Computable convergence bounds of series expansions for infinite dimensional linear-analytic systems and application

Thomas Hélie[a], Béatrice Laroche[b]

[a] Ircam - CNRS STMS UMR 9912 - UPMC, 1 place Igor Stravinsky, 75004 Paris, France
[b] INRA - UR 341 Mathématiques et informatique appliquées, Domaine de Vilvert, 78352 Jouy-en-Josas Cedex, France

Abstract

This paper deals with the convergence of series expansions of trajectories for semi-linear infinite dimensional systems, which are analytic in state and affine in input. A special case of such expansions corresponds to Volterra series which are extensively used for the analysis, the simulation and the control of weakly nonlinear finite dimensional systems. The main results of this paper give computable bounds for both the convergence radius and the truncation error of the series. These results can be used for model simplification and analytic approximation of trajectories with a guaranteed quality. They are available for distributed and boundary control systems. As an illustration, these results are applied to an epidemic population dynamic model. In this example, it is shown that the truncation of the series at order 2 yields an accurate analytic approximation which can be used for time simulation and control issues. The relevance of the method is illustrated by simulations.

Keywords: Nonlinear systems, perturbation analysis, partial differential equations, Volterra series expansions, convergence domain

1. Introduction

This paper addresses the representation of nonlinear systems as a series expansion of linear systems with nonlinear interconnections. It investigates on the well-posedness of such series, the accuracy of truncated sums and their use for application issues.

Such representations were first proposed for finite dimensional systems by Vito Volterra [1] who introduced the series named after him. There exists a vast literature concerning Volterra series. Among others, they were studied in [2, 3, 4, 5] from the geometric control point of view, and in [6, 7, 8] from the input-output representation and realization point of view. For linear analytic finite dimensional systems, they correspond to the Taylor series of Frechet derivatives of the input-to-output operator (see [3] and references therein).

Truncated Volterra series (or their low-order optimized approximations [9, 10]) are very convenient for the modeling, identification, model order reduction and real-time simulation of weakly nonlinear systems. This is why they are widely used in signal processing, control, electronics, electromagnetic waves, mechanics, acoustics, bio-medical engineering, etc. However, only a few results about the convergence and truncation error bound are available. The existence of a non zero convergence radius for complex linear analytic finite dimensional systems with zero initial conditions has been proved in [11]. Other theoretical and local-in-time results are known (see e.g. [5, 12]). Results on fading memory have been investigated in [13]. More recent results have been obtained in [14, 15] for the frequency domain, in [16] based on regular perturbations, and in [17] for interconnected systems defined by Fliess series [4]. Computable convergence bounds have also been established for finite dimensional linear-analytic systems in [18].

Here, we address the series representation and the convergence characterization problem for a general class of semi-linear systems, which are analytic in state, affine in input and infinite dimensional. This includes distributed and boundary control systems. We obtain sufficient conditions on both the input and initial condition, under which the series is convergent. Moreover, we give an estimate of the error on trajectories, when approximating the original system by the truncated series.

The paper is organized as follows. Section 2 describes the class of systems under consideration and the proposed series expansion. The main results of the paper, that is the convergence and truncation error bounds of the series expansion, are detailed in section 3 and proved in section 4. Section 5 points out some additional properties and possible refinements. These results are illustrated on a nonlinear epidemic model in section 6, for which a simplified approximating model is derived.

2. Systems under consideration

The following notations and functional setting are introduced:

- $\mathbb{T}$ denotes the time interval $[0, T]$ with $T > 0$ or $\mathbb{R}_+$.
- $\mathbb{U}$ and $\mathbb{X}$ are Banach spaces on the field $\mathbb{R}$.
- $\mathcal{L}(\mathbb{U}, \mathbb{X})$ and $\mathcal{L}(\mathbb{X})$ are the sets of bounded linear operators from $\mathbb{U}$ to $\mathbb{X}$, and from $\mathbb{X}$ to $\mathbb{X}$, respectively.

Email addresses: thomas.helie@ircam.fr (Thomas Hélie), blaroche@jouy.inra.fr (Béatrice Laroche)

- $\mathcal{ML}_k(\mathbb{X},\mathbb{X})$ ($k \geq 2$) is the set of bounded multilinear operators from $\underbrace{\mathbb{X}\times\cdots\times\mathbb{X}}_{k}$ to $\mathbb{X}$, with norm $\|E\| = \sup_{\substack{(x_1,\dots,x_k)\in\mathbb{X}^k\\ \|x_1\|=\cdots=\|x_k\|=1}} \|E(x_1,\dots,x_k)\|_{\mathbb{X}}$.
- $\mathcal{ML}_{k,1}(\mathbb{X},\mathbb{U},\mathbb{X})$ ($k \geq 1$) is the set of bounded multilinear operators from $\underbrace{\mathbb{X}\times\cdots\times\mathbb{X}}_{k}\times\mathbb{U}$ to $\mathbb{X}$, with norm $\|E\| = \sup_{\substack{(x_1,\dots,x_k,u)\in\mathbb{X}^k\times\mathbb{U}\\ \|x_1\|=\cdots=\|x_k\|=\|u\|=1}} \|E(x_1,\dots,x_k,u)\|_{\mathbb{X}}$.
- $\mathcal{U} = L^\infty(\mathbb{T},\mathbb{U})$ and $\mathcal{X} = L^\infty(\mathbb{T},\mathbb{X})$ are standard Lebesgue spaces.

We consider the class of infinite-dimensional nonlinear control systems on $\mathbb{X}$, having an equilibrium state (shifted to zero without loss of generality), governed by

$$\dot{x} = L(x,u) + P(x) + Q(x,u), \text{ for } t\in\mathbb{T}, \quad (1)$$
$$x(0) = x_{\text{ini}} \in \mathbb{X}. \quad (2)$$

