



# A Newton-like Validation Method for Chebyshev Approximate Solutions of Linear Ordinary Differential Systems

Florent Bréhard

## ► To cite this version:

Florent Bréhard. A Newton-like Validation Method for Chebyshev Approximate Solutions of Linear Ordinary Differential Systems. ISSAC 2018 - 43rd International Symposium on Symbolic and Algebraic Computation, Jul 2018, New York, United States. pp.103-110, 10.1145/3208976.3209000 . hal-01654396v2

**HAL Id: hal-01654396**

**<https://hal.science/hal-01654396v2>**

Submitted on 23 Jul 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A Newton-like Validation Method for Chebyshev Approximate Solutions of Linear Ordinary Differential Systems

Florent Bréhard<sup>1</sup>

<sup>1</sup>École Normale Supérieure de Lyon and LAAS-CNRS, France

February 5, 2018

## Abstract

We provide a new framework for *a posteriori* validation of vector-valued problems with componentwise tight error enclosures, and use it to design a symbolic-numeric Newton-like validation algorithm for Chebyshev approximate solutions of coupled systems of linear ordinary differential equations. More precisely, given a coupled differential system with polynomial coefficients over a compact interval (or continuous coefficients rigorously approximated by polynomials) and componentwise polynomial approximate solutions in Chebyshev basis, the algorithm outputs componentwise rigorous upper bounds for the approximation errors, with respect to the uniform norm over the interval under consideration.

A complexity analysis shows that the number of arithmetic operations needed by this algorithm (in floating-point or interval arithmetics) is proportional to the approximation degree when the differential equation is considered fixed. Finally, we illustrate the efficiency of this fully automated validation method on an example of a coupled Airy-like system.

## 1 Introduction

**Notations.** Let  $p$  be a positive integer for the ambient space  $\mathbb{R}^p$ , whose canonical basis is denoted by  $(e_1, \dots, e_p)$ . For a ring  $\mathbb{A}$ ,  $\mathcal{M}_p(\mathbb{A})$  denotes the set of order  $p$  square matrices, with  $\mathbf{1}$  and  $\mathbf{0}$  the identity and zero matrices.

The order  $\leq$  over  $\mathbb{R}$  is componentwise extended to a (partial) order over  $\mathbb{R}^p$  and  $\mathcal{M}_p(\mathbb{R})$ : for all  $u, v \in \mathbb{R}^p$  (resp.  $A, B$  in  $\mathcal{M}_p(\mathbb{R})$ ),  $u \leq v$  if and only if  $u_i \leq v_i$  for all  $i \in \llbracket 1, p \rrbracket$  (resp.  $A \leq B$  iff  $A_{ij} \leq B_{ij}$  for all  $i, j \in \llbracket 1, p \rrbracket$ ). The uniform norm of a function  $f$ , defined over  $[-1, 1]$ , is  $\|f\|_\infty = \sup_{x \in [-1, 1]} |f(x)|$ .

**General context.** Numerical computing with functions is often done via polynomial approximations. For sufficiently smooth functions, defined over a compact interval, truncated Chebyshev series are well-known for their excellent approximation properties, efficient algorithms and software like Chebfun [6, 21]. Broadly speaking, the principle of working with polynomial approximations instead of functions is analogous to using floating-point instead of real numbers. However, some applications such as safety-critical engineering or computer-assisted proofs in mathematics need *effective* and *safe* (rather than *asymptotic*) error enclosures.

**Problem statement and contributions.** We present a symbolic-numeric *a posteriori* validation algorithm that provides *componentwise* and *tight* error enclosures for Chebyshev approximations to solutions of coupled linear ordinary differential equations (LODEs):

$$Y^{(r)} + A_{r-1}(t) \cdot Y^{(r-1)} + \dots + A_1(t) \cdot Y' + A_0(t) \cdot Y = G(t), \quad (1)$$

of unknown  $Y : [-1, 1] \rightarrow \mathbb{R}^p$ . Coefficients  $A_i$  and  $G$  must be continuous functions, given as polynomials with rigorous error bounds. However, for the sake of simplicity, we mainly focus on the polynomial case, and refer to the solutions as *vector-valued D-finite functions*. Although such functions can be seen as vectors of (scalar) D-finite functions, the decoupling of the system followed by a possible desingularization step may produce hard to validate scalar LODEs (see Section 4). Moreover, in the nonpolynomial case, such techniques do not apply.

Using an appropriate integral transform of the linear differential system, we obtain a Volterra integral equation of the second kind with polynomial kernel, whence the following problem statement:

**Problem 1.** *For a given integral equation of unknown  $\Phi : [-1, 1] \rightarrow \mathbb{R}^p$ :*

$$\Phi(t) + \int_{-1}^t K(t, s) \cdot \Phi(s) ds = \Psi(t),$$

*with a  $p$ -dimensional polynomial kernel  $K(t, s) \in \mathcal{M}_p(\mathbb{R}[t, s])$  and  $\Psi \in \mathbb{R}[t]^p$ , assuming we are given for each component  $\Phi_i^\star$  of the exact solution  $\Phi^\star$  a polynomial approximation  $\Phi_i^\circ$  in Chebyshev basis, compute componentwise error bounds  $\varepsilon_i$ , as tight as desired:*

$$\|\Phi_i^\circ - \Phi_i^\star\|_{\mathbb{Q}^1} \leq \varepsilon_i, \quad \text{for all } i \in \llbracket 1, p \rrbracket.$$

Here,  $\|\cdot\|_{\mathcal{C}_1}$  is a norm for absolutely summable Chebyshev series that upper-bounds the  $\|\cdot\|_{\infty}$  norm over  $[-1, 1]$  (see Section 3.1).

Fixed-point methods are extensively used in the field of functional analysis and differential equations. They provide iterative approximation schemes, like Picard-Chebyshev which integrates nonlinear dynamical systems arising, for instance, in space flight mechanics problems [8, 4]. They also underlie numerous validation methods for function space problems [13, 23].

Many fixed-point validation methods use the Banach fixed-point theorem. Given an equation  $x = \mathbf{T} \cdot x$  with  $\mathbf{T}$  contracting of ratio  $\lambda \in (0, 1)$  over a complete metric space, and an approximation  $x$  to the exact solution  $x^*$ , it provides an error enclosure:

$$\frac{\|x - \mathbf{T} \cdot x\|}{1 + \lambda} \leq \|x - x^*\| \leq \frac{\|x - \mathbf{T} \cdot x\|}{1 - \lambda}. \quad (2)$$

However, in the case we consider,  $x$  belongs to a product space, and the classical method consisting in endowing it with a global norm fails to produce componentwise tight error enclosures. This is particularly annoying when the components of the system are of different nature (e.g., position and speed) or magnitude.

Based on a new refinement with lower bounds for the Perov fixed-point theorem (a vector-valued generalization of Banach fixed-point principle), we propose a validation algorithm to solve Problem 1. It is a generalization of the validation method presented in [7] to vectorial LODEs, within a new general framework for vector-valued fixed-point validation.

**Theorem 1.** *Algorithms 1 and 3 solve together Problem 1 by providing componentwise error enclosures, as tight as desired.*

(i) *Algorithm 1 only depends on the integral equation (not on the provided approximation). It produces and rigorously bounds a Newton-like validation operator and requires  $O(p^3 N_{\text{val}}^2 d)$  arithmetic operations.*

(ii) *Algorithm 3 computes the error enclosures for the approximation and runs in linear time with respect to the maximum degree of the approximations  $\Phi_i^\circ$  and the right-hand sides  $\Psi_i$ . More precisely, its complexity is  $O(p^2 d^2 N_{\text{app}} + p N_{\text{rhs}} + p^2 N_{\text{val}} \min(\max(N_{\text{app}} + d, N_{\text{rhs}}), N_{\text{val}}))$ , where:*

- $N_{\text{app}} = \max_i \deg \Phi_i^\circ$  and  $N_{\text{rhs}} = \max_i \deg \Psi_i$ ;
- $d = 1 + \max_{ij} \deg k_{ij}(t, s)$ ;
- $N_{\text{val}}$  is a truncation index used to rigorously approximate the problem in finite dimension.

We assume a uniform complexity model, i.e., a unit cost for each arithmetic operation  $(+, -, \times, /, \sqrt{\phantom{x}})$ , with, say, floating-point or interval operands.

The previous complexity estimates still involve a truncation index  $N_{\text{val}}$ , which is directly related to how tight the desired error enclosures have to be. As detailed in Theorem 4, its minimal value ensuring a contracting Newton-like operator is potentially exponential with respect to the magnitude of the coefficients of the integral equation, in the case of stiff LODEs for example. In practice however, this method works efficiently and fully automatically. An open source library implementing this validation method (and its extension to the nonpolynomial case) can be found here <sup>1</sup>. It was also recently used for a space flight dynamics application [3].

**Previous work.** In this context, applications of the Banach fixed-point theorem include early works [13, 23], where variations of Newton’s method perform *a posteriori* validation in function spaces. More recent works developed techniques (e.g., *radii polynomials* [10]) to find a stable neighborhood of an approximation  $\varphi$  over which the Banach fixed-point theorem applies. They have the advantage of dealing with nonlinear problems (examples can be found in [14, 22, 10]). However, the above mentioned methods were not fully automated and little emphasis was put on their algorithmic aspects.

By contrast, [5] is a pioneer work towards effective methods for validation of approximations of D-finite functions in Chebyshev basis. At the cost of a more restricted class of functions, namely, D-finite functions, this article introduces a fully automated algorithm together with complexity estimates, based on a Picard iteration scheme. In line with this work, [7] describes another algorithm based on a Newton-like method in an appropriate function space, which is easily extended to the case of continuous coefficients rigorously approximated by Chebyshev polynomials.

The above mentioned validation techniques are usually transposed to the vectorial case by fixing a norm over the vector-valued function space. However, this does not provide componentwise tight error enclosures. To overcome this limitation, we consider the notion of *vector-valued* (or *generalized*) metric spaces and *generalized contractions* (or *P-contractions*) [11, 20, 17]. The Perov fixed-point theorem [11, 18] is a natural extension of the Banach fixed-point theorem and provides componentwise upper bounds for the approximation error. Several works applied this theorem in various settings, for example [24] for the Newton method or [2, 19, 15] for ODEs with nonlocal conditions. To the best of our knowledge, however, none of these works investigate the existence of lower bounds, nor address validation problems.

**Outline.** Section 2 introduces a general framework for componentwise fixed-

---

<sup>1</sup><http://perso.ens-lyon.fr/florent.brehard>

point validation in generalized metric spaces. In Section 3, we design the Newton-like validation algorithm for Chebyshev approximations of vector-valued D-finite functions. Finally, Section 4 details the validation of a two-dimensional highly oscillating system. For completeness, we also provide a comparison with a decoupling technique that boils down to solving scalar LODEs.

## 2 A Framework for Vector-Valued Validation Problems

We address the general problem of componentwise validating an approximation  $x^\circ$  to the exact solution  $x^\star$  of a fixed-point equation  $x = \mathbf{T} \cdot x$ . Section 2.1 gives a rigorous definition of “several components and norms” with the notion of generalized metric spaces, leading to the Perov fixed-point theorem. Section 2.2 presents a new result that complements the Perov theorem with lower bounds on the componentwise approximation errors.

