Volume 1, Number 1, 2001 ISSN 1562-8353

Nonlinear Dynamics and Systems Theory

An International Journal of Research and Surveys

EDITOR-IN-CHIEF A.A.MARTYNYUK
S.P.Timoshenko Institute of Mechanics
National Academy of Sciences of Ukraine, Kiev, Ukraine

REGIONAL EDITORS

Australia, New Zeland and South East Asia
K.L.TEO, Kowloon, Hong Kong

Europe
B.AULBACH, Augsburg, Germany
P.BORNE, Lille, France

Japan
S.OMATU, Osaka, Japan

USA, Central and South America
Y.-H.CHEN, Atlanta, USA
S.SIVASUNDARAM, Daytona Beach, USA

Canada
J.MULDOWNEY, Edmonton, Canada

Academic Periodical Press

Nonlinear Dynamics and Systems Theory

An International Journal of Research and Surveys

EDITOR-IN-CHIEF A.A. MARTYNYUK

S.P. Timoshenko Institute of Mechanics
National Academy of Sciences of Ukraine, Kiev, Ukraine

HONORARY EDITORS

VANGIPURAM LAKSHMIKANTHAM, *Melbourne, USA*
YURY A. MITROPOLSKY, *Kiev, Ukraine*

REGIONAL EDITORS

Australia, New Zeland and South East Asia
K.L. TEO, Kowloon, Hong Kong

Europe
B. AULBACH, Augsburg, Germany
P. BORNE, Lille, France

Japan
S. OMATU, Osaka, Japan

USA, Central and South America
Y.-H.CHEN, Atlanta, USA
S.SIVASUNDARAM, Daytona Beach, USA

Canada
J.MULDOWNEY, Edmonton, Canada

EDITORIAL BOARD

Azbelev, N.V. (Russia)
Corduneanu, C. (USA)
Dauphin-Tanguy, G. (France)
Duan Li (Hong Kong)
Dzhalalow, J.H. (USA)
Freedman, H.I. (Canada)
Hatvani, L. (Hungary)
Izobov, N.A. (Belarussia)
Kats, I.Ya. (Russia)
Kovalev, A.M. (Ukraine)
Kuepper, T. (Germany)
Larin, V.B. (Ukraine)
Leitmann, G. (USA)
Mawhin, J. (Belgium)
Miani, S. (Italy)
Miladzhanov, V.G. (Uzbekistan)
Noldus, E. (Belgium)
Pilipchuk, V.N. (Ukraine)
Pyatnitskiy, E. (Russia)
Richard, J.P. (France)
Shaw, M.D. (USA)
Šiljak, D.D. (USA)
Spinadel, V.W. (Argentina)
Sree Hari Rao, V. (India)
Stavroulakis, I.P. (Greece)
Tonkov, E.L. (Russia)
Trigiante, D. (Italy)
Tzou, H.S. (USA)
Vincent, T. (USA)
Weiyong Yan (Australia)
Yagawa, G. (Japan)
Zhukovskii, V.I. (Russia)

© 2001, Informath Publishing Group ISSN 1562-8353 Printed in Ukraine

No part of this Journal may be reproduced or transmitted in any form or by any means with out permission from Informath Publishing Group.

NONLINEAR DYNAMICS AND SYSTEMS THEORY
An International Journal of Research and Surveys

Volume 1 Number 1 2001

CONTENTS

NONLINEAR DYNAMICS AND SYSTEMS THEORY

An International Journal of Research and Surveys

REGIONAL EDITORS

Australia, New Zeland and South East Asia: K.L. TEO,
Department of Applied Mathematics,
The Hong Kong Polytechnic University
Hung Horn, Kowloon, HONG KONG
E-mail: mateokl@polyu.edu.hk

Europe: B. AULBACH, Institut filr
Mathematik, Universitat Augsburg,
D-86135 Augsburg, GERMANY
E-mail: aulbach@math.uni-augsburg.de
and
P. BORNE, Ecole
Centrale de Lille,
BP 48, 59651 Villeneuve d'Ascq Cedex, FRANCE
E-mail: p.borne@ec-lille.fr or p.borne@ieee.org

Japan: S. 0MATU,
Department of Computer and Systems Sciences, College of Engineering,
Osaka Prefecture University, Sakai, Osaka 599-8531, JAPAN
E-mail: omatu@cs.osakafu-u.ac.jp

USA, Central and South America.: Y.-H. CHEN,
The George W. Woodruff School of Mechanical Engineering,
Georgia Institute of Technology, Atlanta, GA 30332-0405, USA
E-mail: yehwa.chen@me.gatech.edu
and
S. SIVASUNDARAM,
Department of Mathematics, Embry-Riddle Aeronautical University,
Daytona Beach Florida, FL 32114, USA
E-mail: Seenithl@aol.com

Canada: J.S. MULDOWNEY, Department of
Mathematics, University of Alberta,
Edmonton, CANADA E-mail:
jim.muldowney@ualberta.ca

PUBLICATION AND SUBSCRIPTION INFORMATION

The NONLINEAR DYNAMICS AND SYSTEMS THEORY (ISSN 1562-8353) is published four times a year. For 2001, volume 1 is scheduled for publication. Orders, claims, and product inquires: please contact the Regional Editors Office nearest you or the Editor-in-Chief Office at the S.P.Timoshenko Institute of Mechanics NAS of Ukraine, Nesterov str.,3, 03057, Kiev-57, MSP 680, Ukraine. Tel: ++38-044-446-6140; Fax: ++38-044-446-1939.
E-mail:anchern@stability. kiev.ua

PREFACE TO THE JOURNAL

Nonlinear Dynamics and Systems Theory represent a new direction of investigation during the last two decades. The primary goal of the Theory of Systems is to develop unified methods of mathematical modelling of processes and phenomena, in nature and society, so that to establish the conditions of their dynamics and to control them, if necessary. This aim can be attained by a profound and comprehensive analysis of various phenomena occurring in the world that, as a rule, are nonlinear, nonstationary, with random and/or parametric perturbations, taking place at changing of the environment parameters, in near-Earth area or in space.

The hierarchy of theories, at different levels appearing in the result of analytical, qualitative, numerical or experimental studies of such type systems is a corner-stone of the Theory of Systems being developed nowadays.

The constructive results of Nonlinear Dynamics have created necessary premises for the development of the mathematical Theory of Systems characterized by at least two exceptional features:

– the higher degree of generality of the mathematical modelling technique for real processes and phenomena (continuous, discrete, impulsive, system with delay, etc.), and

– "available" transparency for returning from general results to various concrete phenomena and technologies being partial interpretations.

The new Journal presented to the attention of the readers and investigators encompasses a wide branch of natural sciences falling under the scope of Nonlinear Dynamics and Systems Theory:

* Analysis of uncertain systems
* Bifurcations and instability in dynamical behaviors
* Celestial mechanics, variable mass processes, rockets
* Control of chaotic systems
* Controllability, observability, and structural properties
* Deterministic and random vibrations
* Differential games
* Dynamical systems on manifolds
* Dynamics of systems of particles
* Hamilton and Lagrange equations
* Hysteresis
* Identification and adaptive control of stochastic systems
* Modelling of real phenomenon by ODE, FDE and PDE
* Nonlinear boundary problems
* Nonlinear control systems, guided systems
* Nonlinear dynamics in biological systems
* Nonlinear fluid dynamics
* Nonlinear oscillations and waves
* Nonlinear stability in continuum mechanics
* Non-smooth dynamical systems with impacts or discontinuities
* Numerical methods and simulation
* Optimal control and applications
* Qualitative analysis of systems with aftereffect

- * Robustness, sensitivity and disturbance rejection
- * Soft computing: artificial intelligence, neural networks, fuzzy logic, genetic algorithms, etc.
- * Stability of discrete systems
- * Stability of impulsive systems
- * Stability of large-scale power systems
- * Stability of linear and nonlinear control systems
- * Stability theory of intelligence media
- * Stochastic approximation and optimization
- * Symmetry in mechanics.

The aim of the new Journal of Nonlinear Dynamics and Systems Theory is to publish the most significant scientific results in the field obtained at the leading mathematical centres of Europe, in the countries of the former USSR and the rest of the world.

The papers will be thoroughly reviewed by the Regional Editors and the members of the Editorial Board of the Journal.

Our Journal is open to all scientists and experts who consider it suitable for new contributions to Nonlinear Dynamics and Systems Theory and aims at the presentation of such new results to the world scientific community.

I have the pleasure to express my sincere gratitude to Professor, Dr.V.Lakshmikantham for his kind support in settng up new journal under methodological and academic auspices of International Federation of Nonlinear Analysts (USA).

Also, I am thankful to all people who have helped us in one way or another while starting this Journal.

Professor A.A. Martynyuk

Nonlinear Dynamics and Systems Theory, 1(1) (2001) 1–22

Consistent Lyapunov Methodology: Non-Differentiable Non-Linear Systems

Ly.T. Gruyitch

University of Technology Belfort - Montbeliard,
Site de Belfort, 90010 Belfort Cedex, France

Received: February 15, 2000; Revised: June 28, 2000

This invited paper is dedicated to late Professor Wolfgang Hahn and to late Professor Jose P. LaSalle, who contributed fundamentally to the stability theory and supported author's early research on stability domains.

Abstract: The consistent Lyapunov methodology enables us, after its single application, to solve completely the asymptotic (or, exponential) stability problem, to construct a system Lyapunov function and to determine accurately the domain of asymptotic stability. This is achieved in the paper for invariant sets of non-differentiable time-varying non-linear systems. The results (proved in details) presentthe necessary and sufficient conditions for: asymptotic stability, for a determination of a system Lyapunov function and for a set to be the asymptotic stability domain. They are not expressed in terms of existence of a system Lyapunov function. They determine well the procedure how to resolve all the relevant problems.

Keywords: *Asymptotic stability domains; Lyapunov method; Lyapunov functions; non-linear systems; sets; uniform asymptotic stability.*

Mathematics Subject Classification (2000): 34C35, 34D05, 34D20, 34D45, 34H05, 54H20, 93C10, 93C15, 93C50, 93C60, 93D05, 93D20, 93D30.

1 Introduction

The fundamental Lyapunov method [1] is based on two different methodologies, one for time-invariant linear systems and another one for all other systems. The former enables us to effectively apply the method and to get a definite result after its single application. The latter, which will be called the *classical Lyapunov methodology (for non-linear systems)*, does not. The latter lesses us to face two crucial problems unsolved: a) how

© 2001 Informath Publishing Group. All rights reserved.

to construct a system Lyapunov function and b) how to determine the exact asymptotic stability domain. The classical Lyapunov methodology (for non-linear systems) starts with a trial to guess a suitable choice of a positive definite function $v(.)$. Its application continues with the negative (semi-) definiteness test of the total time derivative of $v(.)$ along system motions. The theorems established for time-varying non-linear systems have been expressed only in terms of *existence* of a Lyapunov function $v(.)$ $[u(.)]$ rather than to clarify how to find it for a given non-linear system. If the weak inequlaity in the condition on the Lyapunov function derivative is replaced by the equality, then they do not provide any guideline how to chose a function $p(.)$ in $v^{(1)}(.) = -p(.)$ {or equivalently, in $u^{(1)} = -p(.)[1 - u(.)]$}. Once we understand this, it appears clear that we meet two subproblems: a) what are properties of the system and of the function $p(.)$ to garantee existence of a solution to the differential equation, and b) what are, relative to a selected $p(.)$, the necessary and sufficient conditions for a solution $v(.)$ $\{u(.)\}$, respectively, to guarantee uniform asymptotic stability of an invariant set and/or to determine accurately its domain of uniform asymptotic stability. The former problem is purely mathematical problem that is not related to the stability issue. However, the latter one is crucial for solving the stability problems.

Bhatia [2,3], Bhatia and Lazer [4], Bhatia and Szegö [5], Corne and Rouche [6], Hajek [7,8], Ladde *et al.* [9], Ladde and Leela [10,11], Lakshmikantham and Leela [12,13], LaSalle [14], Yoshizawa [15 − 18] and Zubov [19] extended the classical Lyapunov methodology from the analysis of stability properties of a state and of a motion to the analysis of various stability properties of sets.

A novel Lyapunov methodology for asymptotic stability analysis of the zero equilibrium state of non-linear time-invariant systems was discovered and established in [20 − 32]. It was extended to the asymptotic stability analysis of the zero equilibrium state of non-linear time-varying systems in [33 − 35], as well as of constant sets of non-linear time-invariant systems in [36] and of those time-varying in [37]. It has been aimed at solving the open stability problems. The methodology starts with a determination of a functional family $L(.)$ $[E(.)]$ of functions $p(.)$ that can be used to generate a function $v(.)$ [or, $u(.)$]. An important feature of the novel Lyapunov methodology and of the functional family as its tool is that it permits an *arbitrary selection* of a function $p(.)$ in the family in order to determine exactly a system Lyapunov function. Its another important characteristic is that it provides stability conditions that are not expressed in terms of existence of a system Lyapunov function. The methodology terminates by verifying the properties of $v(.)$ [or, $u(.)$], which are both necessary and sufficient for asymptotic stability of the zero state (or, of a time-invariant set), and/or for a set N to be the domain of its asymptotic stability. This methodolgy is consistent with Lyapunov's original methodlogy for time-invariant linear systems and has been called the *consistent Lyapunov methodology (for both linear and non-linear systems)* [37,38].

The paper [38] further broadened the consistent Lyapunov methodology by presenting the complete solutions for uniform asymptotic stability of invariant sets of time-varying non-linear systems with differentiable motions. The class of systems will be enlarged in what follows by allowing for system motions to be non-differentiable.

The structure is the following: there are eight sections, an appendix and a list of references in the paper. A brief explanation of the notation is available in the next section. The relaxed smoothness properties of the systems are explained in Section 3 that is on the system description. Various stability domains are defined in Section 4. Functional families $L(.)$ and $E(.)$ are introduced in Section 5. The key part are Section 6, which contributes with new criteria for asymptotic stability domains of the sets, and Section 7 that presents the analogous conditions for uniform asymptotic stability of the sets. This order of the Sections 7 and 8 eases significantly their proofs. The criteria expose the consistent Lyapunov methodology. The conclusions compose Section 8. Appendix preceeds the list of references, which terminates the paper.

2 Notation

Capital italic Roman letters are used for sets, lower case block Roman characters for vectors, Greek letters and lower case italic letters denote scalars except for the empty set $\emptyset$ and subscripts. The boundary, interior and closure of a set A are designated by ∂A, $\mathrm{In}\, A$ and $\mathrm{Cl}\, A$, respectively, where A is time-invariant set. If $A(.)\colon R \to 2^{R^n}$ is a set-valued function then its instantaneous set value $A(t)$ at an arbitrary time $t \in R$ will be called a time-varying set $A(t)$. Let $\|.\|\colon R^n \to R_+$ be Euclidean norm on R^n, where $R_+ = [0,\infty) = \{\xi\colon \xi \in R,\ 0 \le \xi < \infty\}$. An initial time $t_0 \in R_i$, where $R_i = (\sigma,\infty)$, $\sigma \in [-\infty,\infty)$. It determines $R_0 = [t_0,\infty)$. Let $R^+ = (0,\infty) = \{\xi\colon \xi \in R,\ 0 < \xi < \infty\}$.

A set J, $J \subset R^n$, will be a compact connected invariant set of the system with the boundary ∂J being also an invariant set. Its time-varying neighbourhood at time $t \in R$ will be denoted by $A(t;J)$, $M(t;J)$ or $S(t;J)$, and its δ-neighbourhood will be designated by $B_\delta(J)$, where $\delta \in R^+$ and $B_\delta(J) = \{x\colon \rho(x,J) < \delta\}$ with the distance function $\rho(.)\colon R \times 2^{R^n} \to R_+$ induced by $\|.\|$ as $\rho(x,J) = \inf\{\|x-y\|\colon y \in J\}$. Notice that $J \subset A(t;J)$, $\forall\, t \in R$, and $J \subset B_\delta(J)$. Besides, $M_m(R_i;J) = \cap[M(t;J)\colon t \in R_i]$, $M_M(R_i;J) = \cup[M(t;J)\colon t \in R_i]$ and $S(R_i;J) = \cap[S(t;J)\colon t \in R_i] = S_m(R_i;J)$. The distance between sets $M_1(t;J)$ and $M_2(t;J)$ at time t is the instantaneous value of a set-distance function $\rho(.)$ at time t, $\rho(.)\colon 2^{R^n} \times 2^{R^n} \to R_+$, where $\rho[M_1(t;J),M_2(t;J)] = \max\{\sup[\rho(x,M_1(t;J))\colon x \in M_2(t;J)],\ \sup[\rho(y,M_2(t;J))\colon y \in M_1(t;J)]\}$.

Let $t_k \to \tau$ as $k \to \infty$, where in special cases of an unbounded value of t:

$$\begin{aligned} &t_k < \tau \quad \text{if} \quad \tau = \infty,\\ &t_k > \tau \quad \text{if} \quad \tau = -\infty. \end{aligned}$$

A non-empty set-valued function $M(.)\colon R \times 2^{R^n} \to 2^{R^n}$ is continuous at $\tau \in R$ if and only if for every $\varepsilon \in R^+$ there is $L \in \{1,2,\dots\}$, $L = L(\varepsilon;\tau)$, such that $k > L$ implies $d\{M(t_k;J),M(\tau;J)\} < \varepsilon$. It is continuous on $R_{(.)}$ if and only if it is continuous at every $t \in R_{(.)}$, which is denoted by $M(t;J) \in C(R_{(.)})$. The time-varying set $M(t;J)$ is non-empty, connected and/or compact on $R_{(.)}$ if and only if it is non-empty, connected and/or compact at every $t \in R_{(.)}$, respectively.

$D_a(t;J)$, $D_s(t;J)$ and $D(t;J)$ will represent the (instantaneous) domain of attraction of the set J at time t, its domain of stability at time t and its domain of asymptotic stability at the same time t, respectively. Their definitions are given in Section 4.

Let $\boldsymbol{x}(.;t_0,x_0)$ be motion (solution) of a system through x_0 at an initial time t_0, and let its vector value at time t be $x(t)$, $x(t) = \boldsymbol{x}(t;t_0,x_0)$.

If a function $v(.)\colon R \times R^n \times 2^{R^n} \to R$ is continuous on $R \times R^n$ then we will use its right-hand Dini derivative $D^+v(t,x;J)$ taken along system motions and determined at $(t,x) \in R \times R^n$ with J being fixed:

$$D^+v(t,x;J) = \limsup\left\{\frac{v[t+\theta,\boldsymbol{x}(t+\theta;t,x);J] - v(t,x;J)}{\theta}\colon\ \theta \to 0^+\right\}.$$

Let $\zeta \in R^{+}$ and $p(.)$, $[v(.)]\colon R \times R^{n} \times 2^{R^{n}} \rightarrow R$. Then $P_{\zeta}(t;J)$, $[V_{\zeta}(t;J)]$ is the largest open connected neighbourhood of J at time $t \in R$ such that $p(t,x;J) < \zeta$, $[v(t,x;J) < \zeta]$ for every $x \in P_{\zeta}(t;J)$, $[V_{\zeta}(t;J)]$.

K is the family defined by Hahn [39] of all the comparison functions $\varphi(.)\colon R_{+} \rightarrow R_{+}$ strictly increasing, continuous and vanishing at the origin:

$$\varphi(\zeta_1) < \varphi(\zeta_2), \quad 0 \leq \zeta_1 < \zeta_2, \quad \varphi(\zeta) \in C(R_{+}), \quad \varphi(0) = 0.$$

3 System Description

Time-varying non-linear systems studied herein in general are described by (1),

$$\frac{dx(t)}{dt} = f(t,x(t)), \quad x(.)\colon R \rightarrow R^{n}, \quad f(.)\colon R \times R^{n} \rightarrow R^{n}, \tag{1}$$

and by one of the following features:

Weak smoothness property

(i) There is an open continuous connected neighbourhood $S(t;J)$ of J, $S(t;J) \subseteq R^{n}$, for every $t \in R_i$, such that $S(R_i;J) = \cap[S(t;J)\colon t \in R_i]$ is also open connected neighbourhood of J, and for every $(t_0,x_0) \in R_i \times S(t_0;J)$ the following holds:

a) system (1) has a unique solution $\boldsymbol{x}(.;t_0,x_0)$ through x_0 at t_0 on the largest interval of its existence I_0, $I_0 = I_0(t_0,x_0)$, and

b) $\boldsymbol{x}(t;t_0,x_0)$ is defined and continuous in (t,t_0,x_0).

(ii) For every $(t_0,x_0) \in R_i \times [R^{n} - \mathrm{Cl}\, S(t_0;J)]$ every motion $\boldsymbol{x}(.;t_0,x_0)$ of system (1) is continuous in $t \in I_0$.

Strong smoothness property

(i) System (1) obeys the weak smoothness property.

(ii) If the boundary $\partial S(t;J)$ of $S(t;J)$ is non-empty at any time $t \in R_i$ then every motion of system (1) passing through $x_0 \in \partial S(t_0;J)$ at $t_0 \in R_i$ satisfies $\inf\{\rho[\boldsymbol{x}(t;t_0,x_0),J]\colon t \in I_0\} > 0$ for every $(t_0,x_0) \in R_i \times \partial S(t_0;J)$.

Any of the above system smothness properties permits non-differentiability of system motions $\boldsymbol{x}(t;t_0,x_0)$ with respect to (t,t_0,x_0). This makes the difference between what follows and the results established in [38]. The smoothness properties are expressed directly in terms of smoothness of system motions rather than indirectly via smoothness of the function $f(.)$ for the following reasons. Dealing with physical systems we can often conclude on smoothness of their motions for physical reasons. We know only sufficient mathematical conditions on $f(.)$, which guarantee smoothness of system motions. Such conditions can be too conservative.

4 Asymptotic Stability Domains

The notions of various stability domains of states [19, 21 – 28, 32, 39 – 46], and of sets [36], of time-invariant systems were broadened to stability domains of sets of time-varying systems in [38] as follows.

Definition 4.1 A set J of system (1) has:

(a) *the domain of attraction at* $t_0 \in R$ denoted by $D_a(t_0; J)$, $D_a(t_0; J) \subseteq R^n$, if and only if:

1) for every $\zeta \in R^+$, there exists $\tau = \tau(t_0, x_0; \zeta; J) \in R_+$ such that

$$\rho[\boldsymbol{x}(t; t_0, x_0), J] < \zeta \quad \text{for all} \quad t \in (t_0 + \tau, \infty)$$

is valid provided only that $x_0 \in D_a(t_0; J)$,

2) the set $D_a(t_0; J)$ is a neighbourhood of J.

(b) *the domain* $D_a(R_i; J)$ *of uniform attraction on* R_i, $D_a(R_i; J) \subseteq R^n$, if and only if 1)$-$4) hold:

1) it has the domain $D_a(t_0; J)$ of attraction at every $t_0 \in R_i$,

2) $\cap[D_a(t; J) \colon t \in R_i]$ is a neighbourhood of J,

3) $D_a(R_i; J) = \cap[D_a(t; J) \colon t \in R_i]$,

4) the minimal $\tau(t_0, x_0; \zeta; J)$ obeying 1) of (a) and denoted by $\tau_m(t_0, x_0; \zeta; J)$ obeys

$$\sup[\tau_m(t_0, x_0; \zeta; J) \colon t_0 \in R_i] < +\infty \quad \text{for every} \quad (x_0, \zeta) \in D_a(R_i; J) \times R^+.$$

The expression "on R_i" is to be omitted if and only if $R_i = R$. Then and only then $D_a(R_i; J)$ will be denoted by $D_a(J)$, $D_a(J) = D_a(R; J)$.

Definition 4.2 A set J of system (1) has:

(a) *the domain of stability at* $t_0 \in R_i$ denoted by $D_s(t_0; J)$, $D_s(t_0; J) \subseteq R^n$, if and only if:

1) for every $\varepsilon \in R^+$ the motion $\boldsymbol{x}(.; t_0, x_0)$ satisfies $\rho[\boldsymbol{x}(t; t_0, x_0), J] < \varepsilon$ for all $t \in R_0$ provided only that $x_0 \in D_s(t_0, \varepsilon; J)$,

2) the set $D_s(t_0, \varepsilon; J)$ is a neighbourhood of J for every $\varepsilon \in R^+$,

3) the set $D_s(t_0; J)$ is the union of all the sets $D_s(t_0, \varepsilon; J)$ over $\varepsilon \in R^+$:

$$D_s(t_0; J) = \cup[D_s(t_0, \varepsilon; J) \colon \varepsilon \in R^+].$$

(b) *the domain* $D_s(R_i; J)$ *of uniform stability on* R_i if and only if:

1) J has the domain of stability $D_s(t_0; J)$ at every $t_0 \in R_i$,

2) $\cap[D_s(t, \varepsilon; J) \colon t \in R_i]$ is a neighbourhood of J for any $\varepsilon \in R^+$,

3) $D_s(R_i; J) = \cap[D_s(t; J) \colon t \in R_i]$.

The expression "on R_i" is to be omitted if and only if $R_i = R$. Then and only then $D_s(R_i; J)$ will be denoted by $D_s(J)$, $D_s(J) = D_s(R; J)$.

Definition 4.3 A set J of system (1) has:

(a) *the domain of asymptotic stability at* $t_0 \in R_i$ denoted by $D(t_0; J)$, $D(t_0; J) \subseteq R^n$, if and only if it has both $D_a(t_0; J)$ and $D_s(t_0; J)$, and $D(t_0; J) = D_a(t_0; J) \cap D_s(t_0; J)$.

(b) *the domain* $D(R_i; J)$ *of uniform asymptotic stability on* R_i if and only if it has both $D_a(R_i; J)$ and $D_s(R_i; J)$, and $D(R_i; J) = D_a(R_i; J) \cap D_s(R_i; J)$.

The expression "on R_i" is to be omitted if and only if $R_i = R$. Then and only then $D(R_i; J)$ will be denoted by $D(J)$, $D(J) = D(R; J)$.

Qualitative features of the stability domains of an invariant set J of system (1) are discovered in Appendix. They are used for the proofs of the results of the paper (Section 6).

5 Families $L(.)$ and $E(.)$ of Functions $p(.)$

A fundamental problem which has not been solved in the classical Lyapunov methodology is that of the generation of a system Lyapunov function. The theorems based on the classical Lyapunov methodology (including also the converse theorems) express conditions on the Lyapunov function derivative in the inequality form: $v^{(1)}(.) \leq -p(.)$. They do not specify how to select the function $p(.)$ in order to get a system Lyapunov function obeying the weak inequality that may be replaced by the equality in order to ease the function generation. Various forms of families $P(.)$ and $P^1(.)$ of functions $p(.)$ were introduced in [21 − 33, 36, 37] in order to generate Lyapunov functions $v(.)$ from $v^{(1)}(.) = -p(.)$ {or, to determine Lyapunov functions $u(.)$ as solutions of $u^{(1)}(.) = -[1-u(.)]p(.)$} in the framework of time invariant systems, and for time-varying systems in [34, 35, 38]. They will be replaced by families $L(.)$ and $E(.)$ of functions $p(.)$ in the sequel. One role of these families is to separate the problem of existence of the differential equation solution from the stability problem. Another their role is to enable an exact determination of a family of system Lyapunov functions [47, 48].

Definition 5.1 A *function* $p(.)\colon R_i \times R^n \times 2^{R^n} \to R$ *belongs to the family* $L(R_i, S; f; J)$ if and only if:

1) $p(.)$ is continuous on $R_i \times S(t;J)\colon p(t,x;J) \in C[R_i \times S(t;J)]$;
2) the equations (2) with (2a) taken along motions of system (1),

$$D^+ v(t,x;J) = -p(t,x;J), \tag{2a}$$

$$v(t,x;J) = 0, \quad \forall\, x \in \partial J, \quad \forall\, t \in R_i, \tag{2b}$$

have a solution $v(.)\colon R_i \times R^n \times 2^{R^n} \to R$ that is continuous in $(t,x) \in R_i \times \operatorname{Cl} B_\mu(J)$ for an arbitrarily small $\mu \in R^+$, $\mu = \mu(f,p;J)$, and which obeys (3) for some $w_\mu(x;J) \in C[\operatorname{Cl} B_\mu(J)]$:

$$v(t,x;J) \leq w_\mu(x;J), \quad \forall\, (t,x) \in R_i \times [\operatorname{Cl} B_\mu(J) - \operatorname{In} J]; \tag{3}$$

3) the following holds for any $\zeta \in R^+$ fulfilling $\operatorname{Cl} P_\zeta(t;J) \subset S(t;J)$ for all $t \in R_i$:

$$\min\{p(t,x;J)\colon (t,x) \in R_i \times [S(t;J) - P_\zeta(t;J)]\} = \alpha, \qquad \alpha = \alpha(\zeta;p) \in R^+.$$

Definition 5.2 A *function* $p(.)\colon R_i \times R^n \times 2^{R^n} \to R$ *belongs to the family* $E(R_i, S; f; J)$ if and only if:

1) $p(.)$ is continuous on $R_i \times S(t;J)\colon p(t,x;J) \in C[R_i \times S(t;J)]$;
2) the equations (4) with (4a) taken along motions of system (1),

$$D^+ u(t,x;J) = -[1 - u(t,x;J)]p(t,x;J), \tag{4a}$$

$$u(t,x;J) = 0, \quad \forall\, x \in \partial J, \quad \forall\, t \in R_i, \tag{4b}$$

have a solution $u(.)\colon R_i \times R^n \times 2^{R^n} \to R$ that is continuous in $(t,x) \in R_i \times \operatorname{Cl} B_\mu(J)$ for an arbitrarily small $\mu \in R^+$, $\mu = \mu(f,p;J)$, and which obeys (5) for some $w_\mu(x;J) \in C[\operatorname{Cl} B_\mu(J)]$:

$$u(t,x;J) \le w_\mu(x;J), \quad \forall\,(t,x) \in R_i \times [\operatorname{Cl} B_\mu(J) - \operatorname{In} J]; \tag{5}$$

3) the following holds for any $\zeta \in R^+$ fulfilling $\operatorname{Cl} P_\zeta(t;J) \subset S(t;J)$ for all $t \in R_i$:

$$\min\{p(t,x;J)\colon (t,x) \in R_i \times [S(t;J) - P_\zeta(t;J)]\} = \alpha, \qquad \alpha = \alpha(\zeta;p) \in R^+.$$

Comment 5.1 Notice that $p(.) \in L(R_i,S;f;J)$ if and only if $p(.) \in E(R_i,S;f;J)$. If $p(.) \in L(R_i,S;f;J)$, hence $p(.) \in E(R_i,S;f;J)$, then solutions $v(.)$ and $u(.)$ to (2) and (4), respectively, are interrelated by (6),

$$u(t,x;J) = 1 - \exp[-v(t,x;J)], \tag{6}$$

which was pointed out by Vanelli and Vidyasagar [45]. Besides, $u(t,x;J) = 0$ if and only if $v(t,x;J) = 0$, and $u(t,x;J) \to 1$ if and only if $v(t,x;J) \to \infty$.

Comment 5.2 No stability condition is imposed on the system and no definiteness requirement is imposed on $p(.)$, $v(.)$ and $u(.)$ in Definition 5.1 and Definition 5.2. Therefore, $L(R_i,S;f;J)$ and $E(R_i,S;f;J)$ are not dependent on a stability property of the system.

6 Domains of Asymptotic Stability Properties of Invariant Sets

The notions of a positive definite function and of a decrescent function relative to a set will be used in the usual sense (c.f. Lyapunov [1], Bhatia and Szegö [5], Yoshizawa [18], Zubov [19], Gruyitch [38], Hahn [39], Krasovskii [49], Rouche *et al.* [50]). Let $\varphi_i(.)$ be a comparison function from the class K defined by Hahn [39]: $\varphi_i(.) \in K$, $i = 1,2$.

Definition 6.1 A *function* $v(.)\colon R \times R^n \times 2^{R^n} \to R$

(a) *is positive definite on* $R_i \times M(t;J)$ *with respect to* J if and only if $M(t;J)$ is open connected neighbourhood of J for all $t \in R_i$ such that there exist $w_1(.)\colon R^n \times 2^{R^n} \to R$ and $\varphi_1(.) \in K$ obeying the following:

1) $v(t,x;J)$ and $w_1(x;J)$ are uniquely determined by $(t,x) \in R_i \times M(t;J)$ and continuous on $R_i \times M(t;J)$; that is that $v(t,x;J) \in C[R_i \times M(t;J)]$ and $w_1(x;J) \in C[M_M(R_i;J)]$,
2) $v(t,x;J) = 0$ and $w_1(x;J) = 0$ for all $(t,x) \in R_i \times \partial J$,
3) $v(t,x;J) \le 0$ for all $(t,x) \in R_i \times \operatorname{In} J$,
4) $v(t,x;J) \ge w_1(x;J) \ge \varphi_1[\rho(x,J)]$ for all $(t,x) \in R_i \times [M(t;J) - \operatorname{In} J]$.

(b) *is decrescent on* $R_i \times M(t;J)$ *with respect to* J if and only if $M_m(R_i;J)$ is open connected neighbourhood of J, and there exist $w_2(.)\colon R^n \times 2^{R^n} \to R$ and $\varphi_2(.) \in K$ obeying the following:

1) $v(t,x;J)$ and $w_2(x;J)$ are uniquely determined by $(t,x) \in R_i \times M(t;J)$ and continuous on $R_i \times M(t;J)$, that is that $v(t,x;J) \in C[R_i \times M(t;J)]$ and $w_2(x;J) \in C[M_M(R_i;J)]$, hence $w_2(x;J) \in C[M_m(R_i;J)]$,
2) $v(t,x;J) \le w_2(x;J) \le \varphi_2[\rho(x,J)]$ for all $(t,x) \in R_i \times [M_m(R_i;J) - \operatorname{In} J]$.

The expression "$R_i\times$" is to be omitted if and only if $R_i = R$, and the expression "$\times M(t;J)$" is to be omitted if and only if $M(t;J)$ is an arbitrarily small open connected neighbourhood of J for all $t \in R_i$.

The functions $w_i(.)$, $i = 1, 2$, can have the following form: $w_i(x;J) = \varphi_i[\rho(x,J)]$.

The form of problem solutions to be established depends on the smoothness properties of system (1) as well as whether a function $p(.)$ generating a system Lyapunov function is selected from $L(R_i, S; f; J)$ or from $E(R_i, S; f; J)$.

Theorem 6.1 *In order for a compact connected invariant set J of system (1) with the strong smoothness property to have the domain $D(R_i;J)$ of uniform asymptotic stability on R_i, for a set $N(t_0)$, $N(t_0) \subseteq R^n$, to be the domain of its asymptotic stability at any $t_0 \in R_i$: $N(t_0) \equiv D(t_0;J)$, and for a set N, $N \subseteq R^n$, to be the domain of its uniform asymptotic stability on R_i, $N = D(R_i;J)$, it is both necessary and sufficient that:*

1) *the set $N(t)$ is open continuous neighbourhood of J and $N(t) \subseteq S(t;J)$ for every $t \in R_i$,*
2) *the set N is a connected neighbourhood of J such that $N = \cap[N(t)\colon t \in R_i] = N_m(R_i;J) \subseteq S(R_i;J)$,*
3) *$f(t,x) = 0$ for all $t \in R_i$ is possible only for $x \notin [N(t) - J]$,*
   and
4) *for any decrescent positive definite function $p(.)$ on $R_i \times S(t;J)$ with respect to J, which obeys:*
   (a) *$p(.) \in L(R_i, S; f; J)$, the equations (2) have the unique solution function $v(.)$ with the following properties:*
      (i) *$v(.)$ is decrescent positive definite function on $R_i \times N(t)$ with respect to J, and*
      (ii) *if the boundary $\partial N(t)$ of $N(t)$ is nonempty then $x \to \partial N(t)$, $x \in N(t)$, implies $v(t,x;J) \to \infty$ for every $t \in R_i$,*

   or obeying:
   (b) *$p(.) \in E(R_i, S; f; J)$, the equations (4) have the unique solution function $u(.)$ with the following properties:*
      (i) *$u(.)$ is decrescent positive definite function on $R_i \times N(t)$ with respect to J, and*
      (ii) *if the boundary $\partial N(t)$ of $N(t)$ is nonempty then $x \to \partial N(t)$, $x \in N(t)$, implies $u(t,x;J) \to 1$ for every $t \in R_i$.*

Proof The proof will be a modification and generalization of the proof of Theorem 1 in [38]. The modification results from non-differentiability of system motions, which was requested in [38].

Necessity. Let the invariant set J of system (1) possessing the strong smoothness property have the uniform asymptotic stability domain $D(R_i;J)$ on R_i. Hence, it has also the asymptotic stability domain $D(t_0;J)$ at every $t_0 \in R_i$ (Definition 4.1). Definitions 4.1 and 4.3 show that it has also the uniform attraction domain $D_a(R_i;J)$ and the instantaneous attraction domain $D_a(t_0;J)$ at every $t_0 \in R_i$. By the definition (Definition 4.3), $D_a(t_0;J) \supseteq D(t_0;J)$ for all $t_0 \in R_i$ and $D_a(R_i;J) \supseteq D(R_i;J)$. Besides, $D_a(t_0;J)$ is a neighbourhood of J at every $t_0 \in R_i$ and $D_a(R_i;J)$ is also a neighbourhood of J (Definition 4.1). The set $S(t_0;J)$ is a neighbourhood of J at every $t_0 \in R_i$ and $S(R_i;J)$ is also a neighbourhood of J (the weak smoothness property).

Hence, $D_a(t_0; J) \cap S(t_0; J) \neq \emptyset$ for all $t_0 \in R_i$ and $D_a(R_i; J) \cap S(R_i; J) \neq \emptyset$. Let us prove $S(t_0; J) \supseteq D_a(t_0; J)$ for every $t_0 \in R_i$. If $S(t_0; J) \supseteq D_a(t_0; J)$ were not true for all $t_0 \in R_i$ then there would exist $t_0 \in R_i$ and $z \in [D_a(t_0; J) - S(t_0; J)]$, which would mean either $z \in [D_a(t_0; J) \cap \partial S(t_0; J)]$ or $z \in [D_a(t_0; J) - \operatorname{Cl} S(t_0; J)]$ due to $D_a(t_0; J) \cap S(t_0; J) \neq \emptyset$ and the fact that $S(t_0; J)$ is open [(i) of the weak smoothness property and (i) of the strong smoothness property]. If $z \in [D_a(t_0; J) \cap \partial S(t_0; J)]$ then $\inf\{\rho[\boldsymbol{x}(t; t_0, z), J] \colon t \in I_0\} > 0$ due to (ii) of the strong smoothness property, which would mean $z \notin D_a(t_0; J)$ and would contradict $z \in [D_a(t_0; J) \cap \partial S(t_0; J)]$. Hence, $z \notin [D_a(t_0; J) \cap \partial S(t_0; J)]$ and $D_a(t_0; J) \cap \partial S(t_0; J) = \emptyset$. If $z \in [D_a(t_0; J) - \operatorname{Cl} S(t_0; J)]$ then $\lim\{\rho[\boldsymbol{x}(t; t_0, z), J] \colon t \to \infty\} = 0$, which together with (i) of the strong smoothness property, (ii) of the weak smoothness property and with $S(t; J) \in C(R_i)$ would imply existence of $t^* \in (t_0, \infty)$ such that $\boldsymbol{x}(t^*; t_0, z) \in \partial S(t^*; J)$. This is impossible as shown above. Assumed t^* does not exist. Hence, $[D_a(t_0; J) - \operatorname{Cl} S(t_0; J)] = \emptyset$, which together with $D_a(t_0; J) \cap \partial S(t_0; J) = \emptyset$ and $D_a(t_0; J) \cap S(t_0; J) \neq \emptyset$ implies $S(t_0; J) \supseteq D_a(t_0; J)$ by having in mind that both $D_a(t_0; J)$ and $S(t_0; J)$ are open neighbourhoods of J [a-1) of Lemma A.1 (Appendix), (i) of the weak smoothness property and (i) of the strong smoothness property]. Therefore, $S(t_0; J) \supseteq D(t_0; J)$ due to $D_a(t_0; J) \equiv D(t_0; J)$ (Lemma A.2, Appendix). Let $N(t_0) \equiv D(t_0; J)$ so that $S(t_0; J) \supseteq N(t_0)$. Hence, $N(t)$ is open continuous neighbourhood of J for all $t \in R_i$ [a-2) of Lemma A.1] and $N = D(R_i; J)$ is connected neighbourhood of J [a-3) of Lemma A.1]. Besides, $N = \cap[N(t) \colon t \in R_i]$ because of $N(t) \equiv D(t; J)$ and $N = D(R_i; J) = \cap[D(t; J) \colon t \in R_i]$. They prove necessity of the conditions 1) and 2). From $D_s(t; J) \supseteq D_a(t; J) \equiv D(t; J) \equiv N(t)$ [a) of Lemma A.2] and Definitions 4.1 − 4.3 it results that there is not an equilibrium state of system (1) in $[N(t) - J]$, $\forall\, t \in R_i$, which implies that $f(t, x) = 0$ for all $t \in R_i$ is possible only for $x \notin [N(t) - J]$ (Proposition 7 in [44]). This proves necessity of the condition 3). From $N(t_0) \equiv D(t_0; J)$ it follows that the interval I_0 of existence of $\boldsymbol{x}(.; t_0, x_0)$ satisfies $I_0 \supseteq R_0$, $\forall\, (t_0, x_0) \in R_i \times N(t_0)$ due to Definitions 4.1 through 4.3. Let $p(.) \in L(R_i, S; f; J)$ be arbitrarily selected positive definite decrescent function on $R_i \times S(t; J)$ with respect to J. Hence, there is $\mu > 0$ such that there exists a solution function $v(.)$ to the equations (2), which is continuous in $(t, x) \in R_i \times B_\mu(J)$ and satisfies (3). Therefore,

$$|v(t, x; J)| < \infty, \quad \forall\, (t, x) \in R_i \times \operatorname{Cl} B_\mu(J). \tag{7}$$

Let $\beta \in (1, \infty)$ and $\zeta \in R^+$ be such that

$$\operatorname{Cl} B_\beta(J) \cap \operatorname{Cl} B_\mu(J) \cap S(t; J) \supset P_\zeta(t; J), \quad \forall\, t \in R_i. \tag{8}$$

Existence of such β and ζ is guaranteed by positive definiteness of $p(.)$ on $R_i \times S(t; J)$ and by the fact that $S(R_i; J)$ is a neighbourhood of J. Let $t_0 \in R_i$ be arbitrary and $\tau \in R_+$, $\tau = \tau(t_0, x_0; \zeta; J; p)$, be such that for any $x_0 \in N(t_0)$ the condition (9) holds,

$$\boldsymbol{x}(t; t_0, x_0) \in \operatorname{Cl} P_\zeta(t; J), \quad \forall\, t \in [t_0 + \tau, \infty). \tag{9}$$

Such τ exists due to Definitions 4.1 and 4.3, $x_0 \in N(t_0)$ and $D_a(t_0; J) \equiv D(t_0; J) \equiv N(t_0)$. Notice that $x_0 \in N(t_0)$ yields also

$$\rho[\boldsymbol{x}(\infty; t_0, x_0), J] = 0. \tag{10}$$

Let (2a) be integrated from $t \in R_0$ to ∞,

$$v[\infty, \boldsymbol{x}(\infty; t_0, x_0); J] - v[t, \boldsymbol{x}(t; t_0, x_0); J] = -\int_t^{\infty} p[\sigma, \boldsymbol{x}(\sigma; t_0, x_0); J]\, d\sigma, \qquad \forall\, (t, x_0) \in R_0 \times N(t_0). \tag{11}$$

Now, invariance of ∂J (by the definition, Section 2), (2b) and (10) enable the transformation of (11) to the next form,

$$v[t, \boldsymbol{x}(t; t_0, x_0); J] = \int_t^{t_0+\tau} p[\sigma, \boldsymbol{x}(\sigma; t_0, x_0); J]\, d\sigma + \int_{t_0+\tau}^{\infty} p[\sigma, \boldsymbol{x}(\sigma; t_0, x_0); J]\, d\sigma, \qquad \forall\, (t, x_0) \in R_0 \times N(t_0). \tag{12}$$

Invariance of $D_a(t; J)$ with respect to system motions on R_i [a-1) of Lemma A.1], $S(t; J) \supseteq D(t; J) \equiv D_a(t; J) \equiv N(t)$, continuity of $\boldsymbol{x}(t; t_0, x_0)$ in $(t; t_0, x_0) \in I_0 \times R_i \times S(t_0; J)$ [(i-b) of the weak smoothness property], continuity, positive definiteness and decrescency of $p(.)$ on $R_i \times S(t; J)$, the definition of τ, (9), and compactness of $[t, t_0 + \alpha]$ for any $\alpha \in R_+$ imply

$$\left| \int_t^{t_0+\alpha} p[\sigma, \boldsymbol{x}(\sigma; t_0, x_0); J]\, d\sigma \right| < \infty, \qquad \forall\, (\alpha, t, t_0, x_0) \in R_+ \times R_0 \times R_i \times N(t_0). \tag{13}$$

Now, (7) – (9), (12) and (13) for $\alpha = \tau$ yield

$$|v[t, \boldsymbol{x}(t; t_0, x_0); J]| < \infty, \qquad \forall\, (t, t_0, x_0) \in R_0 \times R_i \times N(t_0). \tag{14}$$

