_\delta]$*, $\lim_{\delta\to 0} T_\delta = \infty$, and there exists t_δ, $t_\delta \in (0, T_\delta)$, not necessarily unique, such that*

$$\|F(u_\delta(t_\delta)) - f_\delta\| = C_1\delta^\zeta, \quad \lim_{\delta\to 0} t_\delta = \infty, \tag{54}$$

where $C_1 > 1$ and $0 < \zeta \le 1$ are constants. If $\zeta \in (0, 1)$ and t_δ satisfies (54)*, then*

$$\lim_{\delta\to 0} \|u_\delta(t_\delta) - y\| = 0. \tag{55}$$

Remark 18 One can easily choose u_0 satisfying inequality (83). Note that inequality (83) is a sufficient condition for (86) to hold. In our proof inequality (86) is used at $t = t_\delta$. The stopping time t_δ is often sufficiently large for the quantity $e^{-\varphi(t_\delta)}h_0$ to be small. In this case inequality (86) with $t = t_\delta$ is satisfied for a wide range of u_0.

Proof. [Proof of Theorem 17] Denote

$$C := \frac{C_1 + 1}{2}. \tag{56}$$

Let

$$w := u_\delta - V_\delta, \quad g(t) := \|w\|.$$

One has

$$\dot{w} = -\dot{V}_\delta - \big[F(u_\delta) - F(V_\delta) + a(t)w\big]. \tag{57}$$

Multiplying (57) by w and using (2) one gets

$$g\dot{g} \le -ag^2 + \|\dot{V}_\delta\| g. \tag{58}$$

Let $t_0 > 0$ be such that

$$\frac{\delta}{a(t_0)} = \frac{1}{C-1}\|y\|, \qquad C > 1. \tag{59}$$

This t_0 exists and is unique since $a(t) > 0$ monotonically decays to 0 as $t \to \infty$. By Lemma 4, there exists t_1 such that

$$\|F(V_\delta(t_1)) - f_\delta\| = C\delta, \quad F(V_\delta(t_1)) + a(t_1)V_\delta(t_1) - f_\delta = 0. \tag{60}$$

We claim that $t_1 \in [0, t_0]$.

Indeed, from (7) and (16) one gets

$$C\delta = a(t_1)\|V_\delta(t_1)\| \le a(t_1)\bigg(\|y\| + \frac{\delta}{a(t_1)}\bigg) = a(t_1)\|y\| + \delta, \quad C > 1,$$

so

$$\delta \le \frac{a(t_1)\|y\|}{C-1}.$$

Thus,

$$\frac{\delta}{a(t_1)} \le \frac{\|y\|}{C-1} = \frac{\delta}{a(t_0)}.$$

Since $a(t) \searrow 0$, the above inequality implies $t_1 \le t_0$. Differentiating both sides of (7) with respect to t, one obtains

$$A_{a(t)}\dot{V}_\delta = -\dot{a}V_\delta, \quad A := F'(V_\delta), \quad A_a := A + aI.$$

This implies

$$\|\dot{V}_\delta\| \le |\dot{a}|\|A_{a(t)}^{-1}V_\delta\| \le \frac{|\dot{a}|}{a}\|V_\delta\| \le \frac{|\dot{a}|}{a}\bigg(\|y\| + \frac{\delta}{a}\bigg) \le \frac{|\dot{a}|}{a}\|y\|\bigg(1 + \frac{1}{C-1}\bigg), \quad \forall t \le t_0. \tag{61}$$

Since $g \ge 0$, inequalities (58) and (61) imply

$$\dot{g} \le -a(t)g(t) + \frac{|\dot{a}(t)|}{a(t)}c_1, \quad c_1 = \|y\|\bigg(1 + \frac{1}{C-1}\bigg). \tag{62}$$

Inequality (62) is of the type (25) with

$$\gamma(t) = a(t), \quad \alpha(t) = 0, \quad \beta(t) = c_1\frac{|\dot{a}(t)|}{a(t)}.$$

Let us check assumptions (22)–(24). Take

$$\mu(t) = \frac{\lambda}{a(t)}, \quad \lambda = \text{const}.$$

By Lemma 11 there exist λ and $a(t)$ such that conditions (22)–(24) hold. Thus, Lemma 10 yields

$$g(t) < \frac{a(t)}{\lambda}, \quad \forall t \le t_0. \tag{63}$$

Therefore,

$$\begin{aligned} \|F(u_\delta(t)) - f_\delta\| \le & \|F(u_\delta(t)) - F(V_\delta(t))\| + \|F(V_\delta(t)) - f_\delta\| \\ \le & M_1 g(t) + \|F(V_\delta(t)) - f_\delta\| \\ \le & \frac{M_1 a(t)}{\lambda} + \|F(V_\delta(t)) - f_\delta\|, \qquad \forall t \le t_0. \end{aligned} \tag{64}$$

It follows from Lemma 3 that $\|F(V_\delta(t)) - f_\delta\|$ is decreasing. Since $t_1 \le t_0$, one gets

$$\|F(V_\delta(t_0)) - f_\delta\| \le \|F(V_\delta(t_1)) - f_\delta\| = C\delta. \tag{65}$$

This, inequality (64), the inequality $\frac{M_1}{\lambda} \le \|y\|$ (see (33)), the relation (59), and the definition $C_1 = 2C - 1$ (see (56)) imply

$$\begin{aligned} \|F(u_\delta(t_0)) - f_\delta\| \le & \frac{M_1 a(t_0)}{\lambda} + C\delta \\ \le & \frac{M_1 \delta (C-1)}{\lambda \|y\|} + C\delta \le (2C-1)\delta = C_1\delta. \end{aligned} \tag{66}$$

Thus, if

$$\|F(u_\delta(0)) - f_\delta\| \ge C_1 \delta^\zeta, \quad 0 < \zeta \le 1,$$

then there exists $t_\delta \in (0, t_0)$ such that

$$\|F(u_\delta(t_\delta)) - f_\delta\| = C_1 \delta^\zeta \tag{67}$$

for any given $\zeta \in (0, 1]$, and any fixed $C_1 > 1$.