**A toy example in the plane** illustrates the vector-valued validation framework. Consider the trigonometric equation  $\sin^3 \vartheta + \cos 3\vartheta = 0$  for  $\vartheta \in \mathbb{R}$ . By introducing  $c = \cos x$  and  $s = \sin x$ , this is equivalent to finding the roots of the following polynomial system in the plane  $(c, s)$ :

$$\mathbf{F} \cdot (c, s) = \begin{pmatrix} s^3 + 4c^3 - 3c \\ c^2 + s^2 - 1 \end{pmatrix} = 0. \quad (3)$$

Let  $x^\star = (c^\star, s^\star)$  be an exact solution and  $x^\circ := (c^\circ, s^\circ) = (0.84, 0.55)$  an approximation of it. In order to validate this solution with respect to a given norm  $\|\cdot\|$  on  $\mathbb{R}^2$ , we define a Newton-like operator  $\mathbf{T} \cdot (c, s) = (c, s) - \mathbf{A} \cdot \mathbf{F} \cdot (c, s)$  with  $\mathbf{A} := \begin{pmatrix} 0.25 & -0.20 \\ -0.37 & 1.2 \end{pmatrix} \approx (\mathbf{D}\mathbf{F}_{x^\circ})^{-1} \in \mathcal{M}_2(\mathbb{R})$  an approximate inverse of the Fréchet derivative  $\mathbf{D}\mathbf{F}_{x^\circ}$  of  $\mathbf{F}$  at  $x^\circ$ . Since  $\mathbf{A}$  is injective, its fixed points are exactly the roots of  $\mathbf{F}$ . In this example,  $\mathbf{F}$  is nonlinear, so one must find a *stable closed neighborhood* over which  $\mathbf{T}$  is contracting, for the Banach theorem to apply. It suffices to determine a radius  $r > 0$  satisfying the following two conditions:

- (i)  $\lambda := \sup_{\|x - x^\circ\| \leq r} \|\mathbf{1} - \mathbf{A} \cdot \mathbf{D}\mathbf{F}_x\| < 1$ ;
- (ii)  $\|x^\circ - \mathbf{T} \cdot x^\circ\| + kr \leq r$ .

If such a radius exists, then by the Banach fixed-point theorem, we have  $\|x^\circ - x^\star\| \leq \|\mathbf{A} \cdot \mathbf{F} \cdot x\| / (1 - \lambda)$ . However, such a bound captures a “global” error, which may not be what we expect, if, for example, the two components are of different nature (e.g., position and velocity), or differ by several orders of magnitude.

## 2.1 Generalized Metric Spaces and Perov Fixed-Point Theorem

**Definition 1.** Let  $X$  be a set (resp.  $E$  a linear space). A function  $d : X \times X \rightarrow \mathbb{R}_+^p$  (resp.  $\|\cdot\| : E \rightarrow \mathbb{R}_+^p$ ) is a vector-valued or generalized metric (resp. norm) if for all  $x, y, z$  in  $X$  or  $E$  and  $\lambda \in \mathbb{R}$ :

- $d(x, y) = 0$  iff  $x = y$ ,                      resp.  $\|x\| = 0$  iff  $x = 0$ ;
- $d(x, y) = d(y, x)$ ,                              resp.  $\|\lambda x\| = |\lambda| \|x\|$ ;
- $d(x, y) \leq d(x, z) + d(z, y)$ ,              resp.  $\|x + y\| \leq \|x\| + \|y\|$ .

Then  $(X, d)$  (resp.  $(E, \|\cdot\|)$ ) is a vector-valued or generalized metric space (resp. linear space).

A straightforward example is the product of  $p$  metric spaces  $(X_i, d_i)$ ,  $i \in \llbracket 1, p \rrbracket$  (resp.  $p$  normed linear spaces  $(E, \|\cdot\|_i)$ ) and the vector-valued metric  $d(x, y) = (d_1(x_1, y_1), \dots, d_p(x_p, y_p))$  (resp. the vector-valued norm  $\|x\| = (\|x_1\|_1, \dots, \|x_p\|_p)$ ).

**Remark 1.** A vector-valued metric space (respectively a vector-valued normed linear space) can be trivially seen as a metric space (respectively a normed linear space) by taking the maximum of all the components of the vector-valued metric (respectively norm). We therefore recover all the useful topological notions of convergence, limit, neighborhood, completeness, etc.

In the context of vector-valued metric spaces, the notion of contracting map needs to be generalized. Let  $\mathcal{M}_p^{\rightarrow 0}(\mathbb{R}) \subseteq \mathcal{M}_p(\mathbb{R})$  denote the *convergent to zero matrices*, that is the matrices  $M$  such that  $M^k \rightarrow 0$  as  $k \rightarrow \infty$ . Equivalently, these are matrices  $M$  with spectral radius  $\rho(M) < 1$ . Then,  $\mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+) = \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}) \cap \mathcal{M}_p(\mathbb{R}_+)$  denotes those among them with nonnegative coefficients.

**Definition 2.** Let  $(X, d)$  be a vector-valued metric space and  $\mathbf{T} : X \rightarrow X$  an operator.

- $\mathbf{T}$  is  $\Lambda$ -Lipschitz for some  $\Lambda \in \mathcal{M}_p(\mathbb{R}_+)$  if:

$$d(\mathbf{T} \cdot x, \mathbf{T} \cdot y) \leq \Lambda \cdot d(x, y), \quad \text{for all } x, y \in X.$$

- If moreover  $\Lambda$  is convergent to 0 ( $\Lambda \in \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+)$ ), then  $\mathbf{T}$  is said to be a generalized contraction.

Using these definitions, the Perov fixed-point theorem<sup>2</sup> is a generalization of the Banach fixed-point theorem.

---

<sup>2</sup>Although commonly attributed to Perov [18] (in Russian), the idea of generalizing the Banach fixed-point theorem to generalized norms for investigating the componentwise errors in an iterative process first appeared in Kantorovich's work [11] (in Russian).

**Theorem 2** (Perov). *Let  $(X, d)$  be a complete vector-valued metric space and  $\mathbf{T} : X \rightarrow X$  a generalized contraction with a Lipschitz matrix  $\Lambda \in \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+)$ . Then:*

- (i)  $\mathbf{T}$  admits a unique fixed-point  $x^* \in X$ ;
- (ii) for every  $x^\circ \in X$ , the iterated sequence defined by  $x_0 = x^\circ$  and  $x_{n+1} = \mathbf{T} \cdot x_n$  converges to  $x^*$  with the following upper bound on the approximation error:

$$d(x_n, x^*) \leq \Lambda^n \cdot (\mathbf{1} - \Lambda)^{-1} \cdot d(x^\circ, \mathbf{T} \cdot x^\circ), \quad \text{for all } n \in \mathbb{N}. \quad (4)$$

A proof of this theorem is given in Appendix A.1 or [17].

**Perov theorem applied to the toy example.** Endowing  $\mathbb{R}^2$  with the vector-valued norm  $\|(c, s)\| := (|c|, |s|)$  does not change the definition of  $\mathbf{T}$ . The two conditions needed to apply the Banach fixed-point theorem are adapted to the Perov theorem as follows. Choose a *multi-radius*  $r = (r_1, r_2)$  such that

- (i)  $\Lambda := \left( \sup_{\|x - x^\circ\| \leq r} |(D\mathbf{T}_x)_{ij}| \right)_{1 \leq i, j \leq 2}$  satisfies  $\rho(\Lambda) < 1$ ;
- (ii)  $\|x^\circ - \mathbf{T} \cdot x^\circ\| + \Lambda \cdot r \leq r$ .

For  $r = (0.005, 0.005)$ , one obtains:

$$\Lambda = \begin{pmatrix} 5.81 & 1.31 \\ 5.63 & 3.40 \end{pmatrix} \cdot 10^{-2}, \quad \rho(\Lambda) = 7.57 \cdot 10^{-2},$$

which satisfies (i) and (ii). Hence, Theorem 2 gives:

$$|c^\circ - c^*| \leq 2.90 \cdot 10^{-3}, \quad |s^\circ - s^*| \leq 3.65 \cdot 10^{-3}.$$

To assess the tightness of these bounds, we provide lower bounds on the componentwise approximation errors.

## 2.2 Lower Bounds and Error Enclosures

Let  $\varepsilon = d(x^\circ, x^*) \in \mathbb{R}_+^p$  be the vector of unknown errors and  $\eta = d(x^\circ, \mathbf{T} \cdot x^\circ) \in \mathbb{R}_+^p$ . By the triangle inequality,  $\varepsilon$  is circumscribed into a polytope of  $\mathbb{R}_+^p$ :

$$\begin{aligned} (\mathbf{1} - \Lambda) \cdot \varepsilon &\leq \eta, \\ (\mathbf{1} + \Lambda) \cdot \varepsilon &\geq \eta, \\ \varepsilon &\geq 0. \end{aligned} \quad (5)$$



The first inequality gives the upper bounds  $\varepsilon^+ = (\mathbf{1} - \Lambda)^{-1} \cdot \eta$ , as stated by Theorem 2 (with  $n = 0$ ). However, the second one does not directly give the desired lower bounds, say  $\varepsilon^-$ , because the inverse  $(\mathbf{1} + \Lambda)^{-1} = \sum_{k \geq 0} (-\Lambda)^k$  is not nonnegative in general. It is clear that each  $\varepsilon_i^-$  is given by the  $i$ -th coordinate of some vertex of this polytope. Instead of testing its  $2^p$  vertices, the following theorem identifies the correct one.

**Theorem 3** (Lower bounds for the Perov theorem). *With the above notations, for each  $i \in \llbracket 1, p \rrbracket$ , the lower bound  $\varepsilon_i^-$  on the  $i$ -th component  $\varepsilon_i$  of the approximation error of  $x^\circ$  to  $x^\star$  is given by the  $i$ -th component of the vertex defined by the intersection of the  $i$ -th lower-bound constraint together with all the  $j$ -th upper-bound constraints with  $j \neq i$  from (5). Formally:*

$$\varepsilon_i \geq \varepsilon_i^- \quad \text{with} \quad \varepsilon_i^- = e_i^T \cdot (\mathbf{1} - D_i \cdot \Lambda)^{-1} \cdot \eta,$$

where  $D_i$  is the order  $p$  diagonal matrix defined by  $(D_i)_{ii} = -1$  and  $(D_i)_{jj} = 1$  for  $j \neq i$ .