The notation $L(x,u)$ stands for the linear part of the system. We assume that the linearized system

$$\dot{x}_1 = L(x_1,u), \quad x_1(0) = x_{\text{ini}} \in \mathbb{X}, \quad (3)$$

is a distributed or boundary control system in the sense of [19]. This implies that $A = L(.,0)$ generates a strongly continuous semigroup on $\mathbb{X}$, denoted V, with $\alpha\in\mathbb{R}$ and $\beta>0$ such that for all $t\in\mathbb{T}$, $\|V(t)\| \leq \beta e^{\alpha t}$. The growth bound α is assumed to be strictly negative if $\mathbb{T} = \mathbb{R}_+$.
Moreover, the linearized system (3) is assumed to be well-posed, that is, for all $u\in\mathcal{U}$ and $x_{\text{ini}}\in\mathbb{X}$, (3) has a unique mild solution $x_1\in\mathcal{X}$. P and Q are nonlinear terms such that

$$P(x) = \sum_{k=2}^{+\infty} A_k(\underbrace{x,\dots,x}_{k}), \quad (4)$$
$$Q(x,u) = \sum_{k=2}^{+\infty} B_k(\underbrace{x,\dots,x}_{k-1},u), \quad (5)$$

where $A_k\in\mathcal{ML}_k(\mathbb{X},\mathbb{X})$ and $B_k\in\mathcal{ML}_{k-1,1}(\mathbb{X},\mathbb{U},\mathbb{X})$ are multilinear bounded operators. The complex functions

$$a: z\longmapsto \sum_{k=2}^{+\infty}\|A_k\|\,z^k, \quad b: z\longmapsto \sum_{k=2}^{+\infty}\|B_k\|\,z^k, \quad (6)$$

are assumed to be analytic at $z=0$.
A series expansion of the trajectories of (1)-(2) is defined in the following way. For all $m\geq 2$, x_m is the mild solution of

$$\dot{x}_m = Ax_m + \chi_m, \quad x_m(0) = 0, \quad (7)$$

$$\text{where } \chi_m(\tau) = \sum_{k=2}^{m}\sum_{p\in\mathbb{M}_m^k} A_k\Big(x_{p_1}(\tau),\dots,x_{p_k}(\tau)\Big) + \sum_{k=2}^{m}\sum_{\substack{q\in\mathbb{M}_m^k\\ q_k=1}} B_k\big(x_{q_1}(\tau),\dots,x_{q_{k-1}}(\tau),u(\tau)\big), \quad (8)$$

where $\mathbb{M}_m^K$ is defined for all $m\in\mathbb{N}^*$ and $K\in\mathbb{N}^*$ by

$$\mathbb{M}_m^K = \Big\{p\in(\mathbb{N}^*)^K \;\Big|\; p_1+\cdots+p_K = m\Big\}.$$

As a well known result [19, 20], we have

$$x_m(t) = \int_0^t V(t-\tau)\chi_m(\tau)\,\mathrm{d}\tau. \quad (9)$$

The series expansion of the trajectories is

$$x(t) = \sum_{m=0}^{+\infty} x_m(t). \quad (10)$$

It provides a formal solution of (1)-(2).
In the case of finite dimensional linear analytic systems with zero initial conditions, the semigroup associated with the linearized system is $V(t) = e^{At}$, the solution x_1 is the convolution of the input by the impulse response matrix VB. Moreover, (9) corresponds to a multiple convolution (of order m) by a multi-variate kernel and (10) exactly coincides with a standard Volterra series expansion (see e.g. [1, 6, 21]). It is shown in [3] that for such systems, this expansion is indeed the Taylor series of Frechet derivatives of the input-to-state operator.
A realization of the partial sum of order three of (10) is displayed in figure 1. Each term is built as a cascade of linear systems (*lin*, V) and static nonlinear interconnections (A_k, B_k), which provides an easily implementable simplified model. Moreover, this realization is directly expressed in terms of the original system parameters, which constitutes an appealing feature for design issues and physical interpretations.
From a general point of view, approximations by low-order truncated Taylor series are well-adapted to "weakly nonlinear systems for sufficiently small inputs": section 3 provides quantitative assessment criteria for this statement, based on the system parameters. More precisely, we establish (i) a guaranteed convergence domain of (10) with respect to the input and the initial conditions, and (ii) an estimate of the remainder with respect to the tuncation order.
For applications requiring a low-order approximation even for large inputs, optimal approximations may be preferred to truncation. Although this is beyond the scope of this paper, results (i-ii) can still be helpful in this case. Indeed, result (i) provides a range over which optimal approximations defined in e.g. [9, 10] are guaranteed to make sense. Moreover, result (ii) provides a guaranteed estimate to which the optimal approximation error can be compared.

3. Main results

Our first main result is a sufficient condition on x_{ini} and u for the convergence of the series (10), for which we need to introduce the following definitions.
For all $t\in\mathbb{T}$, we set

$$f(t) = \max\left[\sup_{\substack{k\geq 2 \text{ s.t.}\\ \|A_k\|\neq 0}} \frac{\|V(t)\,A_k\|}{\|A_k\|}\,,\ \sup_{\substack{k\geq 2 \text{ s.t.}\\ \|B_k\|\neq 0}} \frac{\|V(t)\,B_k\|}{\|B_k\|}\right].$$

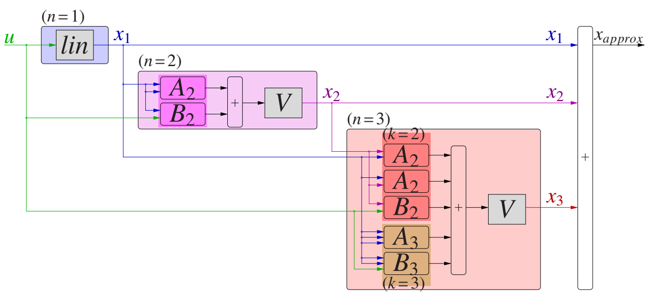


Figure 1: Realization of the first three terms of the series expansion (10). For $n = 1$, *lin* denotes a realization of the linearized system.

By construction, f is bounded by $t \mapsto \beta e^{\alpha t}$. Hence, we can define $\gamma^\star$ and choose γ such that

$$\gamma \geq \gamma^\star = \int_{\mathbb{T}} f(t)\,\mathrm{d}t. \tag{11}$$

For all $\omega \geq 0$, function F_ω is defined on $\mathbb{C}$ by

$$F_\omega(z) = \frac{z + \gamma\,\omega\, b(z)}{z - \gamma\, a(z)}. \tag{12}$$

¿From lemma 2-i in section 4.2, F_ω is analytic with convergence radius $r \in \mathbb{R}_+^* \cup \{+\infty\}$ at $z = 0$. ¿From 2-ii, the characteristic equation

$$x\,F_\omega'(x) - F_\omega(x) = 0 \tag{13}$$

has at most one solution in $]0, r[$. If it has one solution, denoted σ_ω, we define $\rho_\omega > 0$ as

$$\rho_\omega = \frac{\sigma_\omega}{F_\omega(\sigma_\omega)}, \quad \text{(case 1)}, \tag{14}$$

otherwise, we set

$$\rho_\omega = \lim_{x \to r^-} \frac{x}{F_\omega(x)}, \quad \text{(case 2)}. \tag{15}$$

¿From 2-iii, there exists a unique function $z \mapsto \Phi_\omega(z)$ analytic at $z = 0$ and such that $\Phi_\omega(z) = z\,F_\omega(\Phi_\omega(z))$, whose convergence radius is bounded from below by ρ_ω.