Let us prove (55)*. If this is done, then Theorem 17 is proved.*

First, we prove that $\lim_{\delta \to 0} \frac{\delta}{a(t_\delta)} = 0$.

From (64) with $t = t_\delta$, and from (16), one gets

$$\begin{aligned} C_1 \delta^\zeta & \le M_1 \frac{a(t_\delta)}{\lambda} + a(t_\delta)\|V_\delta(t_\delta)\| \\ & \le M_1 \frac{a(t_\delta)}{\lambda} + \|y\| a(t_\delta) + \delta. \end{aligned}$$

Thus, for sufficiently small δ, one gets

$$\tilde{C}\delta^\zeta \le a(t_\delta)\Big(\frac{M_1}{\lambda} + \|y\|\Big), \quad \tilde{C} > 0,$$

where $\tilde{C} < C_1$ is a constant. Therefore,

$$\lim_{\delta\to 0}\frac{\delta}{a(t_\delta)} \le \lim_{\delta\to 0}\frac{\delta^{1-\zeta}}{\tilde{C}}\left(\frac{M_1}{\lambda} + \|y\|\right) = 0, \quad 0 < \zeta < 1. \tag{68}$$

Secondly, we prove that

$$\lim_{\delta\to 0} t_\delta = \infty. \tag{69}$$

Using (53), one obtains:

$$\frac{d}{dt}\big(F(u_\delta) + au_\delta - f_\delta\big) = A_a\dot{u}_\delta + \dot{a}u_\delta = -A_a\big(F(u_\delta) + au_\delta - f_\delta\big) + \dot{a}u_\delta,$$

where $A_a := F'(u_\delta) + a$. This and (7) imply:

$$\frac{d}{dt}\big[F(u_\delta) - F(V_\delta) + a(u_\delta - V_\delta)\big] = -A_a\big[F(u_\delta) - F(V_\delta) + a(u_\delta - V_\delta)\big] + \dot{a}u_\delta. \tag{70}$$

Denote

$$v := F(u_\delta) - F(V_\delta) + a(u_\delta - V_\delta), \quad h = \|v\|.$$

Multiplying (70) by v and using monotonicity of F, one obtains

$$\begin{aligned} h\dot{h} &= -\langle A_a v, v\rangle + \langle v, \dot{a}(u_\delta - V_\delta)\rangle + \dot{a}\langle v, V_\delta\rangle \\ &\le -h^2 a + h|\dot{a}|\|u_\delta - V_\delta\| + |\dot{a}|h\|V_\delta\|, \qquad h \ge 0. \end{aligned} \tag{71}$$

Again, we have used the inequality $\langle F'(u_\delta)v, v\rangle \ge 0$ which follows from the monotonicity of F. Thus,

$$\dot{h} \le -ha + |\dot{a}|\|u_\delta - V_\delta\| + |\dot{a}|\|V_\delta\|. \tag{72}$$

Since $\langle F(u_\delta) - F(V_\delta), u_\delta - V_\delta\rangle \ge 0$, one obtains two inequalities

$$a\|u_\delta - V_\delta\|^2 \le \langle v, u_\delta - V_\delta\rangle \le \|u_\delta - V_\delta\|h, \tag{73}$$

and

$$\|F(u_\delta) - F(V_\delta)\|^2 \le \langle v, F(u_\delta) - F(V_\delta)\rangle \le h\|F(u_\delta) - F(V_\delta)\|. \tag{74}$$

Inequalities (73) and (74) imply:

$$a\|u_\delta - V_\delta\| \le h, \quad \|F(u_\delta) - F(V_\delta)\| \le h. \tag{75}$$

Inequalities (72) and (75) imply

$$\dot{h} \le -h\left(a - \frac{|\dot{a}|}{a}\right) + |\dot{a}|\|V_\delta\|. \tag{76}$$

Since $a - \frac{|\dot a|}{a} \geq \frac{a}{2}$ by the last inequality in (52), it follows from inequality (76) that

$$\dot h \leq -\frac{a}{2}h + |\dot a| \|V_\delta\|. \tag{77}$$

Inequality (77) implies:

$$h(t) \leq h(0)e^{-\int_0^t \frac{a(s)}{2}ds} + e^{-\int_0^t \frac{a(s)}{2}ds} \int_0^t e^{\int_0^s \frac{a(\xi)}{2}d\xi} |\dot a(s)| \|V_\delta(s)\| ds. \tag{78}$$

Denote

$$\varphi(t) := \int_0^t \frac{a(s)}{2} ds.$$

From (78) and (75), one gets

$$\|F(u_\delta(t)) - F(V_\delta(t))\| \leq h(0)e^{-\varphi(t)} + e^{-\varphi(t)} \int_0^t e^{\varphi(s)} |\dot a(s)| \|V_\delta(s)\| ds. \tag{79}$$

Therefore,

$$\begin{aligned} \|F(u_\delta(t)) - f_\delta\| &\geq \|F(V_\delta(t)) - f_\delta\| - \|F(V_\delta(t)) - F(u_\delta(t))\| \\ &\geq a(t)\|V_\delta(t)\| - h(0)e^{-\varphi(t)} - e^{-\varphi(t)} \int_0^t e^{\varphi(s)} |\dot a| \|V_\delta\| ds. \end{aligned} \tag{80}$$

From Lemma 9 it follows that there exists an $a(t)$ such that

$$\frac{1}{2}a(t)\|V_\delta(t)\| \geq e^{-\varphi(t)} \int_0^t e^{\varphi(s)} |\dot a| \|V_\delta(s)\| ds. \tag{81}$$

For example, one can choose

$$a(t) = \frac{d}{(c+t)^b}, \quad b \in (0, \frac{1}{2}], \quad dc^{1-b} \geq 6b, \tag{82}$$

where $d, c > 0$. Moreover, one can always choose u_0 such that

$$h(0) = \|F(u_0) + a(0)u_0 - f_\delta\| \leq \frac{1}{4}a(0)\|V_\delta(0)\|, \tag{83}$$

because the equation

$$F(u_0) + a(0)u_0 - f_\delta = 0$$

is solvable.