Let $t = t_0$ and $x = x_0$ be set in (14). Then,

$$|v(t, x; J)| < \infty, \qquad \forall\, (t, x) \in R_i \times N(t). \tag{15}$$

Continuity of $p(.)$ on $R_i \times S(t; J)$, $p(.) \in L(R_i, S; f; J)$, Definition 5.1, $S(t; J) \supseteq N(t)$, (12) and (15) prove

$$v(t, x; J) \in C[R_i \times N(t)] = C[R_i \times D(t)]. \tag{16}$$

Invariance of $D_a(t; J)$ [a-1) of Lemma A.1], $D_a(t; J) \equiv D(t; J) \equiv N(t)$, continuity of $\boldsymbol{x}(t; t_0, x_0)$ in $(t; t_0, x_0) \in I_0 \times R_i \times D(t_0; J)$, positive definiteness and decrescency of $p(.)$ on $R_i \times N(t)$, $p(.) \in L(R_i, S; f; J)$, (3), the definition of τ and compactness of $[t, t_0 + \tau]$ guarantee existence of $\zeta_k(.)\colon R^n \times 2^{R^n} \to R$, $k = 1, 2$, $\zeta_1(x; J) \in C[N_M(R_i; J)]$ and $\zeta_2(x; J) \in C[N_m(R_i; J)]$, where $N_M(R_i; J) = \cup[N(t; J)\colon t \in R_i]$, $N_m(R_i; J) = \cap[N(t; J)\colon t \in R_i]$ and $\psi_k(.)\colon R^n \times 2^{R^n} \to R$, $k = 1, 2$, such that

$$0 < \varsigma_1(x_0, J) \le \int_t^{t_0+\tau} \psi_1[\boldsymbol{x}(\sigma; t_0, x_0); J]\, d\sigma, \qquad \forall\, (t, t_0, x_0) \in R_0 \times R_i \times [N(t_0) - \operatorname{Cl} B_\mu(J)], \tag{17a}$$

$$\infty > \varsigma_2(x_0, J) \ge \int_t^{t_0+\tau} \psi_2[\boldsymbol{x}(\sigma; t_0, x_0); J]\, d\sigma, \qquad \forall\, (t, t_0, x_0) \in R_0 \times R_i \times [N_m(R_i; J) - \operatorname{Cl} B_\mu(J)], \tag{17b}$$

and

$$\psi_1(x;J) \in C[N_M(R_i;J)], \qquad \psi_2(x;J) \in C[N_m(R_i;J)], \tag{18a}$$

$$\psi_k(x;J) = 0, \quad \forall\, x \in \partial J, \quad k = 1,2, \tag{18b}$$

$$\begin{aligned} &\psi_1(x;J) > 0, \quad \forall\, x \in [N_M(R_i;J) - J], \\ &\psi_2(x;J) > 0, \quad \forall\, x \in [N_m(R_i;J) - J], \end{aligned} \tag{18c}$$

$$\psi_1(x;J) \le p(t,x), \quad \forall\, (t,x) \in R_i \times [N_M(R_i;J) - \operatorname{In} J], \tag{18d}$$

$$p(t,x) \le \psi_2(x;J), \quad \forall\, (t,x) \in R_i \times [N_m(R_i;J) - \operatorname{In} J]. \tag{18e}$$

Such functions $\psi_k(.)$ exist due to decrescency and positive definiteness of $p(.)$ on $R_i \times S(t;J)$, $p(.) \in L(R_i, S; f; J)$ and $S(t;J) \supseteq N(t)$. They can be of the form $\psi_k(x;J) = g_k[\rho(x,J)]$, $k = 1,2$, together with $g_k(.)$ in the class K: $g_k(.) \in K$. Let $w_k(.)\colon R^n \times 2^{R^n} \to R$, $k = 1,2$, obey (19),

$$w_k(x;J) \in C(R^n) \quad \text{and} \quad w_k(x;J) = 0, \quad \forall\, x \in \partial J, \quad k = 1,2, \tag{19a}$$

$$0 < w_1(x;J) \le \begin{cases} \varsigma_1(x;J), & \forall\, x \in [N_M(R_i;J) - \operatorname{Cl} B_\mu(J)], \\ w_\mu(x;J), & \forall\, x \in [\operatorname{Cl} B_\mu(J) - J], \end{cases} \tag{19b}$$

$$w_2(x;J) \ge \begin{cases} \begin{array}{c} \varsigma_2(x;J) + w_\mu(x_\tau;J), \\ x_\tau = \boldsymbol{x}(\tau;t,x), \end{array} & \forall\, (t,x) \in R_i \times [N_m(R_i;J) - \operatorname{Cl} B_\mu(J)], \\ w_\mu(x;J), & \forall\, x \in [\operatorname{Cl} B_\mu(J) - J], \end{cases} \tag{19c}$$

where $w_\mu(.)$ is defined by (3). Now (3), (12), positive definiteness of $p(.)$ on $R_i \times S(t;J)$ with respect to J, invariance of J and (17) – (19) yield the following for $(t_0, x_0) = (t, x)$:

$$w_1(x;J) \le v(t,x;J), \quad \forall\, (t,x) \in R_i \times [N(t) - \operatorname{In} J], \tag{20a}$$

$$v(t,x;J) \le w_2(x;J), \quad \forall\, (t,x) \in R_i \times [N_m(R_i;J) - \operatorname{In} J], \tag{20b}$$

$$v(t,x;J) \le 0, \qquad \forall\, (t,x) \in R_i \times J, \tag{20c}$$

$$v(t,x;J) = 0, \qquad \forall\, (t,x) \in R_i \times \partial J. \tag{20d}$$

From $p(.) \in L(R_i, S; f; J)$, (2b), (16) and (20) it follows that a solution function $v(.)$ to (2) is decrescent, positive definite and continuous on $R_i \times N(t)$ with respect to J. Let be assumed that there exist two such solutions $v_1(.)$ and $v_2(.)$ of (2). Hence,

$$v_1(t_0,x_0;J) - v_2(t_0,x_0;J) = \int_{t_0}^{\infty} \big\{ p[\sigma, \boldsymbol{x}_1(\sigma;t_0,x_0);J] - p[\sigma, \boldsymbol{x}_2(\sigma;t_0,x_0);J] \big\}\, d\sigma, \tag{21}$$
$$\forall\, (t,x_0) \in R_0 \times N(t_0).$$

Uniqueness of the motions $\boldsymbol{x}(.;t_0,x_0)$, $\forall\, (t_0,x_0) \in R_i \times N(t_0)$ (the weak smoothness property), $S(t_0;J) \supseteq N(t_0)$ and uniqueness of $p(t,x)$ for every $(t,x) \in R_i \times S(t;J)$ [due to positive definiteness of $p(.)$ on $S(t;J)$] imply

$$\int_{t_0}^{\infty} \{p[\sigma, \boldsymbol{x}_1(\sigma;t_0,x_0);J] - p[\sigma, \boldsymbol{x}_2(\sigma;t_0,x_0);J]\}\, d\sigma$$

$$= \int_{t_0}^{\infty} \{p[\sigma, \boldsymbol{x}(\sigma;t_0,x_0);J] - p[\sigma, \boldsymbol{x}(\sigma;t_0,x_0);J]\}\, d\sigma = 0, \quad \forall\,(t,x_0) \in R_0 \times N(t_0).$$

This and (21) prove

$$v_1(t_0,x_0;J) \equiv v_2(t_0,x_0;J).$$

Hence, the function $v(.)$ is the unique solution to (2). This completes the proof of necessity of the condition 4-a-i). If $\partial N(t_0) \neq \emptyset$ then let $t_0 \in R_i$ be arbitrary and x_k, $k = 1,2,\ldots$, be a sequence converging to u, $x_k \to u$ as $k \to \infty$, $x_k \in N(t_0)$, for all $k = 1,2,\ldots$, and $u \in \partial N(t_0)$. Let $\zeta \in R^+$ be arbitrarily chosen so that $N \supset \operatorname{Cl} P_\zeta(t;J)$ for all $t \in R_i$. Such ζ exists because $p(.)$ is positive definite on $S(t;J)$, $N \subseteq S(t;J)$, $\forall\, t \in R_i$, and defines $\operatorname{Cl} P_\zeta(t;J)$, and because N is a neighbourhood of J. Let τ_k, $\tau_k = \tau(t_0,x_k;\zeta;J) \in R_+$, be the first instant satisfying (22),

$$\boldsymbol{x}(t;t_0,x_0) \in \operatorname{Cl} P_\zeta(t;J), \quad \forall\, t \in [t_0 + \tau_k, \infty). \tag{22}$$

Existence of such τ_k is ensured by $x_k \in N(t_0)$, $N(t) \equiv D(t)$ and by the fact that $\cap[P_\zeta(t;J)\colon t \in R_i]$ is a neighbourhood of J due to decrescency of $p(.)$ on $R_i \times N(t)$ [42]. Continuity of $\boldsymbol{x}(t;t_0,x_0)$ in $(t;t_0,x_0) \in I_0 \times R_i \times S(t_0;J)$ (the weak smoothness property), $S(t_0;J) \supseteq D(t_0;J) \equiv N(t_0)$, positive invariance of $D(t;J)$ [a) of Lemma A.1], the fact that $\cap[D(t;J)\colon t \in R_i] = D(R_i;J)$ is neighbourhood of J [(b) of Definition 4.1 through Definition 4.3] and $x_k \in N(t_0)$ imply

$$\tau_k \to \infty \quad \text{as} \quad k \to \infty. \tag{23}$$

Let $m \in \{1,2,\ldots\}$ be such that $x_k \in \{N(t_0) - \operatorname{Cl} P_\zeta(t_0;J)\}$ for all $k = m, m+1, \ldots$, and $x_k \to \partial N(t_0)$ as $k \to \infty$. Such x_k exists because $N(t_0) = D(t_0)$ is open [a-2) of Lemma A.1] and $N(t_0) \supset \operatorname{Cl} P_\zeta(t_0;J)$.

Let α be defined by

$$\alpha = \min\{p(t,x)\colon (t,x) \in R_i \times [S(t;J) - P_\zeta(t;J)]\}. \tag{24}$$

Since $p(.) \in L(R_i,S;f;J)$ then $\alpha \in R^+$. Hence, (12), (22), (24) and the definitions of α and τ_k yield $v(t_0,x_k;J) \geq \alpha\tau_k$, $\forall\, t_0 \in R_i$, which together with (23) proves necessity of the condition 4-a-ii). The conditions under 4-b) follow from 4-a) due to (2), (4) and (6). This completes the proof of the necessity part.

Sufficiency. Let all the conditions of Theorem 6.1 hold. Since the function $v(.)$ is the solution to (2), [or, $u(.)$ is the solution to (4)], and it is positive definite and decrescent on $R_i \times N(t)$ with respect to J, $p(.) \in L(R_i,S;f;J)$, [or, $p(.) \in E(R_i,S;f;J)$], and $p(.)$ is decrescent positive definite on $R_i \times N(t)$ with respect to J, then J is uniformly asymptotically stable set on R_i, which is easy to verify by using Definition 4.1 through Definition 4.3 and by following Lyapunov [1], Lakshmikantham and

Leela [13], Yoshizawa [18], Zubov [19], Hahn [39], Grujić *et al.* [42], Miller and Michel [46], Krasovskii [49], Rouche *et al.* [50], Demidovich [51], Halanay [52], Hale [53], Kalman and Bertram [54]. Hence, J has both $D(t_0;J)$ at $t_0 \in R_i$ and $D(R_i;J)$ (Definitions 4.1 − 4.3) so that $D_a(t_0;J) \equiv D(t_0;J)$ and $D_a(R_i;J) = D(R_i;J)$ (Lemma A.2). In order to show that $N(t_0) \equiv D(t_0;J)$ and $N = D(R_i;J)$ we proceed as follows. The condition (ii) of the strong smoothness property guarantees $D(t_0;J) \subseteq S(t_0;J)$, $t_0 \in R_i$. Let $t_0 \in R_i$ be arbitrary and fixed. If $\partial N(t_0) = \emptyset$ then $N(t_0) = R^n$. Hence, $D(t_0;J) \subseteq N(t_0)$ is then only possible. If $D(t_0;J) \subset N(t_0)$ then $\partial D(t_0;J) \cap N(t_0) \neq \emptyset$ that implies $v(t_0,x;J) \to \infty$ as $x \to \partial D(t_0;J)$ (because the function $v(.)$ is the solution to (2), as shown in the proof of necessity), which contradicts the condition 4-a,i) because of $N(t_0) = R^n$. This implies $\partial D(t_0;J) \cap N(t_0) = \partial D(t_0;J) \cap R^n = \emptyset$. Since $D(t_0;J)$ is an open neighbourhood of J and J is compact connected set, then $D(t_0;J) = R^n$, i.e. $D(t_0;J) = N(t_0)$. Let it be now supposed that $\partial N(t_0) \neq \emptyset$, i.e. $N(t_0) \subset R^n$. If we assume now $\partial D(t_0;J) = \emptyset$, then $D(t_0;J) = R^n$ that implies $\partial N(t_0) \cap D(t_0;J) \neq \emptyset$. This and the condition 4-a,ii) show that there is a set $L \subseteq \partial N(t_0) \cap D(t_0;J)$ such that $v(t_0,x;J) \to \infty$ as $x \to L \subseteq \partial N(t_0) \cap D(t_0;J)$, which is impossible because the function $v(.)$ is the unique solution of (2), which is continuous on $R_i \times D(t;J)$, as shown in details in the necessity part. Assumed $\partial D(t_0;J) = \emptyset$ fails. Let $\partial D(t_0;J) \neq \emptyset$ be considered. If $\partial D(t_0;J) \cap \partial N(t_0) = \emptyset$ then either $D(t_0;J) = N(t_0)$ or $D(t_0;J) \subset N(t_0)$ or $N(t_0) \subset D(t_0;J)$ because both are open neighbourhoods of the set J and their boundaries are nonempty. The last two cases are not possible due to positive definiteness of the function $v(.)$ on $R_i \times N(t)$ and its construction via the equations (2) as shown above. If $\partial D(t_0;J) \cap \partial N(t_0) \neq \emptyset$ then either $\partial D(t_0;J) = \partial N(t_0)$, which implies $D(t_0;J) = N(t_0)$, or $\partial D(t_0;J) \cap N(t_0) \neq \emptyset$ and/or $D(t_0;J) \cap \partial N(t_0) \neq \emptyset$. If $\partial D(t_0;J) \cap N(t_0) \neq \emptyset$ then it means that the function $v(.)$ blows up (to ∞) on $N(t_0)$, which contradicts its continuity on $N(t_0)$ due to the condition 4-a,i). If $D(t_0;J) \cap \partial N(t_0) \neq \emptyset$ then it means that the function $v(.)$ blows up on $D(t_0;J)$ that is impossible due to (16) because $v(.)$ is generated by (2). Hence, $\partial D(t_0;J) = \partial N(t_0)$ that implies $D(t_0;J) = N(t_0)$, which holds as the overall result. Now, $N = \cap[N(t)\colon t \in R_i]$ (the condition 2) and the conditions b-3 of Definitions 4.1 and 4.2 together with b) of Definition 4.3 imply $D(R_i;J) = N$. Positive definiteness of $p(.)$ on $S(t;J)$, $p(.) \in L(R_i,S;f;J)$, the equation (2a), $N(t) \subseteq S(t;J)$ for all $t \in R_i$, the condition 4-a,i) and a) of Lemma A.1 imply

$$v[t_0+\tau, \boldsymbol{x}(t_0+\tau,t_0,x_0);J] \leq v(t_0,x_0;J) - \xi(\varsigma;p;v;N;R_i)\tau(t_0,x_0;\varsigma;J;p),$$

where $\zeta \in R^+$ is arbitrarily small,

$$\xi(\varsigma;p;v;N;R_i) = \min\{p(t,x;J)\colon (t,x) \in R_i \times [N - V_\psi(R_i;J)]\} \in R^+, \quad \psi = \varphi_1(\varsigma),$$

$$V_\psi(R_i;J) = \cap[V_\psi(t;J)\colon t \in R_i],$$

$$\varphi_1(\|x\|) \leq v(t,x;J), \quad \forall\,(t,x) \in R_i \times N(t), \quad \varphi_1(.) \in K,$$

$$v(t,x;J) \leq \varphi_2(\|x\|), \quad \forall\,(t,x) \in R_i \times N, \quad \varphi_2(.) \in K,$$

so that

$$\tau(t_0,x_0;\varsigma;J;p) \leq [\varphi_2(\|x_0\|) - \varphi_1(\varsigma)]\xi^{-1}(\varsigma;p;v;N;R_i),$$

$$\sup[\tau_m(t_0,x_0;\varsigma;J;p)\colon t_0 \in R_i] \leq [\varphi_2(\|x_0\|) - \varphi_1(\varsigma)]\xi^{-1}(\varsigma;p;v;N;R_i) < \infty, \quad \forall\,x_0 \in N,$$

and, therefore, the conditions under (b) of Definitions 4.1 through 4.3 are satisfied. This completes the proof of sufficiency of the conditions 1-4a). Sufficiency of the conditions 1-3,4b) is implied by sufficiency of 1-4a) and (6), which completes the proof.

If system (3.1) possesses the weak smoothness property then conditions of Theorem 6.1 change.

Theorem 6.2 *In order for a compact connected invariant set J of system (1) with the weak smoothness property to have the domain $D(R_i;J)$ of uniform asymptotic stability on R_i, for a set $N(t_0)$, $N(t_0) \subseteq S(t_0;J)$ for all $t_0 \in R_i$, to be the domain of its asymptotic stability at $t_0 \in R_i$: $N(t_0) \equiv D(t_0;J)$, and for a set N, $N \subseteq S(R_i,J)$, to be the domain of its uniform asymptotic stability on R_i, $N = D(R_i;J)$, it is both necessary and sufficient that the following holds:*

1) *the set $N(t)$ is open continuous neighbourhood of J for every $t \in R_i$,*
2) *the set N is a connected neighbourhood of J such that $N = \cap[N(t)\colon t \in R_i] = N_m(R_i;J)$,*
3) *$f(t,x) = 0$ for all $t \in R_i$ is possible only for $x \notin [N(t) - J]$,*
   and
4) *for any decrescent positive definite function $p(.)$ on $R_i \times R^n$ with respect to J, which obeys:*
   - (a) *$p(.) \in L(R_i,R^n;f;J)$, the equations (2) have the unique solution function $v(.)$ with the following properties:*
     - (i) *$v(.)$ is decrescent positive definite function on $R_i \times R^n$ with respect to J,*
     - (ii) *if the boundary $\partial N(t)$ of $N(t)$ is nonempty then $x \to \partial N(t)$, $x \in N(t)$, implies $v(t,x;J) \to \infty$ for every $t \in R_i$,*

   or obeying:
   - (b) *$p(.) \in E(R_i,R^n;f;J)$, the equations (4) have the unique solution function $u(.)$ with the following properties:*
     - (i) *$u(.)$ is decrescent positive definite function on $R_i \times N(t)$ with respect to J,*
     - (ii) *if the boundary $\partial N(t)$ of $N(t)$ is nonempty then $x \to \partial N(t)$, $x \in N(t)$, implies $u(t,x;J) \to 1$ for every $t \in R_i$.*

Proof The proof will be a modification and generalization of that of Theorem 2 in [38]. The modification is caused by non-differentiability of system motions, which was assumed in [38].

Necessity. Let system (1) possess the weak smoothness property. Let the invariant set J have the uniform asymptotic stability domain $D(R_i;J)$ on R_i so that it has also the asymptotic stability domain $D(t_0;J)$ at every $t_0 \in R_i$. Let $N(t_0) = D(t_0;J) \subseteq S(t_0;J)$ for all $t_0 \in R_i$ so that $D(R_i;J) \subseteq S(J)$ and $N = D(R_i;J)$. Let a positive definite decrescent function $p(.)$ on $R_i \times R^n$ with respect to J be arbitrarily selected so that $p(.) \in L(R_i,R^n;f;J)$, {or, $p(.) \in E(R_i,R^n;f;J)$}. From now on we should repeat the proof of necessity of the conditions of Theorem 6.1 in order to complete this proof.

Sufficiency. Let system (1) possess the weak smoothness property and let the conditions 1) − 4) hold. The set J is uniformly asymptotically stable on R_i, which can be easily verified by following Lyapunov [1], Lakshmikantham and Leela [13], Yoshizawa [18], Zubov [19], Hahn [39], Grujić *et al.* [42], Miller and Michel [46], Krasovskii [49], Rouche *et al.* [50], Demidovich [51], Halanay [52], Hale [53], Kalman and Bertram [54]. Hence, J has both the domain $D(R_i; J)$ of uniform asymptotic stability and the domain $D(t_0; J)$ of asymptotic stability at $t_0 \in R_i$ (Definition 4.3). Let $x_0 \in [R^n - N(t_0)]$ and $t_0 \in R_i$ be arbitrary. Continuity of $\boldsymbol{x}(t; t_0, x_0)$ in $t \in R_0$ (the weak smoothness property), continuity of $p(.)$ on $R_i \times R^n$ due to its positive definiteness on $R_i \times R^n$, the generation of $v(.)$ via (2) and the condition 4-a-ii), [4-b-ii)] imply $\boldsymbol{x}(t; t_0, x_0) \in [R^n - N(t)]$ for all $t \in I_0$. Therefore, $D(t_0; J) \subseteq \operatorname{Cl} N(t_0)$ and $D(R_i; J) \subseteq \operatorname{Cl} N$. Since $v(.)$ is generated via (2) then (as shown in the proof of the necessity part of Theorem 6.1) $v(t, x) \to \infty$ as $x \to \partial D(t; J)$, $x \in D(t; J)$, for every $t \in R_i$, which, together with the condition 4-a-1) proves $\partial D(t; J) \cap N(t) = \emptyset$ for every $t \in R_i$. This result, $D(t; J) \subseteq \operatorname{Cl} N(t)$, and the fact that both $N(t)$ and $D(t; J)$ are open neighbourhoods of J [condition 1) and a-2) of Lemma A.1] imply $N(t) \equiv D(t; J)$ and $N = D(R_i; J)$. By repeating the end of the proof of sufficiency of Theorem 6.1 we show that

$$\sup[\tau_m(t_0, x_0, \zeta; J)\colon t_0 \in R_i] < +\infty \quad \text{for every} \quad (x_0, \zeta) \in D_a(R_i; J) \times R^+,$$

which completes the proof.

Theorems 6.1 and 6.2 are based on the usage of $p(.) \in L(.)$, $\{p(.) \in E(.)\}$. The function $p(.)$ should obey the condition 3) of Definition 5.1, [3) of Definition 5.2], if we wish to determine exactly $D(t; J)$ and $D(R_i; J)$. Such a condition is not necessary for the test of only uniform asymptotic stability of J.

7 Uniform Asymptotic Stability of Invariant Sets

Uniform stability properties of time-varying systems are interesting for their independence of the initial moment t_0, which is a characteristic of stability properties of time-invariant systems.

Theorem 7.1 *In order for a compact connected invariant set J of system (1) possessing the weak smoothness property to be uniformly asymptotically stable on R_i it is both necessary and sufficient that there is an open connected neighbourhood A of J such that the following is valid:*

1) *$f(t, x) = 0$ for all $t \in R_i$ is possible only for $x \notin (A - J)$,*
2) *for any decrescent positive definite function $p(.)$ on $R_i \times A$ with respect to J, which obeys the conditions 1) and 2) of Definition 5.1, the equations (2) have a unique solution function $v(.)$ that is decrescent positive definite function on $R_i \times A$ with respect to J.*

Proof The proof will be a modification of that of Theorem 3 in [38]. The modification is due to non-differentiability of system motions, which was demanded in [38].

Necessity. Let system (1) possess the weak smoothness property. Let the invariant set J be uniformly asymptotically stable on R_i so that it has the domain $D(R_i; J)$ of uniform asymptotic stability (Definitions 4.1 through 4.3). Necessity of the condition 1) is proved in the same way as in the proof of Theorem 6.1. Since $D(R_i; J)$ and $S(R_i; J)$ are neighbourhoods of J then $D(R_i; J) \cap S(R_i; J) \neq \emptyset$. Let M be an open connected neighbourhood of J, which obeys $M \subseteq D(R_i; J) \cap S(R_i; J)$, and let $p(.)$ be an arbitrary decrescent positive definite function on $R_i \times M$ obeying the conditions 1) and 2) of

Definition 5.1. Hence, there exist positive definite functions $\psi_k(.)$ with respect to J, $\psi_k(.)\colon R^n \times 2^{R^n} \to R,\ k=1,2$, which satisfy (25),

$$\psi_1(x;J) \le p(t,x;J) \le \psi_2(x;J), \quad \forall\,(t,x) \in R_i \times (M - \operatorname{In} J). \tag{25}$$

From the conditions 1) and 2) of Definition 5.1 it results that there is a solution $v(.)$ to (2), which is well defined and continuous on $\operatorname{Cl} B_\mu(J)$ and obeys (3). The set $L = M \cap B_\mu(J)$ is also open and connected neighbourhood of J and $L \subseteq D(R_i;J)$. Let $\varepsilon \in R^+$ be arbitrarily selected so that $B_\varepsilon(J) \subseteq L$. Hence, $B_\varepsilon(J) \subseteq D(R_i;J)$. Let $\rho \in R^+$ obeying $B_\rho(J) \subseteq D_s(\varepsilon;J)$, $D_s(\varepsilon;J) = \cap\{D_s(t_0,\varepsilon;J)\colon t_0 \in R_i\}$ (Definitions 4.2 and 4.3), be arbitrarily selected. By following the proofs of (15) and (16) we prove that the function $v(.)$ has the next property since $B_\rho(J) \subseteq D_s(\varepsilon;J) \subseteq B_\varepsilon(J) \subseteq L \subseteq M$ [the second inclusion is implied by the definition of $D_s(\varepsilon;J)$],

$$|v(t,x;J)| < \infty, \quad \forall\,(t,x) \in R_i \times B_\rho(J), \quad v(t,x;J) \in C[R_i \times B_\rho(J)]. \tag{26}$$

By following the proof of (20) we show that there are $w_k(x;J) \in C[B_\rho(J)]$, $w_k(x;J) = 0$ for every $x \in \partial J$ and $w_k(x;J) > 0$, for every $x \in [B_\rho(J) - J]$, $k = 1,2$, such that

$$w_1(x;J) \le v(t,x;J) \le w_2(x;J), \quad \forall\,(t,x) \in R_i \times [B_\rho(J) - \operatorname{In} J]. \tag{27}$$

The results (26), (27), $w_k(x;J) \in C[B_\rho(J)]$, and $w_k(x;J) = 0$ for every $x \in \partial J$ and $w_k(x;J) > 0$, for every $x \in [B_\rho(J) - J]$, $k = 1,2$, prove that the solution $v(.)$ to (2) is decrescent positive definite function on $R_i \times A$, for $A = B_\rho(J)$. Its uniqueness is proved in the same way as in the proof of the necessity part of Theorem 6.1. Hence, all the conditions are necessary for uniform asymptotic stability of J on R_i.

Sufficiency. Sufficiency of the conditions of Theorem 7.1 for uniform asymptotic stability of J on R_i of system (1) is easy to verify by following Lyapunov [1], Lakshmikantham and Leela [13], Yoshizawa [18], Zubov [19], Hahn [39], Grujić *et al.* [42], Miller and Michel [46], Krasovskii [49], Rouche *et al.* [50], Demidovich [51], Halanay [52], Hale [53], Kalman and Bertram [54], or by following the proof of sufficiency of the conditions of Theorem 6.2.

Comment 7.1 The theorems are valid for global uniform asymptotic stability of J if $S(t;J) \equiv R^n$ without demanding radial unboundedness of $v(.)$ due to (2) and the properties of $p(.)$.

8 Conclusion

The consistent Lyapunov methodology enables us to construct exactly a system Lyapunov function and to determine accurately the domain of asymptotic stability of invariant sets. This is achieved for non-differentiable time-varying non-linear systems. The results provide the conditions that are both necessary and sufficient, and which are not expressed in terms of existence of a system Lyapunov function. They permit *an arbitrary choice* of a non-differentiable decrescent positive definite function $p(.)$ from the functional family $L[R_i, S; f; J)]$, {or from $E[R_i, S; f; J)]$}. They are formulated in terms of properties of a solution function $v(.)$ to $D^+v(.) = -p(.)$ (2), {or in terms of properties of a solution function $u(.)$ to $D^+u(.) = -[1 - u(.)]p(.)$, (4)}, respectively, which are obtained for a selected function $p(.)$. Definitions 5.1 and 5.2 determine the families $L[R_i, S; f; J)]$ and $E[R_i, S; f; J)]$ so that they are independent of a stability property of the system. If an obtained solution $v(.)$, $\{u(.)\}$, is also decrescent positive definite then (Theorem 7.1) the invariant set IS uniformly asymptotically stable. If $v(.)$, $\{u(.)\}$, does not possess any of these features then the invariant set IS NOT uniformly asymptotically stable. The solution to the problem of uniform asymptotic stability is obtained under a *single* application of Theorem 7.1. The same holds for the determination of both the domain of asymptotic stability of the invariant set J at any initial time $t_0 \in R_i$ and for its domain of uniform asymptotic stability (Theorems 6.1 and 6.2). These results generalize those of [38].

The consistent Lyapunov methodology for the non-linear systems is inverse to Lyapunov's original methodology for the non-linear systems. The former is consistent due to its consistency with Lyapunov's methodology for time-invariant linear systems and generalizes it in the framework of both linear and non-linear systems, while the latter is not.

The consistent Lyapunov methodology provides the complete solution to the uniform asymptotic stability problem after its *single* application, which is not guaranteed by Lyapunov's original methodology for non-linear systems. No repetition of the procedure is needed in the former case if the test result is negative.

The consistent Lyapunov methodology can be further developed to other classes of dynamical systems such as discrete-time systems, stochastic systems and those governed by functional-differential or partial differential equations.

Appendix

Lemma A.1 *Let system (1) possess the weak smoothness property and let a compact connected invariant set J be uniformly attractive on R_i with the instantaneous domain $D_a(t;J)$ of attraction obeying $D_a(t;J) \subseteq S(t;J)$ for all $t \in R_i$ and with the domain $D_a(R_i;J)$ of uniform attraction on R_i.*

a) *If $R_i \subset R$ then*
   1) *$(t_0, x_0) \in R_i \times D_a(t_0;J)$ implies $x(t;t_0,x_0) \in D_a(t_0;J)$ for all $t \in R_i$, which means that $D_a(t;J)$ is invariant on R_i,*
   2) *$D_a(t;J)$ is open continuous neighbourhood of J at any $t \in R_i$: $D_a(t;J) \equiv \operatorname{In} D_a(t;J)$, $D_a(t;J) \in C(R_i)$,*
   3) *$D_a(R_i;J)$ is connected neighbourhood of J. If $D_a(t;J) = D_a(R_i;J)$ for all $t \in R_i$ then $D_a(R_i;J)$ is also invariant on R_i.*

b) *If $R_i = R$ then*
   1) *$D_a(t;J)$ is invariant, that is that $(t_0, x_0) \in R_i \times D_a(t_0;J)$ implies $x(t;t_0,x_0) \in D_a(t_0;J)$ for all $t \in R$,*
   2) *$D_a(t;J)$ is open continuous neighbourhood of J at any $t \in R$: $D_a(t;J) \equiv \operatorname{In} D_a(t;J)$, $D_a(t;J) \in C(R)$,*
   3) *$D_a(J)$ is connected neighbourhood of J. If $D_a(t;J) = D_a(J)$ for all $t \in R$ then $D_a(J)$ is also invariant.*

Proof We will follow the proof of Lemma A.1 of [38] in order to show its validity also for non-differentiable time-varying non-linear systems.

Let system (1) possess the weak smoothness property and let a compact connected invariant set J be uniformly attractive on R_i with the instantaneous domain $D_a(t;J)$ of attraction obeying $D_a(t;J) \subseteq S(t;J)$ for all $t \in R_i$ and with the domain $D_a(R_i;J)$ of uniform attraction on R_i. Hence, $D_a(R_i;J) = \cap[D_a(t_0;J)\colon t_0 \in R_i]$ (Definition 4.1).

a) Let t_0 and $t^* \in R_i$, $t_0 \neq t^*$. Let $x^* = \boldsymbol{x}(t^*;t_0,x_0)$ for any $x_0 \in D_a(t_0;J)$. Then, $\boldsymbol{x}(t;t_0,x_0) \to J$ as $t \to \infty$. Since $\boldsymbol{x}(t;t^*,x^*) \equiv \boldsymbol{x}[t;t^*,\boldsymbol{x}(t^*;t_0,x_0)] \equiv \boldsymbol{x}(t;t_0,x_0)$, which is true due to (i) of the weak smoothness property and $D_a(t_0;J) \subseteq S(t_0;J)$, then $\boldsymbol{x}(t;t^*,x^*) \to J$ as $t \to \infty$. Hence, $x^* = \boldsymbol{x}(t^*;t_0,x_0) \in D_a(t^*;J)$ that proves the statement under a-1). Let $\zeta \in R^+$ be such that $B_{2\zeta}(J) \subset D_a(R_i;J)$. It exists due to Definition 4.1b. Let be assumed that $D_a(t;J)$ is not open for all $t \in R_i$. Let there exist $t_0' \in R_i$ and $x_0' \in \partial D_a(t_0';J) \cap D_a(t_0';J)$. Let $\varepsilon \in (0,\zeta/2)$. Then, (i) of the weak smoothness property and $D_a(t_0;J) \subseteq S(t_0;J)$, $t_0 \in R_i$, imply existence of $\theta \in R^+$, $\theta = \theta(t_0',x_0',\varepsilon)$, such that $\|x_0 - x_0'\| < \theta$ ensures $\|\boldsymbol{x}(t_0'+2\sigma';t_0',x_0) - \boldsymbol{x}(t_0'+2\sigma';t_0',x_0')\| < \varepsilon$, where $\sigma' = \tau(t_0',x_0',\zeta)$ (Definition 4.1a). Since $\varepsilon < \zeta/2$ and $\rho[\boldsymbol{x}(t_0'+2\sigma';t_0',x_0');J] < \zeta$ then $\boldsymbol{x}(t_0'+2\sigma';t_0',x_0) \in B_{2\zeta}(J) \subset D_a(R_i;J)$. Hence, $x_0 \in D_a(t_0';J)$. Any x_0 obeying $\|x_0 - x_0'\| < \theta$ may be selected in a θ-neighbourhood of x_0' out of $D_a(t_0';J)$, which is contradicted by the obtained $x_0 \in D_a(t_0';J)$. The former is true and the latter is wrong showing that there are not $t_0' \in R_i$ and $x_0' \in \partial D_a(t_0';J) \cap D_a(t_0';J)$. If $x_0' \in \partial D_a(t_0';J)$ then $x_0' \notin D_a(t_0';J)$. The set $D_a(t_0;J)$ is open for all $t_0 \in R_i$ and it is neighbourhood of J due to Definition 4.1. Therefore, $D_a(t;J) \equiv \operatorname{In} D_a(t;J)$ and it is a neighbourhood of J on R_i. Altogether, $D_a(t;J)$ is open neighbourhood of J on R_i. In order to prove $D_a(t;J) \in C(R_i)$ we will use a contradiction. Let there exist $t_0^* \in R_i$ such that $D_a(t;J)$ is discontinuous at t_0^*. As a consequence, there are $\varepsilon^* \in R^+$ and a sequence $K^* \subseteq \{1,2,\ldots,n,\ldots\}$ such that $t_k \to t_0^*$, $k \to \infty$, $k \in K^*$, and that there is $z^* \in D_a(t_0^*;J)$ for which $\rho[z^*,D_a(t_k;J)] \geq \varepsilon^*$, $\forall\, k \in K^*$, and/or there is $w^* \in D_a(t_k;J)$, $\forall\, k \in K^*$, which obeys $\rho[w^*,D_a(t_0^*;J)] \geq \varepsilon^*$. Let $\xi \in R^+$ obey both $\xi < \varepsilon/2$ and $B_\xi(J) \subseteq D_a(t;J)$ for all $t \in R_i$, which is possible due to uniform attraction of J on R_i [b-2) of Definition 4.1]. Let $m \in K^*$ be such that $t_m > t_0^* + \tau(t_0^*,z^*,\xi/2)$, $t_m \in R_i$. This guarantees (Definition 4.1): $\boldsymbol{x}(t_m;t_0^*,z^*) \in B_{\xi/2}(J)$. Let $\delta = \delta(t_0^*;z^*;m;\xi/2) \in R^+$, $\delta < \xi/2$, and $\psi = \psi(t_0^*;z^*;m;\xi/2) \in R^+$ obey that

$$|t_j - t_0^*| < \psi \quad \text{and} \quad \|x_0 - z^*\| < \delta,\ j \in K^* \quad \text{imply} \quad \|\boldsymbol{x}(t_m;t_j,x_0) - \boldsymbol{x}(t_m;t_0^*,z^*)\| < \xi/2,$$

which is possible due to continuity of the system motions [the weak smoothness property: (i-b) and (ii)]. Hence, $\boldsymbol{x}(t_m;t_0^*,z^*) \in B_{\xi/2}(J)$ implies $\boldsymbol{x}(t_m;t_j,x_0) \in B_\xi(J)$. This further yields $\boldsymbol{x}(t_m;t_j,x_0) \in D_a(t_m;J)$ and $x_0 \in D_a(t_j;J)$. Besides, $\|x_0 - z^*\| < \delta < \xi/2 < \varepsilon^*/4$ and $x_0 \in D_a(t_j;J)$ imply $\rho[z^*,D_a(t_j;J)] < \varepsilon^*$ that contradicts $\rho[z^*,D_a(t_k;J)] \geq \varepsilon^*$, $\forall\, k \in K^*$, and disproves existence of time $t_0^* \in R_i$ and $z^* \in D_a(t_0^*;J)$ for which $\rho[z^*,D_a(t_k;J)] \geq \varepsilon^*$, $\forall\, k \in K^*$. In the analogous way we show that there are not w^* and t_0^* as defined above. This proves continuity of $D_a(t;J)$ on R_i. The statement under a-2) is correct. Furthermore, $D_a(R_i;J)$ is neighbourhood of $x = 0$ by definition (Definition 4.1). Its connectedness is proved by contradiction. Let us assume that it is not connected. Then, there are disjoint sets D_{ak}, $k = 1,2,\ldots,N$, such that $D_a(R_i;J) = \cup[D_{ak}\colon k = 1,2,\ldots,N]$. One of D_{ak} is not a neighbourhood of J. Let it be D_{a1} and let D_{am}, $D_{am} \subset D_a(R_i;J)$, $m \in \{2,3,\ldots,N\}$, be connected neighbourhood of J that is possible because J is a compact connected set. Then $x_0 \in D_{a1}$ implies $\boldsymbol{x}(t;t_0,x_0) \to J$ as $t \to \infty$, $\forall\, t_0 \in R_i$. There is $t_1 \in R_0$ such that $\boldsymbol{x}(t_1;t_0,x_0) \notin D_a$ because of continuity of $\boldsymbol{x}(t;t_0,x_0)$ in $t \in R_0$, $\forall\, t_0 \in R_i$, and because D_{a1} is disjoint subset of $D_a(R_i;J)$, which is not neighbourhood of J. However, this is impossible due to $\boldsymbol{x}[t;t_1,x(t_1;t_0,x_0)] \equiv \boldsymbol{x}(t;t_0,x_0) \to J$ as $t \to \infty$. Hence, the assumption on disconnectedness of $D_a(R_i;J)$

is incorrect, which completes the proof of all the statements under a) by noting that invariance of $D_a(R_i;J)$ on R_i results directly from 1) in case $D_a(t;J) = D_a(R_i;J)$ for all $t \in R_i$.

b) The assertions under b) directly follow from those under a) in case $R_i = R$.

Lemma A.2

a) *If a compact connected invariant set J of system (1) possessing the weak smoothness property is asymptotically stable at $t_0 \in R_i$ and its domain of attraction $D_a(t_0;J)$ at $t_0 \in R_i$ obeys $D_a(t_0;J) \subseteq S(t_0;J)$ then its domains $D_a(t_0;J)$, $D_s(t_0;J)$ and $D(t_0;J)$ are interrelated by $D_a(t_0;J) \subseteq D_s(t_0;J)$ and $D(t_0;J) = D_a(t_0;J)$ for all $t_0 \in R_i$.*

b) *If a compact connected invariant set J of system (1) possessing the weak smoothness property is uniformly asymptotically stable on R_i and its domain $D_a(R_i;J)$ of uniform attraction on R_i satisfies $D_a(R_i;J) \subseteq S(R_i;J)$ then its domains $D_a(R_i;J)$, $D_s(R_i;J)$ and $D(R_i;J)$ are interrelated by $D_a(R_i;J) \subseteq D_s(R_i;J)$ and $D(R_i;J) = D_a(R_i;J)$.*

Proof We will follow the proof of Lemma A.2 of [38] in order to verify its validity also for non-differentiable time-varying non-linear systems.

Let system (1) possess the weak smoothness property and J be its compact connected invariant set.

a) Let the set J be asymptotically stable at t_0 and $D_a(t_0;J) \subseteq S(t_0;J)$. Let $x_0 \in D_a(t_0;J)$ be arbitrary. Time-invariance of J and continuity of $\boldsymbol{x}(t;t_0,x_0)$ in $(t,t_0,x_0) \in R_0 \times R_i \times S(t_0;J)$, $D_a(t_0;J) \subseteq S(t_0;J)$ and $x_0 \in D_a(t_0;J)$ imply $\max\{\rho[\boldsymbol{x}(t;t_0,x_0),J]\colon t \in R_0\} < \infty$. Let $\varepsilon = 2\max\{\rho[\boldsymbol{x}(t;t_0,x0),J]\colon t \in R_0\}$. Hence, $x_0 \in D_s(t_0,\varepsilon;J)$ due to (a-1) of Definition 4.2, which implies $x_0 \in D_s(t_0;J)$ in view of (a-3) of Definition 4.2. Altogether, $x_0 \in D_a(t_0;J)$ yields $x_0 \in D_s(t_0;J)$ that proves $D_a(t_0;J) \subseteq D_s(t_0;J)$ for all $t_0 \in R_i$. This result and (a) of Definition 4.3 complete the proof of the statement under (a).

b) Let the set J be uniformly asymptotically stable on R_i and $D_a(R_i;J) \subseteq S(R_i;J)$. Let $x_0 \in D_a(R_i;J)$ be arbitrary. Hence, $\max\{\rho[\boldsymbol{x}(t;t_0,x_0),J]\colon (t;t_0) \in R_0 \times R_i\} < \infty$ due to time invariance of J and continuity of $\boldsymbol{x}(t;t_0,x_0)$ in $(t,t_0,x_0) \in R_0 \times R_i \times D_a(R_i;J)$. Let $\varepsilon = 2\max\{\rho[\boldsymbol{x}(t;t_0,x_0),J]\colon (t;t_0) \in R_0 \times R_i\} \in R^+$ so that obviously $x_0 \in D_s(\varepsilon,R_i;J) = \cap[D_s(t_0,\varepsilon;J)\colon t_0 \in R_i] \neq \emptyset$. Therefore $x_0 \in D_s(R_i;J)$ (Definition 4.3). The result that $x_0 \in D_a(R_i;J)$ implies $x_0 \in D_s(R_i;J)$ proves $D_a(R_i;J) \subseteq D_s(R_i;J)$ and $D(R_i;J) = D_a(R_i;J)$ (due to Definition 4.3). This completes the proof.

References

[1] Lyapunov, A.M. *The General Problem of the Stability of Motion.* Mathematical Society, Kharkov, Russia, 1892. (Russian); *Academician A.M. Lyapunov: "Collected Papers".* USSR Academy of Science, Moscow, **II** (1956) 5–263. (Russian). [French translation: Problème général de la stabilité du mouvement. *Ann. Fac. Toulouse*, 2nd Ser., **9** (1907) 203–420; English translation: *Int. J. Control*, (1992) 531-719; also the book: Taylor & Francis, London, England, 1992].

[2] Bhatia, N.P. On asymptotic stability in dynamical systems. *Mathematical Systems Theory* **1**(2) (1967) 113–127.

[3] Bhatia, N.P. Attraction and nonsaddle sets in dynamical systems. *Journal of Differential Equations* **8**(2) (1970) 229–249.

[4] Bhatia, N.P. and Lazer, A.C. On global weak attractors in dynamical systems. In: *Differential Equations and Dynamical Systems*, (Eds.: J.K. Hale and J.P. LaSalle), Academic Press, New York, 1967, 321–325.

[5] Bhatia, N.P. and Szegö, G.P. *Dynamical Systems: Stability Theory and Applications.* Springer Verlag, Berlin, 1967.

[6] Corne, J.L. and Rouche, N. Attractivity of closed sets proved by using a family of Liapunov functions. *Journal of Differential Equations* **13**(2) (1973) 231–246.

[7] Hajek, O. Compactness and asymptotic stability. *Mathematical Systems Theory* **4**(2) (1970) 154–159.

[8] Hajek, O. Ordinary and asymptotic stability of noncompact sets. *Journal of Differential Equations* **11**(1) (1972) 49–65.

[9] Ladde, G.S., Lakshmikantham, V. and Leela, S. Conditionally asymptotically invariant sets and perturbed systems. *Annali di Matematica pura ed applicata, Ser. IV*, **XCIV**, Bologna, 1972, 33–40.

[10] Ladde, G.S. and Leela, S. Analysis of invariant sets. *Annali di Matematica pura ed applicata, Ser. IV*, **XCIV**, Bologna, 1972, 283–289.

[11] Ladde, G.S. and Leela, S. Global results and asymptotically self-invariant sets. *Rendiconti Della Classe di Scienze Fisiche, Matematiche e Naturali*, Academia Nazionale dei Lincei, (Ser. VIII), **LIV**(3), Roma, 1973, 321–327.

[12] Lakshmikantham, V. and Leela, S. Asymptotically self-invariant sets and conditional stability. In: *Differential Equations and Dynamical Systems*, (Eds.: J.K. Hale and J.P. LaSalle), Academic Press, New York, 1967, 363–319.

[13] Lakshmikantham, V. and Leela, S. *Differential and Integral Inequalities.* Academic Press, New York, 1969.

[14] LaSalle, J.P. *The Stability of Dynamical Systems.* SIAM, Philadelphia, 1976.

[15] Yoshizawa, T. Asymptotic behaviour of solutions of ordinary differential equations near sets. *Proc. International Symposium on Non-linear Oscillations*, **1**, Kiev, Ukraine, 1963, 213–225.

