Rigorous numerics for nonlinear operators with tridiagonal dominant linear parts

Maxime Breden * Laurent Desvillettes † Jean-Philippe Lessard‡

Abstract

We propose a method to compute solutions of infinite dimensional nonlinear operators $f(x) = 0$ with tridiagonal dominant linear parts. We recast the operator equation into an equivalent Newton-like equation $x = T(x) = x - Af(x)$, where A is an approximate inverse of the derivative $Df(\bar{x})$ at an approximate solution $\bar{x}$. We develop rigorous computer-assisted bounds to show that T is a contraction near $\bar{x}$, which yields existence of a solution. Since $Df(\bar{x})$ does not have an asymptotically diagonal dominant structure, the computation of A is not straightforward. This paper provides a method to obtain A and proposes a new rigorous computational method to prove existence of solutions of nonlinear operators with tridiagonal dominant linear parts.

Keywords

Tridiagonal operator · Contraction mapping · Rigorous numerics · Fourier series

Mathematics Subject Classification (2010)

47H10 · 42A10 · 65L10 · 34B08

1 Introduction

Tridiagonal operators arise naturally in the theory of orthogonal polynomials, ordinary differential equations (ODEs), continued fractions, numerical analysis of partial differential equations (PDEs), integrable systems, quantum mechanics and solid state physics. Some differential operators can be represented by infinite tridiagonal matrices acting in sequence spaces, as it is the case for instance for differentiation in frequency space of the Hermite functions. Other examples come from the study of ODEs like the Mathieu equation, the spheroidal wave equation, the Whittaker-Hill equation and the Lamé equation.

While there exist many well developed methods and efficient algorithms in the literature for solving linear tridiagonal matrix equations and computing their inverses, our proposed method has a different flavour. We aim at developing a computational method to prove, in a mathematically rigorous and constructive sense, existence of solutions of infinite dimensional nonlinear operators of the form

$$f(x) = \mathcal{L}(x) + N(x) = 0, \tag{1}$$

*CMLA, ENS Cachan & CNRS, 61 avenue du Président Wilson, 94235 Cachan, France. mbreden@ens-cachan.fr

†CMLA, ENS Cachan & CNRS, 61 avenue du Président Wilson, 94235 Cachan, France. desville@cmla.ens-cachan.fr

‡Département de Mathématiques et de Statistique, Université Laval, 1045 avenue de la Médecine, Québec, QC, G1V0A6, Canada. jean-philippe.lessard@mat.ulaval.ca

where $\mathcal{L}$ is a tridiagonal linear operator and N is a nonlinear operator. The domain of the operator f is the space of algebraically decaying sequences

$$\Omega^s \stackrel{\text{def}}{=} \left\{ x = (x_k)_{k\geq 0} : \|x\|_s \stackrel{\text{def}}{=} \sup_{k\geq 0}\{|x_k|k^s\} < \infty \right\}. \tag{2}$$

The assumptions on the linear and nonlinear parts of (1) are that $\mathcal{L} : \Omega^s \to \Omega^{s-s_L}$ and $N : \Omega^s \to \Omega^{s-s_N}$, for some $s_L > s_N$. Intuitively, this means that the linear part *dominates* the nonlinear part. Since $\Omega^{s_1} \subset \Omega^{s_2}$ for $s_1 > s_2$, one has that $f : \Omega^s \to \Omega^{s-s_L}$.

General nonlinear operators $f(x) = 0$ defined on Ω^s arise in the study of bounded solutions of finite and infinite dimensional dynamical systems. For instance, $x = (x_k)_{k\geq 0}$ may be the infinite sequence of Fourier coefficients of a periodic solution of an ODE, a periodic solution of a delay differential equation (DDE) or an equilibrium solution of a PDE with Dirichlet, periodic or Neumann boundary conditions. The unknown x may also be the infinite sequence of Chebyshev coefficients of a solution of a boundary value problem (BVP), the Hermite coefficients of a solution of an ODE defined on an unbounded domain or the Taylor coefficients of the solution of a Cauchy problem. In case the differential equation is smooth, the decay rate of the coefficients of x will be algebraic or even exponential [1]. In this paper, we chose to solve (1) in the weighed ℓ^∞ Banach space Ω^s which corresponds to C^k solutions. In order to exploit the analyticity of the solutions, we could follow the idea of [2] and solve (1) in weighed ℓ^1 Banach spaces. This choice of space is not considered in the present paper.

In recent years, there has been several successful attempts in solving $f(x) = 0$ in Ω^s using the field of *rigorous numerics*. This field aims at constructing algorithms that provide an approximate solution to the problem together with precise bounds within which the exact solution is guaranteed to exist in the mathematically rigorous sense. Equilibria of PDEs [3, 4, 5], periodic solutions of DDEs [6], fixed points of infinite dimensional maps [7] and periodic solutions of ODEs [8, 9] have been computed using such methods.

One popular idea in rigorous numerics is to recast the problem $f(x) = 0$ as a problem of looking for the fixed points of a Newton-like equation of the form $T(x) = x - Af(x)$, where A is an approximate inverse of $Df(\bar{x})$, where $\bar{x}$ is a numerical approximations obtained by computing on a finite dimensional projection of f. In [3, 4, 6, 7, 9, 5], the nonlinear equations under study have asymptotically diagonal or block-diagonal dominant linear parts which facilitates the computation of approximate inverses. In contrast, the present work considers problems with tridiagonal dominant linear parts. To the best of our knowledge, this is first attempt to compute rigorously solutions of such problems. While our proposed approach is designed for the moment for a specific class of operators (see assumptions (4) and (5)), we believe it is a first step toward solving rigorously more complicated nonlinear operators with tridiagonal dominant linear parts.

The paper is organized as follows. In Section 2, we present the method to compute, with the help of the computer, pseudo-inverses of tridiagonal operators of a certain class. In Section 3, we recast the problem $f(x) = 0$ as a fixed point problem $T(x) = x - Af(x)$ where A is a pseudo-inverse, and we present the rigorous computational method to prove existence of fixed points of T. In Section 4, we present an application and finally in Section 5, we conclude by presenting some interesting future directions.

2 Computing pseudo-inverses of tridiagonal operators

Given three sequences $(\lambda_k)_{k\geq 0}$, $(\mu_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}$ and $x \in \Omega^s$, define formally the tridiagonal linear operator $\mathcal{L}(x) = (\mathcal{L}_k(x))_{k\geq 0}$ of (1) by

$$\mathcal{L}_k(x) = \lambda_k x_{k-1} + \mu_k x_k + \beta_k x_{k+1}, \quad k \geq 1 \tag{3}$$

and $\mathcal{L}_0(x) = \mu_0 x_0 + \beta_0 x_1$. Assume that there exist real numbers $s_L > 1$, $0 < C_1 \leq C_2$ and an integer k_0 such that

$$\forall\ k \geq 1, \quad \left|\frac{\lambda_k}{k^{s_L}}\right|, \left|\frac{\mu_k}{k^{s_L}}\right|, \left|\frac{\beta_k}{k^{s_L}}\right| \leq C_2 \quad \text{and} \quad \forall\ k \geq k_0, \quad C_1 \leq \left|\frac{\mu_k}{k^{s_L}}\right|. \tag{4}$$

Assume further the existence of $\delta \in \left(0, \frac{1}{2}\right)$ and $k_0 \geq 0$ such that

$$\forall\ k \geq k_0, \quad \left|\frac{\lambda_k}{\mu_k}\right|, \left|\frac{\beta_k}{\mu_k}\right| \leq \delta. \tag{5}$$

Therefore under assumptions (4) and (5), $\mathcal{L}$ defined by (3) is a tridiagonal operator such that $\mathcal{L} : \Omega^s \to \Omega^{s-s_L}$. Indeed, for $x \in \Omega^s$, then

$$\begin{aligned} \|\mathcal{L}(x)\|_{s-s_L} &= \sup_{k\geq 0}\{|\mathcal{L}_k(x)|k^{s-s_L}\} \\ &\leq C_2\left(\sup_{k\geq 0}\{|x_{k-1}|k^s\} + \sup_{k\geq 0}\{|x_k|k^s\} + \sup_{k\geq 0}\{|x_{k+1}|k^s\}\right) < \infty. \end{aligned}$$

From now on, assume for sake of simplicity that $s_N = 0$, that is the nonlinear part N of (1) satisfies $N : \Omega^s \to \Omega^s$. Note that any combination of convolutions satisfies this property by the algebra property of Ω^s for $s > 1$ [5, 10]. Assume that using a finite dimensional projection $f^{(m)} : \mathbb{R}^m \to \mathbb{R}^m$ of (1), we computed a numerical approximation $\bar{x}$, that is $f^{(m)}(\bar{x}) \approx 0$. We identify $\bar{x} \in \mathbb{R}^m$ and $\bar{x} = (\bar{x}, 0, 0, 0, 0, \dots) \in \Omega^s$. The idea is to construct a ball

$$B_{\bar{x}}(r) = \bar{x} + B_0(r) = \bar{x} + \{x \in \Omega^s : \|x\|_s \leq r\} = \{x \in \Omega^s : \|x - \bar{x}\|_s \leq r\}$$

centered at $\bar{x}$ containing a unique solution of (1) by showing that a certain Newton-like operator $T(x) = x - Af(x)$ is a contraction on $B_{\bar{x}}(r)$. This requires constructing A an approximate inverse of $Df(\bar{x}) = \mathcal{L}(\bar{x}) + DN(\bar{x})$. In order to do so, the structures of $\mathcal{L}(\bar{x})$ and $DN(\bar{x})$ need to be understood. From (3) and (4), $\mathcal{L}(\bar{x})$ is a tridiagonal operator with entries growing to infinity at the rate k^{s_L}. Moreover, since $DN(\bar{x}) : \Omega^s \to \Omega^s$, then it is a bounded linear operator. As mentioned above, the expectation is that the coefficients of $\bar{x}$ decay fast to zero. Therefore, a reasonable approximation $A^\dagger$ for $Df(\bar{x})$ is given by

$$A^\dagger = \begin{pmatrix} D & & & 0 & \\ & & \beta_{m-1} & & \\ & \lambda_m & \mu_m & \beta_m & \\ 0 & & \lambda_{m+1} & \mu_{m+1} & \beta_{m+1} \end{pmatrix}, \tag{6}$$

with $D \overset{\text{def}}{=} Df^{(m)}(\bar{x})$ for m large enough. Again we assume that $m \geq k_0$. We wish to find its inverse in terms of D, $(\beta_k)_{k\geq m-1}$, $(\mu_k)_{k\geq m}$ and $(\lambda_k)_{k\geq m}$. We assume therefore that

$$A^\dagger x = y, \tag{7}$$

where x and y are the infinite vectors

$$x = \begin{pmatrix} x_0 \\ x_1 \\ . \\ . \end{pmatrix}, \qquad y = \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \end{pmatrix}.$$