If (83) holds, then

$$h(0)e^{-\varphi(t)} \leq \frac{1}{4}a(0)\|V_\delta(0)\|e^{-\varphi(t)}, \qquad t \geq 0. \tag{84}$$

If (82) holds, $c \geq 1$ and $2b \leq d$, then it follows that

$$e^{-\varphi(t)}a(0) \leq a(t). \tag{85}$$

Indeed, inequality $a(0) \le a(t)e^{\varphi(t)}$ is obviously true for $t = 0$, and $\big(a(t)e^{\varphi(t)}\big)'_t \ge 0$, provided that $c \ge 1$ and $2b \le d$.

Inequalities (84) and (46) imply

$$e^{-\varphi(t)}h(0) \le \frac{1}{4}a(t)\|V_\delta(0)\| \le \frac{1}{4}a(t)\|V_\delta(t)\|, \quad t \ge 0. \tag{86}$$

where we have used the inequality $\|V_\delta(t)\| \le \|V_\delta(t')\|$ for $t \le t'$, established in Lemma 3. From (67) and (80)–(86), one gets

$$C\delta^\zeta = \|F(u_\delta(t_\delta)) - f_\delta\| \ge \frac{1}{4}a(t_\delta)\|V_\delta(t_\delta)\|.$$

Thus,

$$\lim_{\delta\to 0} a(t_\delta)\|V_\delta(t_\delta)\| \le \lim_{\delta\to 0} 4C\delta^\zeta = 0.$$

Since $\|V_\delta(t)\|$ is increasing, this implies $\lim_{\delta\to 0} a(t_\delta) = 0$. Since $0 < a(t) \searrow 0$, it follows that (69) holds.

From the triangle inequality and inequalities (63) and (14) one obtains:

$$\begin{aligned}\|u_\delta(t_\delta) - y\| &\le \|u_\delta(t_\delta) - V_\delta\| + \|V(t_\delta) - V_\delta(t_\delta)\| + \|V(t_\delta) - y\| \\ &\le \frac{a(t_\delta)}{\lambda} + \frac{\delta}{a(t_\delta)} + \|V(t_\delta) - y\|.\end{aligned} \tag{87}$$

From (68), (69), inequality (87) and Lemma 1, one obtains (55). Theorem 17 is proved. □

3.2 An iterative scheme

Let $V_{n,\delta}$ solve the equation:

$$F(V_{n,\delta}) + a_n V_{n,\delta} - f_\delta = 0.$$

Denote $V_n := V_{n,\delta}$.

Consider the following iterative scheme:

$$u_{n+1} = u_n - \alpha_n[F(u_n) + a_n u_n - f_\delta], \quad u_0 = u_0, \tag{88}$$

where u_0 is chosen so that inequality (51) holds, and $\{\alpha_n\}_{n=1}^\infty$ is a positive sequence such that

$$0 < \tilde{\alpha} \le \alpha_n \le \frac{2}{a_n + (M_1 + a_n)}, \qquad M_1 = \sup_{u\in B(u_0,R)} \|F'(u)\|. \tag{89}$$

It follows from this condition that

$$\|1 - \alpha_n(J_n + a_n)\| = \sup_{a_n \le \lambda \le M_1 + a_n} |1 - \alpha_n\lambda| \le 1 - \alpha_n a_n. \tag{90}$$

Here, J_n is an operator in H such that $J_n \geq 0$ and $\|J_n\| \leq M_1$, $\forall u \in B(u_0, R)$. A specific choice of J_n is made in formula (96) below.

Let a_n and λ satisfy conditions (37)–(40). Assume that equation $F(u) = f$ has a solution $y \in B(u_0, R)$, possibly nonunique, and y is the minimal-norm solution to this equation. Let f be unknown but f_δ be given, and $\|f_\delta - f\| \leq \delta$. We prove the following result:

Theorem 19 *Assume $a_n = \frac{d}{(c+n)^b}$ where $c \geq 1$, $0 < b \leq \frac{1}{2}$, and d is sufficiently large so that conditions (37)–(40) hold. Let u_n be defined by (88). Assume that u_0 is chosen so that (51) holds. Then there exists a unique n_δ such that*

$$\|F(u_{n_\delta}) - f_\delta\| \leq C_1\delta^\zeta, \quad C_1\delta^\zeta < \|F(u_n) - f_\delta\|, \quad \forall n < n_\delta, \tag{91}$$

where $C_1 > 1$, $0 < \zeta \leq 1$.

