Lecture Notes in Electrical Engineering 343

Nikos Mastorakis
Aida Bulucea
George Tsekouras
Editors

Computational Problems in Science and Engineering

Springer

Lecture Notes in Electrical Engineering

Volume 343

More information about this series at http://www.springer.com/series/7818

About this Series

"Lecture Notes in Electrical Engineering (LNEE)" is a book series which reports the latest research and developments in Electrical Engineering, namely:

- Communication, Networks, and Information Theory
- Computer Engineering
- Signal, Image, Speech and Information Processing
- Circuits and Systems
- Bioengineering

LNEE publishes authored monographs and contributed volumes which present cutting edge research information as well as new perspectives on classical fields, while maintaining Springer's high standards of academic excellence. Also considered for publication are lecture materials, proceedings, and other related materials of exceptionally high quality and interest. The subject matter should be original and timely, reporting the latest research and developments in all areas of electrical engineering.

The audience for the books in LNEE consists of advanced level students, researchers, and industry professionals working at the forefront of their fields. Much like Springer's other Lecture Notes series, LNEE will be distributed through Springer's print and electronic publishing channels.

Nikos Mastorakis • Aida Bulucea
George Tsekouras
Editors

Computational Problems in Science and Engineering

Springer

Editors
Nikos Mastorakis
Technical University of Sofia
Sofia, Bulgaria

George Tsekouras
Naval Academy
Piraeus, Greece

Aida Bulucea
Departamentul de Inginerie Electrica
University of Craiova
Craiova, Romania

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-3-319-15764-1 ISBN 978-3-319-15765-8 (eBook)
DOI 10.1007/978-3-319-15765-8

Library of Congress Control Number: 2015942582

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Generalized Fuzzy Measurability

Anca Croitoru and Nikos Mastorakis

Abstract In this paper we introduce two concepts of generalized measurability for set-valued functions, namely φ-μ-total-measurability and φ-μ-measurability relative to a non-negative function $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty)$ and a non-negative set function $\mu : \mathcal{A} \to [0, +\infty)$ and present some relationships between them. We also define different types of convergences for sequences of set-valued functions and prove some relationships among them and a theorem of Egorov type. Finally, we introduce two semi-metrics on a space of set-valued functions and then compare them.

Keywords Measurable • Totally-measurable • Set-valued function • Theorem of Egorov type • Almost everywhere convergent • Pseudo-almost everywhere convergent • Almost uniformly convergent • Pseudo-almost uniformly convergent

1.1 Introduction

The theory of set-valued functions has interesting and important applications in many theoretical or practical domains such as economy, theory of control, statistics, fixed point theory, tochastic processes, information sciences, optimization (e.g., [1–32]). In the present work we focus on measurable set-valued functions. Thus we introduce two notions of generalized measurability, namely φ-μ-total-measurability and φ-μ-measurability relative to a non-negative function $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty)$ and a non-negative set function $\mu : \mathcal{A} \to [0, +\infty)$. These concepts generalize the classic definitions from the theory of measurable functions. This work is organized as follows: Sect. 1.1 is for introduction and some preliminaries. In Sect. 1.2 we present the two concepts of generalized measurability and give some

A. Croitoru (✉)
Faculty of Mathematics, "Al. I. Cuza" University, Iaşi, Romania
e-mail: croitoru@uaic.ro

N. Mastorakis
Technical University of Sofia, Sofia, Bulgaria
e-mail: mastor@tu-sofia.bg

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_1

relationships between them. In the last Sect. 1.3 we introduce different types of convergences for sequences of set-valued functions and prove some relationships among them. A theorem of Egorov type is also established. Moreover, we define two semi-metrics on a space of set-valued functions and obtain a comparative result between them.

Let T be a non-empty set, $\mathcal{A}$ a σ-algebra of subsets of T and $\mathcal{P}(T)$ the family of all subsets of T.

Definition 1. A set function $\mu : \mathcal{A} \to [0, +\infty]$ is called:

(i) *monotone* if $\mu(A) \leq \mu(B)$, for every $A, B \in \mathcal{A}$, with $A \subseteq B$.
(ii) *fuzzy* if μ is monotone and $\mu(\emptyset) = 0$.
(iii) *strongly order-continuous* (shortly *strongly o-continuous*) if $\lim_{n\to\infty} \mu(A_n) = 0$, for every $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ with $A_{n+1} \subseteq A_n$ for all $n \in \mathbb{N}$, $\bigcup_{n=0}^{\infty} A_n = A$, $A \in \mathcal{A}$ and $\mu(A) = 0$.
(iv) *autocontinuous from above* if $\lim_{n\to\infty} \mu(A \cup B_n) = \mu(A)$, for every $A \in \mathcal{A}$ and $(B_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ with $\lim_{n\to\infty} \mu(B_n) = 0$.
(v) *null-additive* if $\mu(A \cup B) = \mu(A)$ for every $A, B \in \mathcal{A}$ with $\mu(B) = 0$.
(vi) *decreasing continuous* if for every $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$, $A_0 \supseteq A_1 \supseteq A_2 \supseteq \ldots A_n \supseteq A_{n+1} \supseteq \ldots$ it results $\lim_{n\to\infty} \mu(A_n) = \mu(\bigcap_{n=0}^{\infty} A_n)$.
(vii) *increasing continuous* if for every $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$, $A_0 \subseteq A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \subseteq A_{n+1} \subseteq \ldots$ it results $\lim_{n\to\infty} \mu(A_n) = \mu(\bigcup_{n=0}^{\infty} A_n)$.
(viii) *continuous* if μ is both increasing continuous and decreasing continuous.

Definition 2.

(i) Consider a set function $\mu : \mathcal{A} \to [0, +\infty]$ or $\mu : \mathcal{A} \to X$, where $(X, \|\cdot\|)$ is a normed space. To μ we associate the set function $\overline{\mu} : \mathcal{P}(T) \to [0, +\infty]$ defined by $\overline{\mu}(E) = \sup\{\sum_{i=1}^{n} \|\mu(A_i)\|; A_i \subseteq E, A_i \in \mathcal{A}$ for any $i \in \{1, \ldots, n\}$, $A_i \cap A_j = \emptyset$ for $i \neq j\}$. $\overline{\mu}$ is called the variation of μ.
(ii) To a set function $\mu : \mathcal{A} \to [0, +\infty]$ we associate the set function $\mu^* : \mathcal{P}(T) \to [0, +\infty]$ called *the semivariation of* μ, defined by:

$$\mu^*(E) = \inf\{\mu(A) | E \subseteq A, A \in \mathcal{A}\}, \forall E \in \mathcal{P}(T).$$

(iii) Let $\mu : \mathcal{A} \to [0, +\infty]$ be a set function. We say that a property (P) about the points of T *holds almost everywhere* (denoted μ-a.e.) if there exists $A \in \mathcal{P}(T)$ so that $\mu^*(A) = 0$ and (P) holds on $T \setminus A$.

Remark 1. I. $\mu(A) \leq \overline{\mu}(A)$, $\forall A \in \mathcal{A}$.
II. $\overline{\mu}(A) = 0 \Rightarrow \mu(A) = 0$, $\forall A \in \mathcal{A}$.
III. $\overline{\mu}$ is monotone.
IV. If μ is fuzzy, then $\overline{\mu}$ is fuzzy.
V. $\overline{\mu}$ is superadditive on $\mathcal{P}(T)$ that is, $\overline{\mu}(A \cup B) \geq \overline{\mu}(A) + \overline{\mu}(B)$, $\forall A, B \in \mathcal{P}(T)$.
VI. μ^* is monotone and $\mu^*(\emptyset) = \inf_{A \in \mathcal{A}} \mu(A)$.
VII. If μ is monotone, then $\mu^*(A) = \mu(A)$ for every $A \in \mathcal{A}$.

Let X be a non-empty set and $\mathcal{P}_0(X)$ the family of non-empty subsets of X.

Definition 3. Let $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty]$ be a non-negative function.

(i) φ is called *symmetric* if $\varphi(C, D) = \varphi(D, C)$, for every $C, D \in \mathcal{P}_0(X)$.
(ii) We say that φ *satisfies the triangle inequality* (shortly TI) if $\varphi(C, D) \leq \varphi(C, E) + \varphi(E, D)$, for any $C, D, E \in \mathcal{P}_0(X)$.

Example 1. I. Let (X, d) be a metric space and $\varphi(A, B) = \inf_{\substack{x \in A \\ y \in B}} d(x, y)$, for every $A, B \in \mathcal{P}_0(X)$.

Then φ is symmetric and $\varphi(A, A) = 0$, for every $A \in \mathcal{P}_0(X)$.

II. Let (X, d) be a metric space and the Hausdorff metric

$$h(A, B) = \max\{\sup_{x \in A} d(x, B), \sup_{y \in B} d(y, A)\},$$

where $d(x, B) = \inf_{y \in B} d(x, y)$, for every $A, B \in \mathcal{P}_0(X)$. Then h is symmetric, h satisfies TI and $h(A, A) = 0$, for every $A \in \mathcal{P}_0(X)$.

III. Let $(X, \|\cdot\|)$ be a normed space and $\varphi(A, B) = \sup_{x \in A, y \in B} \|x - y\|$, for every $A, B \in \mathcal{P}_0(X)$. Then φ is symmetric, φ satisfies TI and $\varphi(A, A) = 0$, for every $A \in \mathcal{P}_0(X)$.

Definition 4. I. A function $f : T \to [0, +\infty)$ is called $\mathcal{A}$-*measurable* (shortly *measurable*) if $f^{-1}([\alpha, +\infty)) = \{t \in T; f(t) \geq \alpha\} \in \mathcal{A}$, for every $\alpha \in [0, +\infty)$.
II. If $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty]$ is a non-negative function and $F, G : T \to \mathcal{P}_0(X)$ are set-valued functions, then $\varphi(F, G)$ is called measurable if the function $t \to \varphi(F(t), G(t))$ is measurable.

Definition 5 ([36]). Let $\mathcal{A}$ be a ring of subsets of T, $\mu : \mathcal{A} \to [0, +\infty)$ a fuzzy measure and $f : T \to [0, +\infty)$ be a measurable function. *The Sugeno integral* of f on $A \in \mathcal{A}$ with respect to μ, denoted by $\int_A f d\mu$, is defined by $\int_A f d\mu = \sup_{\alpha \in [0, +\infty]} \min\{\alpha, \mu(A \cap f^{-1}([\alpha, +\infty)))\}$.

1.2 Generalized Fuzzy Measurability

In this section we introduce two concepts of generalized measurability for set-valued functions (particularly for functions) and present some relationships between them.

Let $T \neq \emptyset$, $\mathcal{A}$ a non-empty family of subsets of T and a set function $\mu : \mathcal{A} \to [0, +\infty]$. Suppose X is a non-empty set and $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty]$ is a non-negative function.

Definition 6.

(i) A finite family $\{A_i\}_{i=1}^n \subset \mathcal{A}$ is called an $\mathcal{A}$-*partition* of a set $B \in \mathcal{A}$ if $B = \bigcup_{i=1}^n A_i$ and $A_i \cap A_j = \emptyset$ for $i \neq j, i, j \in \{1, 2, \ldots, n\}$.

(ii) A set-valued function $F : T \to \mathcal{P}_0(X)$ is called $\mathcal{A}$-*simple* (shortly *simple*) if $F = \sum_{i=1}^n C_i \chi_{A_i}$, where $C_i \in \mathcal{P}_0(X)$, $C_i \neq C_j$ for $i \neq j, i, j \in \{1, \ldots, n\}$ and $\{A_i\}_{i=1}^n \subset \mathcal{A}\backslash\{\emptyset\}$ is an $\mathcal{A}$-partition of T (we denote by χ_{A_i} the characteristic function of A_i).

(iii) Let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}$. We say that (F_n) is φ-μ-convergent to F if for every $\varepsilon > 0$, we have $\lim_{n\to\infty} \mu^*(E_n(\varepsilon)) = 0$, where $E_n(\varepsilon) = \{t \in T | \varphi(F_n(t), F(t)) > \varepsilon\}$, for every $n \in \mathbb{N}$ and $\varepsilon > 0$. Denote this by $F_n \overset{\alpha-\mu}{\longrightarrow} F$.

(iv) A set-valued function $F : T \to \mathcal{P}_0(X)$ is called φ-μ-*totally measurable* (shortly μ-totally measurable) (on T) if there exists a sequence (F_n) of $\mathcal{A}$-simple set-valued functions such that $F_n \overset{\alpha-\mu}{\longrightarrow} F$.

(v) A set-valued function $F : T \to \mathcal{P}_0(X)$ is called φ-μ-*measurable* (on T) (shortly μ-*measurable*) if for every $\varepsilon > 0$, there is $\{A_i\}_{i=0}^n$ an $\mathcal{A}$-partition of T with $A_1, \ldots, A_n \in \mathcal{A}\backslash\{\emptyset\}$, so that $\mu(A_0) < \varepsilon$ and

$$\operatorname{osc}_\varphi(F, A_i) = \sup_{t,s\in A_i} \varphi(F(t), F(s)) < \varepsilon, \forall i \in \{1, \ldots, n\}.$$

Remark 2. Let $\mu : \mathcal{A} \to [0, +\infty]$ be a set function such that $\mu(\emptyset) = 0$ and let $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, \infty]$ so that $\varphi(C, C) = 0$, $\forall C \in \mathcal{P}_0(X)$. If $F : T \to \mathcal{P}_0(X)$ is simple, then F is μ-measurable. Indeed, let $F = \sum_{i=1}^n C_i \cdot \chi_{A_i}$, $\{C_i\}_{i=1}^n \subset \mathcal{P}_0(X)$, $\{A_i\}_{i=1}^n$ is an $\mathcal{A}$-partition of T. Considering now the partition $\{A_0, A_1, \ldots, A_n\}$, with $A_0 = \emptyset$, we have $\mu(A_0) = 0$ and

$$\operatorname{osc}_\varphi(F, A_i) = \sup_{t,s\in A_i} \varphi(F(t), F(s)) = 0 < \varepsilon, \quad \forall i \in \{1, \ldots, n\}.$$

So F is μ-measurable.

Proposition 1. *Suppose $\mu : \mathcal{A} \to [0, +\infty]$ is null-additive and $\mu(\emptyset) = 0$. If $F, G : T \to \mathcal{P}_0(X)$ are set-valued functions so that F is μ-measurable and $F = G$ μ-a.e., then G is μ-measurable.*

Proof. Let $C = \{t \in T | F(t) \neq G(t)\}$. Since $F = G\mu$-a.e., it follows $C \in \mathcal{A}$ and $\mu(C) = 0$.

Let be $\varepsilon > 0$. Since F is μ-measurable, there is $\{A_i\}_{i=1}^n$ an $\mathcal{A}$-partition of T such that $\mu(A_0) < \varepsilon$ and

$$\operatorname{osc}_\varphi(F, A_i) < \varepsilon, \quad \forall i \in \{1, \ldots, n\}. \tag{1.1}$$

Now, consider the following $\mathcal{A}$-partition of T : $B_0 = A_0 \cup C$, $B_i = A_i \cap (T \backslash C)$, for every $i \in \{1, \ldots, n\}$. By the null-additivity of μ, we have $\mu(B_0) = \mu(A_0) < \varepsilon$ and from (1.1) we obtain

$$\operatorname{osc}_\varphi(G, B_i) = \sup_{t,s \in B_i} \varphi(G(t), G(s)) \leq \operatorname{osc}_\varphi(F, A_i) < \varepsilon, \quad \forall i \in \{1, \ldots, n\}.$$

Thus G is μ-measurable. □

Theorem 1. *Suppose $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty]$ satisfies TI and $\mathcal{A}$ is an algebra of subsets of T. Let $\mu : \mathcal{A} \to [0, +\infty]$ be a set function and $F : T \to \mathcal{P}_0(X)$ a set-valued function. Then F is μ-totally-measurable if and only if F is μ-measurable. Moreover, if μ is monotone and F is μ-measurable on T, then F is μ-measurable on every set $B \in \mathcal{A}$.*

Proof. Suppose F is μ-totally measurable and let $\varepsilon > 0$. Then there exists a sequence $(F_n)_{n \in \mathbb{N}}$ of simple set-valued functions so that $F_n \xrightarrow{\alpha-\mu} F$. Then there is $n_0 \in \mathbb{N}$ such that

$$\mu^*(E_{n_0}(\frac{\varepsilon}{3})) < \frac{\varepsilon}{3},$$

where $E_{n_0}(\frac{\varepsilon}{3}) = \{t \in T | \varphi(F_n(t), F(t)) > \frac{\varepsilon}{3}\}$.

According to Definition 2 -(ii), there is $A_0 \in \mathcal{A}$, such that $E_{n_0}(\frac{\varepsilon}{3}) \subseteq A_0$ and

$$\mu^*(E_{n_0}(\frac{\varepsilon}{3})) \leq \mu(A_0) < \frac{\varepsilon}{3}. \tag{1.2}$$

Suppose $F_{n_0} = \sum_{i=1}^{p_{n_0}} C_i^{p_{n_0}} \chi_{B_i^{n_0}}$, where $C_i^{n_0} \in \mathcal{P}_0(X)$ for every $i \in \{1, \ldots, p_{n_0}\}$ and $(B_i^{n_0})_{i=1}^{p_{n_0}}$ is an $\mathcal{A}$-partition of T.

Now, if $t, s \in T \backslash A_0$, then

$$\varphi(F_{n_0}(t), F(t)) \leq \frac{\varepsilon}{3}. \tag{1.3}$$

We can write $T \backslash A_0 = \bigcup_{i=1}^{p_{n_0}} A_i$, where $A_i = B_i^{n_0} \cap (T \backslash A_0)$, $\forall i \in \{1, \ldots, p_{n_0}\}$. We remark that the sets $(A_i)_{i=0}^{p_{n_0}}$ are mutually disjoint. Let $i \in \{1, \ldots, p_{n_0}\}$ and

$t, s \in A_i$. It follows $F_{n_0}(t) = F_{n_0}(s) = C_i^{n_0}$ and from (1.3) we have:

$$\varphi(F(t), F(s)) \leq \varphi(F(t), C_i^{n_0}) + \varphi(C_i^{n_0}, F(s)) =$$

$$= \varphi(F(t), F_{n_0}(t) + \varphi(F_{n_0}(s), F(s)) \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \frac{2\varepsilon}{3},$$

that implies $\sup_{t,s \in A_i} \varphi(F(t), F(s)) \leq \frac{2\varepsilon}{3} < \varepsilon$. So for every $\varepsilon > 0$, there is $(A_i)_{i=0}^{p_{n_0}}$ an $\mathcal{A}$-partition of T, such that $\mu(A_0) < \varepsilon$ (due to (1.2)) and $\text{osc}_\varphi(F, A_i) < \varepsilon$, for every $i \in \{1, \ldots, p_{n_0}\}$, which shows that F is μ-measurable.

Suppose now F is μ-measurable. Considering $\varepsilon = \frac{1}{n}$, $n \in \mathbb{N}^*$, it results there is $\{A_i^n\}_{i=0}^{p_n}$, an $\mathcal{A}$-partition of T, such that:

$$\mu(A_0^n) < \frac{1}{n} \text{ and} \tag{1.4}$$

$$\text{osc}_\varphi(F, A_i^n) < \frac{1}{n}, \forall i \in \{1, \ldots, p_n\}. \tag{1.5}$$

For every $i \in \{0, 1, 2, \ldots, p_n\}$ let $t_i^n \in A_i^n$ and define the simple set-valued function $F_n = \sum_{i=0}^{p_n} F(t_i^n)\chi_{A_i^n}$. Now, let $\varepsilon > 0$. Then there is $n_0 \in \mathbb{N}^*$ so that $\frac{1}{n} < \varepsilon$, for every $n \geq n_0$. Denote $E_n(\varepsilon) = \{t \in T | \varphi(F_n(t), F(t)) > \varepsilon\}$, $\varepsilon > 0$, $n \in \mathbb{N}^*$. For every $t \in \bigcup_{i=1}^{p_n} A_i^n$, there exists $i_0 \in \{1, \ldots, p_n\}$ so that $t \in A_{i_0}^n$. From (1.5) it results:

$$\varphi(F_n(t), F(t)) = \varphi(F(t_{i_0}^n), F(t)) \leq \text{osc}_\varphi(F, A_{i_0}^n) < \frac{1}{n} < \varepsilon.$$

So $\bigcup_{i=1}^{p_n} A_i^n \subseteq T \backslash E_n(\varepsilon)$, which is equivalent to $E_n(\varepsilon) \subseteq A_0^n$. From (1.4) and from the monotonicity of μ^*, we obtain $\mu^*(E_n(\varepsilon)) < \frac{1}{n}$, for every $n \in \mathbb{N}$, $n \geq n_0$, which shows that $\lim_{n\to\infty} \mu^*(E_n(\varepsilon)) = 0$. So, there is $(F_n)_{n\in\mathbb{N}}$, a sequence of simple set-valued functions so that $F_n \xrightarrow{\varphi-\mu} F$. In other words, F is μ-totally-measurable.