The infinite part of (7) writes

$$\begin{pmatrix} \mu_m & \beta_m & 0 & 0 & ... \\ \lambda_{m+1} & \mu_{m+1} & \beta_{m+1} & 0 & ... \\ 0 & \lambda_{m+2} & \mu_{m+2} & \beta_{m+2} & ... \\ . & . & . & . & ... \end{pmatrix} \begin{pmatrix} x_m \\ x_{m+1} \\ . \\ . \end{pmatrix} = \begin{pmatrix} y_m - \lambda_m\, x_{m-1} \\ y_{m+1} \\ . \\ . \end{pmatrix}. \tag{8}$$

We introduce the notations of the book of P.G. Ciarlet from Theorem 4.3-2 on page 142 in [11]. Note that only the δ_n are really useful:

$$a_2 = \lambda_{m+1}, \quad a_3 = \lambda_{m+2}, ..., \qquad b_1 = \mu_m, \quad b_2 = \mu_{m+1}, ..., \qquad c_1 = \beta_m, \quad c_2 = \beta_{m+1},$$

and $(\delta_n)_{n\in\mathbb{N}}$ defined by the induction formula

$$\delta_0 = 1, \quad \delta_1 = b_1, \quad \text{and} \quad \delta_n = b_n\,\delta_{n-1} - a_n\,c_{n-1}\,\delta_{n-2}, \quad \text{for} \quad n \geq 2.$$

Let define the tridiagonal operator T by

$$T = \begin{pmatrix} b_1 & c_1 & 0 & 0 & ... \\ a_2 & b_2 & c_2 & 0 & ... \\ 0 & a_3 & b_3 & c_3 & ... \\ . & . & . & . & ... \end{pmatrix}. \tag{9}$$

For any infinite vector $x = (x_0, \ldots, x_k, \ldots)^T$, we introduce the notation

$$x_F = (x_0, \ldots, x_{m-1})^T \quad \text{and} \quad x_I = (x_m, \ldots, x_{m+k}, \ldots)^T.$$

Using the notation $\mathbf{e_1} = (1, 0, 0, 0, 0, \cdots)^T$, the system (8) becomes

$$Tx_I = y_I - \lambda_m\, x_{m-1}\mathbf{e_1}.$$

From Theorem 4.3-2 in [11], we compute an LU-decomposition of the above tridiagonal operator in (9) as $T = L_I U_I$, where

$$L_I = \begin{pmatrix} 1 & 0 & 0 & ... \\ a_2\,\frac{\delta_0}{\delta_1} & 1 & 0 & ... \\ 0 & a_3\,\frac{\delta_1}{\delta_2} & 1 & ... \\ . & . & . & ... \end{pmatrix} \quad \text{and} \quad U_I = \begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & ... \\ 0 & \frac{\delta_2}{\delta_1} & c_2 & ... \\ 0 & 0 & \frac{\delta_3}{\delta_2} & ... \\ . & . & . & ... \end{pmatrix}. \tag{10}$$

Hence, the system (8) becomes $L_I z_I = y_I - \lambda_m\, x_{m-1}\mathbf{e_1}$ combined with $U_I x_I = z_I$, that is

$$\begin{pmatrix} 1 & 0 & 0 & ... \\ a_2\,\frac{\delta_0}{\delta_1} & 1 & 0 & ... \\ 0 & a_3\,\frac{\delta_1}{\delta_2} & 1 & ... \\ . & . & . & ... \end{pmatrix} \begin{pmatrix} z_m \\ z_{m+1} \\ . \\ . \end{pmatrix} = \begin{pmatrix} y_m - \lambda_m\, x_{m-1} \\ y_{m+1} \\ . \\ . \end{pmatrix}, \tag{11}$$

combined with

$$\begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & ... \\ 0 & \frac{\delta_2}{\delta_1} & c_2 & ... \\ 0 & 0 & \frac{\delta_3}{\delta_2} & ... \\ . & . & . & ... \end{pmatrix} \begin{pmatrix} x_m \\ x_{m+1} \\ . \end{pmatrix} = \begin{pmatrix} z_m \\ z_{m+1} \\ . \end{pmatrix}. \tag{12}$$

Both infinite systems (11) and (12) can be solved explicitly.

System (11) leads to

$$z_m = y_m - \lambda_m x_{m-1},$$

and for any $k \geq 1$

$$z_{m+k} = y_{m+k} + \sum_{l=1}^{k} (-1)^l \, a_{k-l+2} \, .. \, a_{k+1} \, \frac{\delta_{k-l}}{\delta_k} \, y_{m+k-l} + (-1)^{k+1} \, a_2 \, .. \, a_{k+1} \, \frac{\delta_0}{\delta_k} \, \lambda_m x_{m-1}$$

which we rewrite with infinite matrix/vectors notations as

$$z_I = {L_I}^{-1}[y_I - \lambda_m \, x_{m-1}\mathbf{e_1}] = {L_I}^{-1} y_I - \lambda_m x_{m-1} \, v_I, \tag{13}$$

where

$$z_I = \begin{pmatrix} z_m \\ z_{m+1} \\ z_{m+2} \\ \\ \vdots \end{pmatrix}, \quad y_I = \begin{pmatrix} y_m \\ y_{m+1} \\ y_{m+2} \\ \\ \vdots \end{pmatrix}, \quad v_I = {L_I}^{-1}\mathbf{e_1} = \begin{pmatrix} 1 \\ -a_2 \, \frac{\delta_0}{\delta_1} \\ a_3 \, a_2 \, \frac{\delta_0}{\delta_2} \\ -a_4 \, a_3 \, a_2 \, \frac{\delta_0}{\delta_3} \\ \vdots \end{pmatrix}.$$

The second system (12) leads to the infinite sum (for any $k \geq 0$)

$$x_{m+k} = \frac{\delta_k}{\delta_{k+1}} z_{m+k} + \sum_{l=1}^{\infty} (-1)^l \, \frac{\delta_k}{\delta_{k+l+1}} \, c_{k+1} \, .. \, c_{k+l} \, z_{m+k+l},$$

which we also rewrite with infinite matrix/vector notations as

$$x_I = {U_I}^{-1} z_I. \tag{14}$$

Coupling (13) and (14), we end up with

$$x_I = {U_I}^{-1} z_I = {U_I}^{-1}[{L_I}^{-1} y_I - \lambda_m x_{m-1} \, v_I] = {U_I}^{-1} {L_I}^{-1} y_I - \lambda_m x_{m-1} w_I, \tag{15}$$

where $w_I = U_I^{-1} v_I$. Denoting $\left({U_I}^{-1} {L_I}^{-1}\right)_{l0}$ the first line of the infinite matrix ${U_I}^{-1} {L_I}^{-1}$ (remember that $(w_I)_0$ denotes the first element of w_I), we can rewrite the first line of (15) as

$$x_m = \left({U_I}^{-1} {L_I}^{-1}\right)_{l0} y_I - \lambda_m x_{m-1} \, (w_I)_0 \, . \tag{16}$$

Then, we investigate the finite part of the linear system (7). It writes

$$D \begin{pmatrix} x_0 \\ x_1 \\ . \\ . \\ x_{m-2} \\ x_{m-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ \beta_{m-1} \, x_m \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \\ y_{m-2} \\ y_{m-1} \end{pmatrix},$$

or, according to (16),

$$D\begin{pmatrix} x_0 \\ x_1 \\ . \\ . \\ x_{m-2} \\ x_{m-1} \end{pmatrix} + \beta_{m-1}\begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ \left(U_I{}^{-1}L_I{}^{-1}\right)_{l0} y_I - \lambda_m x_{m-1}\left(w_I\right)_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \\ y_{m-2} \\ y_{m-1} \end{pmatrix}.$$

Defining

$$K = D - \beta_{m-1}\lambda_m \begin{pmatrix} 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & \left(w_I\right)_0 \end{pmatrix},$$

together with its inverse K^{-1} and resp. last column $(K^{-1})_{c\,m-1}$, last line $(K^{-1})_{l\,m-1}$, and last ("south-east") element $(K^{-1})_{m-1,m-1}$ of this inverse, we end up with

$$\begin{aligned} x_F &= K^{-1}y_F - \beta_{m-1}\left\{\left(U_I{}^{-1}L_I{}^{-1}\right)_{l0} y_I\right\}\left(K^{-1}\right)_{c\,m-1} \\ &= K^{-1}y_F - \beta_{m-1}\left(\left\{\left(K^{-1}\right)_{c\,m-1}\right\} \otimes \left\{\left(U_I^{-1}L_I{}^{-1}\right)_{l0}\right\}\right) y_I, \end{aligned} \tag{17}$$

with the tensor product notation. The last line of this identity reads

$$x_{m-1} = \left(K^{-1}\right)_{l\,m-1} y_F - \beta_{m-1}\left\{\left(U_I{}^{-1}L_I{}^{-1}\right)_{l0} y_I\right\}\left(K^{-1}\right)_{m-1,m-1}. \tag{18}$$

Coming back to (15) and using (18) we see that

$$\begin{aligned} x_I &= U_I{}^{-1}L_I{}^{-1}y_I - \lambda_m x_{m-1} w_I \\ &= U_I{}^{-1}L_I{}^{-1}y_I \\ &\quad -\lambda_m\left[\left(K^{-1}\right)_{l\,m-1} y_F - \beta_{m-1}\left\{\left(U_I{}^{-1}L_I{}^{-1}\right)_{l0} y_I\right\}\left(K^{-1}\right)_{m-1,m-1}\right] w_I \\ &= U_I{}^{-1}L_I{}^{-1}y_I - \lambda_m\, w_I\left\{\left(K^{-1}\right)_{l\,m-1} y_F\right\} \\ &\quad +\beta_{m-1}\lambda_m\left(K^{-1}\right)_{m-1,m-1} w_I\left\{\left(U_I{}^{-1}L_I{}^{-1}\right)_{l0} y_I\right\} \\ &= -\lambda_m\left(\left\{w_I\right\} \otimes \left\{\left(K^{-1}\right)_{l\,m-1}\right\}\right) y_F \\ &\quad +\left(U_I{}^{-1}L_I{}^{-1} + \beta_{m-1}\lambda_m\left(K^{-1}\right)_{m-1,m-1}\left\{w_I\right\} \otimes \left\{\left(U_I{}^{-1}L_I{}^{-1}\right)_{l0}\right\}\right) y_I. \end{aligned} \tag{19}$$