Now, defining $\delta_Q = 0$ if $Q = 0$ in (1) and $\delta_Q = 1$ otherwise, we are in position to state our first main result.

Theorem 1 (Convergence criterion)**.** *Let $\omega \in \mathbb{R}_+$, $x_{\text{ini}} \in \mathbb{X}$ and $u \in \mathcal{U}$ be such that*

$$\delta_Q \|u\|_{\mathcal{U}} \leq \omega\,\|x_1\|_{\mathcal{X}} \quad \text{and} \quad \|x_1\|_{\mathcal{X}} < \rho_\omega. \tag{16}$$

Then, the series $x = \sum_{m \in \mathbb{N}^*} x_m$ converges in norm in $\mathcal{X}$ and

$$\big\|x\big\|_{\mathcal{X}} \leq \Phi_\omega(\|x_1\|_{\mathcal{X}}).$$

Our second main result is a guaranteed bound for the truncation error of the series.

Theorem 2 (Remainder estimates)**.** *Let $\omega \geq 0$ and $M \in \mathbb{N}^*$. Let the remainder function be defined on $\{z \in \mathbb{C}$ s.t. $|z| < \rho_\omega\}$ by*

$$R_M\Phi_\omega(z) = \sum_{m=M+1}^{+\infty} \varphi_m(\omega)\,z^m = \Phi_\omega(z) - \sum_{m=1}^{M} \varphi_m(\omega)\,z^m, \tag{17}$$

where $\varphi_m(\omega)$ is the Taylor coefficient of order m of Φ_ω. Then, for all $(u, x_{\text{ini}}) \in \mathcal{U} \times \mathbb{X}$ satisfying (16),

$$\Big\|x - \sum_{m=1}^{M} x_m\Big\|_{\mathcal{X}} \leq R_M\Phi_\omega(\|x_1\|_{\mathcal{X}}) < +\infty. \tag{18}$$

The general principle of the proof is based on three key steps. First, we derive a majorizing series of (10) under the form of a power series of $\|x_1\|_{\mathcal{X}}$. Second, we exhibit an appropriate functional equation satisfied by the power series. Third, we relate the functional equation satisfied by the power series to the asymptotic behavior of its coefficients, providing its convergence radius. This is done using standard tools of combinatorial analysis and more particularly the singular inversion theorem (see e.g. [22]). The resulting convergence bound and remainder estimate constitute the desired results.

This framework can be adapted to particular situations (see section 5) but the key steps remain the same.

4. Proof of main results

4.1. Proof of theorems 1-2

The three steps of the proof are detailed below.

Step 1: majorizing series. Let $\omega \geq 0$. Define $\varphi_1(\omega) = 1$ and, for $m \geq 2$,

$$\varphi_m(\omega) = \gamma \sum_{k=2}^{m} \Bigg[\|A_k\| \sum_{p \in \mathbb{M}_m^k} \prod_{i=1}^{k} \varphi_{p_i}(\omega) + \omega\,\|B_k\| \sum_{\substack{q \in \mathbb{M}_m^k \\ q_k = 1}} \prod_{i=1}^{k-1} \varphi_{q_i}(\omega) \Bigg]. \tag{19}$$

Then, for all $(u, x_{\text{ini}}) \in \mathcal{U} \times \mathbb{X}$, $\|x_1\|_{\mathcal{X}} = \varphi_1(\omega)\,\|x_1\|_{\mathcal{X}}$ and, by induction from lemma 1 below,

$$\forall m \geq 2, \quad \|x_m\|_{\mathcal{X}} \leq \varphi_m(\omega)\,\|x_1\|_{\mathcal{X}}^m. \tag{20}$$

Hence, introducing the generating function $\Phi_\omega(X) = \sum_{n \in \mathbb{N}^*} \varphi_m(\omega) X^m$, it follows that $\Phi_\omega(\|x_1\|)$ is a majorizing series of $\sum_{m \in \mathbb{N}^*} x_m$.

Step 2: functional equation. We proceed by noticing that

$$\begin{aligned}
&\gamma \left(a(\Phi_\omega(X)) + \omega\, X\, \frac{b(\Phi_\omega(X))}{\Phi_\omega(X)} \right) \\
&= \gamma \left(\sum_{k=2}^{+\infty} \|A_k\| (\Phi_\omega(X))^k + \omega\, X \sum_{k=2}^{+\infty} \|B_k\| (\Phi_\omega(X))^{k-1} \right) \\
&= \gamma \sum_{m=2}^{+\infty} X^m \left(\sum_{k=2}^{m} \|A_k\| \sum_{p \in \mathbb{M}_m^k} \prod_{i=1}^{k} \varphi_{p_i}(\omega) + \omega\,\|B_k\| \sum_{\substack{q \in \mathbb{M}_m^k \\ q_k = 1}} \prod_{i=1}^{k-1} \varphi_{q_i}(\omega) \right) \\
&= \sum_{m=2}^{+\infty} X^m \varphi_m(\omega) = \Phi_\omega(X) - X,
\end{aligned}$$

leading to $\Phi_\omega(X) = X\,F_\omega(\Phi_\omega(X))$.

Step 3: asymptotic estimates. ¿From lemma 2-iii, Φ_ω is the unique solution of this equation that is analytic in the open disk with radius ρ_ω. So, if $\|x_1\|_{\mathcal{X}} < \rho_\omega$, the series $\sum_{m \in \mathbb{N}^*} \varphi_m(\omega)\,\|x_1\|_{\mathcal{X}}^m$

converges. This proves that $\sum_{m\in\mathbb{N}^*} x_m$ converges in norm and is bounded by $\Phi_\omega(\|x_1\|_{\mathcal{X}})$.

Finally, theorem 2 is an immediate consequence of theorem 1 and (20). This concludes the proof.

4.2. Technical lemmas

The estimate (20) of $\|x_m\|_{\mathcal{X}}$ used in the proof is a consequence of the the following lemma.