Suppose now μ is monotone. Let F be μ-measurable on T and $\varepsilon > 0$. Then there exists $\{A_i\}_{i=0}^n$ an $\mathcal{A}$-partition of T, so that $\mu(A_0) < \varepsilon$ and $\text{osc}_\varphi(F, A_i) < \varepsilon$, $\forall i \in \{1, \ldots, n\}$. Let $C_i = A_i \cap B$, for every $i \in \{0, 1, \ldots, n\}$. It results $\text{osc}_\varphi(F, C_i) \leq \text{osc}_\varphi(F, A_i) < \varepsilon$, for every $i \in \{1, \ldots, n\}$ and from the monotonicity of μ we obtain $\mu(C_0) \leq \mu(A_0) < \varepsilon$. So F is μ-measurable on B. □

Remark 3. We can obtain similar measurability results for functions $f : T \to X$, considering a non-negative function $\varphi : X \times X \to [0, +\infty]$.

1.3 Fuzzy Convergences

In this section we define different types of convergences for sequences of set-valued functions, prove some relationships among them and a theorem of Egorov type. We also introduce two semi-metrics on a space of set-valued functions and obtain a comparison between them.

Let be $\emptyset \neq \mathcal{A} \subseteq \mathcal{P}(T)$, $\mu : \mathcal{A} \to [0, +\infty]$ a non-negative set function and X a non-empty set.

Definition 7. Let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}$.

I. The sequence of multifunctions $(F_n)_{n\in\mathbb{N}}$ is called *φ-almost everywhere convergent* to F (denoted by $F_n \overset{\varphi-ae}{\to} F$) if $\{t \in T; \varphi(F_n(t), F(t)) \nrightarrow 0\} \in \mathcal{A}$ and

$$\mu(\{t \in T; \varphi(F_n(t), F(t)) \nrightarrow 0\}) = 0.$$

II. The sequence $(F_n)_{n\in\mathbb{N}}$ is called *pseudo-almost everywhere convergent* to F (denoted by $F_n \overset{\varphi-pae}{\longrightarrow} F$) if there exists $A \in \mathcal{A}$ with $\mu(T\backslash A) = \mu(T)$ and $\varphi(F_n(t), F(t)) \to 0$ for every $t \in T\backslash A$.

III. We say that (F_n) *converges uniformly* to F on $A \in \mathcal{A}$ (denoted $F_n \underset{A}{\overset{u}{\to}} F$) if $\varphi(F_n, F) \underset{A}{\overset{u}{\to}} 0$ (i.e. $\forall \varepsilon > 0,\ \exists n_\varepsilon \in \mathbb{N}$ such that $\varphi(F_n(t), F(t)) < \varepsilon,\ \forall t \in A$).

IV. (F_n) is called *φ-almost uniformly convergent* to F (denoted $F_n \overset{\varphi-au}{\to} F$) if there is $(A_p)_{p\in\mathbb{N}} \subset \mathcal{A}$, with $\lim\limits_{p\to\infty} \mu(A_p) = 0$ such that $F_n \xrightarrow[T\backslash A_p]{u} F$, for every fixed $p \in \mathbb{N}$.

V. (F_n) is called *φ-pseudo-almost uniformly convergent* to F (denoted $F_n \overset{pau}{\to} F$) if there is $(A_p)_{p\in\mathbb{N}} \subset \mathcal{A}$, with $\lim\limits_{p\to\infty} \mu(T\backslash A_p) = \mu(T)$ such that $F_n \xrightarrow[T\backslash A_p]{u} F$, for every fixed $p \in \mathbb{N}$.

In the sequel, we outline some relationships among the convergences introduced in Definition 7.

Theorem 2. *Suppose $\mu : \mathcal{A} \to [0, +\infty]$ is null-additive and let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}$. If $F_n \overset{\varphi-ae}{\to} F$, then $F_n \overset{\varphi-pae}{\to} F$.*

Proof. Denote $A = \{t \in T; \varphi(F_n(t), F(t)) \nrightarrow 0\} \in \mathcal{A}$. Since $F_n \overset{\varphi-ae}{\to} F$, we have $\mu(A) = 0$. Since μ is null-additive, we obtain $\mu(T) = \mu((T\backslash A) \cup A) = \mu(T\backslash A)$ and it is obvious that $\varphi(F_n(t), F(t)) \to 0$ for every $t \in T\backslash A$. So, $F_n \overset{\varphi-pae}{\to} F$. □

Theorem 3. *Suppose $\mu : \mathcal{A} \to [0, +\infty)$ is auto-continuous from below and let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}$. If $F_n \overset{\varphi-au}{\to} F$, then $F_n \overset{\varphi-pau}{\to} F$.*

Proof. Since $F_n \overset{\varphi-au}{\to} F$, there is $(A_p)_{p\in\mathbb{N}} \subset \mathcal{A}$ with $\lim_{p\to\infty} \mu(A_p) = 0$ such that $F_n \xrightarrow[T\setminus A_p]{u} F$, for every fixed $p \in \mathbb{N}$. Since μ is autocontinuous from below, we obtain

$$\lim_{p\to\infty} \mu(T\setminus A_p) = \mu(T).$$

So, $F_n \overset{\varphi-pau}{\to} F$. □

Theorem 4. *Suppose $\mathcal{A}$ is a σ-algebra, $\mu : \mathcal{A} \to [0,\infty]$ is a fuzzy measure and let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}$.*

I. If $F_n \overset{\varphi-au}{\to} F$, then $F_n \overset{\varphi-ae}{\to} F$.
II. If $F_n \overset{\varphi-pau}{\to} F$, then $F_n \overset{\varphi-pae}{\to} F$.

Proof. I. Since $F_n \overset{\varphi-au}{\to} F$, there is $(A_p)_{p\in\mathbb{N}} \subset \mathcal{A}$, with $\lim_{p\to\infty} \mu(A_p) = 0$ such that $F_n \xrightarrow[T\setminus A_p]{u} F$, for every fixed $p \in \mathbb{N}$. Considering $B = \bigcap_{p=1}^{\infty} A_p$, since μ is fuzzy we obtain $\mu(B) = 0$. Now, for every $t \in T\setminus B = \bigcup_{p=1}^{\infty} (T\setminus A_p)$, there exists $p \in \mathbb{N}^*$, so that $t \in T\setminus A_p$. Consequently, $\varphi(F_n(t), F(t)) \to 0$. Thus, $F_n \overset{\varphi-ae}{\to} F$.
II. We proceed similarly to I. □

Theorem 5 (Egorov Type). *Suppose $\mathcal{A}$ is a σ-algebra, $\mu : \mathcal{A} \to [0,+\infty)$ is fuzzy continuous and let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}$. If $\varphi(F_n(t), F(t)) \to 0$, for every $t \in T$, then $F_n \overset{\varphi-au}{\to} F$ and $F_n \overset{\varphi-pau}{\to} F$.*

Proof. Let $A_n^p = \bigcap_{k=n}^{\infty} \{t \in T; \varphi(F_k(t), F(t)) < \frac{1}{p}\}$, for every $p \in \mathbb{N}^*$. Then $A_n^p \subseteq A_{n+1}^p$, for every $p, n \in \mathbb{N}^*$ and $T = \bigcap_{p=1}^{\infty} \bigcup_{n=1}^{\infty} A_n^p$. It results $\lim_{n\to\infty} \mu(T\setminus A_n^p) = \emptyset$, for every fixed $p \in \mathbb{N}^*$. Let $\varepsilon > 0$ be arbitrarily. Since μ is continuous from above, there exists $n_1 \in \mathbb{N}$ so that $\mu(T\setminus A_{n_1}^1) < \frac{\varepsilon}{2}$. Now there is $n_2 \in \mathbb{N}$, $n_1 < n_2$, so that $\mu((T\setminus A_{n_1}^1) \cup (T\setminus A_{n_2}^2)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2^2}$. Consequently, there exists $(A_{n_k}^k)_{k\in\mathbb{N}}$ so that $\mu(\bigcup_{k=1}^{p} (T\setminus A_{n_k}^k)) < \sum_{k=1}^{p} \frac{\varepsilon}{2^k} = (1 - \frac{1}{2^p})\varepsilon < \varepsilon$. Since μ is continuous from below, we have $\mu(\bigcup_{k=1}^{\infty} (T\setminus A_{n_k}^k)) < \varepsilon$. We have to prove that $F_n \xrightarrow[\bigcap_{k=1}^{\infty} A_{n_k}^k]{u} F$. For every $\delta > 0$ there exists $k_0 \in \mathbb{N}^*$ so that $\frac{1}{k_0} < \delta$. For every $t \in \bigcap_{k=1}^{\infty} A_{n_k}^k$, we have:

$$\varphi(F_k(t), F(t)) < \frac{1}{k_0} < \delta, \forall k \geq n_{k_0}.$$

So, $F_n \overset{\varphi-au}{\to} F$. The fact that $F_n \overset{\varphi-pau}{\to} F$ follows in a similar way. □

Corollary 1. *Let be $\mathcal{A}$ a σ-algebra, $\mu : \mathcal{A} \to [0, \infty)$ a null-additive continuous fuzzy measure and $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}$. If $F_n \overset{\varphi-ae}{\to} F$, then $F_n \overset{\varphi-au}{\to} F$ and $F_n \overset{\varphi-pau}{\to} F$.*

Theorem 6. *Let $\mu : \mathcal{A} \to [0, +\infty]$ be so that μ^* is strongly o-continuous and let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}^*$, such that $F_n \overset{\varphi-ae}{\longrightarrow} F$. Then $F_n \overset{\varphi-\mu}{\longrightarrow} F$.*

Proof. If $F_n \overset{\varphi-ae}{\longrightarrow} F$, then $\mu^*(E) = 0$, where $E = \{t \in T; \varphi(F_n(t), F(t)) \nrightarrow 0\}$. We can write $E = \bigcup_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{t \in T; \varphi(F_k(t), F(t)) \geq \frac{1}{m}\}$.

Denoting

$$E_n^m = \bigcup_{k=n}^{\infty} \{t \in T; \varphi(F_k(t), F(t)) \geq \frac{1}{m}\}$$

and $E^m = \bigcap_{n=1}^{\infty} E_n^m$, for every $m \in \mathbb{N}^*$, it follows $\mu^*(E^m) = 0$ and $E_n^m \searrow E^m$. Since μ^* is strongly o-continuous, $\lim_{n\to\infty} \mu^*(E_n^m) = 0$, for every $m \in \mathbb{N}^*$.

But

$$0 \leq \mu^*(\{t \in T; \varphi(F_n(t), F(t)) \geq \frac{1}{m}\}) \leq \mu^*(E_n^m),$$

for every $m, n \in \mathbb{N}^*$. This implies $\lim_{n\to\infty} \mu^*(\{t \in T; \varphi(F_n(t), F(t)) \geq \frac{1}{m}\}) = 0$, for every $m \in \mathbb{N}^*$, which shows that $F_n \overset{\varphi-\mu}{\longrightarrow} F$. □

Definition 8. Let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}^*$. Suppose $\varphi(F_n, F)$ is measurable for every $n \in \mathbb{N}^*$. The sequence $(F_n)_{n\in\mathbb{N}}$ is called *φ-mean convergent* to F (denoted by $F_n \overset{\varphi-m}{\longrightarrow} F$) if $\lim_{n\to\infty} \int_T \varphi(F_n, F) d\mu = 0$.

Theorem 7. *Let $F, F_n : T \to \mathcal{P}_0(X)$, $n \in \mathbb{N}^*$, so that $\varphi(F_n, F)$ is measurable for every $n \in \mathbb{N}^*$. Then $F_n \overset{\varphi-\mu}{\longrightarrow} F$ if and only if $F_n \overset{\varphi-m}{\longrightarrow} F$.*

Proof. Suppose $F_n \overset{\varphi-\mu}{\longrightarrow} F$. Then for every $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ so that

$$\mu(\{t \in T; \varphi(F_n(t), F(t)) \geq \frac{\varepsilon}{2}\}) < \varepsilon,$$

for every $n \geq n_0$. By Lemma 7.5 [32], it results $(S) \int_T \varphi(F_n, F) d\mu < \varepsilon$, which shows that $F_n \overset{\varphi-m}{\longrightarrow} F$.

Conversely, suppose $F_n \xrightarrow{\varphi-m} F$ and $F_n \overset{\varphi-\mu}{\nrightarrow} F$. Then there exist $\varepsilon > 0,\ \delta > 0$ and a sequence $(k_n)_{n\in\mathbb{N}^*} \subset \mathbb{N}$ so that

$$\mu(\{t \in T; \varphi(F_{k_n}(t), F(t)) \geq \varepsilon\}) > \delta$$

for every $n \in \mathbb{N}^*$. Now, we obtain:

$$\begin{aligned}
&(S)\int_T \varphi(F_{k_n}, F)d\mu = \\
&= \sup_{\alpha\in[0,+\infty]} \min\{\alpha, \mu(\varphi^{-1}([\alpha, +\infty)))\} \geq \\
&\geq \min\{\varepsilon, \mu(\{t \in T; \varphi(F_{k_n}(t), F(t)) \geq \varepsilon\})\} \geq \\
&\geq \min\{\varepsilon, \delta\} > 0, \forall n \in \mathbb{N}^*,
\end{aligned}$$

which contradicts the fact that $F_n \xrightarrow{\varphi-m} F$. □

In the sequel, T and X are non-empty sets, $\mathcal{A}$ is a σ-algebra of subsets of T, $\mu : \mathcal{A} \to [0, +\infty]$ is a fuzzy measure and $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty]$ is a non-negative function.

Let $\mathcal{F} = \{F | F : T \to \mathcal{P}_0(X)\}$ and $\mathcal{M} \subseteq \mathcal{F}$ satisfying the following properties:

$(*)$ $\varphi(F(t), G(t)) < +\infty,\ \forall F, G \in \mathcal{M}, \forall t \in T$;
$(**)$ $\varphi(F, G)$ is measurable, $\forall F, G \in \mathcal{M}$.

Denote $d_\varphi(F, G) = \int_T \frac{\varphi(F,G)}{1+\varphi(F,G)} d\mu$, for every $F, G \in \mathcal{M}$ and $B(F, \varepsilon) = \{G \in \mathcal{M}; d_\varphi(G, F) < \varepsilon\} \cup \{F\}$, for every $F \in \mathcal{M}$. We shortly denote $d_\varphi(F, G)$ by $d(F, G)$, $\forall F, G \in \mathcal{M}$.

Let $F, F_n \in \mathcal{M}, n \in \mathbb{N}$. We denote $F_n \xrightarrow{d} F$ iff $\lim_{n\to\infty} d(F_n, F) = 0$.

Remark 4. I. $d(F, G) \geq 0,\ \forall F, G \in \mathcal{M}$.
II. If F and G are measurable functions, then d is a semi-metric.
III. If φ is symmetric, then d is symmetric.
IV. If φ satisfies TI, then $d(F, H) \leq d(F, G) + d(G, H)$, for every $F, G, H \in \mathcal{M}$.

Example 2. Suppose $X = \mathbb{R}$, φ is the Hausdorff metric on $\mathcal{P}_0(\mathbb{R})$ and $\mathcal{M} = \{F | F : T \to \mathcal{P}_0(\mathbb{R}),\ F(t) = [f(t), g(t)],\ \forall t \in T,\ f, g : T \to \mathbb{R}$ are measurable$\}$. Then $\mathcal{M}$ satisfies the above conditions $(*)$ and $(**)$.

Following Wu et al. [34], we obtain the next results:

Theorem 8. *Let $F, F_n \in \mathcal{M}, n \in \mathbb{N}$. Then $F_n \xrightarrow{d} F \Leftrightarrow F_n \overset{\varphi-\mu}{\to} F$.*

Proof. Suppose $\lim_{n\to\infty} d(F_n, F) = 0$ and let $\varepsilon > 0$. Then there is $n_0 \in \mathbb{N}$ so that

$$d(F_n, F) < \frac{\varepsilon}{1+\varepsilon}, \quad \forall n \geq n_0. \tag{1.6}$$

Since the function $u : [0, +\infty) \to \mathbb{R}$, $u(x) = \frac{x}{1+x}$, is increasing and μ is fuzzy, we have

$$\mu(\{t \in T; \varphi(F_n(t), F(t)) \geq \varepsilon\}) \leq x_n, \forall n \in \mathbb{N},$$

where $x_n = \mu(\{x \in T; \frac{\varphi(F_n(t),F(t))}{1+\varphi(F_n(t),F(t))} \geq \frac{\varepsilon}{1+\varepsilon}\})$. Suppose there exists $n \geq n_0$ such that $x_n > \frac{\varepsilon}{1+\varepsilon}$. Then

$$d(F_n, F) = \int_T \frac{\varphi(F_n, F)}{1 + \varphi(F_n, F)} d\mu \geq \min\{\frac{\varepsilon}{1+\varepsilon}, x_n\} = \frac{\varepsilon}{1+\varepsilon}$$

which contradicts (1.6). It follows that $x_n \leq \frac{\varepsilon}{1+\varepsilon}$ for every $n \geq n_0$. Now, we have

$$\begin{aligned} &\mu(\{t \in T; \varphi(F_n(t), F(t)) \geq \varepsilon\}) \leq x_n = \min\{\frac{\varepsilon}{1+\varepsilon}, x_n\} \leq \\ &\leq \sup_{\alpha \in [0,\infty)} \min\{\alpha, \mu(\{t \in T; \frac{\varphi(F_n(t), F(t))}{1 + \varphi(F_n(t), F(t))} \geq \alpha\})\} = \\ &= \int_T \frac{\varphi(F_n, F)}{1 + \varphi(F_n, F)} d\mu = d(F_n, F), \quad \forall n \geq n_0. \end{aligned} \tag{1.7}$$

By (1.7) and the fact that $\lim_{n\to\infty} d(F_n, F) = 0$, it results $F_n \overset{\varphi-\mu}{\to} F$. Now, suppose $F_n \overset{\varphi-\mu}{\to} F$ and denote $f(t) = \frac{\varphi(F_n(t),F(t))}{1+\varphi(F_n(t),F(t))}$ for every $t \in T$. Then for every $\varepsilon > 0$ there is $n_0 \in \mathbb{N}$ so that

$$\mu(\{t \in T; \varphi(F_n(t), F(t)) \geq \varepsilon\}) < \varepsilon, \quad \forall n \geq n_0. \tag{1.8}$$

Since μ is fuzzy and by (1.8) we obtain

$$\begin{aligned} d(F_n, F) &= \int_T \frac{\varphi(F_n, F)}{1 + \varphi(F_n, F)} d\mu = \sup_{\alpha \in [0,\infty)} \min\{\alpha, \mu(\{t \in T; f(t) \geq \alpha\})\} = \\ &= \max\{\sup_{0 \leq \alpha \leq \varepsilon} \min\{\alpha, \mu(\{t \in T; f(t) \geq \alpha\})\}, \sup_{\alpha > \varepsilon} \min\{\alpha, \mu(\{t \in T; f(t) \geq \alpha\})\}\} \leq \\ &\leq \max\{\varepsilon, \sup_{\alpha > \varepsilon} \min\{\alpha, \mu(\{t \in T; f(t) \geq \varepsilon\})\}\} \leq \\ &\leq \max\{\varepsilon, \sup_{\alpha > \varepsilon} \min\{\alpha, \mu(\{t \in T; \varphi(F_n(t), F(t)) \geq \varepsilon\})\}\} = \varepsilon, \ \forall n \geq n_0 \end{aligned}$$

that is, $F_n \overset{d}{\to} F$.