Putting together (17) and (19), we end up with

$$\left(A^\dagger\right)^{-1} = \begin{pmatrix} K^{-1} & -\beta_{m-1}\left(\left\{\left(K^{-1}\right)_{c\,m-1}\right\} \otimes \left\{\left(U_I^{-1}L_I{}^{-1}\right)_{l0}\right\}\right) \\ -\lambda_m\left\{w_I\right\} \otimes \left\{\left(K^{-1}\right)_{l\,m-1}\right\} & U_I{}^{-1}L_I{}^{-1} + \tilde{\Lambda} \end{pmatrix}.$$

where

$$\tilde{\Lambda} = \beta_{m-1}\lambda_m \left(K^{-1}\right)_{m-1,m-1} \left\{ w_I \right\} \otimes \left\{ \left({U_I}^{-1} {L_I}^{-1} \right)_{l0} \right\}.$$

Now to get an approximate (pseudo) inverse of $A^\dagger$ we would like to get a numerical approximation of K^{-1}. However the definition of K involves $(w_I)_0$ which cannot be computed explicitly. By definition $w_I = U_I^{-1} L_I^{-1} \mathbf{e_1}$ so using again the computations made in this section we get

$$\begin{aligned}
(w_I)_0 &= \left(U_I^{-1} v_I \right)_0 \\
&= \frac{\delta_0}{\delta_1} v_m + \sum_{l=1}^{\infty} (-1)^l \frac{\delta_0}{\delta_{l+1}} c_1 \dots c_l \, v_{m+l} \\
&= \frac{\delta_0}{\delta_1} + \sum_{l=1}^{\infty} \frac{\delta_0^2}{\delta_l \delta_{l+1}} c_1 \dots c_l \, a_2 \dots a_{l+1}.
\end{aligned}$$

So given a computational parameter L, we define

$$\tilde{w} = \frac{\delta_0}{\delta_1} + \sum_{l=1}^{L-1} \frac{\delta_0^2}{\delta_l \delta_{l+1}} c_1 \dots c_l \, a_2 \dots a_{l+1}, \tag{20}$$

and

$$\tilde{K} = D - \beta_{m-1}\lambda_m \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \tilde{w} \end{pmatrix}.$$

Now we can consider A_m a numerically computed inverse of $\tilde{K}$ and then define the approximate (pseudo) inverse of $A^\dagger$ as

$$A = \begin{pmatrix} A_m & -\beta_{m-1} \left(\left\{ (A_m)_{c\, m-1} \right\} \otimes \left\{ \left(U_I^{-1} {L_I}^{-1} \right)_{l0} \right\} \right) \\ -\lambda_m \left\{ w_I \right\} \otimes \left\{ (A_m)_{l\, m-1} \right\} & {U_I}^{-1} {L_I}^{-1} + \Lambda \end{pmatrix}, \tag{21}$$

where

$$\Lambda = \beta_{m-1}\lambda_m \left(A_m\right)_{m-1,m-1} \left\{ w_I \right\} \otimes \left\{ \left({U_I}^{-1} {L_I}^{-1} \right)_{l0} \right\}.$$

Lemma 2.1. *Assuming $m \geq k_0$ and $\delta < \frac{1}{2}$, $U_I^{-1} : \Omega^s \to \Omega^{s+s_L}$.*

Proof. Let $z_I \in \Omega^s$ and $x_I = U_I^{-1} z_I$. Using (14) and the formula above, we get

$$\begin{aligned}
|x_{m+k}| &\leq \frac{|\delta_k|}{|\delta_{k+1}|} |z_{m+k}| + \sum_{l=1}^{\infty} \frac{|\delta_k|}{|\delta_{k+l+1}|} \, |c_{k+1}| \, .. \, |c_{k+l}| \, |z_{m+k+l}| \\
&\leq \frac{|\delta_k|}{|\delta_{k+1}|} |z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{|\delta_k|}{|\delta_{k+l+1}|} \, |b_{k+1}| \, .. \, |b_{k+l}| \, |z_{m+k+l}| \, .
\end{aligned} \tag{22}$$

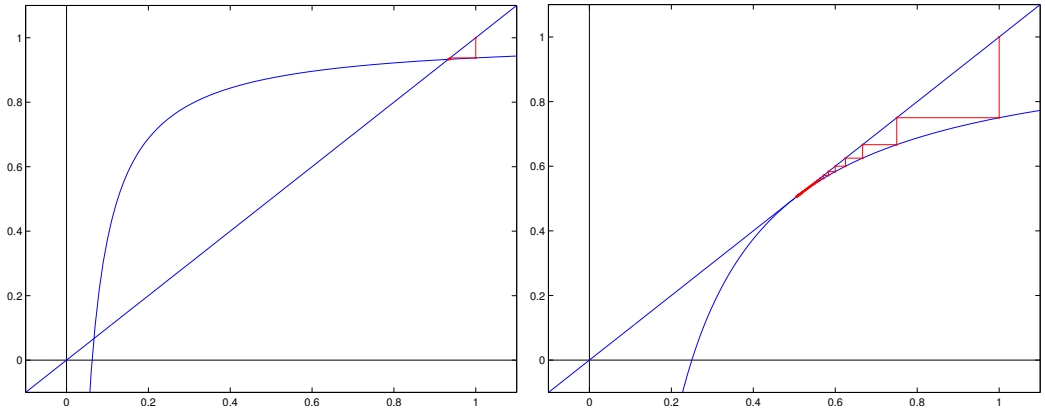
Now remember that for all $k \geq 2$, $\delta_k = b_k\,\delta_{k-1} - a_k\,c_{k-1}\delta_{k-2}$, so

$$\begin{aligned}\frac{|\delta_k|}{|\delta_{k-1}|\,|b_k|} &\geq 1 - \frac{|a_k|\,|c_{k-1}|\,|\delta_{k-2}|}{|b_k|\,|\delta_{k-1}|} \\ &\geq 1 - \frac{\delta^2\,|b_{k-1}|\,|\delta_{k-2}|}{|\delta_{k-1}|}.\end{aligned}$$

We introduce $u_k = \dfrac{|\delta_k|}{|\delta_{k-1}|\,|b_k|}$ which then satisfies

$$\begin{cases} u_1 = 1 \\ u_k \geq 1 - \dfrac{\delta^2}{u_{k-1}}, \quad \forall\ k \geq 2. \end{cases}$$

The study of the inductive sequence defined as above in the equality case yields that for any k, $\gamma \leq u_k \leq 1$, where $\gamma = \frac{1}{2} + \sqrt{\frac{1}{4} - \delta^2}$ is the largest root of $x = 1 - \frac{\delta^2}{x}$ (see figure 1).



(a) $\delta = 0.2$ (b) $\delta = 0.5$

Figure 1: The iterations of $u_{n+1} = 1 - \delta^2/u_n$ with $u_1 = 1$.

We can then rewrite (22) to get

$$\begin{aligned}|x_{m+k}| &\leq \frac{|\delta_k|}{|\delta_{k+1}|}\,|z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{|\delta_k|\,..\,|\delta_{k+l}|}{|\delta_{k+1}|\,..\,|\delta_{k+l+1}|}\,|b_{k+1}|\,..\,|b_{k+l}|\,|z_{m+k+l}| \\ &\leq \frac{|\delta_k|}{|\delta_{k+1}|}\,|z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{1}{u_{k+1}}\,..\,\frac{1}{u_{k+l}}\,\frac{|\delta_{k+l}|}{|\delta_{k+l+1}|}\,|z_{m+k+l}| \\ &\leq \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{|\delta_{k+l}|}{|\delta_{k+l+1}|}\,|z_{m+k+l}| \\ &\leq \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{1}{\gamma\,|b_{k+l+1}|}\,|z_{m+k+l}| \\ &\leq \frac{\|z_I\|_s}{C_1\gamma} \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{1}{(k+l+1)^{s_L}\,(m+k+l)^s}.\end{aligned}$$

Finally, since $\delta < \dfrac{1}{2} < \gamma$,

$$|x_{m+k}|\,(m+k)^{s+s_L} \leq \frac{\|z_I\|_s}{C_1\gamma}\,\frac{1}{1-\frac{\delta}{\gamma}}\,\frac{(m+k)^{s+s_L}}{(k+1)^{s_L}\,(m+k)^s}$$

and $x_I \in \Omega^{s+s_L}$. □

Lemma 2.2. *Assuming $m \geq k_0$ and $\delta < \frac{1}{2}$, $L_I^{-1} : \Omega^s \to \Omega^s$.*