Lemma 1 (Regularity and norm estimates)**.** *Let $(u, x_{\text{ini}}) \in \mathcal{U} \times \mathbb{X}$. Then, for all $m \geq 2$, x_m belongs to $C_0(\mathbb{T}, \mathbb{X}) \cap \mathcal{X}$ and*

$$\|x_m\|_{\mathcal{X}} \leq \gamma \sum_{k=2}^{m} \Big[\|A_k\| \sum_{p\in\mathbb{M}_m^k} \prod_{i=1}^{k} \|x_{p_i}\|_{\mathcal{X}} + \|B_k\| \sum_{\begin{cases} q\in\mathbb{M}_m^k \\ q_k = 1\end{cases}} \Big(\prod_{i=1}^{k-1} \|x_{q_i}\|_{\mathcal{X}} \Big) \|u\|_{\mathcal{U}} \Big]. \tag{21}$$

Proof. By assumption, $x_1 \in \mathcal{X}$. Let $m \geq 2$ and assume that for $1 \leq m' \leq m-1$, $x_{m'} \in \mathcal{X}$. Then $\chi_m \in \mathbb{X}$ and x_m belongs to $C_0(\mathbb{T}, \mathbb{X})$. From (8), (9) and (11), for all $t \in \mathbb{T}$,

$$\int_0^t \|V(t-\tau)\chi_m(\tau)\|_{\mathbb{X}}\, d\tau \leq \sum_{k=2}^{m} \gamma \Big[\|A_k\| \sum_{p\in\mathbb{M}_m^k} \prod_{i=1}^{k} \|x_{p_i}\|_{\mathcal{X}} + \|B_k\| \sum_{\begin{cases} q\in\mathbb{M}_m^k \\ q_k = 1\end{cases}} \Big(\prod_{i=1}^{k-1} \|x_{q_i}\|_{\mathcal{X}} \Big) \|u\|_{\mathcal{U}} \Big],$$

This proves by induction that for all $m \geq 1$, $x_m \in \mathcal{X}$. Moreover, for all $m \geq 2$, (21) holds. □

The tools from combinatorial analysis used in the proof (third part of section 4.1) are summarized below.

Lemma 2. *Let $A(X) = \sum_{k=1}^{+\infty} \alpha_k X^k$ and $B(X) = \sum_{k=1}^{+\infty} \beta_k X^k$ be analytic functions at $X = 0$ with non-negative coefficients. Define $F(X) = \dfrac{1 + B(X)}{1 - A(X)}$ and let $r \in \mathbb{R}_+^* \cup \{+\infty\}$ be the radius of convergence of F at $x = 0$. Then, the following results hold:*

(i) At $x = 0$, F is nonzero and analytic with nonnegative Taylor coefficients.

(ii) Equation $x F'(x) - F(x) = 0$ has either one solution denoted σ (case 1) or zero solution (case 2), in $]0, r[$.

(iii) There exists a unique function $z \mapsto \Phi(z)$, analytic at $z = 0$ such that $\Phi(z) = z F(\Phi(z))$. Its convergence radius ρ at $z=0$ is such that

$$\text{(case 1)} \qquad \rho = \frac{\sigma}{F(\sigma)}, \tag{22}$$

$$\text{(case 2)} \qquad \rho \geq \lim_{x\to r^-} \frac{x}{F(x)}. \tag{23}$$

Proof. **Assertion (i):** If $A = 0$, (i) is straightforward. Otherwise, A has at least one positive Taylor coefficients so that, for all $z \in \mathbb{C}$ such that $|z| < r$, $|A(z)| \leq A(|z|) < \lim_{x\to r^-} A(x) \leq 1$ and $F(z) = (1 + B(z)) \sum_{n=0}^{+\infty} (A(z))^n$, which proves (i).
Assertion (ii): Define $H(x) = x F'(x) - F(x)$ for $x \in [0, r[$. If F is affine then $H(x) = -1$ so that $x F'(x) - F(x) = 0$ has no solution. Otherwise, H is a strictly increasing function on $]0, r[$ from $H(0) < 0$ to $\ell = \lim_{x\to r^-} H(x) \in \mathbb{R} \cup \{+\infty\}$ since for all $x \in]0, r[$, $H'(x) = x F''(x) > 0$. Therefore, if $\ell > 0$, then $x F'(x) - F(x) = 0$ has a unique solution on $[0, r[$ (case 1), otherwise ($\ell \leq 0$), it has no solution (case 2).
Assertion (iii): In case 1, the hypotheses of the singular inversion theorem (see e.g. proposition IV.5. and theorem VI.6. in [22]) are met, and its application proves (iii). In case 2, (iii) is a direct consequence of the analytic inversion lemma (see e.g. lemma 4.2. in [22]). □

5. Additional results and refinements

5.1. Parameter ω

The influence of parameter ω in theorem 1 can be further investigated. This parameter accounts for the effect of the (input dependent) term Q in the system dynamics.

If $\delta_Q = 0$, then $\rho_\omega = \rho_0$ does not depend on ω. Otherwise, the following proposition holds.

Proposition 1. *If $Q \neq 0$, function $\omega \mapsto \rho_\omega$ is strictly decreasing.*

Proof. The result is straightforward in case 2 (see (15)) since r does not depend on ω and b is positive. In case 1, denoting $\rho(\omega) = \rho_\omega$ and $\sigma(\omega) = \sigma_\omega$, we have

$$\rho'(\omega) = -\frac{\sigma'(\omega)}{F_\omega(\sigma(\omega))^2} \Big[\sigma(\omega) F'_\omega(\sigma(\omega)) - F_\omega(\sigma(\omega)) \Big] - \frac{\sigma(\omega)}{F_\omega(\sigma(\omega))^2} [\partial_\omega F_\omega](\sigma(\omega)).$$

The first term is zero from (13) and because $\frac{\sigma'(\omega)}{F_\omega(\sigma(\omega))^2}$ is finite. Indeed, $1 < F_\omega(\sigma(\omega)) < +\infty$ since $0 < \sigma(\omega) < r$, $\big(1 - \gamma\, a(\sigma(\omega))/\sigma(\omega)\big)^{-1} > 1$ and $1 + \gamma\, \omega\, b(\sigma(\omega))/\sigma(\omega) \geq 1$. Moreover, $\sigma'(\omega) = \frac{[\partial_\omega F_\omega](\sigma(\omega)) - \sigma(\omega)\,[\partial_\omega F'_\omega](\sigma(\omega))}{\sigma(\omega)\, F''_\omega(\sigma(\omega))}$ is finite since $z \mapsto F_\omega(z)$ is non-affine with positive Taylor coefficients in case 1 (see lemma 2).