Theorem 9. *If μ satisfies the condition*

$$\forall m \in \mathbb{N}, m > 1, \forall A \in \mathcal{A} \text{with} \mu(A) < \frac{1}{m}, \forall (B_n) \subset \mathcal{A}, \mu(B_n) \to 0 \Rightarrow \tag{1.9}$$

$$\exists n_0 \in \mathbb{N} \text{s.t.} \mu(A \cup B_n) < \frac{1}{m}, \forall n \geq n_0,$$

then there is a unique topology τ on $\mathcal{M}$, defined by the neighborhood system $\mathcal{V}(F) = \{V \in \mathcal{P}(\mathcal{M}) | \exists n \in \mathbb{N}^*$ so that $B(F, \frac{1}{n}) \subseteq V\}$ for every $F \in \mathcal{M}$.

Proof. Let $F \in \mathcal{M}$. From the definitions it results:

(i) $F \in V, \forall V \in \mathcal{V}(F)$.

(ii) For every $V \in \mathcal{V}(F)$, we have $U \in \mathcal{V}(F)$ for every $U \supset V$.

(iii) Let $V_1, V_2 \in \mathcal{V}(F)$. Then there exist $n_1, n_2 \in \mathbb{N}$ so that $B(F, \frac{1}{n_1}) \subseteq V_1$ and $B(F, \frac{1}{n_2}) \subseteq V_2$. Taking $n = \max\{n_1, n_2\}$, we have

$$B(F, \frac{1}{n}) \subseteq B(F, \frac{1}{n_1}) \cap B(F, \frac{1}{n_2}) \subseteq V_1 \cap V_2$$

that is, $V_1 \cap V_2 \in \mathcal{V}(F)$.

(iv) Let $V \in \mathcal{V}(F)$. Then there is $n \in \mathbb{N}$ so that $B(F, \frac{1}{n}) \subseteq V$. Taking $W = B(F, \frac{1}{n})$, it follows that $W \in \mathcal{V}(F)$ and we have to show that $W \in \mathcal{V}(G)$, for every $G \in W$. In order to show this, we prove

$$\forall p \in \mathbb{N}^*, \forall G \in B(F, \frac{1}{p}), \exists n_0 \in \mathbb{N}^* \text{ suchthat } B(G, \frac{1}{n_0}) \subseteq B(F, \frac{1}{p}). \quad (1.10)$$

Suppose by contrary that there are $p_0 \in \mathbb{N}^*$, $G_0 \in B(F, \frac{1}{p_0})$ and for every $n \in \mathbb{N}^*$, there is $G_n \in B(G_0, \frac{1}{n})$ so that

$$G_n \notin B(F, \frac{1}{p_0}). \quad (1.11)$$

Since $G_n \in B(G_0, \frac{1}{n})$, it results $d(G_n, G_0) < \frac{1}{n}$, for every $n \in \mathbb{N}^*$. This implies $G_n \xrightarrow{d} G$ and by Theorem 8, it follows that $G_n \xrightarrow{\varphi-\mu} G$. That is, $\lim_{n\to\infty} \mu(B_n) = 0$, where $B_n = \{t \in T; \varphi(G_n(t), G_0(t)) \geq b\}$, $\forall n \in \mathbb{N}^*$ and $b = \frac{1}{3p_0} - \frac{1}{3} d(G_0, F)$. Denote $A = \{t \in T; \frac{\varphi(G_0(t), F(t))}{1+\varphi(G_0(t), F(t))} \geq \frac{1}{p_0} - 2b\}$. Notice that $0 < b < \frac{1}{p_0}$ and $\frac{1}{p_0} - 2b > d(G_0, F) \geq 0$. Thus we obtain

$$\begin{aligned} d(G_0, F) &= \int_T \frac{\varphi(G_0, F)}{1 + \varphi(G_0, F)} d\mu = \\ &= \sup_{\alpha \geq 0} \min\{\alpha, \mu(t \in T; \frac{\varphi(G_0(t), F(t))}{1 + \varphi(G_0(t), F(t))} \geq \alpha)\} \geq \quad (1.12) \\ &\geq \min\{\frac{1}{p_0} - 2b, \mu(A)\} = \mu(A). \end{aligned}$$

From (1.12) it results $\mu(A) < \frac{1}{p_0}$.

Denote $f(t) = \varphi(G_n(t), G_0(t)) + \frac{\varphi(G_0(t),F(t))}{1+\varphi(G_0(t),F(t))}$, for every $t \in T$ and $x_n = \mu(\{t \in T; f(t) \geq \frac{1}{p_0} - b\})$, for every $n \in \mathbb{N}$. Since μ is fuzzy and by (1.9), it results that there exists $n_0 \in \mathbb{N}$ so that

$$x_n \leq \mu(A \cup B_n) < \frac{1}{p_0}, \quad \forall n \geq n_0. \tag{1.13}$$

From (1.13) we have

$$d(G_n, F) = \int_T \frac{\varphi(G_n, F)}{1 + \varphi(G_n, F)} d\mu =$$

$$= \sup_{0\leq\alpha<1} \min\{\alpha, \mu(\{t \in T; \frac{\varphi(G_n(t), F(t))}{1 + \varphi(G_n(t), F(t))} \geq \alpha\})\} \leq$$

$$\leq \sup_{0\leq\alpha<1} \min\{\alpha, x_n\} = \max\{\sup_{0\leq\alpha<\frac{1}{p_0}-b} \min\{\alpha, x_n\}, \sup_{\frac{1}{p_0}-b\leq\alpha<1} \min\{\alpha, x_n\}\} \leq$$

$$\leq \max\{\frac{1}{p_0} - b, \sup_{\frac{1}{p_0}-b\leq\alpha<1} \min\{\alpha, x_n\}\} \leq \max\{\frac{1}{p_0} - b, x_n\} < \frac{1}{p_0}, \ \forall n \geq n_0.$$

So $d(G_n, F) < \frac{1}{p_0}$, for every $n \geq n_0$, that contradicts (1.11).

So (1.10) hold. Now, for n and G, according to (1.10), there is $n_0 \in \mathbb{N}^*$ such that

$$B(G, \frac{1}{n_0}) \subseteq B(F, \frac{1}{n}) = W,$$

that is $W \in \mathcal{V}(G)$ and the theorem is proved. □

Remark 5. If μ is autocontinuous from above, then μ satisfies (1.9).

In the sequel we will introduce another semi-metric on the space $\mathcal{M}$.

Definition 9. For every $F, G \in \mathcal{M}$, let $D(F, G)$ be defined by

$$D(F, G) = \inf\{\varepsilon > 0 | \mu(\{t \in T; \varphi(F(t), G(t)) > \varepsilon\}) \leq \varepsilon\}.$$

I. $0 \leq D(F, G) \leq \mu(T), \forall F, G \in \mathcal{M}$.
II. $\mu(\{t \in T; \varphi(F, G) > D(F, G)\}) \leq D(F, G), \forall F, G \in \mathcal{M}$.

In the next theorem we present some properties of D, including a comparison with d, following Florescu [35].

Theorem 10. *The application* $D : \mathcal{M} \times \mathcal{M} \to [0, \infty]$ *has the following properties:*

(i) if φ *is symmetric, then* D *is symmetric;*
(ii) if φ *satisfies TI and* μ *is subadditive, then* D *also satisfies TI;*
(iii) if $\mu : \mathcal{A} \to [0, 1]$ *is a continuous fuzzy measure, then* $D^2 \leq (1 + \mu(T))d$.

Proof. (i) It results from the definitions.

(ii) Let be $F, G, H \in \mathcal{M}$ and every $\alpha < D(F, G) + D(G, H)$. Then there exist $\beta,\ \gamma$ such that $\alpha = \beta + \gamma,\ \beta > D(F, G), \gamma > D(G, H)$. Let be $0 < \varepsilon < \beta,\ 0 < \delta < \gamma$ so that:

$$\mu(\{t \in T; \varphi(F(t), G(t)) > \varepsilon\}) \leq \varepsilon \text{ and } \mu(\{t \in T; \varphi(G(t), H(t)) > \delta\}) \leq \delta.$$

Since φ satisfies TI and μ is subadditive we obtain $\mu(\{t \in T; \varphi(F(t), H(T)) > \varepsilon + \delta\}) \leq \varepsilon + \delta$. So $D(F, H) \leq \varepsilon + \delta < \beta + \gamma = \alpha$. Thus $D(F, H) \leq D(F, G) + D(G, H)$.

(iii) Let be $F, G \in \mathcal{M}$. For every $\alpha < D(F, G)$, we have $\mu(A) > \alpha$, where $A = \{t \in T; \varphi(F(t), G(t)) > \alpha\}$. Hence

$$d(F, G) = \int_T \frac{\varphi(F, G)}{1 + \varphi(F, G)} d\mu \geq \int_A \frac{\varphi(F, G)}{1 + \varphi(F, G)} d\mu \geq \frac{\alpha}{1 + \alpha} \min\{1, \mu(A)\}$$

$$= \frac{\alpha}{1 + \alpha} \mu(A) > \frac{\alpha^2}{1 + \alpha} > \frac{\alpha^2}{1 + D(F, G)} \geq \frac{\alpha^2}{1 + \mu(T)}.$$

This implies $\sqrt{(1 + \mu(T))d(F, G)} \geq \alpha$ for every $\alpha < D(F, G)$. So $D^2(F, G) \leq (1 + \mu(T))d(F, G)$ and the proof is complete. □

Corollary 2. *If* φ *is symmetric and satisfies TI, and* μ *is a submeasure, then* D *is a semi-metric on* $\mathcal{M}$.

Remark 6. The theory of fuzzy convergences of set-valued functions sequences and set-valued integrals have interesting and important applications in:

- modeling subjective evaluation or subjective preference [36, 37];
- (multi-criteria) decision making theory [38, 39];
- stochastic processes [40];
- information sciences and theory of aggregation operators [32, 41, 42];
- study of efficiency or relevance of different processes in medicine, accident statistics, work technology [43–45].

1.4 Conclusion

We introduced two concepts of generalized measurability for set-valued functions, namely φ-μ-total-measurability and φ-μ-measurability, relative to a non-negative function $\varphi : \mathcal{P}_0(X) \times \mathcal{P}_0(X) \to [0, +\infty)$ and a non-negative set function

$\mu : \mathcal{A} \to [0, +\infty)$, and present some relationships between them. We also defined different types of convergences for sequences of set-valued functions such as almost everywhere convergence, pseudo-almost everywhere convergence, almost uniformly convergence, pseudo-almost uniformly convergence, and prove some relationships among them and a theorem of Egorov type. Finally, we introduced two semi-metrics on a space of set-valued functions and then compared them. As future work, we are going to study the topologies induced by the two semi-metrics.

References

1. Apreutesei, G.: Families of subsets and the coincidence of hypertopologies. An. Şt. Univ. "Al. I. Cuza" Iaşi **49**(1), 3–18 (2003)
2. Candeloro, D., Pucci, S.: Radon-Nikodym derivatives and conditioning in fuzzy measure theory. Stochastica **XI-2,3**, 107–120 (1987)
3. Chiţescu, I.: Finitely purely atomic measures: Coincidence and rigidity properties. Rendiconti Circolo Matematico Palermo Ser II, **Tomo L**, 455–476 (2001)
4. Choquet, G.: Theory of capacities. Ann. Inst. Fourier (Grenoble) **5**, 131–292 (1953–1954)
5. Croitoru, A., Gavriluţ, A., Mastorakis, N.E., Gavriluţ, G.: On different types of non-additive set multifunctions. WSEAS Trans. Math. **8**, 246–257 (2009)
6. Croitoru, A.: Set-norm continuity of set multifunctions. ROMAI J. **6**, 47–50 (2010)
7. Croitoru, A.: Fuzzy integral of measurable multifunctions. Iran. J. Fuzzy Syst. **9**, 133–140 (2012)
8. Croitoru, A.: Strong integral of multifunctions relative to a fuzzy measure. Fuzzy Set. Syst. **244**, 20–33 (2014)
9. Croitoru, A., Godet-Thobie, C.: Set-valued integration in seminorm I. Ann. Univ. Craiova Math. Comp. Sci. Ser. **33**, 16–25 (2006)
10. Croitoru, A., Văideanu, C.: On pointwise measurability of multifunctions. An. Şt. Univ. "Ovidius" Constanţa **17**(1), 69–75 (2009)
11. Croitoru, A., Mastorakis, N.E.: Estimations, convergences and comparisons on fuzzy integrals of Sugeno, Choquet and Gould type. In: Proceedings of the 2014 IEEE International Conferences on Fuzzy Systems (FUZZ-IEEE), pp. 1205–1212. IEEE Press, New York (2014)
12. De Korvin, A., Kleyle, R.: A convergence theorem for convex set valued supermartingales. Stoch. Anal. Appl. **3**, 433–445 (1985)
13. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. **38**, 325–339 (1967)
14. Dubois, D., Prade, H.: Fuzzy Sets and Systems. Theory and Applications. Academic, New York (1980)
15. Gavriluţ, A.: Properties of regularity for multisubmeasures with respect to the Vietoris topology. An. Şt. Univ. "Al. I. Cuza" Iaşi **52**, 389–400 (2006)
16. Gavriluţ, A.: Regularity and o-continuity for multisubmeasures. An. Şt. Univ. "Al. I. Cuza" Iaşi **50**, 393–406 (2004)
17. Gavriluţ, A., Iosif, A.E., Croitoru, A.: The Gould integral in Banach lattices. Positivity (2014). doi:10.10007/s11117-014-0283-7
18. Mastorakis, N.E.: General fuzzy systems as extensions of the Takagi-Sugeno methodology. WSEAS Trans. Syst. **3**(2), 795–800 (2004)

19. Mastorakis, N.E.: Modeling dynamical systems via the Takagi-Sugeno fuzzy model. WSEAS Trans. Syst. **3**(2), 668–675 (2004)
20. Mastorakis, N.E., Gavriluţ, A., Croitoru, A., Apreutesei, G.: On Darboux property of fuzzy multimeasures. In: Proceedings of the 10-th WSEAS International Conference on Fuzzy Systems (FS '09), pp. 54–58. WSEAS Press, Prague (2009)
21. Oberguggenberger, M.: The mathematics of uncertainty: Models, methods and interpretations. In: Fellin, W., Lessman, H., Oberguggenberger, M., Veider, R. (eds.) Analyzing Uncertainty in Civil Engineering, pp. 51–72. Springer, Berlin (2005)
22. Pap, E.: Null-Additive Set Functions. Kluwer Academic, Dordrecht (1995)
23. Papageorgiou, N.S.: On the efficiency and optimality of allocations II. SIAM J. Control Optim. **24**, 452–479 (1986)
24. Patriche, M.: Equilibrium of Bayesian fuzzy economies and quasi-variational inequalities with random fuzzy mappings. J. Inequal. Appl. **2013**, 374 (2013)
25. Precupanu, A.M.: On the set valued additive and subadditive set functions. An. Şt. Univ. "Al. I. Cuza" Iaşi **29**, 41–48 (1984)
26. Precupanu, A.M., Gavriluţ, A., Croitoru, A.: A fuzzy Gould type integral. Fuzzy Set. Syst. **161**, 661–680 (2010)
27. Sambucini, A.R.: A survey on multivalued integration. Atti. Sem. Mat. Fis. Univ. Modena **50**, 53–63 (2002)
28. Satco, B.: A Vitali type theorem for the set-valued Gould integral. An. Şt. Univ. "Al. I. Cuza" Iaşi **51**, 191–200 (2005)
29. Schmelzer, B.: Set-valued assessments of solutions to stochastic differential equations with random set parameters. J. Math. Anal. Appl. **400**, 425–438 (2013)
30. Stamate, C., Croitoru, A.: Non-linear integrals, prperties and relationships. In: Recent Advances in Telecommuncations, Signals and Systems, pp. 118–123. WSEAS Press, Lemesos (2013)
31. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
32. Mesiar, R., De Baets, B.: New construction methods for aggregation operators. In: Proceedings of the IPMU '2000, Madrid, pp. 701–707 (2000)
33. Wang, Z., Klir. G.J.: Fuzzy measure theory, Plenum Press, New York (1992)
34. Wu, C., Ren, X., Wu, C.: A note on the space of fuzzy measurable functions for a monotone measure. Fuzzy Set. Syst. **182**, 2–12 (2011)
35. Florescu, L.C.: The Centennial of Convergence in Measure. Series Mathematics and Informatics, vol. 19, no. 2, pp. 221–240. "Vasile Alecsandri" University of Bacău (2009)
36. Sugeno, M.: Theory of fuzzy integrals and its applications. Ph.D. Thesis, Tokyo Institute of Technology (1974)
37. Tversky, A., Kahneman, D.: Advances in prospect theory: Cummulative representation of uncertainty. J. Risk Uncertain. **5**, 297–323 (1992)
38. Grabisch, M.: Fuzzy integral in multicriteria decision making. Fuzzy Set. Syst. **69**, 279–298 (1995)
39. Grabisch, M., Labreuche, C.: A decade of applications of the Choquet and Sugeno integrals in multi-criteria decision aid. **40R 6**, 1–44 (2008)
40. Aletti, G., Bongiorno, E.G., Capasso, V.: Statistical aspects of fuzzy monotone set-valued stochastic processes. Application to birth-and-growth processes. Fuzzy Set. Syst. **160**, 3140–3151 (2009)
41. Cao, Y.: Aggregating multiple classification results using Choquet integral for financial distress early warning. Expert Syst. Appl. **38**(7), 8285–8292 (2011)
42. Wang, Z., Yang, R., Leung, K.S.: Nonlinear integrals and their applications in data mining. In: Advances in Fuzzy Systems-Applications and Theory. World Scientific, Singapore (2010)

43. Liu, W.L., Song, X.Q., Zhang, Q.Z., Zhang, S.B.: (T) Fuzzy integral of multi-dimensional function with respect to multi-valued measure. Iran. J. Fuzzy Syst. **9**(3), 111–126 (2012)
44. Mahmoud, Muhammad, M.A.S., Moraru, S.: Accident rates estimation modelling based on human factors using fuzzy c-means clustering algorithm. World Acad. Sci. Eng. Technol. (WASET) **64**, 1209–1219 (2012)
45. Pham, T.D., Brandl, M., Nguyen, N.D., Nguyen, T.V.: Fuzzy measure of multiple risk factors in the prediction of osteoporotic fractures. In: Proceedings of the 9-th WSEAS International Conference on Fuzzy Systems (FS '08), pp. 171–177. WSEAS Press, Sofia (2008)

Chapter 2
Short Term Load Forecasting in Electric Power Systems with Artificial Neural Networks

G.J. Tsekouras, F.D. Kanellos, and N. Mastorakis

Abstract The demand in electric power should be predicted with the highest possible accuracy as it affects decisively many of power system's operations. Conventional methods for load forecasting were built on several assumptions, while they had to cope with relations between the data used that could not be described analytically. Artificial Neural Networks (ANNs) gave answers to many of the above problems and they became the predominant load forecasting technique. In this chapter the reader is first introduced to Artificial Neural Networks and their usage in forecasting the load demand of electric power systems. Several of the major training techniques are described with their pros and cons being discussed. Finally, feed- forward ANNs are used for the short-term forecasting of the Greek Power System load demand. Various ANNs with different inputs, outputs, numbers of hidden neurons etc. are examined, techniques for their optimization are proposed and the obtained results are discussed.

Keywords Artificial neural networks • ANN evaluation • Load Forecasting • Short term load forecasting • Training methods

2.1 Introduction

In a deregulated electricity market, the electric load has to be predicted with the highest possible accuracy for different time periods: very short-term (few minutes), short-term (few hours up to 1 week), midterm (few weeks up to few months) and long-term (few months up to years). Especially, the short-term load forecasting is very crucial as it affects decisively several power systems operations such as unit

G.J. Tsekouras (✉) • N. Mastorakis
Department of Electrical Engineering and Computer Science, Hellenic Naval Academy, Terma Hatzikiriaku, PIRAEUS, Athens 18539, Greece
e-mail: tsekouras@snd.edu.gr; mastor@snd.edu.gr

F.D. Kanellos
School of Production Engineering and Management, Technical University of Crete, University Campus, Chania 73100, Greece
e-mail: fkanellos@dpem.tuc.gr

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_2

commitment [1], spinning reserve scheduling [2], estimation of available transfer capability [3] and stability margins [3], load shedding decisions, etc. Consequently, the accurate load forecasting ensures higher reliability in power system operation while it facilitates the minimization of its operation cost by providing accurate input to day-ahead scheduling.