Let $0 < (\delta_m)_{m=1}^\infty$ be a sequence such that $\delta_m \to 0$. If the sequence $\{n_m := n_{\delta_m}\}_{m=1}^\infty$ is bounded, and $\{n_{m_j}\}_{j=1}^\infty$ is a convergent subsequence, then

$$\lim_{j\to\infty} u_{n_{m_j}} = \tilde{u}, \tag{92}$$

where $\tilde{u}$ is a solution to the equation $F(u) = f$. If

$$\lim_{m\to\infty} n_m = \infty, \tag{93}$$

where $\zeta \in (0, 1)$, then

$$\lim_{m\to\infty} \|u_{n_m} - y\| = 0. \tag{94}$$

Proof. Denote

$$C := \frac{C_1 + 1}{2}. \tag{95}$$

Let

$$z_n := u_n - V_n, \quad g_n := \|z_n\|.$$

One has

$$F(u_n) - F(V_n) = J_n z_n, \qquad J_n = \int_0^1 F'(u_0 + \xi z_n)d\xi. \tag{96}$$

Since $F'(u) \geq 0$, $\forall u \in H$ and $\|F'(u)\| \leq M_1, \forall u \in B(u_0, R)$, it follows that $J_n \geq 0$ and $\|J_n\| \leq M_1$. From (88) and (96) one obtains

$$\begin{aligned} z_{n+1} &= z_n - \alpha_n[F(u_n) - F(V_n) + a_n z_n] - (V_{n+1} - V_n) \\ &= (1 - \alpha_n(J_n + a_n))z_n - (V_{n+1} - V_n). \end{aligned} \tag{97}$$

From (97) and (90), one gets

$$\begin{aligned} g_{n+1} &\le g_n \|1-\alpha_n(J_n+a_n)\| + \|V_{n+1}-V_n\| \\ &\le g_n(1-\alpha_n a_n) + \|V_{n+1}-V_n\|. \end{aligned} \tag{98}$$

Since $0 < a_n \searrow 0$, for any fixed $\delta > 0$ there exists n_0 such that

$$\frac{\delta}{a_{n_0+1}} > \frac{1}{C-1}\|y\| \ge \frac{\delta}{a_{n_0}}, \qquad C > 1. \tag{99}$$

By (37), one has $\frac{a_n}{a_{n+1}} \le 2$, $\forall\, n \ge 0$. This and (99) imply

$$\frac{2}{C-1}\|y\| \ge \frac{2\delta}{a_{n_0}} > \frac{\delta}{a_{n_0+1}} > \frac{1}{C-1}\|y\| \ge \frac{\delta}{a_{n_0}}, \qquad C > 1. \tag{100}$$

Thus,

$$\frac{2}{C-1}\|y\| > \frac{\delta}{a_n}, \quad \forall n \le n_0 + 1. \tag{101}$$

The number n_0, satisfying (101), exists and is unique since $a_n > 0$ monotonically decays to 0 as $n \to \infty$. By Remark 5, there exists a number n_1 such that

$$\|F(V_{n_1+1}) - f_\delta\| \le C\delta < \|F(V_{n_1}) - f_\delta\|, \tag{102}$$

where V_n solves the equation $F(V_n) + a_n V_n - f_\delta = 0$.

We claim that $n_1 \in [0, n_0]$.

Indeed, one has $\|F(V_{n_1}) - f_\delta\| = a_{n_1}\|V_{n_1}\|$, and $\|V_{n_1}\| \le \|y\| + \frac{\delta}{a_{n_1}}$ (cf. (16)), so

$$C\delta < a_{n_1}\|V_{n_1}\| \le a_{n_1}\left(\|y\| + \frac{\delta}{a_{n_1}}\right) = a_{n_1}\|y\| + \delta, \quad C > 1. \tag{103}$$

Therefore,

$$\delta < \frac{a_{n_1}\|y\|}{C-1}. \tag{104}$$

Thus, by (100),

$$\frac{\delta}{a_{n_1}} < \frac{\|y\|}{C-1} < \frac{\delta}{a_{n_0+1}}. \tag{105}$$

Here the last inequality is a consequence of (100). Since a_n decreases monotonically, inequality (105) implies $n_1 \le n_0$. One has

$$\begin{aligned} a_{n+1}\|V_n - V_{n+1}\|^2 &= \langle (a_{n+1}-a_n)V_n - F(V_n) + F(V_{n+1}), V_n - V_{n+1}\rangle \\ &\le \langle (a_{n+1}-a_n)V_n, V_n - V_{n+1}\rangle \\ &\le (a_n - a_{n+1})\|V_n\|\|V_n - V_{n+1}\|. \end{aligned} \tag{106}$$

By (16), $\|V_n\| \le \|y\| + \frac{\delta}{a_n}$, and, by (101), $\frac{\delta}{a_n} \le \frac{2\|y\|}{C-1}$ for all $n \le n_0 + 1$. Therefore,

$$\|V_n\| \le \|y\| \biggl(1 + \frac{2}{C-1}\biggr), \qquad \forall n \le n_0 + 1, \tag{107}$$

and, by (106),

$$\|V_n - V_{n+1}\| \le \frac{a_n - a_{n+1}}{a_{n+1}} \|V_n\| \le \frac{a_n - a_{n+1}}{a_{n+1}} \|y\| \biggl(1 + \frac{2}{C-1}\biggr), \quad \forall n \le n_0 + 1. \tag{108}$$

Inequalities (98) and (108) imply

$$g_{n+1} \le (1 - \alpha_n a_n) g_n + \frac{a_n - a_{n+1}}{a_{n+1}} c_1, \qquad \forall\, n \le n_0 + 1, \tag{109}$$

where the constant c_1 is defined in (62).

By Lemma 4 and Remark 14, the sequence $(a_n)_{n=1}^{\infty}$, satisfies conditions (37)–(40), provided that a_0 is sufficiently large and $\lambda > 0$ is chosen so that (42) holds. Let us show by induction that

$$g_n < \frac{a_n}{\lambda}, \qquad 0 \le n \le n_0 + 1. \tag{110}$$

Inequality (110) holds for $n = 0$ by Remark 16. Suppose (110) holds for some $n \ge 0$. From (109), (110) and (40), one gets

$$\begin{aligned} g_{n+1} &\le (1 - \alpha_n a_n) \frac{a_n}{\lambda} + \frac{a_n - a_{n+1}}{a_{n+1}} c_1 \\ &= -\frac{\alpha_n a_n^2}{\lambda} + \frac{a_n}{\lambda} + \frac{a_n - a_{n+1}}{a_{n+1}} c_1 \\ &\le \frac{a_{n+1}}{\lambda}. \end{aligned} \tag{111}$$

Thus, by induction, inequality (110) holds for all n in the region $0 \le n \le n_0 + 1$.