Proof. Let $y_I \in \Omega^s$ and $z_I = L_I^{-1} y_I$. Using (13) and the formula above (without the last term since we do not consider here $L_I^{-1}(y_I - \lambda_m\, x_{m-1}\mathbf{e_1})$), we get

$$\begin{aligned}
|z_{m+k}| &\leq |y_{m+k}| + \sum_{l=1}^{k} \frac{|\delta_{k-l}|}{|\delta_k|}\ |a_{k-l+2}|\ ..\ |a_{k+1}|\ |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^{k} \delta^l\,\frac{|\delta_{k-l}|}{|\delta_k|}\ |b_{k-l+2}|\ ..\ |b_{k+1}|\ |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^{k} \delta^l\,\frac{|\delta_{k-l}|\ ..\ |\delta_{k-1}|}{|\delta_{k-l+1}|\ ..\ |\delta_k|}\ \frac{|b_{k-l+1}|\,|b_{k-l+2}|\ ..\ |b_{k+1}|}{|b_{k-l+1}|}\ |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^{k} \delta^l\,\frac{1}{u_{k-l+1}} .. \frac{1}{u_k}\,\frac{|b_{k+1}|}{|b_{k-l+1}|}\,|y_{m+k-l}|\,,
\end{aligned}$$

where we use the sequence u_k introduced in the previous proof. So we get

$$|z_{m+k}| \leq \sum_{l=0}^{k} \left(\frac{\delta}{\gamma}\right)^l \frac{|b_{k+1}|}{|b_{k-l+1}|}\,|y_{m+k-l}|\,,$$

and

$$|z_{m+k}|\,(m+k)^s \leq \frac{C_2\,\|y\|_s}{C_1} \sum_{l=0}^{k} \left(\frac{\delta}{\gamma}\right)^l \left(\frac{k+1}{k+1-l}\right)^{s_L} \left(\frac{m+k}{m+k-l}\right)^s,$$

and it is enough to show that given $0 < \theta < 1$ and $q \geq 0$, $\displaystyle\sum_{l=1}^{k-1} \theta^l \left(\frac{k}{k-l}\right)^q$ is bounded uniformly in k to prove that $z_I \in \Omega^s$. Indeed,

$$\begin{aligned}
\sum_{l=1}^{k-1} \theta^l \left(\frac{k}{k-l}\right)^q &\leq \sum_{l=1}^{\left[\frac{k}{2}\right]} \theta^l \left(\frac{k}{k-l}\right)^q + \sum_{l=\left[\frac{k}{2}\right]+1}^{k-1} \theta^l \left(\frac{k}{k-l}\right)^q \\
&\leq 2^q \sum_{l=1}^{\left[\frac{k}{2}\right]} \theta^l + \theta^{\left[\frac{k}{2}\right]+1} \frac{k}{2} k^q \\
&\leq \frac{2^q}{1-\theta} + \theta^{\left[\frac{k}{2}\right]+1} \frac{k^{q+1}}{2},
\end{aligned}$$

which is bounded uniformly in k since the last term goes to 0 when k goes to $+\infty$, and the proof is complete.

□

Proposition 2.3. *Assuming $m \geq k_0$ and $\delta < \frac{1}{2}$, $A : \Omega^s \to \Omega^{s+s_L}$.*

Proof. Consider $y = (y_F, y_I)^T \in \Omega^s$. Let $x = (x_F, x_I)^T = Ay$. Then, by definition of the operator A in (21),

$$\begin{aligned} x_F &= A_m y_F - \beta_{m-1} \left(\left\{ (A_m)_{c\, m-1} \right\} \otimes \left\{ \left({U_I}^{-1} {L_I}^{-1} \right)_{l0} \right\} \right) y_I \\ &= A_m y_F - \beta_{m-1} \left\{ \left({U_I}^{-1} {L_I}^{-1} \right)_{l0} y_I \right\} (A_m)_{c\, m-1}. \end{aligned}$$

By the previous lemmas, ${U_I}^{-1} {L_I}^{-1} y_I \in \Omega^{s+sL}$, in particular $\left({U_I}^{-1} {L_I}^{-1} \right)_{l0} y_I = \left({U_I}^{-1} {L_I}^{-1} y_I \right)_0$ is well defined and so is x_F.

Again using (21),

$$x_I = -\lambda_m \left(\left\{ w_I \right\} \otimes \left\{ (A_m)_{l\, m-1} \right\} \right) y_F + {U_I}^{-1} {L_I}^{-1} y_I + \Lambda y_I.$$

Remember that $w_I = U_I^{-1} {L_I}^{-1} \mathbf{e_1}$, and therefore $w_I \in \Omega^s$ for any s, so according to the previous lemmas and the definition of Λ (see (21)), $x_I \in \Omega^{s+s_L}$. □

3 Computations of fixed points of the operator T

Our main motivation for computing approximate inverses is to prove existence, in a mathematically rigorous sense, of a fixed point of the Newton-like operators T in a set centered at a numerical approximation $\bar{x}$. The Newton-like operator has the form

$$T(x) = x - Af(x), \tag{23}$$

where A is the approximate inverse (21) of $Df(\bar{x})$ computed using the theory of Section 2. By Proposition 2.3, $A : \Omega^s \to \Omega^{s+s_L}$. Since $f : \Omega^s \to \Omega^{s-s_L}$, then $T : \Omega^s \to \Omega^s$.

Before proceeding further, we endow Ω^s with the operation of discrete convolution. More precisely, given $x = (x_k)_{k\geq 0}, y = (y_k)_{k\geq 0} \in \Omega^s$, extend x, y symmetrically by $\tilde{x} = (x_k)_{k\in\mathbb{Z}}, \tilde{y} = (y_k)_{k\in\mathbb{Z}}$ where $\tilde{x}_{-k} = x_k$, $\tilde{y}_{-k} = y_k$, for $k \geq 1$. The discrete convolution of x and y is denoted by $x * y$ and defined by

$$(x * y)_k = \sum_{\substack{k_1+k_2=k \\ k_1,k_2\in\mathbb{Z}}} \tilde{x}_{k_1} \tilde{y}_{k_2}.$$

It is known that for $s > 1$, $(\Omega^s, *)$ is an algebra under discrete convolution (e.g. see [10]), that is, given $x, y \in \Omega^s$, $x * y \in \Omega^s$. This fact will be especially useful when computing in practice the Z bounds as defined in the following results.

Theorem 3.1. *For fixed $s > 1$, consider $T\colon \Omega^s \to \Omega^s$ with $T = (T_k)_{k\geq 0}$, $T_k \in \mathbb{R}$. Assume that there exists a point $\bar{x} \in \Omega^s$ and vectors $Y = \{Y_k\}_{k\geq 0}$ and $Z = \{Z_k(r)\}_{k\geq 0}$ with $Y_k, Z_k(r) \in \mathbb{R}$ satisfying*

$$|(T(\bar{x}) - \bar{x})_k| \leq Y_k \tag{24}$$

and

$$\sup_{b_1,b_2\in B_0(r)} \left| \left[DT(\bar{x} + b_1)b_2 \right]_k \right| \leq Z_k(r), \quad \forall\ k \geq 0. \tag{25}$$

If there exists $r > 0$ such that $\|Y + Z(r)\|_s < r$, then the operator T is a contraction in $B_{\bar{x}}(r)$ and there exists a unique $\hat{x} \in B_{\bar{x}}(r)$ that satisfies $T(\hat{x}) = \hat{x}$.

Proof. We perform the proof in two parts:

i) $\|T(x) - \bar{x}\|_s < r$ for all $x \in B_{\bar{x}}(r)$. This implies that $T(B_{\bar{x}}(r)) \subset B_{\bar{x}}(r)$;

ii) T is a contraction, that is there exists $\kappa \in (0,1)$ such that for every $x, y \in B_{\bar{x}}(r)$, one has that $\|T(x) - T(y)\|_s \leq \kappa \|x - y\|_s$.

For all $k \geq 0$ and for any $x, y \in B_{\bar{x}}(r)$, the Mean Value Theorem implies that

$$T_k(x) - T_k(y) = DT_k(z)(x - y)$$

for some $z = z(k) \in \{tx + (1-t)y : t \in [0,1]\} \subset B_{\bar{x}}(r)$. Since $r\frac{(x-y)}{\|x-y\|_s} \in B_0(r)$ then from (25)

$$|T_k(x) - T_k(y)| = \left| DT_k(z) \frac{r(x-y)}{\|x-y\|_s} \right| \frac{1}{r} \|x - y\|_s \leq \frac{Z_k(r)}{r} \|x - y\|_s. \tag{26}$$

The triangular inequality applied component-wise using $y = \bar{x}$ above gives

$$|T_k(x) - \bar{x}_k| \leq |T_k(x) - T_k(\bar{x})| + |T_k(\bar{x}) - \bar{x}_k| \leq Z_k(r) + Y_k$$

and hence

$$|T_k(x) - \bar{x}_k| \leq |Y_k + Z_k(r)|.$$

Therefore for any $x \in B_{\bar{x}}(r)$

$$\|T(x) - \bar{x}\|_s = \sup_{k \geq 0} \{|T_k(x) - \bar{x}_k| \omega_k^s\} \leq \sup_{k \geq 0} \{|Y_k + Z_k(r)| \omega_k^s\} = \|Y + Z(r)\|_s < r.$$

This proves that $T(B_{\bar{x}}(r)) \subset B_{\bar{x}}(r)$. From (26), we obtain that for any $x, y \in B_{\bar{x}}(r)$, $|T_k(x) - T_k(y)| \leq \frac{|Z_k(r)|}{r} \|x - y\|_s$ and thus

$$\|T(x) - T(y)\|_s \leq \frac{\|Z(r)\|_s}{r} \|x - y\|_s. \tag{27}$$

Since $\|Z(r)\|_s \leq \|Y + Z(r)\|_s < r$, then

$$\kappa \stackrel{\text{def}}{=} \frac{\|Z(r)\|_s}{r} < 1,$$

and we can conclude that $T : B_{\bar{x}}(r) \to B_{\bar{x}}(r)$ is a contraction. An application of the Contraction Mapping Theorem on the complete metric space $B_{\bar{x}}(r)$ gives the existence and unicity of a solution $\hat{x}$ of the equation $T(x) = x$ in $B_{\bar{x}}(r)$. □

Now we are going to see how to get such bounds Y (Section 3.2) and $Z(r)$ (Section 3.3) as well as how to find in an efficient way a $r > 0$ such that $\|Y + Z(r)\|_s < r$ (Section 3.4). We start by computing some estimate to control the action of $U_I^{-1} L_I^{-1}$.