[16] Yoshizawa, T. Some notes on stability of sets and perturbed system. *Funkcialaj Ekvacioj* **6**(1) (1964) 1–11.

[17] Yoshizawa, T. Eventual properties and quasi-asymptotic stability of a non-compact set. *Funkcialaj Ekvacioj* **8**(2) (1966) 25–90.

[18] Yoshizawa, T. *Stability Theory by Liapunov's Second Method.* The Mathematical Society of Japan, Tokyo, 1966.

[19] Zubov, V.I. *Methods of A.M. Liapunov and their Applications.* P. Noordhoff Ltd., Groningen, 1964.

[20] Dauphin-Tanguy, G. and Grujić, Lj.T. Asymptotic stability via energy and power. Part II: Bond-graph bridging for non-linear systems. *Proc. IFAC Conference System Structure and Control*, Nantes, France, 1995, 96–101.

[21] Grujić, Lj.T. Solutions to Lyapunov stability problems: Non-linear systems with globally differentiable motions. In: *The Lyapunov Functions Method and Applications*, (Eds.: P. Borne and V. Matrosov), IMACS, J.C.Baltzer AG, Scientific Publishing Co., Basel, 1990, 19–27.

[22] Grujić, Lj.T. Solutions to Lyapunov stability problems: non-linear systems with differentiable motions. *Proc. 13th IMACS World Congress on Computation and Applied Mathematics*, **3**, 1991, 1228–1231.

[23] Grujić, Lj.T. The necessary and sufficient conditions for the exact construction of a Lyapunov function and the asymptotic stability domain. *Proc. 30th IEEE Conference on Decision and Control*, **3**, 1991, 2885–2888.

[24] Grujić, Lj.T. On solutions of Lyapunov stability problems. *Facta Universitatis, Ser.: Mechanics, Automatic Control and Robotics*, University of Nish, Nish, Serbia, **I**(2) (1992) 121–138.

[25] Grujić, Lj.T. Solutions to Lyapunov stability problems: non-linear systems with differentiable motions. In: *Computational and Applied Mathematics*, (Eds.: W.F. Ames and P.J. Van Der Houwen), **II**, Elsevier Science Publishers B.V. (North Holland), 1992, 39–47.

[26] Grujić, Lj.T. Exact both construction of Lyapunov function and asymptotic stability domain determination. *IEEE International Conference on Systems, Man and Cybernetics*, Le Touquet, France, **1** (1993) 331–336.

[27] Grujić, Lj.T. Solutions to Lyapunov stability problems: non-linear systems with continuous motions. *International Journal of Mathematics and Mathematical Sciences* **17**(3) (1994) 587–596.

[28] Grujić, Lj.T. Solutions to Lyapunov stability problems: Time-invariant systems. *Proc. 14th IMACS World Congress on Computation and Applied Mathematics*, **1**, 1994, 203–205.

[29] Grujić, Lj.T. Exact solutions for asymptotic stability: Non-linear systems. *Int. J. Non-Linear Mechanics* **30**(1) (1995) 45–56.

[30] Grujić, Lj.T. Solutions to Lyapunov stability problems via O-uniquely bounded sets. *Control-Theory and Advanced Technology 10*(4) (1995) 1069–1091.

[31] Grujić, Lj.T. New Lyapunov method based methodology for asymptotic stability. *Proc. IFAC Conference System Structure and Control*, Nantes, 1995, 536–541.

[32] Grujić, Lj.T. and Dauphin-Tanguy, G. Asymptotic stability via energy and power. Part I: New Lyapunov methodology for non-linear systems. *Proc. IFAC Conference System Structure and Control.*, Nantes, 1995, 548–553.

[33] Grujić, Lj.T. Complete exact solution to the Lyapunov stability problem: Time-varying non-linear systems with differentiable motions. *Non-linear Analysis, Theory, Methods & Applications* **22**(8) (1994) 971–981.

[34] Grujić, Lj.T. Time-varying continuous non-linear systems: uniform asymptotic stability. *International Journal of Systems Science* **26**(5) (1995) 1103–1127. Corrigendum, *Ibid* **27**(7) (1996) 689.

[35] Grujić, Lj.T. New approach to asymptotic stability: time-varying non-linear systems. *International Journal of Mathematics and Mathematical Sciences* **20**(2) (1997) 347–366.

[36] Grujić, Lj.T. Solutions to Lyapunov stability problems of sets: non-linear systems with differentiable motions. *International Journal of Mathematics and Mathematical Sciences* **17**(1) (1994) 103–112.

[37] Gruyitch, Ly.T. Consistent Lyapunov methodology for time-invariant non-linear systems. *Avtomatika i Telemehanika* **12** (1997) 35–73. (Russian).

[38] Gruyitch, Ly.T. Consistent Lyapunov methodology, time-varying non-linear systems and sets. *Nonlinear Analysis* **39** (2000) 413–446.

[39] Hahn, W. *Stability of Motion.* Springer Verlag, Berlin, 1967.

[40] Grujić, Lj. Novel development of Lyapunov stability of motion. *Int. Journal of Control* **22**(4) (1975) 525–549.

[41] Grujić, Lj.T. Concepts of stability domains. *Automatika (Zagreb)* **26**(1-2) (1985), 5–10 (Serbo-Croatian).

[42] Grujić, Lj.T., Martynyuk, A.A. and Ribbens-Pavella, M. *Large-Scale Systems Stability under Singular and Structural Perturbations.* Naukova Dumka, Kiev, 1984. (Russian); English translation: Springer Verlag, New York, 1987.

[43] Grujić, Lj.T. and Michel, A.N. Exponential stability and trajectory bounds of neural networks under structural variations. *IEEE Transactions on Circuits and Systems* **38**(10) (1991) 1182–1192.

[44] Grujić, Lj.T. and Ribbens-Pavella, M. Asymptotic stability of large-scale systems with application to power systems: Part I: domain estimation. *Elec. Power and Energy Systems* **1**(3) (1979) 151–157.

[45] Vanelli, A. and Vidyasagar, M. Maximal Lyapunov functions and domains of attraction for autonomous non-linear systems. *Automatica* **21**(1) (1985) 69–26.

[46] Miller, R.K. and Michel, A.N. *Ordinary Differential Equations.* Academic Press, New York, 1982.

[47] Grujić, Lj. On absolute stability and the Aizerman conjecture. *Automatica* **17**(2) (1981) 335–349.

[48] Grujić, Lj.T. Reply to "Comments on 'On absolute stability and the Aizerman conjecture'". *Automatica* **29**(2) (1993) 561.

[49] Krasovskii, N.N. *Stability of Motion.* Stanford University Press, Stanford, 1963.

[50] Rouche, N., Habets, P. and Laloy, M. *Stability Theory by Liapunov's Direct Method.* Springer Verlag, New York, 1977.

[51] Demidovich, B.P. *Lectures on the Mathematical Theory of Stability.* Nauka, Moscow, 1967. (Russian).

[52] Halanay, A. *Differential Equations.* Academic Press, New York, 1966.

[53] Hale, J.K. *Ordinary Differential Equations.* Wiley-Interscience, New York, 1969.

[54] Kalman, R.E. and Bertram, J.E. Control system analysis and design via the 'second method' of Lyapunov, Part I. *Trans. of ASME-J. Basic Eng.* **82** (1960) 317–393.

Nonlinear Dynamics and Systems Theory, 1(1) (2001) 23–37

On Three Definitions of Chaos

Bernd Aulbach and Bernd Kieninger

Department of Mathematics, University of Augsburg, D-86135 Augsburg, Germany

Received: April 20, 2000; Revised: July 18, 2000

Abstract: We discuss in this paper three notions of chaos which are commonly used in the mathematical literature, namely those being introduced by Li & Yorke, Block & Coppel and Devaney, respectively. We in particular show that for continuous mappings of a compact interval into itself the notions of chaos due to Block & Coppel and Devaney are equivalent while each of these is sufficient but not necessary for chaos in the sense of Li & Yorke. We also give an example indicating that in the general context of continuous mappings between compact metric spaces the relation between these three notions of chaos is more involved.

Keywords: *Iteration; chaos; chaotic map.*

Mathematics Subject Classification (2000): 26A18, 58F13, 58F08.

1 Introduction

As a mathematical notion the term *chaos* has first been used in 1975 by Li & Yorke in their paper [13] "Period three implies chaos", but even before it has been observed that very simple functions may give rise to very complicated dynamics. One of the cornerstones in the development of *chaotic dynamics* is the 1964 paper [15] "Coexistence of cycles of a continuous mapping of the line into itself" (in Russian) by Šarkovskii. During the seventies and eighties the interest in *chaotic dynamics* has been exploding and various attempts have been made to give the notion of *chaos* a mathematically precise meaning. Outstanding works in this context are the 1980 book [6] "Iterated Maps on the Interval as Dynamical Systems" by Collet & Eckmann, the 1989 book [16] "Dynamics of One-dimensional Mappings" (in Russian) by Šarkovskii, Kolyada, Sivak & Fedorenko and the 1992 Lecture Notes [4] "Dynamics in One Dimension" by Block & Coppel. While up to the end of the eighties the subject of *chaotic dynamics* was restricted mainly to research oriented publications, the 1986 book [7] "An Introduction to Chaotic Dynamical Systems" by Devaney marked the point where *chaos* (as a mathematical notion) became popular and began to enter university textbooks such as [9] "A First Course in Discrete Dynamical Systems" by Holmgren (1994) or [8] "Discrete Chaos" by Elaydi (1999).

© 2001 Informath Publishing Group. All rights reserved.

The different definitions of *chaos* being around at the turn of the century have been designed to meet different purposes and they are based on very different backgrounds and levels of mathematical sophistication. Therefore it is not obvious how these notions of *chaos* relate to each other and whether there is a chance that – in the long run – a universally accepted definition of *chaos* might evolve. With this paper we want to make a contribution to this question by picking three of the most popular definitions of *chaos* and investigating their mutual interconnections. After listing the technical prerequisites in Section 2, in Section 3 we give the precise definitions of the notions of *chaos* due to Li & Yorke [13], Block & Coppel [4] and Devaney [7], respectively. In Section 4 we then show that in the case of continuous maps on a compact interval the notions of *chaos* in the sense of Block & Coppel and Devaney are equivalent and that, on the other hand, each of these two notions is sufficient for *chaos* in the sense of Li & Yorke. In Section 5 we discuss the familiy of so-called truncated tent maps by means of which we in particular demonstate that *chaos* in the sense of Block & Coppel and Devaney is not necessary for *chaos* in the sense of Li & Yorke. Finally, in Section 6 we indicate by means of an example that the previously described simple relations for interval maps do not carry over to maps between general compact metric spaces. In fact, we exhibit a map which is chaotic both in the sense of Li & Yorke and Block & Coppel but not chaotic in the sense of Devaney.

Before going into detail we want to mention that the majority of results we describe and prove in this paper can be found – more or less explicitly stated – in the literature, in particular in the Lecture Notes [4] of Block & Coppel. Since those results, on the other hand, appear in the broad context of *discrete dynamics* and since they partly are formulated with notations which subtly differ from each other it is the purpose of this paper to narrow the view and sketch a clearer picture of the notion of *chaos* in a unified way.

2 Preliminaries

In this section we collect some notation and a few facts from topological and symbolic dynamics which are used in this paper.

Throughout this paper let (X, d) be a compact metric space and $f\colon X \to X$ a continuous mapping. A nonempty subset Y of X satisfying $f(Y) \subseteq Y$ is said to be (f-)*invariant* and it is called *strongly* (f-)*invariant* if $f(Y) = Y$. The set Y is called *minimal* if it is compact and does not contain any nonempty compact invariant proper subset. By f^n, $n \in \mathbb{N}$, we denote the mapping which is defined recursively by $f \circ f^{n-1}$ and $f^0 = \mathrm{id}$. The sequence $\gamma(x, f) := (f^n(x))_{n \geq 0}$ is the *trajectory* of $x \in X$, the set $\mathrm{O}(x, f) := \{f^n(x) \mid n \geq 0\}$ the (*forward*-)*orbit* of x and $\omega(x, f) := \bigcap\limits_{m \geq 0} \overline{\mathrm{O}(f^m(x), f)}$ is the (ω-)*limit set* of x.

A point $x \in X$ and its orbit is *periodic* if $f^n(x) = x$ for some $n \in \mathbb{N}$, x is called *recurrent* if $x \in \omega(x, f)$. $\mathrm{P}(f)$ denotes the set of periodic points in X and $\mathrm{R}(f)$ the set of recurrent points. A point $x \in X$ is *finally periodic* if $f^n(x)$ is periodic for some $n \in \mathbb{N}$, it is *asymptotically periodic* if there exists a periodic point $z \in X$ such that $\lim\limits_{n \to \infty} d(f^n(x), f^n(z)) = 0$, and it is *approximately periodic* if for any $\varepsilon > 0$ there exists a periodic point $z \in X$ such that $\limsup\limits_{n \to \infty} d(f^n(x), f^n(z)) < \varepsilon$.

If $\omega(x, f) = X$ for some $x \in X$, the mapping f is called (*topologically*) *transitive.* The map f is said to have *sensitive dependence on initial conditions* if there exists a $\delta > 0$

such that for any $x \in X$ and any $\varepsilon > 0$ there exists a $y \in \{z \in X \mid d(x,z) < \varepsilon\}$ and an $n \in \mathbb{N}$ with $d(f^n(x), f^n(z)) > \delta$.

In the sequel we need a few results on the notions just introduced. They are described in the following remarks.

Remarks 2.1

(1) X always contains a minimal subset Y (see [4, V Lemma 3]) and $Y \subseteq X$ is minimal if and only if $\omega(x,f) = Y$ for all $x \in Y$ (see [4, V Lemma 1]).

(2) Any limit set $\omega(x,f)$ is nonempty, compact and strongly invariant (see [4, IV Lemma 2]), $\omega(x,f) = \bigcup\limits_{i=0}^{n-1} \omega(f^i(x), f^n)$ and $f(\omega(x,f^n)) = \omega(f(x), f^n)$ for any $n \in \mathbb{N}$ (see [4, p.70/71]).

(3) $\mathrm{P}(f) = \mathrm{P}(f^n)$ and $\mathrm{R}(f) = \mathrm{R}(f^n)$ for all $n \in \mathbb{N}$ (see [4, I Lemma 10 and IV Lemma 25]).

(4) f is transitive if and only if for any two open sets U and V in X there exists an $n \in \mathbb{N}$ such that $f^n(U) \cap V \neq \emptyset$ (see [4, VI Prop. 39]).

Let $\Sigma := \{\alpha = (a_0, a_1, \dots) \mid a_i \in \{0,1\}\}$ be the space of sequences with entries 0 or 1. The map $d_\Sigma \colon \Sigma \times \Sigma \to \mathbb{R}$ defined by $(\alpha,\beta) \mapsto 0$ if $\alpha = \beta$ and $(\alpha,\beta) \mapsto 2^{-j}$ if $\alpha = (a_0, a_1, \dots) \neq (b_0, b_1, \dots) = \beta$ and $a_i = b_i$ for $i = 0, \dots, j-1$ and $a_j \neq b_j$ defines a metric on Σ under which Σ becomes a cantor set [4, p.34]. The *shift* operator $\sigma \colon \Sigma \to \Sigma$ is defined by $(a_0, a_1, \dots) \mapsto (a_1, a_2, \dots)$. This shift operator is continuous and surjective (see e.g. [7, Proposition 6.5]).

Let (Y, d_Y) be another compact metric space and let $g \colon Y \to Y$ be continuous. If there exists a continuous surjection $h \colon X \to Y$ such that $h \circ f = g \circ h$ on X then f is said to be (*topologically*) *semi-conjugate* to g via the (*topological*) *semi-conjugacy* h.

3 Three Definitions of Chaos

We now describe three commonly used definitions of chaos for continuous maps. To begin with we consider the general case of maps from a compact metric space (X,d) into itself and later we concentrate on the special case where X is a real interval.

Definition 3.1 [L/Y-chaos] A continuous map $f \colon X \to X$ on a compact metric space (X,d) is called *chaotic in the sense of Li and Yorke* – or just *L/Y-chaotic* – if there exists an uncountable subset S (called a *scrambled set*) of X with the following properties:

(i) $\limsup\limits_{n\to\infty} d(f^n(x), f^n(y)) > 0$ for all $x, y \in S,\ x \neq y$,

(ii) $\liminf\limits_{n\to\infty} d(f^n(x), f^n(y)) = 0$ for all $x, y \in S,\ x \neq y$,

(ii) $\limsup\limits_{n\to\infty} d(f^n(x), f^n(p)) > 0$ for all $x \in S,\ p \in X,\ p$ periodic.

Remarks 3.1

(1) In Li & Yorke's original definition of chaos there is the additional condition that f has periodic points of any period in $\mathbb{N}$ [13, Theorem 1]. In the literature, however, most authors (see e.g. [12, Definition. 1.1]) refer to chaos in the sense of Li & Yorke without this condition.

(2) That condition (iii) in the definition of L/Y-chaos is redundant can be seen as follows. Two approximately periodic points x, y cannot satisfy both (i) and (ii) in the definition of a scrambled set (see [4, VI Lemma 28]). Consequently there exists at most one approximately periodic point in any set satisfying conditions (i) and (ii) of Definition 3.1. Removing this point the new set also satisfies (iii).

Definition 3.2 [B/C-chaos] A continuous map $f\colon X \to X$ on a compact metric space X is called *chaotic in the sense of Block and Coppel* – or just *B/C-chaotic* – if there exists an $m \in \mathbb{N}$ and a compact f^m-invariant subset Y of X such that $f^m|_Y$ is semi-conjugate to the shift on Σ, i.e. if there exists a continuous surjection $h\colon Y \to \Sigma$ satisfying

$$h \circ f^m = \sigma \circ h \quad \text{on } Y.$$

Remarks 3.2

(1) In [4, p.127/128] it has been described that the definition of B/C-chaos is equivalent to the property that there exist an $m \in \mathbb{N}$ and two compact disjoint sets X_0, X_1 in X such that given any $(a_0, a_1, \dots) \in \Sigma$ there is an $x \in X$ such that $f^{mn}(x) \in X_{a_n}$ for all $n \in \mathbb{N}_0$ (see also [10, Theorem 2.2.3]).

(2) If $m = 1$ in Defintion3.2 then this notion of chaos is also known as *chaos in the sense of coin tossing* [11, Definition 1]. It should be noted here that not every B/C-chaotic map is also chaotic in the sense of coin tossing. In order to see this consider the subset $T := \{(0, a_0, 0, a_1, 0, \dots) \mid (a_0, a_1, \dots) \in \Sigma\} \cup \{(a_0, 0, a_1, 0, a_2, \dots) \mid (a_0, a_1, \dots) \in \Sigma\}$ of Σ. This set is σ-invariant and the map $\sigma^2|_T$ is semi-conjugate to $\sigma\colon \Sigma \to \Sigma$. However, there is no σ-invariant subset W of T such that $\sigma|_W$ is semi-conjugate to $\sigma\colon \Sigma \to \Sigma$ (see [10, Example 2.2.5] for more details).

Definition 3.3 [D-chaos] A continuous map $f\colon X \to X$ on a compact metric space X is called *chaotic in the sense of Devaney* – or just *D-chaotic* – if there exists a compact invariant subset Y (called a *D-chaotic set*) of X with the following properties:

(i) $f|_Y$ is transitiv,

(ii) $\overline{\mathrm{P}(f|_Y)} = Y$,

(iii) $f|_Y$ has sensitive dependence on initial conditions.

Remarks 3.3

(1) In [7, Definition 8.5] Devaney originally defined f to be chaotic if $Y = X$ in Definition 3.3. In the literatur, however, D-chaos is usually meant in the more general sense with $Y \subseteq X$ (see e.g. [12, Definition 1.3]).

(2) Condition (iii) in Definition 3.3 turned out to be redundant in the nontrivial case where Y is infinite (see [2]).

The three notions of chaos just described have the nice property of being invariant under conjugation and iteration. The corresponding statements are as follows.

Proposition 3.1 *Let (X, d_X) and (Y, d_Y) be compact metric spaces and suppose that a continuous map $f\colon X \to X$ is conjugate to a continuous map $g\colon Y \to Y$. Then f is D-chaotic (or B/C-chaotic or L/Y-chaotic) if and only if g is D-chaotic (or B/C-chaotic or L/Y-chaotic, respectively).*

Proof See [11, Proposition 1].

Proposition 3.2 *Let (X,d) be compact and suppose $f\colon X \to X$ is continuous. Then for any $n \in \mathbb{N}$ the map f is D-chaotic (or B/C-chaotic or L/Y-chaotic) if and only if f^n is D-chaotic (or B/C-chaotic or L/Y-chaotic, respectively).*

Proof D-chaos: If f is D-chaotic with D-chaotic set $Y \subseteq X$ then there is an $x \in Y$ such that $\omega(x, f|_Y) = Y$. Since $\mathrm{R}(f|_Y) = \mathrm{R}(f^n|_Y)$ we have $x \in W := \omega(x, f^n|_Y)$, so $f^n|_W$ is transitive. Since $\mathrm{P}(f|_Y) = \mathrm{P}(f^n|_Y)$ and $Y = \bigcup\limits_{i=0}^{n-1} f^i(W)$ we get $\overline{\mathrm{P}(f^n|_W)} = \overline{\mathrm{P}(f^n|_Y)} \cap W = Y \cap W = W$ and W is infinite because Y is infinite ($f^n|_Y$ has sensitive dependence on initial conditions). So also $f^n|_W$ sensitively depends on initial conditions. Conversely, let $f^n|_W$ satisfy the conditions in the definition of D-chaos for an f^n-invariant compact set W in X. Then it is easy to see that also $f|_Y$ with $Y := \bigcup\limits_{i=0}^{n-1} f^i(W)$ satisfies these conditions (see also [12, Proposition 4.10]).

B/C-chaos: Let f be B/C-chaotic, $m \in \mathbb{N}$, $Y \subseteq X$ compact and f^m-invariant and let $f^m|_Y$ be semi-conjugate to σ via $h\colon Y \to \Sigma$. Then defining the continuous surjection $t\colon \Sigma \to \Sigma$ by $(a_0, a_1, a_2, \dots) \mapsto (a_0, a_n, a_{2n}, \dots)$ and $\bar{h} := t \circ h$ we get $\bar{h} \circ (f^n)^m = \sigma \circ \bar{h}$ on Y and f^n is B/C-chaotic. The converse immediately follows from the definition of B/C-chaos.

L/Y-chaos: It is easy to see that a set $S \subseteq X$ is a scrambled set with respect to f if and only if it is a scrambled set with respect to f^n (see also [10, Proposition 2.3.8]). This completes the proof of Proposition 3.2.

For the remainder of this section we concentrate on the special case where the compact metric space X is a nontrivial real interval I (i.e. nonempty and not a singleton). In this case B/C-chaos originally (see [4, p.33]) has been defined differently from Definition 3.2. In fact, the original definition of a B/C-chaotic map was based on the notion of *tubulence* which is defined as follows (see [4, p.25]). A map $f\colon I \to I$ is called *turbulent* if there exist compact subintervals J, K of I with at most one common point such that $J \cup K \subseteq f(J) \cap f(K)$. If J and K can be chosen disjoint then f is said to be *strictly turbulent.* The relation between turbulence and B/C-chaos is described in the following result (see [4, p.33/128]).

Proposition 3.3 *A continuous map $f\colon I \to I$ on a nontrivial compact interval I is B/C-chaotic if and only if one of the following equivalent conditions is satisfied:*

(i) *f^m is turbulent for some $m \in \mathbb{N}$,*
(ii) *f^m is strictly turbulent for some $m \in \mathbb{N}$,*
(iii) *f has a periodic point whose period is not a power of* 2.

Three more results are needed in order to reach the goals of this paper.

Proposition 3.4 *f is L/Y-chaotic if and only if not every point in I is approximately periodic.*

Proof See [4, p.145]).

Lemma 3.1 *f is B/C-chaotic if and only if there exists a $c \in I$ such that $\omega(c, f)$ contains a periodic orbit as a proper subset.*

Proof See [4, VI Proposition 6]).

Lemma 3.2 *Let J and K be two compact subintervals of I having the property $K \subseteq f(J)$. Then there exists a compact subinterval L of J such that $f(L) = K$ and that f maps the endpoints of L onto the endpoints of K.*

Proof Let $K = [a, b]$ for two points $a, b \in I$ and let c be the largest point in J with $f(c) = a$. If there exists an $x \in J$, $x > c$, with $f(x) = b$, let d be the smallest x with this property. Then with $L := [c, d]$ the claim follows. On the other hand, if there exists an $x \in J$, $x < c$, with $f(x) = b$ we define $\tilde{c}$ as the largest x with this property. Let $\tilde{d}$ be the smallest $x \in (\tilde{c}, c]$ $(\subset J)$ satisfying $f(x) = a$. Then the interval $L := [\tilde{c}, \tilde{d}]$ has the claimed property and the proof of the lemma is complete.

4 The Mutual Relations for Interval Maps

In this section we discuss the mutual relations between the three notions of chaos described in the previous section for the special case of interval maps. In fact, throughout the present section we consider continuous maps $f\colon I \to I$ from a nontrivial compact interval $I = [a, b]$, $a < b$, into itself. The main results of this section say that in this case B/C-chaos and D-chaos are equivalent while, on the other hand, B/C-chaos and D-chaos are sufficient for L/Y-chaos. For the lacking necessity in the last statement we refer to the counterexample g_{λ^*} presented in the next section.

Theorem 4.1 *A continuous map $f\colon I \to I$ on an interval I is D-chaotic if and only if it is B/C-chaotic.*

Proof ($\Rightarrow$) Let f be D-chaotic with compact D-chaotic set $Y \subseteq I$. Then Y is infinite since $f|_Y$ has sensitive dependence on initial conditions. Furthermore, since $f|_Y$ is transitive there is a $c \in Y$ with $\omega(c, f) = Y$, and because of the relation $\overline{\mathrm{P}(f|_Y)} = Y$ the map $f|_Y$ has a periodic orbit. As a finite set this periodic orbit is a proper subset of $Y = \omega(c, f)$, and this implies (by Lemma 3.1) that f is B/C-chaotic.

($\Leftarrow$) Now suppose f is B/C-chaotic. Then because of Proposition 3.3 the map f^m is strongly turbulent for some $m \in \mathbb{N}$, i.e. there exist two disjoint compact subintervals X_0 and X_1 of I with the property that for $g := f^m$ we have

$$X_0 \cup X_1 \subseteq g(X_0) \cap g(X_1). \tag{1}$$

The idea of proceeding from here is to first derive from (1) the existence of a compact g-invariant subset X of $X_0 \cup X_1$ with the property that the map $g|_X\colon X \to X$ is semi-conjugate to the shift σ via a continuous surjection $s\colon X \to \Sigma$ and then to show that there exists a compact g-invariant subset Z of X on which g is D-chaotic. We carry out this program in 5 steps.

Step 1. Construction of X and s: Starting with the above X_0 and X_1 and using mathematical induction, for each $\alpha = (a_1, a_2, \dots) \in \Sigma$ Lemma 3.2 yields a sequence of compact, pairwise disjoint intervals $X_{a_1 a_2 \dots a_k}$, $(a_1, a_2, \dots, a_k) \in \{0,1\}^k$, $k \geq 1$ in $X_0 \cup X_1$ having the following properties:

$$\begin{gathered} X_{a_1 a_2 \dots a_k} \subseteq X_{a_1 a_2 \dots a_{k-1}}, \quad g(X_{a_1 a_2 \dots a_k}) = X_{a_2 a_3 \dots a_k} \quad \text{and} \\ g \text{ maps endpoints of } X_{a_1 a_2 \dots a_k} \text{ onto endpoints of } X_{a_2 a_3 \dots a_k}. \end{gathered} \tag{2}$$

Then for each $\alpha=(a_1,a_2,\dots)\in\Sigma$ the set

$$X_\alpha := \bigcap_{k=1}^{\infty} X_{a_1\dots a_k} \tag{3}$$

is either a singleton or a nontrivial compact interval. Furthermore we have

$$X_\alpha \cap X_\beta = \emptyset \quad \text{for all} \quad \alpha,\, \beta\in\Sigma, \quad \alpha\neq\beta \tag{4}$$

since the sets $X_{a_1a_2\dots a_k}$, $(a_1,a_2,\dots,a_k)\in\{0,1\}^k$, are pairwise disjoint and

$$g(X_\alpha) = X_{\sigma(\alpha)} \quad \text{for all} \quad \alpha\in\Sigma. \tag{5}$$

Next we define the set

$$\tilde{X} := \bigcup_{\alpha\in\Sigma} X_\alpha \tag{6}$$

which turns out to be strongly g-invariant and compact. Also the set

$$X := \{x\in I \mid x \ \text{ is an endpoint of } X_\alpha \text{ for some } \alpha\in\Sigma\}$$

is compact (even if $X_\alpha=\{x\}$ for some x we call x an endpoint of X_α). From (2) and (5) we conclude that for any $\alpha\in\Sigma$ the map g maps the endpoints of X_α onto the endpoints of $X_{\sigma(\alpha)}$ and that X is strongly g-invariant. On X we define the map

$$s\colon X\to\Sigma, \quad x\mapsto\alpha \quad \text{if} \quad x\in X_\alpha.$$

Obviously, this map is well defined, continuous and onto and each point of Σ is the s-image of at most two points of X. Finally, because of (5) and the definition of s we have

$$s\circ g|_X = \sigma\circ s \quad \text{on } X. \tag{7}$$

Step 2. Construction of Z: For any $\alpha\in\Sigma$ the set X_α defined in (3) is a nonempty compact interval. Since the X_α's are pairwise disjoint (see (4)) there exist at most countably many α's in Σ such that X_α is not a singleton. Therefore the set

$$R := \{x\in X \mid X_\alpha=\{x\} \ \text{for some } \alpha\in\Sigma\}$$

is nonempty and consists of all but countably many points of X. Because of (5) the set R is g-invariant and the set

$$Z := \bar{R}$$

and the map $g|_Z\colon Z\to Z$ are well defined.

Step 3. Transitivity of $g|_Z$: Let U be an arbitrary open nonempty subset of Z. Then there exists a point $x\in U\cap R$ and some $\alpha=(a_1,a_2,\dots)\in\Sigma$ with $X_\alpha=\{x\}$. Because of definition (3) of X_α and the openness of U in Z there exists a $k\in\mathbb{N}$ with $Z\cap X_{a_1a_2\dots a_k}\subseteq U$. Therefore, in order to prove the transitivity of $g|_Z$ it suffices to prove the relation

$$g^k(Z\cap X_{a_1a_2\dots a_k}) = Z. \tag{8}$$

Since $Z = \bar{R}$ and since the set $Z \cap X_{a_1 a_2 \dots a_k}$ is compact, it even suffices to find a g^k-preimage of an arbitrary point $y \in R$ in the set $Z \cap X_{a_1 a_2 \dots a_k}$. Due to the definition of R, for any $y \in R$ there exists a $\beta = (b_1, b_2, \dots) \in \Sigma$ with $\{y\} = X_\beta$. With the aid of this β we define

$$\gamma := (a_1, a_2, \dots , a_k, b_1, b_2, \dots) \in \Sigma$$

and use (5) to get the relation

$$g^k(X_\gamma) = X_\beta = \{y\}. \tag{9}$$

If X_γ consists of a single point we get the inclusion $X_\gamma \subseteq Z$ and the claim (8) is proved, since X_γ is a subset of $X_{a_1 a_2 \dots a_k}$. If, on the other hand, X_γ is a nontrivial interval then at least one of its endpoints is contained in Z. This can be shown as follows: For any $n \in \mathbb{N}$ there exists (because of (3)) a number $m_n \in \mathbb{N}$ with

$$X_{a_1 a_2 \dots a_k b_1 b_2 \dots b_{m_n}} \subseteq \{x \in I \mid \operatorname{dist}(x, X_\gamma) < \tfrac{1}{n}\}, \tag{10}$$

and since the set

$$\{(a_1, a_2, \dots , a_k, b_1, b_2, \dots , b_{m_n}, *, *, \dots) \in \Sigma \mid * \in \{0, 1\}\}$$

is uncountable we can find a point γ_n in this set such that $X_{\gamma_n} = \{y_n\}$. By (4) the sets X_γ and X_{γ_n} are disjoint and by (10) the distance of the point y_n from at least one of the endpoints of X_γ is less than $\frac{1}{n}$ (since $y_n \in X_{a_1 a_2 \dots a_k b_1 b_2 \dots b_{m_n}}$). Because the relation $X_{\gamma_n} \subset R$ holds for all $n \in \mathbb{N}$, the sequence $(y_n)_{n \in \mathbb{N}}$ in R converges, w.l.o.g., to one of the endpoints of X_γ. On the other hand, because of $Z = \bar{R}$ this endpoint is contained in Z and it is mapped via g^k to y according to (9). In both cases we thus can find a g^k-preimage of the point y in $Z \cap X_{a_1 a_2 \dots a_k}$, and this proves claim (8).

Step 4. $\overline{\mathrm{P}(g|_Z)} = Z$: Let U again be an arbitrary open nonempty set in Z and x a point in $U \cap R$ with $\{x\} = X_\alpha$ for some $\alpha = (a_1, a_2, \dots) \in \Sigma$. As in the proof of Step 3, given any $n \in \mathbb{N}$ there is an $m_n \in \mathbb{N}$ with

$$X_{a_1 a_2 \dots a_{m_n}} \subseteq \{x \in I \mid \operatorname{dist}(x, X_\alpha) < \tfrac{1}{n}\}. \tag{11}$$

We now consider the periodic point

$$\gamma_n := (\overline{a_1, a_2, \dots , a_{m_n}}, \dots) \in \Sigma$$

and notice that because of $\sigma^{m_n}(\gamma_n) = \gamma_n$ and (5) we get $g^{m_n}(X_{\gamma_n}) = X_{\gamma_n}$. Furthermore, the two endpoints of X_{γ_n} are periodic with respect to g, since g^{m_n} maps the endpoints of X_{γ_n} onto the endpoints of $g^{m_n}(X_{\gamma_n})$ $(= X_{\gamma_n})$. In case X_{γ_n} is a nontrivial interval then at least one of its (periodic) endpoints is contained in Z. This can be seen as in the previous Step 3. So in any case, for any $n \in \mathbb{N}$ we get a g-periodic point $x_n \in X_{\gamma_n} \cap Z$ and the sequence $(x_n)_{n \in \mathbb{N}}$ converges to x because of $X_{\gamma_n} \subseteq X_{a_1 \dots a_{m_n}}$ and (11). This implies the relation $x \in \overline{\mathrm{P}(g|_Z)}$ and completes the proof of Step 4.

Step 5. Conclusion: The set Z is infinite (because R is infinite), and therefore the map g is D-chaotic on Z. By Proposition 3.2 then f is D-chaotic on $Y := \bigcup_{i=0}^{n-1} f^i(Z)$. This completes the proof of Theorem 4.1.

Remark 4.1 Theorem 4.1 can also be proved by using the notion of *positive entropy* (see [4, VIII] and [12]). The argument is as follows: The map f is D-chaotic if and only if it has positive entropy (see [12]) and positive entropy in turn is equivalent to B/C-chaos (see e.g. [4, VII Theorem 24]).

Theorem 4.2 *If a continuous map $f\colon I \to I$ on an interval I is B/C-chaotic then it is also L/Y-chaotic.*

Proof See [4, VI Proposition 27]).

As mentioned above the implication of Theorem 4.2 cannot be reversed. In fact, in the next section we present an example of a map which is L/Y-chaotic but not B/C-chaotic. Before doing this, however, we want to show that all maps which are L/Y- but not B/C-chaotic have an interesting property in common. In fact, we show that for any map of this kind there exists an infinite compact invariant set such that the restriction of the map to this set is transitive but does not have periodic points. This shows that the result "On intervals transitivity = chaos" [3] (earlier proved in [4, VI Lemma 41] and stating that $\overline{\mathrm{P}(f)} = I$ if f is transitive) cannot be generalized from intervals to disconnected compact subsets of $\mathbb{R}$.

Proposition 4.1 *If a continuous map $f\colon I \to I$ on a compact interval I is L/Y-chaotic but not B/C-chaotic then there exists a compact infinite invariant subset Y of I such that $f|_Y\colon Y \to Y$ is transitive but does not have periodic points.*

Proof Since f is L/Y-chaotic there exists (by Proposition 3.4) a point $x \in I$ which is not approximately periodic. So the limit set $\omega(x, f)$ of x is infinite by [4, IV Lemma 4]. According to [4, VI Proposition 7] there exists a unique minimal set Y in $\omega(x, f)$ such that

$$Y = \omega(y, f) \quad \text{for some} \quad y \in Y. \tag{12}$$

Again using [4, IV Lemma 4] the set Y (as a subset of $\omega(x, f)$) is infinite. This is because f is not B/C-chaotic and therefore the infinite set $\omega(x, f)$ contains no periodic points by Proposition 3.4. So finally the well defined map $f|_Y\colon Y \to Y$ has no periodic points but is transitive by (12).

5 The Family of Truncated Tent Maps

In order to analyse the so-called *truncated tent maps* we need a famous theorem due to Šarkovskii and two lemmas. In any case we consider a continuous map $f\colon I \to I$ of a compact interval into itself.

Theorem 5.1 [Šarkovskii] *Let $\mathbb{N}$ be totally ordered in the following way:*

$$3 \prec 5 \prec 7 \prec \ldots \prec 3 \cdot 2 \prec 5 \cdot 2 \prec 7 \cdot 2 \prec \ldots \prec 3 \cdot 2^2 \prec 5 \cdot 2^2 \prec \ldots \prec 2^3 \prec 2^2 \prec 2 \prec 1.$$

Then if f has a periodic orbit of period $n \in \mathbb{N}$ and if $m \in \mathbb{N}$ with $n \prec m$, then f also has a periodic orbit of period m.

Proof See [15], also [4, I Theorem 1].

Lemma 5.1 *Suppose $f\colon I \to I$ is not B/C-chaotic. Then for any $x \in I$ with infinite $\omega(x, f)$ and any $s \in \mathbb{N}$ the intervals*

$$J_i^s := [\min \omega(f^i(x), f^{2^s}),\ \max \omega(f^i(x), f^{2^s})], \quad i = 0, 1, \ldots, 2^s - 1$$

have the following properties:

(i) *$J_i^s \cap J_k^s = \emptyset$ for all $i, k \in \{0, 1, \ldots, 2^s - 1\}$ with $i \neq k$,*

(ii) *J_i^s contains a 2^s-periodic point for $i = 0, 1, \ldots, 2^s - 1$.*

Proof See [4, VI Lemma 14].

Lemma 5.2 *If f is L/Y-chaotic then f has infinitely many periodic points.*

Proof We distinguish the two cases of f being B/C-chaotic or not.

If f is B/C-chaotic then by Proposition 3.3 f has an n-periodic point with n not being a power of 2. By Šarkovskii's Theorem then f has periodic points with periods 2^n for all $n \in \mathbb{N}$. So f has infinitely many periodic points.

In case f is not B/C-chaotic then there exists a point $x \in I$ which is not approximately periodic by Proposition 3.4. Therefore the limit set $\omega(x, f)$ of x is infinite (see [4, IV Lemma 4]). Now take any $s \in \mathbb{N}$ and define

$$J_i^s := [\min \omega(f^i(x), f^{2^s}),\ \max \omega(f^i(x), f^{2^s})] \quad \text{for} \quad i = 0, 1, \ldots, 2^s - 1.$$

Then the intervals J_i^s are pairwise disjoint and each of them contains a 2^s-periodic point by Lemma 5.1. Therefore f has at least 2^s distinct periodic points and hence infinitely many since $s \in \mathbb{N}$ was arbitrary. This completes the proof of Lemma 5.2.

Now we are prepared to investigate the announced family of maps one member of which shows that the statement of Theorem 4.2 cannot be reversed.

Example 5.1 The piecewise linear map $g\colon [0,1] \to [0,1]$ with

$$g(0) = 0, \quad g\big(\tfrac{1}{2}\big) = 1, \quad g(1) = 0$$

is known as the (*standard*) *tent map.* Its graph is a "tent" with peak of height 1 at the point $\frac{1}{2}$. In order to modify this map to get a family of maps suitable for our purposes we cut the peak at any height $\lambda \in [0,1]$ and consider the family of *truncated tent maps* defined by

$$g_\lambda\colon [0,1] \to [0,1], \quad x \mapsto \min\{\lambda, g(x)\}, \quad \lambda \in [0,1].$$

It is apparent that for any $0 \le \lambda < \gamma \le 1$ the maps g_λ and g_γ coincide on the set

$$J_\lambda := \big[0, \tfrac{\lambda}{2}\big] \cup \big[1 - \tfrac{\lambda}{2}, 1\big], \quad \lambda \in [0,1]$$

and that (periodic) orbits of g_γ in J_λ are also (periodic) orbits of g_λ and vice versa. Furthermore, since g_λ is constant on the open interval

$$K_\lambda := \big(\tfrac{\lambda}{2}, 1 - \tfrac{\lambda}{2}\big), \quad \lambda \in [0,1],$$

the map g_λ has at most one periodic point in $\bar{K}_\lambda$.

For the original tent map $g\ (= g_1)$ the set $\big\{\frac{2}{7}, \frac{4}{7}, \frac{6}{7}\big\}$ is obviously a 3-periodic orbit and therefore, by Šarkovskii's Theorem, it has 2^n-periodic points for all $n \in \mathbb{N}$. Furthermore, it is easy to see that

$$|\{x \in [0,1] \mid x \text{ is } m\text{-periodic with respect to } g\}| \le 2^m \text{ for all } m \in \mathbb{N}. \tag{13}$$

Therefore the number

$$\lambda_n := \min\{\lambda \in [0,1] \mid g \text{ has a } 2^n\text{-periodic orbit in } [0,\lambda]\}$$

is well defined and λ_n is a 2^n-periodic point of g for any $n \in \mathbb{N}$. Because of the relation $g(K_{\lambda_n}) = (\lambda_n, 1]$ we have $\mathrm{O}(\lambda_n, g) \subseteq J_{\lambda_n}$, and therefore λ_n is also periodic with respect to g_{λ_n} having the same periodic orbit as for g. By Šarkovskii's Theorem we have the identity

$$\{2^i \mid i = 0, 1, \dots, n\} = \{k \in \mathbb{N} \mid x \text{ is } k\text{-periodic w.r. to } g_{\lambda_n} \text{ for some } x \in [0,1]\} \tag{14}$$

because otherwise there were an m-periodic orbit M of g_{λ_n} for some $m \in \mathbb{N}$ with $m \prec 2^n$. Since g_{λ_n} has at most one periodic point in $\bar{K}_{\lambda_n}$ (the point λ_n) the inclusion $M \subseteq J_{\lambda_n}$ holds and with $\rho := \max M < \lambda_n$ the map g_ρ and hence also g_{λ_n} has a 2^n-periodic orbit in $[0, \rho] \cap J_\rho$. This contradicts the minimality of λ_n.

The sequence $(\lambda_n)_{n \in \mathbb{N}}$ is strongly increasing because otherwise there would exist numbers $n, m \in \mathbb{N}$, $m > n$ with $\lambda_m \le \lambda_n$ such that the map g_{λ_m} has a 2^n-periodic orbit in $[0, \lambda_m)$ and this would again contradict the minimality of λ_n. On the other hand, the sequence $(\lambda_n)_{n \in \mathbb{N}}$ is bounded above by 1 and therefore it has a limit

$$\lambda^* := \lim_{n \to \infty} \lambda_n$$

which is smaller then $\frac{6}{7}$ since the map $g_{\frac{6}{7}}$ has periodic points of any period $n \in \mathbb{N}$ (by Šarkovskii's Theorem). In addition, λ^* is greater than $\frac{4}{5}$, since $\lambda_2 = \frac{4}{5}$. Indeed, in [14, Remark 4] it has been mentioned that $\lambda^* = 0.8249080\dots$

We now determine for each member of the family of truncated tent maps which kind of chaos prevails.

For $0 \le \lambda < \lambda^*$ the map g_λ is not L/Y-chaotic:

For any $\lambda \in [0, \lambda^*)$ there exists an $n \in \mathbb{N}$ with $\lambda < \lambda_n$. Therefore, any periodic orbit of g_λ in J_λ is also a periodic orbit of g_{λ_n}. On the other hand, the map g_{λ_n} has finitely many periodic points because of (13) and (14), and therefore also g_λ has only finitely many periodic points, since at most one periodic orbit of g_λ has nonempty intersection with K_λ. So g_λ is not L/Y-chaotic by Lemma 5.2.

The map g_{λ^*} is L/Y-chaotic but not B/C-chaotic:

In [4, VI Example 29] it has been shown that not all points in $[0,1]$ are approximately periodic with respect to g_{λ^*}, and therefore g_{λ^*} is L/Y-chaotic by Proposition 3.4. On the other hand, assuming to the contrary that g_{λ^*} is B/C-chaotic, by Proposition 3.3 there exists an odd number $q > 1$ such that g_{λ^*} has a $q\,2^k$-periodic orbit P for some $k \ge 0$. In case $p := \max P < \lambda^*$ there is an $n \in \mathbb{N}$ with $\lambda_n > p$ such that P is a periodic orbit of g_{λ_n}. This contradicts (14). If, on the other hand, $p = \lambda^*$, by Šarkovskii's Theorem the map g_{λ^*} has a $(q+2)\,2^k$-periodic orbit Q. Because of $\max Q < \lambda^*$ this again leads to a contradiction.

For $\lambda^* < \lambda \le 1$ the map g_λ is B/C-chaotic:

The original tent map g $(= g_1)$ is D-chaotic on $[0,1]$ (see e.g. [7, III Example 9]) and therefore g has a periodic point $\rho \in [\lambda^*, \lambda]$.