The second term is strictly negative since $\sigma(\omega) > 0$ and for all $z > 0$, $\partial_\omega F_\omega(z) = \frac{\gamma\, b(z)}{z - \gamma a(z)} \geq \frac{\gamma b(z)}{z} > \frac{\gamma b(z)}{z}\Big|_{z=0} = 0$, which concludes the proof. □

5.2. Bound tightness

Another issue is the tightness of the convergence and truncation error bounds given in theorems 1-2. We do not expect these bounds to be optimal in general, since the estimate in lemma 1 can be rather coarse.

For instance, for the system $\dot{x} = a_2 x^2$, $x(0) = x_0$, $a_2 > 0$, on $\mathbb{T} = [0, T]$ (example 1), we get $F_\omega(z) = 1/(1 - T a_2 z)$

and $\rho_\omega = 1/(4Ta_2)$. For $\dot{x} = b_2 x u$, $x(0) = 0$, $b_2 > 0$ on $\mathbb{T} = [0, T]$ (example 2), we get $F_\omega(z) = 1 + \omega T b_2 z$ and $\rho_\omega = 1/(\omega T b_2)$. In both cases, a direct computation of the series expansion and its convergence radius shows that the convergence bound in theorem 1 is underestimated. However, for system $\dot{x} = \alpha x + u + a_2 x^2$, $x(0) = 0$, this bound is optimal [23]. This bound was also tested on a Euler-Bernoulli beam with a higher order nonlinearity (order three): it reveals to be accurate [24].

As mentioned in section 3, it is possible to improve the bounds given in theorems 1-2 in particular situations. As an example, we show how this can be done for quadratic-bilinear systems on a finite time horizon, namely systems for which $P(x) = A_2(x, x)$, $Q(x, u) = B_2(x, u)$ and $\mathbb{T} = [0, T]$.

For these systems, a simple but useful refinement of theorem 2 is given below. It used in section 6.

Proposition 2. *Consider a quadratic-bilinear system with input and initial condition such that theorems 1 and 2 apply. Assume that we can find $\kappa \leq 1$ such that*

$$\|x_2\|_X \leq \kappa\, \varphi_2\, \|x_1\|_X^2. \tag{24}$$

Then, for all $M \geq 2$, $\left\| x - \sum_{m=1}^{M} x_m \right\|_X \leq \kappa R_M \Phi_\omega(\|x_1\|_X)$.

Proof. Since the system is quadratic, a straightforward induction shows that $\|x_m\|_X \leq \kappa\varphi_m \|x_1\|_X^m$ hold for all $m \geq 2$. Hence $(1-\kappa)z + \kappa\Phi(z)$ is a dominating series of $\sum_{m\geq 1} x_m$ and the result follows. □

To improve the convergence bound in theorem 1, we introduce $K_T = \beta T \max(1, e^{\alpha T})$ and define, for $\omega \in \mathbb{R}_+$,

$$\widetilde{\Phi}_\omega(X) = \frac{X\, e^{K_T \omega b_2 X}}{1 - a_2 \frac{1 - e^{K_T \omega b_2 X}}{\omega b_2}}, \quad \widetilde{\rho}_\omega = \frac{\ln\left(1 + \frac{\omega b_2}{a_2}\right)}{K_T \omega b_2}, \quad \text{if } \delta_Q \neq 0, \tag{25}$$

$$\widetilde{\Phi}(X) = \frac{X}{1 - K_T\, a_2\, X}, \qquad \widetilde{\rho} = \frac{1}{K_T a_2}, \qquad \text{otherwise}, \tag{26}$$

where $a_2 = \|A_2\|$, $b_2 = \|B_2\|$. Then the following holds.

Proposition 3. *Let $(u, x_{\text{ini}}) \in \mathcal{U} \times \mathbb{X}$ be such that $\delta_Q \|u\|_{\mathcal{U}} \leq \omega \|x_1\|_X$. A sufficient condition for the series $x = \sum_{m\in\mathbb{N}^*} x_m$ to converge in norm in X is $\|x_1\|_X < \widetilde{\rho}_\omega$. Moreover, the truncation error of order $M > 0$ is bounded in X by $R_M \widetilde{\Phi}_\omega(\|x_1\|_X)$.*

Proof. We use the same key steps as for theorem 1, with different estimates.

Step 1: majorizing series. By definition, $\|S(t)\| \leq K_T/T$ for all $t \in \mathbb{T}$. Let us set $\widetilde{\varphi}_1(\omega) = 1$ and

$$\widetilde{\varphi}_m(\omega) = \frac{K_T}{m-1}\left[a_2 \sum_{j=1}^{m-1} \widetilde{\varphi}_j(\omega)\widetilde{\varphi}_{m-j}(\omega) + \omega b_2\, \widetilde{\varphi}_{m-1}(\omega)\right], \tag{27}$$

where ω is omitted in the sequel. We prove by induction that for all $m \in \mathbb{N}$ and for all $t \in \mathbb{T}$, $\|x_m(t)\|_{\mathbb{X}} \leq \widetilde{\varphi}_m \|x_1\|_X^m (t/T)^{m-1}$. The claim is true for $m = 1$. For $m > 1$, if it holds for all $m' < m$, we have for all $t \in \mathbb{T}$,

$$\begin{aligned}
\|x_m(t)\|_{\mathbb{X}} &\leq \int_0^t \|S(t-\tau)\|\, \|\chi_m(\tau)\|_{\mathbb{X}}\, \mathrm{d}\tau \\
&\leq \int_0^t \|S(t-\tau)\| \Big[a_2 \sum_{j=1}^{m-1} \|x_j(\tau)\|_{\mathbb{X}}\, \|x_{m-j}(\tau)\|_{\mathbb{X}} \\
&\qquad + b_2 \|x_{m-1}\|_{\mathbb{X}}\, \|u\|_{\mathcal{U}}\Big] \mathrm{d}\tau, \\
&\leq \frac{K_T}{T} \|x_1\|_X^m \left[a_2 \sum_{j=1}^{m-1} \widetilde{\varphi}_j \widetilde{\varphi}_{m-j} + \omega b_2\, \widetilde{\varphi}_{m-1}\right] \int_0^t \left(\frac{\tau}{T}\right)^{m-2} \mathrm{d}\tau \\
&\leq \widetilde{\varphi}_m\, \|x_1\|_X^m \left(\frac{t}{T}\right)^{m-1},
\end{aligned}$$

which proves the induction.