3.1 Some preliminary computations

Suppose $y_I = (y_m, y_{m+1}, \ldots)^T$ is a given infinite vector. We want to bound component wise $x_I = U_I^{-1} L_I^{-1} y_I$. Let $\theta = \frac{\delta}{\gamma}$. Again introducing $z_I = L_I^{-1} y_I$ and using the computations made in the proof of lemma 2.1 and lemma 2.2, we get

$$|x_{m+k}| \leq \frac{1}{\gamma} \sum_{l=0}^{+\infty} \theta^l \frac{1}{|b_{k+l+1}|} |z_{m+k+l}|$$

and

$$|z_{m+k}| \leq \sum_{l=0}^{k} \theta^{k-l} \frac{|b_{k+1}|}{|b_{l+1}|} |y_{m+l}| .$$

Putting the two together,

$$\begin{aligned}
|x_{m+k}| &\leq \frac{1}{\gamma} \sum_{l=0}^{+\infty} \sum_{j=0}^{k+l} \theta^{k+2l-j} \frac{|y_{m+j}|}{|b_{j+1}|} \\
&= \frac{1}{\gamma} \left(\sum_{j=0}^{k} \frac{|y_{m+j}|}{|b_{j+1}|} \sum_{l=0}^{+\infty} \theta^{k+2l-j} + \sum_{j=k+1}^{+\infty} \frac{|y_{m+j}|}{|b_{j+1}|} \sum_{l=j-k}^{+\infty} \theta^{k+2l-j} \right) \\
&= \frac{1}{\gamma} \left(\sum_{j=0}^{k} \frac{|y_{m+j}|}{|b_{j+1}|} \frac{\theta^{k-j}}{1-\theta^2} + \sum_{j=k+1}^{+\infty} \frac{|y_{m+j}|}{|b_{j+1}|} \frac{\theta^{j-k}}{1-\theta^2} \right) \\
&= \frac{1}{\gamma(1-\theta^2)} \left(\sum_{j=0}^{k} \theta^{k-j} \frac{|y_{m+j}|}{|b_{j+1}|} + \sum_{j=k+1}^{+\infty} \theta^{j-k} \frac{|y_{m+j}|}{|b_{j+1}|} \right) \\
&= \eta \left(\sum_{j=0}^{k} \theta^{k-j} \frac{|y_{m+j}|}{|\mu_{m+j}|} + \sum_{j=k+1}^{+\infty} \theta^{j-k} \frac{|y_{m+j}|}{|\mu_{m+j}|} \right),
\end{aligned}$$

with

$$\eta = \frac{1}{\gamma(1-\theta^2)}.$$

In particular, for $w_I = (w_m, w_{m+1}, \ldots)^T \stackrel{\text{def}}{=} U_I^{-1} L_I^{-1} \mathbf{e_1}$ we have for all $k \geq 0$

$$|w_{m+k}| \leq \eta \theta^k \frac{1}{|\mu_m|}. \tag{28}$$

More generally, if y is such that $y_{m+k} = 0$ for any $k \geq K$, then

$$\forall\ k \leq K-2, \quad |x_{m+k}| \leq \eta \left(\sum_{l=0}^{k} \theta^{k-l} \frac{|y_{m+l}|}{|\mu_{m+l}|} + \sum_{l=k+1}^{K-1} \theta^{l-k} \frac{|y_{m+l}|}{|\mu_{m+l}|} \right) \tag{29}$$

and

$$\forall\ k \geq K-1, \quad |x_{m+k}| \leq \eta \theta^k \sum_{l=0}^{K-1} \frac{|y_{m+l}|}{\theta^l |\mu_{m+l}|}. \tag{30}$$

We will also need a bound of $|x_{m+k}| (m+k)^{s+s_L}$ that is uniform in k for k large enough.

$$\begin{aligned}
|x_{m+k}| (m+k)^{s+s_L} &\leq \frac{\eta \|y_I\|_s}{C_1} \left(\sum_{l=0}^{k} \theta^{k-l} \left(\frac{m+k}{m+l} \right)^{s+s_L} + \sum_{l=k+1}^{+\infty} \theta^{l-k} \left(\frac{m+k}{m+l} \right)^{s+s_L} \right) \\
&\leq \frac{\eta \|y_I\|_s}{C_1} \left(\sum_{l=0}^{k} \theta^{k-l} \left(\frac{m+k}{m+l} \right)^{s+s_L} + \frac{\theta}{1-\theta} \right).
\end{aligned}$$

We assume here that $m \geq 2$ (which will always be the case in practice), fix a computational parameter $M \in \mathbb{N}$ such that

$$M \geq \max\left(\frac{-m\ln\sqrt{\theta} - s - s_L - 1 - \sqrt{(m\ln\sqrt{\theta} + s + s_L + 1)^2 - 4m\ln\sqrt{\theta}}}{2\ln\sqrt{\theta}}, \frac{4}{(\ln\theta)^2}, m\right), \tag{31}$$

and say that for all $k < M$,

$$|x_{m+k}|\,(m+k)^{s+s_L} \leq \frac{\eta\|y_I\|_s}{C_1}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \frac{\theta}{1-\theta}\right). \tag{32}$$

Then for $k \geq M$, we split the remaining sum

$$\begin{aligned}\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} &= \sum_{l=0}^{\left[\frac{k}{2}\right]-1}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \sum_{l=\left[\frac{k}{2}\right]}^{k-\left[\sqrt{k}\right]-1}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \sum_{l=k-\left[\sqrt{k}\right]}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} \\ &\leq \theta^{\frac{k}{2}}\frac{k}{2}\left(\frac{m+k}{m}\right)^{s+s_L} + \theta^{\sqrt{k}}\frac{k}{2}2^{s+s_L} + \frac{1}{1-\theta}\left(\frac{m+k}{m+k-\sqrt{k}-1}\right)^{s+s_L} \\ &\leq \theta^{\frac{M}{2}}\frac{M}{2}\left(\frac{m+M}{m}\right)^{s+s_L} + \theta^{\sqrt{M}}\frac{M}{2}2^{s+s_L} + \frac{1}{1-\theta}\left(\frac{m+M}{m+M-\sqrt{M}-1}\right)^{s+s_L}.\end{aligned} \tag{33}$$

Let us detail a bit why the last inequality holds. First consider, for $x > 0$, $\varphi_1(x) = \theta^{\frac{x}{2}}x(m+x)^{s+s_L}$, whose derivative is

$$\begin{aligned}\varphi_1'(x) &= \sqrt{\theta}^x\left(\left(\ln\sqrt{\theta}\right)x(m+x)^{s+s_L} + (m+x)^{s+s_L} + (s+s_L)x\,(m+x)^{s+s_L-1}\right) \\ &= (m+x)^{s+s_L-1}\sqrt{\theta}^x\left(\left(\ln\sqrt{\theta}\right)(m+x)x + (m+x) + (s+s_L)x\right) \\ &= (m+x)^{s+s_L-1}\sqrt{\theta}^x\left(\left(\ln\sqrt{\theta}\right)x^2 + \left(m\ln\sqrt{\theta} + s + s_L + 1\right)x + m\right).\end{aligned}$$

Notice that since $0 < \theta < 1$, the discriminant of $\ln\sqrt{\theta}x^2 + \left(m\ln\sqrt{\theta} + s + s_L + 1\right)x + m$

$$\Delta = \left(m\ln\sqrt{\theta} + s + s_L + 1\right)^2 - 4m\ln\sqrt{\theta},$$

is positive but still, by the definition of M in (31), for any $x \geq M$ $\varphi_1'(x) \leq 0$ and so $\varphi_1(k) \leq \varphi_1(M)$ for all $k \geq M$. Then consider $\varphi_2(x) = \theta^{\sqrt{x}}x$.

$$\begin{aligned}\varphi_2'(x) &= \theta^{\sqrt{x}}\left(\frac{\ln\theta}{2\sqrt{x}}x + 1\right) \\ &= \frac{\theta^{\sqrt{x}}}{2}\left(\sqrt{x}\ln\theta + 2\right),\end{aligned}$$

so for $x \geq \frac{4}{(\ln\theta)^2}$, $\varphi_2'(x) \leq 0$ and so $\varphi_2(k) \leq \varphi_2(M)$ for all $k \geq M$. Finally we consider

$\varphi_3(x) = \dfrac{m+x}{m+x-\sqrt{x}-1}$ and then,

$$\begin{aligned}\varphi_3'(x) &= \frac{m+x-\sqrt{x}-1-(m+x)\left(1-\frac{1}{2\sqrt{x}}\right)}{(m+x-\sqrt{x}-1)^2} \\ &= -\frac{x+2\sqrt{x}-m}{2\sqrt{x}\,(m+x-\sqrt{x}-1)^2}\end{aligned}$$

so for $x \geq m$, $\varphi_3'(x) \leq 0$ and $\varphi_3(k) \leq \varphi_3(M)$ for all $k \geq M$. Putting all this together we get (33). So we can define

$$\chi = \chi(\theta, m, M, s, s_L) \stackrel{\text{def}}{=} \theta^{\frac{M}{2}} \frac{M}{2} \left(\frac{m+M}{m}\right)^{s+s_L} + \theta^{\sqrt{M}} \frac{M}{2} 2^{s+s_L} + \frac{1}{1-\theta} \left(\frac{m+M}{m+M-\sqrt{M}-1}\right)^{s+s_L}, \tag{34}$$

and state that for all $k \geq M$

$$|x_{m+k}|\,(m+k)^{s+s_L} \leq \frac{\eta \|y_I\|_s}{C_1} \left(\chi + \frac{\theta}{1-\theta}\right). \tag{35}$$

Finally, we need to bound the error made by using $\tilde{w}$ instead of $(w_I)_0$ for the definition (21) of A. Using (5) together with the sequence (u_l) introduced in the proof of Lemma 2.1, we get

$$\begin{aligned}|(w_I)_0 - \tilde{w}| &\leq \sum_{l=L}^{\infty} \frac{|\delta_0|^2}{|\delta_l|\,|\delta_{l+1}|} |c_1| \ldots |c_l|\ |a_2| \ldots |a_{l+1}| \\ &\leq \frac{|\delta_0|}{|\delta_1|} \sum_{l=L}^{\infty} \delta^{2l} \left(\frac{1}{u_1} \cdots \frac{1}{u_l}\right) \left(\frac{1}{u_2} \cdots \frac{1}{u_{l+1}}\right) \\ &\leq \frac{1}{|\mu_m|} \sum_{l=L}^{\infty} \theta^{2l} \\ &= \frac{\theta^{2L}}{|\mu_m|\,(1-\theta^2)}.\end{aligned} \tag{36}$$

3.2 Computation of the Y bounds

Now and for the rest of this paper we assume for the sake of clarity that the nonlinearity N of f in (1) is a polynomial of degree two. The generalization to a polynomial nonlinearity of higher degree requires only the use of the estimates developed in [5] to bound terms like

$$\left(x^1 * \ldots * x^p\right)_n$$

where $x^1, \ldots, x^p \in B_0(r)$. Moreover, as long as one is interested in problems with nonlinearities built from elementary functions of mathematical physics (powers, exponential, trigonometric functions, rational, Bessel, elliptic integrals, etc.), our method is applicable. Indeed, since these nonlinearities are themselves solutions of first or second order linear ODEs, they can be appended to the original problem of interest in order to obtain a strictly polynomial nonlinearity, albeit in a higher number of variables. This standard trick is explained in more details in [12].