From (16) one has $\|V_n\| \le \|y\| + \frac{\delta}{a_n}$. This and the triangle inequality imply

$$\|u_0 - u_n\| \le \|u_0\| + \|z_n\| + \|V_n\| \le \|u_0\| + \|z_n\| + \|y\| + \frac{\delta}{a_n}. \tag{112}$$

Inequalities (107), (110), and (112) guarantee that the sequence u_n, generated by the iterative process (88), remains in the ball $B(u_0, R)$ for all $n \le n_0 + 1$, where $R \le \frac{a_0}{\lambda} + \|u_0\| + \|y\| + \frac{\delta}{a_n}$. This inequality and the estimate (101) imply that the sequence u_n, $n \le n_0 + 1$, stays in the ball $B(u_0, R)$, where

$$R \le \frac{a_0}{\lambda} + \|u_0\| + \|y\| + \|y\| \frac{C+1}{C-1}. \tag{113}$$

By Remark 15, one can choose a_0 and λ so that $\frac{a_0}{\lambda}$ is uniformly bounded as $\delta \to 0$ even if $M_1(R) \to \infty$ as $R \to \infty$ at an arbitrary fast rate. Thus, the sequence u_n stays in the ball $B(u_0, R)$ for $n \le n_0 + 1$ when $\delta \to 0$. An upper bound on R is given above. It does not depend on δ as $\delta \to 0$.

One has:

$$\begin{aligned}\|F(u_n) - f_\delta\| \leq &\|F(u_n) - F(V_n)\| + \|F(V_n) - f_\delta\| \\ \leq &M_1 g_n + \|F(V_n) - f_\delta\| \\ \leq &\frac{M_1 a_n}{\lambda} + \|F(V_n) - f_\delta\|, \qquad \forall n \leq n_0 + 1,\end{aligned} \tag{114}$$

where (110) was used and M_1 is the constant from (3). Since $\|F(V_n) - f_\delta\|$ is decreasing, by Lemma 3, and $n_1 \leq n_0$, one gets

$$\|F(V_{n_0+1}) - f_\delta\| \leq \|F(V_{n_1+1}) - f_\delta\| \leq C\delta. \tag{115}$$

From (39), (114), (115), the relation (99), and the definition $C_1 = 2C - 1$ (see (95)), one concludes that

$$\begin{aligned}\|F(u_{n_0+1}) - f_\delta\| \leq &\frac{M_1 a_{n_0+1}}{\lambda} + C\delta \\ \leq &\frac{M_1 \delta (C-1)}{\lambda \|y\|} + C\delta \leq (2C-1)\delta = C_1\delta.\end{aligned} \tag{116}$$

Thus, if

$$\|F(u_0) - f_\delta\| > C_1 \delta^\zeta, \quad 0 < \zeta \leq 1,$$

then one concludes from (116) *that there exists* n_δ*,* $0 < n_\delta \leq n_0 + 1$*, such that*

$$\|F(u_{n_\delta}) - f_\delta\| \leq C_1 \delta^\zeta < \|F(u_n) - f_\delta\|, \quad 0 \leq n < n_\delta, \tag{117}$$

for any given $\zeta \in (0, 1]$*, and any fixed* $C_1 > 1$*.*

Let us prove (92)*.*

If $n > 0$ is fixed, then $u_{\delta,n}$ is a continuous function of f_δ. Denote

$$\tilde{u} := \tilde{u}_N = \lim_{\delta \to 0} u_{\delta, n_{m_j}}, \tag{118}$$

where

$$\lim_{j \to \infty} n_{m_j} = N.$$

From (118) and the continuity of F, one obtains:

$$\|F(\tilde{u}) - f_\delta\| = \lim_{j \to \infty} \|F(u_{n_{m_j}}) - f_\delta\| \leq \lim_{\delta \to 0} C_1 \delta^\zeta = 0.$$

Thus, $\tilde{u}$ is a solution to the equation $F(u) = f$, and (92) is proved.

Let us prove (94) *assuming that* (93) *holds.*

From (91) and (114) with $n = n_\delta - 1$, and from (117), one gets

$$C_1\delta^\zeta \le M_1\frac{a_{n_\delta - 1}}{\lambda} + a_{n_\delta-1}\|V_{n_\delta-1}\| \le M_1\frac{a_{n_\delta-1}}{\lambda} + \|y\|a_{n_\delta-1} + \delta.$$

If $\delta > 0$ is sufficiently small, then the above equation implies

$$\tilde{C}\delta^\zeta \le a_{n_\delta-1}\biggl(\frac{M_1}{\lambda} + \|y\|\biggr), \quad \tilde{C} > 0,$$

where $\tilde{C} < C_1$ is a constant. Therefore, by (37),

$$\lim_{\delta\to 0}\frac{\delta}{2a_{n_\delta}} \le \lim_{\delta\to 0}\frac{\delta}{a_{n_\delta-1}} \le \lim_{\delta\to 0}\frac{\delta^{1-\zeta}}{\tilde{C}}\biggl(\frac{M_1}{\lambda} + \|y\|\biggr) = 0, \quad 0 < \zeta < 1. \tag{119}$$

In particular, for $\delta = \delta_m$, one gets

$$\lim_{\delta_m\to 0}\frac{\delta_m}{a_{n_m}} = 0. \tag{120}$$

From the triangle inequality, inequalities (14) and (110), one obtains

$$\begin{aligned}\|u_{n_m} - y\| &\le \|u_{n_m} - V_{n_m}\| + \|V_n - V_{n_m,0}\| + \|V_{n_m,0} - y\|\\ &\le \frac{a_{n_m}}{\lambda} + \frac{\delta_m}{a_{n_m}} + \|V_{n_m,0} - y\|.\end{aligned} \tag{121}$$