The efficiency of load forecasting is highly affected by the used input data. The most significant data used for short-term load forecasting are the hourly average values of the load for time periods extending from few past hours up to some weeks before the day the load is forecasted. The type of the day e.g. weekday, special day, weekend etc. plays also a key role in load forecasting accuracy. For example, load profile is different in weekdays and weekend, while load is more difficult to be forecasted in special days. Electric load can be also attributed with different behavior over the epochs of the year e.g. in Greece load demand is higher in summer due to the increased touristic activity and energy demand for air-conditioning. The above issues can be addressed with careful and methodical selection of the input data used for the load forecasting and the time period they are referred to.

Before applying a load forecasting model the load demand should be carefully decomposed in its components e.g. deterministic load, weather-dependent load component etc. Several load decomposition methods have been proposed so far in the literature. Namely, load decomposition:

(a) in four components: Deterministic load, weather-independent load, weather-dependent load and noise component for the remaining load [4].
(b) in three components: Yearly, seasonal and daily loads [5] that can be exploited in autoregressive models. Alternatively, the daily, weekly and cyclic components of the load can be used [6].
(c) in two components: the basic and the weather-dependent load components [7].

Several methods have been used for short-term load forecasting with different levels of success, such as ARMAX models [8], regression [5], Artificial Neural Networks (ANNs) [9], fuzzy logic [10], expert systems etc. Among the variety of load forecasting methods ANNs have been proved the most effective [9]. In this chapter introduction to ANNs and their application to load forecasting is provided, special issues concerning the optimization of ANN training and validation are discussed, while results obtained from the exploitation of various types of ANNs for Greek power system load forecasting are presented and discussed.

2.2 Review of Short Term Load Forecasting Methods

A large number of modern load forecasting techniques are based on the exploitation of ANNs as there are no well–defined relations between the load and several factors affecting it, such as temperature, humidity, time, load values at previous hours etc. Despite the fact that ANNs were initially avoided [11, 12] as they did not help to understand the nature of the problem they finally dominated due to the same reason i.e. they allow to tackle extremely complex problems without fully understanding them.

Pioneers in this field were the members of the EPRI research team, Khotanzad et al., that developed various methodologies exploiting separate ANNs, for the prediction of time components of the load (weekly, daily, hourly) [13], for each prediction hours [14] and the load types (weather non-dependent and dependent load components) [15]. Papalexopoulos et al. introduced the concept of load seasonality through the use of sinusoidal functions of period equal to the number of days of the year [16].

Rapid developments in load forecasting took place in the following period. More specifically, significant advances occurred in, the forecasting of load increase [17], learning techniques [18–21]. New network structures such as: radial basis function networks [22–24], recurrent networks [25], not fully connected networks [26], abductive networks [27], probabilistic networks [28, 29], and networks using similar days [30] were also explored. Moreover, several new techniques like fuzzy logic [31–39], wavelet decomposition [40], support vector machines [24, 41], genetic algorithms [42], harmony search [43] etc. were combined with ANNs to enhance their performance. Especially for Greece, pioneers in load forecasting were Bakirtzis et al., that proposed load forecasting techniques for interconnected power systems [44] as well as autonomous systems [45] followed by Kalaitzakis et al. [20] and Tsekouras et al. [46–48].

Hippert et al. [9], Choueiki et al. [49], and other researchers [50–54] have proved that the short-term load forecasting with classical multilayer artificial neural networks trained by error back propagation algorithm lead to mean absolute forecasting error ranging between 1.5 and 2.5 %. Usually, these techniques incorporate simple correction algorithms of data irregularities e.g. load behavior in holidays, measurement errors etc. However, the conclusions drawn can be hardly generalized as the influence of the inherent characteristics of the power systems is very strong and differs from system to system.

2.3 Multilayer Feed Forward Artificial Neural Networks

2.3.1 Introduction

A widely used type of multilayer artificial neural networks is the multilayer perceptron (MLP). The structure of a MLP is presented in Fig. 2.1. The neurons of the network are organized into three layers: the input, the hidden and the output layer. MLP is characterized as a feed forward network as the information flows only on the direction from the input to the output layer. According to Kolmogorov's theorem [55] an ANN can solve a problem by using one hidden layer provided that it comprises adequate number of neurons. Under these circumstances one hidden layer is used, but the number of its neurons should be properly selected. The number of the output layer neurons is equal to the number of the model output variables, while the input nodes correspond to the input variables of the model.

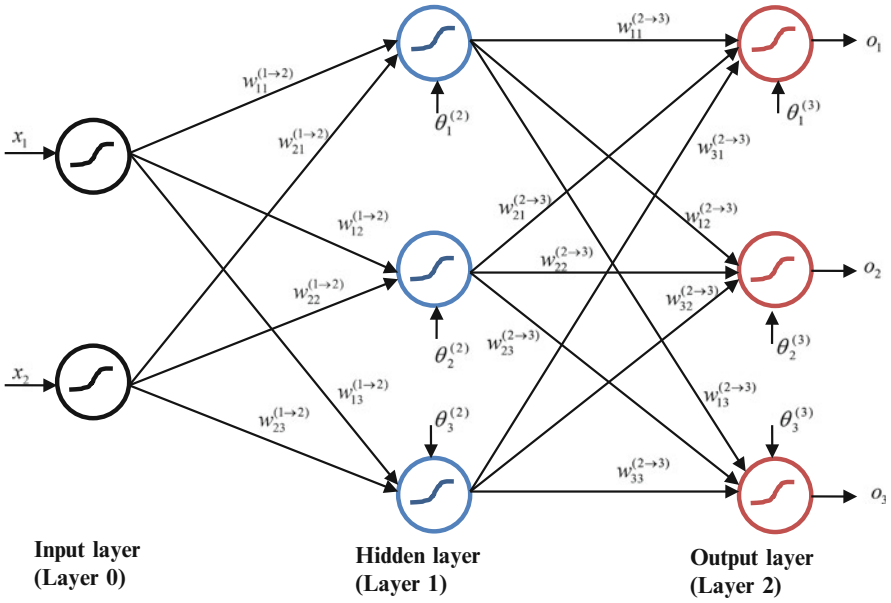


Fig. 2.1 2-3-3 (with 2 input, 3 hidden and 3 output neurons) feed forward artificial neural network

The increased computational capability of a MLP stems from the inherent nonlinear nature of its neurons, the complete interconnection between successive layers and its ability to learn after proper training. Research interest in MLPs first appeared by Rosenblatt in his work on the perceptron and also by Widrow who presented Madaline network in 1962. However, an efficient training algorithm for these networks was missing until 1985. Then, the error back-propagation algorithm was proposed for application to multilayer ANNs and it is still one of the most widely used ANN training methods. It should be noticed that error back propagation algorithm was first proposed by Werbos in his Ph.D. thesis in 1974. From 1985 until 1986 it was used by Rumelhart, Hinton, Williams, McClelland, Parker and LeCun, while it has been used since then in numerous applications [55, 56].

Next some basic information on ANN models is provided. The activation signal (input) of the k-th neuron of the ℓ-th layer of an ANN is:

$$u_k^{(\ell)}(n) = \sum_{v=0}^{p_{\ell-1}} w_{kv}^{(\ell)}(n) y_v^{(\ell-1)}(n) \tag{2.1}$$

where $w_{kv}^{(\ell)}$ is the interconnection weight between the k-th neuron of the ℓ-th layer and the v-th neuron of the $(\ell-1)$ -th layer, $p_{\ell-1}$ is the total number of the neurons of the $(\ell-1)$ -th layer and $y_v^{(\ell-1)}$ is the output of the v-th neuron of the $(\ell-1)$ -th layer.

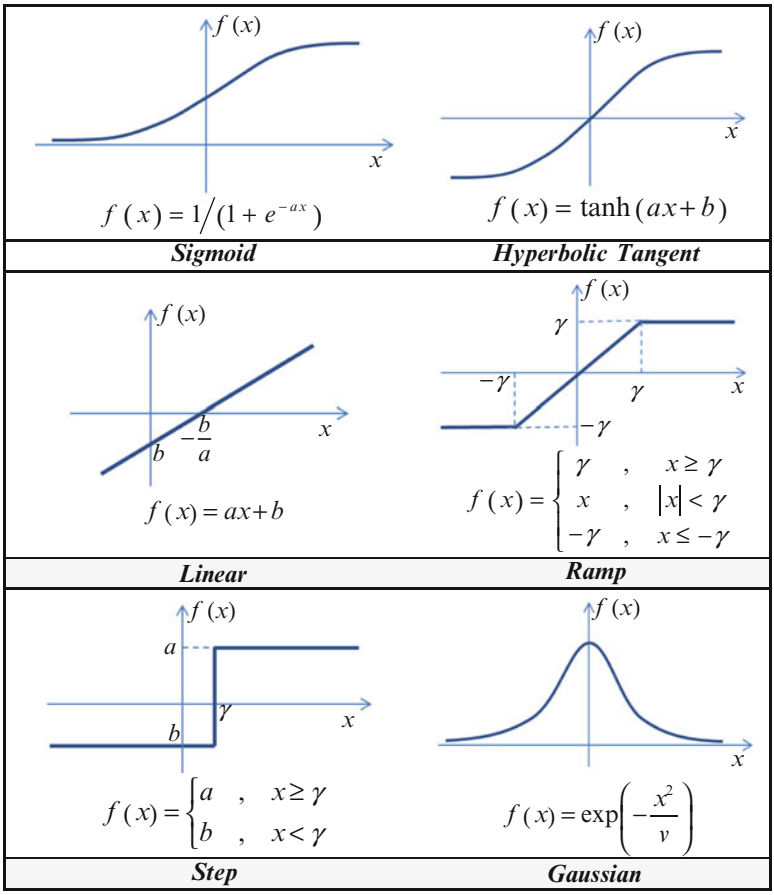


Fig. 2.2 Activation functions

The activation function of the neurons is denoted next with f. Hence, the output the k-th neuron of the ℓ -th layer is:

$$y_k^{(\ell)}(n) = f\left(u_k^{(\ell)}(n)\right) \tag{2.2}$$

The most common activation functions are shown in Fig. 2.2.

Nonlinear activation functions are preferred in nonlinear problems. However, in that case saturation problems may occur. These problems can be attributed to the use of sigmoid activation functions which take values in a bounded area and are also highly nonlinear outside the region [−1, 1]. In order to avoid saturation problems the input and the output variables of the ANN are normalized as in the following relation:

$$\widehat{x} = a + \frac{b-a}{x_{\max} - x_{\min}}(x - x_{\min}) \tag{2.3}$$

where $\widehat{x}$ is the normalized value of variable x, x_{min} and x_{max} are the lower and the upper bounds of x, and a, b are the respective values of the normalized variable.

Alternatively, the input and output variables can be normalized by using their mean and standard deviation values as shown in Eq. (2.4).

$$\widehat{x} = \frac{x - \mu}{\sigma} \tag{2.4}$$

However, the second method does not necessarily maintain the normalized variables outside the highly nonlinear region of the activation function.

2.3.2 Steepest Descent Error Back-Propagation Algorithm

Appropriate data sets are needed for the training, evaluation, validation and optimization of the network. Let us assume that m_1 vectors are used for training, m_2 for optimization-evaluation of ANN parameters and m_3 for load forecasting purposes. The data set comprising the m_2 vectors used for network evaluation and optimization may be a subset of the training data set.

The basic steps of the steepest descent error back-propagation training algorithm [55, 56] are as follows:

(a) *Initialization*: Parameters like the number of the neurons in each layer, the training rate etc. are defined. Connection weights are initialized to small random values uniformly distributed in the interval $[-0.1, 0.1]$.

(b) *Use of the training set*: In each training epoch all the training patterns are randomly used. For each input vector c and d steps are applied.

(c) *Forward pass calculations*: The n-th training pattern is defined as $\{\overrightarrow{x}_{in}(n), \overrightarrow{t}(n)\}$, where $\overrightarrow{x}_{in}(n)$ is the q_{in}x1 input vector comprising the normalized input variables x_j and $\overrightarrow{t}(n)$ the respective q_{out}x1 normalized output vector. The activation signal of the k-th neuron of the ℓ-th layer is:

$$u_k^{(\ell)}(n) = \sum_{v=0}^{p_{\ell-1}} w_{kv}^{(\ell)}(n) y_v^{(\ell-1)}(n) \tag{2.5}$$

where $w_{kv}^{(\ell)}$ is the weight assigned to the connection of the k-th neuron of the ℓ-th layer with the v-th neuron of the $(\ell-1)$ -th layer, $p_{\ell-1}$ is the total number of neurons of the $(\ell-1)$ -th layer and $y_v^{(\ell-1)}$ is the output of the v-th neuron of the $(\ell-1)$ -th layer. For $v=0$, the bias is defined as $\theta_k = w_{k0}$, while $y_0^{(\ell-1)} = -1$. The activation function f at each layer can be the hyperbolic tangent, sigmoid or linear. The output of the neuron is:

$$y_k^{(\ell)}(n) = f\left(u_k^{(\ell)}(n)\right) \tag{2.6}$$

The output of the v-th neuron of the input layer is:

$$y_v^{(0)}(n) = x_v(n), \quad \forall v \tag{2.7}$$

where $x_v(n)$ is the v-th element of network input vector $\overrightarrow{x}_{in}(n)$.

The output of the k-th neuron of the output layer (L) is:

$$y_k^{(L)}(n) = o_k(n), \quad \forall k \tag{2.8}$$

where $o_k(n)$ is the k-th element of the output vector $\overrightarrow{o}(n)$, estimated by the ANN.

Hence, the error obtained at output of the k-th neuron of the output layer is:

$$e_k(n) = t_k(n) - o_k(n) \tag{2.9}$$

(d) *Reverse pass calculations*: The weights are updated by using the delta-rule:

$$w_{kv}^{(\ell)}(n+1) = w_{kv}^{(\ell)}(n) + \eta \cdot \delta_k^{(\ell)}(n) \cdot y_v^{(\ell-1)}(n) \tag{2.10}$$

where, n is the constant training rate and $\delta_k{}^{(\ell)}(n)$ the local descent of the k-th neuron estimated for the output and hidden layers as following:

$$\delta_k^{(L)}(n) = e_k^{(L)}(n) \cdot f^{/}\left(u_k^{(L)}(n)\right) \tag{2.11}$$

$$\delta_k^{(\ell)}(n) = f^{/}\left(u_k^{(\ell)}(n)\right) \cdot \sum_i \delta_i^{(\ell+1)}(n) w_{ik}^{(\ell+1)}(n) \tag{2.12}$$

(e) *Stopping criteria*: Steps b up to d are repeatedly executed until the weights of neuron interconnections are stabilized or the output error function does not improve or the maximum number of epochs is exceeded.

Neurons' connection weights stabilization criterion is formulated as:

$$\left| w_{kv}^{(\ell)}(ep) - w_{kv}^{(\ell)}(ep-1) \right| < \ell limit_1, \forall k, v, \ell \tag{2.13}$$

where, $\ell limit_1$ is the upper limit of the absolute weight change and ep is the current epoch of training algorithm.

The output error function is the root mean square error $RMSE_{va}$ estimated for the evaluation data set according to:

$$RMSE_{va} = \sqrt{\frac{1}{m_2 \cdot q_{out}} \sum_{m=1}^{m_2} \sum_{k=1}^{q_{out}} e_k^2(m)} \tag{2.14}$$

The respective stopping criterion is formulated as:

$$|RMSE_{va}(ep) - RMSE_{va}\,(ep-1)| < \ell imit_2 \tag{2.15}$$

where $\ell imit_2$ is the upper bound of the absolute change of RMSE in two successive epochs.

The maximum number of epochs exceedance criterion is formulated as:

$$ep \geq max\_epochs \tag{2.16}$$

If one of the above criteria becomes true, the main core of error back-propagation algorithm is ended. Otherwise the number of epochs is increased by one and the algorithm returns to step *b*. The above criterions are used on one hand to avoid data overfitting and on the other to enable the training algorithm to converge.

(f) *Validation criteria*: The mean absolute percentage error (MAPE) is calculated for the evaluation data set as following:

$$MAPE_{va} = 100\% \cdot \sum_{i=1}^{m_2} |(t_k(i) - o_k(i))\,/\,t_k(i)|/m_2 \tag{2.17}$$

In this training process each input vector is used randomly per epoch (stochastic training) to minimize the error function, $G(n) = \frac{1}{2}\sum_{j=1}^{q_{out}} e_j^2(n)$.

Alternatively, all input vectors can be used in a series during the forward process and afterwards the weights are updated minimizing the average output error function, $G_{av} = \frac{1}{m_1}\sum_{n=1}^{m_1} G(n) = \frac{1}{2m_1}\sum_{n=1}^{m_1}\sum_{j=1}^{q_{out}} e_j^2(n)$. This training process is called batch mode and the respective weights update term is calculated as:

$$\Delta\vec{w}(ep) = -\eta \cdot \nabla G\left(\vec{w}(ep)\right) \tag{2.18}$$

If a momentum term, a, is added then the respective equation becomes:

$$\Delta\vec{w}(ep) = -\eta \cdot \nabla G\left(\vec{w}(ep)\right) + a \cdot \Delta\vec{w}\,(ep-1) \tag{2.19}$$

In steepest descent algorithm the parameters of learning rate and momentum are kept constant. Alternatively, decreasing exponential functions, as described in Eqs. (2.20) and (2.21) can be used:

$$\eta(ep) = \eta_0 \cdot \exp\left(-ep/T_\eta\right) \tag{2.20}$$

$$a(ep) = a_0 \cdot \exp\left(-ep/T_a\right) \tag{2.21}$$

where, T_n, T_a are time parameters and n_0, a_0 the initial values of training rate and momentum term, respectively. Faster convergence can be achieved through the proper calibration of T_n, n_0, T_a, a_0 and it becomes even faster especially if the initial values n_0, a_0 are large.

2.3.3 *Other Training Methods Based on Error Back-Propagation Algorithm*

Variations of the error back-propagation algorithm are presented in the following paragraphs.

2.3.3.1 Adaptive Error Back Propagation

In this method both the training rate and the momentum term are adaptively changed as described in Eqs. (2.22) and (2.23) in order to achieve rapid convergence.

$$\eta(ep) = \begin{cases} \eta\,(ep-1)\,, & RMSE_{tr}(ep) > RMSE_{tr}\,(ep-1) \\ \eta\,(ep-1)\cdot\exp\left(-1/T_\eta\right), & RMSE_{tr}(ep) \le RMSE_{tr}\,(ep-1) \end{cases} \tag{2.22}$$

$$a(ep) = \begin{cases} a\,(ep-1)\,, & RMSE_{tr}(ep) \le RMSE_{tr}\,(ep-1) \\ a\,(ep-1)\cdot\exp\left(-1/T_a\right), & RMSE_{tr}(ep) > RMSE_{tr}\,(ep-1) \end{cases} \tag{2.23}$$

where $\eta_0 = \eta(0)$, $a_0 = a(0)$ and $RMSE_{tr}(ep)$ is the root mean square of the output error estimated for the training data set in training epoch, ep. If $RMSE_{tr}(ep-1)$ is larger than $RMSE_{tr}(ep)$, which means that neurons' connection weights were updated in the correct direction, then training rate is decreased while the momentum term values remains the same in the next epoch. In this way, the previous successful update of the weights is rewarded. Otherwise, if $RMSE_{tr}(ep) > RMSE_{tr}(ep-1)$ it is reasonable to reduce the momentum term and keep the learning rate constant as a penalty to the previous unsuccessful update of the weights. It should be noted that increasing the momentum term or the learning rate, as proposed in [57], may lead the training process to instability.

2.3.3.2 Resilient Algorithm

In resilient algorithm only the sign of the derivative of the error function with respect w_{ij} is used for the estimation of the connection weight change direction. The weights are updated by using the following relations:

$$\Delta w_{ij}(ep) = \begin{cases} \delta_1 \cdot \Delta w_{ij}(ep-1), & \frac{\partial G_{av}}{\partial w_{ij}}(ep) \cdot \frac{\partial G_{av}}{\partial w_{ij}}(ep-1) > 0 \\ \Delta w_{ij}(ep-1), & \frac{\partial G_{av}}{\partial w_{ij}}(ep) \cdot \frac{\partial G_{av}}{\partial w_{ij}}(ep-1) = 0 \\ \frac{1}{\delta_2} \cdot \Delta w_{ij}(ep-1), & \frac{\partial G_{av}}{\partial w_{ij}}(ep) \cdot \frac{\partial G_{av}}{\partial w_{ij}}(ep-1) < 0 \end{cases} \tag{2.24}$$

where, δ_1, δ_2 are increasing and decreasing factors of the weight change over two successive epochs, respectively. If the derivative of the error function G with respect to the weight w_{ij} maintains the same sign during two sequential epochs, the weight change over epoch *ep* is the multiple of the respective change in epoch, *ep*-1 (multiplied by δ_1). Respectively, if the sign of the derivative changes then the weight change is decreased else if the derivative is zero then the same weight change with the one of the previous epoch is applied.