We first prove now that the map g_ρ is B/C-chaotic. To this end we notice that ρ is a periodic point of g_ρ. This is due to the fact that either $\mathrm{O}(\rho, g) \cap K_\rho = \emptyset$, and thus $\mathrm{O}(\rho, g_\rho) = \mathrm{O}(\rho, g)$, or $g^j(\rho) \in K_\rho$ for some minimal $j \in \mathbb{N}$, and therefore $g_\rho^{j+1}(\rho) = \rho$. The case where ρ is m-periodic with respect to g_ρ for some $m \in \mathbb{N}$, $m \notin \{2^n \mid n \in \mathbb{N}_0\}$ is easily settled because in this case g_ρ is B/C-chaotic by Proposition 3.3. So from now on we may assume that ρ is a 2^n-periodic point of g_ρ for some $n \in \mathbb{N}_0$. We now define the intervals

$$K := [g_\rho^{2^n}(\lambda_{n+1}), \lambda_{n+1}] \quad \text{and} \quad J := [\lambda_{n+1}, \rho]$$

and prove the existence of an $N \in \mathbb{N}$ with the property

$$K \cup J \subseteq g_\rho^N(K) \cap g_\rho^N(J) \tag{15}$$

which in turn implies, by Proposition 3.3, that g_ρ is B/C-chaotic. To this end we first notice that the intervals K and J are well defined since the relations $\rho > \lambda^*$ and $\mathrm{O}(\lambda_{n+1}, g_\rho) = \mathrm{O}(\lambda_{n+1}, g)$ are valid. Since ρ is 2^n-periodic we get

$$K \cup J \subseteq g_\rho^{2^n}(J). \tag{16}$$

In order to prove that for some $r \in \mathbb{N}$ we have

$$K \cup J \subseteq g_\rho^r(K) \tag{17}$$

we state the following property which is valid for all $\tau \in [0,1]$ and $j \in \mathbb{N}$:

$$\begin{gathered} |g_\tau^j(x) - g_\tau^j(y)| = 2^j|x-y| \quad \text{for all} \quad x,\, y \in [0,1] \quad \text{such that} \\ g_\tau^i(x),\ g_\tau^i(y) \geq 1 - \tfrac{\tau}{2} \quad \text{or} \quad g_\tau^i(x),\ g_\tau^i(y) \leq \tfrac{\tau}{2} \quad \text{for} \quad i = 0, 1, \ldots, j. \end{gathered} \tag{18}$$

Using (18) and the fact that both λ_{n+1} and $g_\rho^{2^n}(\lambda_{n+1})$ are 2^{n+1}-periodic with respect to g_ρ, we see that there is some $j \in \{0, 1, \ldots, 2^n - 1\}$ with the following property:

$$\begin{aligned} & g_\rho^j(\lambda_{n+1}) < \tfrac{\rho}{2} && \text{and} \quad g_\rho^{2^n+j}(\lambda_{n+1}) > 1 - \tfrac{\rho}{2} \\ \text{or} \quad & g_\rho^j(\lambda_{n+1}) > 1 - \tfrac{\rho}{2} && \text{and} \quad g_\rho^{2^n+j}(\lambda_{n+1}) < \tfrac{\rho}{2}. \end{aligned}$$

So we get $\rho \in g_\rho^{j+1}(K)$ and therefore $[\lambda_{n+1}, \rho] \subseteq g_\rho^{j+1}(K)$. Thus with $r := 2^n + j + 1$ condition (17) is satisfied. Using (16), (17) and the definition $N := r + 2^n$ the claim (15) follows and g_ρ^N is turbulent. By Proposition 3.3 then g_ρ is B/C-chaotic.

Finally, since g_ρ is now known to be B/C-chaotic, there is a p-periodic orbit P of g_ρ in J_ρ for some $p \in \mathbb{N}$, $p \notin \{2^n \mid n \in \mathbb{N}_0\}$ (compare with the above proof that g_{λ^*} is not B/C-chaotic). Because of $\lambda > \rho$ the orbit P is also p-periodic with respect to g_λ and therefore g_λ is B/C-chaotic by Proposition 3.3.

Combining the previous considerations with the results of Section 4 we get the following summary for the family of truncated tent maps:

- For each $\lambda \in [0, \lambda^*)$ the map g_λ is not chaotic in any of the three senses considerd in this paper.
- The particular map g_{λ^*} is chaotic in the sense of Li & Yorke but neither in the sense of Block & Coppel nor Devaney.
- For each $\lambda \in (\lambda^*, 1]$ the map g_λ is chaotic in any of the three senses considered in this paper.

We conclude this section with a few additional remarks on the family of truncated tent maps.

Remarks 5.1

(1) The map g_{λ^*} first appeared in [14, Remark 4] as an example of a map of *typ* 2^∞ (i.e. the set of periods of its periodic points is $\{2^i \mid i \in \mathbb{N}_0\}$) having a scrambled set. Block and Coppel [4, VI Example 29] have proved that g_{λ^*} is L/Y-chaotic by showing that $\frac{\lambda^*}{2},\, 1-\frac{\lambda^*}{2} \in \overline{\mathrm{P}(g_{\lambda^*})}$, but $\bigl[\frac{\lambda^*}{2},\, 1-\frac{\lambda^*}{2}\bigr] \cap \mathrm{P}(g_{\lambda^*}) = \emptyset$. To this end additional results [4, VI Lemma 17 and Theorem 24] have been used.

(2) The fact that g_λ is B/C-chaotic for all $\lambda \in (\lambda^*, 1]$ can also be proved by using *kneading theory* for *unimodal* maps (see [6]). In this context one considers the restriction $g_\lambda|_{[0,\lambda]}$ of g_λ on $[0,\lambda]$ and defines the intervals $L := \bigl[0, \frac{\lambda}{2}\bigr)$, $C := \bigl[\frac{\lambda}{2},\, 1-\frac{\lambda}{2}\bigr]$ and $R := \bigl(1-\frac{\lambda}{2}, \lambda\bigr]$, where $\{L, C, R\}^{\mathbb{N}}$ is the symbol space of the *itineraries* of g_λ. Then the corresponding results from [6] for unimodal maps are also valid for $g_\lambda|_{[0,\lambda]}$.

Using *kneading theory* one can also prove that g_{λ^*} is L/Y-chaotic. To this end one shows that λ^* is not finally periodic with respect to g_{λ^*} and hence not approximately periodic (see (18)). In addition one can see that the set of periodic points of g_{λ^*} is $\bigcup\limits_{n\in\mathbb{N}_0} \mathrm{O}(\lambda_n, g) \cup \{\frac{2}{3}\}$.

(3) Example 5.1 suggests that the set of B/C-chaotic maps on I is open in the set $\mathrm{C}_0(I,I)$ of continuous self maps of I (in the topology defined by the supremum norm). That this is indeed true can be seen in [4, Corrolary 20].

Example 5.1 might also suggest that the set of L/Y-chaotic maps on I is closed in $\mathrm{C}_0(I,I)$. This, however, is not true. In [5] the family $T_\lambda\colon [0,1] \to [0,1]$, $0 \le \lambda \le 2$, of tent maps is defined by $T_\lambda(x) := \lambda x$ for $x \in \bigl[0, \frac{1}{2}\bigr]$ and $T_\lambda(x) := \lambda(1-x)$ for $x \in \bigl[\frac{1}{2}, 1\bigr]$. It is proved then that for any $\lambda > 1$ the map T_λ is B/C- and therefore L/Y-chaotic. On the other hand, for $\lambda = 1$ all points of $[0,1]$ are obviously mapped on fixpoints in $\bigl[0, \frac{1}{2}\bigr]$ and this means that T_1 is not L/Y-chaotic.

6 An Example in a General Compact Metric Space

We finally show by means of an example that in the context of general compact metric spaces we canot expect the close relations between the three definitions of chaos as they appear in the particular case of interval maps. In fact, our example shows that even L/Y-chaos together with B/C-chaos does not imply D-chaos.

Example 6.1 The *adding machine* $\tau\colon \Sigma \to \Sigma$ is defined as follows: To any point $\alpha = (a_0, a_1, a_2, \dots) \in \Sigma$ it "adds" the particular point $(1,0,0,0,\dots)$ according to the following rule: If $\alpha = (1,1,1,\dots)$ define $\tau(\alpha) := (0,0,0,\dots)$, otherwise let all entries of α unchanged except the first a_n which vanishes; change this a_n to 1. It is well known that τ is a homeomorphism without periodic points (see e.g. [4, p.133/134]). So if we define the continuous map

$$f\colon \Sigma\times\Sigma \to \Sigma\times\Sigma, \quad (\alpha,\, \beta) \mapsto (\sigma(\alpha), \tau(\beta))$$

we see that f is semi-conjugate to σ via the projection $h\colon \Sigma\times\Sigma \to \Sigma$ from $\Sigma\times\Sigma$ to its first component. Therefore f is B/C-chaotic. Furthermore, if $S \subset \Sigma$ is a scrambled set for σ then for any $\alpha \in \Sigma$ the set $S_\alpha := \{(s, \alpha) \mid s \in S\}$ is obviously a scrambled set for f. Hence f is L/Y-chaotic. On the other hand, since τ has no periodic points f has none either and therefore f is not D-chaotic.

Remarks 6.1

(1) The previous example shows that condition (ii) of the Definition 3.3 of D-chaos (or more general, the existence of periodic points) may by too restrictive. Indeed, for a B/C-chaotic map $f \colon X \to X$ (X any compact metric space) there exists a compact invariant set $Y \subseteq X$ such that $f|_Y$ is B/C-chaotic, transitive and has sensitive dependence on initial conditions (use [1, Theorem. 3]).

(2) For maps defined on non-compact metric spaces in [10] we have given examples of maps which are D- but neither L/Y- nor B/C-chaotic (see [10, Theorem 3.3.3]) or D- and L/Y-chaotic but not B/C-chaotic (see [10, Theorem 3.3.5]).

We want to conclude this paper with raising the question about the relations between the three definitions of chaos considered above in the general case of mappings on arbitrary compact metric spaces. It is widely believed that B/C-chaos implies L/Y-chaos but we do not know if this is really true. And what is known about the other relations?

Acknowledgement

We would like to thank Louis Block for pointing out that the truncated tent map g_λ should not be chaotic for $\lambda < \lambda^*$ but chaotic in every sense for $\lambda > \lambda^*$. The idea of using kneading theory for the proof of L/Y-chaos for g_{λ^*} and B/C-chaos for g_λ, $\lambda > \lambda^*$ is due to Michał Misiurewicz whome we also thank.

References

[1] Auslander, J. and Yorke, J.A. Interval maps, factors of maps, and chaos. *Tohoku Math. J.* II. Ser. 32 (1980) 177–188.

[2] Banks, J., Brooks, J., Cairns, G., Davis, G. and Stacey, P. On Devaney's definition of chaos. *Amer. Math. Monthly* **99** (1992) 332–334.

[3] Berglund, R. and Vellekop, M. On intervalls transitivity chaos. *Amer. Math. Monthly* **101** (1994) 353–355.

[4] Block, L.S. and Coppel, W.A., *Dynamics in One Dimension.* Springer Lecture Notes, **1513**, Springer Verlag, Berlin, 1992.

[5] Butler, G.J. and Piagiani, G. Periodic points and chaotic functions in the unit interval. *Bull. Austral. Math. Soc.* **18** (1978) 255–265.

[6] Collet, P. and Eckmann, J.-P. *Iterated Maps on the Interval as Dynamical Systems.* Progress in Physics, **1**, Birkhäuser, Basel, 1980.

[7] Devaney, R.L. *An Introduction to Chaotic Dynamical Systems.* Benjamin/ Cummings, Menlo Park CA, 1986.

[8] Elaydi, S. *Discrete Chaos.* CRC Press, Boca Raton, 1999.

[9] Holmgren, R.A. *A First Course in Discrete Dynamical Systems.* Springer, New York, 1994.

[10] Kieninger, B. *Analysis of three definitions of chaos for continuous maps on metric spaces.* Diploma Thesis, University of Augsburg, 1998. (German).

[11] Kirchgraber, U. and Stoffer, D. On the definition of chaos. *Z. angew. Math. Mech.* **69** (1989) 175–185.

[12] Li, S. ω-chaos and topological entropy. *Trans. Amer. Math. Soc.* **339** (1993) 243–249.

[13] Li, T.J. and Yorke, J.A. Period three implies chaos. *Amer. Math. Monthly* **82** (1975) 985–992.

[14] Misiurewicz, M. and Smital, J. Smooth chaotic maps with zero topological entropy. *Ergodic Theory Dyn. Syst.* **8** (1988) 421–424.

[15] Šarkovskii, A.N. Coexistence of cycles of a continuous mapping of the line into itself. *Ukrain. Mat. Ž.* **16** (1964) 61–71. (Russian).

[16] Šarkovskii, A.N., Kolyada, S.F., Sivak, A.G. and Fedorenko, V.V. *Dynamics of One-dimensional Mappings.* Naukova Dumka, Kiev, 1989. (Russian).

Nonlinear Dynamics and Systems Theory, **1**(1) (2001) 39–58

Input-Output Decoupling with Stability for Bond Graph Models

J.M. Bertrand, C. Sueur and G. Dauphin-Tanguy

L.A.I.L., U.P.R.E.S.A. C.N.R.S. 8021, Ecole Centrale de Lille,
B.P. 48, 59651 Villeneuve d'Ascq cedex, France

Received: February 17, 2000;   Revised: July 6, 2000

Abstract: In this paper, the geometric approach and the bond-graph methodology are combined to characterize the structure of square linear systems modeled by bond-graph. A new concept is defined to emphasize the symbolic expressions of the fixed modes of the decoupled model and to design decoupling state feedback laws.

Keywords: *Bond graphs; linear systems; non-interacting control; stability properties.*

Mathematics Subject Classification (2000): 34D20, 93C15, 93D25.

1 Introduction

The bond graph is an appreciated tool for physical systems modelling. Based on power flows representation, it enables the description of the system through energy storage and dissipative elements [10, 16]. In a control objective, the structure of the chosen model is also of greatest importance: closed loop requirements may depend on groups of elements of the open loop model. Refining these parts of the model would enable to meet the control goals more efficiently, provided that these refinements also improve the model accuracy. In an input-output decoupling objective, the aim of this work is to identify, on the bond graph model describing the system, the elements involved in major properties of the control solution.

Suitable tools for both structural analysis and synthesis of input-output decoupling control laws are defined by the geometric approach [1, 22]. In particular, many contributions have been brought about input-output decoupling by regular static state feedback, in which the structure of the open loop model is of greatest interest. This structure specially enables to know whether the model is decouplable [5 − 8, 11, 13]. If so, some poles of the decoupled model are also shown to be independent of the control law, so-called fixed modes [9, 12]. Suitable tools for the structural synthesis of such input-output decoupling

© 2001 Informath Publishing Group. All rights reserved.

control laws are defined by the geometric approach. Using particular state space subspaces [4], the designer may choose the number of degrees of freedom introduced by the control law. An unstable unassigned mode would lead to an unstable decoupled model, making this control strategy unrealistic.

In this paper, thanks to geometric concepts, structural analysis methods are emphasized for the input-output decoupling of linear square bond graph models by regular static state feedbacks. Graphical methods are first developed to determine, in terms of fixed modes, if a stable solution exists for the regular input-output decoupling problem. If so, the bond graph methodology is then used to compute state feedbacks insuring stability of the decoupled model.

2 Basic Concepts for Model Analysis

In this part, the basic concepts for model analysis are recalled with different approaches. These concepts are used in the main part of this paper for the characterization of feedback laws.

Consider square linear time-invariant systems $(\sum) = (A, B, C)$ described by equation

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t), \end{cases} \tag{1}$$

where $x(t) \in \mathcal{X} \approx \mathbb{R}^n$ is the state, $u(t) \in \mathcal{U} \approx \mathbb{R}^m$ is the control input, $y(t) \in \mathcal{Y} \approx \mathbb{R}^m$ is the output to be controlled. The same notation is used for maps and their matrix representations in particular bases $A\colon \mathcal{X} \to \mathcal{X}$, $B\colon \mathcal{U} \to \mathcal{X}$, $C\colon \mathcal{X} \to \mathcal{Y}$. $\mathcal{B}$ is the image of B and $\mathcal{K}$ the kernel of C. System (1) is supposed to be invertible.

2.1 Algebraic approach

The infinite structure allows us to express whether a model is decouplable by a regular static state feedback. The stability property of the decoupled model is deduced from the finite structure. It means that the controlled model can be made stable if the fixed modes are stable. The algebraic way for the study of these two structures is now recalled.

2.1.1 Infinite structure. The infinite structure is characterized by the row and global infinite zero orders.

Definition 2.1 Let n_i be the smallest integer verifying $c_i A^{k-1} B = 0$, $k < n_i$ and $c_i A^{n_i - 1} B \neq 0$, with c_i the i-th row of matrix C.

Definition 2.2 [5] n_i is called the i^{th} row infinite zero order, associated with the i^{th} output variable.

Property 2.1 *n_i is the number of derivations of the i^{th} output variable necessary to make appear explicitly at least one of the control input variables.*

Definition 2.3 [20] Let $T(s)$ be the transfer matrix of system (1). The Smith-McMillan form at infinity of $T(s)$ is the unique matrix $\Phi(s)$ defined by equation

$$T(s) = B_1(s).\Phi(s).B_2(s), \tag{2}$$

with

$$\Phi(s) = \operatorname{diag}\left\{s^{-n'_1}, s^{-n'_2}, \dots, s^{-n'_m}\right\},$$

and B_1 and B_2 are two non unique bicausal matrices.

Definition 2.4 [5] The set of non increasing integers $\{n'_1, n'_2, \dots, n'_m\}$ is the set of global infinite zero orders of $(\sum)$. It contains m numbers because the system is invertible.

2.1.2 Finite structure.

Definition 2.5 [17] Let $P(s)$,

$$P(s) = \begin{bmatrix} sI - A & B \\ -C & 0 \end{bmatrix}, \tag{3}$$

be the system matrix of $(\sum)$. The invariant zeros of $(\sum)$ are the zeros of the Smith form of $P(s)$. They also are the roots of $\det P(s)$ because the system is square.

2.2 Geometric approach

Some geometric results are now recalled that define invariant subspaces used in input-output decoupling.

Definition 2.6 [23] A subspace $\mathcal{W}$ is $(A, \mathcal{B})$ invariant subspace if it satisfies the inclusion $A\mathcal{W} \subset \mathcal{W} + \mathcal{B}$.

In terms of state feedback, $\mathcal{W}$ is an $(A, \mathcal{B})$ invariant subspace iff there exists a set $\mathcal{F}(A, B; W)$ of state feedback matrices F such as $(A + BF)\mathcal{W} \subset \mathcal{W}$. Let $\mathcal{L}(A, \mathcal{B}; \Psi)$ be the set of $(A, \mathcal{B})$ invariant subspaces included in the subspace Ψ. This subspace is closed for addition, it thus contains a supremal element.

Property 2.2 *The subspace $\mathcal{L}(A, \mathcal{B}; \Psi)$ contains a unique supremal element denoted as $\mathcal{V}^*(\Psi) = \sup \mathcal{L}(A, \mathcal{B}; \Psi)$.*

The subspace $\mathcal{V}^*(\Psi)$ is the limit of the algorithm (4)

$$\begin{cases} \mathcal{V}^0 = \mathcal{X}, \\ \mathcal{V}^\mu = \Psi \cap A^{-1}(\mathcal{B} + \mathcal{V}^{\mu-1}), \end{cases} \tag{4}$$

called “Controlled Invariant Subspace Algorithm” [23].

For control purposes, a particular set of subspaces is used: $(A, \mathcal{B})$ invariant subspaces included in the kernel of the output matrix, denoted $\mathcal{L}(A, \mathcal{B}; \mathcal{K})$. The supremal element is usually denoted as $\mathcal{V}^* = \sup \mathcal{L}(A, \mathcal{B}; \mathcal{K})$. It can be obtained by using equation (4), with $\Psi = \mathcal{K}$. For control purposes, the orthogonal complement of the subspace $\mathcal{V}^*$ is used in this paper. It is the limit of algorithm (5):

$$\begin{cases} \mathcal{V}^{0\perp} = 0, \\ (\mathcal{V}^{\mu-1})^\perp = \mathcal{K}^\perp + A^t(\mathcal{B}^\perp \cap (\mathcal{V}^{\mu-1})^\perp). \end{cases} \tag{5}$$

There is a fundamental property between $\mathcal{V}^*$ and the observability property of the controlled system. This property can be expressed as:

Property 2.3 *Subspace $\mathcal{V}^*$ is the greatest non observable subspace built with state feedback.*

Stable dynamics are associated with a second set of $(A, \mathcal{B})$ invariant subspaces: stabilizable subspaces.

Definition 2.7 $\mathcal{W}$ is a stabilizable $(A,\mathcal{B})$ invariant subspace iff there exists a set of state feedback matrices $F \in \mathcal{F}(A,B;\mathcal{W})$ verifying equation

$$\sigma(\mathcal{W}|A+BF|\mathcal{W}) \subset C_-. \tag{6}$$

C_- is a set of negative real part eigenvalues. A stabilizable subspace $\mathcal{W}$ is thus an $(A,\mathcal{B})$ invariant subspace with which a state feedback matrix $F \in \mathcal{F}(A,B;\mathcal{W})$ is built such as $(A+BF)$ is stable on $\mathcal{W}$. Suppose $\mathcal{L}^-(A,\mathcal{B};\mathcal{K})$ the set of stabilizable subspaces included in the kernel of the output matrix. This subspace verifies the following property:

Property 2.4 *The set $\mathcal{L}^-(A,\mathcal{B};\mathcal{K})$ contains a unique supremal element denoted as $\mathcal{V}^*_{\text{stab}} = \sup\mathcal{L}^-(A,\mathcal{B},\mathcal{K})$. It satisfies equation*

$$\mathcal{V}^*_{\text{stab}} \subset \mathcal{V}^*. \tag{7}$$

Among all output nulling trajectories, subspace $\mathcal{V}^*_{\text{stab}}$ only characterizes those which are stable. Guarantying in the same time decoupling and stability property of the decoupled system, it will be used for the control law synthesis. Among the set of output nulling trajectories, free dynamics and fixed dynamics are pointed out. They will be characterized for bond graph models.

Other subspaces are briefly used in this paper: $(\mathcal{K},A)$-invariant subspaces.

Definition 2.8 $\mathcal{W}$ is a $(\mathcal{K},A)$-invariant subspace iff it satisfies equation (8).

$$A(\mathcal{W}\cap\mathcal{K}) \subset \mathcal{W}. \tag{8}$$

Let $\mathcal{S}(\mathcal{K},A;\mathcal{B})$ the set of $(\mathcal{K},A)$-invariant subspaces containing subspace $\mathcal{B}$. This set contains a minimal element denoted as $S^* = \inf\mathcal{S}(\mathcal{K},A;\mathcal{B})$. It is the limit of algorithm

$$\begin{cases} S^0 = 0, \\ S^\mu = \mathcal{B} + A(\mathcal{K}\cap S^{\mu-1}), \end{cases} \tag{9}$$

called “Conditional Invariant Subspace Algorithm”.

As described by equations (4) and (9), subspaces $\mathcal{V}^*$ and S^* are obtained with dual algorithms. The following relation can be written:

Property 2.5 $\sup\mathcal{L}(A,\mathcal{B};\mathcal{K}) = (\inf\mathcal{S}(\mathcal{B}^\perp,A^t;\mathcal{K}^\perp))^\perp$.

Property 2.6 *For invertible systems, subspaces $\mathcal{V}^*$ and S^* satisfy equation*

$$\mathcal{V}^* + S^* = \mathcal{X}. \tag{10}$$

Property 2.7 *For invertible systems, equation*

$$\dim\mathcal{V}^* = n - \sum_i n'_i \tag{11}$$

is satisfied.

According to equation (11), if $\mathcal{V}^*_i$ is the supremal subspace of subsystem $(\sum_i) = (A,B,c_i)$ included in $\ker c_i$, a basis for each subspace $\mathcal{V}^{*\perp}_i$ is the limit of algorithm (5) with $\mathcal{K} = \ker c_i$ and is given by equation

$$\mathcal{V}^{*\perp}_i = \text{vect}\left\{c_i^t, ..., (c_iA^{n_i-1})^t\right\}. \tag{12}$$

2.3 Bond graph approach

Let us consider, in the following, bond graph models with complete integral causality assignment. The minimal state vector thus deduced is x whereas the state space equation is described by equation (1). The previous results can be applied on the state space representation. The object of this part is to recall some results about infinite zeros and invariant zeros of such models, directly with a graphical approach. Particularly, the equivalence between null invariant zeros of bond graph models with an integral causality assignment, denoted BGI, and infinite zeros of bond graph models with a derivative causality assignment, denoted BGD, is emphasized.

2.3.1 Infinite structure. Consider bond graph models with an integral causality assignment.

Definition 2.9 The length of a causal path is equal to the number of dynamical elements met when following the path.

Definition 2.10 When they contain at least one dynamical element, two causal paths are said to be different if they do not have any common dynamical element.

Property 2.8 [19] *n_i is equal to the length of the shortest causal path between the i^{th} output detector and all the input sources.*

Property 2.9 [19] *The number of global infinite zeros is equal to the number of different input-output causal paths. Their orders are computed as in equation*

$$\begin{cases} n'_m = L_1, \\ n'_{m-k+1} = L_k - L_{k-1}, \end{cases} \tag{13}$$

where L_k is the sum of the lengths of the k shortest input-output different causal paths.

If there are several choices of m different shortest input-output causal paths, the gains of the shortest different causal paths from at least two output detectors to all the input sources may be proportional. It means that, in this case, the control inputs do not appear independently in the output derivatives. Hence, the integers computed according to equations (13) do not define the global infinite zero orders of the model. For independence between control inputs and output derivatives to be performed, at least one output variable must be derived more times. The order of the i^{th} global infinite zero is thus greater than the length of the shortest causal path from the associated output detector to the input sources. For any invertible bond graph model with m inputs and m outputs, there exists at least one choice of m different input-output causal paths.

2.3.2 Finite structure. Graphical methods allow the characterization of the invariant zeros of $(\sum)$ straight from its bond graph model. Particularly, considering the bond graph model obtained by removing from the initial one each choice of m different input-output causal paths and expressing each characteristic polynomial, one determines the invariant zeros of the global square model [19]. Null invariant zeros can be derived straightforward with a causal approach. A new concept is now defined on the BGD: the i^{th} output infinite zero order n_{id}, associated with the i^{th} output variable. It will be pointed out that n_{id} is equal to the number of null invariant zeros of the i^{th} row subsystem.

Let us assign the derivative causality on the bond graph model of the system. As the derivative causality assignment can be performed, the state matrix A is invertible. A more general approach is proposed in [3]. Hence, the associated mathematical representation is given by equation

$$\left\{ \begin{array}{l} x = A^{-1}\dot{x} - A^{-1}Bu, \\ y = CA^{-1}\dot{x} - CA^{-1}Bu. \end{array} \right. \tag{14}$$

Definition 2.11 Let n_{id} be the smallest integer verifying $c_i A^{-(k+1)}B = 0,\ k < n_{id}$ and $c_i A^{-(n_{id}+1)}B \neq 0$.

n_{id} is thus the number of integrations of the i^{th} output variable necessary to make appear explicitly at least one of the control input variables.

Property 2.10 *n_{id} is equal to the length of the shortest causal path between the i^{th} output detector and all the input sources on the BGD.*

Definition 2.12 n_{id} is called the order of the i^{th} row infinite zero associated with the i^{th} output variable on the BGD.

Extending the previous result to the whole system, let us now define for the BGD the new concept of global infinite zero orders, noted $\{n'_{1d}, ..., n'_{md}\}$.

Definition 2.13 Let $\{n'_{1d}, ..., n'_{md}\}$ the integer set verifying equation

$$\left\{ \begin{array}{l} n'_{md} = L_{1d}, \\ n'_{(m-k+1)d} = L_{kd} - L_{(k-1)d}, \end{array} \right. \tag{15}$$

where L_{kd} is the sum of the lengths of the k shortest different input-output causal paths on the BGD. These integers are called global infinite zero orders of the BGD.

These integers are obtained directly on the BGD with the same approach as the set $\{n'_1, ..., n'_m\}$ on the BGI.

Property 2.11 *n_{id} is equal to the number of null invariant zeros of the i^{th} row subsystem.*

Theorem 2.1 *Let $\{n'_{1d}, ..., n'_{md}\}$ be the set of global infinite zero orders of the BGD. The number of null invariant zeros of the BGI is equal to $\sum_{k=1}^{m} n'_{kd}$.*

The proof of this theorem is proposed in appendix. Note that the BGD has frequently direct input-output causal paths. In that case, several choices of m different shortest input-output causal paths are often found. Then, when computing the integers from the bond graph model with derivative causality assignment, take care of the proportionality between the gains of these causal paths.

3 Regular Static State Feedback Decoupling with Stability

In this part, $(\sum)$ is supposed to be invertible, controllable and observable [18]. A static state feedback control law, described as equation

$$u = Fx + Gv, \tag{16}$$

is applied on equation (1). It is called regular when matrix G is square invertible.

3.1 Algebraic and geometric approaches

Let $\{n_i\}$ be the set of row infinite zero orders and $\{n'_i\}$ the set of global infinite zero orders. If $(\sum)$ is decouplable by a regular static state feedback, this control strategy is called rssf in the next [9].

3.1.1 Structural condition for decoupling with stability. Let Ω be the decoupling matrix defined as in equation

$$\Omega = \begin{bmatrix} c_1 A^{n_1-1} B \\ \vdots \\ c_m A^{n_m-1} B \end{bmatrix}. \tag{17}$$

Property 3.1 [5, 15] *n_i is invariant under rssf.*

Property 3.2 *$(\sum)$ is decouplable by rssf iff Ω is invertible.*

Theorem 3.1 [5] *$(\sum)$ is decouplable by rssf iff equivalent equations*

$$\{n_i\} = \{n'_i\} \tag{18}$$

and

$$\mathcal{V}^* = \bigcap_{i=1}^{m} \mathcal{V}_i^* \tag{19}$$

are satisfied.

When decoupling $(\sum)$ by state feedback, some poles of the decoupled model are unobservable and independent of the control law. They are called fixed modes. These fixed modes are defined straight from the open-loop model [9]: they are all or only some of the invariant zeros of the open-loop model. In order to achieve decoupling with stability, a second set of conditions must be satisfied. Let us denote $Z^+(c_i, A, B)$ the set of unstable invariant zeros of system (c_i, A, B).

Theorem 3.2 [12] *$(\sum)$ is decouplable with stability by rssf iff equations*

$$\begin{cases} \{n_i\} = \{n'_i\}\,, \\ Z^+(C, A, B) = \sum\limits_{i=1}^{m} Z^+(c_i, A, B) \end{cases} \tag{20}$$

are satisfied.

Theorem 3.3 [12] *$(\sum)$ is decouplable with stability by rssf iff equations*

$$\begin{cases} \mathcal{V}^* = \bigcap\limits_{i=1}^{m} \mathcal{V}_i^*, \\ \mathcal{V}^*_{\text{stab}} = \bigcap\limits_{i=1}^{m} \mathcal{V}^*_{i\,\text{stab}} \end{cases} \tag{21}$$

are satisfied.

3.1.2 Decoupling and disturbance rejection. In this section, symbolic expressions of regular static state feedback control laws $u(t) = Fx(t) + Gv(t)$ insuring input output decoupling are recalled. The geometric approach consists on identifying state subspaces with adequate properties for the control law goal [21].

These methods use geometric supports derived on the control law synthesis for disturbance rejection, and particularly on the concept of decoupling subspace [4]. In a first step, the methodology for disturbance rejection is recalled. Then, this concept is used in order to achieve input-output decoupling, by considering all the control inputs, except one, as a disturbance input for each output variable. Two sets of decoupling subspaces are used in this paper: $\mathcal{V}_i^* = \sup \mathcal{L}\,(A, B; \ker c_i)$ and $\mathcal{V}_{i\,\text{stab}}^* = \sup \mathcal{L}^-\,(A, B; \ker c_i)$. The properties of the associated decoupled system are recalled.

Consider the SISO system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Ed(t), \\ y(t) = cx(t), \end{cases} \tag{22}$$

where $d(.) \in \mathcal{D} \approx \mathcal{R}^q$ is the disturbance.

The goal is to find a control law such that the transfer function matrix from $d(s)$ to $y(s)$ be zero. This goal is achievable if the following theorem is satisfied.

Theorem 3.4 [22] *The output variable $y(t)$ of system (22) can be decoupled from the disturbance vector $d(t)$ iff there exists a $(A, \mathcal{B})$ invariant subspace D satisfying equation*

$$\operatorname{Im} E \subset D \subset \mathcal{V}^* \subset \ker c. \tag{23}$$

D is called decoupling subspace of the disturbance $d(t)$.

The supremal decoupling subspace is $\mathcal{V}^* = \sup \mathcal{L}\,(A, \mathcal{B}; \ker c)$. The controlled system is described by equation

$$\begin{cases} \dot{x}(t) = (A + BF)x(t) + Bv(t) + Ed(t), \\ y(t) = cx(t), \end{cases} \tag{24}$$

where $v(t)$ is the new control input variable.

The state feedback matrix is calculated by considering the following property.

Proposition 3.1 [4] *Consider the SISO system (22) satisfying Theorem 3.4. Suppose that its infinite zero order is such that $n_0 \geq 1$. The feedback matrices F which render $(A + BF)$ invariant each decoupling subspace D are calculated following the equations*

$$\begin{cases} cA^{(n_0-1)}(A + BF) = h, \\ h.D = 0. \end{cases} \tag{25}$$

The column matrix h^t is a linear combination of subspace $D^{\perp}$ basis vectors. Parameters defining this linear combination are the degrees of freedom in the control law. The number of degrees of freedom is thus equal to $\dim D^{\perp}$.

According to Theorem 3.4, the supremal decoupling subspace is $\mathcal{V}^* = \sup \mathcal{L}(A, \mathcal{B}; \ker c)$. This solution minimizes the number of degrees of freedom for the control law. A better solution is given by subspace $\mathcal{V}_{\text{stab}}^* \subset \mathcal{V}^*$ – Property 2.4.

Theorem 3.5 [22] *Suppose that system (22) is stabilizable. The output variable $y(t)$ can be decoupled from the disturbance $d(t)$ while guarantying stability iff equation*

$$\operatorname{Im} E \subset \mathcal{V}_{\text{stab}}^* \subset \mathcal{V}^* \tag{26}$$

is satisfied.

In that case, subspace $\mathcal{V}_{\text{stab}}^*$ is used for the calculus of matrix F. Matrix F satisfies the following property.

Property 3.3 [4] *Each state feedback matrix F such $(A+BF)\mathcal{V}^*_{\rm stab} \subset \mathcal{V}^*_{\rm stab}$ satisfies equation*

$$\sigma(\mathcal{V}^*_{\rm stab}|A+BF|\mathcal{V}^*_{\rm stab}) \subset C_-. \tag{27}$$

These results are right for multivariable systems. Consider now system $(\sum)$ described by equation (1). The decoupled system using a rssf $u(t) = Fx(t) + Gv(t)$ is described by equation

$$\begin{cases} \dot{x}(t) = (A+BF)x(t) + BGv(t), \\ y(t) = Cx(t), \end{cases} \tag{28}$$

with $v(t)$ the new input control vector.

Denote $\overline{v}_i(t)$ the vector $v(t)$ without its ith variable. Given that system (1) is decoupled, each output variable $y_i(t)$ is decoupled from the disturbance vector $\overline{v}_i(t)$, for $i = 1,\dots,m$ at the same time. The following property can than be written.

Property 3.4 [4] *Each $(A+BF)$ invariant subspace D_i satisfying equation*

$$\operatorname{Im} E_i \subset D_i \subset \mathcal{V}^*_i \subset \ker c_i, \qquad i = 1,\dots,m \tag{29}$$

is associated with each output variable $y_i(t)$ of the decoupled system (28).

E_i is the ith column of matrix E. The supremal decoupling subspace is $\mathcal{V}^*_i = \sup \mathcal{L}\,(A,\mathcal{B};\ker c_i)$. For each of the m SISO subsystems, the control law $u(t) = Fx(t) + Gv(t)$ is such that the disturbance $\overline{v}_i(t)$ is included in a decoupling subspace D_i – equation (29). This subspace is $(A+BF)$-invariant. Properties 2.1 and 3.4 allow the definition of the decoupling control law $u(t) = Fx(t) + Gv(t)$.

Property 3.5 [4] *Consider a system which can be decoupled by a rssf. Let Ω be the decoupling matrix, $\{n_i\}$ its row infinite structure and $\{D_i\}$ a set of subspaces solution for the decoupling problem. A decoupled system is obtained with matrices F and G and the control law $u(t) = Fx(t) + Gv(t)$ following equations*

$$\begin{cases} h_i.D_i = 0, \qquad i = 1,\dots,m, \\ \Omega F = [h_i - c_i A^{n_i}]_{i=1,\dots,m}\,, \\ \Omega G = \operatorname{diag}\,[g_i]_{i=1,\dots,m}. \end{cases} \tag{30}$$

A formal expression of matrices F and G using Maple is derived from the set of decoupling subspaces. g_i, $i = 1,\dots,m$, are freely assignable parameters to choose static gains of the closed loop system. Each row matrix h_i is a linear combination of subspace $D_i^\perp$ basis vector. The number of degrees of freedom is thus function of the choice of the decoupling subspace. Two sets of decoupling subspaces are used in this paper: $\{\mathcal{V}^*_1,\dots,\mathcal{V}^*_m\}$ and $\{\mathcal{V}^*_{1\,\rm stab},\dots,\mathcal{V}^*_{m\,\rm stab}\}$. These subspaces characterize the properties of the decoupled system.

Property 3.6 [22] *Consider a square controllable system decoupled by a rssf. Choosing $\{\mathcal{V}^*_1,\dots,\mathcal{V}^*_m\}$ as the set of decoupling subspaces, the unassignable modes of the decoupled systems are all the invariant zeros of the open loop system. If decoupling with stability is possible, choosing $\{\mathcal{V}^*_{1\,\rm stab},\dots,\mathcal{V}^*_{m\,\rm stab}\}$ as the set of decoupling subspaces, unassignable modes of the decoupled system are strictly stable invariant zeros of the open loop system.*

A given rssf is thus associated with a given set of decoupling subspaces that introduces degrees of freedom used to assign some closed loop modes. The decoupling subspaces enable the choice of the number of the decoupled model unassignable modes.

The control law giving the maximum number of unassignable modes is obtained when taking as decoupling subspaces the greatest ones. The associated unassignable modes are all the invariant zeros of $(\sum)$: one unstable invariant zero makes unstable the decoupled model. However, if decoupling with stability is possible, a stable decoupled model may be designed. In this case, for bond graph models, the set of fixed modes is only composed of all the strictly stable invariant zeros [12].

A graphical necessary and sufficient condition is derived in the next section for the existence of at least a control law insuring stability of the decoupled model. This control law is associated with decoupling subspaces $\{\mathcal{V}^*_{1\,\mathrm{stab}}, \dots, \mathcal{V}^*_{m\,\mathrm{stab}}\}$. The formal expressions of these decoupling subspaces are then expressed from the bond graph model of $(\sum)$.

Remark 3.1 The state feedback control law creates an unobservable subspace contained in $\mathcal{V}^*$. Given that the system is square, the controllable part in $\mathcal{V}^*$ is empty. All modes which are unobservable are also non controllable modes for the control law. They are non assignable modes in Property 3.6.

3.2 Bond graph approach

If the bond graph model has no invariant zero and if it is rssf, then it is rssf with stability because there are no fixed mode in that case. If some of the invariant zeros are strictly unstable, the problem does not have any solution for bond graph models because these invariant zeros are fixed modes. If none of these invariant zeros are strictly unstable, the only unstable invariant zeros are the null ones: the following study is dealing with this case.

A method allows us to determine if there exists a stable solution for the input-output decoupling problem. This method is based on the study of infinite zeros structures, whose main concepts are now recalled [2].

Then, it is shown how the bond graph formalism allows us to determine the invariant subspaces symbolic expression and then the control law symbolic expression, directly with a graphical approach.

An example is then proposed.

3.2.1 Structural approach analysis. Combining the previous results, Theorem 2.1 and Theorem 3.2 enable to derive a graphical necessary and sufficient condition for $(\sum)$ to be decouplable by rssf with stability.

Theorem 3.6 *Assume that $(\sum)$ does not have any strictly unstable invariant zero. A stable solution for the input-output decoupling of $(\sum)$ thus exists iff the infinite zeros structures of BGI and BGD verify equations*

$$\begin{cases} \{n_i\} = \{n'_i\}\,, \\ \{n_{id}\} = \{n'_{id}\}\,. \end{cases} \tag{31}$$

According to Property 2.11 and Theorem 3.2, the proof is immediate.

Hence, the bond graph model of $(\sum)$ enables to know graphically if a decoupling rssf exists insuring closed loop stability. In the next part, regular decoupling with stability is supposed to be possible.

3.2.2 Control law. As expressed by Property 3.5, the decoupling control law associated with a set of decoupling subspaces is computed thanks to the symbolic expressions of their orthogonal complements. No simple algorithm allows the calculation of these subspaces with a symbolic expression. The bond graph methodology gives a different way. Causal path length concepts on the bond graph models with integral and derivative causality assignment are now used to determine the expressions of the two sets of useful subspaces $\{(\mathcal{V}_i^*)^\perp, \ldots, (\mathcal{V}_m^*)^\perp\}$ and $\{(\mathcal{V}_{1\text{stab}}^*)^\perp, \ldots, (\mathcal{V}_{m\text{stab}}^*)^\perp\}$.

Consider the bond graph model with integral causality assignment. Let DE_i- resp. DE_{id}- be the i^{th} dynamical element in integral – resp. derivative – causality, associated with the i^{th} state vector component $x_i(t)$ on the bond graph model with integral – resp. derivative – causality assignment. Let be $G_k(DE_i, D_j)$ the constant term, without Laplace operator s, of the gain of a causal path of length k between the i^{th} dynamical element in integral causality DE_i and the j^{th} output detector D_j. Let $g(DE_i)$ be equal to $1/I$ for an I-element and $1/C$ for a C-element.

Property 3.7 $c_j A^k I^i = \sum G_k(DE_i, D_j).g(DE_i)$.

I^i is the identity matrix i^{th} column. From Property 3.7, the formal expressions of the subspaces $\{(\mathcal{V}_i^*)^\perp, \ldots, (\mathcal{V}_m^*)^\perp\}$ can be obtained with a graphical manner. Consider now the bond graph model with derivative causality assignment. For $n_{id} \geq 1$, let $\mathcal{V}_{id}^*$ be such that:

$$(\mathcal{V}_{id}^*)^\perp = \text{span}\,\{(c_i A^{-1})^t, ..., (c_i A^{-n_{id}})^t\}. \tag{32}$$

Let $G_{kd}(DE_{id}, D_j)$ be the constant term of the gain of a causal path of length k between the ith dynamical element in derivative causality DE_{id} and the jth output detector D_j.

Property 3.8 $c_j A^{-k} I^i = \sum G_{k-1}(DE_{id}, D_j)$.

From the same graphical way as in Property 3.7, the formal expressions of the subspaces $\{(\mathcal{V}_{1d}^*)^\perp, \ldots, (\mathcal{V}_{md}^*)^\perp\}$ can be obtained. The Property 3.9 is deduced from equation (12) and (32).

Property 3.9 $\dim(\mathcal{V}_i^*)^\perp = n_i$ *and* $\dim(\mathcal{V}_{id}^*)^\perp = n_{id}$.

Finally, the symbolic expression of the subspaces $(\mathcal{V}_{i\,\text{stab}}^*)^\perp$ can be derived.

Property 3.10 $(\mathcal{V}_{i\,\text{stab}}^*)^\perp = (\mathcal{V}_i^*)^\perp \oplus (\mathcal{V}_{id}^*)^\perp,\ i = 1, ..., m$.

$\{(\mathcal{V}_{i\,\text{stab}}^*), \ldots, (\mathcal{V}_{m\,\text{stab}}^*)\}$ is the set of greatest decoupling subspaces insuring closed loop stability if stable decoupling is possible. The proof of Property 3.10, rather technical, is detailed in the appendix. The bond graph model of $(\sum)$ thus allows us to derive graphically the symbolic expressions of the subspaces needed for the synthesis of input-output (stable) decoupling rssf. An example using the previous analysis and computation methods is now presented.

4 Example

Let us define, Figure 4.1, the bond graph model $BG1$ containing two input sources $\{E_1, E_2\}$, two output detectors $\{D_1, D_2\}$ associated with outputs variable to be controlled

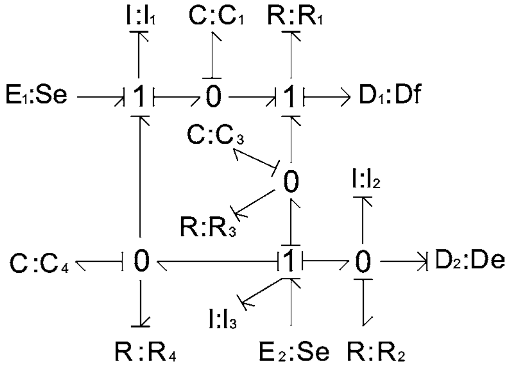

Figure 4.1. Bond-graph model $BG1$.

and six dynamical elements, each with integral causality assignment. This model is invertible, controllable, observable and decouplable by rssf [2].

Assigning a derivative causality on the whole set of dynamical elements leads to the bond graph model $BG2$ described Figure 4.2. The derivative causality can be assigned to each dynamical element. It means that the state matrix is invertible.