The first step is to bound

$$|T(\bar{x}) - \bar{x}| = |Af(\bar{x})|,$$

where here and in the following, $|\cdot|$ applied to vectors or matrices must be understood component-wise. Note that since we suppose that f is at most quadratic and $\bar{x}$ is constructed such that $\bar{x}_k = 0$ for all $k \geq m$, we have that $(f(\bar{x}))_{m+k} = 0$ for all $k \geq m-1$. According to (21),

$$|(Af(\bar{x}))_F| \leq |A_m (f(\bar{x}))_F| + |\beta_{m-1}| \left|\left(U_I^{-1} L_I^{-1} (f(\bar{x}))_I\right)_0\right| \left|(A_m)_{c\, m-1}\right|,$$

so using (29) with $K = m-1$, we can set

$$Y_F = |A_m (f(\bar{x}))_F| + |\beta_{m-1}| \, \eta \left(\sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) \left|(A_m)_{c\, m-1}\right|.$$

Using (21) again,

$$|(Af(\bar{x}))_I| \leq |\lambda_m| \left(\left|(A_m)_{l\, m-1} f(\bar{x})_F\right| + \left|\beta_{m-1} (A_m)_{m-1,m-1} \left(U_I^{-1} L_I^{-1} f(\bar{x})_I\right)_0\right| \right) |w_I| + \left|U_I^{-1} L_I^{-1} f(\bar{x})_I\right|,$$

so using (28), (29) and (30) (again with $K = m-1$), we can set

$$\begin{aligned} Y_{m+k} = & \left(\left|(A_m)_{l\, m-1} f(\bar{x})_F\right| + \left|\beta_{m-1} (A_m)_{m-1,m-1}\right| \eta \left(\sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} \\ & + \eta \sum_{l=0}^{k} \theta^{k-l} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} + \eta \sum_{l=k+1}^{m-2} \theta^{l-k} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}, \quad \forall\ 0 \leq k \leq m-3 \end{aligned}$$

and

$$\begin{aligned} \tilde{Y}_{m+k} = & \left(\left|(A_m)_{l\, m-1} f(\bar{x})_F\right| + \left|\beta_{m-1} (A_m)_{m-1,m-1}\right| \eta \left(\sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} \\ & + \eta \theta^k \sum_{l=0}^{m-2} \frac{|f(\bar{x})|_{m+l}}{\theta^l |\mu_{m+l}|}, \quad \forall\ k \geq m-2. \end{aligned}$$

We then take an integer M such that

$$M \geq \max\left(m-2, \frac{-s}{\ln \theta} - m\right). \tag{37}$$

This yields that

$$\forall\ k \geq M, \quad \tilde{Y}_{m+k} \leq \tilde{Y}_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s}.$$

Therefore we can set

$$Y_{m+k} = \tilde{Y}_{m+k}, \quad \forall\ m-2 \leq k \leq M \quad \text{and} \quad Y_{m+k} = Y_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall\ k > M.$$

You will see in Section 3.4 the rationale behind this choice.

3.3 Computation of the Z bounds

For $y, z \in B_0(r)$, we need to bound

$$DT(\bar{x}+y)\, z = (I - ADf(\bar{x}+y))\, z = \left(I - AA^\dagger\right) z - A\left(Df(\bar{x}+y) - A^\dagger\right) z.$$

3.3.1 Estimates for $(I - AA^\dagger)z$

According to (6) and (21),

$$\begin{aligned}\left(AA^\dagger z\right)_F &= A_m \left(Dz_F + \begin{pmatrix} 0 \\ \vdots \\ \beta_{m-1} z_m \end{pmatrix} \right) - \beta_{m-1} \left(U_I^{-1} L_I^{-1} \left(T z_I + \lambda_m z_{m-1} \mathbf{e_1} \right) \right)_0 \left(A_m\right)_{c\, m-1} \\ &= A_m D z_F + \beta_{m-1} z_m \left(A_m\right)_{c\, m-1} - \beta_{m-1} \left(z_m + \lambda_m z_{m-1} \left(w_I\right)_0 \right) \left(A_m\right)_{c\, m-1} \\ &= A_m \tilde{K} z_F + \beta_{m-1} \lambda_m \left(\tilde{w} - \left(w_I\right)_0 \right) z_{m-1} \left(A_m\right)_{c\, m-1},\end{aligned}$$

and so

$$\left(\left(I - AA^\dagger \right) z \right)_F = \left(I - A_m \tilde{K} \right) z_F + \beta_{m-1} \lambda_m \left(\tilde{w} - \left(w_I\right)_0 \right) z_{m-1} \left(A_m\right)_{c\, m-1}.$$

Again using (6) and (21), we get

$$\begin{aligned}\left(AA^\dagger z\right)_I &= -\lambda_m \left(A_m\right)_{l\, m-1} \left(Dz_F + \begin{pmatrix} 0 \\ \vdots \\ \beta_{m-1} z_m \end{pmatrix} \right) w_I + \left(U_I^{-1} L_I^{-1} + \Lambda \right) \left(T z_I + \lambda_m z_{m-1} \mathbf{e_1} \right) \\ &= z_I + \lambda_m w_I \\ &\left(- \left(A_m\right)_{l\, m-1} D z_F - \beta_{m-1} \left(A_m\right)_{m-1,m-1} z_m + z_{m-1} + \beta_{m-1} \left(A_m\right)_{m-1,m-1} \left(z_I + \lambda_m z_{m-1} w_I \right)_0 \right) \\ &= z_I + \lambda_m \left(- \left(A_m\right)_{l\, m-1} D z_F + z_{m-1} + \beta_{m-1} \lambda_m \left(A_m\right)_{m-1,m-1} z_{m-1} \left(w_I\right)_0 \right) w_I \\ &= z_I + \lambda_m \left(z_{m-1} - \left(A_m\right)_{l\, m-1} \tilde{K} z_F + \beta_{m-1} \lambda_m \left(A_m\right)_{m-1,m-1} z_{m-1} \left(\tilde{w} - \left(w_I\right)_0 \right) \right) w_I \\ &= z_I + \lambda_m \left(\left(I - A_m \tilde{K} \right)_{l\, m-1} z_F + \beta_{m-1} \lambda_m \left(A_m\right)_{m-1,m-1} z_{m-1} \left(\tilde{w} - \left(w_I\right)_0 \right) \right) w_I\end{aligned}$$

and so

$$\left(\left(I - AA^\dagger \right) z \right)_I = -\lambda_m \left(\left(I - A_m K \right)_{l\, m-1} z_F + \beta_{m-1} \lambda_m \left(A_m\right)_{m-1,m-1} z_{m-1} \left(\tilde{w} - \left(w_I\right)_0 \right) \right) w_I.$$

We introduce $W^s = \left(\dfrac{1}{\omega_0^s}, \ldots, \dfrac{1}{\omega_k^s}, \ldots \right)^T$. For $z \in B_0(r)$ we have, using (36),

$$\left| \left(I - AA^\dagger \right) z \right|_F \le \left(\left| I - A_m \tilde{K} \right| W_F^s + \frac{|\beta_{m-1}|\, |\lambda_m|\, \theta^{2L}}{|\mu_m|\, \omega_{m-1}^s (1-\theta^2)} \left| A_m \right|_{c\, m-1} \right) r$$

and, using also (28),

$$\left| \left(I - AA^\dagger \right) z \right|_{m+k} \le \left(\left| I - A_m \tilde{K} \right|_{l\, m-1} W_F^s + \frac{|\beta_{m-1}|\, |\lambda_m|\, \theta^{2L}}{|\mu_m|\, \omega_{m-1}^s (1-\theta^2)} \left| A_m \right|_{m-1,m-1} \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} r, \quad \forall\, k \ge 0.$$

As in Section 3.2, we then assume (37) and define Z^1 by

$$Z_F^1 = \left(\left| I - A_m \tilde{K} \right| W_F^s + \frac{|\beta_{m-1}|\, |\lambda_m|\, \theta^{2L}}{|\mu_m|\, \omega_{m-1}^s (1-\theta^2)} \left| A_m \right|_{c\, m-1} \right) r,$$

$$Z^1_{m+k} = \left(\left| I - A_m \tilde{K} \right|_{l\, m-1} W^s_F + \frac{|\beta_{m-1}|\,|\lambda_m|\,\theta^{2L}}{|\mu_m|\,\omega^s_{m-1}(1-\theta^2)}\,|A_m|_{m-1,m-1} \right) \eta\theta^k \frac{|\lambda_m|}{|\mu_m|} r, \quad \forall\, 0 \le k \le M,$$

and

$$Z^1_{m+k} = Z^1_{m+M} \frac{\omega^s_{m+M}}{\omega^s_{m+k}}, \quad \forall\ k > M.$$

By the definition of M, $\left|\left(I - AA^\dagger\right) z\right| \le Z^1$, for all $z \in B_0(r)$.