2.3.3.3 Conjugate Gradient Algorithms

In conjugate gradient (CG) algorithm [21] a search is performed along conjugate directions leading generally to faster convergence than following the steepest descent direction. The basic steps of the CG algorithm (Fletcher-Reeves and Polak-Ribiere) are as follows:

(a) In the first iteration the search direction $\vec{p}_0$ is determined by the opposite of the output error function gradient:

$$\vec{p}_0 = -\nabla G\left(\vec{w}\right)\Big|_{\vec{w}=\vec{w}_0} \tag{2.25}$$

(b) At *k*-th iteration the weights of the network are updated towards the search direction $\vec{p}_k$ as following:

$$\Delta \vec{w}_k = a_k \cdot \vec{p}_k \tag{2.26}$$

The positive parameter a_k is calculated by numerical methods such as the golden section, bisection etc.

(c) In the next iteration the search direction $\vec{p}_{k+1}$ is calculated by:

$$\vec{p}_{k+1} = -\nabla G\left(\vec{w}\right)\Big|_{\vec{w}=\vec{w}_{k+1}} + \beta_{k+1} \cdot \vec{p}_k \tag{2.27}$$

β_{k+1} is calculated either by the Fletcher-Reeves [58] or Polak-Ribiere equation [59] respectively:

$$\beta_{k+1} = \frac{\nabla G\left(\vec{w}\right)\Big|_{\vec{w}=\vec{w}_{k+1}}^{T} \cdot \nabla G\left(\vec{w}\right)\Big|_{\vec{w}=\vec{w}_{k+1}}}{\nabla G\left(\vec{w}\right)\Big|_{\vec{w}=\vec{w}_k}^{T} \cdot \nabla G\left(\vec{w}\right)\Big|_{\vec{w}=\vec{w}_k}} \tag{2.28}$$

$$\beta_{k+1} = \frac{\Delta\left(\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_k}^{T}\right)\cdot\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_{k+1}}}{\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_k}^{T}\cdot\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_k}} \tag{2.29}$$

(d) If the algorithm has not converged then step (b) is repeated. It is mentioned that the k-th iteration usually coincides with the respective epoch, but this is not necessarily the case. The iterative process must be occasionally restarted in order to avoid a constant convergence rate. It is usual to restart it every N_w or $(N_w + 1)$ iterations, where N_w is the number of the variables (weights and biases). Powell and Beale [60] proposed to restart the process of Polak-Ribiere algorithm occasionally or when the orthogonality between $\vec{p}_k$ and $\vec{p}_{k-1}$ is quite small:

$$\left|\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_k}^{T}\cdot\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_{k+1}}\right| \geq \lim_{orthogonality}\cdot\left\|\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_{k+1}}\right\|^2 \text{ with } k \geq 1 \tag{2.30}$$

The limit, $\lim_{orthogonality}$, may take values within the interval (0.1, 0.9). Usually, it is set equal to 0.2.

The basic drawback of CG algorithm is the complexity of the evolved calculations per iteration as a linear search is performed to determine the appropriate step size. In scaled conjugate gradient algorithm (SCGA) the search process is avoided by using the Levenberg-Marquardt approach. The basic steps of SCGA are as follows [61]:

(a) $\vec{p}_0$ is initialized by Eq. (2.25) and the vector of the weights and biases $\vec{w}_0$ is properly chosen. The rest parameters of the algorithm (σ, λ_0, $\bar{\lambda}_0$, *flag*) are set as following:

$$0 < \sigma \leq 10^{-4} \qquad 0 < \lambda_0 \leq 10^{-6} \qquad \bar{\lambda}_0 = 0 \qquad flag = 1$$

(b) If *flag* is 1 then the following additional information is calculated:

$$\sigma_k = \sigma / \left\|\vec{p}_k\right\| \tag{2.31a}$$

$$\vec{s}_k = \left(\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_k+\sigma_k\cdot\vec{p}_k} - \nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_k}\right)/\sigma_k \tag{2.31b}$$

$$\delta_k = \vec{p}_k^{\,T}\cdot\vec{s}_k \tag{2.31c}$$

(c) The parameter δ_k is scaled according to:

$$\delta_k = \delta_k + \left(\lambda_k - \bar{\lambda}_k\right)\cdot\left\|\vec{p}_k\right\|^2 \tag{2.32}$$

(d) If $\delta_k \leq 0$, then the Hessian matrix is made positive by setting:

$$\overline{\lambda}_k = 2\left(\lambda_k - \delta_k / \left\| \overrightarrow{p}_k \right\|^2\right) \tag{2.33a}$$

$$\delta_k = -\delta_k + \lambda_k \cdot \left\| \overrightarrow{p}_k \right\|^2 \tag{2.33b}$$

$$\lambda_k = \overline{\lambda}_k \tag{2.33c}$$

(e) The step size is calculated as:

$$\mu_k = -\overrightarrow{p}_k^T \cdot \nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_k} \tag{2.34a}$$

$$a_k = \mu_k / \delta_k \tag{2.34b}$$

(f) The comparison parameter Δ_k is calculated as:

$$\Delta_k = 2 \cdot \delta_k \cdot \left(G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_k} - G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_k + a_k \cdot \overrightarrow{p}_k} \right) / \mu_k^2 \tag{2.35}$$

(g) If $\Delta_k \geq 0$ then a successful reduction in error can be achieved by applying the following equations:

$$\Delta \overrightarrow{w}_k = a_k \cdot \overrightarrow{p}_k \tag{2.36a}$$

$$\overrightarrow{r}_{k+1} = -\nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_{k+1}} \tag{2.36b}$$

$$\overline{\lambda}_k = 0 \tag{2.36c}$$

$$flag = 1 \tag{2.36d}$$

(h) If the number of iterations is multiple of the population of the weights and biases, N_w, then the algorithm is restarted:

$$\overrightarrow{p}_{k+1} = -\nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_{k+1}} \tag{2.37}$$

else:

$$\beta_{k+1} = \left(\left\| \nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_{k+1}} \right\|^2 - \nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_{k+1}}^T \cdot \nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_k} \right) / \mu_k \tag{2.38}$$

$$\overrightarrow{p}_{k+1} = -\nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_{k+1}} + \beta_{k+1} \cdot \overrightarrow{p}_k \tag{2.39}$$

If $\Delta_k \geq 0.75$, then $\lambda_k = 0.25 \cdot \lambda_k$, else $\overline{\lambda}_k = \lambda_k, flag = 0$.
If $\Delta_k < 0.25$, then $\lambda_k = \lambda_k + \delta_k (1 - \Delta_k) / \left\| \overrightarrow{p}_k \right\|^2$.

(i) If $\nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_{k+1}} \neq \overrightarrow{0}$, then $k = k + 1$ and the step (b) is repeated else the training process is completed.

The basic drawback of the SCGA algorithm is the increased complexity of the calculations within an iteration that is in the order of $O(6{N_w}^2)$ instead of $O(3{N_w}^2)$ of the basic steepest descent method. If the scale parameter λ_k is zero, then the SCGA coincides with the CGA. SCGA's basic advantage is that the error decreases monotonically as error increase is not allowed. If the error is constant for one or two iterations then the Hessian matrix has not been positive definite and λ_k has been increased. It is also recommended that the value of parameter σ should be as small as possible in order to constrain its effect on the performance of the algorithm.

2.3.3.4 Newton Algorithm

In *Newton* method the inverse of the Hessian matrix, $\nabla^2 G\left(\overrightarrow{w}\right)$, is used to update the connection weights and biases of the network as follows:

$$\Delta \overrightarrow{w}_k = -\nabla^2 G\left(\overrightarrow{w}\right)\big|^{-1}_{\overrightarrow{w}=\overrightarrow{w}_k} \cdot \nabla G\left(\overrightarrow{w}\right)\big|_{\overrightarrow{w}=\overrightarrow{w}_k} \tag{2.40}$$

Usually, the convergence of this algorithm is more rapid than the aforementioned algorithms if the size of the problem is small. The calculation and the inversion of Hessian matrix are complex and computationally demanding processes. Hessian and Jacob matrices are estimated according to the following equations:

Hessian matrix:

$$\nabla^2 G\left(\overrightarrow{w}\right) = J\left(\overrightarrow{w}\right)^T \cdot J\left(\overrightarrow{w}\right) + \sum_{j=1}^{m_1} e_j\left(\overrightarrow{w}\right) \cdot \nabla^2 e_j\left(\overrightarrow{w}\right) \tag{2.41}$$

Jacob matrix:

$$J\left(\overrightarrow{w}\right) = \begin{bmatrix} \frac{\partial e_1}{\partial w_1} & \frac{\partial e_1}{\partial w_2} & \cdots & \frac{\partial e_1}{\partial w_{N_w}} \\ \frac{\partial e_2}{\partial w_1} & \frac{\partial e_2}{\partial w_2} & \cdots & \frac{\partial e_2}{\partial w_{N_w}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial e_{m_1}}{\partial w_1} & \frac{\partial e_{m_1}}{\partial w_2} & \cdots & \frac{\partial e_{m_1}}{\partial w_{N_w}} \end{bmatrix}_{m_1 \times N_w} \tag{2.42}$$

The *quasi-Newton* method belongs to this family of Newton's algorithms. In this method the second term of Hessian matrix is omitted as it usually takes small values that are not significant.

Alternatively, in the *one step secant* algorithm only the diagonal elements of Hessian matrix are used to simplify its inversion. It needs more iterations than the basic Newton method but the required computational effort is significantly compressed.

The *Levenberg-Marquardt* method [62, 63] is another commonly used variation of Newton algorithm. The weights and the biases of the network are updated as follows:

$$\begin{aligned}\Delta\vec{w}_k &= -\left(J^T\cdot J+\lambda\cdot diag\left[J^T\cdot J\right]\right)^{-1}\cdot\nabla G\left(\vec{w}\right)\big|_{\vec{w}=\vec{w}_k}\\ &= -\left(J^T\cdot J+\lambda\cdot diag\left[J^T\cdot J\right]\right)^{-1}\cdot J^T\cdot\vec{e}\left(\vec{w}_k\right)\end{aligned}\tag{2.43}$$

Factor λ is updated with respect to the change of the output error function as following:

$$\lambda\left(k+1\right)=\begin{cases}\lambda(k)\cdot\beta, & G_{av}(k)>G_{av}\left(k-1\right)\\ \lambda(k), & G_{av}(k)=G_{av}\left(k-1\right)\\ \lambda(k)/\beta, & G_{av}(k)<G_{av}\left(k-1\right)\end{cases}\tag{2.44}$$

The parameter β usually takes values near 10. The *Levenberg-Marquardt* algorithm is not the optimal one but it performs extremely well if the number of the weights and biases is smaller than few thousands.

2.3.4 Bias-Variance Dilemma

One significant issue closely related with the efficiency of the ANNs is the selection of the number of the neurons at hidden layer. Using many neurons leads the network to memorize the training data and degradation of its generalization properties occurs i.e. the ANN does not perform satisfactorily on data not belonging to the training set. On the contrary, small number of neurons reduces the capability of the ANN to identify complex relations between the input data while generalization capability may be satisfactory.

Therefore, it is crucial to determine the optimal ANN structure that leads to accurate results and ensures at the same time satisfactory generalization properties. This is known as the bias-variance dilemma where the generalization error is expressed by the terms of bias and variance. Simple models are characterized by increased bias while complex models comprising large number of parameters lead to increased variance. According to bias-variance dilemma, if the bias increases then variance decreases and vice versa. Consequently, the best model is the one obtained by the optimal proportion of the two terms.

There are two major approaches to achieve balance between bias and variance. In the first, known as the "structural" approach, the number of the neurons of the

hidden layer is gradually increased until the optimal generalization is achieved. Generalization becomes optimal when MAPE of the test data set starts to increase and overtraining occurs.

The second approach is based on the idea of network parameters regularization. The simplest way to achieve regularization is the introduction of a penalty term at the output error function of the network. This term prevents the parameters (weights and biases) from taking large values. In this way, parameters of small significance are nullified and the number of the parameters of the network eventually decreases. The sum of the squared values of the network parameters is often used as the regularization term. Assuming that p_i are the parameters of the network and G the output error function of the network then the following function is minimized during the training process:

$$G_f = G + \lambda \cdot \sum_i p_i^2 \tag{2.45}$$

where, λ is the parameter that determines the significance of the two minimization goals that are the minimization of the output error function and the reduction of network parameters values.

2.3.4.1 Confidence Interval Estimation

In case of ANNs the confidence interval is not directly estimated unlike to the classical forecasting models. Three techniques have been proposed so far [64]:

(a) *error output*: According to this technique the ANN model uses two outputs for each output variable, the first one is the forecasted mean value and the second one the respective absolute percentage error. After the training process a larger confidence interval is determined by multiplying the initial one by a proper factor in order to reach to the required confidence level.

(b) *re-sampling*: The prediction and the respective error are calculated for each set and for all available m input vectors and are sorted in ascending order. The cumulative distribution function of the prediction errors can be estimated by:

$$S_m(z) = \begin{cases} 0, & z < z_1 \\ r/m, & z_r \le z < z_{r+1} \\ 1, & z_m \le z \end{cases} \tag{2.46}$$

When m is large enough, $S_m(z)$ is a good approximation of the true cumulative probability distribution $F(z)$. The confidence interval is estimated by keeping the intermediate value z_r and discarding the extreme values, according to the desired confidence degree. The obtained intervals are equal in probability (not necessarily symmetric in z). The number of cases to discard in each tail of the prediction error distribution is $n \cdot p$, where p is the probability in each tail. If $n \cdot p$

is not an integer then $\lfloor n \cdot p \rfloor$ cases are discarded in each tail. If the cumulative probability distribution $F(Z_p)$ equals to p, then an error is less than or equal to Z_p with a p probability which indicates that Z_p is the lower confidence limit. Consequently, Z_{1-p} is the upper limit and there is a $(1-2p)$ confidence interval for future errors.

(c) *multi-linear regression model adapted to ANN*: This technique can be applied only if linear activation functions are used at the output layer. In this technique, a multi-linear regression model can be implemented for each neuron of the output layer. The inputs of the regression model are the outputs of the hidden neurons, while regression coefficients are their connection weights with the output neuron. The computation of the confidence interval is accomplished through the estimation of the prediction error variance:

$$\sigma^2 = \sum_{i=1}^{m_1} (t_i - o_i)^2 / (m_1 - p_c) \tag{2.47}$$

where, p_c is the number of the regression model coefficients. The confidence interval at a prediction point τ can be computed using the prediction error variance estimated in Eq. (2.47), the ANN inputs and the desired confidence degree that follows t-Student's distribution with $(m_1 - p_c)$ degrees of freedom [64].

Re-sampling technique is usually preferred to the other two techniques as the doubling of the ANN's outputs and the use of linear activation function at the output layer are voided. Finally, Silva et al. [65] propose the re-sampling technique as the most suitable for the estimation of the confidence interval with a high degree of confidence.

It is noted that the second method has been the most applied [65] with some empirical enhancements in some cases [66] while other similar probabilistic methods have been recently proposed [67, 68].

2.3.5 Compression of the Input Data

Usually, a large number of inputs is used in short-term load forecasting models. Compression techniques, such as the principal component analysis (PCA), can be applied [21, 69] to suppress the inputs of the model. Let us assume that X is a $N\text{x}p_{in}$ matrix whose rows contain the input vectors while the p_{in} columns of X represent the properly transformed input variables so as their means equal to zero. Also, vector $\overrightarrow{a}$ is defined as the vector that maximizes variance of X projection on it $(X \cdot \overrightarrow{a})$. The variance of $\overrightarrow{a}$ is defined as:

$$\sigma_a^2 = \left(X \cdot \overrightarrow{a}\right)^T \cdot \left(X \cdot \overrightarrow{a}\right) = \overrightarrow{a}^T \cdot X^T \cdot X \cdot \overrightarrow{a} = \overrightarrow{a}^T \cdot V \cdot \overrightarrow{a}$$

where $V = X^T \cdot X$ is the covariance matrix of X.

${\sigma_a}^2$ the variance is a function of both $\overrightarrow{a}$ and X and it must be maximized. In order to avoid unconstraint increase of $\overrightarrow{a}$ it is normalized according to the constraint, $\overrightarrow{a}^T \cdot \overrightarrow{a} = 1$. In this case the optimization problem is transformed to the maximization of the quantity $f_{opt} = \overrightarrow{a}^T \cdot V \cdot \overrightarrow{a} - \lambda \cdot \left(\overrightarrow{a}^T \cdot \overrightarrow{a} - 1\right)$, where λ is a Lagrange multiplier. By setting the derivative of f_{opt} with respect to $\overrightarrow{a}$ equal to zero the maximization problem is finally reduced to the estimation of the eigenvalues of the covariance matrix V as follows:

$$\frac{\partial f_{opt}}{\partial \overrightarrow{a}} = 2 \cdot V \cdot \overrightarrow{a} - 2 \cdot \lambda \cdot \overrightarrow{a} = 0 \Rightarrow (V - \lambda \cdot I) \cdot \overrightarrow{a} = 0$$

If the eigenvalues of V are ranked in descending order, the first principal component $\overrightarrow{a}_1$ is the eigenvector associated with the largest eigenvalue λ_1 of the covariance matrix V, the second principal component $\overrightarrow{a}_2$ is the eigenvector associated with the second largest eigenvalue λ_2 etc. The obtained eigenvectors are orthogonal each other because the V matrix is real and symmetric.

A basic property of this method is that if the data are projected to the first k eigenvectors then the variance of the projected data can be expressed as the sum of the eigenvalues, $\sum_{j=1}^{k} \lambda_j$. Equivalently, the squared error in matrix X approximation by using only k eigenvectors can be expressed as $\sum_{j=k+1}^{p_{in}} \lambda_j / \sum_{j=1}^{p_{in}} \lambda_j$. Increasing the number, k, of the principal components the respective mean square error $J_k = \sum_{m=1}^{N} \left\| \sum_{j=1}^{k} \lambda_j \cdot \overrightarrow{a}_j - \overrightarrow{x}_m \right\|^2$ decreases, where $\overrightarrow{x}$ are the column-vectors of X^T.

In case of multidimensional data sets with strongly correlated elements, 5 up to 10 principal elements are necessary to achieve at least a 90 % accumulated percentage of explained variance.

2.4 Short Term Foresting of the Greek Power System Load Demand by Using ANNs

In this paragraph feed forward multilayer ANNs are used for load forecasting purposes in Greek electric power system. Different configurations of ANNs and issues related with their optimization are examined.

2.4.1 Basic ANN Configuration

The basic ANN configuration used in the following analysis comprises 71 input variables [44, 46]. Assuming that the hourly average values of the load of the d-th day of the year are to be forecasted the input and output variables of the basic ANN model are grouped as in Table 2.1.

It is noted that according to Kolmogorov's theorem [56] ANNs comprising one hidden layer with adequate number of neurons can describe any nonlinear system. Hence, ANNs with one hidden layer are used and the optimal number of the neurons of the hidden layer should be found. The basic ANN configuration is shown in Fig. 2.3.

In the next paragraphs several case studies are presented concerning the use of different training methods, inputs/outputs of the ANN, confidence interval estimation etc. In each of the examined cases the general heuristic methodology shown in Fig. 2.4 is used to determine the optimal parameters of the ANN model.

A brief description of the basic steps of the methodology of Fig. 2.4 is provided next.

(a) *Data selection*: The input and output variables of the forecasting model are selected. The inputs and the outputs of the basic ANN configuration are as described Table 2.1 while scenarios concerning the use of different inputs and outputs will be examined in the next paragraphs.

(b) *Data pre-processing*: Data normality is examined and outliers are modified or extracted from the data (noise suppression). Next, input and output variables are normalized according to Eq. (2.3) in order to avoid saturation.