From (93), (120), inequality (121) and Lemma 1, one obtains (94). Theorem 19 is proved. □

4 Numerical experiments

Let us do a numerical experiment solving nonlinear equation (1) with

$$F(u) := B(u) + \frac{u^3}{6} := \int_0^1 e^{-|x-y|}u(y)dy + \frac{u^3}{6}, \quad f(x) := \frac{13}{6} - e^{-x} - \frac{e^x}{e}. \tag{122}$$

One can check that $u(x) \equiv 1$ solves the equation $F(u) = f$. The operator B is compact in $H = L^2[0,1]$. The operator $u \longmapsto u^3$ is defined on a dense subset D of of $L^2[0,1]$, for example, on $D := C[0,1]$. If $u, v \in D$, then

$$\langle u^3 - v^3, u - v\rangle = \int_0^1 (u^3 - v^3)(u - v)dx \ge 0.$$

Moreover,

$$e^{-|x|} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{e^{i\lambda x}}{1+\lambda^2}d\lambda.$$

Therefore, $\langle B(u - v), u - v\rangle \ge 0$, so

$$\langle F(u) - F(v), u - v\rangle \ge 0, \qquad \forall u, v \in D.$$

Note that D does not contain subsets, open in $H = L^2[0,1]$, i.e., it does not contain interior points of H. This is a reflection of the fact that the operator $G(u) = \frac{u^3}{6}$ is unbounded on any open subset of H. For example, in any ball $\|u\| \leq C$, $C = const > 0$, where $\|u\| := \|u\|_{L^2[0,1]}$, there is an element u such that $\|u^3\| = \infty$. As such an element one can take, for example, $u(x) = c_1 x^{-b}$, $\frac{1}{3} < b < \frac{1}{2}$. here $c_1 > 0$ is a constant chosen so that $\|u\| \leq C$. The operator $u \longmapsto F(u) = G(u) + B(u)$ is maximal monotone on $D_F := \{u : u \in H,\, F(u) \in H\}$ (see [2, p.102]), so that equation (7) is uniquely solvable for any $f_\delta \in H$.

The Fréchet derivative of F is:

$$F'(u)h = \frac{u^2 h}{2} + \int_0^1 e^{-|x-y|} h(y)dy. \tag{123}$$

If $u(x)$ vanishes on a set of positive Lebesgue's measure, then $F'(u)$ is obviously not boundedly invertible. If $u \in C[0,1]$ vanishes even at one point x_0, then $F'(u)$ is not boundedly invertible in H.

Let us use the iterative process (88):

$$\begin{aligned} u_{n+1} &= u_n - \alpha_n (F(u_n) + a_n u_n - f_\delta), \\ u_0 &= 0. \end{aligned} \tag{124}$$

We stop iterations at $n := n_\delta$ such that the following inequality holds

$$\|F(u_{n_\delta}) - f_\delta\| < C\delta^\zeta, \quad \|F(u_n) - f_\delta\| \geq C\delta^\zeta, \quad n < n_\delta, \quad C > 1, \quad \zeta \in (0,1). \tag{125}$$

Integrals of the form $\int_0^1 e^{-|x-y|} h(y)dy$ in (122) and (123) are computed by using the trapezoidal rule. The noisy function used in the test is

$$f_\delta(x) = f(x) + \kappa f_{noise}(x), \quad \kappa > 0.$$

The noise level δ and the relative noise level are determined by

$$\delta = \kappa \|f_{noise}\|, \quad \delta_{rel} := \frac{\delta}{\|f\|}.$$

In the test, κ is computed in such a way that the relative noise level δ_{rel} equals to some desired value, i.e.,

$$\kappa = \frac{\delta}{\|f_{noise}\|} = \frac{\delta_{rel}\|f\|}{\|f_{noise}\|}.$$

We have used the relative noise level as an input parameter in the test.

The version of DSM, developed in this paper and denoted by DSMS, is compared with the version of DSM in [3], denoted by DSMN. Indeed, the DSMN is the following iterative scheme

$$u_{n+1} = u_n - A_n^{-1}(F'(u_n) + a_n u_n - f_\delta), \quad u_0 = u_0, \qquad n \geq 0, \tag{126}$$

where $a_n = \frac{a_0}{1+n}$. This iterative scheme is used with a stopping time n_δ defined by (91). The existence of this stopping time and the convergence of the method is proved in [3].

As we have proved, the DSMS converges when $a_n = \frac{a_0}{(1+n)^b}$, $b \in (0, \frac{1}{2}]$, and a_0 is sufficiently large. However, in practice, if we choose a_0 too large then the method will use too many iterations before reaching the stopping time n_δ in (125). This means that the computation time is large. Since

$$\|F(V_\delta) - f_\delta\| = a(t)\|V_\delta\|,$$

and $\|V_\delta(t_\delta) - u_\delta(t_\delta)\| = O(a(t_\delta))$, we have

$$C\delta^\zeta = \|F(u_\delta(t_\delta)) - f_\delta\| \sim a(t_\delta).$$

Thus, we choose

$$a_0 = C_0\delta^\zeta, \qquad C_0 > 0.$$

The parameter a_0 used in the DSMN is also chosen by this formula.

In all figures, the x-axis represents the variable x. In all figures, by *DSMS* we denote the numerical solutions obtained by the DSMS, by *DSMN* we denote solutions by the DSMN and by *exact* we denote the exact solution.