3.3.2 Estimates for $A\left(Df\left(\bar{x}+y\right) - A^\dagger\right) z$

We assumed that the nonlinear part N was polynomial of degree 2 so $Df\left(\bar{x}+y\right)$ can be written as a finite Taylor expansion:

$$Df\left(\bar{x}+y\right) = Df\left(\bar{x}\right) + D^2 f\left(\bar{x}\right)(y)$$

and

$$\left(Df\left(\bar{x}+y\right) - A^\dagger\right) z = \left(Df\left(\bar{x}\right) - A^\dagger\right) z + D^2 f\left(\bar{x}\right)(y,z).$$

If we denote by σ the coefficient of degree 2 of f, we have that $D^2 f\left(\bar{x}\right)(y,z) = 2\sigma(y * z)$. We then bound the convolution product using

Lemma 3.2. *Let $s \ge 2$, $x, y \in \Omega^s$, $K \ge 6$ and $L \ge 1$ computational parameters.*

$$\forall\ k \ge 0, \quad |(x*y)_k| \le \alpha^s_k(K) \frac{\|x\|_s \|y\|_s}{\omega^s_k},$$

where

$$\alpha^s_k(K) = \begin{cases} 1 + 2\displaystyle\sum_{l=1}^{L} \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}}, & k = 0 \\ 2 + 2\displaystyle\sum_{l=1}^{L} \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}} + \sum_{l=1}^{k-1} \frac{k^s}{l^s(k-l)^s}, & 1 \le k < K \\ 2 + 2\displaystyle\sum_{l=1}^{L} \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}} + 2\left(\frac{K}{K-1}\right)^s + \left(\frac{4\ln(K-2)}{K} + \frac{\pi^2-6}{3}\right)\left(\frac{2}{K} + \frac{1}{2}\right)^s, & k \ge K. \end{cases}$$

Proof. See [13] for a proof of this bound and [10] for a similar bound for $1 < s < 2$. □

It is important to notice here that $\alpha^s_k(K) = \alpha^s_K(K)$ for all $k \ge K$. For the rest of this paper we assume that m is taken ≥ 6 which will allow us to use Lemma 3.2 with $K = m$. For $y, z \in B_0(r)$, we get

$$\left|D^2 f\left(\bar{x}\right)(y,z)\right| \le 2\,|\sigma|\,\alpha^s(m) W^s r^2,$$

We define

$$C^2 = 2\,|\sigma|\,\alpha^s(m) W^s$$

so that for all $y, z \in B_0(r)$

$$\left|D^2 f\left(\bar{x}\right)(y,z)\right| \le C^2 r^2.$$

Now we focus on the order one term. According to the definition (6) of $A^\dagger$, we have

$$\left(\left(Df(\bar{x}) - A^\dagger\right) z\right)_F = (Df(\bar{x})z)_F - Df^{(m)}(\bar{x})z_F - \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ \beta_{m-1}\, z_m \end{pmatrix}$$
$$= 2\sigma \left((\bar{x} * z)_F - (\bar{x} * z_F)_F\right),$$

where in the convolution product, z_F must be understood as the infinite vector $(z_F, 0, \ldots, 0, \ldots)^T$. So

$$\left(\left(Df(\bar{x}) - A^\dagger\right) z\right)_0 = 0$$

and for all $z \in B_0(r)$,

$$\left|\left(Df(\bar{x}) - A^\dagger\right) z\right|_k \leq 2\,|\sigma|\, r \sum_{l=m-k}^{m-1} \frac{|\bar{x}_l|}{\omega^s_{k+l}}, \quad \forall \;\; 1 \leq k \leq m-1.$$

Then, remembering that $Df(\bar{x}) = \mathcal{L} + DN(\bar{x})$ and (6), we get that

$$\left(\left(Df(\bar{x}) - A^\dagger\right) z\right)_I = (DN(\bar{x})z)_I = 2\sigma\,(\bar{x} * z)_I,$$

so using Lemma 3.2, for all $z \in B_0(r)$,

$$\left|\left(Df(\bar{x}) - A^\dagger\right) z\right|_{m+k} \leq \frac{2\,|\sigma|\,\alpha^s_{m+k}(m)\,\|\bar{x}\|_s}{\omega^s_{m+k}} r, \quad \forall\; k \geq 0.$$

Putting both together we define

$$C^1_0(\bar{x}) = 0, \quad C^1_k(\bar{x}) = 2\,|\sigma| \sum_{l=m-k}^{m-1} \frac{|\bar{x}_l|}{\omega^s_{k+l}}, \;\forall\; 1 \leq k \leq m-1,$$

and

$$C^1_{m+k}(\bar{x}) = \frac{2\,|\sigma|\,\alpha^s_m(m)\,\|\bar{x}\|_s}{\omega^s_{m+k}}, \quad \forall\; k \geq 0,$$

to get that for all $z \in B_0(r)$

$$\left|\left(Df(\bar{x}) - A^\dagger\right) z\right| \leq C^1(\bar{x}) r.$$

Finally,

$$\left|A\left(Df\,(\bar{x}+y) - A^\dagger\right) z\right| \leq |A|\left(C^1(\bar{x})r + C^2 r^2\right),$$

and we are left to bound $|A|\,C^1(\bar{x})$ and $|A|\,C^2$.

According to (21),

$$\left(|A|\,C^1(\bar{x})\right)_F \leq |A_m|\,C^1_F(\bar{x}) + |\beta_{m-1}|\left|U_I^{-1}L_I^{-1}C^1_I(\bar{x})\right|_0 |A_m|_{c\,m-1},$$

and using (32)

$$\left|U_I^{-1}L_I^{-1}C^1_I(\bar{x})\right|_0 \leq \frac{\eta\|C^1_I(\bar{x})\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \leq \frac{2\eta\,|\sigma|\,\alpha^s_m(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}.$$

so we set

$$D_F^1(\bar{x}) = |A_m|\, C_F^1(\bar{x}) + \frac{2\,|\beta_{m-1}|\,\eta\,|\sigma|\,\alpha_m^s(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\,|A_m|_{c\,m-1}\,.$$

Still according to (21),

$$\begin{aligned}\left(|A|\,C^1(\bar{x})\right)_I &\le |\lambda_m|\,|A_m|_{l\,m-1}\,C_F^1(\bar{x})\,|w_I| + \left|U_I^{-1}L_I^{-1}C_I^1(\bar{x})\right| + |\lambda_m|\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\left|U_I^{-1}L_I^{-1}C_I^1(\bar{x})\right|_0|w_I| \\ &\le |\lambda_m|\left(|A_m|_{l\,m-1}\,C_F^1(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha_m^s(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)|w_I| + \left|U_I^{-1}L_I^{-1}C_I^1(\bar{x})\right|.\end{aligned}$$

Now we take some M as in (31) and use (28), (32) and (35) to set

$$\begin{aligned}D_{m+k}^1(\bar{x}) &= \left(|A_m|_{l\,m-1}\,C_F^1(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha_m^s(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^k \\ &\quad + \frac{2\eta\,|\sigma|\,\alpha_m^s(m)\|\bar{x}\|_s}{C_1\omega_{m+k}^{s+s_L}}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \frac{\theta}{1-\theta}\right), \quad \forall\ 0 \le k < M,\end{aligned}$$

$$\begin{aligned}D_{m+M}^1(\bar{x}) &= \left(|A_m|_{l\,m-1}\,C_F^1(\bar{x}) + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha_m^s(m)\|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^M \\ &\quad + \frac{2\eta\,|\sigma|\,\alpha_m^s(m)\|\bar{x}\|_s}{C_1\omega_{m+M}^{s+s_L}}\left(\chi + \frac{\theta}{1-\theta}\right),\end{aligned}$$

and

$$D_{m+k}^1(\bar{x}) = D_{m+M}^1(\bar{x})\frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall\ k > M$$

so that

$$|A|\,C^1(\bar{x}) \le D^1(\bar{x}).$$

Similarly, we set

$$D_F^2 = |A_m|\,C_F^2 + \frac{2\,|\beta_{m-1}|\,\eta\,|\sigma|\,\alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}}\,|A_m|_{c\,m-1}\,,$$

$$\begin{aligned}D_{m+k}^2 &= \left(|A_m|_{l\,m-1}\,C_F^2 + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^k \\ &\quad + \frac{2\eta\,|\sigma|\,\alpha_m^s(m)}{C_1\omega_{m+k}^{s+s_L}}\left(\sum_{l=0}^{k}\theta^{k-l}\left(\frac{m+k}{m+l}\right)^{s+s_L} + \frac{\theta}{1-\theta}\right), \quad \forall\ 0 \le k < M,\end{aligned}$$

$$D_{m+M}^2 = \left(|A_m|_{l\,m-1}\,C_F^2 + \frac{2\,|\beta_{m-1}|\,|A_m|_{m-1,m-1}\,\eta\,|\sigma|\,\alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}}\right)\eta\frac{|\lambda_m|}{|\mu_m|}\theta^M + \frac{2\eta\,|\sigma|\,\alpha_m^s(m)}{C_1\omega_{m+M}^{s+s_L}}\left(\chi + \frac{\theta}{1-\theta}\right),$$

and

$$D_{m+k}^2 = D_{m+M}^2\frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall\ k > M$$

so that

$$|A|\,C^2 \le D^2.$$

Finally we can set

$$Z^2 = D^1(\bar{x})r + D^2 r^2$$

and $Z = Z^1 + Z^2$ is such that, for all $y, z \in B_0(r)$

$$|DT\,(\bar{x}+y)\,z| \le Z.$$

3.4 Radii polynomials and interval arithmetic

We are now left to find a radius $r > 0$ such that for every $k \geq 0$, the *radii polynomials* $\{P_k(r)\}_k$ satisfy

$$P_k(r) \stackrel{\text{def}}{=} Y_k + Z_k(r) - \frac{r}{\omega_k^s} < 0.$$

Note that since we constructed Y and Z so that for every $k \geq M$

$$Y_{m+k} = Y_{m+M}\frac{\omega_{m+M}^s}{\omega_{m+k}^s} \quad \text{and} \quad Z_{m+k} = Z_{m+M}\frac{\omega_{m+M}^s}{\omega_{m+k}^s},$$

it is enough to find a $r > 0$ such that for all $0 \leq k \leq m + M$, $P_k(r) < 0$. To do so, we compute numerically for each $0 \leq k \leq m + M$

$$I_k \stackrel{\text{def}}{=} \{r > 0 \mid P_k(r) < 0\},$$

and

$$I \stackrel{\text{def}}{=} \bigcap_{k=0}^{m+M} I_k.$$

If I is empty then the proof fails, and we should try again with some larger parameters m and M. If I is non empty, we pick an $r \in I$ and check rigorously, using the interval arithmetic package INTLAB [14], that for all $0 \leq k \leq m+M$, $P_k(r) < 0$ which according to Theorem 3.1 proves that T defined in (23) is a contraction on $B_s(\bar{x}, r)$, yielding the existence of a unique solution of $f(x) = 0$ in $B_s(\bar{x}, r)$.