**Remark 2.** *Contrary to the one-dimensional case,  $\varepsilon_i^-$  may be negative (we then round it to 0): the overestimation factor of  $\varepsilon_i^+$  provided by Theorem 2 is not controlled. A tighter enclosure can be obtained with a more contracting  $\mathbf{T}$  (see Appendix A.2).*

*Proof of Theorem 3.* Among the Inequalities (5), take the  $p$  upper-bound constraints and replace the  $i$ -th one by the corresponding lower-bound constraint. Multiply these  $p - 1$  upper-bound constraints by  $-1$  to obtain the following system of inequalities:

$$(\Lambda - D_i) \cdot \varepsilon \geq -D_i \cdot \eta. \tag{6}$$

From Lemma 1,  $\Lambda - D_i$  is nonsingular and its inverse has nonnegative coefficients on its  $i$ -th row. Hence we can multiply (6) by  $(\Lambda - D_i)^{-1}$  and only keep the resulting  $i$ -th constraint:

$$\begin{aligned} \varepsilon_i &= e_i^T \cdot (\Lambda - D_i)^{-1} \cdot (\Lambda - D_i) \cdot \varepsilon \\ &\geq e_i^T \cdot (\Lambda - D_i)^{-1} \cdot (-D_i) \cdot \eta = e_i^T \cdot (\mathbf{1} - D_i \cdot \Lambda)^{-1} \cdot \eta. \end{aligned}$$

**Lower bounds for the toy example.** The polytope given by the linear constraints (5) is depicted in Figure 1. The top right vertex corresponds to  $(\varepsilon_1^+, \varepsilon_2^+)$ . Also, the  $\varepsilon_1^-$  (resp.  $\varepsilon_2^-$ ) is given by the top left (resp. bottom right) vertex, which is consistent with Theorem 3. This gives the following numerical enclosures:

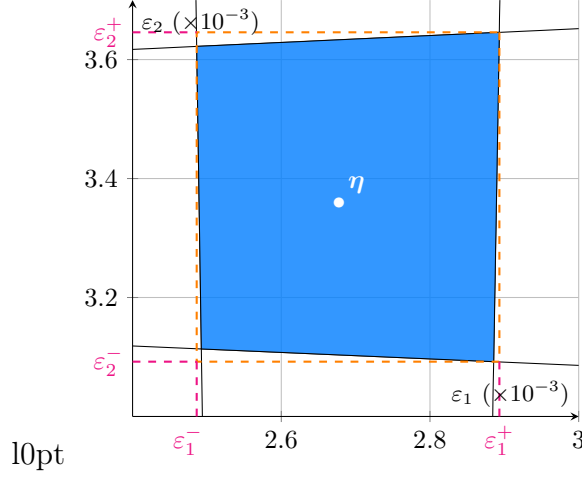


Figure 1: Error polytope for the toy example.

$$\begin{aligned} \varepsilon_1^- = 2.48 \cdot 10^{-3} \leq \varepsilon_1 = |c^\circ - c^\star| \leq \varepsilon_1^+ = 2.90 \cdot 10^{-3}, \\ \varepsilon_2^- = 3.09 \cdot 10^{-3} \leq \varepsilon_2 = |s^\circ - s^\star| \leq \varepsilon_2^+ = 3.65 \cdot 10^{-3}. \end{aligned} \quad (7)$$

The tightness of these enclosures is discussed in Appendix A.2. Roughly speaking, the ratio  $\varepsilon_i^+/\varepsilon_i^-$  depends not only on  $\Lambda$  (like in the univariate case), but also on  $\eta = d(x^\circ, \mathbf{T} \cdot x^\circ)$ .

### 3 Componentwise Validation of Chebyshev Approximations

We present the validation method to solve Problem 1. Section 3.1 contains reminders about Chebyshev approximation theory and LODEs. This leads to an efficient approximating procedure (Section 3.2). Section 3.3 presents **Algorithms 1** and **2** to create and bound a Newton-like operator associated to a given vectorial LODE, then **Algorithm 3** to compute componentwise error enclosures for any Chebyshev approximation  $(\Phi_i^\circ)_{1 \leq i \leq p}$ .

#### 3.1 Reminders on Chebyshev Approximations

**Chebyshev series and  $\mathcal{V}^1$  space.** The Chebyshev family of polynomials is defined by the three-term recurrence  $T_{n+2} = 2XT_{n+1} - T_n$  with initial terms  $T_0 = 1$  and  $T_1 = X$ . They satisfy the fundamental trigonometric relation  $T_n(\cos \vartheta) = \cos(n\vartheta)$ , from which we deduce some of their basic algebraic

properties:

$$T_n T_m = \frac{1}{2}(T_{n+m} + T_{|n-m|}), \quad \int T_n = \frac{T_{n+1}}{2(n+1)} - \frac{T_{n-1}}{2(n-1)} = T_n \quad (n \geq 2), \quad (8)$$

and that  $|T_n(t)| \leq 1$  for  $x \in [-1, 1]$ .

Let  $L^2_{\mathbb{Q}} = L^2(1/\sqrt{1-t^2})$  denote the space of real-valued measurable functions  $f$  over  $[-1, 1]$  such that  $\int_{-1}^1 f(t)^2/\sqrt{1-t^2} dt < \infty$ . The inner product

$$\langle f, g \rangle = \int_{-1}^1 f(t)g(t)/\sqrt{1-t^2} dt = \int_0^\pi f(\cos \vartheta)g(\cos \vartheta) d\vartheta,$$

defines a Hilbert space structure over  $L^2_{\mathbb{Q}}$ , for which the Chebyshev polynomials form a complete orthogonal system. To any continuous function  $f$  in this space we can associate its Chebyshev coefficients:

$$[f]_n = \begin{cases} \frac{1}{\pi} \int_0^\pi f(\cos \vartheta) d\vartheta, & \text{if } n = 0, \\ \frac{2}{\pi} \int_0^\pi f(\cos \vartheta) \cos(n\vartheta) d\vartheta, & \text{if } n > 0. \end{cases}$$

Hence, the truncated Chebyshev series  $f^{[N]} = \Pi_N \cdot f := \sum_{n=0}^N [f]_n T_n$  of  $f$  is the orthogonal projection of  $f$  onto the finite-dimensional subspace spanned by  $T_0, \dots, T_N$ . In addition to the  $L^2_{\mathbb{Q}}$  convergence, and analogously to Fourier series, Chebyshev series have excellent approximation properties [6]. For example, if  $f$  is of class  $\mathcal{C}^r$  over  $[-1, 1]$  with  $r \geq 1$ , then  $f^{[N]}$  uniformly converges to  $f$  in  $O(N^{-r})$ , and the convergence is even exponential for analytic functions.

We call  $\mathbb{Q}^1$  the Banach space of continuous functions with absolutely summable Chebyshev series, with norm  $\|f\|_{\mathbb{Q}^1} = \sum_{n \geq 0} |[f]_n|$ . Note that  $\mathbb{Q}^1$  is analogous to the Wiener algebra  $A(\mathbb{T})$  of absolutely convergent Fourier series [12, §I.6]: for  $f \in \mathbb{Q}^1$ , we have  $\|f\|_{\mathbb{Q}^1} = \|f(\cos)\|_{A(\mathbb{T})}$ . We obtain a Banach algebra structure:  $\|fg\|_{\mathbb{Q}^1} \leq \|f\|_{\mathbb{Q}^1} \|g\|_{\mathbb{Q}^1}$ . Moreover, this norm is a safe overestimation of the uniform norm:

$$\|f\|_{\mathbb{Q}^1} \geq \sup_{-1 \leq t \leq 1} \sum_{n \geq 0} |[f]_n T_n(t)| \geq \sup_{-1 \leq t \leq 1} |f(t)| = \|f\|_{\infty}.$$

Given an endomorphism  $\mathbf{F} : \mathbb{Q}^1 \rightarrow \mathbb{Q}^1$ , the operator norm induced by the  $\mathbb{Q}^1$  norm is given by:

$$\|\mathbf{F}\|_{\mathbb{Q}^1} = \sup_{\|f\|_{\mathbb{Q}^1} \leq 1} \|\mathbf{F} \cdot f\|_{\mathbb{Q}^1} = \sup_{n \geq 0} \|\mathbf{F} \cdot T_n\|_{\mathbb{Q}^1}. \quad (9)$$

This corresponds to the maximum sum of the coefficients in absolute value over all columns of the matrix representation of  $\mathbf{F}$ .

***D-finite equations and integral transforms.*** We consider a generic  $p$ -dimensional order  $r$  system of LODEs over the compact interval  $[-1, 1]$ :

$$Y^{(r)} + A_{r-1}(t) \cdot Y^{(r-1)} + \cdots + A_1(t) \cdot Y' + A_0(t) \cdot Y = G(t), \quad (10)$$

of unknown  $Y = (Y_1, \dots, Y_p) : [-1, 1] \rightarrow \mathbb{R}^p$ , with polynomial coefficients  $A_k = (a_{kij})_{1 \leq i, j \leq p} \in \mathcal{M}_p(\mathbb{R}[t])$  and right-hand side  $G = (G_1, \dots, G_p) \in \mathbb{R}[t]^p$ . We also fix initial conditions at  $-1$ :

$$Y^{(i)}(-1) = v_i, \quad v_i \in \mathbb{R}^p, \quad \text{for all } i \in \llbracket 0, r-1 \rrbracket. \quad (11)$$

Together, (1) and (11) form an *Initial Value Problem* (IVP).

Several barriers arise when working directly on a differential equation (1): the differentiation of Chebyshev polynomials does not admit a compact formula, whence a dense linear system to solve, and, from the theoretical point of view, the space  $\mathfrak{U}^1$  is not stable under differentiation. A common way to circumvent these limitations is to apply an integral transform onto the IVP problem so as to obtain an equivalent *Volterra integral equation of the second kind* over  $[-1, 1]$ :

$$\Phi + \mathbf{K} \cdot \Phi = \Psi, \quad \text{with} \quad \mathbf{K} \cdot \Phi(t) = \int_{-1}^t K(t, s) \cdot \Phi(s) ds, \quad (12)$$

with a bivariate polynomial kernel  $K = (k_{ij})_{1 \leq i, j \leq p} \in \mathcal{M}_p(\mathbb{R}[t, s])$  and right-hand side  $\Psi = (\Psi_1, \dots, \Psi_p) \in \mathbb{R}[t]^p$ . Depending on the integral transform, the unknown function  $\Phi = (\Phi_1, \dots, \Phi_p) : [-1, 1] \rightarrow \mathbb{R}^p$  can be either  $Y$  or one of its derivatives. For example, [5] acts over  $Y$ , whereas [7] considers the last derivative  $Y^{(r)}$ .

In any case,  $\mathbf{K} : \Phi \mapsto \int_{-1}^t K(t, s) \cdot \Phi(s) ds$  is a bounded linear operator from  $(\mathfrak{U}^1)^p$  to itself. We may describe it by blocks  $\mathbf{K} = (\mathbf{K}_{ij})_{1 \leq i, j \leq p}$ , where each  $\mathbf{K}_{ij}$  is a one-dimensional integral operator of kernel  $k_{ij}(t, s)$ . By decomposing  $k_{ij}(t, s)$  in Chebyshev basis with respect to  $s$ , we obtain unique polynomials  $b_{ijk}(t)$  such that

$$k_{ij}(t, s) = \sum_{k=0}^{\kappa_{ij}} b_{ijk}(t) T_k(s), \quad \mathbf{K}_{ij} \cdot \varphi(t) = \sum_{k=0}^{\kappa_{ij}} b_{ijk}(t) \int_{-1}^t T_k(s) \varphi(s) ds.$$

Consequently to the multiplication and integration formulas (8), the (infinite dimensional) matrix representation of  $\mathbf{K}_{ij} : \mathfrak{U}^1 \rightarrow \mathfrak{U}^1$  has a so-called  $(h_{ij}, d_{ij})$  *almost-banded* structure [16], meaning that the nonzero entries are located on the  $h_{ij}$  first rows (*horizontal band with initial entries*) and the diagonal plus the first  $d_{ij}$  upper and lower diagonals (*diagonal band with diagonal entries*), with  $h_{ij} = \max_{0 \leq k \leq \kappa_{ij}} \deg b_{ijk}(t)$  and  $d_{ij} = 1 + \deg k_{ij}(t, s) = 1 + \max_{0 \leq k \leq \kappa_{ij}} (k + \deg b_{ijk}(t))$  (see Figure 2(a)).