In experiments, we found that the DSMS works well with $a_0 = C_0\delta^\zeta$, $C_0 \in [0.5, 2]$. Indeed, in the test the DSMS is implemented with $a_n := C_0\frac{\delta^{0.99}}{(n+1)^{0.5}}$, $C_0 = 1$ while the DSMN is implemented with $a_n := C_0\frac{\delta^{0.99}}{(n+1)}$, $C_0 = 1$. For $C_0 > 3$ the convergence rate of DSMS is much slower while the DSMN still works well if $C_0 \in [1, 4]$. In all experiments, the noise function f_{noise} is a vector with random entries normally distributed of mean 0 and variant 1.

Figure 12 plots the solutions using relative noise levels $\delta = 0.01$ and $\delta = 0.001$. The exact solution used in these experiments is $u = 1$. In the test the DSMS is implemented with $\alpha_n = 1$, $C = 1.01$, $\zeta = 0.99$ and $\alpha_n = 1$, $\forall n \geq 0$. The number of iterations of the DSMS for $\delta = 0.01$ and $\delta = 0.001$ were 98 and 99 while the number of iteration for the DSMN are 10 and 10, respectively. The CPU time for the DSMS are 0.0139 and 0.0147 second while the CPU time for the DSMN are 0.0153 and 0.0169 corresponding to $\delta_{rel} = 0.01$ and $\delta_{rel} = 0.001$. The number of node points used

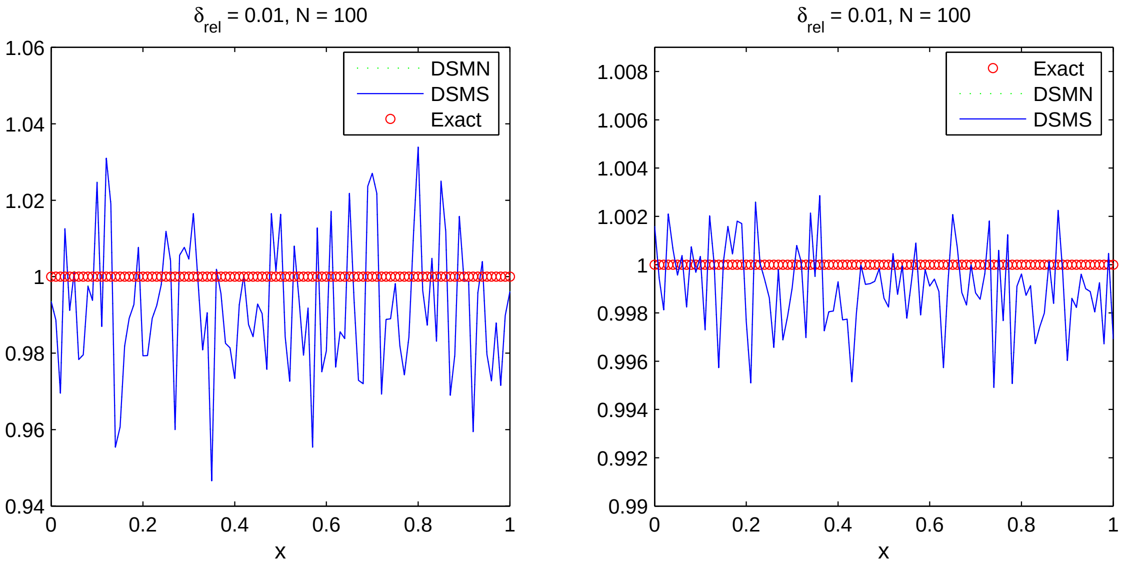

Figure 12: Plots of solutions obtained by the DSMN and DSMS when $N = 100$, $u = 1$, $x \in [0, 1]$, $\delta_{rel} = 0.01$ (left) and $N = 100$, $u = 1$, $x \in [0, 1]$, $\delta_{rel} = 0.001$ (right).

in computing integrals in (122) and (123) was $N = 100$. Figure 12 shows that the solutions by the DSMN and DSMS are nearly the same in this figure.

Figure 13 presents the numerical results when $N = 100$ with $\delta = 0.01$ $u(x) = \sin(2\pi x)$, $x \in [0, 1]$ (left) and with $\delta = 0.001$, $u(x) = \sin(\pi x)$, $x \in [0, 1]$ (right). In these cases, the DSMN took 10 and 12 iterations to give the numerical solutions while the DSMS took 56 and 67 iterations for $\delta = 0.01$ and $\delta = 0.001$, respectively. The computation time for the DSMS are 0.0102 and 0.0132 second while those for the DSMN are 0.0169 and 0.0186 second for $\delta = 0.01$ and $\delta = 0.001$, respectively. For larger number of node points experiments show that the DSMS is much faster than the DSMN. Figure 13 show that the numerical results of the DSMS are better than those of the DSMN.

In our experiments, the DSMS requires about the same or less time of computation than the DSMN. For larger number of node points, we found out that the DSMS runs faster than the DSMN. Moreover, the DSMS yields numerical results with the same accuracy as the DSMN does.

All the computations were carried out using MATLAB in double-precision arithmetic on a PC computer with an Intel Centrino Duo CPU of 1.62 GHz and 3 GB RAM.