(c) *Main procedure*: A large number of model parameters combinations are used in order to optimize the performance of the ANN model. The obtained forecasting models are tested and evaluated. Mean absolute percentage error (*MAPE*) is used for evaluation purposes.

MAPE is estimated by:

$$MAPE_{ev} = 100\% \cdot \frac{1}{m_{ev}} \cdot \sum_{d=1}^{m_{ev}} \sum_{i=1}^{24} \frac{\left| \widehat{L}(d,i) - L(d,i) \right|}{L(d,i)} \tag{2.48}$$

Where, $L(d,i)$ is the measured value of load demand at the i-th hour of d-th day of the evaluation set, $\widehat{L}(d,i)$ the respective forecasted load and m_{ev} is the number of the data vectors of the evaluation set.

(d) *ANN test*: The load demand is finally estimated for the days of the test set by using the optimal values of the parameters that lead to the lowest MAPE.

Next, the above optimization method is applied to the basic ANN configuration of Fig. 2.3. Several parameters of the ANN model and the training method are optimized and they are listed next:

Table 2.1 Inputs and outputs of the basic ANN configuration

ANN_Inp. 1	The hourly load values of the two previous days: $L(\mathrm{d}-1,1), \ldots, L(\mathrm{d}-1,24), L(\mathrm{d}-2,1), \ldots, L(\mathrm{d}-2,24)$ (in MW),
ANN_Inp. 2	The maximum value of the 3-h average temperature measurements in Athens for the current day (d) and the previous one (d − 1): $\max\_Temp_{\mathrm{Ath}}(\mathrm{d})$, $\min\_Temp_{\mathrm{Ath}}(\mathrm{d})$, $\max\_Temp_{\mathrm{Ath}}(\mathrm{d}-1)$, $\min\_Temp_{\mathrm{Ath}}(\mathrm{d}-1)$, respectively (°C),
ANN_Inp. 3	The maximum value of the 3-h average temperature measurements in Thessalonica for the current day and the previous one (d − 1): $\max\_Temp_{\mathrm{Th}}(\mathrm{d})$, $\min\_Temp_{\mathrm{Th}}(\mathrm{d})$, $\max\_Temp_{\mathrm{Th}}(\mathrm{d}-1)$, $\min\_Temp_{\mathrm{Th}}(\mathrm{d}-1)$, respectively (°C),
ANN_Inp. 4	The difference between the maximum value of the 3-h average temperature measurements of the current day (d) and the previous one (d − 1) for Athens and Thessalonica: $\Delta Temp_{\mathrm{Ath}} = \max\_Temp_{\mathrm{Ath}}(\mathrm{d}) - \max\_Temp_{\mathrm{Ath}}(\mathrm{d}-1)$ $\Delta Temp_{\mathrm{Th}} = \max\_Temp_{\mathrm{Th}}(\mathrm{d}) - \max\_Temp_{\mathrm{Th}}(\mathrm{d}-1)$
ANN_Inp. 5	The temperature dispersion from the temperature of the comfortable living conditions for Athens, for the current and the previous day ${T_{Ath}}^2(d)$, ${T_{Ath}}^2(d-1)$, respectively. Temperature dispersion is calculated as: $T^2_{dispersion} = \begin{cases} (T_c - T)^2, & T < T_c \\ 0, & T_c < T < T_h \\ (T - T_h)^2, & T_h < T \end{cases}$ Where, $T_c = 18$ °C, $T_h = 25$ °C
ANN_Inp. 6	The temperature dispersion from comfortable living conditions temperature for Thessalonica for the current and the previous day ${T_{Th}}^2(d)$, ${T_{Th}}^2(d-1)$, respectively,
ANN_Inp. 7	Seven digit binary numbers for the coding of the weekdays, e.g. Monday is coded with 1000000, Tuesday with 0100000, etc
ANN_Inp. 8	Two sinusoidal functions ($\cos(2\pi d/T)$, $\sin(2\pi d/T)$), which express the seasonal behavior of the current day. Where, T is the total number of the days of the year
ANN_Output	The output variables are the 24 hourly average values of the load demand of the current day, $\widehat{L}(\mathrm{d},1), \ldots, \widehat{L}(\mathrm{d},24)$

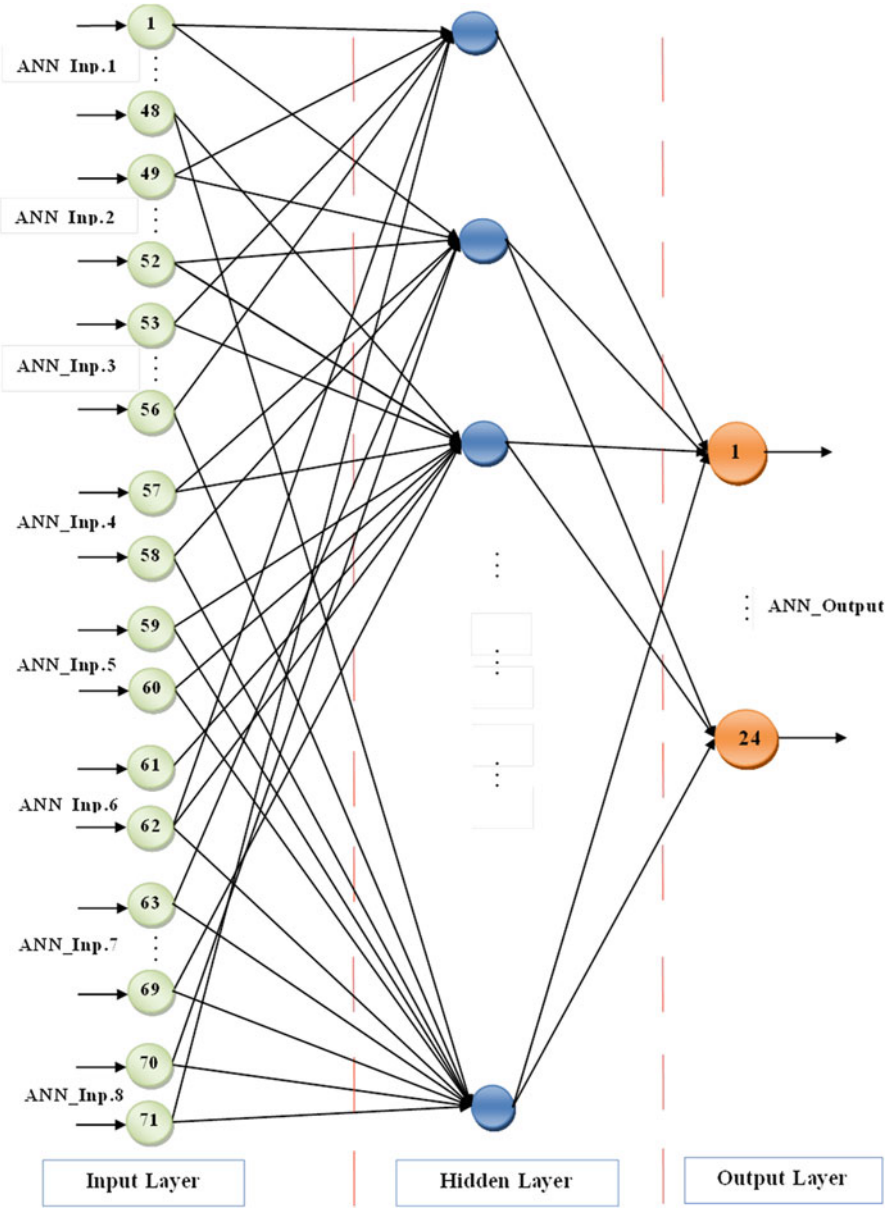


Fig. 2.3 Basic ANN configuration

- the number of the neurons of the hidden layer ranging from 20 up to 70.
- the initial value of n, n_0, and the time parameter T_n of the training rate, which get values from the sets {0.1, 0.2, ..., 0.9} and {1,000, 1,200, ..., 2,000}, respectively.

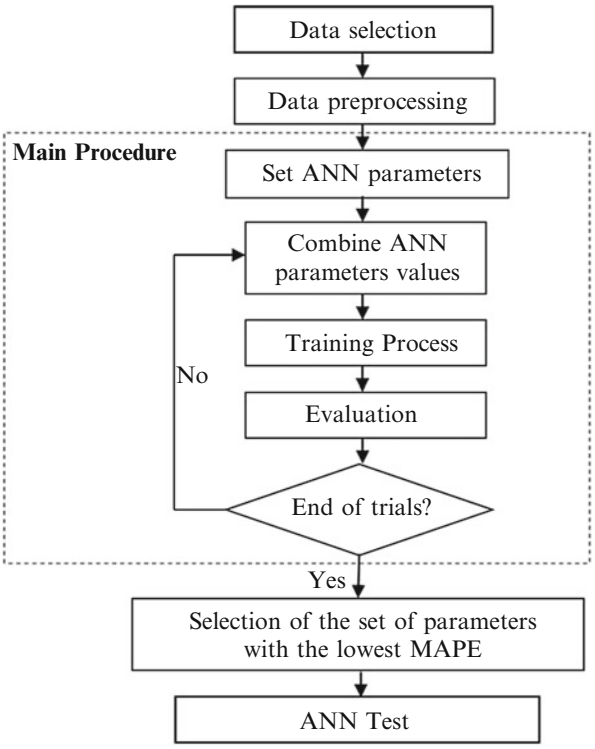


Fig. 2.4 Flowchart of the ANN optimization and the evaluation process

- the initial value of parameter a and the time parameter T_a of the momentum term, which get values from the sets {0.1,0.2, . . . ,0.9} and {1,000, 1,200, . . . , 2,000}, respectively,
- the type and the parameters of the activation functions used at the hidden and output layers. The type of the activation functions can be the *hyperbolic tangent*, *linear or logistic*, while their parameters a_1, a_2 get values from the set {0.1,0.2, . . . , 0.5} and parameters b_1, b_2 from the set {0.0, ±0.1,±0.2}.

Stopping criteria are defined after a few trials as max_*epochs* = 10,000, $limit_1 = 10^{-5}$, $limit_2 = 10^{-5}$.

In this study the combinations resulting from the above assumptions account to 836,527,500 and they cannot be practically examined. To this end, calibration process through successive variations of the parameters values is applied in order to determine the optimal or nearly optimal values of the parameters.

First, the number of the hidden neurons is varied from 20 up to 70, while the remaining parameters are assigned with fixed values ($a_0 = 0.4$, $T_a = 1{,}800$, $\eta_0 = 0.5$, $T_\eta = 2{,}000$, the type of activation functions at hidden and output layers is hyperbolic tangent with $a_1 = a_2 = 0.25$, $b_1 = b_2 = 0.0$). *MAPE* of the training, evaluation and

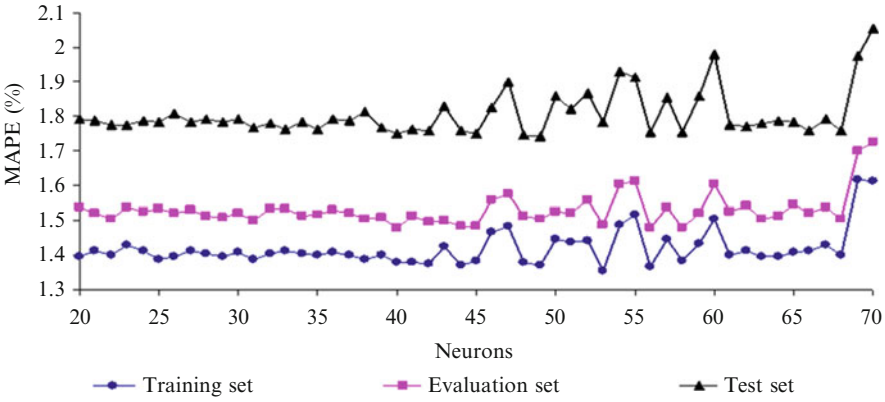


Fig. 2.5 *MAPE* (%) of all sets versus the number of hidden neurons. ($a_0 = 0.4$, $T_a = 1{,}800$, $\eta_0 = 0.5$, $T_\eta = 2{,}000$, activation functions in both layers: hyperbolic tangent, $a_1 = a_2 = 0.25$, $b_1 = b_2 = 0.0$)

test sets versus the number of the hidden neurons are presented in Fig. 2.5. The *MAPE* of the evaluation and tests set keep step with the respective one of the training set and the following relationship is valid within the entire examined range of hidden neurons number:

$$MAPE_{training} < MAPE_{evaluation} < MAPE_{test} \tag{2.49}$$

MAPE of the evaluation set is minimized for 45 neurons at hidden layer while it rapidly increases when the number of the hidden neurons increases.

Next, the initial value η_0 and the time parameter T_η of the training rate are varied while the other parameters remain constant (hidden neurons = 45, $a_0 = 0.4$, $T_a = 1{,}800$, hyperbolic tangent activation functions are used at hidden and output layers with $a_1 = a_2 = 0.25$, $b_1 = b_2 = 0.0$). It appears in Fig. 2.6 that the results obtained for the *MAPE* of the evaluation set are satisfactory for $0.5 \leq \eta_0 \leq 0.8$ and $1{,}000 \leq T_\eta \leq 1{,}400$. The lowest MAPE is obtained for $\eta_0 = 0.5$, $T_\eta = 2{,}000$. It is mentioned that *MAPE* increases dramatically for $\eta_0 \geq 0.7$ and $T_\eta \geq 1{,}600$.

In the next step of the calibration process, the initial value a_0 and the time parameter T_a of the momentum term are simultaneously varied, while the rest of the parameters are kept constant (hidden neurons = 45, $\eta_0 = 0.5$, $T_\eta = 2{,}000$, activation functions at hidden and output layers: hyperbolic tangent with $a_1 = a_2 = 0.25$, $b_1 = b_2 = 0.0$). In this case, the obtained *MAPE* of the evaluation set is satisfactory for $a_0 \geq 0.6$ and $T_a \geq 1{,}600$, while the lowest MAPE is obtained for $a_0 = 0.9$, $T_a = 2{,}000$. It is mentioned that *MAPE* increases dramatically for $a_0 \leq 0.5$.

Similarly, it is found that the ANN gives better results if the hyperbolic tangent activation function with parameters $a_1 = a_2 = 0.25$ and $b_1 = b_2 = 0.0$, is used at both layers. The results obtained by using different activation functions are registered in Table 2.2.

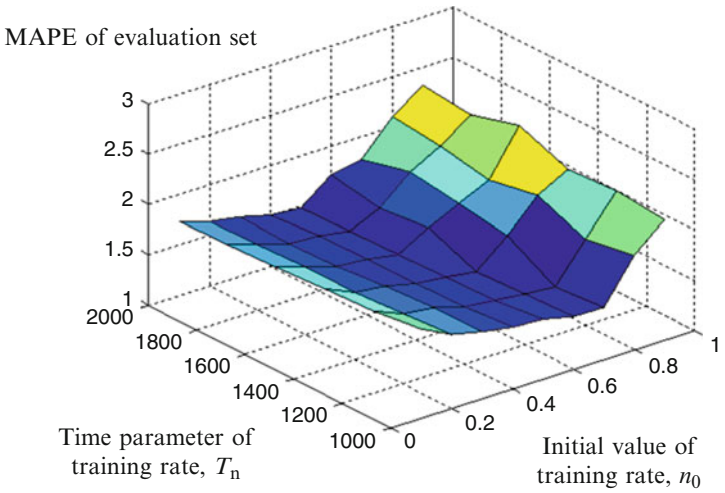


Fig. 2.6 *MAPE* (%) of the evaluation set versus parameters T_n and n_0. ($\eta_0 = \{0.1, 0.2, \ldots, 0.9\}$, $T_\eta = \{1{,}000, 1{,}200, \ldots, 2{,}000\}$, hidden neurons: 45, $a_0 = 0.4$, $T_a = 1{,}800$, activation function used in both layers: hyperbolic tangent, $a_1 = a_2 = 0.25$, $b_1 = b_2 = 0.0$)

Table 2.2 MAPE (%) of (A) training set, (B) evaluation set, (C) test set for different activation functions

Activation function of output layer	Activation function of hidden layer								
	Hyperbolic sigmoid			Hyperbolic tangent			Linear		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
Hyperbolic sigmoid	2.030	2.048	2.625	1.510	1.621	1.817	1.788	1.850	2.091
Hyperbolic tangent	1.671	1.737	2.042	1.383	1.482	1.749	1.900	1.987	2.200
Linear	1.603	1.691	1.903	1.390	1.522	1.747	1.936	2.023	2.194

Number of hidden neurons: 45, $a_0 = 0.4$, $T_a = 1{,}800$, $\eta_0 = 0.5$, $T_\eta = 2{,}000$, $a_1 = a_2 = 0.25$, $b_1 = b_2 = 0.0$

The above described process finally leads to the following optimal intervals of the parameters values: 40 up to 50 neurons at the hidden layer, $a_0 = 0.8 - 0.9$, $T_a = 1{,}800$-$2{,}000$-$2{,}200$, $\eta_0 = 0.5 - 0.6$, $T_\eta = 1{,}000$-$1{,}200$-$1{,}400$, activation functions of both layers: hyperbolic tangent with $a_1 = a_2 = 0.20$-0.25-0.30, $b_1 = b_2 = 0$.

The minimum *MAPE* of the evaluation set is 1.48 % and it is obtained for an ANN with 45 neurons in the hidden layer, $a_0 = 0.9$, $T_a = 2{,}000$, $\eta_0 = 0.5$, $T_\eta = 2{,}000$, $a_1 = a_2 = 0.25$ and $b_1 = b_2 = 0$ with hyperbolic tangent activation functions in hidden and output layer.

2.4.2 Comparison of Different Training Algorithms

In this case study the basic ANN structure presented in paragraph 2.4.1 is used while parameters, such as the number of neurons of the hidden layer, activation functions, weighting factors, learning rate, momentum term, etc. are determined by a calibration methodology through an extensive search. The performance of each set of parameters is evaluated by the MAPE index of the evaluation data set. The ANN parameter calibration process is based on the philosophy of the heuristic process outlined in Fig. 2.4 and it is repeated for 14 different training algorithms. The main goals are:

- the determination of the optimal ANN structure (number of neurons in the hidden layer, learning rate initial value, etc.) for each of the examined training algorithms,
- the comparison of the training algorithms in terms of MAPE and computation time minimization
- the selection of the training algorithm with the best performance.

The examined training algorithms are registered in Table 2.3. Several parameters of the ANN model should be optimized for each of the case studies included in Table 2.3. The common parameters in all cases are:

- the number of the hidden neurons, N_n,
- the type of the activation functions (hyperbolic tangent, logistic, linear),

Table 2.3 Examined training algorithms

No.	Training algorithms
1	Stochastic training with learning rate and momentum term (decreasing exponential functions)
2	Stochastic training, use of adaptive rules for the learning rate and the momentum term
3	Stochastic training, constant learning rate
4	Batch mode, constant learning rate
5	Batch mode with learning rate and momentum term (decreasing exponential functions)
6	Batch mode, use of adaptive rules for the learning rate and the momentum term
7	Batch mode, conjugate gradient algorithm with Fletcher-Reeves equation
8	Batch mode, conjugate gradient algorithm with Fletcher-Reeves equation and Powell-Beale restart
9	Batch mode, conjugate gradient algorithm with Polak-Ribiere equation
10	Batch mode, conjugate gradient algorithm with Polak-Ribiere equation and Powell-Beale restart
11	Batch mode, scaled conjugate gradient algorithm
12	Batch mode, resilient algorithm
13	Batch mode, quasi-Newton algorithm
14	Batch mode, Levenberg-Marquardt algorithm

- the parameters of the activation functions,
- the maximum number of the training epochs.

The additional parameters used in the each of the examined training methods are listed next.

Methods *1–6*:

- the time parameter, T_n, and the initial value of the learning rate, n_0,
- the time parameter, T_α, and the initial value of the momentum term, α_0 (not applied to methods 3, 4).

M*ethods 7–10*:

- the initial value of s,
- the number of iterations, T_{bn}, for the calculation of the basic vector,
- the number of iterations, T_{trix}, used for the trisection according to the golden section method for output error minimization.

M*ethod 11*:

- Parameters σ and λ_0,

Method *12*:

- the increasing and decreasing factors δ_1 and δ_2 (see Eq. (2.24))

Method *14*:

- the initial value of λ, $\lambda(0)$, and the multiplicative parameter β.