### 3.2 Efficient numerical solving

The integral equation (12) is an infinite-dimensional linear system over the Chebyshev coefficients of the unknown function  $\Phi$ . The *projection method* (also sometimes called *Galerkin method* [9]) consists in truncating for a given index  $N_{\text{app}}$  and solving the obtained finite-dimensional linear system. In our case, this can be efficiently done by taking advantage of its sparse structure.

Define the  $N_{\text{app}}$ -th truncation of  $\mathbf{K}$  as  $\mathbf{K}^{[N_{\text{app}}]} = (\mathbf{K}_{ij}^{[N_{\text{app}}]})_{1 \leq i, j \leq p}$ , where  $\Pi_{N_{\text{app}}} \cdot \mathbf{K}_{ij} \cdot \Pi_{N_{\text{app}}}$  (see Figure 2(b)). It is represented by the order  $p(N_{\text{app}} + 1)$  square matrix depicted by blocks in Figure 2(c). By permuting the natural basis  $\mathcal{B}_{p, N_{\text{app}}}$  of  $(\Pi_{N_{\text{app}}} \cdot \Upsilon^1)^p$  into  $\mathcal{B}'_{p, N_{\text{app}}}$ :

$$\begin{aligned} \mathcal{B}_{p, N_{\text{app}}} &= (T_0 e_1, \dots, T_{N_{\text{app}}} e_1, \dots, T_0 e_p, \dots, T_{N_{\text{app}}} e_p), \\ \mathcal{B}'_{p, N_{\text{app}}} &= (T_0 e_1, \dots, T_0 e_p, \dots, T_{N_{\text{app}}} e_1, \dots, T_{N_{\text{app}}} e_p), \end{aligned} \quad (13)$$

$\mathbf{K}^{[N_{\text{app}}]}$  recovers a  $(ph, pd)$  almost-banded structure, where  $h = \max_{ij} h_{ij}$  and  $d = \max_{ij} d_{ij}$  (see Figure 2(d)).

Hence, solving the approximate problem:

$$\Phi + \mathbf{K}^{[N_{\text{app}}]} \cdot \Phi = \Psi$$

requires  $O(p^3 N_{\text{app}} d^2)$  operations, using the algorithm of [16] for solving almost-banded linear systems.

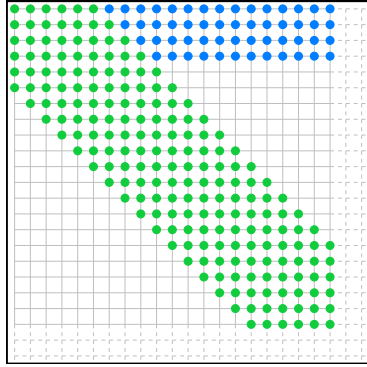
### 3.3 Validation Procedure

We extend the validation procedure of [7] to the vectorial case. We prove the main Theorem 1 in order to solve Problem 1 in two steps: (1) a Newton-like validation operator is created and bounded by **Algorithm 1**. This first step is independent of the approximation degree  $N_{\text{app}}$ . (2) The error enclosure of the given approximation is computed by **Algorithm 3**, following Theorems 2 and 3.

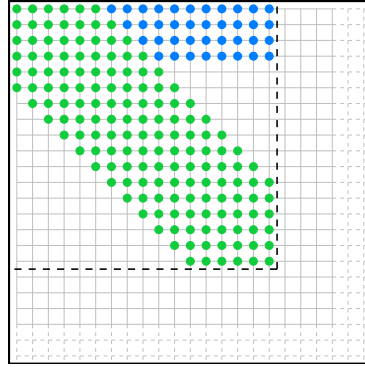
**Newton-like validation operator.** Following the idea of Newton's method and similar approaches, Equation (12) is transformed into the fixed-point equation:

$$\mathbf{T} \cdot \Phi = \Phi, \quad \mathbf{T} \cdot \Phi := \Phi - \mathbf{A} \cdot (\Phi + \mathbf{K} \cdot \Phi - \Psi), \quad (14)$$

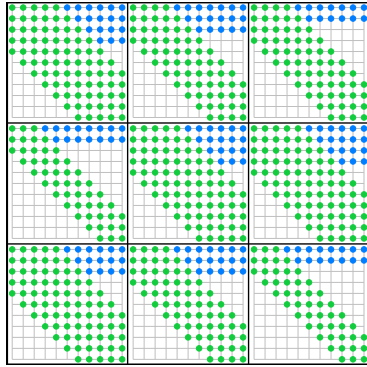
which is equivalent to (12) as soon as  $\mathbf{A} : (\Upsilon^1)^p \rightarrow (\Upsilon^1)^p$  is injective. Moreover,  $\mathbf{T}$  is an affine operator of linear part  $D\mathbf{T} = \mathbf{1} - \mathbf{A} \cdot (\mathbf{1} + \mathbf{K})$ . The main challenge is to efficiently compute  $\mathbf{A}$  and bound  $\|D\mathbf{T}\|_{(\Upsilon^1)^p}$ . This is



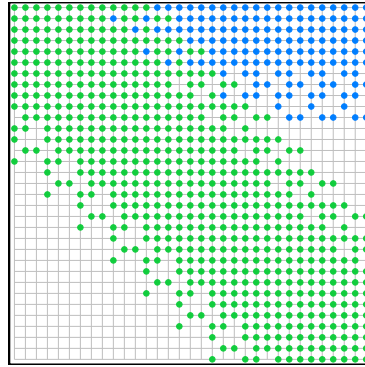
(a) Almost-banded structure of  $\mathbf{K}_{ij}$



(b) Truncation  $\mathbf{K}_{ij}^{[N_{\text{app}}]}$  of  $\mathbf{K}_{ij}$



(c)  $\mathbf{K}^{[N_{\text{app}}]}$  in  $\mathcal{B}_{p,N_{\text{app}}}$ , by blocks ( $p = 3$ )



(d)  $\mathbf{K}^{[N_{\text{app}}]}$  in  $\mathcal{B}'_{p,N_{\text{app}}}$  is almost-banded

Figure 2: Almost-banded structure of integral operators

handled by **Algorithm 1**. Similarly to numerical solving,  $\mathbf{A}$  approximates  $(\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]})^{-1}$ , for some truncation order  $N_{\text{val}}$ . Choosing  $N_{\text{val}}$  is a trade-off between proving  $\mathbf{T}$  is contracting ( $N_{\text{val}}$  must be large enough so that  $\|\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]}\|_{(\mathbb{Q}^1)^p}$  is rigorously proved to be sufficiently small) and efficiency requirements (see [7] for heuristics to find  $N_{\text{val}}$ ).

Once  $N_{\text{val}}$  is fixed, **Algorithm 1** first computes an approximate inverse  $A$  in floating-point (lines 1-4). Since  $\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]}$  is almost-banded in  $\mathcal{B}'_{p, N_{\text{val}}}$ , its numerical inverse can be either computed with the numerically stable algorithm of [16], or approximated by a  $(ph', pd')$  almost-banded matrix [7, **Algorithm 5**], in  $O(p^3 N_{\text{val}}(h' + d')(h + d))$  operations. The operator  $\mathbf{A}$  is defined by extending  $A$  to the whole space  $(\mathbb{Q}^1)^p$  by the identity.

Second, **Algorithm 1** bounds a Lipschitz matrix for  $\mathbf{T}$ , as  $\|\mathbf{DT}\|_{(\mathbb{Q}^1)^p} = (\|(\mathbf{DT})_{ij}\|_{\mathbb{Q}^1})_{1 \leq i, j \leq p}$ , block by block, using the triangle inequality:

$$\|\mathbf{DT}\|_{\mathbb{Q}^1} \leq \|\mathbf{1} - \mathbf{A} \cdot (\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]})\|_{\mathbb{Q}^1} + \|\mathbf{A} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]})\|_{\mathbb{Q}^1}. \quad (15)$$

The first part of (15) is the *approximation error*, measuring how far  $\mathbf{A}$  is from the inverse of  $\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]}$ . This is straightforwardly bounded as  $\Lambda^A$  by **Algorithm 1** (lines 5-9) using  $O(p^3 N_{\text{val}}(h' + d')(h + d))$  interval arithmetic operations, and the resulting bound takes into account all sources of errors: rounding errors, sparse approximation, etc. Since only additions and multiplications of matrices are involved, the use of interval arithmetics is not critical. However, if needed, the underlying floating-point precision can be increased.

The second part of (15) is the *truncation error*, because the truncated operator  $\mathbf{K}^{[N_{\text{val}}]}$  only approximates  $\mathbf{K}$ . Let  $\mathbf{E}_{ij}$  be the  $(i, j)$  block of  $\mathbf{E} := \mathbf{A} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]})$ :

$$\mathbf{E}_{ij} = \sum_{k=1}^p \mathbf{A}_{ik} \cdot (\mathbf{K}_{kj} - \mathbf{K}_{kj}^{[N_{\text{val}}]}). \quad (16)$$

**Algorithm 1** (lines 10-16) computes  $\Lambda^T \geq \|\mathbf{E}\|_{(\mathbb{Q}^1)^p}$  by blocks, with the triangle inequality: each subterm of (16) is rigorously bounded by **Algorithm 2**. This algorithm, detailed below, requires  $O((h' + d')(h + d)^2)$  interval arithmetic operations. Hence the computation of  $\Lambda^T$  is in  $O(p^3(h' + d')(h + d)^2)$ .

Finally, **Algorithm 1** computes  $\Lambda = \Lambda^A + \Lambda^T$  and checks that this Lipschitz matrix is convergent to zero, in which case the constructed Newton-like operator  $\mathbf{T}$  is contracting. The eigenvalues of  $\Lambda$  can be safely computed with interval arithmetics, for the dimension  $p$  is usually small (typically,  $p \leq 100$ ).

*Proof of Theorem 1(i).* The detailed description of **Algorithm 1** above proves its correctness, and the given complexity estimates for lines 1-4, 5-9 and 10-16 sum to a global complexity of  $O(p^3 N_{\text{val}}(h' + d')(h + d))$  operations. In

the worst case, when  $A$  is dense ( $h' + d' \approx N_{\text{val}}$ ), we recover the estimate of Theorem 1(i).  $\square$

**Truncation error bounding.** From Equation (16), one needs to bound  $\|\mathbf{A}_{ik} \cdot (\mathbf{K}_{kj} - \mathbf{K}_{kj}^{[N_{\text{val}}]})\|_{\mathcal{U}^1}$ , where  $\mathbf{A}_{ik}$  is the extension to  $\mathcal{U}^1$  of the order  $N_{\text{val}} + 1$  matrix  $A_{ik}$  by the identity if  $i = k$ , and zero otherwise. This computation is handled by **Algorithm 2**, which is a modification of [7, **Algorithm 6**], that only treats the case  $i = k$ .