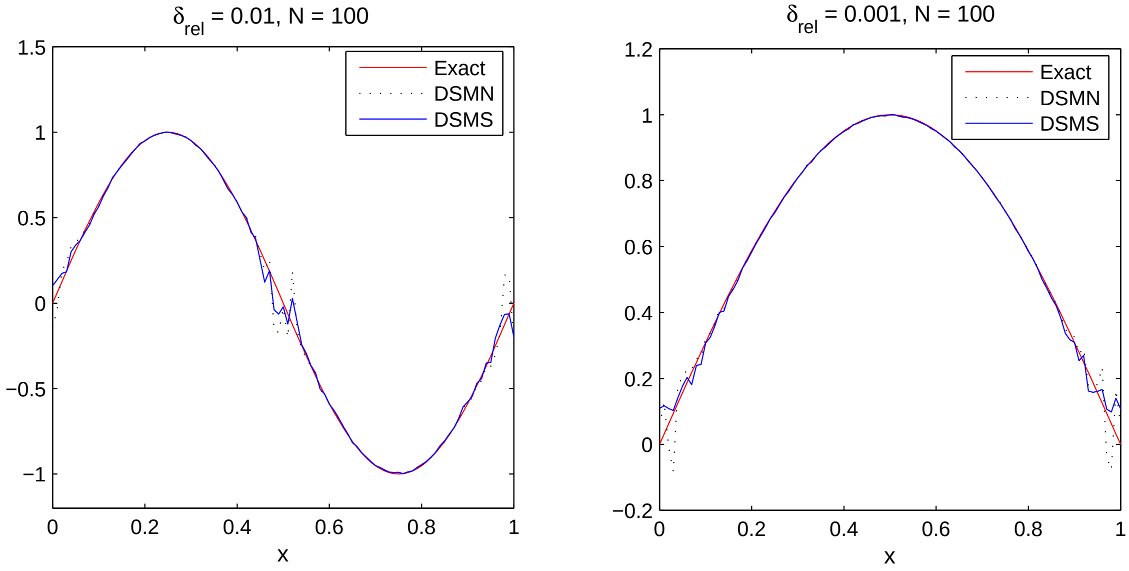

Figure 13: Plots of solutions obtained by the DSMN and DSMS when $N = 100$, $u(x) = \sin(2\pi x)$, $x \in [0, 1]$, $\delta_{rel} = 0.01$ (left) and $N = 100$, $u(x) = \sin(\pi x)$, $x \in [0, 1]$, $\delta_{rel} = 0.001$ (right).

5 Concluding remarks

Numerical experiments agree with the theory that the convergence rate of the DSMS is slower than that of the DSMN. This is because the rate of decay of the sequence $\{\frac{1}{(1+n)^{\frac{1}{2}}}\}_{n=1}^{\infty}$ is much slower than that of the sequence $\{\frac{1}{1+n}\}_{n=1}^{\infty}$. However, since the cost of one iteration of the DSMS is $O(N^2)$, and is much smaller than that of the DSMN (the cost of one iteration of the DSMN is $O(N^3)$), the DSMS requires less time to get a numerical result than the DSMN. Here, N is the number of the nodal points. Thus, for large scale problems, the DSMS may be an alternative to the DSMN. Also, as it is shown in Figure 13, the DSMS may yield more accurate solutions.

Experiments show that the DSMN still works with $a_n = \frac{a_0}{(1+n)^b}$ for $\frac{1}{2} \le b \le 1$. So, in practice one may use faster decaying sequence a_n to reduce the time of computation.

From the numerical results we conclude that the proposed DSM with the discrepancy-type stopping rule is a good alternative for the DSMN for large scale problems.

Remark. After the completion of this work, we saw the paper [1] in which an iterative process for solving equation (1) with monotone operator is proposed. In [1] some unnatural assumptions are made. For example, assumption (2.4) in [1] implies that the growth of the nonlinearity is not faster than linear, assumption (2.5) is not verifiable practically, in Theorem 2.1 the existence of $N(\delta)$ is not proved, so the result is actually not proved. A "generalized discrepancy principle" (2.8) in [1] is therefore not justified.

References

[1] A.Bakushinsky and A.Smirnova, Iterative regularization and generalized discrepancy principle for monotone operators, *Numer. Funct. Anal. and Optimization*, 28, (2007), 13-25.

[2] K. Deimling, *Nonlinear functional analysis*, Springer Verlag, Berlin, 1985.

[3] N. S. Hoang and A. G. Ramm, An iterative scheme for solving nonlinear equations with monotone operators, *BIT*, 48, (2008), N4, 725–741.

[4] N. S. Hoang and A. G. Ramm, Dynamical Systems Gradient method for solving nonlinear equations with monotone operators, *Acta Appl. Math.*, 106 (2009), N3, 473–499.

[5] V. Ivanov, V. Tanana and V. Vasin, *Theory of ill-posed problems*, VSP, Utrecht, 2002.

[6] V. A. Morozov, *Methods of solving incorrectly posed problems*, Springer Verlag, New York, 1984.

[7] A. G. Ramm, *Dynamical systems method for solving operator equations*, Elsevier, Amsterdam, 2007.

[8] A. G. Ramm, Global convergence for ill-posed equations with monotone operators: the dynamical systems method, *J. Phys A*, 36, (2003), L249-L254.

[9] A. G. Ramm, Dynamical systems method for solving nonlinear operator equations, International Jour. of *Applied Math. Sci.*, 1, N1, (2004), 97-110.

[10] A. G. Ramm, Dynamical systems method for solving operator equations, *Communic. in Nonlinear Sci. and Numer. Simulation*, 9, N2, (2004), 383-402.

[11] A. G. Ramm, DSM for ill-posed equations with monotone operators, *Comm. in Nonlinear Sci. and Numer. Simulation*, 10, N8, (2005),935-940.

[12] A. G. Ramm, Discrepancy principle for the dynamical systems method, *Communic. in Nonlinear Sci. and Numer. Simulation*, 10, N1, (2005), 95-101

[13] A. G. Ramm, Dynamical systems method (DSM) and nonlinear problems, in the book: *Spectral Theory and Nonlinear Analysis*, World Scientific Publishers, Singapore, 2005, 201-228. (ed J. Lopez-Gomez).

[14] A. G. Ramm, *Dynamical systems method (DSM) for unbounded operators*, Proc. Amer. Math. Soc., 134, N4, (2006), 1059-1063.

[15] E. Zeidler, *Nonlinear functional analysis*, Springer, New York, 1985.