Remove from $BG1$ the two shortest different input-output causal paths $D_1 \to R_1 \to C_1 \to I_1 \to E_1$ and $D_2 \to R_2 \to I_3 \to E_2$. The remaining bond graph model contains three dynamical elements: the global model has thus three invariant zeros [19]. Some of these invariant zeros may be null. Studying the global infinite zero structure of $BG2$ allows us to determine graphically their number. The shortest causal path from the output detector D_1 to the input sources does not meet any dynamical element. Thus $n'_{2d} = 0$. Furthermore, there are causal paths of length 1 from the output detector D_2 to the input sources. Due to the R-element R_4, these causal paths are independent of those of length 0 defining n'_{2d}. Hence $n'_{1d} = 1$. Theorem 2.1 so allows to state that the global model has one null invariant zero.

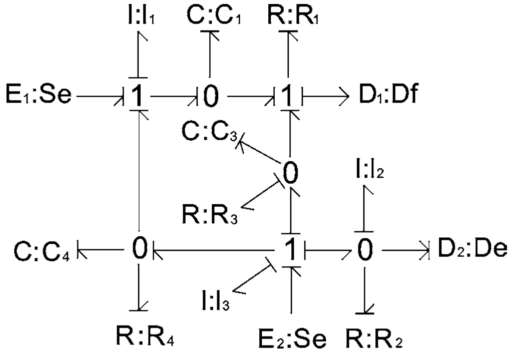

Figure 4.2. Bond-graph model $BG2$.

On $BG1$, finally removing each couple of different input-output causal paths, one computes the two remaining invariant zeros: $s = -1/R_4C_4$ and $s = -1/R_4C_4$ [19]. The three invariant zeros of $BG1$ define the unassignable modes when $(\mathcal{V}_1^*)^\perp$ and $(\mathcal{V}_2^*)^\perp$ are chosen as decoupling subspaces. None of the invariant zeros are strictly unstable.

Computing the row infinite zero orders of $BG2$ thus allows us to know if a stable solution exists for the input-output decoupling of $BG1$. The shortest causal paths from each output detector to the input sources are $D_1 \to R_1 \to E_1$ and $D_2 \to I_2 \to R_4 \to E_2$. Thus $n_{1d} = 0$ and $n_{2d} = 1$. Hence, according to Theorem 3.6, a stable decoupled model may be designed. The suitable rssf leads to a set of fixed modes composed of the only strictly stable invariant zeros.

The decoupling subspaces associated with the two previous decoupling strategies are the subspaces $(\mathcal{V}_i^*)$ and $(\mathcal{V}_{i\,\mathrm{stab}}^*)$, $i = 1, 2$. The expressions of their orthogonal complements are determined according to equations (12), (32) and Property 3.10. Symbolic computations with MAPLE enable the derivation from Property 3.5 of the two associated rssf and the two closed loop transfer matrices, where (a_1, b_1, b_2, b_3) depend of the bond graph parameters. For the first decoupling control law, the closed loop transfer matrix is given by equation

$$T(s) = \begin{bmatrix} g_1/(s^2 + p_1^1 s + a_1 p_0^1) & 0 \\ 0 & g_2/(s + b_1 p_0^2) \end{bmatrix}. \tag{33}$$

As expected, it is a third order matrix: the three remaining modes have been made unassignable. They are the invariant zeros of $BG1$. For the second decoupling control law, the closed loop transfer matrix is given by equation

$$T(s) = \begin{bmatrix} g_1/(s^2 + p_1^1 s + a_1 p_0^1) & 0 \\ 0 & g_2 s/(s^2 + b_2 p_0^2 s + b_3 p_0^2 + p_1^2) \end{bmatrix}. \tag{34}$$

It is a 4^{th} order matrix. As determined by the previous analysis, $s = -1/R_4C_4$ and $s = -1/R_4C_4$ are the fixed modes. According to Property 3.5, p_k^i are degrees of freedom available to tune closed loop dynamics and g_i are parameters used to assign closed loop static gains, $k = 0, 1$, $i = 1, 2$. Note that the model obtained by removing the R-element R_4 would still be decouplable; but closed loop stability could not be performed. Indeed, proportionality between the gains of the shortest different input-output causal paths would change the set of global infinite zero orders: $n'_{2d} = 0$ and $n'_{1d} = 2$. The row infinite zero orders staying unchanged, Theorem 3.6 states that no decoupling rssf exists, achieving closed loop stability.

5 Conclusion

In this paper, structural analysis methods are developed for the input-output decoupling of linear square bond graph models by regular static state feedback.

The poles of the decoupled model are first studied. Due to the non-interaction constraints, some of these poles are fixed: these modes are some of the invariant zeros of the open loop model. Input-output causal path concepts on both bond graph models with integral and derivative causality assignment are used to characterize the symbolic expressions of these invariant zeros. A graphical interpretation of a necessary and sufficient condition is also derived for the input-output decoupling problem with stability to be solvable.

The bond graph methodology is then used to compute a decoupling state feedback insuring stability of the decoupled model, when it is possible. An example is finally presented to detail these analysis and computation methods.

In this paper, it is recalled that the input-output decoupling problem is often achieved with algebraic and geometrical approaches. The bond graph approach is principally based on graphical manipulations and at each step of the procedures information on the model, thus on the physical process, can be analyzed.

References

[1] Basile, G. and Marro, G. *Controlled Invariants and Conditioned Invariants in Linear System Theory.* Englewood Cliffs, Prentice Hall, New Jersey, 1992.

[2] Bertrand, J.M., Sueur, C. and Dauphin-Tanguy, G. Bond graph modelling and geometric approach: input-output decoupling with stability. *IFAC Conference on Control of Industrial Systems*, Belfort, France, 1997, 317–322.

[3] Bertrand, J.M. *Analyse structurelle et commande par decouplage entrée sortie des modèles bond graphs.* PhD Thesis, Lille University, 1997.

[4] Claude, D. *Automatique – Cours de Maîtrise.* Service de Publications Paris Onze édition, Université Paris-Sud, Orsay, 1992.

[5] Descusse, J. and Dion, J.M. On the structure at infinity of linear square decoupled systems. *IEEE Transactions on Automatic Control* **AC-27**(4) (1982) 971–974.

[6] Descusse, J., Lafay, J.F. and Malabre, M. Solution to Morgan's problem. *IEEE Transactions on Automatic Control* **AC-33**(8) (1988) 732–739.

[7] Falb, P.L. and Wolowich, W.A. Decoupling in the design and synthesis of multivariable control systems. *IEEE Transactions on Automatic Control* **AC-12** (1967) 651–659.

[8] Gilbert, E.G. The decoupling of multivariable systems by state feedback. *SIAM Journal of Control and Optimization* **7** (1969) 50–63.

[9] Icart, S., Lafay, J.F. and Malabre, M. A Unified study of the fixed modes of systems decoupled via regular static state feedback. *Joint Conference on New Trends in System Theory, Genova, Italy*, Birkhäuser, Boston, 1990, 425–432.

[10] Karnopp, D.C., Margolis, D.L. and Rosenberg, R.C. *System Dynamics: a Unified Approach.* John Wiley & Sons, Inc., 1990.

[11] Koussiouris, T. A frequency domain approach to the block decoupling problem II: pole assignment while block decoupling a minimal system by state feedback and a constant non singular input transformation and the observability of the block decoupled system. *International Journal of Control* **32** (1980) 443–464.

[12] Martinez Garcia, J.C. and Malabre, M. The row by row decoupling problem with stability: a structural approach. *IEEE Transactions on Automatic Control* **AC-39**(12) (1994) 2457–2460.

[13] Morgan, B.S. The synthesis of linear multivariable systems by state feedback. *Joint Automatic Control Conference*, 1964, 468–472.

[14] Morse, A.S. and Wonham, W.M. Status of noninteracting control. *IEEE Transactions on Automatic Control* **AC-16**(6) (1971) 568–581.

[15] Morse, A.S. Structural invariants of linear multivariable systems. *SIAM Journal of Control and Optimization* **11**(3) (1973) 446–465.

[16] Paynter, H.M. *Analysis Design of Engineering Systems.* Cambridge, Mass. MIT Press, 1960.

[17] Rosenbrock, H.H. *State Space and Multivariable Theory.* Nelson, London, 1970.

[18] Sueur, C. and Dauphin-Tanguy, G. Bond graph approach for structural analysis of MIMO linear systems. *Journal of the Franklin Institute* **328**(1) (1991) 55–70.

[19] Sueur, C. and Dauphin-Tanguy, G. Poles and zeros of multivariable linear systems: a bond graph approach. In: *Bond Graphs for Engineers.* (Eds.: P.C. Breedveld and G. Dauphin-Tanguy), Elsevier Science Publishers B. V., 1992, 211–228.

[20] Verghese, G.C. *Infinite frequency behaviour of generalized dynamical systems.* PhD. Thesis, Electrical Engineering Department, Stanford University, 1978.

[21] Wonham, W.M. Geometric state-space theory in linear multivariable control: a status report. *Automatica* **15** (1979) 5–13.

[22] Wonham, W.M. *Linear Multivariable Control: A Geometric Approach.* Springer Verlag, third edition, New York, 1985.

[23] Wonham, W.M. and Morse, A.S. Decoupling and pole assignment in linear multivariable systems: a geometric approach. *SIAM Journal of Control and Optimization* **8**(1) (1970) 1–18.

Appendix

A Row invariant zeros

Proof of Property 2.11.

□ Suppose $G_i(s)$ the transfer matrix of the subsystem $\sum(c_i, A, B)$ – equations (35)

$$\begin{cases} G_i(s) = N_i(s)/D(s), \\ N_i(s) = [N_{i1}(s) \dots N_{im}(s)]. \end{cases} \tag{35}$$

The state matrix is invertible, thus for bond graph models the subsystem $\sum(c_i, A, B)$ is structurally controllable and observable. The invariant zeros of the subsystem $\sum(c_i, A, B)$ are therefore its null transmission zeros and the null zeros of matrix $G_i(s)$.

The null zeros of matrix $G_i(s)$ are all the zeros of the polynomial matrix $N_i(s)$. These null zeros are zeros of matrix $N_i(s)$ invariant polynomials. As $N_i(s)$ is a row matrix, it has only one invariant polynomial, denoted $\lambda_1^i(s)$. $\lambda_1^i(s)$ is the gcd of the polynomials $\{N_{i1}(s), \dots, N_{im}(s)\}$. The row subsystem $\sum(c_i, A, B)$ null invariant zeros are the common null roots of the transfer matrix numerators. The transfer matrix $G_i(s)$ is given by equation

$$G_i(s) = c_i\,(sI - A)^{-1}\,B. \tag{36}$$

An equivalent expression is equation

$$G_i(s) = c_i\left(sA^{-1} - I\right)^{-1} A^{-1}B. \tag{37}$$

Around $s = 0$, equation (38) can be written

$$\left[\left(sA^{-1} - I\right)^{-1}\right]_{s\to 0} = -\left[I + sA^{-1} + s^2A^{-2} + \dots\right]. \tag{38}$$

A representation of $G_i(s)$ deduced from equations (37) and (38) is given by equation

$$[G_i(s)]_{s\to 0} = -\left[\left(c_iA^{-1}B\right) + \left(c_iA^{-2}B\right)s + \left(c_iA^{-3}B\right)s^2 + \dots\right]. \tag{39}$$

With the new Definition 2.11 of the integer n_{id}, the expression of the matrix $[G_i(s)]_{s\to 0}$ is given by equation

$$[G_i(s)]_{s\to 0} = -\left[(c_iA^{-(n_{id}+1)}B)s^{n_{id}} + (c_iA^{-(n_{id}+2)}B)s^{(n_{id}+1)} + \dots\right]. \tag{40}$$

The expression of matrix $G_i(s)$ around $s = 0$ is given by equation

$$[G_i(s)]_{s\to 0} = -s^{n_{id}}\left[(c_iA^{-(n_{id}+1)}B) + (c_iA^{-(n_{id}+2)}B)s + \dots\right]. \tag{41}$$

n_{id} is equal to the number of common null roots of each transfer matrix $G_i(s)$ numerator. The number of row subsystem $\sum(c_i, A, B)$ null invariant zeros is thus equal to n_{id}. □

B Global invariant zeros

□ Consider an invertible square system $\sum(C, A, B)$. Suppose $P(s)$ its system matrix and $G(s)$ its transfer matrix. These two matrices satisfy equation

$$\det[P(s)] = \det[sI - A] \cdot \det[G(s)]. \tag{42}$$

System $\sum(C, A, B)$ invariant zeros are the roots of $\det[P(s)]$. Given that the state matrix is invertible, the number of null invariant zeros of $\sum(C, A, B)$ is equal to the number of $\det[G(s)]$ null roots.

Consider the bond graph model with a derivative causality assignment (BGD), and its transfer matrix $G_d(s)$ deduced from equations (14). Around $s = 0$, this matrix satisfies equation

$$[G_d(s)]_{s\to 0} = \left[\left(-CA^{-1}B\right) + \left(-CA^{-2}B\right)s + \ldots\right]. \tag{43}$$

Suppose θ'_k the constant coefficient matrix of input output causal path gains of length k in the BGD. Equation (43) can be rewritten as equation

$$[G_d(s)]_{s\to 0} = \sum_{k=0}^{\infty} \theta'_k s^k \tag{44}$$

or equivalently as equation

$$[G_d(1/s)]_{s\to\infty} = \sum_{k=0}^{\infty} \theta'_k / s^k. \tag{45}$$

Suppose a bond graph model with direct transmission between the input sources and the output detectors. The output equation is given by

$$y(t) = Cx(t) + Du(t). \tag{46}$$

On the BGI, around $s \to \infty$ the transfer matrix $G(s)$ is written as equation

$$[G(s)]_{s\to\infty} = \left[D + \frac{CB}{s} + \frac{CAB}{s^2} + \ldots\right]. \tag{47}$$

Suppose θ_k the constant coefficient matrix of input output causal path gains of length k in the BGI. Equation (47) can be rewritten as equation

$$[G(s)]_{s\to\infty} = \sum_{k=0}^{\infty} \theta_k / s^k. \tag{48}$$

With equations (44), (45) and (48), matrices $[G_d(s)]_{s\to 0}$ and $[G(s)]_{s\to\infty}$ can be written with the same formalism. It allows to conclude that the set of integers $\{n'_{1d}, \ldots, n'_{md}\}$

are obtained from matrix $G_d(1/s)$ Smith McMillan form at infinity. This matrix, denoted as $\Phi'(s)$ satisfies equations

$$\begin{cases} G_d(1/s) = J_1'(s) \cdot \Phi'(s) \cdot J_2'(s), \\ \det\left[\lim_{s\to\infty} \{J_k'(s)\}\right] \neq 0 \quad \text{with} \quad k = 1, 2, \\ \Phi'(s) = \operatorname{diag}\left\{s^{-n'_{1d}}, \dots, s^{-n'_{md}}\right\}. \end{cases} \tag{49}$$

From equations (49) follows equation

$$\det\left[G_d(1/s)\right]_{s\to\infty} \approx K_1' \cdot K_2' \cdot 1/\left(s^{\left(\sum_{i=1}^m n'_{id}\right)}\right), \tag{50}$$

with K_1' and K_2' non zero constants, or equivalently equation

$$\det\left[G_d(s)\right]_{s\to 0} \approx K_1' \cdot K_2' \cdot s^{\left(\sum_{i=1}^m n'_{id}\right)}. \tag{51}$$

According that matrices $G(s)$ and $G_d(s)$ are equal, can be written equation

$$\det[G(s)]_{s\to 0} \approx K_1' \cdot K_2' \cdot s^{\left(\sum_{i=1}^m n'_{id}\right)}. \tag{52}$$

From equation (52), it comes that the number of null invariant zeros in BGI is equal to the sum of the infinite zero orders of the BGD for square models. The property remains valid for non square models, that is with $m > p$. □

C Stabilizing decoupling subspace

The proof is divided in three parts. At first, it is shown that the two subspaces $\mathcal{V}_i^{*\perp}$ and $\mathcal{V}_{id}^{*\perp}$ are such as $\mathcal{V}_i^{*\perp} \oplus \mathcal{V}_{id}^{*\perp} = \mathcal{V}_{is}^{*\perp}$ – **step 1**, then that $\mathcal{V}_{is}^*$ is a $(A, \mathcal{B})$ invariant subspace included in the subspace $\mathcal{V}_i^*$ – **step 2**. It is then shown that $\mathcal{V}_{is}^*$ is equal to $\mathcal{V}_{i\,\text{stab}}^*$ – **step 3**.

Step 1: $\mathcal{V}_{is}^{*\perp} = \mathcal{V}_i^{*\perp} \oplus \mathcal{V}_{id}^{*\perp}$.

Consider the $\mathcal{V}_i^{*\perp}$ subspace basis defined by equation

$$\mathcal{V}_i^{*\perp} = \operatorname{vect}\left\{(c_i)^t, \dots, \left(c_i A^{n_i - 1}\right)^t\right\}, \qquad n_i \geq 1. \tag{53}$$

$\mathcal{V}_{id}^{*\perp}$ subspace basis is defined by equation

$$\mathcal{V}_{id}^{*\perp} = \operatorname{vect}\left\{\left(c_i A^{-1}\right)^t, \dots, \left(c_i A^{-n_{id}}\right)^t\right\}, \qquad n_{id} \geq 1 \tag{54}$$

if $n_{id} \geq 1$, else $\mathcal{V}_{id}^{*\perp} = 0$. n_{id} is the smallest integer satisfying equations

$$\begin{cases} c_i A^{-(k+1)} B = 0, \quad k < n_{id}, \\ c_i A^{-(n_{id}+1)} B \neq 0. \end{cases} \tag{55}$$

Consider S_i^* the smallest (c_i, A) invariant subspace containing $\mathcal{B}$, defined by the following algorithm

$$\begin{cases} S_i^0 = 0, \\ S_i^\mu = \mathcal{B} + A((\ker c_i) \cap S_i^{\mu-1}). \end{cases} \tag{56}$$

$S_i^{*\perp}$ and $\mathcal{V}_{id}^{*\perp}$ are related as equation

$$\mathcal{V}_{id}^{*\perp} \subset S_i^{*\perp}. \tag{57}$$

Indeed, suppose the product $(\mathcal{V}_{id}^{*\perp})^t \cdot S_i^*$, according to equations (54), (55) and (56), the first basis vector of subspace $\mathcal{V}_{id}^{*\perp}$ satisfies equation

$$c_i A^{-1} \cdot S_i^* = 0. \tag{58}$$

The same equation can be written for each $\mathcal{V}_{id}^{*\perp}$ basis vector. The last one satisfies equation

$$c_i A^{-n_{id}} \cdot S_i^* = 0. \tag{59}$$

It is thus possible to deduce equation

$$(\mathcal{V}_{id}^{*\perp})^t \cdot S_i^* = 0, \tag{60}$$

which implies equation (57).

Consider now the subspace $\mathcal{V}_{is}^{*\perp}$ satisfying equation

$$\mathcal{V}_{is}^{*\perp} = \mathcal{V}_i^{*\perp} + \mathcal{V}_{id}^{*\perp}. \tag{61}$$

$\sum(C, A, B)$ is right invertible and thus row subsystems $\sum(c_i, A, B)$ are in the same way right invertible. Equation (62) can be written

$$\mathcal{V}_i^* + S_i^* = \mathcal{X}. \tag{62}$$

Equation (63) can be deduced

$$\mathcal{V}_i^{*\perp} \cap S_i^{*\perp} = 0. \tag{63}$$

From equations (63) and (57) it comes:

$$\mathcal{V}_i^{*\perp} \cap \mathcal{V}_{id}^{*\perp} = 0. \tag{64}$$

According to equations (61) and (64), subspaces $\mathcal{V}_i^{*\perp}$ and $\mathcal{V}_{id}^{*\perp}$ satisfy equation

$$\mathcal{V}_{is}^{*\perp} = \mathcal{V}_i^{*\perp} \oplus \mathcal{V}_{id}^{*\perp}. \tag{65}$$

Step 2: $\mathcal{V}_{is}^*$ is a $(A, \mathcal{B})$ invariant subspace included in $\mathcal{V}_i^*$.

$\mathcal{V}_{is}^*$ is a $(A, \mathcal{B})$ invariant subspace iff it satisfies equation

$$A\mathcal{V}_{is}^* \subset \mathcal{V}_{is}^* + \mathcal{B}. \tag{66}$$

It is sufficient to prove that for each vector $x \in \mathcal{V}_{is}^*$, equation

$$\left\{\mathcal{V}_{is}^{*\perp} \cap \mathcal{B}^{\perp}\right\}^t \cdot Ax = 0 \tag{67}$$

is satisfied.

Consider a subspace of $\left\{\mathcal{V}_{is}^{*\perp} \cap \mathcal{B}^{\perp}\right\}$. According to equation (65), a subspace basis is the union of subspace basis $\left\{\mathcal{V}_{i}^{*\perp} \cap \mathcal{B}^{\perp}\right\}$ and $\left\{\mathcal{V}_{id}^{*\perp} \cap \mathcal{B}^{\perp}\right\}$. Basis vectors of subspace $\mathcal{V}_{i}^{*\perp}$ belonging to subspace $\mathcal{B}^{\perp}$ are identified thanks to the integer n_i which satisfies equation

$$\begin{cases} c_i A^{(k-1)} B = 0, \quad k < n_i, \\ c_i A^{(n_i-1)} B \neq 0. \end{cases} \tag{68}$$

This equation can be rewritten as equation

$$\begin{cases} (c_i A^{(k-1)})^t \in \mathcal{B}^{\perp}, \quad k < n_i, \\ (c_i A^{(n_i-1)})^t \notin \mathcal{B}^{\perp}. \end{cases} \tag{69}$$

According to equations (53) and (69), subspace $\left\{\mathcal{V}_{i}^{*\perp} \cap \mathcal{B}^{\perp}\right\}$ can be describes by the following equation:

$$\left\{\mathcal{V}_{i}^{*\perp} \cap \mathcal{B}^{\perp}\right\} = \operatorname{vect}\left\{(c_i)^t, \dots, \left(c_i A^{n_i-2}\right)^t\right\}. \tag{70}$$

With the same manner, basis vectors of subspace $\mathcal{V}_{id}^{*\perp}$ belonging to subspace $\mathcal{B}^{\perp}$ are identified thanks to the integer n_{id} which satisfies equation

$$\begin{cases} (c_i A^{-(k+1)})^t \in \mathcal{B}^{\perp}, \quad k < n_{id}, \\ (c_i A^{-(n_{id}+1)})^t \notin \mathcal{B}^{\perp}. \end{cases} \tag{71}$$

According to equations (54) and (71), a basis of subspace $\left\{\mathcal{V}_{id}^{*\perp} \cap \mathcal{B}^{\perp}\right\}$ is described by equation

$$\left\{\mathcal{V}_{id}^{*\perp} \cap \mathcal{B}^{\perp}\right\} = \operatorname{vect}\left\{(c_i A^{-1})^t, \dots, (c_i A^{-n_{id}})^t\right\}. \tag{72}$$

Equations (70) and (72) allow to write a basis for subspace $\left\{\mathcal{V}_{is}^{*\perp} \cap \mathcal{B}^{\perp}\right\}$:

$$\left\{\mathcal{V}_{is}^{*\perp} \cap \mathcal{B}^{\perp}\right\} = \operatorname{vect}\left\{(c_i)^t, \dots, (c_i A^{n_i-2})^t \,\middle|\, (c_i A^{-1})^t, \dots, (c_i A^{-n_{id}})^t\right\}. \tag{73}$$

Let us prove that each vector $x \in \mathcal{V}_{is}^*$ satisfies equation (67). This equation can be rewritten as equation

$$\mathcal{V}_{is}^* = \mathcal{V}_{i}^* \cap \mathcal{V}_{id}^*. \tag{74}$$

It means that each vector x belonging to subspace $\mathcal{V}_{is}^*$ also belongs to $\mathcal{V}_{i}^*$. It satisfies equation

$$(\mathcal{V}_{i}^{*\perp})^t \cdot x = 0. \tag{75}$$

From equations (53) and (75), it comes that each vector $x \in \mathcal{V}_{is}^*$ satisfies equations

$$\begin{cases} c_i x = 0, \\ c_i A x = 0, \\ \vdots \\ c_i A^{(n_i-1)} x = 0. \end{cases} \tag{76}$$

According to equation (74), if x belong to subspace $\mathcal{V}_{is}^*$ it also belongs to subspace $\mathcal{V}_{id}^*$, and satisfies equation

$$(\mathcal{V}_{id}^{*\perp})^t \cdot x = 0. \tag{77}$$

Thus, each vector $x \in \mathcal{V}_{is}^*$ satisfies equations

$$\begin{cases} c_i A^{-1} x = 0, \\ c_i A^{-2} x = 0, \\ \vdots \\ c_i A^{-n_{id}} x = 0. \end{cases} \tag{78}$$

For each vector basis z belonging to subspace $\left\{\mathcal{V}_{is}^{*\perp} \cap \mathcal{B}^\perp\right\}$, expression $z^t A x$ is calculated with $x \in \mathcal{V}_{is}^*$. For each vector $v_k = (c_i A^k)^t$, $k = 0, \dots, (n_i - 2)$, from equation (76) it comes:

$$v_k^t A x = 0, \qquad x \in \mathcal{V}_{is}^*. \tag{79}$$

For each vector $w_k = (c_i A^{-k})^t$, $k = 1, \dots, n_{id}$, from equation (76) and (78) it comes equation

$$w_k^t A x = 0. \tag{80}$$

Thus, each basis vector z belonging to subspace $\left\{\mathcal{V}_{is}^{*\perp} \cap \mathcal{B}^\perp\right\}$ satisfies equation

$$z^t A x = 0, \qquad x \in \mathcal{V}_{is}^*. \tag{81}$$

For each vector $x \in \mathcal{V}_{is}^*$, it comes equation

$$\left\{\mathcal{V}_{is}^{*\perp} \cap \mathcal{B}^\perp\right\}^t \cdot A x = 0. \tag{82}$$

Thus, $\mathcal{V}_{is}^*$ is a $(A, \mathcal{B})$ invariant subspace and from equation (74) it can be concluded that this subspace is included in subspace $\mathcal{V}_i^*$.

Step 3: $\mathcal{V}_{is}^* = \mathcal{V}_{i\,\mathrm{stab}}^*$.

Step 1 and Step 2 allow to prove that $\mathcal{V}_{is}^*$ satisfies the following properties:

$$\begin{cases} \mathcal{V}_{is}^{*\perp} = \mathcal{V}_i^{*\perp} \oplus \mathcal{V}_{id}^{*\perp}, \\ \mathcal{V}_{is}^* \text{ is a } (A, \mathcal{B}) \text{ invariant subspace,} \\ \dim(\mathcal{V}_{is}^{*\perp}) = n_i + n_{id}. \end{cases} \tag{83}$$

If the row subsystem $\sum(c_i, A, B)$ does not contain any strictly instable invariant zero, it is possible to write equation

$$\dim(\mathcal{V}_{is}^{*\perp}) = n_i + C^+(c_i, A, B). \tag{84}$$

Then, subspace $\mathcal{V}_{is}^*$ satisfies equation

$$\dim(\mathcal{V}_{is}^*) = \dim(\mathcal{V}_i^*) - C^+(c_i, A, B). \tag{85}$$

Equations (83) and (85) allow to conclude that subspace $\mathcal{V}_{is}^*$ satisfies the following properties:

$$\begin{cases} \mathcal{V}_{is}^* \text{ is a } (A, \mathcal{B}) \text{ invariant subspace,} \\ \mathcal{V}_{is}^* \subset \mathcal{V}_i^*, \\ \dim(\mathcal{V}_{is}^*) = \dim(\mathcal{V}_i^*) - C^+(c_i, A, B). \end{cases} \tag{86}$$

From equation (86) it follows the conclusion: if the row subsystem $\sum(c_i, A, B)$ does not contain any strictly instable invariant zero, subspace $\mathcal{V}_{is}^*$ is the greatest internally stable $(A, \mathcal{B})$ invariant subspace included in $(\ker c_i)$. It is thus equal to subspace $\mathcal{V}_{i\,\mathrm{stab}}^*$. □

Nonlinear Dynamics and Systems Theory, **1**(1) (2001) 59–79

Preconditioning and Conditioning of Systems Arising from Boundary Value Methods*

F. Iavernaro[1] and D. Trigiante[2]

[1] *Dipartimento di Matematica, Università di Bari, Via Orabona 4, I-70125 Bari, Italy*
[2] *Dipartimento di Energetica, Università di Firenze, via C. Lombroso 6/17, I-50134 Firenze, Italy*

Received: February 29, 2000; Revised: July 13, 2000

Abstract: The application of Boundary Value Methods to several classes of Differential Equations requires the solution of large dimension and sparse linear systems having (block) quasi-Toeplitz coefficient matrices. This has naturally suggested the use of Krylov subspace methods in combination with well known preconditioners suitable for Toeplitz matrices. However, the behaviour of such methods is closely related to the continuous problem (in the simplest case the system to be solved depends on a complex parameter) and some aspects need to be carefully studied in order to determine the effectiveness of these preconditioners and even their compatibility with some basic concepts in this area. Considerations about the choice of an optimal preconditioner are also presented.

Keywords: *Circulant preconditioners; Toeplitz-like matrices; initial value problems; linear multistep formulae; boundary value methods.*

Mathematics Subject Classification (2000): 65F10, 65L05, 65L20, 65F15, 15A18.

1 Introduction

Boundary Value Methods (BVMs) are a relatively recent class of methods for the numerical treatment of a wide variety of differential equations (IVPs, BVPs, DAEs, PDEs) (see for example [2, 3, 7, 10–13, 17]). Their application transforms a continuous differential problem of dimension m into a discrete one of dimension mn, represented by a system of the form

$$(A_n \otimes I_m)Y - h(B_n \otimes I_m)F(Y) = \boldsymbol{\delta}. \tag{1}$$

*Work supported by MURST and GNIM.

© 2001 Informath Publishing Group. All rights reserved.

The matrices A_n, B_n are square of dimension n, $\boldsymbol{\delta}$ is a known vector of length mn, I_m is the identity matrix of dimension m, $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$ is a block vector of length mn whose components $\mathbf{y}_i \in \mathbb{R}^m$ are approximations to the true solution at given mesh points. The vector $F(Y) = [\mathbf{f}(\mathbf{y}_1), \ldots, \mathbf{f}(\mathbf{y}_n)]^T$ contains the evaluations over $\mathbf{y}_i$ of a function $\mathbf{f}\colon \mathbb{R}^m \to \mathbb{R}^m$ which is typically defined by the continuous problem, while the step-length h depends on n and the time integration interval. For instance, choosing $\boldsymbol{\delta} = -\mathbf{a}_0 \otimes I_m \mathbf{y}_0 + h\mathbf{b}_0 \otimes I_m \mathbf{f}(t_0, \mathbf{y}_0)$ ($\mathbf{a}_0$ and $\mathbf{b}_0$ are vectors of length n), the system (1) may be considered as the discrete counterpart of the IVP

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(\mathbf{y}), & t \in [t_0, t_0 + T], \\ \mathbf{y}(t_0) = \mathbf{y}_0, \end{cases} \tag{2}$$

where now $h = T/n$. By definition BVMs give to the matrices A_n and B_n a banded quasi Toeplitz structure with bandwidth k independent of n. Applied to the problem (2) they take the form

$$A_n = \begin{pmatrix} \alpha_1^{(1)} & \alpha_2^{(1)} & \ldots & \alpha_k^{(1)} & & & \\ \vdots & \vdots & \vdots & \vdots & & & \\ \alpha_1^{(k_1-1)} & \alpha_2^{(k_1-1)} & \ldots & \alpha_k^{(k_1-1)} & & & \\ \alpha_1 & \alpha_2 & \ldots & \alpha_k & & & \\ & \alpha_0 & \alpha_1 & \ldots & \alpha_k & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & \alpha_0 & \alpha_1 & \ldots & \alpha_k \\ & & & \alpha_0^{(k_1)} & \alpha_1^{(k_1)} & \ldots & \alpha_k^{(k_1)} \\ & & & \vdots & \vdots & \vdots & \vdots \\ & & & \alpha_0^{(k-1)} & \alpha_1^{(k-1)} & \ldots & \alpha_k^{(k-1)} \end{pmatrix}_{n\times n},$$

$$B_n = \begin{pmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \ldots & \beta_k^{(1)} & & & \\ \vdots & \vdots & \vdots & \vdots & & & \\ \beta_1^{(k_1-1)} & \beta_2^{(k_1-1)} & \ldots & \beta_k^{(k_1-1)} & & & \\ \beta_1 & \beta_2 & \ldots & \beta_k & & & \\ & \beta_0 & \beta_1 & \ldots & \beta_k & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & \beta_0 & \beta_1 & \ldots & \beta_k \\ & & & \beta_0^{(k_1)} & \beta_1^{(k_1)} & \ldots & \beta_k^{(k_1)} \\ & & & \vdots & \vdots & \vdots & \vdots \\ & & & \beta_0^{(k-1)} & \beta_1^{(k-1)} & \ldots & \beta_k^{(k-1)} \end{pmatrix}_{n\times n},$$

while $\mathbf{a}_0 = \left[\alpha_0^{(1)}, \alpha_0^{(k_1-1)}, \alpha_0, 0, \ldots, 0\right]$ and $\mathbf{b}_0 = \left[\beta_0^{(1)}, \beta_0^{(k_1-1)}, \beta_0, 0, \ldots, 0\right]$. The i-th component of (1) is actually a k-step linear formula with k_1 initial and $k_2 = k - k_1$ final conditions. Its coefficients $\alpha_i,\ \beta_i,\ i = 0,\dots,k$ are determined imposing that $\mathbf{y}_i$ is an approximation of order p to the true solution $\mathbf{y}(t_i)$. In such a case p is also the order of the BVM, and the local truncation error assumes the form

$$\boldsymbol{\tau}(h) \equiv A_n\hat{Y} - hB_nF(\hat{Y}) + \mathbf{a}_0 \otimes I_m\mathbf{y}_0 - h\mathbf{b}_0 \otimes I_m\mathbf{f}(t_0,\mathbf{y}_0) = h^{p+1}G(\boldsymbol{\xi}), \tag{3}$$

where $\hat{Y} = [\mathbf{y}(t_1),\dots,\mathbf{y}(t_n)]^T$ is the vector of evaluations of the true solution $\mathbf{y}(t)$ of (2) at the internal mesh times t_i and $G(\boldsymbol{\xi}) = [c_1\mathbf{y}^{(p+1)}(\boldsymbol{\xi}_1),\dots,c_n\mathbf{y}^{(p+1)}(\boldsymbol{\xi}_n)]^T$, with c_i the error constant of the i-th formula. The first $k_1 - 1$ and the final k_2 components of (1) are called respectively initial and final methods, and they cause the loss of Toeplitz structure which is instead conferred by the main method in the remaining rows. We remark that similar arguments are also valid for the other class of evolutionary problems to which BVMs have been applied.

The system (1) is nonlinear if $\mathbf{f}$ is so and its solution Y is therefore obtained as the limit of a sequence of vectors Y^k computed as solution of suitable linear systems. Here we suppose to linearize (1) in a neighborhood of its solution according to a simplified Newton iteration that gives rise to the scheme

$$(A_n \otimes I_m - hB_n \otimes J_k)\big(Y^{k+1} - Y^k\big) = G(Y^k), \tag{4}$$

where $G(Y^k) = \boldsymbol{\delta} - (A_n \otimes I_m)Y^k + h(B_n \otimes I_m)F(Y^k)$ and J_k is the Jacobian of $\mathbf{f}(y)$ evaluated at a suitable component of the current vector Y^k (in the simplest case J_k is independent of k). We observe that a similar system as (4) is to be solved when the continuous problem is linear and autonomous, namely $\mathbf{f}(y) = J\mathbf{y} + \mathbf{b}$. In this paper we are interested in analysing the properties of some Krylov subspace methods (see [14]) such as GMRES or BICGSTAB as applied to such linear systems subject to preconditioning and hence, until the convergence of the procedure (4) will be considered, it is reasonable to confine our analysis to linear problems only. The block quasi-Toeplitz and banded structure of the matrix $M_n = (A_n \otimes I_m - hB_n \otimes J)$, has suggested the use of preconditioners that normally work well when applied to Toeplitz or block Toeplitz matrices. In [8] the authors compare the efficiency of some preconditioning techniques showing, on the basis of their experiments, that good results, in terms of computational complexity, is achieved considering the block circulant preconditioner $S_n = C_n^A \otimes I_m - hC_n^B \otimes J$, where C_n^A and C_n^B are the Strang circulant preconditioners generated by the main method [15]:

$$C_n^A = \begin{pmatrix}
\alpha_{k_1} & \dots & \alpha_k & & & & \alpha_0 & \dots & \alpha_{k_1-1} \\
\vdots & \ddots & & \ddots & & & & \ddots & \vdots \\
\alpha_1 & & \ddots & & \ddots & & & & \alpha_0 \\
& \ddots & & \ddots & & \ddots & & & \\
& & \ddots & & \ddots & & \ddots & & \\
& & & \ddots & & \ddots & & \ddots & \\
\alpha_k & & & & \ddots & & \ddots & & \alpha_{k-1} \\
\vdots & \ddots & & & & \ddots & & \ddots & \vdots \\
\alpha_{k_1+1} & \dots & \alpha_k & & & & \alpha_0 & \dots & \alpha_{k_1}
\end{pmatrix}_{n\times n},$$

and analogously for C_n^B, with β_i instead of α_i (for simplicity, S_n will also be referred to as the Strang preconditioner).

For a convergent BVM $(p \geq 1)$, one has from (3), $\sum_{i=0}^{k} \alpha_i = 0$ and hence $C_n^A \mathbf{e} = \mathbf{0}$, with $\mathbf{e} = [1, \ldots, 1]^T$. It follows that C_n^A is singular for all values of n and this causes the singularity of the preconditioner S_n when $\det(J) = 0$. Indeed, if $\mathbf{x} \in \mathbb{R}^m - \{\mathbf{0}\}$ is such that $Jx = \mathbf{0}$, one also has $S_n X = \mathbf{0}$, with $X = \mathbf{e} \otimes \mathbf{x}$. It is not difficult to realize that this fact produces undesirable effects also when $\det(J) \simeq 0$ due to a bad conditioning of the matrix S_n. Despite the good behaviour presented in [8] (which has favourably impressed the present authors), other elements must be considered that show how the use of S_n as preconditioner of M_n could be unappropriate in several cases. A comparison of preconditioners in terms of their conditioning is in our case indispensable but not new (see for example [16]); in [5, 6] the present problem is outlined and solved by P-circulant preconditioners.

In Sections 2 and 3 we weigh up in more details the pros and cons of this strategy and propose (Section 4) a modification in S_n that prevents a number of drawbacks. Lately (Section 5), we also introduce a modification in the method itself that allow to the Strang preconditioner to work well when $\det(J) = 0$. The properties of all these preconditioners are analysed to show their effectiveness.

2 Circulant Preconditioners for BVMs

As seen for the Strang preconditioner, in general a circulant matrix is a Toeplitz matrix (that is its entries are constant along diagonals) for which the last entry in each row is the first one in the subsequent row. Multiplication of a circulant matrix of dimension n by a vector requires only $O(n \log(n))$ arithmetic operations if the Fast Fourier Transform (FFT) is performed. A circulant matrix $\mathcal{C}$ is in fact similar to a diagonal matrix D via a Fourier transformation matrix V. More precisely we have $\mathcal{C} = VDV^H$, where the diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$ contains the eigenvalues of $\mathcal{C}$ and the Fourier matrix $V = \{v_{jk}\}$ has elements (i is the imaginary unit):

$$v_{jk} = \frac{1}{\sqrt{n}} e^{\frac{2\pi i}{n} jk}, \qquad j,\, k = 0, \ldots, n-1. \tag{5}$$

A consequence of (5) is that

$$\|\mathcal{C}\| = \max_j |d_j|, \qquad \|\mathcal{C}^{-1}\| = \frac{1}{\min\limits_j |d_j|} \quad \text{and} \quad \mu(\mathcal{C}) = \frac{\max\limits_j |d_j|}{\min\limits_j |d_j|},$$

where here and in the rest of the paper $\|\cdot\|$ will denote the 2-norm and $\mu(\mathcal{C}) = \|\mathcal{C}\|\,\|\mathcal{C}^{-1}\|$ is the conditioning number (in 2-norm) of $\mathcal{C}$. Concerning the basic properties of circulant matrices that we will exploit during our discussion, we refer to [9].

To account for the choice of S_n as preconditioner of the matrix M_n, it is sufficient to observe that the preconditioned matrix P_n may be recast as

$$P_n \equiv S_n^{-1} M_n = I_{nm} + S_n^{-1} E_n,$$

with $E_n = M_n - S_n$. Since the rank of E_n is at most km, it follows that, for n large, most of the eigenvalues of P_n coincide with 1, which allows fast convergence of iterative methods like GMRES or BICGSTAB. However the other eigenvalues of P_n also play a role that cannot be neglected. For example, it is not possible to bound them inside a finite region of the complex plane independently of the function $\mathbf{f}$, a circumstance that may be critical when dealing with some classes of problems. To go into the question we will consider, here and in the rest of the paper, a class of BVMs called Generalized Backward Differentiation Formulae (GBDFs) over which a test problem will be performed and mathematical results will be derived. In passing, we emphasize that similar considerations may be easily extended to other classes of methods. The k-step GBDF is defined by choosing B_n as the identity matrix I_n, $\mathbf{b}_0 \equiv \mathbf{e}_1 = (1, 0, \dots, 0)^T$, the index $k_1 = \nu$ according to the formula

$$\nu = \begin{cases} (k+2)/2, & \text{for even } k, \\ (k+1)/2, & \text{for odd } k, \end{cases} \tag{6}$$

and all the coefficients α_i in order that the formula has the highest possible order $p = k$.

As test problem we consider the linear pendulum system

$$\mathbf{y}' = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} \mathbf{y} \tag{7}$$

in the time interval $[0, 2\pi]$ and study the numerical solution obtained by the order 5 GBDF for different values of the frequency $\omega/(2\pi)$ and dimension $n = 100$ (the stepsize is therefore $h = 2\pi/100$).

The behaviour of this simple problem is also typical of more general dynamical systems in a neighborhood of marginally stable equilibrium points or even in a small time interval during which an equilibrium point loses or acquires stability due to the occurrence of a Hopf bifurcation.

The linear system originated by the BVM is solved by the GMRES routine of MATLAB using 10^{-12} as control of the relative residual and the Strang preconditioner S_n as input parameter. To state the inefficiency of S_n for small values of $|h \det(J)|$, we set $\omega = 10^{-m}$, $m = 1, 2, \dots, 8$ and consider, for each value of the frequency, the number of iterations needed to get the numerical solution; this in fact is proportional to the overall cost of the algorithm (numbers of floating point operations).

Figure 2.1 shows an unexpected increase of the computational cost while ω decreases (the smaller the frequency the easier the numerical treatment of the problem should be). The reason of that may be understood looking at the three columns of Table 2.1 that report the conditioning numbers (in 2-norm) of the matrices M_n, S_n and P_n. It is seen that while the conditioning of the GBDF formula (the matrix M_n) stays constant independently of ω, the same is not true for the Strang preconditioner S_n and consequently for the preconditioned matrix P_n. They are indeed proportional to $1/\omega^2$ and as ω decreases, the use of finite precision arithmetic causes a drop in the convergence properties of GMRES and a loss of accuracy in the results. For instance the error is $1.5 \cdot 10^{-12}$ at $\omega = 10^{-1}$ and $5.7 \cdot 10^{-1}$ at $\omega = 10^{-8}$. Such problems also occur fixing a small value for ω and decreasing the stepsize $h = 2\pi/n$. In such a case the global error should decrease as $O(h^p)$ but once again, since $\mu(S_n)$ is proportional to n, loss of accuracy is experienced.

A modified Strang preconditioner $\bar{S}_n$, to be defined in the sequel, has also been used with the same set of parameters. The fifth and sixth columns of Table 2.1 tell us that

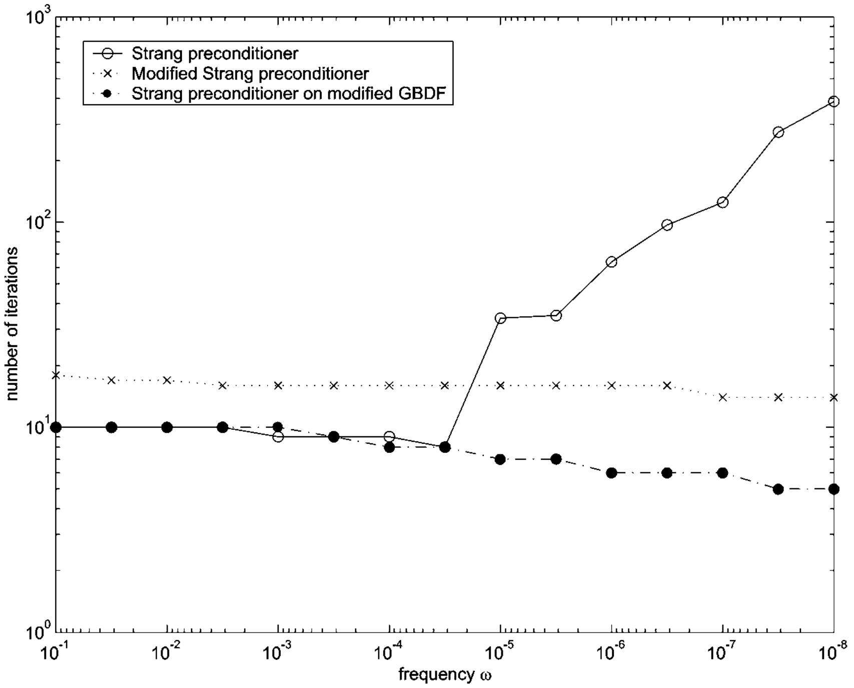

Figure 2.1. Computational cost of GMRES applied to problem (7).

the conditioning of $\bar{S}_n$ and $\bar{P}_n \equiv \bar{S}_n^{-1} M_n$ is comparable to that of M_n and what's more, they are independent of ω which make $\bar{S}_n$ suitable for small values of the frequency. A different but more appealing approach consists in modifying the GBDF formula via a similarity transformation (see Section 5). The new matrix $\widehat{M}_n$ generates a nonsingular circulant matrix $\widehat{S}_n$ even if $h \det(J) = 0$. Figure 2.1 and the conditioning of $\widehat{M}_n$, $\widehat{S}_n$ and $\widehat{P}_n \equiv \widehat{S}_n^{-1} \widehat{M}_n$ in Table 2.1 prove the good behaviour of this technique.