Step 2: equation. From (27), a straightforward computation shows that $\Psi(X) = \sum_{m\in\mathbb{N}^*} \widetilde{\varphi}_m X^{m-1}$ satisfies a differential (instead of functional) equation

$$\mathrm{d}\Psi/\mathrm{d}X = K_T a_2 \Psi^2 + K_T \omega b_2 \Psi \text{ with } \Psi(0) = 1.$$

Step 3: estimates. Solving this differential equation shows that $X\Psi(X) = \sum_{m\in\mathbb{N}^*} \widetilde{\varphi}_m X^m$ coincides with $\widetilde{\Phi}_\omega(X)$ given in (25). So, the majorizing series $\widetilde{\Phi}_\omega$ converges provided that $\|x_1\|_X$ is less than $\widetilde{\rho}_\omega$, which concludes the proof. □

Obviously, the best results obtained from theorems 1-2 $(\rho_\omega, \Phi_\omega)$ or proposition 3 $(\widetilde{\rho}_\omega, \widetilde{\Phi}_\omega)$ can be chosen for each value of $\|x_1\|_X$. Table 1 displays some comparison for examples 1 and 2: here, proposition 3 provides better results (optimal for these examples), but this is not true in general.

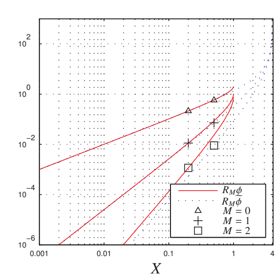
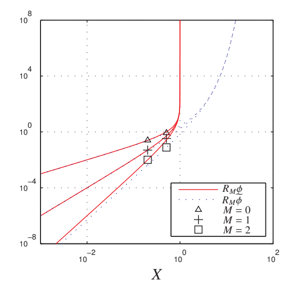
	Example 1 $\dot{x} = a_2 x^2 \quad (a_2 > 0)$	Example 2 $\dot{x} = b_2 x u \quad (b_2 > 0)$
ρ_ω	$1/(4Ta_2)$	$1/(\omega T b_2)$
Φ_ω	$2\rho(1 - \sqrt{1 - X/\rho})$	$X/(1 - X/\rho_\omega)$
$\widetilde{\rho}_\omega$	$1/(Ta_2) = 4\rho$	$+\infty$
$\widetilde{\Phi}_\omega$	$X/(1 - X/\widetilde{\rho})$	$X \exp(X/\widetilde{\rho}_\omega)$
remainders $R_M\phi$ and $R_M\widetilde{\phi}$		

Table 1: Comparison between results of theorems 1-2 ($R_M\Phi$: solid lines) and proposition 3 ($R_M\widetilde{\Phi}$, dotted lines). Illustrations are given for log-scales and $\rho_\omega = 1$.

6. Application to an epidemic model in a flock

6.1. Problem description

We consider a model of an epidemic spread in a flock (see [25] for a detailed description). The disease is characterized by a long and variable incubation period, during which individuals are infectious but cannot be detected. The flock population (assumed to be perfectly mixed) is described by population densities structured with respect to status (susceptible

S, infected I), age ($a \in [0, \bar{A}]$) and, for infected animals, with respect to infection load ($\theta \in [0, 1]$). Newly infected individuals are distributed along θ according to a probability density function Θ with support $[0, 1]$. The load then grows exponentially with time during the incubation period, which ends when θ reaches 1. Infected individuals for which $\theta = 1$ show detectable clinical signs and are removed from the flock. The population densities $S(t, a)$ and $I(t, a, \theta)$ are positive functions, governed by the following system of transport-reaction integro-differential PDE, for all $(a, \theta) \in [0, \bar{A}]\times[0, 1]$ and $t \in \mathbb{T} = [0, T]$,

$$\partial_t S + \partial_a S + \mu S = -\delta K(I) S, \tag{28}$$
$$\partial_t I + \partial_a I + \partial_\theta(c\,\theta I) + \mu I = \delta\,\Theta(\theta)\,K(I)\,S, \tag{29}$$

where $K(I)(t) = \int_0^{\bar{A}} \int_0^1 I(t, a, \theta)\,\mathrm{d}\theta\,\mathrm{d}a$ denotes the total number of infected individuals at time t. Positive parameters μ, δ, and c respectively denote the basic mortality rate, the transmission rate, and the infection load growth rate. The boundary conditions are $S(t, a = 0) = b(t)$ and $I(t, a = 0, \theta) = I(t, a, \theta = 0) = 0$, where the (positive) birth inflow b defines the input of the system. This means that the new born are susceptible, and that infection occurs after birth.

The issue here is to derive a simplified, easily tractable model for flock management policy design.

6.2. Linearized problem and well-posedness

The linearized problem defines a decoupled boundary control system with input $u = b$, state $x = (S, I)^T$, for $\mathbb{U} = \mathbb{R}$ and

$$\mathbb{X} = L^1\big(0, \bar{A}\big)\times L^1\big((0, \bar{A})\times(0, 1)\big) \text{ with } \big\|(S, I)^T\big\|_{\mathbb{X}} = \sqrt{\|S\|_1^2 + \|I\|_1^2}.$$

The associated strongly continuous semigroup on $\mathbb{X}$ is given by

$$V(t)\begin{pmatrix} S \\ I \end{pmatrix} = \begin{pmatrix} V_S(t) & 0 \\ 0 & V_I(t) \end{pmatrix}\begin{pmatrix} S \\ I \end{pmatrix},$$

where, denoting h^+ the Heaviside function,

$$[V_S(t)\,S](a) = S(a - t)\,e^{-\mu t}\,h^+(a - t), \tag{30}$$
$$[V_I(t)\,I](a, \theta) = I(a - t, e^{-ct}\theta)\,e^{-(\mu+c)t}\,h^+(a - t). \tag{31}$$

For all $t \in [0, T]$, $\|V(t)\|_{\mathcal{L}} = \sqrt{\|V_S(t)\|_1^2 + \|V_I(t)\|_1^2} \le e^{-\mu t}$ so that $\alpha = -\mu$ and $\beta = 1$. The mild solution $x_1 = (S_1, I_1)^T$ of the linearized problem with initial condition $x_{\text{ini}} = (S_0, I_0)^T \in \mathbb{X}$ is

$$S_1(t, a) = S_0(a - t)e^{-\mu t}h^+(a - t) + b(t - a)e^{-\mu a}h^+(t - a), \tag{32}$$
$$I_1(t, a, \theta) = I_0(a - t, e^{-ct}\theta)e^{-(\mu+c)t}h^+(a - t). \tag{33}$$

Obviously, if $b \in \mathcal{U}$, then x_1 is in $\mathcal{X} \cap C_0(\mathbb{T}, \mathbb{X})$.