Specifically, let  $\mathbf{K}$  denote here a *one-dimensional*  $(h, d)$  almost-banded integral operator, and  $\mathbf{A} : \mathcal{U}^1 \rightarrow \mathcal{U}^1$  the extension of an order  $N_{\text{val}} + 1$  matrix  $A$  by the identity or zero. We have:

$$\|\mathbf{A} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]})\|_{\mathcal{U}^1} = \sup_{\ell \geq 0} B(\ell), \quad \text{with } B(\ell) = \|\mathbf{A} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]}) \cdot T_\ell\|_{\mathcal{U}^1}.$$

Indices  $\ell$  are divided into four groups, reflecting how the initial and diagonal coefficients are impacted by the action of  $\mathbf{A}$ :  $\llbracket 0, N_{\text{val}} - d \rrbracket$ ,  $\llbracket N_{\text{val}} - d + 1, N_{\text{val}} \rrbracket$ ,  $\llbracket N_{\text{val}} + 1, N_{\text{val}} + d \rrbracket$  and  $\llbracket N_{\text{val}} + d + 1, +\infty \rrbracket$ . The  $T_\ell$  in the first group lie in the kernel. The second and third ones are explicitly computed, yielding bounds  $\delta^{(1)}$  and  $\delta^{(2)}$  in **Algorithm 2** (lines 1-7 and 8-13). For the infinite last group,  $B(\ell)$  is decomposed as  $B_I(\ell) + B_D(\ell)$ , the contribution of the initial and diagonal coefficients. **Algorithm 2** uses the efficient bounding strategy of [7]. First, it computes the image of  $T_{N_{\text{val}}+d+1}$  for the initial and diagonal coefficients (lines 16 and 22). Then, it bounds the difference between the images of  $T_{N_{\text{val}}+d+1}$  and the remaining  $T_\ell$  for  $\ell > N_{\text{val}} + d + 1$  to finally deduce bounds  $\delta^{(3)}$  and  $\delta^{(4)}$  (lines 17 and 23).

**Error enclosures.** Finally, **Algorithm 3** implements the validation procedure of Theorems 2 and 3 by applying the operator  $\mathbf{T}$  to the candidate approximation  $\Phi^\circ$ , bounding the distance of the resulting polynomial to  $\Phi^\circ$  and producing componentwise error enclosures to  $\Phi^*$  with respect to the  $\mathcal{U}^1$  norm.

*Proof of Theorem 1(ii).* **Algorithm 3** computes  $\Phi^\circ - \mathbf{T} \cdot \Phi^\circ = \mathbf{A} \cdot (\Phi^\circ + \mathbf{K} \cdot \Phi^\circ - \Psi)$ . Each  $P_k$  (line 1) is a polynomial of degree at most  $\max(N_{\text{app}} + d, N_{\text{rhs}})$ , and computing its Chebyshev coefficients is in  $O(pd^2 N_{\text{app}} + N_{\text{rhs}})$ . Then, the computation of the coefficients of each  $A_{ik} \cdot \Pi_{N_{\text{val}}} \cdot P_k$  (line 3) is in  $O((h' + d') \deg(\Pi_{N_{\text{val}}} \cdot P_k)) = O((h' + d') \min(\max(N_{\text{app}} + d, N_{\text{rhs}}), N_{\text{val}}))$ . Finally, the complexity of computing the enclosures (lines 6-7) only depends on  $p$ , and is therefore negligible. The overall complexity is:

$$O(p^2 d^2 N_{\text{app}} + p N_{\text{rhs}} + p^2 (h' + d') \min(\max(N_{\text{app}} + d, N_{\text{rhs}}), N_{\text{val}})),$$

which gives the estimate of Theorem 1(ii) when  $h', d' \approx N_{\text{val}}$ .  $\square$



---

**Algorithm 1** Create and bound a Newton-like operator  $\mathbf{T}$ 

---

**Require:** A polynomial integral operator  $\mathbf{K} = (\mathbf{K}_{ij})_{1 \leq i, j \leq p}$  given by the  $(b_{ijk})_{0 \leq k \leq \kappa_{ij}}^{1 \leq i, j \leq p}$ , and a truncation order  $N_{\text{val}}$ .

**Ensure:** An approximate inverse  $\mathbf{A}$  of  $\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]}$  and a certified Lipschitz matrix  $\Lambda$  for  $\mathbf{1} - \mathbf{A} \cdot (\mathbf{1} + \mathbf{K})$ , or fail if  $N_{\text{val}}$  not large enough.

▷ *Compute an approximate inverse matrix  $A$ .*

1:  $M = (M_{ij})_{1 \leq i, j \leq p} \leftarrow \mathbf{1} + \mathbf{K}^{[N_{\text{val}}]}$ , by blocks

2:  $M' \leftarrow M$  in basis  $\mathcal{B}'_{p, N_{\text{val}}}$

3:  $A' \leftarrow$  a numerical approximate inverse of  $M'$ , either dense or almost-banded.

4:  $A = (A_{ij})_{1 \leq i, j \leq p} \leftarrow A'$  in basis  $\mathcal{B}_{p, N_{\text{val}}}$ , by blocks

▷ *Compute the approx error  $\Lambda^A = (\lambda_{ij}^A)$  in interval arith.*

5: **for**  $i = 1$  **to**  $p$  **and**  $j = 1$  **to**  $p$  **do**

6:    $C \leftarrow \sum_{1 \leq k \leq p} A_{ik} \cdot M_{kj}$

7:   **if**  $i = j$  **then**  $C \leftarrow C - \mathbf{1}_{N_{\text{val}}+1}$

8:    $\lambda_{ij}^A \leftarrow \|C\|_{\mathbb{Q}^1}$

9: **end for**

▷ *Compute the trunc error  $\Lambda^T = (\lambda_{ij}^T)$  in interval arith.*

10: **for**  $i = 1$  **to**  $p$  **and**  $j = 1$  **to**  $p$  **do**

11:    $\lambda_{ij}^T \leftarrow 0$

12:   **for**  $k = 1$  **to**  $p$  **do**

13:      $\delta \leftarrow$  **Algorithm 2** on  $\mathbf{K}_{jk}$ ,  $A_{ik}$  and  $\text{diag} := (i = k)$ .

14:      $\lambda_{ij}^T \leftarrow \lambda_{ij}^T + \delta$

15:   **end for**

16: **end for**

▷ *Compute  $\Lambda$  and check if  $\mathbf{T}$  contracting.*

17:  $\Lambda \leftarrow \Lambda^A + \Lambda^T$

18: **if**  $\rho(\Lambda) < 1$  **then**

19:   **return**  $A, \Lambda$

20: **else**

21:   **print** "Fail,  $\Lambda$  is not convergent to 0"

22: **end if**

---

---

**Algorithm 2** Bound the truncation error

---

**Require:** A polynomial (one-dimensional) integral operator  $\mathbf{K}$  given by the  $(b_k)_{0 \leq k \leq \kappa}$ , a truncation order  $N_{\text{val}}$ , a  $N_{\text{val}} + 1$  order square matrix  $A$ , and a Boolean *diag*.

**Ensure:** An upper bound  $\delta$  for  $\|\mathbf{A} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]})\|_{\mathbb{Q}^1}$ , where  $\mathbf{A}$  is the extension of  $A$  to the whole space  $\mathbb{Q}^1$  by the identity if *diag* = *true*, and by zero otherwise.

▷ *All operations are performed in interval arithmetics*

▷ Compute  $\delta^{(1)} \geq \sup_{\ell \in \llbracket N_{\text{val}} - d + 1, N_{\text{val}} \rrbracket} B(\ell)$

```
1:  $\delta^{(1)} \leftarrow 0$ 
2: if diag then
3:   for  $\ell = N_{\text{val}} - d + 1$  to  $N_{\text{val}}$  do
4:      $P \leftarrow (\mathbf{1} - \Pi_{N_{\text{val}}}) \cdot \mathbf{K} \cdot T_\ell$ 
5:     if  $\|P\|_{\mathbb{Q}^1} > \delta^{(1)}$  then  $\delta^{(1)} \leftarrow \|P\|_{\mathbb{Q}^1}$ 
6:   end for
7: end if
   ▷ Compute  $\delta^{(2)} \geq \sup_{\ell \in \llbracket N_{\text{val}} + 1, N_{\text{val}} + d \rrbracket} B(\ell)$ 
8:  $\delta^{(2)} \leftarrow 0$ 
9: for  $\ell = N_{\text{val}} + 1$  to  $N_{\text{val}} + d$  do
10:   $P \leftarrow A \cdot \Pi_{N_{\text{val}}} \cdot \mathbf{K} \cdot T_\ell$ 
11:  if diag then  $P \leftarrow P + (\mathbf{1} - \Pi_{N_{\text{val}}}) \cdot \mathbf{K} \cdot T_\ell$ 
12:  if  $\|P\|_{\mathbb{Q}^1} > \delta^{(2)}$  then  $\delta^{(2)} \leftarrow \|P\|_{\mathbb{Q}^1}$ 
13: end for
   ▷ Compute  $\delta^{(3)} \geq \sup_{\ell \geq N_{\text{val}} + d + 1} B_D(\ell)$ 
14:  $\ell_0 \leftarrow N_{\text{val}} + d + 1$  and  $B \leftarrow \sum_{k=0}^{\kappa} \|b_k\|_{\mathbb{Q}^1}$ 
15: if diag then
16:   $P \leftarrow (\mathbf{1} - \Pi_{N_{\text{val}}}) \cdot \mathbf{K} \cdot T_{\ell_0}$ 
17:   $\delta^{(3)} \leftarrow \|P\|_{\mathbb{Q}^1} + \frac{(\kappa+1)B}{(\ell_0 - (\kappa-1))^2}$ 
18: else
19:   $\delta^{(3)} \leftarrow 0$ 
20: end if
   ▷ Compute  $\delta^{(4)} \geq \sup_{\ell \geq N_{\text{val}} + d + 1} B_I(\ell)$ 
21:  $B' \leftarrow \sum_{k=0}^{\kappa} \|A \cdot b_k\|_{\mathbb{Q}^1}$ 
22:  $P \leftarrow A \cdot \Pi_{N_{\text{val}}} \cdot \mathbf{K} \cdot T_{\ell_0}$ 
23:  $\delta^{(4)} \leftarrow \|P\|_{\mathbb{Q}^1} + \frac{(\kappa+1)^3 B'}{(\ell_0^2 - (\kappa+1)^2)^2}$ 
24:  $\delta \leftarrow \max(\delta^{(1)}, \delta^{(2)}, \delta^{(3)} + \delta^{(4)})$ 
25: return  $\delta$ 
```

---

---

**Algorithm 3** Validate a candidate solution of an integral equation

---

**Require:** A polynomial integral operator  $\mathbf{K} = (\mathbf{K}_{ij})_{1 \leq i, j \leq p}$  given by the  $(b_{ijk})_{0 \leq k \leq \kappa_{ij}}$ , a polynomial right-hand side  $\Psi = (\Psi_1, \dots, \Psi_p)$ , a truncation order  $N_{\text{val}}$ ,  $(A, \Lambda)$  obtained from **Algorithm 1** with  $\Lambda$  convergent to 0, and a candidate solution  $\Phi^\circ = (\Phi_1^\circ, \dots, \Phi_p^\circ)$ .

**Ensure:** Two vectors of upper and lower bounds  $\varepsilon^+$  and  $\varepsilon^-$  such that  $\|\Phi_i^\circ - \Phi_i^*\|_{\mathbb{Q}^1} \in [\varepsilon_i^-, \varepsilon_i^+]$  for  $1 \leq i \leq p$ .