Table 2.1. Comparison of conditioning numbers of the matrices M_n, S_n, P_n, $\bar{S}_n$, $\bar{P}_n$, $\widehat{M}_n$, $\widehat{S}_n$, $\widehat{P}_n$.

ω	$\mu(M_n)$	$\mu(S_n)$	$\mu(P_n)$	$\mu(\bar{S}_n)$	$\mu(\bar{P}_n)$	$\mu(\widehat{M}_n)$	$\mu(\widehat{S}_n)$	$\mu(\widehat{P}_n)$
10^{-1}	$3.3 \cdot 10^3$	$2.6 \cdot 10^3$	$4.3 \cdot 10^5$	$7.6 \cdot 10^2$	$1.2 \cdot 10^5$	$1.7 \cdot 10^3$	$7.6 \cdot 10^2$	$6.2 \cdot 10^4$
$5 \cdot 10^{-2}$	$3.4 \cdot 10^3$	$2.6 \cdot 10^4$	$4.5 \cdot 10^6$	$1.0 \cdot 10^3$	$1.7 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$8.9 \cdot 10^4$
10^{-2}	$3.4 \cdot 10^3$	$2.6 \cdot 10^5$	$4.5 \cdot 10^7$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
$5 \cdot 10^{-3}$	$3.4 \cdot 10^3$	$2.6 \cdot 10^6$	$4.5 \cdot 10^8$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
10^{-3}	$3.4 \cdot 10^3$	$2.6 \cdot 10^7$	$4.5 \cdot 10^9$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
$5 \cdot 10^{-4}$	$3.4 \cdot 10^3$	$2.6 \cdot 10^8$	$4.5 \cdot 10^{10}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
10^{-4}	$3.4 \cdot 10^3$	$2.6 \cdot 10^9$	$4.5 \cdot 10^{11}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
$5 \cdot 10^{-5}$	$3.4 \cdot 10^3$	$2.6 \cdot 10^{10}$	$4.5 \cdot 10^{12}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
10^{-5}	$3.4 \cdot 10^3$	$2.6 \cdot 10^{11}$	$4.5 \cdot 10^{13}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
$5 \cdot 10^{-6}$	$3.4 \cdot 10^3$	$2.6 \cdot 10^{12}$	$4.5 \cdot 10^{14}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
10^{-6}	$3.4 \cdot 10^3$	$2.6 \cdot 10^{13}$	$4.5 \cdot 10^{15}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
$5 \cdot 10^{-7}$	$3.4 \cdot 10^3$	$2.6 \cdot 10^{14}$	$4.5 \cdot 10^{16}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
10^{-7}	$3.4 \cdot 10^3$	$2.5 \cdot 10^{15}$	$4.3 \cdot 10^{17}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
$5 \cdot 10^{-8}$	$3.4 \cdot 10^3$	$1.6 \cdot 10^{17}$	$1.5 \cdot 10^{19}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$
10^{-8}	$3.4 \cdot 10^3$	$1.8 \cdot 10^{16}$	$1.5 \cdot 10^{21}$	$1.0 \cdot 10^3$	$1.8 \cdot 10^5$	$1.8 \cdot 10^3$	$1.0 \cdot 10^3$	$9.3 \cdot 10^4$

In the next section the dependence of the conditioning of P_n on both the problem and the dimension n is analysed. It is custom in numerical analysis, to carry out the study of the discrete problem as applied to the scalar test equation

$$y' = \lambda y, \quad \lambda \in \mathbb{C}. \tag{8}$$

This approach reduces the complexity of calculus and may be easily generalized to the vector case in many cases of interest (for example when the system has a complete set of eigenfunctions).

3 Preconditioning and Conditioning

The discrete problem corresponding to (8) has dimension n and is now defined for the GBDFs by the matrix $M_n = A_n - h\lambda I_n$, with $h = T/n$. From the arbitrariness of λ, it follows that it is not a restriction to consider $T = 1$.

Of particular interest in the following is the main method of the GBDF, defined by the polynomial pair (ρ, σ):

$$\rho(z) = \sum_{j=0}^{k} \alpha_j z^j, \qquad \sigma(z) = z^{\nu}.$$

A link between the method and the algebraic properties of the preconditioner is in the function $g(z) = \rho(z)/\sigma(z)$ which generates the boundary locus of the former when evaluated at $z = e^{i\theta}$, $\theta \in [0, 2\pi]$ (i is the imaginary unit), and represents the symbol of the latter apart from a translation of size $-\lambda/n$ in the complex plane. Figure 3.1 reports the boundary loci of the main method of GBDFs up to the order 7. These curves also approximate the boundaries of the A-stability regions of the methods when n is large and state that GBDFs are indeed A-stable methods.

A necessary condition for A-stability is that all the eigenvalues of the matrix M_n have positive real part, when $\lambda \in \mathbb{C}^-$, where $\mathbb{C}^-$ is the left half of the complex plane. It follows that the solution of the equivalent method identified by the matrix $P_n = S_n^{-1} M_n$ will retain all the stability properties of the original one if none of the eigenvalues of S_n lies in $\mathbb{C}^-$ when $\lambda \in \mathbb{C}^-$. However the eigenvalues of the circulant matrix S_n are $g(e^{(2\pi i/n)j}) + \lambda/n$, $j = 0, \ldots, n-1$, and since $Re(g(e^{i\theta})) \geq 0$ they actually have nonnegative real part. Unfortunately the matrix S_n has $d_1 = \lambda/n$ as eigenvalue of minimum real part and consequently $d_1 = 0$ if $\lambda = 0$. Taking into account that the conditioning number of a circulant matrix is the ratio between the maximum and minimum modulus of its eigenvalues, it follows that $\mu(S_n)$ behaves at least as $O(n/\lambda)$. This means that, although both $\mu(M_n)$ and $\mu(S_n)$ are proportional to their dimension n, the latter cannot be bounded from below by a quantity independent of the problem: despite M_n, the preconditioner S_n may become ill conditioned if $\lambda \simeq 0$. In Figure 3.2, the location of the eigenvalues of M_n and S_n is displayed for $n = 80$, $\lambda = -1$ and order $p = 5$. We see that all the eigenvalues of M_n (except two) are inside the region delimited by the boundary locus and away from zero (see [4] for a characterization of the asymptotic spectra of banded quasi-Toeplitz matrices), whereas the eigenvalues of S_n place themselves on the boundary locus which in turn passes near zero for small values of λ/n.

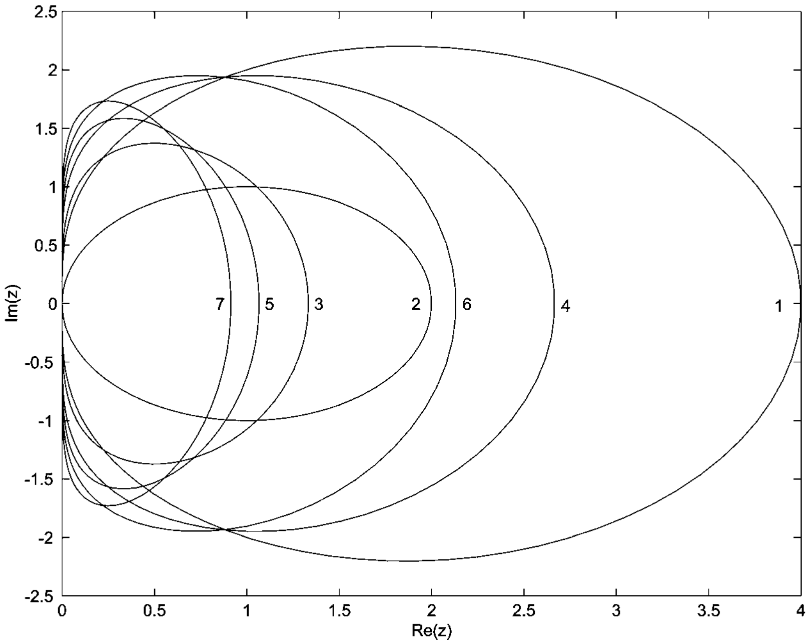

Figure 3.1. Boundary loci of the main formulae of GBDFs up to the order 7.

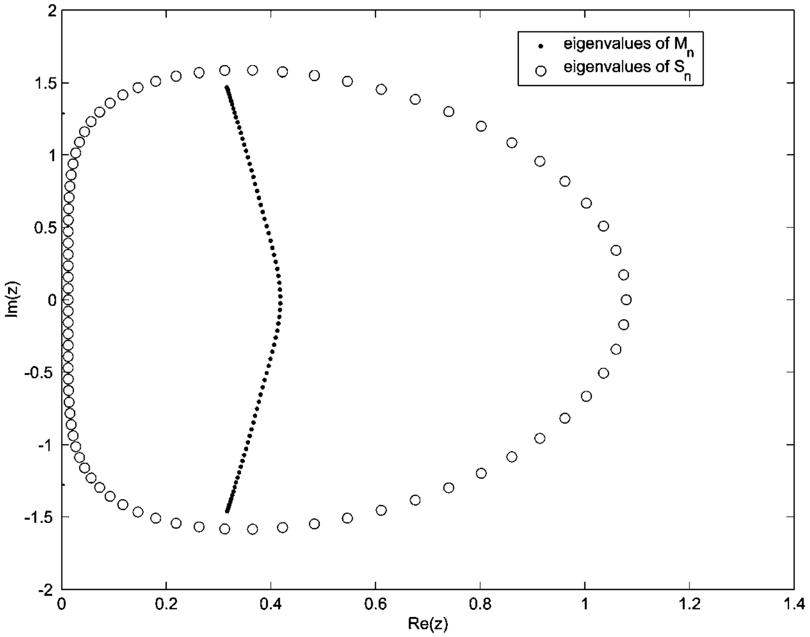

Figure 3.2. Eigenvalues of M_n and S_n of the order 5 GBDF.

From the relation

$$\mu(S_n) = \mu(M_n M_n^{-1} S_n) = \mu(M_n P_n^{-1}) \le \mu(M_n)\mu(P_n)$$

it follows that $\mu(P_n) \ge \mu(S_n)/\mu(M_n)$ and the same considerations hold true for the preconditioned matrix P_n. As seen above for the pendulum problem, the possible dangerous effects of that, are the loss of accuracy in the numerical computation and the weakening of the convergence properties of the iteration procedure used to determine the solution of the linear system.

As concerns nonlinear dynamical systems, the overall integration interval is usually partitioned into adjacent subintervals in each of which a scheme of the form (4), based on the Newton method, is performed to get the solution. It is then clear that analogous problems may be encountered in the convergence properties of (4) when $\det(J_k) \simeq 0$ and S_n is used as preconditioner.

For the sake of simplicity, we shall suppose in the following $\lambda \in (-\epsilon, 0]$, $\epsilon > 0$. The restriction to the real case makes the calculation easier and, using a continuity argument, it describes as well the behaviour of the complex problem in a neighborhood of zero, which is the primary objective of the present analysis. In the rest of the paper the matrix M_n will therefore assume the expression

$$M_n = A_n + \frac{|\lambda|}{n} I_n, \tag{9}$$

and since $C_n^B = I_n$, to simplify the notation C_n will stand for C_n^A.

4 A Modified Strang Preconditioner

We focus now our attention to the conditioning of the preconditioned matrix $P_n = S_n^{-1} M_n$. The final purpose is to introduce a family of preconditioners depending on a real parameter γ in order that for the new preconditoned matrix $\bar{P}_n(\gamma)$ the inequality

$$\mu(\bar{P}_n(\gamma)) \le c\mu(M_n) \tag{10}$$

may hold true with the constant $c \ge 1$ independent of n and of moderate size. To begin, we introduce the family of preconditioners

$$\mathcal{S}_n(\gamma) = C_n + \gamma/n I_n,$$

and the associated preconditioned matrices

$$\mathcal{P}_n(\gamma) = (\mathcal{S}_n(\gamma)^{-1}) M_n,$$

which will be related later on to the family $\bar{P}_n(\gamma)$.

Lemma 4.1 *For the main method (ρ, σ) of a GBDF of order $p \ge 1$, the functions $\varphi(\theta) = Re(g(e^{i\theta}))$ and $\xi(\theta) = Im(g(e^{i\theta}))$ satisfy:*

(a) $\varphi(\theta) = \begin{cases} O(\theta^{p+2}), & \textit{if } p \textit{ is even}, \\ O(\theta^{p+1}), & \textit{if } p \textit{ is odd}; \end{cases}$

(b) $\xi(\theta) = \begin{cases} \theta + O(\theta^{p+1}), & \textit{if } p \textit{ is even}, \\ \theta + O(\theta^{p+2}), & \textit{if } p \textit{ is odd}. \end{cases}$

Proof The order conditions for the main method of a p order GBDF are:

$$\sum_{j=0}^{k} j^s \alpha_j = s\nu^{s-1}, \qquad s = 0, \ldots, p, \tag{11}$$

where ν is as in (6). For $s = 0, 1, \ldots,$ define the quantities

$$c_s = \sum_{j=0}^{k} (j-\nu)^s \alpha_j .$$

The $p+1$ independent conditions (11) are seen to be equivalent to the following ones:

$$c_0 = 0, \quad c_1 = 1, \quad c_s = 0, \quad s = 2, \ldots, p. \tag{12}$$

Indeed, by direct comparison, $c_0 = 0$ and $c_1 = 1$ are equivalent to (11) for $s = 0, 1$. Consider now $s \in \{2, \ldots, p\}$. We have

$$\begin{aligned} c_s =& \sum_{j=0}^{k} (j-\nu)^s \alpha_j = \sum_{j=0}^{k} \alpha_j \sum_{t=0}^{s} (-1)^{s-t} \binom{s}{t} j^t \nu^{s-t} = \sum_{t=0}^{s} (-1)^{s-t} \nu^{s-t} \binom{s}{t} \sum_{j=0}^{k} j^t \alpha_j \\ =& \sum_{t=0}^{s} (-1)^{s-t} \nu^{s-t} \binom{s}{t} t\nu^{t-1} = \sum_{t=0}^{s} (-1)^{s-t} \binom{s}{t} t\nu^{s-1} = \nu^{s-1} \sum_{t=1}^{s} (-1)^{s-t} \binom{s}{t} t. \end{aligned}$$

Exploiting the equality

$$\binom{s}{t} t = \binom{s-1}{t-1} s,$$

it follows that

$$c_s = s\nu^{s-1} \sum_{t=1}^{s} (-1)^{s-t} \binom{s-1}{t-1} = s\nu^{s-1} \sum_{t=0}^{s} (-1)^{s-t-1} \binom{s-1}{t} = s\nu^{s-1} (1-1)^{s-1} = 0.$$

The assertion follows considering that the Taylor expansion of $\varphi(\theta)$ and $\xi(\theta)$ in a neighborhood of zero are respectively

$$\varphi(\theta) = \sum_{j=0}^{k} \alpha_j \cos(j-\nu)\theta = \sum_{j=0}^{k} \alpha_j \sum_{n=0}^{\infty} (-1)^n \frac{(j-\nu)^{2n}}{(2n)!} \theta^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} c_{2n} \theta^{2n},$$

and

$$\xi(\theta) = \sum_{j=0}^{k} \alpha_j \sin(j-\nu)\theta = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} c_{2n+1} \theta^{2n+1}.$$

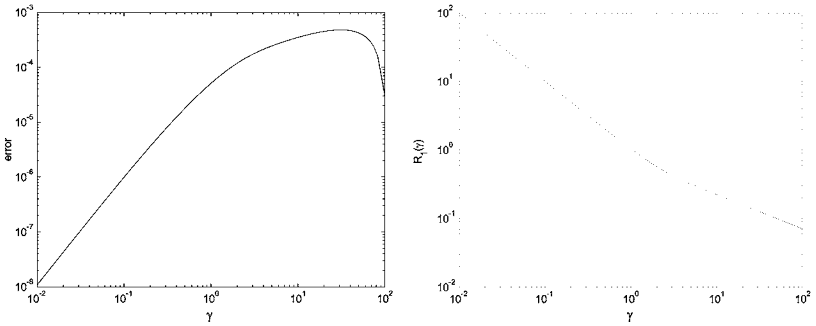

Figure 4.1. Error in the estimation (14) (left), and a plot of the function $R(\gamma)$ (right).

Lemma 4.2 *Consider the family of circulant matrices* $\mathcal{S}_n(\gamma) = C_n + \gamma/nI_n$*, with* γ *a positive parameter and denote by* d_j *the eigenvalues of* $\mathcal{S}_n(\gamma)$ *defined as*

$$d_j = \frac{\gamma}{n} + g\big(e^{\frac{2\pi i}{n}j}\big), \qquad j = 0, \dots, n-1. \tag{13}$$

For n *sufficiently large, the following estimation holds true:*

$$\frac{1}{n^2}\sum_{j=0}^{n}\frac{1}{|d_j|^2} \simeq \frac{i}{2\pi\gamma}\left[\Psi\Big(-\frac{i}{2\pi}\,\gamma\Big) - \Psi\Big(\frac{i}{2\pi}\,\gamma\Big)\right] - \frac{1}{\gamma^2}, \tag{14}$$

where Ψ *is the digamma function* (i *is the imaginary unit*)*.*

Proof The increase of the dimension n reduces the shift of the boundary locus of the method (ρ, σ) of a term γ/n and, as Figures 3.1, 3.2 and formula (13) suggest, gives rise to the accumulation of a number of eigenvalues, proportional to n, into a neighborhood of the origin. All of these eigenvalues (say $\pm d_i$, $i = 1, \dots, c(n)$, by the symmetry of the distribution), will provide the significative contribution to the sum in the left hand side of (14) and therefore neglecting the remaining terms will not produce a consistent error. The neighborhood may be chosen so that in the expressions (a) and (b) of Lemma 4.1 we can also neglect the higher order terms. Under these assumptions we have

$$\begin{aligned} \sum_{j=0}^{n}\frac{1}{|d_j|^2} &= \sum_{j=0}^{n}\frac{1}{\big(\frac{\gamma}{n} + \varphi\big(\frac{2\pi}{n}j\big)\big)^2 + \xi^2\big(\frac{2\pi}{n}j\big)} \\ &\simeq 2\sum_{j=0}^{c(n)}\frac{1}{\big(\frac{\gamma}{n}\big)^2 + \big(\frac{2\pi}{n}\big)^2 j^2} - \frac{n^2}{\gamma^2} \le 2n^2\sum_{j=0}^{\infty}\frac{1}{\gamma^2 + \big(2\pi\big)^2 j^2} - \frac{n^2}{\gamma^2}. \end{aligned} \tag{15}$$

The assertion follows noting that the last series converges to half the first term in the right hand side of (14).

To check for the reliability of the estimation (14) we report in Figure 4.1 the relative error of the computed values that the expressions in its left and right hand side assume in a wide range of values of γ of interest.

A plot of the function

$$R(\gamma) \equiv \left(\frac{i}{2\pi\gamma}\left[\Psi\Big(-\frac{i}{2\pi}\gamma\Big) - \Psi\Big(\frac{i}{2\pi}\gamma\Big)\right] - \frac{1}{\gamma^2}\right)^{1/2}$$

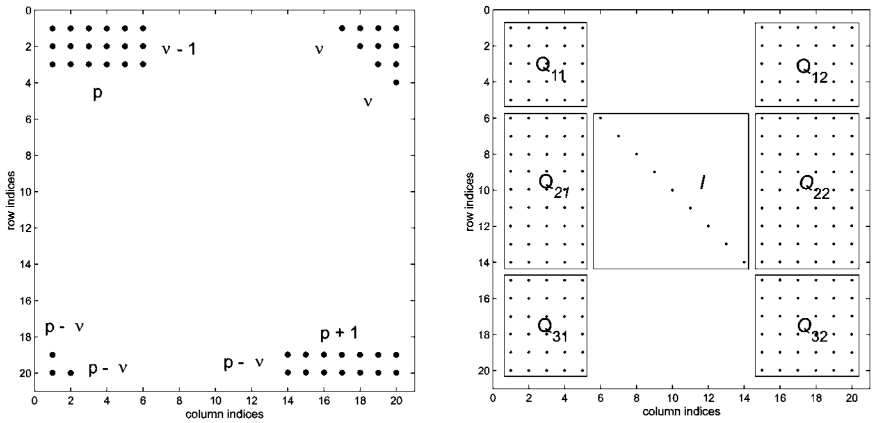

Figure 4.2. Structures of the error matrix E_n (left) and of a matrix $Q \in \mathcal{H}$ (right).

is also reported. This function is strictly decreasing for $\gamma \geq 0$ and its range is $(0, \infty)$; furthermore the principal part in its Laurent expansion is $1/\gamma$. It will be used in the sequel to obtain the main result of this approach.

In the proof of the main result, expressed by Theorem 4.1, the structure of the error matrix $E_n = A_n - C_n$ plays an important role. For a GBDF of order p, E_n has rank p and its nonzero elements are located in the four corners as sketched in Figure 4.2 for the case $p = 6$ and $n = 20$. It is easy to realize that the 2-norm of E_n remains constant for $n \geq 2p+1$; such constant has been computed and reported in Table 4.1 for the GBDFs up to the order 9. Multiplication of a square matrix W_n of dimension n by E_n satisfies the property $W_n E_n = W_n^* E_n$, where W_n^* has all zero columns apart from the first ν and the last $p - \nu$ ones that agree with those of W_n. The asterisk upon a generic square matrix will assume hereafter the same meaning as for W_n^*.

Table 4.1. Norm of the matrix E_n for $p = 1, \ldots, 9$ and $n \geq 2p+1$.

p	1	2	3	4	5	6	7	8	9
$\lVert E_n \rVert$	1	2.62	2.90	4.17	7.59	11.69	19.76	32.51	55.55

For reasons that will be clear in the sequel, we are interested in investigating some properties of the set of square matrices of dimension n

$$\mathcal{H} = \big\{ Q \mid Q = I + W, \ \ W \text{ has zero columns except the first } r \text{ and the last } s \text{ ones} \big\},$$

defined by the integers r and s, $r + s \leq n$. A picture of how an element of $\mathcal{H}$ looks like is in Figure 4.2 in the case $n = 20$, $r = 5$, $s = 6$. It is easy to verify that the product of two matrices in $\mathcal{H}$ belongs to $\mathcal{H}$. This also holds true for the inverse, as the following lemma states.

Lemma 4.3 *The inverse of a nonsingular matrix $Q \in \mathcal{H}$ belongs to $\mathcal{H}$.*

Proof We refer to the partition of Q by means of the blocks Q_{ij} given in Figure 4.2 (to simplify the notation subscripts describing the dimension of the blocks have been omitted), and consider a matrix $H \in \mathcal{H}$:

$$Q = \begin{pmatrix} Q_{11} & 0 & Q_{12} \\ Q_{21} & I & Q_{22} \\ Q_{31} & 0 & Q_{32} \end{pmatrix}_{n\times n}, \qquad H = \begin{pmatrix} H_{11} & 0 & H_{12} \\ H_{21} & I & H_{22} \\ H_{31} & 0 & H_{32} \end{pmatrix}_{n\times n}.$$

The condition $QH = I$ is expressed in terms of the blocks Q_{ij} and H_{ij} by means of the following six equations:

$$\begin{cases} Q_{11}H_{11} + Q_{12}H_{31} = I, \\ Q_{11}H_{12} + Q_{12}H_{32} = 0, \\ Q_{21}H_{11} + H_{21} + Q_{22}H_{31} = 0, \\ Q_{21}H_{12} + H_{22} + Q_{22}H_{32} = 0, \\ Q_{31}H_{11} + Q_{32}H_{31} = 0, \\ Q_{31}H_{12} + Q_{32}H_{32} = I. \end{cases} \tag{16}$$

The first two and the last two equations may be recast as $\tilde{Q}\tilde{H} = I$, where

$$\tilde{Q} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{31} & Q_{32} \end{pmatrix}, \qquad \text{and} \qquad \tilde{H} = \begin{pmatrix} H_{11} & H_{12} \\ H_{31} & H_{32} \end{pmatrix}.$$

Since $\det(\tilde{Q}) = \det(Q)$, from the invertibility of Q we conclude that the blocks H_{11}, H_{12}, H_{31} and H_{32} are uniquely determined from the relation $\tilde{H} = \tilde{Q}^{-1}$. The remaining blocks H_{21} and H_{22}, come from the third and the fourth equations in (16).

We recall (see for example [7]) that a polynomial $p(z) = \sum\limits_{j=0}^{k} z^k$ is of type (s, u, l) (s, u and l are integers such that $k = s + u + l$), if it has s zeros with modulus smaller than 1, u zeros with unit modulus and l zeros with modulus larger than 1.

Lemma 4.4 *Consider the matrix $M_n = A_n + |\lambda|/nI_n$. Constants $\eta > 0$ and $0 < \zeta < 1$ independent of n and λ exist such that the following two statements hold true:*

(a) *The matrix $|M_n^{-1}|$, whose entries are the absolute values of the corresponding ones in M_n^{-1} satisfies the componentwise bound*

$$|M_n^{-1}| \le \eta(I_n + \Omega_n + \Delta_n^T), \tag{17}$$

where

$$\Omega_n = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \ddots & 1 & 0 \end{pmatrix}_{n\times n}, \qquad \Delta_n = \begin{pmatrix} 0 & 0 & \dots & 0 \\ \zeta & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \zeta^{n-1} & \ddots & \zeta & 0 \end{pmatrix}_{n\times n};$$

(b) $\|M_n^{-1}\|_\infty$, $\|M_n^{-1}\|_1$, $\|M_n^{-1}\| \le \eta n$.

Proof (a). The starting point is Theorem 4 in [1] which states the analogous result for Toeplitz matrices. More explicitly, let T_n be the Toeplitz matrix with associated symbol $g(z)+|\lambda|/n$. From A-stability of GBDFs, its characteristic polynomial $p(z) = z^{\nu}(g(z)+|\lambda|/m)$ (of degree k), turns out to be of type $(\nu, 0, k-\nu)$ for $|\lambda| \neq 0$ and of type $(\nu-1, 1, k-\nu)$ if $|\lambda| = 0$. As a result of the above mentioned theorem, a bound, uniform with respect to λ, of the form (17) holds true for T_n, and is attained at $\lambda = 0$. The residual matrix $R_n = M_n - T_n$ differs from the zero matrix in the $(\nu-1)\times p$ upper left block and in the $(p-\nu)\times(p+1)$ lower right one. In terms of T_n and R_n, the matrix M_n^{-1} is recast as

$$M_n^{-1} = (I_n + T_n^{-1}R_n)^{-1}T_n^{-1}. \tag{18}$$

The matrix $Q_n \equiv (I_n + T_n^{-1}R_n)$ belongs to $\mathcal{H}$; using Lemma 4.3, we have also that $H_n \equiv (I_n + T_n^{-1}R_n)^{-1} \in \mathcal{H}$. Hence, considering (18), the assertion will follow if we prove that the entries of H_n are bounded with respect to n and λ. This is true for the matrix Q_n, as a direct consequence of its definition. As concerns H_n, we will see this in the simpler case $Q_{12} = 0$; this does not represent a loss of generality since actually $Q_{12} = O(\sigma^n)$ for some $\sigma \in (0,1)$ (the elements of Q_{12} are in fact combinations of entries of the upper right corresponding block in T_n^{-1}) and a continuity argument may be considered. Exploiting the results presented in [1], it is possible to deduce that the blocks Q_{11}, Q_{31} and Q_{32} essentially remain the same independently of the dimension n (actually they are exponentially convergent as n tends to infinity). Consequently the blocks $H_{11} = Q_{11}^{-1}$, $H_{12} = 0$, $H_{32} = Q_{32}^{-1}$ and $H_{31} = -Q_{32}^{-1}Q_{31}Q_{11}^{-1}$ also have bounded coefficients. Finally, equations three and four in (16) lead to the same conclusions for the coefficients in the blocks H_{21} and H_{22}.

(b). The bound of the norm of the inverse of M_n is a consequence of (17):

$$\begin{gathered} \|M_n^{-1}\|_\infty = \|\,|M_n^{-1}|\,\|_\infty \le \eta n, \\ \|M_n^{-1}\|_1 = \|\,|M_n^{-1}|\,\|_1 \le \eta n, \\ \|M_n^{-1}\| \le \sqrt{\|M_n^{-1}\|_\infty \|M_n^{-1}\|_1} \le \eta n. \end{gathered}$$

A consequence of this lemma is that, as mentioned in Section 3, the matrix M_n is weakly well conditioned, uniformly with respect to λ, that is $\mu(M_n) \le cn$, with $c > 0$ independent of n and λ. This is in general not the case when the Strang preconditioner is used, unless some adjustment is introduced. The following theorem, that reports the main result, is in this direction.

Theorem 4.1 *The conditioning of the preconditioned matrix* $\mathcal{P}_n(\gamma) = \mathcal{S}_n(\gamma)^{-1}M_n$, *with* $\mathcal{S}_n(\gamma) = C_n + \gamma/nI_n$, *satisfies:*

$$\mu(\mathcal{P}_n(\gamma)) \le p^2\eta\|E_n\|^2\ R(\gamma)n, \tag{19}$$

where η *is a positive constant independent of* n *and* γ.

Proof An upper bound for the quantities $\|\mathcal{S}_n(\gamma)^{-1}M_n\|$ and $\|M_n^{-1}\mathcal{S}_n(\gamma)\|$ is derived in the two following steps.

Step 1. From

$$\mathcal{S}_n(\gamma)^{-1}M_n = \mathcal{S}_n(\gamma)^{-1}\left(\mathcal{S}_n(\gamma) + E_n + \frac{|\lambda|-\gamma}{n}\,I_n\right) = \mathcal{S}_n(\gamma)^{-1}E_n + I_n + \frac{|\lambda|-\gamma}{n}\,\mathcal{S}_n(\gamma)^{-1},$$

we deduce

$$\|\mathcal{S}_n(\gamma)^{-1}M_n\| \le \|\mathcal{S}_n(\gamma)^{-1}E_n\| + \left\| I_n + \frac{|\lambda|-\gamma}{n}\,\mathcal{S}_n(\gamma)^{-1}\right\|.$$

We separately analyse the two terms in the right hand side. Introducing the decomposition $C_n = V_n D_n V_n^H$ in the first one yields

$$\begin{aligned}
\|\mathcal{S}_n(\gamma)^{-1}E_n\| &\le \|(C_n + \tfrac{\gamma}{n}\,I_n)^{-1}E_n\| = \|V_n(D_n + \tfrac{\gamma}{n}\,I_n)^{-1}V_n^H E_n\| \\
&= \|(D_n + \tfrac{\gamma}{n}\,I_n)^{-1}(V_n^H)^* E_n\| \le \|E_n\|\,\|(D_n + \tfrac{\gamma}{n}\,I_n)^{-1}(V_n^H)^*\| \\
&= \|E_n\| \max_{\|\mathbf{y}\|=1} \|(D_n + \tfrac{\gamma}{n}\,I_n)^{-1}(V_n^H)^*\mathbf{y}\| \\
&= \|E_n\| \max_{\|\mathbf{y}\|=1} \|y_1\mathbf{z}_1 + \dots + y_\nu\mathbf{z}_\nu + y_{n-p+\nu+1}\mathbf{z}_{n-p+\nu+1} + \dots + y_n\mathbf{z}_n\| \\
&\le p\|E_n\| \max\{\|\mathbf{z}_1\|, \dots, \|\mathbf{z}_\nu\|, \|\mathbf{z}_{n-p+\nu+1}\|, \dots, \|\mathbf{z}_n\|\},
\end{aligned}$$

where, $\mathbf{z}_i$, $i \in \{1, \dots, \nu, n-p+\nu+1, \dots, n\}$ are the nonzero columns of $(D_n + \frac{\gamma}{n}\,I_n)^{-1}(V_n^H)^*$. From (5) we deduce that these columns have constant norm

$$\|\mathbf{z}_i\| = \frac{1}{\sqrt{n}}\left(\sum_{j=0}^{n}\frac{1}{|d_j|^2}\right)^{1/2}.$$

Hence Lemma 4.2 leads to

$$\|\mathcal{S}_n(\gamma)^{-1}E_n\| \le p\|E_n\|R(\gamma)\sqrt{n}.$$

For the second term we have

$$\left\| I_n + \frac{|\lambda|-\gamma}{n}\mathcal{S}_n(\gamma)^{-1}\right\| \le 1 + \frac{||\lambda|-\gamma|}{n}\,\|\mathcal{S}_n(\gamma)^{-1}\| = 1 + \frac{||\lambda|-\gamma|}{\gamma}. \tag{20}$$

Since this term is bounded with respect to n, and γ is a fixed positive constant, for large n we can assume

$$\|\mathcal{S}_n(\gamma)^{-1}M_n\| \le p\|E_n\|R(\gamma)\sqrt{n}. \tag{21}$$

Step 2. Observe that

$$M_n^{-1}\mathcal{S}_n(\gamma) = -M_n^{-1}E_n + I_n + \frac{\gamma-|\lambda|}{n}\,M_n^{-1},$$

and consequently

$$\|M_n^{-1}\mathcal{S}_n(\gamma)\| \le \|M_n^{-1}E_n\| + \left\| I_n + \frac{\gamma-|\lambda|}{n}\,M_n^{-1}\right\|. \tag{22}$$

Proceeding analogously as in step 1, we obtain

$$\begin{aligned}
\|M_n^{-1}E_n\| &= \|(M_n^{-1})^*E_n\| \le \|(M_n^{-1})^*\|\,\|E_n\| \\
&\le \|E_n\| \max\{\|\mathbf{w}_1\|, \dots, \|\mathbf{w}_\nu\|, \|\mathbf{w}_{n-p+\nu+1}\|, \dots, \|\mathbf{w}_n\|\},
\end{aligned}$$

where now $\mathbf{w}_i$ are the non-null columns of $(M_n^{-1})^*$. Exploiting point (a) of Lemma 4.4, we deduce that

$$\|M_n^{-1}E_n\| \le p\eta\|E_n\|\sqrt{n}.$$

Point (b) of Lemma 4.4 is invoked to state that

$$\left\| I_n + \frac{\gamma - |\lambda|}{n} M_n^{-1} \right\| \le 1 + |\gamma - |\lambda||\eta, \tag{23}$$

thus for large n we can write

$$\|M_n^{-1}\mathcal{S}_n(\gamma)\| \le p\eta\|E_n\|\sqrt{n}. \tag{24}$$

The bound (19) is finally derived by combining formulae (21) and (24).

Two applications of this result are now discussed in points A_1 and A_2; they may be considered as corollaries of Theorem 4.1 corresponding to two different choices of the parameter γ.

A_1. The Strang preconditioner is obtained choosing $\gamma = |\lambda|$ and the associated preconditioned matrix P_n, defined in Section 2, is consequently $P_n = \mathcal{P}_n(|\lambda|)$ (observe that in this case the bounds (20) and (23) become independent of γ and η). Considering (19) and the behaviour of the function $R(|\lambda|)$ we conclude that this preconditioner performs well when $|\lambda|$ is far off zero. On the contrary, for small values of $|\lambda|$ we have $R(|\lambda|) \simeq 1/|\lambda|$ and the right hand side of (19) reduces to $O(1/|\lambda|)$ showing that, as actually happens in the applications, the conditioning of P_n may arbitrarily increase.

A_2. Consider now the family of preconditioners $\bar{S}_n(\gamma) = \mathcal{S}_n(|\lambda| + \gamma)$, $\gamma \ge 0$; the corresponding preconditioned matrices are $\bar{P}_n(\gamma) = \mathcal{P}_n(|\lambda| + \gamma)$. The choice $\gamma = 0$ leads back to the case reported in A_1. We are rather interested in comparing the conditioning numbers of the two matrices $\bar{P}_n(\gamma)$ and M_n and in particular to solve the inequality (10) which, considering (19) is certainly fulfilled for all values of γ that satisfy

$$R(\gamma + |\lambda|) \le \frac{c}{p^2\eta\|E_n\|^2}\,\frac{\mu(M_n)}{n}. \tag{25}$$

Taking into account that $\mu(M_n) = O(n)$ and that $R(\gamma)$ strictly decreases to zero as γ tends to infinity, we see that (25) has solutions $\gamma \in [\bar{\gamma}(c, |\lambda|), \infty)$ for some $\bar{\gamma}(c, |\lambda|) > 0$.

The approach presented in A_2 proves that a control of the conditioning during the preconditioning procedure is in principle possible, but two question about the setting up of the technique must be addressed:

(i) the dependence of $\bar{\gamma}(c, |\lambda|)$ on $|\lambda|$ should be removed because for more general problems of the form (4), it requires information about the location of the eigenvalues of J;

(ii) the determination of $\bar{\gamma}(c, |\lambda|)$ is impracticable, unless the quantities η and $\mu(M_n)/n$ are estimated in some way.

The problem pointed out in (i) is easily overcome restricting the analysis to the case $\lambda = 0$ and extending, by continuity, the results to the interval $|\lambda| \in (-\epsilon, 0)$ for some positive ϵ sufficiently small. As seen above, the choice of the Strang preconditioner represents the worst possible case when the conditioning of the problem is considered. Of particular interest is therefore the number $\gamma^*(c) = \bar{\gamma}(c, 0)$, which is the solution of the equation

$$R(\gamma) = \frac{c}{p^2\eta\|E_n\|^2}\,\frac{\mu(A_n)}{n}. \tag{26}$$

The question raised in point (ii) is conveniently solved as follows. Instead of searching approximations of η and $\mu(M_n)/n$, we go back to step 2 of Theorem 4.1. From (22), where now $M_n = A_n$, we can assume

$$\|A_n^{-1}\mathcal{S}_n(\gamma)\| \le \|A_n^{-1}E_n\|,$$

and therefore, without going into the inspection of this last term, we can simply conclude that

$$\mu(\mathcal{P}_n(\gamma)) \le p\|E_n\|R(\gamma)\|A_n^{-1}E_n\|n.$$

Now observe that the quantity

$$\chi(p) = \frac{\sqrt{n}\|A_n^{-1}E_n\|}{\mu(A_n)}, \tag{27}$$

only depends on the particular GBDF used (namely on p) and therefore may be estimated and tabulated (see Table 4.2).

Table 4.2. Values of $\chi(p)$ for $p = 1, \dots, 9$.

p	1	2	3	4	5	6	7	8	9
$\chi(p) \simeq$	0.79	2.3	1.7	2.4	1.8	2.8	1.3	1.4	0.50

On the basis of these considerations, equation (26) may be replaced by the following one

$$R(\gamma) = \frac{c}{p\chi(p)\|E_n\|}, \tag{28}$$

where now all the quantities displayed in the right hand side are available. It is useful to notice that $\gamma^*(c)$, as solution of (28), turns out to be decreasing with respect to c (see also Figure 4.1).

The numbers c and $\gamma^*(c)$, which are in a one to one correspondence, are related respectively to conditioning and preconditioning properties of our problem and a pertinent question is what the best choice of c (or equivalently $\gamma^*(c)$) should be, namely how to define our *optimal* preconditioner.

If the preconditioner $\bar{S}_n(\gamma)$ is demanded to optimize the clustering rate of the eigenvalues of the preconditioned matrix $\bar{P}_n(\gamma)$, then $\gamma^*(c) = 0$ (the Strang preconditioner) would be the best choice because $\bar{P}_n(0)$ has almost all of its eigenvalues exactly centered in 1. However $\gamma^*(c) = 0$, gives also $c = \infty$ and the control of the conditioning over the preconditioned matrix is completely loss.

On the other hand, from the point of view of the conditioning, the best choice would be $c = 1$, because if so, the conditioning of the preconditioned system would not become worse than that of the original one. In such a case however the cluster around 1 becomes wide and a slowing down in the convergence speed is noticed.

However if, as it should be, the optimality is linked to the property of the preconditioner of minimizing the algorithm cost, we deduce that we cannot define optimal neither the Strang preconditioner nor the preconditioner $\bar{S}(\gamma^*(1))$; the best choice is rather obtained for an intermediate value of $\gamma \in (0, \gamma^*(1))$. We experienced that choosing $c = 10^j$, with small value of the integer j, provides the right compromise. In the example of Section 2 for example we chose $\gamma^*(c) = 1$ and, considering Tables 4.1 and 4.2, the corresponding value of c is

$$c = 5 \cdot \chi(5) \cdot R(1) \cdot 7.59 \simeq 71,$$

which is the maximum amplification factor of $\mu(M_n)$ for small values of the eigenvalues. Indeed from Table 2.1, we get

$$\mu(M_n) \simeq 3.4 \cdot 10^3, \quad c \cdot \mu(M_n) \simeq 2.4 \cdot 10^5, \quad \mu(\bar{P}_n) = 1.8 \cdot 10^5,$$

in agreement with the obtained results of the test problem in Section 2.

5 The Strang Preconditioner on a Modified GBDF

The ill conditioning of the Strang preconditioner for $\lambda = 0$, is generated by the consistency condition $\sum\limits_{i=0}^{k} \alpha_i = 0$. In the previous section we overcame the problem introducing a modification in the preconditioner. An alternative is to modify the coefficients that define the method. In details we consider hereafter the approach used in [10] to deduce the global contractivity of GBDFs. In that paper, the authors proved that

$$\min_{-\pi \le \theta < \pi} Re\left(\frac{\rho(e^{\frac{1}{n}+i\theta})}{\sigma(e^{\frac{1}{n}+i\theta})}\right) \ge \frac{s}{n}, \tag{29}$$

with s a positive constant independent of n (see Lemma 1.2 and Theorem 5.1). The modified symbol $\hat{g}(z) = g(e^{1/n}z)$ is generated by the matrix $\widehat{A}_n$ defined by a similarity transformation of A_n:

$$\widehat{A}_n = L_n A_n (L_n)^{-1}, \qquad L_n = \begin{pmatrix} e^{-\frac{1}{n}} & & & \\ & e^{-\frac{2}{n}} & & \\ & & \ddots & \\ & & & e^{-1} \end{pmatrix}_{n \times n}.$$

In details, the matrix L_n operates as follows: the original linear system (4) for a GBDF ($B_n = I_n$), is equivalent to

$$(L_n \otimes I_m)(A_n \otimes I_m - hI_n \otimes J)(L_n \otimes I_m)^{-1}(L_n \otimes I_m)(Y^{k+1} - Y^k) = (L_n \otimes I_m)G(Y^k). \tag{30}$$

Introducing the change of variables $Z^k = (L_n \otimes I_m)Y^k$, and considering that

$$(L_n \otimes I_m)(A_n \otimes I_m)(L_n \otimes I_m)^{-1} = (L_n A_n (L_n)^{-1}) \otimes I_m = \widehat{A} \otimes I_m$$

and

$$(L_n \otimes I_m)(I_n \otimes J)(L_n \otimes I_m)^{-1} = (L_n I_n (L_n)^{-1}) \otimes J = I_n \otimes J,$$

the equation (30) becomes

$$(\widehat{A}_n \otimes I_m - h I_n \otimes I_m)(Z^{k+1} - Z^k) = (L_n \otimes I_m) G((L_n \otimes I_m)^{-1} Z^k).$$

The matrix $\widehat{M_n} \equiv (\widehat{A}_n \otimes I_m - h I_n \otimes I_m)$ is therefore similar to M_n via the similarity transformation $(L_n \otimes I_m)$. Taking into account that $\mu(L_n) < e$, the conditioning of $\widehat{M_n}$ is close to that of M_n. However, contrary to what happens for M_n, the Strang circulant preconditioner $\widehat{S}_n$ associated to $\widehat{M_n}$ preserves a good conditioning. The lower bound (29) states in fact that even if $h \det(J) = 0$, $\widehat{S}_n$ is nonsingular and weakly well conditioned. As for the modified Strang preconditioner, we are now interested in studying the conditioning of the preconditioned matrix $\widehat{P}_n = \widehat{S}_n^{-1} \widehat{M_n}$ in the scalar case $(J = -|\lambda|)$. The novelty that will make $\mu(\widehat{P}_n)$ independent of $|\lambda|$ is expressed in the following lemma that is the analogue of Lemma 4.1 for the function $\hat{g}(z)$.

Lemma 5.1 *For the modified GBDF of order $p \geq 1$, the functions $\hat{\varphi}(\rho,\theta) = \mathrm{Re}(g(e^{\rho+i\theta}))$ and $\hat{\xi}(\rho,\theta) = \mathrm{Im}(g(e^{\rho+i\theta}))$, $\rho, \theta \in \mathbb{R}$, satisfy in a neighborhood of the origin:*

(a) $\hat{\varphi}(\rho,\theta) = \rho +$ *higher order terms;*

(b) $\hat{\xi}(\rho,\theta) = \theta +$ *higher order terms.*

Proof The Taylor expansion of $\hat{\varphi}(\rho,\theta)$ and $\hat{\xi}(\rho,\theta)$ about $(0,0)$ are respectively

$$\begin{aligned}\hat{\varphi}(\rho,\theta) &= \sum_{j=0}^{k} \alpha_j e^{(j-\nu)\rho} \cos(j-\nu)\theta = \sum_{j=0}^{k} \alpha_j \sum_{s=0}^{\infty} \frac{(j-\nu)^s}{s!} \rho^s \sum_{n=0}^{\infty} (-1)^n \frac{(j-\nu)^{2n}}{(2n)!} \theta^{2n} \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} \sum_{s=0}^{\infty} \frac{1}{s!} c_{2n+s} \theta^{2n} \rho^s,\end{aligned}$$

and

$$\hat{\xi}(\rho,\theta) = \sum_{j=0}^{k} \alpha_j e^{(j-\nu)\rho} \sin(j-\nu)\theta = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} \sum_{s=0}^{\infty} \frac{1}{s!} c_{2n+s+1} \theta^{2n+1} \rho^s,$$

where the coefficients c_j were defined in Lemma 4.1. From this expressions and (12) we get the assertion.