6.3. Nonlinear system and bound estimates

The system is quadratic, with $P(x) = A_2(x, x)$ and $Q = 0$, where for all $x = (S, I)^T$ and $x' = (S', I')^T$ in $\mathbb{X}$,

$$A_2(x, x') = \frac{\delta}{2}\begin{pmatrix} -K(I)\,S' - K(I')\,S \\ (K(I)\,S' + K(I')\,S)\,\Theta \end{pmatrix}.$$

It follows that $a_2 = \|A_2\| = \delta/\sqrt{2}$ and $b_2 = 0$. The computation of $\gamma^\star$ in (11) provides

$$\gamma = \gamma^\star = \frac{1}{\sqrt{2}}\int_{\mathbb{T}} e^{-\mu t}\sqrt{1 + C_\Theta(e^{-ct})^2}\,\mathrm{d}t,$$

where C_Θ is the cumulative density function of Θ. Then the convergence results of theorem 1 and proposition 3 correspond to $\Phi_\omega(X) = 2\rho(1 - \sqrt{1 - X/\rho})$ and $\widetilde{\Phi}_\omega(X) = X/(1 - X/\widetilde{\rho})$ where

$$\rho = 1/(4\gamma a_2) \text{ and } \widetilde{\rho} = 1/(Ta_2).$$

This makes $\widetilde{\rho}$ a better bound than ρ as long as $T < 4\gamma$ and a worse one otherwise.

In order to compute the truncation error bound we set

$$r = \max_{t\in\mathbb{T}} \|S_1(t)\|_1/\|x_1\|_{\mathcal{X}}, \quad \kappa_S = \min(t, \bar{A}),$$
$$r' = \max_{t\in\mathbb{T}} \|I_1(t)\|_1/\|x_1\|_{\mathcal{X}}, \quad \kappa_I = \int_0^{\min(t,\bar{A})} C_\Theta(e^{-cs})\mathrm{d}s.$$

Equation (30) yields $\|S_1(t)\|_1\|I_1(t)\|_1 \le e^{-\mu t} r r' \|x_1\|_{\mathcal{X}}^2$ so that, from the definition of S_2 and I_2, we obtain

$$\|S_2(t)\|_1 \le \delta\kappa_S\,e^{-\mu t}\,r\,r'\|x_1\|_{\mathcal{X}}^2, \tag{34}$$
$$\|I_2(t)\|_1 \le \delta\kappa_I e^{-\mu t}\,r\,r'\|x_1\|_{\mathcal{X}}^2. \tag{35}$$

Therefore, for all $t \in \mathbb{T}$, $\|x_2(t)\|_1 \le \kappa\,\varphi_2\|x_1\|_{\mathcal{X}}^2$ with

$$\kappa = \min\Big(1,\ \sqrt{2}rr'\Big(\max_{t\in\mathbb{T}} e^{-\mu t}\sqrt{\kappa_S^2 + \kappa_I^2}\Big)/\gamma\Big).$$

Finally, from proposition 2, the error bound of the series truncated at order M is given by $\kappa R_M\Phi(\|x_1\|_{\mathcal{X}})$. In addition, the truncation error on I is shown to be bounded by $(\gamma_I/\gamma)\,\kappa R_M\Phi(\|x_1\|_{\mathcal{X}})$, where $\gamma_I = \frac{1}{\sqrt{2}}\int_{\mathbb{T}} e^{-\mu t}C_\Theta(e^{-ct})\,\mathrm{d}t$.

6.4. Numerical simulations

We consider an initial population size of 600 animals, among which 60 are infected. The basic mortality rate is $\mu = 0.5$ and the infection load growth rate is $c = 1$. The initial density of susceptible is the steady state distribution of the linearized problem. The initial infected density is a peak, with support $[0.25,\ 2.25] \times [e^{-2.5},\ e^{-1.5}]$, described by

$$I_0(a, \theta) = -\frac{C}{\theta}(\ln(\theta) + 1.5)(\ln(\theta) + 2.5)(a - 0.25)(2.25 - a),$$

where C is computed so that the initial number of infected animals is 60. The initial infection load distribution is

$$\Theta(\theta) = \frac{1}{\nu^\alpha\Gamma(\alpha)}(-\ln(\theta))^{\alpha-1}\theta^{1/\nu-1},$$

with $\alpha = 32$ and $\nu = 1/16$. We consider an infection rate $\delta = 1.5 \times 10^{-4}$. All the parameter values correspond to realistic situations.

In this case, the best convergence bound is $\widetilde{\rho} = 2357$ animals, whereas $\rho = 1500$. We consider a constant birth inflow equal to $S_0(0)$, in such a way that, in the absence of infection, the flock population age distribution would be time invariant, with a total population of 540 animals. Therefore the norm of the solution of the linearized system is $\|x_1\|_X = 543.3 < \max(\rho, \widetilde{\rho})$, with $r = 0.99$ and $r' = 0.11$. The best truncation error bound at order 2 on x (resp. I) is equal to 1.17 (resp. 0.66). It is even smaller at order 3 for which it is equal to 0.27 (resp. 0.15). This is consistent with the evolution of the total number of infected simulated and displayed in figure 2. The numerical truncation error at order 2 on I for this example is found to be 0.3, which is close to the computed upper bound.

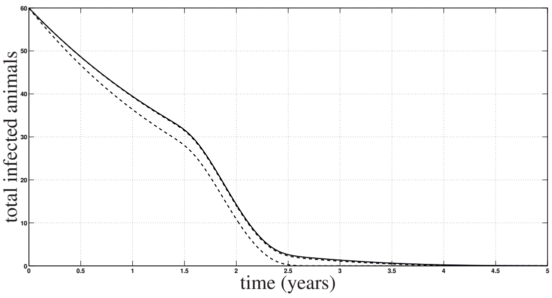


Figure 2: Total infected population $K(I)$ w.r.t. time. The solution of the nonlinear problem (solid line) is compared to the series expansions at order 1, 2 and 3 (dotted lines), with $\widetilde{\rho} = 2357$ and $\|x_1\|_X = 543.3$.