The number of all possible combinations of the values of the parameters that should be examined ranges in the order of some hundreds of millions. Hence, all possible combinations cannot be practically examined and a gradual calibration process is applied. The intervals of the parameters values used in this study are registered in Table 2.4 while the calibrated parameters obtained for the 14 examined scenarios of Table 2.3 are registered in Table 2.6.

The minimum MAPE of the evaluation data set is obtained by the stochastic training algorithm with adaptive rules for the learning rate and the momentum term and the scaled conjugate gradient algorithm. The MAPE of the test data set is also satisfying while the stochastic training algorithm with decreasing learning rate and momentum term leads to slightly better results (see Table 2.5).

The proportion of the computation times the first 11 training algorithms require, is: $3.2 \div 3 \div 1.2 \div 4.0 \div 3.5 \div 1.4 \div 10 \div 12 \div 12 \div 12 \div 1$.

Based on the above the scaled conjugate gradient algorithm is proposed.

An indicative daily load curve and the respective load curve forecasted by the proposed ANN are presented in Fig. 2.7. The MAPE of the specific daily load curve is 1.173 %.

Table 2.4 Intervals of the training parameters values

Training algorithm	Intervals of the training parameters values
1–2	$\alpha_0 = 0.1,0.2,\ldots,0.9$, $T_\alpha = 1{,}000,1{,}200,\ldots,2{,}000$, $\eta_0 = 0.1,0.2,\ldots,0.9$, $T_\eta = 1{,}000,1{,}200,\ldots,2{,}000$
3	$\eta_0 = 0.01,0.02,\ldots,0.1,0.2,\ldots,3$
4	$\eta_0 = 0.1,0.2,\ldots,3$
5–6	$\alpha_0 = 0.1,0.2,\ldots,0.9$, $T_\alpha = 1{,}200,1{,}500,\ldots,6{,}000$, $\eta_0 = 0.1,0.2,\ldots,0.9$, $T_\eta = 1{,}200,1{,}500,\ldots,6{,}000$
7, 9	$s = 0.04,0.1,0.2$, $T_{bv} = 20, 40$, $T_{trix} = 50, 100$, $e_{trix} = 10^{-6}, 10^{-5}$
8, 10	$s = 0.04,0.1,0.2$, $T_{bv} = 20, 40$, $T_{trix} = 50, 100$, $e_{trix} = 10^{-6}, 10^{-5}$, $\lim_{orthogonality} = 0.1,0.5,0.9$
11	$\sigma = 10^{-3}, 10^{-4}, 10^{-5}$, $\lambda_0 = 10^{-6}, 10^{-7}, 5 \cdot 10^{-8}$
12	$\delta_1 = 0.1,0.2,\ldots,0.5$, $\delta_2 = 0,1,0.2,\ldots,2$
13	–
14	$\lambda(0) = 0.1, 0.2, \ldots, 1, 2, \ldots, 5$, $\beta = 2, 3, \ldots, 9, 10, 20, \ldots, 50$
Common parameters	$N_n = 20,21,\ldots,70$, activation function for hidden and output layer = hyperbolic tangent, linear, logistic, $a_1 = 0.1,0.2,\ldots,0.5$, $a_2 = 0.1,0.2,\ldots,0.5$, $b_1 = 0.0, \pm 0.1, \pm 0.2$, $b_2 = 0.0, \pm 0.1, \pm 0.2$

Table 2.5 MAPE (%) of training, evaluation and test sets for all examined ANN training algorithms

	MAPE (%)		
Training algorithm	Training set	Evaluation set	Test set
1	1.383	1.482	1.749
2	1.311	1.475	1.829
3	1.372	1.489	1.833
4	2.356	2.296	2.602
5	2.300	2.294	2.783
6	2.019	2.026	2.475
7	1.798	1.831	2.147
8	2.544	2.595	3.039
9	2.545	2.600	3.035
10	2.544	2.600	3.035
11	1.294	1.487	1.781
12–14	#	#	#

2.4.3 Study of ANN Inputs

In this paragraph, several scenarios regarding the input variables of the model are examined. A methodology based on the heuristic methodology introduced in paragraph 2.4.1 is used to obtain the optimal set of input variables and ANN parameters. The steps followed are briefly described next.

Table 2.6 Number of hidden layer neurons, activation functions and training parameters for all examined ANN training algorithms

Training algorithm	Neurons	Activation functions	Rest parameters
1	45	$f_l = \tanh(0.25x)$, $f_o = \tanh(0.25x)$	$\alpha_0 = 0.4$, $T_\alpha = 1{,}800$, $\eta_0 = 0.5$, $T_\eta = 2{,}000$, $e = 10^{-5}$ max_epochs = 10,000
2	30	$f_l = \tanh(0.40x)$, $f_o = 1/(1 + \exp(-0.25x))$	$\alpha_0 = 0.7$, $T_\alpha = 1{,}800$, $\eta_0 = 0.5$, $T_\eta = 1{,}300$, $e = 10^{-5}$, max_epochs = 12,000
3	48	$f_l = \tanh(0.50x)$, $f_o = \tanh(0.25x)$	$\eta_0 = 0.1$, $e = 10^{-5}$, max_epochs = 10,000
4	48	$f_l = \tanh(0.40x)$, $f_o = 0.40x$	$\eta_0 = 2.0$, $e = 10^{-5}$, max_epochs = 12,000
5	25	$f_l = \tanh(0.50x)$, $f_o = 0.25x$	$\alpha_0 = 0.9$, $T_\alpha = 6{,}000$, $\eta_0 = 0.9$, $T_\eta = 6{,}000$, $e = 10^{-5}$, max_epochs = 12,000
6	22	$f_l = \tanh(0.50x)$, $f_o = 0.25x$	$\alpha_0 = 0.9$, $T_\alpha = 6{,}000$, $\eta_0 = 0.9$, $T_\eta = 6{,}000$, $e = 10^{-5}$, max_epochs = 12,000
7	43	$f_l = \tanh(0.40x)$, $f_o = 0.20x$	$s = 0.04$, $T_{bv} = 20$, $T_{trix} = 50$, $e = 10^{-5}$, max_epochs = 5,000
8	43	$f_l = \tanh(0.40x)$, $f_o = 0.20x$	$s = 0.04$, $T_{bv} = 20$, $T_{trix} = 50$, $e = 10^{-5}$, max_epochs = 5,000, $\lim_{orthogonality} = 0.9$
9	43	$f_l = \tanh(0.40x)$, $f_o = 0.20x$	$s = 0.04$, $T_{bv} = 20$, $T_{trix} = 50$, $e = 10^{-5}$, max_epochs = 5,000
10	43	$f_l = \tanh(0.40x)$, $f_o = 0.20x$	$s = 0.04$, $T_{bv} = 20$, $T_{trix} = 50$, $e = 10^{-5}$, max_epochs = 5,000, $\lim_{orthogonality} = 0.9$
11	52	$f_l = \tanh(0.50x)$, $f_o = \tanh(0.25x)$	$\sigma = 10^{-5}$, $\lambda_0 = 5 \times 10^{-8}$, $e = 10^{-5}$, max_epochs = 10,000
12–14	#	#	No convergence

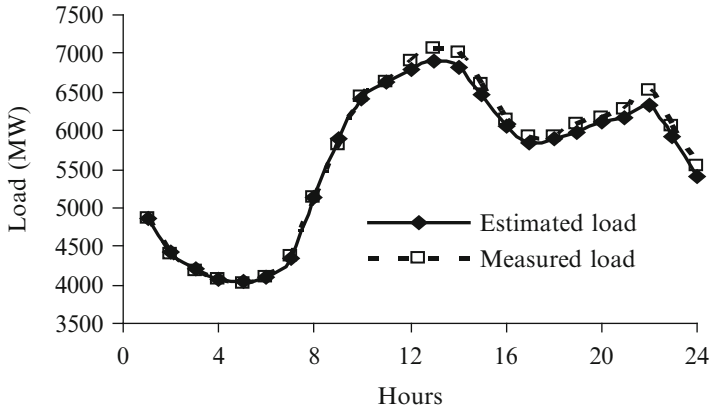


Fig. 2.7 Chronological load curve (measured and estimated load) of 1 day of the test set (Thursday, June 8, 2000)

(a) *Data selection*: In this step the input variables for the short term load forecasting model are defined.

- *1st scenario* (*basic*): The input and output variables of the basic ANN configuration as described in paragraph 2.4.1 are used.
- *2nd scenario*: This scenario is same with the 1st except from the use of a seven-digit binary coding for the weekdays in replacement of the sinusoidal functions ($\cos(2\pi d/7)$, $\sin(2\pi d/7)$).
- *3rd scenario*: This scenario is same with the 2nd scenario except from the use of 3-h average temperature values of the d-th day and the previous one (for Athens and Thessalonica) in replacement of the group of inputs ANN_Inp.2 up to ANN_Inp.6. In this case the input vector of t he ANN comprises 84 elements.
- *4th scenario*: This scenario is same with the 1st scenario except from the fact that only the hourly load measurements of the previous day ($d-1$) are used. In this scenario, the input vector comprises 47 elements.
- *5th scenario*: This scenario is same with the 1st scenario except from the fact that hourly load measurements for 3 days before the prediction day are used. In this case the input vector comprises 95 elements.
- *6th scenario*: In this scenario, PCA is applied to the input data used in the 2nd scenario eventually suppressing the input variables from 66 to 6 with a 99 % accumulated percentage of explained variance according to Kaiser's criterion [21].
- *7th scenario*: In this scenario, PCA is applied to the inputs used in 2nd scenario except from the ones dealing with the week days and the season of the year. The input variables are eventually suppressed from 66 to 19 with a 99 % accumulated percentage of explained variance according to Kaiser's criterion.

(b) *Data pre-processing*: As described in paragraph 2.4.1.
(c) *Main procedure*: The ANN is trained by using the scaled conjugate gradient algorithm (SCGA).

The optimal results obtained for each of the seven examined scenarios are summarized in Table 2.7. The basic scenario (no 1 of Table 2.7) leads to the minimum MAPE of the evaluation set. The second and the third scenario lead to similar results. It is worth mentioning that if MAPE of the test set is considered then the third scenario becomes the optimal while the first and the second scenarios perform similarly. MAPE increases in scenarios 4 and 5 while further increase occurs in scenarios 6 and 7.

The proportion of the computation times of the seven examined scenarios is: $1 \div 0.98 \div 1.1 \div 0.8 \div 1.3 \div 0.4 \div 0.55$. The use of compression techniques decreases the computational time significantly, but MAPE is considerably increased. According to the obtained results the first scenario is suggested for this case study.

2.4.4 Study of ANN Outputs

The outputs of the ANN models examined so far are the 24 hourly load values that should be forecasted. This classical design imposes the use of ANNs with 24 output variables. An alternative solution comprises 24 different ANNs that are separately used for the prediction of each of the 24 hourly load demands. The performance of each of the 24 ANNs can be affected in a different way by the inputs it uses. Hence, various scenarios will be studied next regarding the inputs used. In all cases the scaled conjugate gradient training algorithm is used with its most crucial parameters being properly calibrated. Moreover, a large number of combinations of ANN parameters such as the number of neurons, the type of the activation functions, etc. are explored in order to finally obtain the optimal configuration of the short term load forecasting model and the optimal set of training parameters values. The performance of each of the examined forecasting models is evaluated by using the MAPE of the evaluation data set.

The scenarios examined are:

- 1st *scenario*: The ANN described in paragraph 2.4.1 trained with scaled conjugate gradient training algorithm is used.
- 2nd *scenario*: 24 ANNs are used; one for each of the predicted hourly average loads. The input variables of the ANNs are the same with those used in the 1st scenario except from the seven digit binary coding of the weekdays that is replaced by the two sinusoidal functions $\cos(2\pi d/7)$ and $\sin(2\pi d/7)$.
- 3rd *scenario*: In this scenario, PCA is applied to the input variables used in 2nd scenario except from the ones dealing with the weekdays and the season of the year. The inputs are eventually suppressed from 66 to 19 with a 99 % accumulated percentage of explained variance according to Kaiser's criterion [21].

Table 2.7 Results obtained from the examined scenarios and optimally calibrated ANN parameters

Scenario	MAPE (%)			Optimal number of neurons—range examined	Activation functions
	Training set	Evaluation set	Test set		
1	1.294	1.487	1.781	52 (20–70)	$f_1 = \tanh(0.50x)$, $f_o = \tanh(0.25x)$
2	1.315	1.516	1.776	48 (20–70)	$f_1 = \tanh(0.50x)$, $f_o = \tanh(0.25x)$
3	1.293	1.504	1.717	39 (20–80)	$f_1 = \tanh(0.50x)$, $f_o = \tanh(0.25x)$
4	1.574	1.674	2.224	21 (20–70)	$f_1 = \tanh(0.30x)$, $f_o = 1/(1 + \exp(-0.20x))$
5	1.524	1.804	1.808	76 (30–90)	$f_1 = \tanh(0.25x)$, $f_o = 1/(1 + \exp(-0.25x))$
6	1.901	1.960	2.570	21 (15–60)	$f_1 = \tanh(0.30x)$, $f_o = 1/(1 + \exp(-0.30x))$
7	1.458	1.583	1.900	28 (15–60)	$f_1 = \tanh(0.50x)$, $f_o = \tanh(0.25x)$

- 4th *scenario*: The inputs of the i-*th* ANN (used for the forecasting of the average load demand of the i-*th* day hour) are:
  - The average load demands of the i-*th*, (i − 1)-*th* and (i − 2)-*th* hour of the previous 2 days of the day the load is forecasted.
  - The mean temperatures of Athens and Thessalonica for the current day and the previous one.
  - Two sinusoidal functions representing the week day.
  - Two sinusoidal functions representing the seasonal behaviour of the load.

  In this scenario the input vector comprises 14 elements.
- 5th *scenario*: It is the same with the 4th scenario, but Principal Components Analysis is applied to all inputs except from those dealing with the weekly and seasonal behaviour of the load. The inputs are suppressed in a way that 99 % accumulated percentage of explained variance according to Kaiser's criterion. The number of the remaining inputs after PCA application is different for each ANN.

The *MAPE* estimated for the training, evaluation and test data sets and the output of the ANN used to forecast the load demand at the 12-th hour of the day of the year, are shown in Table 2.8. Moreover, the optimally calibrated parameters of the ANN are registered in the same Table. It is noted here, that the parameters of the training algorithm, the type and the parameters of the activation functions of the hidden and the output layers are the same with those used in the 1st scenario.

Respectively, the *MAPE* of training, evaluation and test sets and for all examined scenarios are shown in Fig. 2.8 and Table 2.9. For scenarios 2–5 the average of the *MAPE* of the 24 ANNs is calculated.

The proportion of the computation times of the examined scenarios is: $1 \div 0.4 \div 0.05 \div 0.15 \div 0.03$. This indicates that the computation time decreases significantly when the dimension of the input vector decreases. Even larger decrease could be achieved in scenarios 2–5 if 24 different computers were used (parallel processing) for the training of each ANN. However, the 1st scenario leads to the minimum *MAPE* of the evaluation data set while the second and the third scenario lead to slightly worse results. *MAPE* of test set remains almost the same in 1st and 2nd scenario. In the 4th scenario, leaving out some of the input data led to worse results. The application of PCA (3rd and 5th scenarios) decreases significantly the computation time but the performance of the forecasting model deteriorates. This becomes even more evident if MAPE of the test data set is considered. Based on the previous observations the data compression is not suggested while the 2nd scenario presents satisfactory behavior regarding MAPE and it could be preferred to the 1st scenario if minimization of the computation time is of high importance.

Table 2.8 MAPE (%) of training, evaluation and test sets for the ANN forecasting the load at 12:00

Scenario	MAPE (%) of training set	MAPE (%) of evaluation set	MAPE (%) of test set	Neurons—Range of examined neurons	Activation functions
2	1.347	1.824	1.847	32 (10–60)	$f_1 = \tanh(0.50x)$, $f_2 = \tanh(0.25x)$
3	1.492	1.876	1.924	17 (2–25)	$f_1 = \tanh(0.50x)$, $f_2 = \tanh(0.25x)$
4	1.747	2.210	2.620	40 (10–60)	$f_1 = \tanh(0.25x)$, $f_2 = \tanh(0.25x)$
5	2.120	2.404	2.953	14 (2–25)	$f_1 = \tanh(0.50x)$, $f_o = \tanh(0.25x)$

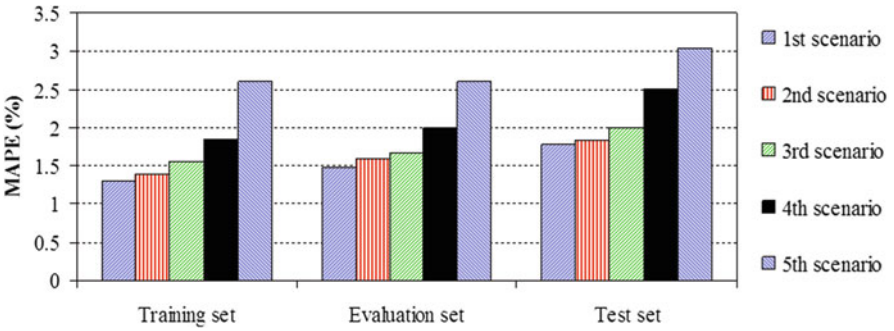


Fig. 2.8 MAPE (%) of training, evaluation and test set for the interconnected Greek power system and the 5 different scenarios of output variables

Table 2.9 MAPE (%) of training, evaluation and test sets for the 24-h load prediction

	MAPE (%)		
Scenario	Training set	Evaluation set	Test set
1	1.294	1.487	1.781
2	1.390	1.603	1.830
3	1.565	1.667	1.994
4	1.850	1.989	2.503
5	2.596	2.598	3.043

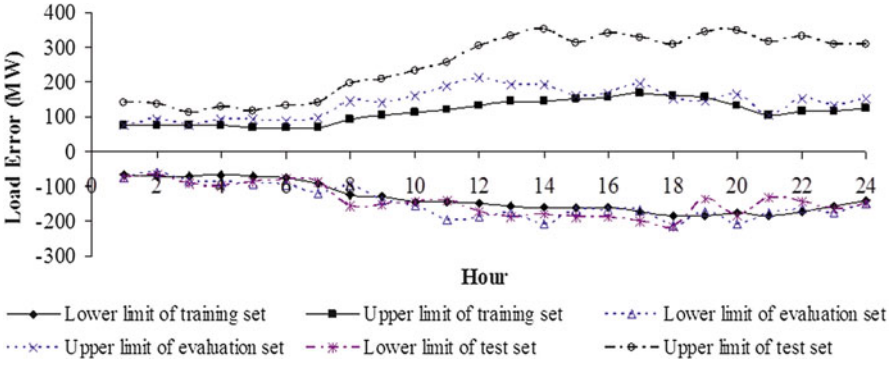


Fig. 2.9 90 % confidence interval limits with respect to the training, evaluation and test sets for the best ANN model for 8-6-2000 in Greek interconnected power system

2.4.5 Estimation of the Confidence Interval

Using the basic ANN configuration described in paragraph 2.4.1 the 90 % confidence interval is estimated using the re-sampling technique with the probability in each tail equal to 5 %. In Fig. 2.9 the prediction errors of a typical summer day for Greek interconnected power system of the year 2000 (Thursday 8-6-2000) are presented for the training, evaluation and test sets respectively, while in Fig. 2.10

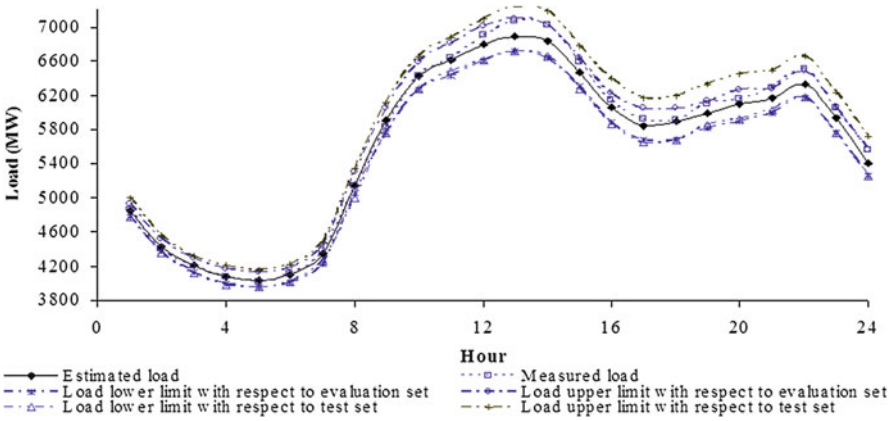


Fig. 2.10 Chronological active load curves of the measured load, the estimated load, the estimated load with the 5 % lower limit with respect to evaluation set, the estimated load with the 5 % upper limit with respect to evaluation set, the estimated load with the 5 % lower limit with respect to test set, the estimated load with the 5 % upper limit with respect to test set for the best ANN model for 8-6-2000 in Greek interconnected power system

the respective measured and estimated load values are presented together with the 90 % confidence intervals of the evaluation and the test data sets.