▷ *All operations are performed in interval arithmetics*

- 1: **for**  $k = 1$  **to**  $p$  **do**  $P_k \leftarrow \Phi_k + \sum_{j=1}^p \mathbf{K}_{kj} \cdot \Phi_j^\circ - \Psi_k$
  - 2: **for**  $i = 1$  **to**  $p$  **do**
  - 3:    $Q_i \leftarrow \sum_{k=1}^p A_{ik} \cdot \Pi_{N_{\text{val}}} \cdot P_k + (\mathbf{1} - \Pi_n) \cdot P_i$
  - 4:    $\eta_i \leftarrow \|Q_i\|_{\mathbb{Q}^1}$
  - 5: **end for**
  - 6:  $\varepsilon^+ \leftarrow (\mathbf{1} - \Lambda)^{-1} \cdot \eta$
  - 7: **for**  $i = 1$  **to**  $p$  **do**  $\varepsilon_i^- \leftarrow ((\mathbf{1} - D_i \cdot \Lambda)^{-1} \cdot \eta)_i$
  - 8: **return**  $\varepsilon^+$  and  $\varepsilon^-$
- 

**Estimating  $N_{\text{val}}$ .** The following theorem (proved in Appendix A.3) provides a worst-case estimate for the minimal value of  $N_{\text{val}}$ . Although theoretically interesting, this exponential bound is overpessimistic for a wide range of examples.

**Theorem 4.** Let  $B_{ij} = \sum_{k=0}^{\kappa_{ij}} \|b_{ijk}\|_{\mathbb{Q}^1}$  and  $B = (B_{ij})_{1 \leq i, j \leq p}$ .

(i) The following bound estimates the minimal possible value for  $N_{\text{val}}$  making **Algorithm 1** produce a contracting Newton-like operator:

$$N_{\text{val}} = O(d\rho(B)^2 \exp(2\rho(B))),$$

where  $\rho(B)$  denotes the spectral radius of  $B$ .

(ii) For a given approximation  $\Phi^\circ$  of  $\Phi^*$  and in order that **Algorithm 3** computes error enclosures  $[\varepsilon_i^-, \varepsilon_i^+]$  for  $\varepsilon_i = \|\Phi_i^\circ - \Phi_i^*\|_{\mathbb{Q}^1}$  with  $\varepsilon_i^+/\varepsilon_i^- \leq \kappa$  (for some  $\kappa > 1$ ),  $N_{\text{val}}$  for **Algorithm 1** must be at least:

$$N_{\text{val}} = O\left(\frac{\nu}{\kappa - 1} d\|B\|_\infty^2 \exp(2\|B\|_\infty)\right),$$

with  $\nu = \max_{1 \leq i \leq p} \|\varepsilon\|_\infty / \varepsilon_i$ ,  $\|\varepsilon\|_\infty = \max_{1 \leq i \leq p} \varepsilon_i$  and  $\|B\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^p B_{ij}$  the associated operator norm.

## 4 Example and Discussion

Consider the following order 1, two-dimensional system, for  $x \in [0, a]$  with  $a > 0$ , whose solutions (depicted in Figure 3) are highly oscillating functions.

Rescale it over  $[-1, 1]$  with the change of variable  $x = \frac{a}{2}(1 + t)$ :

$$\begin{cases} y_1' = -x^n y_2 \\ y_2' = x^m y_1 \\ y_1(0) = 1, y_2(0) = 0 \end{cases} \Rightarrow \begin{cases} Y_1' = -\left(\frac{a}{2}\right)^{n+1} (1+t)^n Y_2 \\ Y_2' = \left(\frac{a}{2}\right)^{m+1} (1+t)^m Y_1 \\ Y_1(-1) = 1, Y_2(-1) = 0 \end{cases} . \quad (17)$$

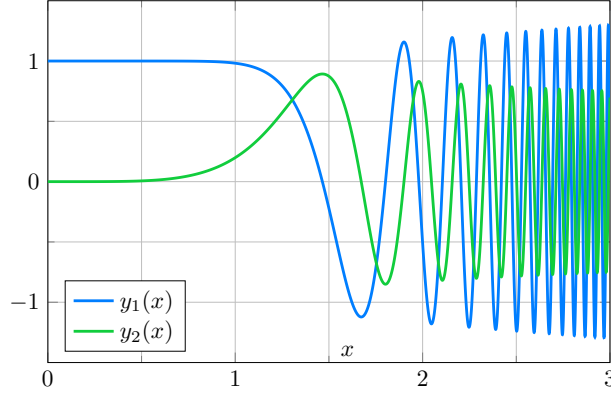


Figure 3: Solution of (17) with  $n = 5$ ,  $m = 4$  and  $a = 3$

We give two different integral transforms associated to this equation. The integral transform described in [5] consists in integrating Equation (17) once, resulting into an integral equation for  $Y$  with polynomial kernel and right-hand side given by:

$$K(t, s) = \begin{pmatrix} 0 & \left(\frac{a}{2}\right)^{n+1} (1+s)^n \\ -\left(\frac{a}{2}\right)^{m+1} (1+s)^m & 0 \end{pmatrix}, \quad \Psi(t) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

$K(t, s)$ , which is of degree 0 in  $t$ , is decomposed over the Chebyshev basis with respect to  $s$  into constant polynomials  $b_{001}, b_{101}, \dots, b_{n01}$  and  $b_{010}, b_{110}, \dots, b_{m10}$ .

On the other side, the integral transform used in [7] allows us to validate the derivative  $\Phi = Y'$ . The polynomial kernel and right-hand side are:

$$K(t, s) = \begin{pmatrix} 0 & \left(\frac{a}{2}\right)^{n+1} (1+t)^n \\ -\left(\frac{a}{2}\right)^{m+1} (1+t)^m & 0 \end{pmatrix}, \quad \Psi(t) = \begin{pmatrix} \left(\frac{a}{2}\right)^{n+1} (1+t)^n \\ 0 \end{pmatrix}.$$

Now,  $K(t, s)$  is of degree 0 with respect to  $s$ , giving two polynomials  $b_{001}$  and  $b_{010}$  of respective degrees  $n$  and  $m$ .

Let's now focus on the first integral transform, with  $n = 5$ ,  $m = 4$ ,  $a = 3$ . Using the spectral method explained in Section 3.1 and implemented in our C library, we fix an approximation degree  $N_{\text{app}} = 100$  and obtain numerical approximations  $Y_1^\circ$  and  $Y_2^\circ$ , that must now be validated. The whole implemented procedure automatically computes and bounds for increasing values

of  $N_{\text{val}}$  the Newton-like operator  $\mathbf{T}$  associated to the truncated operator  $\mathbf{K}^{[N_{\text{val}}]}$ . The approximate inverse is computed as an  $(2h', 2d')$  almost-banded order  $2(N_{\text{val}} + 1)$  matrix. This process stops as soon as the total Lipschitz matrix returned by **Algorithm 1** has a spectral radius less than 1. In case of failure of **Algorithm 1**, the procedure is relaunched with  $N_{\text{val}} \leftarrow 2N_{\text{val}}$ . For this example, we obtain  $N_{\text{val}} = 1664$ ,  $h' = 48$  and  $d' = 304$ , giving the following Lipschitz matrix:

$$\Lambda = \begin{pmatrix} 9.73 \cdot 10^{-4} & 9.89 \cdot 10^{-2} \\ 3.60 \cdot 10^{-2} & 9.92 \cdot 10^{-2} \end{pmatrix}, \quad \rho(\Lambda) = 6.06 \cdot 10^{-2}.$$

The last step is performed by **Algorithm 3**. Given the numerical approximations  $Y_1^\circ$  and  $Y_2^\circ$ , it computes  $\eta = \|Y^\circ - \mathbf{T} \cdot Y^\circ\|_{(\mathbb{Q}^1)^2}$  (the examples gives  $\eta_1 = 3.20 \cdot 10^{-3}$  and  $\eta_2 = 1.91 \cdot 10^{-3}$ ) and outputs the error enclosures given by Theorems 2 and 3:

$$\begin{aligned} \varepsilon_1^- &= 2.99 \cdot 10^{-3}, & \varepsilon_1^+ &= 3.41 \cdot 10^{-3}, \\ \varepsilon_2^- &= 1.78 \cdot 10^{-3}, & \varepsilon_2^+ &= 2.04 \cdot 10^{-3}. \end{aligned}$$

This whole process for this example takes about 30 seconds on a modern computer.

**Comparison with decoupling/desingularization.** In the case of polynomial coefficients, an alternative consists in decoupling the system to obtain  $p$  scalar LODEs of order  $p$ , at the cost of introducing singularities in the equations. As an example, the first component  $y_1$  in (17) satisfies the following differential equation:

$$xy_1'' - ny_1' + x^{n+m+1}y_1 = 0. \quad (18)$$

This equation is *singular* (its leading coefficient vanishes at 0), so our validation method cannot be used. However, with *desingularization* techniques [1], one obtains a higher order but nonsingular equation, whose set of solutions (strictly) contains the ones of the singular equation. In our example, by differentiating Equation (18)  $n$  times and dividing the result by  $x$ :

$$y_1^{(n+2)} + \frac{1}{x} \frac{d^n}{dx^n} (x^{n+m+1}y_1) = 0. \quad (19)$$

By inverting the roles of  $n$  and  $m$ , one obtains a similar equation for  $y_2$ . Hence, validating the approximation  $y$  of (17) can be done with the validation algorithm of [7]. Several caveats must therefore be raised. Applying the integral operator of [7] results into a totally intractable problem, since the minimal value for proving that  $\mathbf{T}$  is contracting is far too large (in practice,

we stopped at  $N_{\text{val}} \simeq 10^6$ ). This is due to the fact that this transform is used to validate the last derivative  $y_1^{(n+2)}$ , which increases very rapidly due to the highly oscillating behavior of  $y_1$ . On the other hand, the integral transform of [5] yields a far more tractable problem: a truncation order  $N_{\text{val}} = 750$  is sufficient for our example. However, Equation (19) is very ill-conditioned because of the factorial terms created by the  $n$  differentiations. For instance, with classical double precision (53 bits), the scalar validation procedure is able to produce and bound a contracting Newton-like operator **T** (**Algorithm 1**), but **Algorithm 3** outputs an upper bound  $\varepsilon_1^+ = 2.57$ , which is 3 orders of magnitude larger than what was found with the vector-valued validation method.

**The non D-finite case.** In the case of nonpolynomial coefficients, there is no general method to decouple and desingularize the system. Moreover, these coefficients may not be known exactly, but only given as polynomial approximations together with rigorous error bounds. We believe that in such a general case, the vector-valued approach presented in this article is essential to approximate and validate the solution. For example, a successful application of our method to a station keeping problem of a satellite is given in [3].

**Future extensions** include: validated expansions in other orthogonal polynomial bases for LODEs; automation and complexity analysis for some classes of nonlinear ODEs; formally proving this method in a proof assistant.

**Heartfelt acknowledgments** are extended to Mioara Joldeş and Nicolas Brisebarre for their extraordinary support in the writing of this article, to Bruno Salvy for instructive discussions about D-finite functions, to Bogdan Pasca for his friendly culinary support.

This work was partially funded by the ANR FastRelax project.