Define now $\widehat{C}_n$ as the Strang preconditioner of $\widehat{A}_n$. The proof of Lemma 4.2 remains the same for the circulant matrix $\widehat{S}_n = \widehat{C}_n + |\lambda|/n I_n$ except that, due to (a) of the previous lemma, the new term $\hat{\varphi}(1/n, 2\pi j/n)$ can not longer be neglected as $\varphi(2\pi j/n)$ in (15), rather it must be replaced by $1/n$. As a consequence, the result expressed by Lemma 4.2 holds as well in this case provided that γ is replaced by $1 + |\lambda|$. To conclude, observing that $\|\widehat{M_n}^{-1}\| \leq e \|M_n^{-1}\|$ and denoting by $\widehat{E}_n = \widehat{M_n} - \widehat{S}_n$, we can reformulate Theorem 4.1 as follows.

Theorem 5.1 *The conditioning of the preconditioned matrix* $\widehat{P}_n = \widehat{S}_n^{-1}\widehat{M}_n$ *satisfies:*

$$\mu(\widehat{P}_n) \leq ep^2\eta\|\widehat{E}_n\|^2 \; R(1+|\lambda|)n, \tag{31}$$

where η *is a positive constant independent of* n *and* $|\lambda|$*.*

The bound (31) proves that $\widehat{P}_n$ is weakly well conditioned and non increasing for $\lambda \simeq 0$ (in particular for $|\lambda| = 0$ we get $R(1) \simeq 1.08$).

Acknowledgements

The authors wish to acknowledge P. Amodio and F. Mazzia for their very helpful suggestions that improved the paper.

References

[1] Amodio, P. and Brugnano, L. The conditioning of Toeplitz band matrices. *Math. Comput. Modelling* **23**(10) (1996) 29–42.

[2] Amodio, P. and Brugnano, L. ParalleloGAM: a parallel code for ODEs. *Appl. Numer. Math.* **28** (1998) 95–106.

[3] Amodio, P. and Mazzia, F. Numerical solution of differential algebraic equations and computation of consistent initial/boundary conditions. *J. Comput. Appl. Math.* **87**(1) (1997) 135–146.

[4] Beam, R.M. and Warming, R.F. The asymptotic spectra of Banded Toeplitz and quasi-Toeplitz matrices. *SIAM J. Sci. Comput.* **14** (1993) 971–1006.

[5] Bertaccini, D. P-circulant preconditioners and the systems of ODE codes. In: *Iterative Methods in Scientific computation II, IMACS Series in Computational and Applied Mathematics.* (Eds.: D.R. Kincaid and A.C. Elster), IMACS, New Brunswick, NJ, 1999, 179–193.

[6] Bertaccini, D. A circulant preconditioner for the systems of LMF-based ODE codes. *SIAM J. Sci. Comput.*, (to appear).

[7] Brugnano, L. and Trigiante, D. *Solving ODEs by Linear Multistep Initial and Boundary Value Methods.* Gordon and Breach Sciences Publishers, Amsterdam, 1998.

[8] Chan, R.H., Ng, M.K. and Jin, X.Q. Circulant preconditioners for solving ordinary differential equations. In: *Structured Matrices: Recent Developments in Theory and Computation.* (Eds.: D. Bini, E. Tyrtyshnikov and P. Yalamov), Nova Science Pub. Inc., 2000.

[9] Davis, P.J. *Circulant Matrices.* John Wiley & Sons, New York, 1979.

[10] Iavernaro, F. and Mazzia, F. Convergence and stability of multistep methods solving nonlinear initial value problems. *SIAM J. Sci. Comput.* **18** (1997) 270–285.

[11] Iavernaro, F. and Mazzia, F. Solving ordinary differential equations by block Boundary Value Methods: properties and implementation techniques. *Appl. Num. Math.* **28**(2-4) (1998) 107–126.

[12] Iavernaro, F. and Mazzia, F. Block-boundary value methods for the solution of ordinary differential equations. *Siam J. Sci. Comput.* **21** (1999) 323–339.

[13] Mazzia, A., Mazzia, F. and Trigiante, D. Boundary value methods for PDEs. In: *Proceedings of the First International Conference on Nonlinear Problems in Aviation and Aerospace (Daytona Beach, FL, 1996).* (Ed.: S. Sivasundaram), Embry-Riddle Aeronaut. Univ. Press, Daytona Beach, FL, 1996, 421–436.

[14] Saad, Y. *Iterative Methods for Sparse Linear Systems.* PWS publishing company, Boston, MA, 1995.

[15] Strang, G. A proposal for Toeplitz matrix calculations. *Stud. Appl. Math.* **74** (1986) 171–176.

[16] Strela, V.V. and Tyrtyshnikov, E.E. Which circulant preconditioner is better? *Math. Comp. Vol.* **65** (1996) 137–150.

[17] Trigiante, D. Multipoint methods for linear Hamiltonian systems. In: *Advances in Nonlinear Dynamics.* (Eds.: S. Sivasundaram and A.A. Martynyuk), Series "Stability and Control: Theory Methods and Applications", Gordon and Breach Sciences Publishers, Reading, UK, 1997, 335–348.

Nonlinear Dynamics and Systems Theory, 1(1) (2001) 81–96

Stability of Stationary Motions of Mehcanical Syatems with a Rigid Body as the Basic Element

A.M. Kovalev

Institute of Applied Mathematics and Mechanics of National Academy of Sciences of Ukraine, Donetsk, Ukraine

Received: November 1, 1999; Revised: May 15, 2000

Abstract: The paper presents a brief description of the problem on the permanent rotations of a rigid body from its statement moment up to factual completion. Stability theory of stationary motions connected with this problem is stated. Their interconnection is shown and the closest generalizations have been considered.

Keywords: *Rigid body systems; stationary motion; stability.*

Mathematics Subject Classification (2000): 70E05, 70E15, 70K20.

1 Introduction

The problem of the permanent rotations of a rigid body with a fixed point in the gravity force field occupies an important place in analytical mechanics and different applications. In rigid body dynamics a complete investigation of the permanent rotations has been made by Staude [1]. This remarkable paper by Staude practically closed this problem unfortunately because specialists on rigid body dynamics lost interest in further investigation of the permanent rotations for many years. However, this problem attracted the attention of the experts in the stability theory in connection with the study of stability of stationary motions of mechanical systems and has played an important role in the development of this problem. Stability problems on the stationary motions of mechanical systems and on the permanent rotations of a rigid body are closely connected, their interinfluence defining their joint development in many respects. The formation of these problems was connected with the Routh theorem [2] and Majevskii criterion [3]. Their systematic investigation started with the appearance of the Chetaev method [4] and Rumyantsev's paper [5] having provided a suitable mathematical apparatus and having defined the direction of research. The introduction in the research domain of the problem of gyrostat motion [6, 7], other new objects [8] and force fields [9] raised the interest in the problem and defined the period of its intensive development. At this

© 2001 Informath Publishing Group. All rights reserved.

time KAM-theory had an essential influence on it and led to the extension of the use of Hamiltonian mechanics methods and the raising of the meaning of necessary conditions. The use of Kolmogorov's idea [10] permits us to characterize stability domains in the phase and parameter space as the domains of the fulfilment of the necessary conditions from which only some subdomains of smaller dimension can be excluded. On this base it is possible to say about practical end of the problems for which the investigation of necessary stability conditions is fulfilled, what is really done for many problems. The registration of this approach signifies the end of an intensive development period of the permanent rotations stability problem. The modern stage is characterized by the study of new objects such as multibody systems with different kinds of joints, a body on a string and others; by the search for new effects and by the movement of interest from stability theory into attractor theory, chaos and other modern topics of dynamical systems theory.

In the presented paper the state of stability theory of stationary motions of mechanical systems, the stability problem of permanent rotations of a rigid body and its generalizations are described. The presentation is in the main based on the results obtained by the Donetsk school of mechanics where these problems were studied the most widely and completely.

2 Objects and Motions

The problem of a motion of a rigid body with a fixed point in a gravity force field occupies the central place in rigid body dynamics. For its study different forms of the motion equations are offered, from which we choose the best-known Euler-Poisson equations

$$A_1\dot{\omega}_1 = (A_2 - A_3)\omega_2\omega_3 + \Gamma(e_2\nu_3 - e_3\nu_2) \quad (123), \tag{1}$$

$$\dot{\nu}_1 = \nu_2\omega_3 - \nu_3\omega_2 \quad (123), \tag{2}$$

where $\omega_2, \omega_2, \omega_3$; ν_1, ν_2, ν_3; e_1, e_2, e_3 are, respectively, projections on the moving axes of an angular velocity, a vertical unit vector and a unit vector leading from the fixed point in the direction of the mass center of a body; A_1, A_2, A_3 are principal moments of inertia; Γ is the product of the body weight and the distance from the fixed point to the mass center; (123) is a symbol of cyclic index permutation.

Equations (1) and (2) allow the integrals

$$\begin{gathered} A_1\omega_1^2 + A_2\omega_2^2 + A_3\omega_3^2 - 2\Gamma(e_1\nu_1 + e_2\nu_2 + e_3\nu_3) = h, \\ A_1\omega_1\nu_1 + A_2\omega_2\nu_2 + A_3\omega_3\nu_3 = k, \\ \nu_1^2 + \nu_2^2 + \nu_3^2 = 1. \end{gathered} \tag{3}$$

Gyrostat. The necessity of accounting for the influence of interior masses motions on the Earth's motion led Volterra [11] to the creation of a new mechanical object named a gyrostat. At the present time by the term gyrostat we understand a rigid body having cavities with liquid performing a potential motion [8] or a body carrying fly-wheels rotating in a definite way [7]: on inertia or with constant relative velocity. Let's write the motion equations and integrals of a gyrostat with fixed point in a gravity force field

$$A_1\dot{\omega}_1 = (A_2 - A_3)\omega_2\omega_3 + \lambda_2\omega_3 - \lambda_3\omega_2 + \Gamma(e_2\nu_3 - e_3\nu_2),$$
$$\dot{\nu}_1 = \nu_2\omega_3 - \nu_3\omega_2 \quad (123), \tag{4}$$

$$A_1\omega_1^2 + A_2\omega_2^2 + A_3\omega_3^2 - 2\Gamma(e_1\nu_1 + e_2\nu_2 + e_3\nu_3) = h,$$
$$(A_1\omega_1 + \lambda_1)\nu_1 + (A_2\omega_2 + \lambda_2)\nu_2 + (A_3\omega_3 + \lambda_3)\nu_3 = k, \tag{5}$$
$$\nu_1^2 + \nu_2^2 + \nu_3^2 = 1.$$

Here in addition to the notations introduced under λ_1, λ_2, λ_3 the projections of gyrostatic moment vector on the moving axes are designated.

A Rigid Body with Vortex Filling. A great number of papers are devoted to the study of the motion of a body with an ellipsoidal cavity completely filled by an ideal uniform incompressible liquid performing the uniform vortex motion. The motion equations of the body-liquid system have the form [8, 12]

$$\dot{\Omega}_1 = (1 - \varepsilon_3)\omega_3\Omega_2 - (1 + \varepsilon_2)\omega_2\Omega_3 + (\varepsilon_2 + \varepsilon_3)\Omega_2\Omega_3,$$
$$\tfrac{d}{dt}(a_1\omega_1 + b_1\Omega_1) = (a_2\omega_2 + b_2\Omega_2)\omega_3 - (a_3\omega_3 + b_3\Omega_3)\omega_2 + \Gamma(e_2\nu_3 - e_3\nu_2), \tag{6}$$
$$\dot{\nu}_1 = \nu_2\omega_3 - \nu_3\omega_2 \quad (123).$$

Here in addition the designations are introduced: Ω_1, Ω_2, Ω_3 are the projections of the vortex vector on the moving axes; a_1, a_2, a_3 are changed inertia moments of the body-liquid system; Γ is the product of the body-liquid system weight and the distance from the mass center to the fixed point divided by $2c^2M/s$; $c^2 = c_1^2 + c_2^2 + c_3^2$; c_1, c_2, c_3 are semiaxes of the cavity-ellipsoid; M is the liquid mass in the cavity;

$$\varepsilon = \frac{c_3^2 - c_2^2}{c_3^2 + c_2^2}, \qquad b_1 = \frac{2c_2^2c_3^2}{c^2(c_2^2 + c_3^2)} \quad (123).$$

Equations (6) allow the integrals

$$\sum_{i=1}^{3}(a_i\omega_i^2 + b_i\Omega_i^2 - 2\Gamma e_i\nu_i) = h, \qquad \sum_{i=1}^{3}(a_i\omega_i + b_i\Omega_i)\nu_i = k, \tag{7}$$
$$\frac{\Omega_1^2}{c_1^2} + \frac{\Omega_2^2}{c_2^2} + \frac{\Omega_3^2}{c_3^2} = m, \qquad \nu_1^2 + \nu_2^2 + \nu_2^2 = 1.$$

Multibody System. The equations of motion of the system of rigid bodies can be obtained in different forms depending on the choice of coordinate systems and main variables. The number of possible forms of equations is increasing because the bodies can be composed in groups in different ways. A form of such equations that are transparent and accessible for further investigation are required. The equations satisfying these

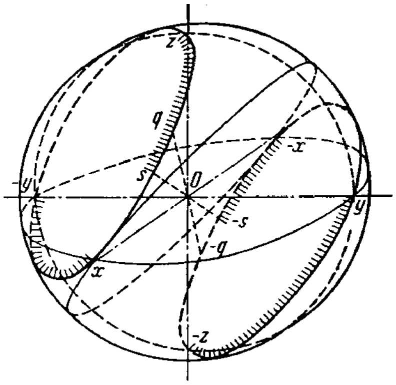

Figure 2.1. Intersection of the Staude cone with the unit sphere.

requirements are offered in [13] for the system of n gyrostats

$$\frac{d}{dt}\left(A^n\omega^n+\lambda^n\right)+m^nc^n\times\left[\sum_{l=1}^{n-1}\frac{d}{dt}\left(\omega^l\times s^l\right)-g\nu\right]=0,$$

$$\frac{d}{dt}\left(A^k\omega^k+\lambda^k\right)+m^kc^k\times\left[\sum_{l=1}^{k-1}\frac{d}{dt}\left(\omega^l\times s^l\right)-g\nu\right]+s^k \tag{8}$$

$$\times\sum_{q=k+1}^{n}m^q\left[\frac{d}{dt}\left(w^q\times c^q\right)+\sum_{l=1}^{q-1}\frac{d}{dt}\left(\omega^l\times s^l\right)-g\nu\right]=0\quad(k=2,3,\ldots,n-1),$$

$$\frac{d}{dt}\left(A^1\omega^1+\lambda^1\right)+s^1\times\sum_{q=2}^{n}m^q\left[\frac{d}{dt}\left(\omega^q\times c^q\right)+\sum_{l=1}^{q-1}\frac{d}{dt}\left(\omega^l\times s^l\right)-g\nu\right]=m^1c^1\times g\nu,$$

$$\dot\nu=\nu\times\omega.$$

Carrier-bodies S_0^k, S_0^{k-1} of gyrostats S^k, S^{k-1} $(k>1)$ have one generic point O^k; where ω^k is the absolute angular velocity of the body S_0^k; vector s^k leads from O^k to the mass center of gyrostat S^k; m^k is the mass of S^k; A^k is the inertia tensor of gyrostat S^k at point O^k; λ^k is the gyrostatic moment of this gyrostat.

The development of equations (8), and their further transformation have played the essential role in the amplification of interest in multibody dynamics and have promoted significant progress in obtaining the exact solutions; their number has more than doubled since the time of their publication.

Permanent Rotations. In rigid body dynamics the most studied stationary motions are permanent rotations which are characterized by the property that the angular velocity vector is constant and leads along the vertical. This provides us with the possibility of representing permanent rotation in the moving (connected with the body) space. The real motion is obtained by the coincidence of the permanent rotation axis with the vertical and rotating the body around the vertical with angular velocity obtained. Under the common values of parameters the set of permanent rotations is one-dimensional and consists of three parts of the curve, lying at the Staude cone (Figure 2.1)

$$(A_2 - A_3)e_1\omega_2\omega_3 + (A_3 - A_1)e_2\omega_3\omega_1 + (A_1 - A_2)e_3\omega_1\omega_2 = 0. \tag{9}$$

Note that in the general situation permanent rotations about principal axes are impossible. They can appear under some conditions on parameters. So they exist in the Lagrange, Kovalevskaya, Hess cases and in the Euler case permanent rotations are admissible only around three principal axes. Note more the singular case: in the Lagrange case the permanent rotations set consists of the principal axis and some surface.

On the whole a similar but more complicated picture exists for a gyrostat [14]. In principle, the picture is different for a body with vortex filling, for which the permanent rotations set forms a solid angle [12]. For systems of the rigid bodies, on the whole, permanent rotations of the Lagrange gyroscopes around principal axes have been considered [12].

3 Stability of Stationary Motions

Two main approaches to the investigation of stationary motions stability have been created. The first is based on the Routh-Lyapunov theorem and Chetaev method, the second – on the Arnold-Moser theorem extended to stationary motions. To form these theorems it is convenient to use Hamilton variables $q_i,\ p_i\ \ (i = 1,\dots,n)$ for the description the motion of a conservative mechanical system with n degrees of freedom. In the presence of cyclic coordinates $q_\alpha\ \ (\alpha = k+1,\dots,n)$ the Hamilton function has the form $H(q_1,\dots,q_k,\, p_1,\dots,p_n)$ and the equations of motion have $n-k$ cyclic integrals $p_\alpha = c_\alpha = \text{const}\ \ (\alpha = k+1,\dots,n)$. The function $H(q_1,\dots,q_k,\, p_1,\dots,p_k,\, c_{k+1},\, c_n)$ defines the system with k degrees of freedom which is called the reduced one.

The stationary motions of the mechanical system are called such motions for which positional coordinates and impulses $q_i,\ p_i\ \ (i = 1,\dots,k)$ and cyclic impulses $p_\alpha\ \ (\alpha = k+1,\dots,n)$ preserve constant values $q_i = q_i^0,\ p_i = p_i^0,\ p_\alpha = c_\alpha$. Constants c_α are arbitrary, values $q_i^0,\, p_i^0$ are obtained from the equations

$$\frac{\partial H}{\partial q_i} = 0, \qquad \frac{\partial H}{\partial p_i} = 0, \qquad i = 1,\dots,k$$

and determine the equilibrium position of the reduced system.

Under stability of stationary motions we understand the stability of these motions with respect to the values $q_i,\ p_i,\ c_\alpha\ \ (i = 1,\dots,k;\ \alpha = k+1,\dots,n)$. The effective tool of investigation of stationary motions stability is the Routh theorem [2] (with the Lyapunov addition [15]) reducing the question about stationary motions stability to the analysis of the extremum of potential energy of the reduced system.

Theorem 3.1 *If potential energy of the reduced system has the minimum both under the given values $p_\alpha = c_\alpha$ responding to the stationary motion considered and under any close to the given values $p_\alpha = c_\alpha + \eta_\alpha$ and also the values q_i inverting it in the minimum are continuous functions of the variables p_α, then stationary motion is stable.*

Different generalizations of Theorem 3.1, its connection with the Chetaev method and application to the investigation of the permanent rotations stability have been considered in the papers [16, 17], where in particular it was pointed out that for the establishment of stationary motions stability it was sufficient to establish by the Lyapunov method the stability of equilibrium position of the reduced system, moreover it was convenient to construct the Lyapunov function by the Chetaev method from the motion integrals.

For Hamiltonian systems in the formulation of Theorem 3.1 the Hamilton function of the reduced system is used instead of the potential energy. Theorem 3.1 does not solve the question about stationary motions stability if the Hamiltonian of the reduced system is not a function of fixed sign in the equilibrium position. For Hamiltonian systems with two-dimensional reduced system, the stability of stationary motions can be obtained with the help of the following theorem [18] extending the known Arnold-Moser theorem [19, 20] to the case of stationary motions.

Theorem 3.2 *Let the Hamiltonian $H(q_1, q_2, p_1, p_2, \dots, p_{2+m})$ be an analytical function of the coordinates and impulses at the point p with the coordinates*

$$q_1 = q_2 = 0, \quad p_1 = p_2 = 0, \quad p_{2+i} = c_i, \quad i = 1, \dots, m \tag{10}$$

defining the stationary motion considered. The Hamiltonian of the reduced system at this point satisfies the following conditions:

1. *Eigenvalues of the linearized reduced system are pure imaginary $\pm i\alpha_1$, $\pm i\alpha_2$.*
2. *For all integers k_1, k_2 satisfying the condition $|k_1| + |k_2| \le 4$ the inequality $k_1\alpha_1 + k_2\alpha_2 \ne 0$ is fulfilled.*
3. *$D = -(\beta_{11}\alpha_2^2 - 2\beta_{12}\alpha_1\alpha_2 + \beta_{22}\alpha_1^2) \ne 0$, where $\beta_{\nu\mu}$ are the coefficients of the fourth order form for the Hamiltonian, transformed into the form*

$$H = \sum_{\nu=1}^{2} \frac{\alpha_\nu}{2} R_\nu + \sum_{\nu,\mu=1}^{2} \frac{\beta_{\nu\mu}}{4} R_\nu R_\mu + O_5, \qquad R_\nu = \xi_\nu^2 + \eta_\nu^2$$

(O_5 is a power series containing the terms of order not less then five).

Then stationary motion (10) is stable.

Condition 1 of this theorem is fulfilled in the domain of fulfilment of necessary stability conditions. In the common situation nonfulfilment of conditions 2 and 3 leads to the exclusion of some sets of lesser dimension from this domain. In the rest domain, which differs little from the domain of the fulfilment of necessary stability conditions, stationary motions are stable. Therefore in the nonsingular case (conditions 2 and 3 don't take identities) it is practically sufficient to study only the necessary stability conditions. For the analysis of singular cases it is possible to apply the theorems on stability of the equilibrium position under the presence of resonances [21] and the vanishing of discriminant D [22] extended to stationary motions.

4 Permanent Rotations of a Rigid Body

One of the first problems solved on the stability of permanent rotations is the problem of the stability of permanent rotations of the Lagrange gyroscope around its principal axis. The conditions of their stability are known as the Majevskii criterion [3] and were obtained during research into projectile motion. Following attempts [23, 24] didn't bring any serious progress in this problem and only the appearance of the Chetaev method gave the possibility of its systematic investigation which was begun in Rumyantsev's paper [5]. Sufficient stability conditions of permanent rotations were obtained in his paper by the construction of the Lyapunov function in the form of the bundle of the integrals of perturbed motion. With their help the stability domains were found both in the common case of mass distribution and in particular cases when the mass center belonged to one of the principal planes or to the principal axis and also when the ellipsoid of inertia was the ellipsoid of rotation. Subsequent investigations can be divided into three groups: the study of permanent rotations in integrable cases; analysis of the permanent rotations around principal axes; research on the general case (permanent rotations around principal axes are impossible). The interest in integrable cases is caused by the fact that the presence of additional integrals permits us to obtain the necessary and sufficient stability conditions by the Lyapunov functions method. Two rest directions are connected with the motions which are of most interest from the theoretical and applied point of views. Let's examine them in more detail.

Permanent Rotations about Principal Axes. Let the mass center belong to the principal axis, then the body can rotate about this axis permanently with arbitrary velocity. The stability of these motions is studied with respect to the variables ω_1, ω_2, ω_3, ν_1, ν_2, ν_3 under the use of equations (1) or with respect to the Euler angles θ, φ and all generalized impulses p_θ, p_φ, p_ψ under the use of Hamiltonian equations. Using the Chetaev method, Rumyantsev [5, 25] obtained sufficient stability conditions of the considered permanent rotations which are equivalent to the conditions of the property of having fixed sign the square part of Hamiltonian H_2. The following investigation of this problem was fulfilled with the help of Theorem 3.2 in paper [26]. Let's look at its main result.

A body motion is described by Hamilton equations in the Euler angles introduced in the usual way. For Hamiltonian to have no singularities on the considered motion we place the mass center on the first principal axis. The following solution corresponds to the permanent rotations studied

$$\theta^0 = \varphi^0 = \frac{\pi}{2}, \quad \psi^0 = \omega_0 t + \psi_0, \quad p^0_\theta = p^0_\varphi = 0, \quad p^0_\psi = A_1\omega,$$

where ω_0 is the angular velocity of the permanent rotation. Introducing the perturbations

$$\theta = \frac{\pi}{2} + y_1', \quad \varphi = \frac{\pi}{2} + y_2', \quad p_\theta = x_1', \quad p_\varphi = x_2'$$

and going over to the dimensionless variables, we obtain the following presentation for the Hamiltonian H

$$H = H_2 + H_4 + O_5, \tag{11}$$

$$2H_2 = ax_1^2 + bx_2^2 + (\omega^2 - e)y_1^2 + [(a-1)\omega^2 - e]y_2^2 + 2(a-1)\omega x_1 y_2 + 2\omega x_2 y_1,$$

$$\begin{aligned} 2H_4 = {} & (1-a)x_1^2 y_2^2 + x_2^2 y_1^2 + \frac{8\omega^2 + e}{12}\, y_1^4 + \frac{4\omega^2(1-a) + e}{12}\, y_2^4 \\ & + \frac{2\omega^2(a-1) + e}{2}\, y_1^2 y_2^2 + \frac{4\omega(1-a)}{3}\, x_1 y_2^3 + \omega(a-1)x_1 y_1^2 y_2 \\ & + \frac{5}{3}\,\omega x_2 y_1^3 + 2\omega(a-1)x_2 y_1 y_2^2 + 2(a-1)x_1 x_2 y_1 y_2, \end{aligned}$$

where

$$e = \begin{cases} 1 & \text{for} \quad \Gamma < 0, \\ -1 & \text{for} \quad \Gamma > 0 \end{cases}.$$

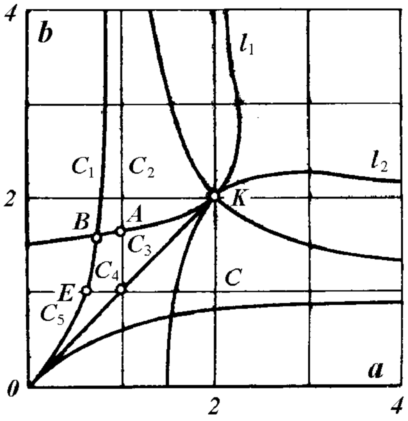

Figure 4.1. Stability domain in the space Oab.

Condition 1 of Theorem 3.2 is fulfilled in the domain G of necessary stability conditions fulfilment which were obtained and analyzed in detail in paper [27]. Condition 2 is reduced to one inequality which is broken under the condition defining the resonance of the third order

$$9\omega^4 + 2(41b - 59)\omega^2 + (9b - 1)(b - 9) = 0. \tag{12}$$

To check the third condition Hamiltonian (11) is transformed to the normal form up to terms of the fourth order inclusively and discriminant D is calculated which has rather simple expression for values $a = 1,\ e = 1$

$$\begin{aligned} D_1 = (b-1)^2\omega^8 + 2(b-1)(b^2+2b-5)\omega^6 + (b-1)(b^3+13b^2 \\ - 41b + 7)\omega^4 + 8(b^4 - 5b^3 + 5b^2 + b + 2)\omega^2 + 4b(b-1)^2. \end{aligned} \tag{13}$$

A conclusion about the stability of permanent rotations is obtained by the application of Theorem 3.2. For the descriptive representation of the obtained results there has been introduced an extended parametric space being the direct product of the space of mechanical system parameters and cyclic constants space, in this case the plane $Ob\omega$. The equation (12) and $D_1 = 0$ determine in the plane $Ob\omega$ resonance curve s_1 and discriminant curve s_2 (Figure 4.1). The theorem is true.

Theorem 4.1 *Let a rigid body having equal inertia moments about two first axes* $(a = 1)$ *be rotated permanently about the first axes carrying the mass center situated higher than fixed point* $(e = 1)$. *Then in extended parametric space-plane* $Ob\omega$*-stability domain represents domain* G_1 *of fulfilment of necessary stability conditions from which curves* s_1, s_2 *are excluded (Figure 4.1).*

Returning to the common case we note from the formulas (12) and (13) that in the space $Oab\omega$ conditions 2 and 3 of Theorem 3.2 are not fulfilled on resonance and discriminant surfaces S_1 and S_2; that is the following result occurs.

Theorem 4.2 *Let a rigid body be rotated permanently around the principal axis carrying the mass center. Then in the extended parametric space* $Oab\omega$ *stability domain represents the domain* G *from which surfaces* S_1 *and* S_2 *are excluded.*

It should be noted that the permanent rotations corresponding to resonance curve s_1 have been studied in paper [28]. They are found to be stable and they should not be excluded from the domain G_1.

Permanent Rotations about Staude Cone Axes. The problem of searching for the stability conditions of the permanent rotations under arbitrary mass distribution seemed at first hopelessly difficult [23]. In 1920 R.Grammel obtained the necessary stability conditions which he couldn't analyze because of their complexity and was forced "to invert" the statement of the problem. Sufficient conditions were obtained by Rumyantsev [5] by the construction of the Lyapunov function in the form of the bundle of integrals (3):

$$\mu\nu_1^2 + \Gamma\,\frac{e_1}{\nu_1} > 0, \qquad \Gamma\,\frac{e_1 e_2}{\nu_1\nu_2} + \mu\left(\frac{e_1}{\nu_1}\,\nu_2^2 + \frac{e_2}{\nu_2}\,\nu_1^2\right) > 0, \tag{14}$$
$$\mu\left(\frac{e_2e_3}{\nu_2\nu_3}\,\nu_1^2 + \frac{e_1e_3}{\nu_1\nu_3}\,\nu_2^2 + \frac{e_1e_2}{\nu_1\nu_2}\,\nu_3^2\right) + \Gamma\frac{e_1e_2e_3}{\nu_1\nu_2\nu_3} > 0.$$

Here the variables ν_1, ν_2, ν_3 satisfy equation (9), where μ is an arbitrary constant.

In paper [5] only preliminary analysis of conditions (14) is fulfilled. On its basis some stability domains of permanent rotations at Staude cones are noted.

The application of Theorem 3.2 to the analysis of these rotations is made difficult by the awkwardness of the calculations. Therefore it is natural to do their analysis by computer. Such research was fulfilled for gyrostats [29] and is described in the subsection below.

A Body in the Newtonian Gravity Force Field. Where there is a significant distance between a body and the attracting center in many cases it is assumed only taking account of the forces containing linear terms of the expansions on the degrees of value inverse to this distance. The force field obtained in this way is called the Newtonian field. Motion equations and integrals of a body with a fixed point have the form [30]

$$A_1\dot\omega_1 = (A_2 - A_3)\omega_2\omega_3 + \Gamma(e_2\nu_3 - e_3\nu_2) - \mu(A_2 - A_3)\nu_2\nu_3,$$
$$\dot\nu_1 = \nu_2\omega_3 - \nu_3\omega_2 \qquad (123), \tag{15}$$

$$A_1\omega_1^2 + A_2\omega_2^2 + A_3\omega_3^2 - 2\Gamma(e_1\nu_1 + e_2\nu_2 + e_3\nu_3)$$
$$+ \mu(A_1\nu_1^2 + A_2\nu_2^2 + A_3\nu_3^2) = h,$$
$$A_1\omega_1\nu_1 + A_2\omega_2\nu_2 + A_3\omega_3\nu_3 = k, \tag{16}$$
$$\nu_1^2 + \nu_2^2 + \nu_3^2 = 1.$$

Here μ is a constant characterizing the force field.

Stability of stationary motions of this system was studied by Kuz'min [9]. He established that as for the case of constant gravity stationary motions are permanent rotations about "vertical" the axes of which belong to the Staude cone. Note the convenient parametrization of the permanent rotations introduced in this paper

$$\nu_1 = \frac{\Gamma e_1}{\Omega(\rho - A_1)} \qquad (123), \qquad \Omega^2 = \Gamma^2\sum_{i=1}^{3}\frac{e_i^2}{(\rho - A_i)^2}, \tag{17}$$

where $\Omega = \omega^2 - \mu$, and ω is the angular velocity of permanent rotation. Stability conditions are obtained in accordance with the Routh-Lyapunov theorem as the conditions of the property of having fixed sign second variation of the first integral (16) under conservation the rest integrals on the perturbed motions

$$\Omega T > 0, \tag{18}$$

$$4\omega^2\Omega T + \Omega^2 IS > 0. \tag{19}$$

Here

$$T = \sum_{(123)} (\rho - A_1)(A_2 - A_3)\nu_2^2\nu_3^2,$$

$$S = \sum_{(123)} (\rho - A_2)(\rho - A_3)\nu_1^2,$$

$$I = \sum_{(123)} A_1\nu_1^2,$$

where symbol $\sum\limits_{(123)}$ means the summation of three terms obtained from the one shown under the sum symbol by the cyclic permutation of indexes.

Condition (19) excluding the boundary is not only one of the sufficient conditions but the necessary one. In addition, setting $\Omega = \omega^2$ in inequalities (18) and (19) we obtain from them the stability conditions for constant gravity softening Rumyantsev conditions (14).

5 Permanent Rotations of a Gyrostat

The investigation of permanent rotations stability of a gyrostat was begun by Volterra [11] who considered in detail the permanent rotations of a gyrostat on inertia. Rumyantsev [6] analyzed these rotations by the Lyapunov method. In this paper the sufficient stability conditions of permanent rotations of a heavy gyrostat around the principal axis with arbitrary angular velocity are also obtained. The case when a gyrostat can rotate around the principal axis only with some fixed velocity was studied by Anchev [31]. The investigation of the permanent rotations around the principal axis was continued with the help of Theorem 3.2 in paper [32]. Here is its main result.

Rotations Around the Principal Axis. Under the assumption that the mass center of a gyrostat belongs to the first principal axis and the vector λ of gyrostatic moment is directed along the same axis the gyrostat can rotate permanently around the first axis with angular velocity ω and this rotation is defined by the following values of the variables

$$p_\theta = p_\varphi = 0, \quad p_\psi = \frac{\omega}{a_1} + \lambda, \quad \theta = \varphi = \frac{\pi}{2}, \quad \psi = \omega t + \psi_0, \tag{20}$$

where a_1 is the first component of tensor A^{-1}.

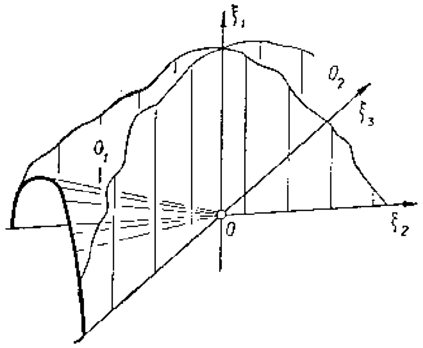

Figure 5.1. The domain of the fulfilment of the necessary stability conditions in the space $O\xi_1\xi_2\xi_3$.

In dimensionless variables the Hamilton function of perturbed motion has the form

$$
\begin{aligned}
&H = H_2 + H_4 + \dots, \\
&2H_2 = ax_1^2 + bx_2^2 + (\omega^2 + \omega\lambda - e)y_1^2 + \big[(\omega+\lambda)\big(a(\omega+\lambda)-\omega\big) - e\big]y_2^2 \\
&\quad +2\big(a(\omega+\lambda)-\omega\big)x_1y_2 + 2\omega x_2y_1, \\
&24H_4 = (3\lambda^2 + 11\lambda\omega + 8\omega^2 + e)y_1^4 + \big[(\omega+\lambda)(4\omega+3\lambda-4a(\omega+\lambda)) + e\big]y_2^4 \\
&\quad +6\big[(\omega+\lambda)(-2\omega-\lambda+2a(\omega+\lambda)) + e\big]y_1^2y_2^2 + 12(1-a)x_1^2y_2^2 + 12x_2^2y_1^2 \\
&\quad +4(4\omega+3\lambda-4a(\omega+\lambda))x_1y_2^3 + 4(5\omega+3\lambda)x_2y_1^3 + 12(a-1)(\omega+\lambda)x_1y_1^2y_2 \\
&\quad +12(-2\omega-\lambda+2a(\omega+\lambda))x_2y_1y_2^2 + 24(a-1)x_1x_2y_1y_2.
\end{aligned}
\tag{21}
$$

To obtain the necessary stability conditions we write the characteristic equation for the linearized system with function H_2

$$
\begin{aligned}
&\mu^4 + \xi\mu^2 + \xi_2\xi_3 = 0, \\
&\xi_1 = ab(\omega+\lambda)^2 - (a+b)(\omega+\lambda)\omega + 2\omega^2 - e(a+b), \\
&\xi_2 = \omega^2(a-1) + a\omega\lambda - ae, \quad \xi_3 = \omega^2(b-1) + b\omega - be.
\end{aligned}
$$

It is convenient to represent the domain D of fulfilment of the necessary stability conditions in the space of parameters ξ_1, ξ_2, ξ_3 (Figure 5.1). In the subdomain G_2 the square form H_2 is of the fixed sign and corresponding permanent rotations are stable. To study the stability of permanent rotations corresponding to the subdomain G_1 Theorem 3.2 is applied. We obtain the expressions for resonance relations and the discriminant by applying the corresponding normalization transformation. Note that under $\lambda = 0$, $a = 1$ the discriminant D has the form (13), i.e. $D \not\equiv 0$. Therefore the equality $D(a, b, \lambda, \omega) = 0$ selects in the space $Oab\lambda\omega$ some manifolds just as resonance relations. Permanent rotations corresponding to the manifolds selected are excluded from consideration. As for the remaining permanent rotations in the domain G_1 on the basis of Theorem 3.2 we conclude that these rotations are stable on Lyapunov.

Comparing these conclusions with the result obtained above for a rigid body we can state that the presence of a rotor in a body renders the stabilization effect on the carrying body motion under the corresponding choice of rotor rotation. Thus, unstable rotation of a body around the middle axis can be made stable under the corresponding choice of a gyrostatic moment. Moreover, any permanent rotation of a rigid body can be made stable under the corresponding choice of a gyrostatic moment.

Rotations Around an Arbitrary Axis. Sufficient stability conditions of a gyrostat around an arbitrary axis of permanent rotations cone were obtained by Anchev [33] and Druzhinin [34]. In paper [34] necessary stability conditions were also obtained. Using these conditions in paper [35] the stability and instability domains are shown on the permanent rotations cone.

It is convenient to describe the permanent rotations considered using the Kuz'min parametrization [9]

$$\nu_1 = \frac{\omega\lambda_1 + \Gamma e_1}{\omega^2(\rho - A_1)} \quad (123), \tag{22}$$

where ρ is an auxiliary parameter and angular velocity satisfies the equation

$$\omega^4 - \sum_{(123)} \frac{(\omega\lambda_1 + \Gamma e_1)^2}{\omega^2(\rho - A_1)^2} = 0.$$

Sufficient stability conditions have the following form [34]

$$D = A_1 A_2 A_3 \omega^6 (L + \omega^2 JM) > 0, \qquad D_1 = \omega^2 L > 0, \tag{23}$$

$$L = \sum_{(123)} (\rho - A_1)[2\omega(A_2 - A_3)\nu_2\nu_3 + \lambda_3\nu_2 - \lambda_2\nu_3]^2,$$

$$M = \sum_{(123)} (\rho - A_2)(\rho - A_3)\nu_1^2, \qquad J = \sum_{(123)} A_1\nu_1^2.$$

Necessary stability conditions are such [34]

$$D > 0, \qquad N > 2\sqrt{D}, \tag{24}$$

$$N = \omega^2 \sum_{(123)} [\omega^2 A_1(\rho - A_1)(A_2\nu_2^2 + A_3\nu_3^2) + A_1(\omega(A_1 - A_2 - A_3)\nu_1 + \lambda_1)^2].$$

Conditions (23) and (24) were analyzed in paper [35]. From the conclusions obtained there we note that the permanent rotations around the axes near to the middle principal axis are unstable. On the stability of the permanent rotations around the axes close to the major axis it is impossible to come to a conclusion on the basis of conditions (23) and (24) because the necessary conditions are fulfilled, but the sufficient conditions (23) are not. For their study Theorem 3.2 is applied.

Analytical difficulties connected with the investigation of the general case have led to the necessity of the use of computer methods for its analysis. The appearance of the computing normalization algorithms for Hamilton systems promotes this analysis. A numerical algorithm of the investigation of stability of the permanent rotations of a gyrostat was created by Chudnenko [29]. Using the parametrization (22) he introduces the partition $\rho_1, \dots, \rho_n$ of the set of permanent rotations. For every ρ_k conditions (23) and (24) are checked. If conditions (24) are fulfilled and conditions (23) are not, conditions of Theorem 3.2 are checked. In addition the values of parameter ρ are noted for which Theorem 3.2 does not solve the stability question. Thus the obtained algorithm permits us to solve practically completely the stability problem of permanent rotations of a gyrostat in the case of general mass distribution.

6 Permanent Rotations of a Rigid Body with Vortex Filling

The permanent rotations of a rigid body with vortex filling are realized only around an axis coincident with the vertical under the condition that vortex components in the moving coordinate system are constant [12]. Because of the complexity of the problem on the distribution of the permanent rotation axes with respect to a body geometrically visual turned out the approach which was based on the construction the set of axes corresponding to the permanent rotation velocity given. These sets are depicted on the plane Ouv defined by the mapping

$$\Gamma^{-1}u = e_1\nu_1^1 - e_3\nu_3^{-1}, \quad \Gamma^{-1}v = e_2\nu_2^{-1} - e_3\nu_3^{-1}, \quad \nu_1^2 + \nu_2^2 + \nu_3^2 = 1.$$

On the basis of the fulfilled analysis interesting effects connected with the vortex presence are obtained. In particular it has been established that a body can rotate permanently around the principal axis when its mass center doesn't belong to the principal axis, and conversely permanent rotations around nonprincipal axes are possible when the mass center belongs to the principal axis.

One of the most interesting effects from the application point of view is the permanent rotation of a shell around the principal axis carring the mass center with angular velocity ω and permanent rotation of a liquid as a rigid body around the same axis with angular velocity Ω. The case of a symmetric body is studied the most completely. On the basis of analysis of necessary stability conditions it is fixed that appearance of codirected vorticity $(\omega\Omega > 0)$ extends with respect to ω the domain of fulfilment of necessary stability conditions and contrarily directed vorticity $(\omega\Omega < 0)$ constricts. Note separately the case when the motion of a body-liquid system is unstable under any value of ω. This takes place for a top with a cavity that is a rotated ellipsoid stretched along the axis of a body symmetry under the fulfilment of some additional conditions. For example, for the case when the mass of a shell is negligibly small with respect to the mass of a liquid in the cavity and the distance from the fixed point to the center of the cavity is equal to the major semiaxis of an ellipsoid c_3 with $c_1 < c_3 < 1.26c_1$, such a "slightly" stretched fluid rotated ellipsoid is unstable no matter with what angular velocity it is rotated. This effect was experimentally found by Lord Kelvin [36].

The sufficient stability conditions and formal stability of permanent rotations around principal and nonprincipal axes are investigated [12]. It has been established that the presence of vortex motion of a liquid in a cavity leads to the appearance of such conditions of shell motion that are absent for a rigid body and a gyrostat. If it is necessary these rotations can be made stable.

7 Permanent Rotations of Multibody Systems

The essential influence on the development of multibody systems dynamics was rendered by the stability problem of permanent rotations of two heavy Lagrange gyros connected by ideal spherical hinges one of which has a fixed point. Ishlinskii called attention to this problem in 1972 at the 13th IUTAM Congress although the characteristic equation for it was obtained in 1898 and was published in Lur'e's monograph [37]. The solution of this problem is given in paper [38] where equations (8) are used for the description of the motion. The permanent rotation of the system of two Lagrange gyros represents such motion in which every from the gyros is rotated with permanent angular velocity about its dynamic symmetry axis collinear to the gravity force direction. Under the stabiliity of this motion is understood the stability of the corresponding solution of equations (8) with respect to some of the variables – namely, to the angular velocities of the bodies S_1, S_2 and to the parameters defining the position in a space of symmetry axes of the bodies S_1 and S_2. Sufficient stability conditions are obtained by the Chetaev method. When both gyros are unbalanced $(c_i > 0)$ they have the form

$$\omega_1\omega_2 > 0, \quad A_2^2\omega_2^2 - 4\mu_2 B_2 > 0, \quad A_1\omega_1^2 - 4\mu_1 B_1 > 0, \tag{24}$$

where $\mu_1 = m_1c_1 + m_2s_1 > 0$, $\mu_2 = m_2c_2 > 0$; where A_1 and B_1 are the axial and equatorial inertia moments of i-th body, respectively. These conditions mean that the bodies are rotated in the same direction. For the second body the Majevskii stability criterion of permanent rotations for one unbalanced Lagrange gyro is fulfilled, the first gyro S_1 with point mass m_2 at the point O_2 has also to satisfy the Majevskii criterion.

The necessary stability conditions are analyzed in detail and are compared with the sufficient ones. Here the interesting stabilization effect is found when one of the remaining unbalanced Lagrange gyros becomes stable under the definite rotation velocity of the second one. This effect calls to mind the stabilization effect of an unbalanced remaining gyro on the oscillating base. But in this case the oscillations of the fixed point arise not at the expense of exterior forces but are, in a definite sense, the self-vibrating ones.

The permanent rotations are also considered for an n-bodies system. It has proved ([13]) to be possible (for $\lambda = 0$) only under the condition when the vectors c_k, s_k, ν are collinear. Equations (8) admit the solution $\omega_k = \omega_k\nu$ which corresponds to the permanent rotations of every body S_k around the axis carring its mass center and coinciding with vertical. In this connection the angular velocities ω_k for each body can be different. Under the additional assumption that the bodies considered are the Lagrange gyros the stability of these rotations was investigated [39] with respect to the angular velocities and the parameters defined the position of the rotation axes in the space. The notion of the "enlarged" body S_k' is introduced which is obtained from the body S_k by adding to it at the point O_k the point mass equals to $\sum\limits_{i=k+1}^{m} m_i$. Such a body is characterized by the parameters A_k, B_k', a_k:

$$B_k' = B_k + s_k^2 \sum_{i=k+1}^{n} m_i, \quad a_k = m_kc_k + s_k \sum_{i=k+1}^{n} m_i, \quad k = 1, \ldots, n-1,$$

where A_k, B_k are the axial and equatorial inertia moments of the body S_k with respect to its suspension point O_k.