4 An example application

Equations of the following form

$$\begin{aligned} &-(2+\cos\xi)u''(\xi) + u(\xi) = -\sigma u(\xi)^2 + g(\xi) \\ &u'(0) = u'(\pi) = 0, \end{aligned} \tag{38}$$

where g is a 2π-periodic even smooth function, fall into the framework developed in Section 2. Indeed consider the cosine Fourier expansions of u and g

$$u(\xi) = \sum_{k\in\mathbb{Z}} x_k \cos(k\xi), \qquad g(\xi) = \sum_{k\in\mathbb{Z}} g_k \cos(k\xi).$$

Then (38) can be rewritten as $f(x) = 0$, where

$$f_0(x) \stackrel{\text{def}}{=} x_0 + x_1 + \sigma\,(x * x)_0 - g_0$$

and for all $k \geq 1$

$$f_k(x) \stackrel{\text{def}}{=} \frac{1}{2}(k-1)^2 x_{k-1} + (1+2k^2)x_k + \frac{1}{2}(k+1)^2 x_{k+1} + \sigma\,(x * x)_k - g_k. \tag{39}$$

Then we do have that the linear part of (39) is as in (3), given by

$$\mathcal{L}_k(x) = \lambda_k x_{k-1} + \mu_k x_k + \beta_k x_{k+1},$$

with

$$\mu_0 \stackrel{\text{def}}{=} 1, \quad \beta_0 \stackrel{\text{def}}{=} 1,$$

and for all $k \geq 1$

$$\lambda_k \stackrel{\text{def}}{=} \frac{1}{2}(k-1)^2, \quad \mu_k \stackrel{\text{def}}{=} (1+2k^2) \quad \text{and} \quad \beta_k \stackrel{\text{def}}{=} \frac{1}{2}(k+1)^2.$$

Let's fix some $m \geq 2$. With

$$C_1 = 2, \quad C_2 = 3 \quad \text{and} \quad \delta = \frac{1}{4}\frac{(m+1)^2}{m^2+\frac{1}{2}},$$

we have

$$\forall\ k \geq 1, \quad \left|\frac{\lambda_k}{k^2}\right|, \left|\frac{\mu_k}{k^2}\right|, \left|\frac{\beta_k}{k^2}\right| \leq C_2,$$

together with

$$\forall\ k \geq m, \quad C_1 \leq \left|\frac{\mu_k}{k^2}\right| \quad \text{and} \quad \left|\frac{\lambda_k}{\mu_k}\right|, \left|\frac{\beta_k}{\mu_k}\right| \leq \delta.$$

We now focus on the example where

$$g(\xi) \stackrel{\text{def}}{=} \frac{1}{2} + 3\cos(\xi) + \frac{1}{2}\cos(2\xi),$$

so that $u(\xi) = \cos(\xi)$ is a trivial solution for $\sigma = 0$. In the next section we are going to use rigorous computation to prove the existence of solutions for $\sigma \neq 0$ and compute those solutions.

4.1 Results

Starting from $\sigma = 0$ we first use standard pseudo-arclength continuation techniques to numerically get some non trivial approximate solutions for $\sigma \neq 0$. We computed 1250 different solutions (half for $\sigma > 0$ and the other half for $\sigma < 0$). See Figure 2 for a diagram summing up those computations, where each point represent a solution of (38).

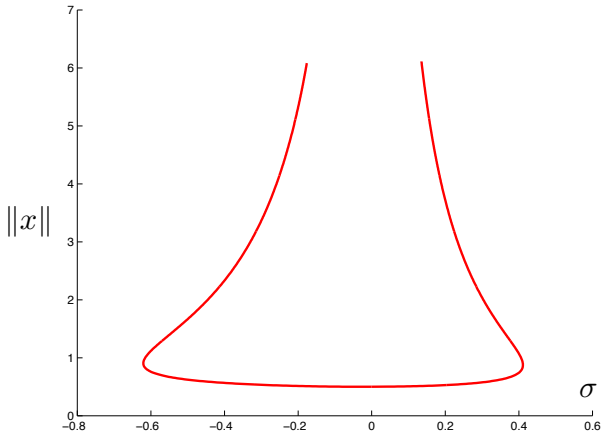


Figure 2: Branch of solutions of (38).

Then we use the rigorous computation method described in this paper to prove, for each numerical solution, the existence of a true solution in a small neighbourhood of the numerical

approximation. We keep $m = 20$ Fourier coefficient for the numerical computation and use $M = 20$ and the decay rate $s = 2$ for the proof. The bounds of Lemma 3.2 as well as the error on $\tilde{\omega}$ (36) are computed with $L = 100$. For each numerical solution the proof succeeds. More precisely, the I defined in Section 3.4 on which all radii polynomials should be negative always contains $[4 \times 10^{-11}, 10^{-4}]$ and we prove rigorously using interval arithmetic that indeed they all are negative for $r = 10^{-10}$. Hence the hypotheses of Theorem 3.1 hold and we have that within a ball of radius $r = 10^{-10}$ in Ω^s centered on the numerical approximation lays a unique solution to (38). Therefore the existence of the solutions represented Figure 2 is rigorously proven, within a margin of error that is to small to be depicted. The codes to perform the proofs can be found at [17].

Notice that existence of solutions of (38) could certainly have been obtained in different and more classical ways, for example using perturbative methods when σ is close to 0, or using a variational approach (that is, considering (38) as the Euler-Lagrange equation related to the critical points of a functional), or even using topological tools such as the Leray-Schauder theory. The advantage of our method is that it gives us more quantitative information than those approaches: indeed it enables to provide more than one solution for some values of σ, and, maybe more importantly, it gives a very precise localization of this (or these) solution(s) in terms of Fourier coefficients (something that looks very hard to obtain with qualitative PDEs methods).

5 Conclusion

A first interesting future direction of research would be to adapt our proposed approach to rigorously compute connecting orbits of ODEs using spectral methods. For instance, we would like to investigate the possibility of combining Hermite spectral methods with our approach to compute homoclinic orbits (e.g. see [15, 16]). Since the differential operator in frequency space of the Hermite functions is tridiagonal, adapting our approach to this class of operator could lead to a new rigorous numerical method for connecting orbits.

It would also be interesting to adapt our method to the case of looking for solutions in the sequence space

$$\ell^1_\nu = \{x = (x_k)_{k\geq 0} : \|x\|_\nu \stackrel{\text{def}}{=} \sum_{k\geq 0} |x_k|\nu^k < \infty\}$$

for some $\nu \geq 1$. With this choice of Banach space, we could exploit the fact that ℓ^1_ν is naturally a Banach algebra under discrete convolutions. This could greatly simplify the nonlinear analysis.

Note that assumption (5) requires the tridiagonal operator to have symmetric ratios between the diagonal terms and the upper and lower diagonal terms. This is a restriction that we hope could be relaxed. Since many interesting problems involve tridiagonal operators with non symmetric ratios (as in the case of differentiation in frequency space of the Hermite functions), we believe that this is a promising route to follow.

Finally, generalizing our approach to problems with block-tridiagonal structures could also be a valuable project.

Acknowledgement

The research leading to this paper was partially funded by the french "ANR blanche" project Kibord: ANR-13-BS01-0004.

References

[1] John P. Boyd. *Chebyshev and Fourier spectral methods.* Dover Publications Inc., Mineola, NY, second edition, 2001.

[2] Allan Hungria, Jean-Philippe Lessard, and Jason D. Mireles-James. Radii polynomial approach for analytic solutions of differential equations: Theory, examples, and comparisons. Submitted, 2014.

[3] Piotr Zgliczyński and Konstantin Mischaikow. Rigorous numerics for partial differential equations: the Kuramoto-Sivashinsky equation. *Found. Comput. Math.*, 1(3):255–288, 2001.

[4] Yasuaki Hiraoka and Toshiyuki Ogawa. Rigorous numerics for localized patterns to the quintic Swift-Hohenberg equation. *Japan J. Indust. Appl. Math.*, 22(1):57–75, 2005.

[5] Marcio Gameiro and Jean-Philippe Lessard. Analytic estimates and rigorous continuation for equilibria of higher-dimensional PDEs. *J. Differential Equations*, 249(9):2237–2268, 2010.

[6] Gábor Kiss and Jean-Philippe Lessard. Computational fixed-point theory for differential delay equations with multiple time lags. *J. Differential Equations*, 252(4):3093–3115, 2012.

[7] S. Day, O. Junge, and K. Mischaikow. A rigorous numerical method for the global analysis of infinite-dimensional discrete dynamical systems. *SIAM J. Appl. Dyn. Syst.*, 3(2):117–160 (electronic), 2004.

[8] Anthony W. Baker, Michael Dellnitz, and Oliver Junge. A topological method for rigorously computing periodic orbits using Fourier modes. *Discrete Contin. Dyn. Syst.*, 13(4):901–920, 2005.

[9] Roberto Castelli and Jean-Philippe Lessard. Rigorous Numerics in Floquet Theory: Computing Stable and Unstable Bundles of Periodic Orbits. *SIAM J. Appl. Dyn. Syst.*, 12(1):204–245, 2013.

[10] Maxime Breden, Jean-Philippe Lessard, and Matthieu Vanicat. Global Bifurcation Diagrams of Steady States of Systems of PDEs via Rigorous Numerics: a 3-Component Reaction-Diffusion System. *Acta Appl. Math.*, 128:113–152, 2013.

[11] Philippe G. Ciarlet. *Introduction to numerical linear algebra and optimisation.* Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1989. With the assistance of Bernadette Miara and Jean-Marie Thomas, Translated from the French by A. Buttigieg.

[12] Donald E. Knuth. *The art of computer programming. Vol. 2.* Addison-Wesley Publishing Co., Reading, Mass., second edition, 1981. Seminumerical algorithms, Addison-Wesley Series in Computer Science and Information Processing.

[13] Marcio Gameiro and Jean-Philippe Lessard. Efficient Rigorous Numerics for Higher-Dimensional PDEs via One-Dimensional Estimates. *SIAM J. Numer. Anal.*, 51(4):2063–2087, 2013.

[14] S.M. Rump. INTLAB - INTerval LABoratory. In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, Dordrecht, 1999. http://www.ti3.tu-harburg.de/rump/.

[15] Valeriy R. Korostyshevskiy and Thomas Wanner. A Hermite spectral method for the computation of homoclinic orbits and associated functionals. *J. Comput. Appl. Math.*, 206(2):986–1006, 2007.

[16] Valeriy R. Korostyshevskiy. *A Hermite spectral approach to homoclinic solutions of ordinary differential equations*. ProQuest LLC, Ann Arbor, MI, 2005. Thesis (Ph.D.)–University of Maryland, Baltimore County.

[17] M. Breden, L. Desvillettes and J.-P. Lessard. MATLAB codes to perform the proofs. http://archimede.mat.ulaval.ca/jplessard/PseudoInverse