## References

- [1] S. A. Abramov, M. A. Barkatou, and M. Van Hoeij. Apparent singularities of linear difference equations with polynomial coefficients. *Appl. Algebra Eng. Commun. Comput.*, 17(2):117–133, 2006.
- [2] R. P. Agarwal. Contraction and approximate contraction with an application to multi-point boundary value problems. *J. Comput. Appl. Math.*, 9(4):315–325, 1983.
- [3] P. R. Arantes Gilz, F. Bréhard, and C. Gazzino. Validated Semi-Analytical Transition Matrix for Linearized Relative Spacecraft Dynam-

- ics via Chebyshev Polynomials. In *2018 Space Flight Mechanics Meeting, AIAA Science and Technology Forum and Exposition*, page 24, 2018.
- [4] X. Bai. *Modified Chebyshev-Picard iteration methods for solution of initial value and boundary value problems*. PhD thesis, Texas A&M University, 2010.
  - [5] A. Benoit, M. Joldeş, and M. Mezzarobba. Rigorous uniform approximation of D-finite functions using Chebyshev expansions. *Math. Comp.*, 86(305):1303–1341, 2017.
  - [6] J. P. Boyd. *Chebyshev and Fourier spectral methods*. Dover Publications, 2001.
  - [7] F. Bréhard, N. Brisebarre, and M. Joldeş. Validated and numerically efficient Chebyshev spectral methods for linear ordinary differential equations. Preprint (<https://hal.archives-ouvertes.fr/hal-01526272/>), May 2017.
  - [8] C. Clenshaw and H. Norton. The solution of nonlinear ordinary differential equations in Chebyshev series. *The Computer Journal*, 6(1):88–92, 1963.
  - [9] D. Gottlieb and S. A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications*, volume 26. Siam, 1977.
  - [10] A. Hungria, J.-P. Lessard, and J. D. Mireles James. Rigorous numerics for analytic solutions of differential equations: the radii polynomial approach. *Math. Comp.*, 85(299):1427–1459, 2016.
  - [11] L. Kantorovich, B. Vulikh, and A. Pinskiĭ. Functional analysis in partially ordered spaces (in Russian). *Gostekhizdat, Moscow*, 1950.
  - [12] Y. Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.
  - [13] E. W. Kaucher and W. L. Miranker. *Self-validating numerics for function space problems: Computation with guarantees for differential and integral equations*, volume 9. Elsevier, 1984.
  - [14] J.-P. Lessard and C. Reinhardt. Rigorous numerics for nonlinear differential equations using Chebyshev series. *SIAM J. Numer. Anal.*, 52(1):1–22, 2014.

- [15] O. M. Nica-Bolojan. *Fixed point methods for nonlinear differential systems with nonlocal conditions*. PhD thesis, Babes-Bolyai University of Cluj-Napoca, 2013.
- [16] S. Olver and A. Townsend. A fast and well-conditioned spectral method. *SIAM Review*, 55(3):462–489, 2013.
- [17] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 1970.
- [18] A. I. Perov. On the Cauchy problem for a system of ordinary differential equations. *Približ. Metod. Rešen. Differencial'. Uravnen. Vyp.*, 2:115–134, 1964.
- [19] R. Precup. The role of matrices that are convergent to zero in the study of semilinear operator systems. *Math. Comput. Model.*, 49(3):703–708, 2009.
- [20] F. Robert. *Étude et utilisation de normes vectorielles en analyse numérique linéaire (in French)*. PhD thesis, Université de Grenoble, 1968.
- [21] L. N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, 2013. See <http://www.chebfun.org/ATAP/>.
- [22] J. B. van den Berg and J.-P. Lessard. Rigorous numerics in dynamics. *Notices of the AMS*, 62(9), 2015.
- [23] N. Yamamoto. A numerical verification method for solutions of boundary value problems with local uniqueness by Banach’s fixed-point theorem. *SIAM J. Numer. Anal.*, 35(5):2004–2013, 1998.
- [24] T. Yamamoto. A unified derivation of several error bounds for Newton’s process. *J. Comput. Appl. Math.*, 12:179–191, 1985.



## Appendix

### A.1 Complementary Proofs for Section 2

*Proof of Theorem 2.* (i) Endow  $X$  with the metric  $d_\infty(x, y) = \|d(x, y)\|_\infty = \max_{1 \leq i \leq p} d_i(x)$ , so that  $(X, d_\infty)$  is a complete metric space. For an order  $p$  square matrix  $A$ , define:

$$\|A\|_\infty = \max_{1 \leq i \leq p} \sum_{1 \leq j \leq p} |A_{ij}| = \sup_{\|x\|_\infty \leq 1} \|A \cdot x\|_\infty.$$

Since  $\Lambda^k \rightarrow 0$ , there is a  $k$  such that  $\mu = \|\Lambda^k\|_\infty < 1$ . Then  $\mathbf{T}^k$  is a  $\mu$ -contraction for the  $d_\infty$  metric, so that the Banach theorem applies and gives  $x^*$  as the unique fixed-point of  $\mathbf{T}^k$ . Hence  $\mathbf{T}$  can have at most one fixed point. From the following inequality:

$$\begin{aligned} d_\infty(x^*, \mathbf{T} \cdot x^*) &= d_\infty(\mathbf{T}^k \cdot x^*, \mathbf{T}^{k+1} \cdot x^*) \\ &\leq \|\Lambda^k\|_\infty d_\infty(x^*, \mathbf{T} \cdot x^*) < d_\infty(x^*, \mathbf{T} \cdot x^*), \end{aligned}$$

we get that  $x^* = \mathbf{T} \cdot x^*$  is the unique fixed point of  $\mathbf{T}$ .

(ii) Let  $x^\circ \in X$ . Since  $d(x^\circ, x^*) \leq d(x^\circ, \mathbf{T} \cdot x^\circ) + d(\mathbf{T} \cdot x^\circ, x^*) \leq d(x^\circ, \mathbf{T} \cdot x^\circ) + \Lambda \cdot d(x^\circ, x^*)$ , we get:

$$(\mathbf{1} - \Lambda) \cdot d(x^\circ, x^*) \leq d(x^\circ, \mathbf{T} \cdot x^\circ). \quad (20)$$

Since  $\Lambda^k \rightarrow 0$  as  $k \rightarrow \infty$ , it is easy to prove that  $\mathbf{1} - \Lambda$  is nonsingular, with nonnegative inverse  $(\mathbf{1} - \Lambda)^{-1} = \sum_{k \geq 0} \Lambda^k \geq \mathbf{0}$ . Therefore, multiplying both members of Inequality (20) by  $(\mathbf{1} - \Lambda)^{-1}$  is licit, so as to obtain the upper bound (4) for  $n = 0$ . The general bound for  $n \geq 0$  follows from the fact that  $\mathbf{T}$  is  $\Lambda$ -Lipschitz. □

**Lemma 1.** *Let  $\Lambda \in \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+)$  be a convergent to zero nonnegative matrix. Then, for every  $i \in \llbracket 1, p \rrbracket$ ,  $\Lambda - D_i$  is nonsingular and the entries on the  $i$ -th row of its inverse are nonnegative.*

In order to prove the technical Lemma 1, we need the following fact:

**Lemma 2.** *Let  $A \in \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+)$  a convergent to zero nonnegative matrix and  $B \in \mathcal{M}_p(\mathbb{R})$  a matrix whose entries are dominated by those of  $A$ :*

$$|B_{ij}| \leq A_{ij}, \quad \text{for all } i, j \in \llbracket 1, p \rrbracket.$$

*Then  $B$  is convergent to zero.*

*Proof.* Since  $A$  has nonnegative entries which bound those of  $B$ , it can be easily shown by the triangle inequality that for any exponent  $k \geq 0$ ,  $|B_{ij}^k| \leq A_{ij}^k$  for all  $i, j \in \llbracket 1, p \rrbracket$ . This directly implies the conclusion of Lemma 2.  $\square$

*Proof of Lemma 1.* First,  $\mathbf{1} - D_i \cdot \Lambda$  is nonsingular because  $D_i \cdot \Lambda$  is convergent to zero by use of Lemma 2, since its entries are clearly dominated by those of  $\Lambda \in \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+)$ . Hence so is  $\Lambda - D_i$ .

Then we prove that  $\mathbf{1} - \Lambda$  and  $\mathbf{1} - D_i \cdot \Lambda$  both have positive determinant. The segment  $\mathbf{1} - \tau\Lambda$  ( $\tau \in [0, 1]$ ) connects  $\mathbf{1}$  to  $\mathbf{1} - \Lambda$ , and all these matrices are nonsingular, because  $\tau\Lambda$  converges to zero according to Lemma 2. Since  $\det(\mathbf{1}) = 1 > 0$ , we get by connectedness that  $\det(\mathbf{1} - \Lambda) > 0$ . A similar argument proves that  $\det(\mathbf{1} - D_i\Lambda) > 0$ , and hence  $\det(D_i - \Lambda) < 0$ .

Remember that for a nonsingular matrix  $M$ , we have  $M^{-1} = (\det M)^{-1} \text{Cof}(M)^T$ , where  $\text{Cof}(M)$  is the cofactor matrix of  $M$ , whose entries are the minors of  $M$ . Noticing that  $\text{Cof}(D_i - \Lambda)_{ji} = \text{Cof}(\mathbf{1} - \Lambda)_{ji}$  for  $j \in \llbracket 1, p \rrbracket$  and using the fact that  $\det(D_i - \Lambda) < 0$ ,  $\det(\mathbf{1} - \Lambda) > 0$  and all entries in  $(\mathbf{1} - \Lambda)^{-1}$  are nonnegative, we conclude that all entries on the  $i$ -th row of  $(D_i - \Lambda)^{-1}$  are non-positive.  $\square$

## A.2 Discussion About the Tightness of Error Enclosures of Section 2

In the one-dimensional case with a contracting operator of Lipschitz constant  $\lambda \in (0, 1)$ , the ratio of the upper bound and the lower bound given by the Banach fixed-point theorem is equal to  $(1 + \lambda)/(1 - \lambda) > 1$ . This quantity does not depend on the approximation  $x^\circ$ , and uniformly tends to 1 as  $\lambda \rightarrow 0$ , justifying the principle: the more contracting the operator is, the tighter the obtained enclosure is.

This section aims at extending this study to the vectorial case, for the bounds obtained from Theorems 2 and 3. The following lemma quantifies how much the obtained enclosure  $[\varepsilon_i^-, \varepsilon_i^+]$  of  $\varepsilon_i = d(x^\circ, x^\star)_i$  may overapproximate it.

**Lemma 3.** *Let  $\mathbf{T}$  be contracting of Lipschitz matrix  $\Lambda \in \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+)$ ,  $x^\star$  its unique fixed-point,  $x^\circ$  an approximation,  $\varepsilon = d(x^\circ, x^\star)$  the approximation error, and  $\eta = d(x^\circ, \mathbf{T} \cdot x^\circ)$ . Then for all  $i \in \llbracket 1, p \rrbracket$ :*

$$\begin{aligned} \varepsilon_i^+ &:= e_i^T \cdot (\mathbf{1} - \Lambda)^{-1} \cdot \eta \leq e_i^T \cdot (\mathbf{1} - \Lambda)^{-1} \cdot (\mathbf{1} + \Lambda) \cdot \varepsilon \\ \varepsilon_i^- &:= e_i^T \cdot (\mathbf{1} - D_i \cdot \Lambda)^{-1} \cdot \eta \geq e_i^T \cdot (\mathbf{1} - D_i \cdot \Lambda)^{-1} \cdot (\mathbf{1} + D_i \cdot \Lambda) \cdot \varepsilon \end{aligned} \quad (21)$$

The overapproximation ratio can be bounded as follows:

$$\frac{\varepsilon_i^+}{\varepsilon_i^-} \leq \frac{1 + \alpha \nu_i}{1 - \alpha \nu_i}, \quad \text{with } \nu_i = \frac{\|\varepsilon\|_\infty}{\varepsilon_i} \text{ and } \alpha = \frac{2\|\Lambda\|_\infty}{1 - \|\Lambda\|_\infty}, \quad (22)$$

provided that  $\alpha\nu_i < 1$ , where  $\|\varepsilon\|_\infty = \max_{1 \leq i \leq p} |\varepsilon_i|$  is the infinite norm over  $\mathbb{R}^p$  and  $\|\Lambda\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^p |\Lambda_{ij}|$  the associated operator norm.