For the case when all "the enlarged" bodies are unbalanced $(a_k > 0)$ the sufficient stability conditions are obtained

$$A_k^2\omega_k^2 - 4B_k^{'}a_kg > 0, \quad \omega_k\omega_i > 0; \quad k,\, i = 1,\dots,n$$

which generalize the Majevskii criterion and conditions (24). Necessary stability conditions of permanent rotations and regular precessions have been considered. More complicated motions named "similar" ones [12] when symmetry axes of Lagrange gyros belong to one plane during the entirety of the motion are found and investigated.

New effects were discovered during the analysis of the influence of the stiffness in the elastic joints on the stability of the permanent rotations of a multibody system. In particular, under specific values of the stiffness instability interval appears in the problem which classical analog is Euler case. However, when the stiffness is rather great this system behaves as a single rigid body and under the fulfilment of the Majevskii criterion for the body obtained from the system considered by the change of the joints on rigid fixings permanent rotations are stable.

In conclusion, we note a new direction in the investigation of the stability of permanent rotations connected with studying the influence of small nonsymmetry on the stability of the motion of the system of the bodies connected by elastic joints. The analysis of the motion of a single unsymmetric body has already showed that its stable permanent rotations about the symmetry axis after introducing the system debalance pass into unstable ones in the neighborhoods of some frequencies named resonance ones [39, 40]. The research into multibody systems discovered similar situations. A general approach for finding the resonance frequencies was offered and a constructive algorithm for finding two groups of such frequencies was created which gave the possibility to obtain the analytical expressions for them in some cases. With its help the motion of multibody systems was studied with different ways of connection and the force action and the critical operating conditions of the elastic objects motion were established.

References

[1] Staude, O. Uber permanente Rotationsaxen bei der Bewegung eines schweren Korpers um einen festen Punkt. *Z. Reine und Angew. Math.* **113**(H4) (1894) 318–334.

[2] Routh, E.J. *A Treatise on the Dynamics of a System of Rigid Bodies.* The advanced part. Macmillan and Co., London, 1884.

[3] Chetaev, N.G. On the stability of a rotation of a rigid body with one fixed point in the Lagrange case. *Prikl. Mat. Mekh.* **18**(1) (1954) 123–124. (Russian).

[4] Chetaev, N.G. *The Stability of a Motion.* Gostehizdat, Moscow, 1954. (Russian).

[5] Rumyantsev, V.V. Stability of permanent rotations of a heavy rigid body. *Prikl. Mat. Mekh.* **20**(1) (1956) 51–66. (Russian).

[6] Rumyantsev, V.V. On the stability of a motion of gyrostats. *Prikl. Mat. Mekh.* **25**(1) (1961) 9–16. (Russian).

[7] Kharlamov, P.V. *Lectures on the Rigid Body Dynamics.* Novosibirsk University Press, Novosibirsk, 1965. (Russian).

[8] Moiseev, N.N. and Rumyantsev, V.V. *Dynamics of a Body with cavities Containing a Liquid.* Nauka, Moscow, 1965. (Russian).

[9] Kuz'min, P.A. Stationary motions of a rigid body and their stability in the central field of gravity. *Proc. Interinst. Conference on Appl. Theory of Stability of a Motion and Analyt. Mech.* Kazan'62, (1964). (Russian).

[10] Kolmogorov, A.N. On the conservation of the conditionally periodic motions under a small variation of the Hamilton function. *Dokl. Akad. Nauk SSSR* **98**(4) (1954) 527–530. (Russian).

[11] Volterra, V. Sur la theorie des variations des latitudes. *Acta Math.* **22** (1899) 201–358.

[12] Savchenko, A.Ya. *Stability of Stationary Motions of Mechanical Systems.* Naukova Dumka, Kiev, 1977. (Russian).

[13] Kharlamov, P.V. On the equations of a motion of a system of two rigid bodies. *Mekh. Tver. Tela* **4** (1972) 52–73. (Russian).

[14] Kovalev, A.M. and Kiselev, A.M. On the cone of the permanent rotation axes of a gyrostat. *Mekh. Tver.Tela* **4** (1972) 36–45. (Russian).

[15] Lyapunov, A.M. On the permanent spiral motions of a body in a liquid. *Reports of the Kharkov Math. Society. Ser.2* **1**(1–2) (1888) 7–60. (Russian).

[16] Rumyantsev, V.V. On the stability of the stationary motions. *Prikl. Mat. Mekh.* **30**(5) (1966) 922–933. (Russian).

[17] Rubanovskii, V.N. and Stepanov, S.Ya. On the Routh theorem and the Chetaev method for the construction Lyapunov function from the integrals of the motion equations. *Prikl. Mat. Mekh.* **33**(5) (1969) 904–912. (Russian).

[18] Kovalev, A.M. and Savchenko, A.Ya. On the stability of the stationary motions of the Hamilton systems. *Dokl. Akad. Nauk UkrSSR, Ser. A* **6** (1975) 521–524. (Russian).

[19] Arnold, V.I. On the stability of equilibrium positions of the Hamilton system of ordinary differential equations in the general elliptic case. *Dokl. Akad. Nauk SSSR* **137**(2) (1961) 255–257. (Russian).

[20] Moser, J. *Lectures on the Hamilton Systems.* Mir, Moscow, 1973.

[21] Markeev, A.P. On the stability of the canonical system with two degrees of freedom under the presence of a resonance. *Prikl. Mat. Mekh.* **32**(4) (1968) 738–744. (Russian).

[22] Markeev, A.P. On the stability of the triangular libration points in the circular boundary problem of three bodies. *Prikl. Mat. Mekh.* **33**(1) (1969) 112–116. (Russian).

[23] Grammel, R. Die Stabilitat der Staudeschen Kreiselbewegungen. *Math. Z.* **6** (1920).

[24] Bottema, O. De stabiliteit van de tolbewegingen van Staude. *Proc. Koninklijke Nederlandsche Akad. van Wetenschappen* **48** (1945) 316–325.

[25] Rumyantsev, V.V. On the stability of the permanent rotations of a rigid body with fixed point. *Prikl. Mat. Mekh.* **21**(3) (1957) 339–346. (Russian).

[26] Kovalev, A.M. and Savchenko, A.Ya. Stability of the permanent rotations of a rigid body around principal axis. *Prikl. Mat. Mekh.* **39**(4) (1975) 650–660. (Russian).

[27] Grammel, R. *The Gyroscope, Its Theory and Applications.* V. 1. Inostr. Literatura, Moscow, 1952. (Russian).

[28] Kovalev, A.M. and Savchenko, A.Ya. Stability of stationary motions of the Hamilton systems under the presence of the resonance of the fourth order. *Mekh. Tver.Tela* **9** (1977) 40–44. (Russian).

[29] Chudnenko, A.N. Numerical algorithm of the investigation of the stability of the permanent rotations of a gyrostat. *Mekh. Tver. Tela* **17** (1985) 61–65. (Russian).

[30] Beletskii, V.V. Some questions of a motion of a rigid body in the Newtonian force field. *Prikl. Mat. Mekh.* **21**(6) (1957) 749–758. (Russian).

[31] Anchev, A. On the stability of the permanent rotations of a heavy gyrostat. *Prikl. Mat. Mekh.* **26**(1) (1962) 22–28. (Russian).

[32] Kovalev, A.M. Stability of the permanent rotations of a heavy gyrostat around the principal axis. *Prikl. Mat. Mekh.* **44**(6) (1980) 994–998. (Russian).

[33] Anchev, A. On the permanent rotations of a heavy gyrostat with fixed point. *Prikl. Mat. Mekh.* **31**(1) (1967) 49–58. (Russian).

[34] Druzhinin, E.I. Stability of the stationary motions of gyrostats. *Proc. Kazan Aviation Inst.* **92** (1966) 12–23. (Russian).

[35] Kovalev, A.M. and Kiselev, A.M. Separation of the stability domains on the cone of the permanent rotations axes of a gyrostat. *Mekh. Tver. Tela* **4** (1972) 46–48. (Russian).

[36] Kelvin Lord *Mathematical and Physical Papers.* Cambridge, Vol. 4, 1882, 193–201.

[37] Lur'e, A.I. *Analytical Mechanics.* Fizmatgiz, Moscow, 1961. (Russian).

[38] Savchenko, A.Ya. Investigation of the stability of the permanent rotations of the system of two Lagrange gyros. *Prikl. Mekh.* **10**(12) (1974) 71–77. (Russian).

[39] Savchenko, A.Ya., Bolgrabskaya, I.A. and Kononyhin, G.A. *Stability of a Motion of the Systems of Connected Rigid Bodies.* Naukova Dumka, Kiev, 1991. (Russian).

[40] Bolgrabskaya, I.A. and Savchenko, A.Ya. Stability of the permanent rotations of the free bundle of n Lagrange gyros. *Mat. Phys. i Nelin. Mekh.* **2** (1984) 9–14. (Russian).

Nonlinear Dynamics and Systems Theory, **1**(1) (2001) 99–110

On the Design of Nonlinear Controllers for Euler-Lagrange Systems

L. Luyckx, M. Loccufier and E. Noldus

Automatic Control Department, University of Ghent,
Technologiepark-Zwijnaarde9, B-9052 Zwijnaarde, Belgium

Received: February 8, 2000; Revised: August 1, 2000

Abstract: The dynamics are studied of nonlinear feedback loops for the set point control of Euler-Lagrange (EL) systems. A class of controllers is considered that possess a linear dynamic component and several nonlinear amplifiers. Frequency domain conditions are presented for nonoscillatory behaviour of the closed loop, by which is meant that for increasing time all bounded solutions converge to one of the system's equilibrium states. The results constitute a systems theoretical basis for a new controller design method for EL systems.

Keywords: *Euler-Lagrange systems; nonlinear control; Liapunov's method; convergence criteria; stability regions.*

Mathematics Subject Classification (2000): 34D20, 70H35, 70K15, 70Q05.

1 Introduction

Euler-Lagrange systems constitute the outcome of a powerful mathematical modelling technique for dynamic processes, the variational method [4]. Their structural properties and constraints have been exploited to develop practically meaningful controller design procedures including Liapunov based methods [9], passivity based control [6] or the stabilization scheme of backstepping [1]. Most of the literature dealing with set point regulation of EL systems concentrates on the global asymptotic stabilization of a unique closed loop equilibrium state, elaborating on such fundamental concepts as potential energy shaping and damping injection [8]. Nevertheless there remain several drawbacks that stymie the utilization of these methodologies in practical applications. For example, the global stabilization of a unique equilibrium point often requires control inputs beyond the physical saturation constraints of the actuators. This has led to the development of saturated controllers which apply to EL systems with limited growth rates of the potential energy functions for large position values [3].

© 2001 Informath Publishing Group. All rights reserved.

In this paper we consider a class of nonlinear feedback controllers for EL systems that allow the existence of several closed loop equilibria. Sufficient conditions are established for nonoscillatory behaviour of the loop, by which is meant that for increasing time every bounded solution converges to one of the equilibria. If for increasing time all solutions remain bounded, nonoscillatory behaviour implies the convergence of all solutions to the set of equilibria, i.e. the set of equilibria is globally convergent. The conditions for nonoscillatory behaviour are much less restrictive than those for the global asymptotic stabilization of a unique equilibrium point. Thus they constitute a systems theoretical basis for an alternative controller design method in cases where a closed loop phase portrait possessing several equilibrium states is acceptable or desirable.

Following some background concepts in Section 2, Section 3 presents the basic conditions which guarantee closed loop nonoscillatory behaviour and global convergence of the set of equilibria. Sections 4 and 5 contain their application to a dynamic output feedback controller possessing several nonlinear amplifiers, some comments regarding controller design and some special cases. Section 6 discusses an application to EL controllers. In Section 7 a simple example is worked to illustrate the proposed concepts. The paper terminates with an overview of possible extensions to the theory and of further work under development.

2 EL Systems: Basic Concepts and Assumptions

Consider an EL system [4]

$$\frac{\partial \mathcal{L}}{\partial q}(q,\dot q) - \frac{d}{dt}\left[\frac{\partial \mathcal{L}}{\partial \dot q}(q,\dot q)\right] - \frac{\partial \mathcal{F}}{\partial \dot q}(\dot q) + Mu = 0, \tag{1}$$

where $q \in R^m$ is a vector of generalized coordinates, $\mathcal{L}(q,\dot q) \triangleq \mathcal{T}(q,\dot q) - \mathcal{V}(q)$ is the Lagrangian and $\mathcal{F}(\dot q)$ is Rayleigh's dissipation function. $u \in R^r$ denotes a linearly entering input force. We assume that the kinetic energy $\mathcal{T}(q,\dot q)$ and the potential energy $\mathcal{V}(q)$ belong to the class of C^1-functions (continuous with continuous partial derivatives w.r.t. their arguments), that rank $M = r \le m$ and that the Lipschitz conditions are satisfied which ensure the existence and the uniqueness of the solutions of (1) for given initial conditions and for a given input $u(.)$. The system is called fully actuated if $r = m$, otherwise it is underactuated. Following Meirovitch [4] we assume that $\mathcal{T}(q,\dot q)$ depends quadratically on the components of $\dot q$:

$$\mathcal{T}(q,\dot q) = \frac{1}{2}\,\dot q' D(q)\dot q + b'(q)\dot q + c(q), \tag{2}$$

where the generalized inertia matrix $D(q) = D'(q) \in R^{m\times m}$ is positive definite, $b(q) \in R^m$ and $c(q) \in R$. Defining the Hamiltonian as

$$\mathcal{H}(q,\dot q) \triangleq \frac{1}{2}\,\dot q' D(q)\dot q + \mathcal{V}(q) - c(q) \tag{3}$$

it is easily verified that along the solutions of (1):

$$\frac{d\mathcal{H}}{dt}(q,\dot q) = -\dot q'\,\frac{\partial \mathcal{F}}{\partial \dot q}(\dot q) + \dot q' M u. \tag{4}$$

We assume that

$$\dot{q}' \frac{\partial \mathcal{F}}{\partial \dot{q}}(\dot{q}) \geq 0; \quad \forall\, \dot{q} \in R^m. \tag{5}$$

If

$$\dot{q}' \frac{\partial \mathcal{F}}{\partial \dot{q}}(\dot{q}) > 0; \quad \forall\, \dot{q} \neq 0 \tag{6}$$

then the EL system is said to be fully damped. Otherwise it is underdamped. Observing that the system state is $x \triangleq \begin{bmatrix} q \\ \dot{q} \end{bmatrix} \in R^{2m}$, (4) shows that (1) is a dissipative system [11] with storage function $\mathcal{H}(q, \dot{q})$ and supply $\dot{q}'Mu$. We shall assume that the output

$$w \triangleq M'q \in R^r \tag{7}$$

is available for feedback. (1), (7) define an EL system with collocated actuator-sensor control [7].

3 Closed Loop Nonoscillatory Behaviour and Global Convergence

Let

$$\dot{z} = \varphi(z, w), \tag{8}$$

$$u = \psi(z, w) \tag{9}$$

with state $z \in R^n$ be a feedback controller for (1), (7). Let $x_c \triangleq \begin{bmatrix} x \\ z \end{bmatrix} \in R^{2m+n}$ be the closed loop state vector. Suppose a scalar function $V(z, w) \in C^1$ can be found such that along the solutions of (8)

$$\dot{z}' \frac{\partial V}{\partial z}(z, w) = \varphi'(z, w) \frac{\partial V}{\partial z}(z, w) \leq 0; \quad \forall\, z \in R^n, \quad \forall\, w \in R^r. \tag{10}$$

Define

$$V_c(x_c) \triangleq \mathcal{H}(q, \dot{q}) + V(z, w). \tag{11}$$

It follows that

$$\begin{aligned} \dot{V}_c(x_c) &= -\dot{q}' \frac{\partial \mathcal{F}}{\partial \dot{q}}(\dot{q}) + \dot{q}'Mu + \dot{z}' \frac{\partial V}{\partial z}(z, w) + \dot{w}' \frac{\partial V}{\partial w}(z, w) \\ &= -\dot{q}' \frac{\partial \mathcal{F}}{\partial \dot{q}}(\dot{q}) + \dot{z}' \frac{\partial V}{\partial z}(z, w) \leq 0; \quad \forall\, x_c \in R^{2m+n}, \end{aligned} \tag{12}$$

if we choose

$$u = \psi(z, w) \triangleq -\frac{\partial V}{\partial w}(z, w). \tag{13}$$

By (12), $V_c(x_c)$ is a global Liapunov function for the closed loop system. Invoking Lasalle's invariance principle [2] it follows that every solution $x_c(t)$ that remains bounded for $t \geq 0$ will for $t \to +\infty$ converge to the largest invariant subset $\mathcal{M}$ of the closed loop state space where

$$-\dot{q}'\,\frac{\partial \mathcal{F}}{\partial \dot{q}}\,(\dot{q}) + \dot{z}'\,\frac{\partial V}{\partial z}\,(z,w) \equiv 0. \tag{14}$$

Suppose we can select $V(z,w)$ such that $\mathcal{M}$ consists of the set of the closed loop equilibria. Then every solution $x_c(t)$ that remains bounded $t \geq 0$ will converge to an equilibrium point. A system possessing this property will be called nonoscillatory. Every solution of a nonoscillatory system either tends to infinity or converges to an equilibrium point as $t \to +\infty$. It cannot perform complicated motions such as periodic oscillations or chaos. As a corollary to the above we have

Lemma 3.1 *If the EL system (1) is fully damped and if in addition to (10),*

$$\dot{z}'\,\frac{\partial V}{\partial z}\,(z,w) = 0 \quad \Longleftrightarrow \quad \dot{z} = 0 \tag{15}$$

then the closed loop (1), (7), (8), (13) is nonoscillatory.

If all solutions of a nonoscillatory system remain bounded for $t \geq 0$, then every solution converges to an equilibrium state as $t \to +\infty$. In other words the set of the equilibria is globally convergent. The boundedness of solutions can often easily be proved, for example using a suitable Liapunov function. Specifically we have

Lemma 3.2 *If in addition to the conditions of Lemma 3.1, $V_c(x_c)$ is radially unbounded then the set of the closed loop equilibria is globally convergent.*

4 Controllers with Several Arbitrary Nonlinear Amplifiers

Consider a controller with state dynamics of the form

$$\dot{z} = Az - Bf(\sigma) + \eta(w), \tag{16}$$

$$\sigma = C'z + \zeta(w), \tag{17}$$

where $A \in R^{n\times n}$ is nonsingular; $B,\ C \in R^{n\times s}$; $\eta \in R^n$; $\zeta \in R^s$; $f(\sigma) = \operatorname{col}[f_i(\sigma_i);\ i = 1,\dots,s]$. Let

$$V(z,w) \triangleq z'Pz + \int\limits_0^\sigma f'(\theta)\bar{\alpha}\,d\theta + z'p(w) + \mu(w)$$

with $P = P' \in R^{n\times n}$; $\bar{\alpha} = \operatorname{diag}(\alpha_i) \in R^{s\times s}$; $\theta \in R^s$; $p(w) \in R^n$ and $\mu(w) \in R$. Partial differentiation of $V(z,w)$ along the solutions of (16), (17) produces

$$\begin{aligned}
\left(\frac{\partial V}{\partial z}\right)'\dot{z} &= \dot{z}'Pz + z'P\dot{z} + f'(\sigma)\bar{\alpha}C'\dot{z} + \dot{z}'p(w)\\
&= \dot{z}'PA^{-1}[\dot{z} + Bf(\sigma) - \eta(w)]\\
&\quad + [\dot{z} + Bf(\sigma) - \eta(w)]'A^{-1\prime}P\dot{z} + f'(\sigma)\bar{\alpha}C'\dot{z} + \dot{z}'p(w)\\
&= \dot{z}'[PA^{-1} + A^{-1\prime}P]\dot{z} + \dot{z}'[2PA^{-1}B + C\bar{\alpha}]f(\sigma)\\
&\quad + \dot{z}'[p(w) - 2PA^{-1}\eta(w)].
\end{aligned}$$

Defining $W \triangleq A^{-1\prime}PA^{-1}$ and choosing

$$A'W + WA = -QQ' - \varepsilon I; \quad \varepsilon > 0, \tag{18}$$

$$2WB + A^{-1\prime}C\bar{\alpha} = 0, \tag{19}$$

$$p(w) = 2A'W\eta(w) \tag{20}$$

results in

$$\left(\frac{\partial V}{\partial z}\right)'\dot{z} = -\dot{z}'QQ'\dot{z} - \varepsilon\dot{z}'\dot{z} \tag{21}$$

which satisfies (10) and (15). By virtue of the Kalman-Yacubovich-Popov main lemma [10] the system (18), (19) has a real solution $W = W' \in R^{n\times n}$, $Q \in R^{n\times s}$ for a sufficiently small $\varepsilon > 0$ if and only if for all real ω:

$$2He[(-A^{-1\prime}C\bar{\alpha})'(j\omega I - A)^{-1}2B] > 0 \tag{22}$$

(positive definite). (22) can readily be transformed into the frequency condition

$$He\,\frac{1}{j\omega}\,\bar{\alpha}[G(j\omega) - G(0)] < 0, \quad \forall\,\omega \in R \tag{23}$$

(negative definite), where

$$G(s) \triangleq C'(sI - A)^{-1}B \tag{24}$$

represents the transfer matrix of the controller's linear dynamic component. Defining $w \triangleq \operatorname{col}[w_i;\ i = 1,\dots,r]$; $\mu_d(w) \triangleq \operatorname{col}\left[\frac{\partial\mu}{\partial w_i};\ i = 1,\dots,r\right]$;

$$\zeta_d(w) \triangleq \begin{bmatrix} \frac{\partial\zeta_1}{\partial w_1} & \cdots & \frac{\partial\zeta_1}{\partial w_r} \\ \vdots & \ddots & \vdots \\ \frac{\partial\zeta_s}{\partial w_1} & \cdots & \frac{\partial\zeta_s}{\partial w_r} \end{bmatrix} \in R^{s\times r}$$

and $\eta_d(w)$ similarly the control law (13) becomes

$$u = -[\zeta_d{}'(w)\bar{\alpha}f(\sigma) + 2\eta_d{}'(w)WAz + \mu_d(w)]. \tag{25}$$

Special cases occur for the choices

$$\zeta(w) \equiv 0; \quad \eta(w) = H'w, \quad \text{hence} \quad \eta_d(w) = H', \tag{26}$$

$$\eta(w) \equiv 0; \quad \zeta(w) = Z'w, \quad \text{hence} \quad \zeta_d(w) = Z'. \tag{27}$$

For the choice (26) the controller dynamics simplify to

$$\dot{z} = Az - Bf(C'z) + H'w, \tag{28}$$

$$u = -[2HWAz + \mu_d(w)] \tag{29}$$

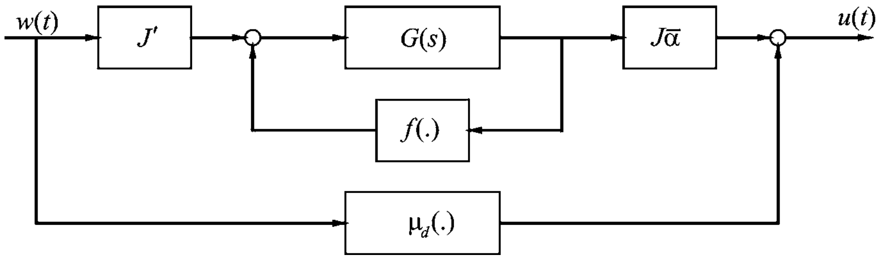

Figure 4.1. Block diagram of the controller (28), (29), (32).

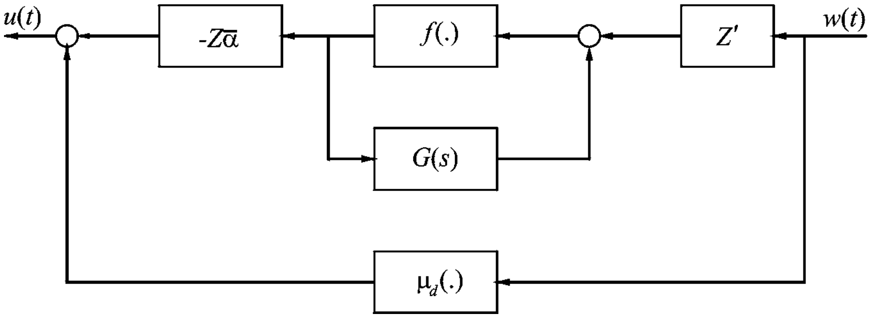

Figure 4.2. Block diagram of the controller (30), (31).

and for the choice (27):

$$\dot{z} = Az - Bf(C'z + Z'w), \tag{30}$$

$$u = -[Z\bar{\alpha}f(C'z + Z'w) + \mu_d(w)]. \tag{31}$$

If we take H of the special form

$$H = JB', \tag{32}$$

then the controller (28), (29) can be represented in the block diagram form of Figure 4.1. The controller (30), (31) has the block diagram representation of Figure 4.2.

5 Discussion

In summary of Section 4 the controller (16), (17), (25) and in particular the controllers of Figures 4.1 and 4.2 ensure closed loop nonoscillatory behaviour provided the EL system (1) is fully damped and the transfer matrix of the controller's linear dynamic component $G(s)$ satisfies condition (23) for some diagonal $\bar{\alpha}$. All other controller components are arbitrary. In the underdamped case nonoscillatory behaviour is still guaranteed if the largest invariant subset $\mathcal{M}$ of the closed loop state space where $\dot{z} \equiv 0$ and $\dot{q}' \frac{\partial \mathcal{F}}{\partial \dot{q}}(\dot{q}) \equiv 0$ consists of the union of the equilibrium points.

Furthermore in (21) we may choose $\varepsilon = 0$, which weakens the negative definiteness condition (23) to negative semidefinite, provided the largest invariant set where $Q'\dot{z} \equiv 0$ and $\dot{q}' \frac{\partial \mathcal{F}}{\partial \dot{q}}(\dot{q}) \equiv 0$ still consists of the union of the equilibrium points.

The condition (23) is relatively mild. In the special case of a single nonlinearity $f(.)$ in the controller structure $(s = 1)$ it simplifies to

$$\frac{\alpha}{\omega} ImG(j\omega) < 0, \quad \forall\, \omega \in R, \tag{33}$$

where $G(s)$ is a scalar transfer function. For example in the second order case $G(s) = 1/(s^2 + as + b)$, (33) is satisfied for $\alpha a > 0$. For $s = 0$, i.e. for controllers without nonlinearities $f(.)$, the conditions (18), (19) simplify to the single condition

$$A'W + WA < 0. \tag{34}$$

It is easy to show that if A has no characteristic values on the imaginary axis there always exists a real symmetric W such that (34) holds.

For nonoscillatory systems the method of the closest unstable equilibrium point is a well known direct method of the Liapunov type for estimating regions of asymptotic stability (RAS) in state space for the system's stable equilibria $\hat{x}_{c,s}$ [5]. The method requires that a global Liapunov function $V_c(x_c) \in C^1$ can be found such that:

1. The associated invariant set $\mathcal{M}$ consists of the union of all equilibrium points.
2. $V_c(x_c)$ possesses an absolute minimum $V_{c,min}$ on the stability boundary of $\hat{x}_{c,s}$. The existence of the minimum is ensured if all solutions of the system remain bounded for $t \geq 0$. Hence the conditions for applicability of the method are exactly those which have been imposed on the control loop in the sections above.

In design problems, once nonoscillatory behaviour has been established the controller's structure must be further specified to implement the control objectives w.r.t. the location of the closed loop equilibria in state space, the linearized system dynamics around the stable equilibria and their RAS. In the next sections we consider the application of EL controllers to an EL system and we present a design example.

6 EL Controllers

In the literature it has been proposed to control EL systems by means of controllers that itself possess an EL structure [6]. Consider an EL controller of the form

$$D_0\ddot{p} + C_0\dot{p} + K_0 p + C_1 f(C_1' p) + \nu(w) = 0, \tag{35}$$

where D_0, K_0 and C_0 are symmetric and positive definite. (35) can be written in the state representation (16), (17) with

$$z \triangleq \begin{bmatrix} p \\ \dot{p} \end{bmatrix}; \quad A \triangleq \begin{bmatrix} 0 & I \\ -D_0^{-1}K_0 & -D_0^{-1}C_0 \end{bmatrix}; \quad B \triangleq \begin{bmatrix} 0 \\ D_0^{-1}C_1 \end{bmatrix};$$

$$C \triangleq \begin{bmatrix} C_1 \\ 0 \end{bmatrix}; \quad \eta(w) \triangleq \begin{bmatrix} 0 \\ -D_0^{-1}\nu(w) \end{bmatrix}; \quad \zeta(w) = 0.$$

Straightforward calculations reveal that $G(s) = G'(s) = {C_1}'[D_0 s^2 + C_0 s + K_0]^{-1} C_1$ such that, observing that $G(0)$ is symmetric and assuming C_1 has full rank s, (23) holds with $\bar{\alpha} = I$ if and only if

$$\frac{1}{2j\omega}\left[(K_0 - D_0\omega^2 + C_0 j\omega)^{-1} - (K_0 - D_0\omega^2 - C_0 j\omega)^{-1}\right]$$
$$= -\{(K_0 - D_0\omega^2) C_0^{-1} (K_0 - D_0\omega^2) + C_0\omega^2\}^{-1} < 0$$

which is true for all $\omega \in R$. (18)$-$(20) where for simplicity we take $\varepsilon = 0$ yields

$$P = \frac{1}{2}\begin{bmatrix} K_0 & \\ & D_0 \end{bmatrix}; \quad Q = \begin{bmatrix} C_0^{\frac{1}{2}} \\ 0 \end{bmatrix}; \quad p(w) = \begin{bmatrix} \nu(w) \\ 0 \end{bmatrix}$$

while

$$\left(\frac{\partial V}{\partial z}\right)' \dot{z} = -\dot{p}' C_0 \dot{p}. \tag{36}$$

The control law (25) becomes

$$u = -[\nu_d(w) p + \mu_d(w)]. \tag{37}$$

Substituting (36) in the left hand side of (14) shows that (37) renders the closed loop nonoscillatory assuming the controlled EL system (1) satisfies the damping conditions discussed in Section 5.

7 Example

Figure 7.1 displays a simple conceptive example of a one-degree-of-freedom system in its set point equilibrium position $y = 0$. Rescaling time as $\tau = \omega_0 t$; $\omega_0 \triangleq \sqrt{\frac{k}{m}}$ and defining $\zeta \triangleq \frac{c}{2\sqrt{km}}$, $\rho \triangleq \frac{\sqrt{l^2 + d^2}}{d}$, $q \triangleq \frac{y}{d}$ and $u \triangleq \frac{f_0}{kd}$ the equation of motion can be written in dimensionless form as:

$$\ddot{q} + 2\zeta\dot{q} + g(q) = u \tag{38}$$

with

$$g(q) \triangleq \left[1 - \frac{\rho}{\sqrt{\rho^2 + q^2 + 2q}}\right](1 + q).$$

There are three open loop equilibria resp. at $q = 0$, $q = -1$ and $q = -2$. A first order controller state equation of the form (16), (17) with $w = q$ reads

$$\dot{z} = -az - f(z) + \eta(q), \quad z \in R, \quad a \neq 0. \tag{39}$$

The frequency condition (33) where $G(s) = \frac{1}{s+a}$ holds for $\alpha > 0$. Now taking $\varepsilon = 0$ some computations yield

$$V_c(x_c) = \frac{1}{2}\dot{q}^2 + \int_0^q g(\theta)\,d\theta + \frac{\alpha a}{2} z^2 + \alpha \int_0^z f(\theta)\,d\theta - \alpha z \eta(q) + \mu(q), \tag{40}$$

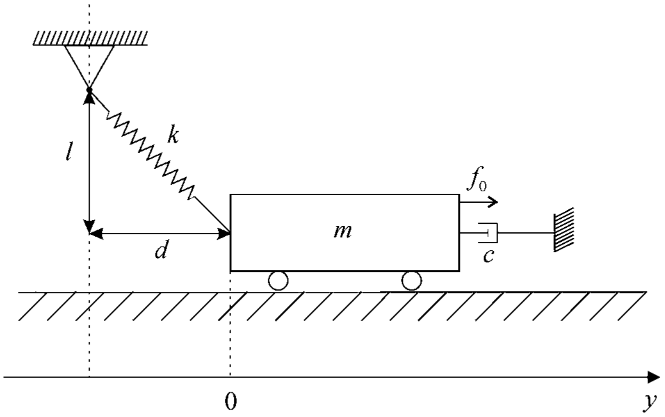

Figure 7.1. A one-degree-of-freedom nonlinear EL system.

where $x_c = [q, \dot{q}, z]'$. Differentation along the solutions of (38), (39) produces

$$\dot{V}_c(x_c) = -2\zeta\dot{q}^2 - \alpha\dot{z}^2 \le 0; \quad \forall\, x_c \in R^3, \tag{41}$$

if according to (13)

$$u = -\frac{\partial V}{\partial q}\,(z, q) = \alpha z \eta_d(q) - \mu_d(q), \tag{42}$$

where the subscript d denotes differentiation w.r.t. q. (41) ensures that in closed loop the state space's largest invariant subset where $\dot{V}_c(x_c) \equiv 0$ consists of the union of the equilibrium points. Next assume that $|\eta(q)|$, $|\eta_d(q)|$ and $|\mu_d(q)|$ are bounded for all $q \in R$ and that

$$\frac{az + f(z)}{z} \ge k_1 > 0; \quad \forall\, z \in R, \quad z \ne 0 \tag{43}$$

with $f(0) = 0$. Then it is an easy exercise to show that:

1. $V_c(x_c)$ is radially unbounded such that the set of the closed loop equilibria is globally convergent.
2. $$\frac{d}{dt}\,z^2 \le 0 \quad \text{for} \quad |z| \ge \frac{|\eta(q)|_{\max}}{k_1} \triangleq n_0 \tag{44}$$

   such that $|z(0)| \le n_0$ implies $|z(t)| \le n_0$ for all $t \ge 0$, hence the control force remains bounded:

$$|u(t)| \le \alpha n_0 |\eta_d(q)|_{\max} + |\mu_d(q)|_{\max}; \quad \forall\, t \ge 0. \tag{45}$$

Let the desired closed loop equilibria be $x_{c0} = [0, 0, 0]'$ (set point); $x_{c1} = [q_1, 0, 0]'$; $x_{c2} = [q_2, 0, 0]'$, where $q_2 < q_1 < -1$ and let $\Lambda = \{\lambda_i,\ i = 1, \dots, 3\}$ be a selected eigenvalue spectrum in $\{Re\, s < 0\}$ for the linearized closed loop dynamics at x_{c0}. As an example choose

$$\mu_d(q) = \frac{m_1 q + m_2 q^2}{1 + m_3 q^2}; \quad m_3 > 0,$$

$$\eta(q) = \frac{m_4 q (1 - \frac{q}{q_1})(1 - \frac{q}{q_2})}{\sqrt{1 + m_5 q^6}}; \quad m_5 > 0,$$

$$f(z) = \gamma z^3; \quad \gamma > 0$$

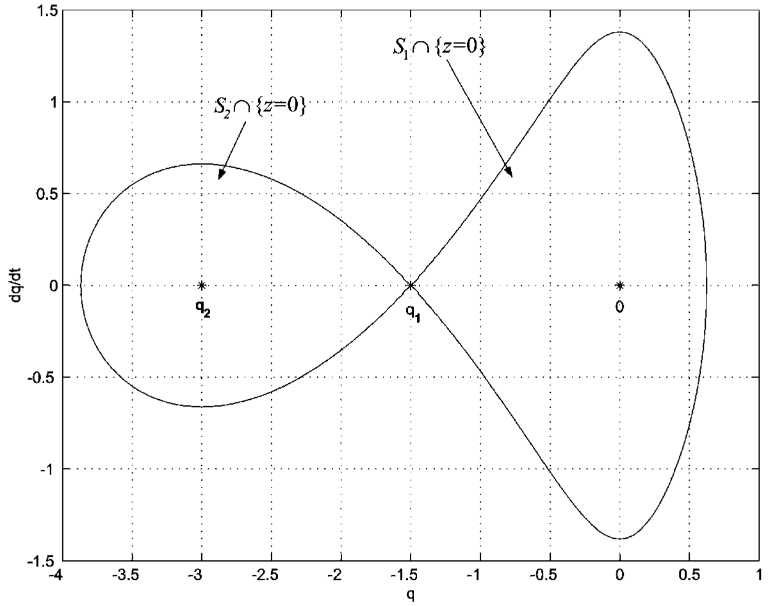

Figure 7.2. Intersections of the estimated stability regions S_1 and S_2 with the plane $\{z = 0\}$. Parameter values: $\rho = 2,\ q_1 = -1.5,\ q_2 = -3,\ \Lambda = \{-1, -1 + j\sqrt{3}, -1 - j\sqrt{3}\},\ a = 3,\ m_1 = 5.75,\ m_2 = 3.6238,\ m_3 = 3.3861$.

which implies that $k_1 = a > 0$. Now a straightforward analysis reveals that:

1. For $\alpha > 0$ and sufficiently small there are no other closed loop equilibria besides x_{c0} (of index 0), x_{c1} (of index 1) and x_{c2} (of index 0).
2. Λ, q_1 and q_2 can be arbitrarily assigned by suitabl tuning the parameters a and $m_1 \to m_4$.

In addition to α the remaining free control parameters are m_5 and γ. Their choice influences the upper bound of the control force $|u(t)|_{\max}$, the extent of the set point's region of attraction in state space and the linearized dynamics at x_{c1} and x_{c2}. The method of the closest unstable equilibrium point produces the set

$$S \triangleq \{x_c \in R^3;\ V_c(x_c) < V_c(x_{c1})\}$$

which consists of two disjoint subsets $S_1 \ni x_{c0}$ and $S_2 \ni x_{c2}$. These constitute estimated regions of attraction for x_{c0} and x_{c2}. The control parameters α, m_5 and γ do not affect the intersections of S_1 and S_2 with the plane $\{z = 0\}$ (Figure 7.2), but they bear an influence on the extent of the stability regions in the z-direction (Figure 7.3).

8 Conclusion: Extensions and Further Work

We have derived sufficient conditions for a class of nonlinear feedback controllers for EL systems to render the closed loop nonoscillatory. The obtained results can be extended in several ways. As to the controlled system, other classes of dissipative processes can be considered possessing various types of nonlinear components. Noncollocal control of EL

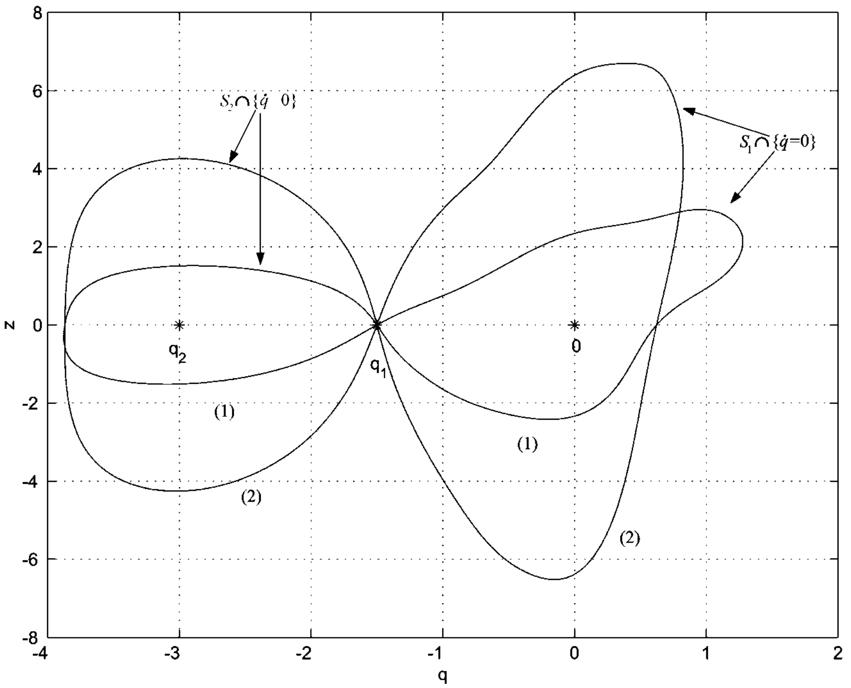

Figure 7.3. Intersections of the estimated stability regions S_1 and S_2 with the plane $\{\dot{q} = 0\}$ for the parameter values of Figure7.2 and for: (1) $\alpha = 0.025$, $m_4 = 23.6643$, $\gamma = 4$, $m_5 = 0.5$; (2) $\alpha = 0.002$, $m_4 = 83.6660$, $\gamma = 1$, $m_5 = 5$.

systems may be studied. As to the structure of the controller the frequency condition on its dynamic component may be further weakened at the expense of imposing some restrictions on the nonlinear amplifier characteristics. For example we may consider monotonous or slope restricted nonlinearities. In view of the fact that wide classes of neural networks have a state description of the form (16), (17), possible applications include neural control of nonlinear systems. Research will be conducted to incorporate the proposed theory in practical controller design procedures and to analyse specific applications.

References

[1] Krstić, M., Kanellakopoulos, I. and Kokotovic, P.V. *Nonlinear and Adaptive Control Design.* Wiley, New York, 1995.

[2] LaSalle, J.P. An invariance principle in the theory of stability. In: *Differential Equations and Dynamical Systems.* (Eds.: J.P. LaSalle and J.K. Hale), Academic Press, New York, 1967, 277–286.

[3] Loria, A., Kelly, R., Ortega, R. and Santibañez, V. On the output feedback control of Euler-Lagrange systems under input constraints. *IEEE Trans. on Automatic Control* **42** (1997) 1138–1143.

[4] Meirovitch, L. *Computational Methods in Structural Dynamics.* Sijthoff and Noordhof, Alphen aan den Rijn, The Netherlands, 1980.

[5] Noldus, E. and Loccufier, M. A comment on the method of the closest unstable equilibrium point in nonlinear stability analysis. *IEEE Trans. on Automatic Control* **40** (1995) 497–500.

[6] Ortega, R., Loria, A., Nicklasson, P.J. and Sira-Ramirez, H. *Passivity-based Control of Euler-Lagrange Systems.* Springer, Berlin, 1998.

[7] Preumont, A. *Vibration Control of Active Structures.* Kluwer, Dordrecht, 1997.

[8] Takegaki, M. and Arimoto, S. A new feedback method for dynamic control of manipulators. *ASME J. Dyn. Syst. Meas. Contr.* **103** (1981) 119–125.

[9] Tomei, P. A simple PD controller for robots with elastic joints. *IEEE Trans. on Automatic Control* **36** (1991) 1208–1213.

[10] Vidyasagar, M. *Nonlinear Systems Analysis.* Prentice-Hall, N.J., 1993.

[11] Willems, J.C. Dissipative dynamical systems. Part I: General theory. *Arch. Rat. Mech. and Analysis* **45** (1972) 321–351.

NONLINEAR DYNAMICS AND SYSTEMS THEORY
An International Journal of Research and Surveys

INSTRUCTIONS FOR CONTRIBUTORS

(1) General. This quarterly International Journal publishes original carefully refereed papers, brief notes and reviews on a wide range of nonlinear dynamics and systems theory problems. Contributions will be considered for publication in ND&ST if they have not been published previously.

(2) Manuscript and Correspondence. Contributions are welcomed from all countries and should be written in English. Two copies of the manuscripts, double spaced one column format and electronic version by AMSTEX, TEX or LATEX program (on diskette) should be sent directly to

Professor A.A. Martynyuk
Institute of Mechanics,
Nesterov str.3, 03057, MSP 680
Kiev-57, Ukraine
(e-mail: anmart@stability.kiev.ua).

The title of the article must include: author(s) name, name of institution, department, address, FAX, and e-mail; an Abstract of 50-100 words should not include any formulas and citations; key words, and AMS subject classifications number(s). The size for regular paper 10-14 pages, survey (up to 24 pages), short papers, letter to the editor and book reviews (2-3 pages).

(3) Tables, Graphs and Illustrations. All figures must be suitable for reproduction without retouched or redrawn and must include a title. Line drawings should include all relevant details and should be drawn in black ink on plain white drawing paper. In addition to a hard copy of the artwork, it is necessary to attach a PC diskette with files of the artwork (preferably in PCX format).

(4) References. Each entry must be cited in the text by author(s) and number or by number alone. All references should listed in their alphabetic order. Use please the following style:

Journal: [1] Poincaré, H. Title of the article. *Title of the Journal* **Vol.1**(No.1) (year) pages. [Language].

Book: [2] Liapunov, A.M. *Title of the book.* Name of the Publishers, Town, year.

Proceeding : [3] Bellman, R. Title of the article. In: *Title of the book.* (Eds.). Name of the Publishers, Town, year, pages. [Language].

(5) Proofs and Reprints. Proofs sent to authors should be returned to the Editor with corrections within three days after receipt. Acceptance of the paper entitles the author to 10 free reprints.

(6) Editorial Policy. Every paper is reviewed by the regional editor, and/or a referee (or referees), and it may be returned for revision or reject if considered unsuitable for publication.

(7) Copyright Assignment. When a paper is accepted for publication, author(s) will be requested to sign a form assigning copyright to Informath Publishing Group. Failure to do it promptly may delay publication.

NONLINEAR DYNAMICS AND SYSTEMS THEORY
An International Journal of Research and Surveys

Volume 1 Number 1 2001

CONTENTS

ISSN 1562-8353
© 2001, Informath Publishing Group
Printed in Ukraine