In this context, the approximation at order 2 is close enough, so that the system can be simplified as $S \simeq S_1 + S_2$ and $I \simeq I_1 + I_2$, as long the conditions of theorems 1-2 are met. This offers quite an interesting perspective from the control point of view. For instance, optimal herd management is usually performed using the birth input and the mortality rate (culling) as piecewise constant control input. Here, S_1, S_2, I_1 and I_2 are explicit functions of the mortality rate and depend linearly on the input flow, which allow the design of optimal management policies by solving simple constrained optimization problems.

When increasing δ to reach the limit case where $\widetilde{\rho} = \|x_1\|_X$ ($\delta \simeq 6.5 \times 10^{-4}$), the series seems to be still convergent on $[0, T = 4]$ and beyond, but 5 terms are needed to approximate the original system. Figure 3 displays trajectories for $\widetilde{\rho} = 295 < \|x_1\|_X$ ($\delta \simeq 12 \times 10^{-4}$). The series exhibits an extremely slow divergent behavior on $[0, T = 4]$, visible only around order 12. This divergent behavior increases very rapidly with δ and corresponds to the onset of a persistent epidemic. Hence, on this example, our method provides a conservative convergence bound estimate. Nonetheless, in practice, these bounds characterize situations where a low order truncation is possible and where the corresponding remainder estimate is accurate.

7. Conclusion

We have established a sufficient convergence criterion and truncation error bound for generalized Volterra series expansions of a class of infinite dimensional systems that are analytic in state and affine in input. We have also established the corresponding algorithms. Although these bounds are not optimal in general, the method provides a general framework that can

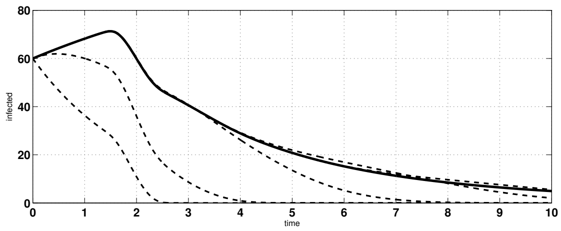


Figure 3: Total infected population $K(I)$ w.r.t. time (solid line) and series expansions at order 1, 2, 5, 9 and 12 (dashed lines), with $\rho = 295 < \|x_1\|_X$.

be adapted to specific systems. This was illustrated in the case of quadratic-bilinear systems. Finally, a simulation example in animal epidemiology was presented that demonstrate the accuracy and utility of the method.

References

[1] V. Volterra, Theory of Functionnals and of Integral and Integro-Differential Equations, Dover Publications, 1959.

[2] R. W. Brockett, Volterra series and geometric control theory, Automatica 12 (1976) 167–176.

[3] E. G. Gilbert, Functional expansions for the response of nonlinear differential systems, IEEE T Automat Contr 22 (1977) 909–921.

[4] M. Fliess, M. Lamnabhi, F. Lamnabhi-Lagarrigue, An algebraic approach to nonlinear functional expansions, IEEE T Circuits Syst 30 (8) (1983) 554–570.

[5] A. Isidori, Nonlinear Control Systems, Springer, 3rd ed. edn., 1995.

[6] W. J. Rugh, Nonlinear System Theory, The Volterra/Wiener approach, The Johns Hopkins University Press, Baltimore, 1981.

[7] P. E. Crouch, P. C. Collingwood, The observation space and realizations of finite Volterra series, SIAM J Control Optim 25 (2) (1987) 316–333.

[8] M. Schetzen, The Volterra and Wiener Theories of Nonlinear Systems, Wiley-Interscience, 1989.

[9] R. De Figueiredo, T. Dwyer, A best approximation framework and implementation for simulation of large-scale nonlinear systems, IEEE Trans. Circuits Syst. 27 (11) (1980) 1005–1014.

[10] G. Chen, Representation, Approximation, and Identification, in: Circuits & Filters Handbook 3e, CRC Press, –, 2009.

[11] R. W. Brockett, Convergence of Volterra Series on Infinite Intervals and Bilinear Approximations, in: V. Lakshmikanthan (Ed.), Nonlinear Systems and Applications, Academic Press, 39–46, 1977.

[12] F. Lamnabhi-Lagarrigue, Analyse des Systemes Non Lineaires, Editions Hermes, iSBN 2-86601-403-0, 1994.

[13] S. Boyd, L. Chua, Fading memory and the problem of approximating nonlinear operators with Voltera series, IEEE T Circuits Syst 32 (11) (1985) 1150–1161.

[14] X. J. Jing, Z. Q. Lang, S. A. Billings, Magnitude bounds of generalized frequency response functions for nonlinear Volterra systems described by NARX model, Automatica 44 (2008) 838–845.

[15] Z. K. Peng, Z. Q. Lang, On the convergence of the Volterra-series representation of the Duffing's oscillators subjected to harmonic excitations, J Sound Vib 305 (2007) 322–332.

[16] F. Bullo, Series expansions for analytic systems linear in control, Automatica 38 (2002) 1425–1432.

[17] M. Thitsa, W. S. Gray, On the Radius of Convergence of Interconnected Analytic Nonlinear Input-Output Systems, SIAM J. Control and Optimization 50 (5) (2012)) 2786–2813.

[18] T. Hélie, B. Laroche, Computation of convergence bounds for Volterra series of linear-analytic single-input systems, IEEE T Automat Contr 56 (9) (2011) 12p.

[19] R. F. Curtain, H. J. Zwart, An introduction to Infinite-dimensional Linear Systems Theory, Springer, 1995.

[20] A. Pazy, Semigroups of Linear Operators and Applications to Partial Differential Equations, vol. 44 of *Applied mathematical sciences*, Springer, 1983.

[21] S. Boyd, L. O. Chua, C. A. Desoer, Analytical Foundations of Volterra Series, IMA J Math Control I 1 (1984) 243–282.

[22] P. Flajolet, R. Sedgewick, Analytic Combinatorics, Cambridge University Press, 2009.

[23] T. Hélie, B. Laroche, On the convergence of Volterra series of finite dimensional quadratic MIMO systems, International Journal of Control 81-3 (2008) 358–370.

[24] T. Hélie, B. Laroche, Series expansion of infinite dimensional linear-analytic systems: application to a nonlinear Euler-Bernoulli Beam. Technical Report 2013-1, Tech. Rep., INRA, 2013.

[25] A. Perasso, B. Laroche, Y. Chitour, S. Touzeau, Identifiability analysis of an epidemiological model in a structured population, J Math Anal Appl 374 (1) (2011) 154–165.