In particular, Lemma 3 shows that the ratio now depends not only on  $\Lambda$ , but also on  $\varepsilon = d(x^\circ, x^\star)$ .

*Proof.* For the first inequality, we have:

$$\varepsilon_i^+ = e_i^T \cdot (1 - \Lambda)^{-1} \cdot \eta \leq e_i^T \cdot (1 - \Lambda)^{-1} \cdot (1 + \Lambda) \cdot \varepsilon,$$

since  $\eta \leq (1 + \Lambda) \cdot \varepsilon$  by Equation (5) and  $(1 - \Lambda)^{-1}$  has nonnegative coefficients.

For the second inequality, consider  $\bar{\eta} = (1 + D_i \cdot \Lambda) \cdot \varepsilon$ , so that:

$$\begin{aligned} \bar{\eta}_i &= e_i^T \cdot (1 - \Lambda) \cdot \varepsilon \leq \eta_i, \\ \bar{\eta}_j &= e_j^T \cdot (1 + \Lambda) \cdot \varepsilon \geq \eta_j, \text{ for } j \neq i \end{aligned}$$

Since by Lemma 1,  $e_i^T \cdot (1 - D_i \cdot \Lambda)^{-1}$  (the  $i$ -th line of  $(1 - D_i \cdot \Lambda)^{-1}$ ) has a positive entry on the  $i$ -th position and nonpositive entries on the other ones, we have:

$$\varepsilon_i^- = e_i^T \cdot (1 - D_i \cdot \Lambda)^{-1} \cdot \eta \geq e_i^T \cdot (1 - D_i \cdot \Lambda) \cdot \bar{\eta}.$$

In order to prove (22), we notice that:

$$\begin{aligned} e_i^T \cdot (1 - \Lambda)^{-1} \cdot (1 + \Lambda) \cdot \varepsilon &= \varepsilon_i + 2e_i^T \cdot \sum_{n \geq 1} \Lambda^n \cdot \varepsilon \\ &\geq \varepsilon_i + 2 \sum_{n \geq 1} \|\Lambda\|_\infty \|\varepsilon\|_\infty = \varepsilon_i + \alpha \|\varepsilon\|_\infty, \end{aligned}$$

under the condition that  $\|\Lambda\|_\infty < 1$ , which is automatic from assumption  $\alpha\nu_i < 1$ . The inequality:

$$e_i^T \cdot (1 - D_i \cdot \Lambda)^{-1} \cdot (1 + D_i \cdot \Lambda) \cdot \varepsilon \geq \varepsilon_i - \alpha \|\varepsilon\|_\infty,$$

is proven similarly by considering  $D_i \cdot \Lambda$  instead of  $\Lambda$  and  $\|D_i \cdot \Lambda\|_\infty = \|\Lambda\|_\infty$ . Taking the quotient of these two inequalities yields the bound (22).  $\square$

**Tightness cones.** For a fixed  $\Lambda$ , the overapproximation ratio depends on the error distribution of the  $\varepsilon_i$ . More specifically, using Inequality (21) of Lemma 3, the constraint  $\varepsilon_i^+ / \varepsilon_i^- \leq \kappa$  for some ratio  $\kappa > 1$  is fulfilled when:

$$e_i^T \cdot (1 - \Lambda)^{-1} \cdot (1 + \Lambda) \cdot \varepsilon \leq \kappa \cdot e_i^T \cdot (1 - D_i \cdot \Lambda)^{-1} \cdot (1 + D_i \cdot \Lambda) \cdot \varepsilon.$$

The intersection of these constraints for  $i \in \llbracket 1, p \rrbracket$  defines a cone  $\mathcal{C}_\kappa$  in  $\mathbb{R}_+^p$  of points  $\varepsilon$  for which the computed overapproximation ratio does not exceed

$\kappa$ . Under a certain value for  $\kappa$ ,  $\mathcal{C}_\kappa$  is empty, meaning that  $\mathbf{T}$  is not contracting enough to achieve this ratio, whatever  $\varepsilon$  is. The cone  $\mathcal{C}_\kappa$  grows to a *limit cone*  $\mathcal{C}_\infty$  as  $\kappa \rightarrow +\infty$ : a point outside  $\mathcal{C}_\infty$  means that the componentwise error distribution is so unbalanced that some lower bound  $\varepsilon_i^-$  is negative (hence rounded to zero). Figure 4 illustrates the cones  $\mathcal{C}_\kappa$  for different values of  $\kappa$  and the limit cone  $\mathcal{C}_\infty$  arising in our toy example.

**Validation up to a given ratio.** When the problem consists in finding an error enclosure  $[\varepsilon_i^-, \varepsilon_i^+]$  for a given approximation  $x^\circ$  of  $x^*$  such that  $\varepsilon_i^+/\varepsilon_i^- \leq \kappa$ , the fixed-point validation operator  $\mathbf{T}$  must be chosen sufficiently contracting. The following lemma characterizes how small the Lipschitz matrix  $\Lambda$  must be in that case.

**Lemma 4.** *Let  $\mathbf{T}$  be contracting of Lipschitz matrix  $\Lambda \in \mathcal{M}_p^{\rightarrow 0}(\mathbb{R}_+)$ ,  $x^\circ$  an approximation of its unique fixed point  $x^*$ ,  $\varepsilon = d(x^\circ - x^*)$  the approximation error, and  $\kappa > 1$ . If*

$$\|\Lambda\|_\infty \leq \frac{\beta}{2 + \beta} \quad \text{with } \beta = \frac{1}{\nu_i} \frac{\kappa - 1}{\kappa + 1} \text{ and } \nu_i \frac{\|\varepsilon\|_\infty}{\varepsilon_i},$$

*then the upper and lower bounds computed using Theorems 2 and 3 satisfy  $\varepsilon_i^+/\varepsilon_i^- \leq \kappa$ .*

*Proof.* By Lemma 3, having:

$$\frac{1 + \alpha \nu_i}{1 - \alpha \nu_i} \leq \kappa, \quad \text{with } \alpha = \frac{2\|\Lambda\|_\infty}{1 - \|\Lambda\|_\infty},$$

is sufficient to ensure  $\varepsilon_i^+/\varepsilon_i^- \leq \kappa$  (note that the assumption  $\|\Lambda\|_\infty \leq \beta/(2 + \beta)$  implies  $\alpha \nu_i < 1$ ). This in turn is equivalent to  $\alpha \leq \beta$  and finally  $\|\Lambda\|_\infty \leq \beta/(2 + \beta)$ .  $\square$

### A.3 Proof of Theorem 4

*Proof.* (i) The value of  $N_{\text{val}}$  must be sufficiently large to ensure that the truncation error  $\|(\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]})^{-1} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]})\|_{(\mathbf{q}^1)^p}$  is convergent to zero matrix.

- We have as a direct consequence of the one-dimensional case [7, Lemma 3.5]:

$$\|\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]}\|_{(\mathbf{q}^1)^p} = O\left(\frac{B}{N_{\text{val}}}\right).$$

- For  $i \geq 0$ , the bound  $\|\mathbf{K}^i\|_{(\mathbf{q}^1)^p} \leq (6di + 1) \frac{(2C)^i}{i!}$  is generalized from the one-dimensional case contained in the proof of [7, Lemma 3.3], where  $C =$

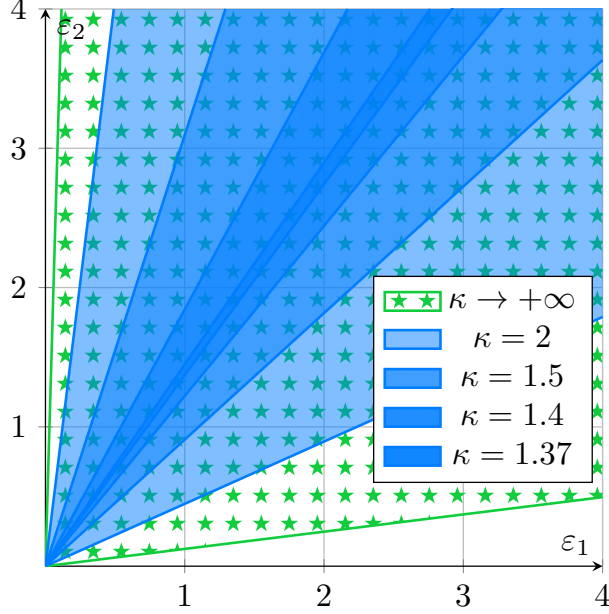


Figure 4: Tightness cones for the toy example

$(C_{ij})_{1 \leq i, j \leq p}$  with  $C_{ij} = \sup_{-1 \leq s, t \leq 1} |k_{ij}(t, s)|$  is bounded by  $B$ . Since  $\mathbf{K}^{[N_{\text{val}}]}$  converges to  $\mathbf{K}$ , we may approximate:

$$\|(\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]})^{-1}\|_{(\mathbb{Q}^1)^p} \approx \|(\mathbf{1} + \mathbf{K})^{-1}\|_{(\mathbb{Q}^1)^p} = O(dB \exp(2B)).$$

- We therefore have:

$$\Lambda = O\left(\frac{dB^2 \exp(2B)}{N_{\text{val}}}\right),$$

where  $\Lambda = \|(\mathbf{1} + \mathbf{K}^{[N_{\text{val}}]})^{-1} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]})\|_{(\mathbb{Q}^1)^p}$ , and by taking the spectral radius (note that  $B$  and  $\exp(2B)$  commute):

$$\rho(\Lambda) = O\left(\frac{d\rho(B)^2 \exp(2\rho(B))}{N_{\text{val}}}\right),$$

which gives the estimate for  $N_{\text{val}}$  to obtain a matrix with spectral radius less than 1.  $\square$

(ii) When a maximal overapproximation ratio  $\kappa > 1$  is fixed, Lemma 4 provides a condition on the infinite norm  $\|\Lambda\|_\infty$  with  $\Lambda = \|\mathbf{D}\mathbf{T}\|_{(\mathbb{Q}^1)^p}$ :

$$\|\Lambda\|_\infty \leq \frac{\beta}{2 + \beta} = O\left(\frac{\kappa - 1}{\nu}\right) \quad \text{with } \beta = \frac{1}{\nu} \frac{\kappa - 1}{\kappa + 1}$$

Again, we focus on the truncation error (since the approximation error can be made as small as desired for a given  $N_{\text{val}}$ ) and we have  $\Lambda = \|(\mathbf{1} - \mathbf{K}^{[N_{\text{val}}]})^{-1} \cdot (\mathbf{K} - \mathbf{K}^{[N_{\text{val}}]})\|_{(\mathbb{Q}^1)^p}$ . Similarly than for the spectral radius, we have:

$$\|\Lambda\|_{\infty} = O\left(\frac{d\|B\|_{\infty}^2 \exp(2\|B\|_{\infty})}{N_{\text{val}}}\right).$$

We therefore obtain the desired estimate.