It is observed in Fig. 2.10 that the lower limits of the confidence intervals for the three data sets are quite similar. The ratio between the lower hourly errors of the test set to the respective one of the evaluation set varies between 0.71 and 1.60, while the mean value is equal to 1.00. However, the upper limit of the confidence intervals for the test set is almost the double of the respective one of the evaluation set. The ratio between the upper hourly errors of the test set to the respective one of the evaluation set varies between 1.22 and 3.02, while the mean value is equal to 1.78. It is observed in Fig. 2.10 that the confidence interval of the test set is broader than the respective one of the evaluation set. Similar behavior is presented for all days studied. In fact the limits of the test set are unknown in real applications, which means that they should be corrected, i.e. using the proper multiplying factor which is calculated by trial executions with historical data [66].

2.4.6 Effect of the Special Days

Load curves of special days differ significantly from the respective ones of the regular days. One approach to dealing with this problem is to ignore the irregularity of the special days and include their data into the training and test sets. However, ANN performance is generally expected to deteriorate in this case.

In a different approach, proposed in [44], the load curve of a special day is decomposed in two components; one representing the load of a normal day and

another representing the special day effect. More specifically, the load is expressed as follows:

$$\overline{L}_{holiday} = \overline{L}_{normal} - \Delta\overline{L}_{holiday} \tag{2.50}$$

Hence, the vectors of the input data set corresponding to special days are increased by the special day corrective term, $\Delta\overline{L}_{holiday}$, before they are used in training process. Finally, the vectors used for the training are of the form $\overline{L}_{normal} = N\left(\overline{X} + \Delta\overline{X}_{holiday}\right)$. Where, $\overline{X}$ is the input vector comprising load data containing, temperature measurements and $\Delta\overline{X}_{holiday}$ the correction term of the effect of the special days. The last term is calculated by reusing load data of respective special days of previous years.

Another way to deal with the effect of the special days is to properly group the input data with emphasis placed on the special days [70].

2.4.7 *The Effect of the Time Period Length Used for the Training of the ANN*

In order to study the effect of the duration of the training data time period seven different time periods were used. More specifically, these time periods extend on the last one, two, ..., and seven years before the reference year, respectively. The MAPE of the training, test and validation data sets obtained for the basic ANN configuration of paragraph 2.4.1 (trained with the scaled conjugate gradient algorithm) is shown in Fig. 2.11. It appears that in case of short term load forecasting in Greek power system using data from the last 3 years leads to the best results regarding MAPE minimization. The use of validation test set is necessary as it helps to achieve better generalization results and it seems to have similar behaviour with the test set rather than the training set.

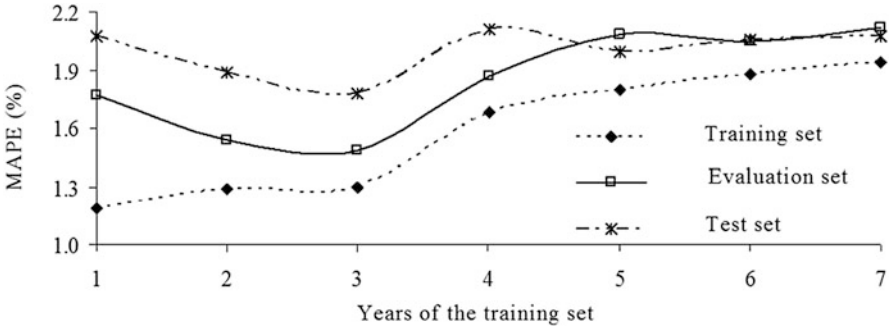


Fig. 2.11 MAPE of training, validation and test sets versus the training years. The scaled conjugate gradient training algorithm is use and the 10 % of the data is used for model evaluation

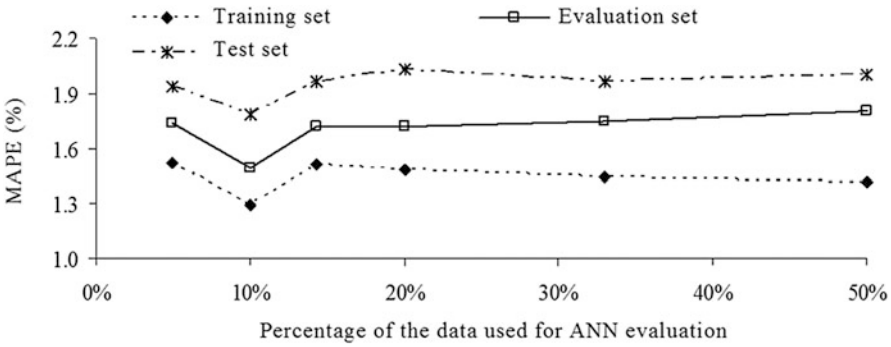


Fig. 2.12 MAPE of training, validation and test sets versus the percentage of the data used for ANN evaluation. Scaled conjugate gradient training algorithm is used

Another important issue that should be carefully studied is the percentage of the data used for the evaluation of the model. Due to the abundance of the available data it was decided that the training and evaluation data sets should not overlap. For the specific case study, it is proved that the minimum MAPE is obtained if the 10 % of the data is used for evaluation purposes. The respective results are shown in Fig. 2.12.

2.5 Conclusions

Short term load forecasting in electric power systems is a very complex problem. ANNs have been proved very efficient in short term load forecasting and they constitute nowadays the predominant method in this field. The basics of ANNs and the major training techniques are described at the beginning of this chapter. Next, several variations concerning the structure of the ANN model, the training method and its parameters have been examined and applied to the Greek power system for short term load forecasting. The results obtained reveal that general conclusions cannot be easily extracted as the performance of each of the examined forecasting models depends highly on the characteristics of each case study. However, it seems that in most cases the basic ANN configuration described in 2.4.1 seems to be the most appropriate for application to Greek power system. Data compression techniques reduce greatly computation time but generally deteriorate performance. Hence, they could be avoided if computation time is not a constraining factor.

Further work on this topic could be the combination of ANNs with other techniques like fuzzy logic, GA algorithms etc. Moreover, short term load forecasting in emerging types of power systems incorporating extensive smart load demand management will be a topic of great importance in the forthcoming years.

References

1. Hobbs, B.F., Jitprapaikulsarn, S., Maratukulam, D.J.: Analysis of the value for unit commitment of improved load forecasts. IEEE Trans. Power Syst. **14**(4), 1342–48 (1999)
2. Gooi, H.B., Mendes, D.P., Bell, K.R.W., Kirschen, D.S.: Optimal scheduling of spinning reserve. IEEE Trans. Power Syst. **14**(4), 1485–1492 (1999)
3. Gravener, M.H., Nwankpa, C.: Available transfer capability and first order sensitivity. IEEE Trans. Power Syst. **14**(2), 512–518 (1999)
4. Fan, J.Y., McDonald, J.D.: A real-time implementation of short-term load forecasting for distribution power systems. IEEE Trans. Power Syst. **9**(2), 988–994 (1994)
5. Haidda, T., Muto, S.: Regression based peak load forecasting using a transformation technique. IEEE Trans. Power Syst. **9**(4), 1788–1794 (1994)
6. Rupanagunta, P., Baughman, M.L., Jones, J.W.: Scheduling of cool storage using non-linear programming techniques. IEEE Trans. Power Syst. **10**(3), 1279–1285 (1995)
7. Kassaei, H.R., Keyhani, A., Woung, T., Rahman, M.: A hybrid fuzzy neural network bus load modeling and prediction. IEEE Trans. Power Syst. **14**(2), 718–724 (1999)
8. Yang, H.T., Huang, C.M., Huang, C.L.: Identification of ARMAX model for short term load forecasting: an evolutionary programming approach. IEEE Trans. Power Syst. **11**(1), 403–408 (1996)
9. Hippert, H.S., Pedreira, C.E., Souza, R.C.: Neural networks for short-term load forecasting: a review and evaluation. IEEE Trans. Power Syst. **16**(1), 44–55 (2001)
10. Mastorocostas, P.A., Theocharis, J.B., Bakirtzis, A.G.: Fuzzy modeling for short-term load forecasting using the orthogonal least squares method. IEEE Trans. Power Syst. **14**(1), 29–36 (1999)
11. Infield, D.G., Hill, D.C.: Optimal smoothing for trend removal in short term electricity demand forecasting. IEEE Trans. Power Syst. **13**(3), 1115–1120 (1998)
12. Charytoniuk, W., Chen, M.S., Van Olinda, P.: Nonparametric regression based short-term load forecasting. IEEE Trans. Power Syst. **13**(3), 725–730 (1998)
13. Khotanzad, A., Hwang, R.C., Abaye, A., Maratukulam, D.: An adaptive modular artificial neural network hourly load forecaster and its implementation at electric utilities. IEEE Trans. Power Syst. **10**(3), 1716–1722 (1995)
14. Khotanzad, A., Afkhami-Rohani, R., Lu, T.L., Abaye, A., Davis, M., Maratukulam, D.: ANNSTLF- neural network-based electric load forecasting system. IEEE Trans. Neural Netw. **8**(4), 835–846 (1996)
15. Khotanzad, A., Afkhami-Rohani, R., Maratukulam, D.: Artificial neural network short-term load forecaster-generation three. IEEE Trans. Power Syst. **13**(4), 1413–1422 (1998)
16. Papalexopoulos, A.D., Hao, S., Peng, T.M.: An implementation of a neural network based load fore-casting model for the EMS. IEEE Trans. Power Syst. **9**(4), 1956–62 (1994)
17. Mohammed, O., Park, D., Merchant, R., Dinh, T., Tong, C., Azeem, A.: Practical experiences with an adaptive neural network short-term load forecasting system. IEEE Trans. Power Syst. **10**(1), 254–265 (1995)
18. Wang, X., Shi, M.: Matlab for forecasting of electric power load based on BP neural network. Proceedings ICICIS 2011, Part I, Series: Communications in Computer and Information Science **134**: 636-642 (2011)
19. Barghinia, S., Ansarimehr, P., Habibi, H., Vafadar, N.: Short-term load forecasting of Iran national power system using artificial neutral network, p. 5. IEEE Porto Power Tech Conference, Porto, Portugal (2001)
20. Kalaitzakis, K., Stavrakakis, G.S., Anagnostakis, E.M.: Short-term load forecasting based on artificial neural networks parallel implementation. Electr. Power Syst. Res. **63**, 185–196 (2002)
21. Saini, L.M., Soni, M.K.: Artificial neural network-based peak load forecasting using conjugate gradient methods. IEEE Trans. Power Syst. **17**(3), 907–912 (2002)
22. Gontar, Z., Hatziargyriou, N.: Short term load forecasting with radial basis function network. 2001 IEEE Porto Power Tech Conference 10th-13th September, p. 4. Porto, Portugal (2001)

23. Yun, Z., Quan, Z., Caixin, S., Shaolan, L., Yuming, L., Yang, S.: RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment. IEEE Trans. Power Syst. **23**(3), 853–858 (2008)
24. Ko, C.N., Lee, C.M.: Short-term load forecasting using SVR (support vector regression)-based radial basis function neural network with dual extended Kalman filter. Energy **49**, 413–422 (2013)
25. Vermaak, J., Botha, E.C.: Recurrent neural networks for short-term load forecasting. IEEE Trans. Power Syst. **13**(1), 126–132 (1998)
26. Chen, S.T., Yu, D.C., Moghaddamjo, A.R.: Weather sensitive short-term load forecasting using nonfully connected artificial neural network. IEEE Trans. Power Syst. **7**(3), 1098–1105 (1992)
27. Abdel-Aal, R.E.: Short-term hourly load forecasting using abductive networks. IEEE Trans. Power Syst. **19**(1), 164–173 (2004)
28. Ranaweera, D.K., Karady, G.G., Farmer, R.G.: Effect of probabilistic inputs on neural network-based electric load forecasting. IEEE Trans. Neural Netw. **7**(6), 1528–1532 (1996)
29. Taylor, J.W., Buizza, R.: Neural network load forecasting with weather ensemble predictions. IEEE Trans. Power Syst. **17**(3), 626–632 (2002)
30. Senjyu, T., Takara, H., Uezato, K., Funabashi, T.: One-hour-ahead load forecasting using neural network. IEEE Trans. Power Syst. **17**(1), 113–118 (2002)
31. Kim, K.H., Park, J.K., Hwang, K.J., Kim, S.H.: Implementation of hybrid short-term load forecasting system using artificial neural networks and fuzzy expert systems. IEEE Trans. Power Syst. **10**(3), 1534–1539 (1995)
32. Kim, K.H., Youn, H.S., Kang, Y.C.: Short-term load forecasting for special days anomalous load conditions using neural networks and fuzzy inference method. IEEE Trans. Power Syst. **15**(2), 559–565 (2000)
33. Srinivasan, D., Tan, S.S., Chang, C.S., Chan, E.K.: Parallel neural network-Fuzzy expert system strategy for short-term load forecasting: system implementation and performance evaluation. IEEE Trans. Power Syst. **14**(3), 1100–1106 (1999)
34. Daneshdoost, M., Lotfalian, M., Bumroonggit, G., Ngoy, J.P.: Neural network with fuzzy set-based classification for short-term load forecasting. IEEE Trans. Power Syst. **13**(4), 1386–1391 (1998)
35. Banik, S., Anwer, M., Khodadad Khan, A.F.M.: Predictive Power of the Daily Bangladeshi Exchange Rate Series based on Markov Model, Neuro Fuzzy Model and Conditional Heteroskedastic Model. Proceedings of 12th International Conference on Computer and Information Technology (ICCIT 2009), Dhaka, Bangladesh (2009)
36. Tanabe, T., Ueda, Y., Suzuki, S., Ito, T., Sasaki, N., Tanaka, T., Funabashi, T., Yokoyama, R.: Optimized operation and stabilization of microgrids with multiple energy resources. 7th International conference on Power Electronics, EXCO, Daegu, Korea (2007)
37. Jaipradidtham, C.: Next day load demand forecasting of future in electrical power generation on distribution networks using Adaptive Neuro-Fuzzy Inference. 1st International Power and Energy Conference PECon, Putrajaya, Malaysia (2006)
38. Malkocevic, D., Konjic, T., Miranda, V.: Preliminary comparison of different neural-fuzzy mappers for load curve short term prediction. 8th Seminar on Neural Network Applications in Electrical Engineering, NEUREL-2006, Faculty of Electrical Engineering, University of Belgrade, Serbia (2006)
39. Wang, X., Hatziargyriou, N., Tsoukalas, L.H.: A new methodology for nodal load forecasting in deregulated power systems. IEEE Power Eng. Rev. **12**, 48–51 (2002)
40. Chen, Y., Luh, P.B., Guan, C., Zhao, Y., Michel, L.D., Coolbeth, M.A., Friedland, P.B., Rourke, S.J.: Short-term load forecasting: similar day-based wavelet neural networks. IEEE Trans. Power Syst. **25**(1), 322–330 (2010)
41. Ferreira, V.H., Alves da Silva, A.P.: Toward estimating autonomous neural network-based electric load forecasters. IEEE Trans. Power Syst. **22**(4), 1554–1562 (2007)
42. Cardenas, J.J., Romeral, L., Garcia, A., Andrade, F.: Load forecasting framework of electricity consumptions for an intelligent energy management system in the user-side. Expert Syst. Appl. **39**, 5557–5565 (2012)

43. Amjady, N., Keynia, F.: A new neural network approach to short term load forecasting of electrical power systems. Energies **4**, 488–503 (2011)
44. Bakirtzis, A.G., Petridis, V., Kiartzis, S.J., Alexiadis, M.C., Maissis, A.H.: A neural network short term load forecasting model for the Greek power system. IEEE Trans. Power Syst. **11**(2), 858–863 (1996)
45. Kiartzis, S.J., Zournas, C.E., Theocharis, J.M., Bakirtzis, A.G., Petridis, V.: Short term load forecasting in an autonomous power system using artificial neural networks. IEEE Trans. Power Syst. **12**(4), 1591–1596 (1997)
46. Tsekouras, G.J., Kanellos, F.D., Kontargyri, V.T., Tsirekis, C.D., Karanasiou, I.S., Elias, C.N., Salis, A.D., Mastorakis, N.E.: A comparison of artificial neural networks algorithms for short term load forecasting in Greek intercontinental power system. WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal Processing, Puerto De La Cruz, Canary Islands, Spain (2008)
47. Tsekouras, G.J., Kanellos, F.D., Kontargyri, V.T., Tsirekis, C.D., Karanasiou, I.S., Elias, C.N., Salis, A.D., Kontaxis, P.A., Mastorakis, N.E.: Short term load forecasting in Greek Intercontinental Power System using ANNs: a Study for Input Variables. 10th WSEAS International Conference on Neural NetworksPrague, Czech Republic (2009)
48. Tsekouras, G.J., Kanellos, F.D., Elias, C.N., Kontargyri, V.T., Tsirekis, C.D., Karanasiou, I.S., Salis, A.D., Kontaxis, P.A., Gialketsi, A.A., Mastorakis, N.E.: Short term load forecasting in Greek interconnected power system using ANN: A study for output variables. 15th WSEAS International Conference on Systems Corfu, Greece (2011)
49. Choueiki, M.H., Mount-Campbell, C.A., Ahalt, S.C.: Implementing a weighted least squares procedure in training a neural network to solve the short-term load forecasting problem. IEEE Trans. Power Syst. **12**, 41689–1694 (1997)
50. Metaxiotis, K., Kagiannas, A., Askounis, D., Psarras, J.: Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. Energy Convers. Manag. **44**, 1525–1534 (2003)
51. Kyriakides, E., Polycarpou, M.: Short term electric load forecasting: a tutorial (Chapter 16). In: Chen, K., Wang, L. (eds.), Trends in neural computation, studies in computational intelligence, 35 pp. 391–418. Springer (2007)
52. Hahn, H., Meyer-Nieberg, S., Pickl, S.: Electric load forecasting methods: tools for decision making. Eur. J. Oper. Res. **199**, 902–907 (2009)
53. Felice, M., Yao, X.: Short-term load forecasting with neural network ensembles: a comparative study. IEEE Comput. Intell. Mag. **6**(3), 47–56 (2011)
54. Rego, L.P., Santana, A.L., Conde, G., Silva, M.S., Francês, C.R.L., Rocha, C.A.: Comparative analyses of computational intelligence models for load forecasting: a case study in the Brazilian Amazon power suppliers. Lecture Notes Comput. Sci. **5553**, 1044–1053 (2009)
55. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall (1994)
56. Likas, A.: Computational Intelligence (in Greek), 1st edn. Ioannina, Greece (1999)
57. Ghosh, P.S., Chakravorti, S., Chatterjee, N.: Estimation of time-to-flashover characteristics of contaminated electrolytic surfaces using a neural network. IEEE Trans Dielectr. Electr. Insul. **2**(6), 1064–1074 (1995)
58. Fletcher, R., Reeves, C.M.: Function minimization by conjugate gradients. Comput. J. **7**, 149–154 (1964)
59. Polak, E.: Computational methods in optimization: a unified approach, 1st edn. Academic Publication, New York (1971)
60. Powell, M.J.: Restart procedures for the conjugate gradient method. Math. Program. **12**, 241–254 (1977)
61. Moller, M.F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. **6**, 525–533 (1993)
62. Levenberg, K.: A method for the solution of certain problems in least squares. Quart. Appl. Math. **2**, 164–168 (1944)
63. Marquardt, D.: An algorithm for least squares estimation of nonlinear parameters. SIAM J. Appl. Math. **11**, 431–441 (1963)

64. Silva, A.P.A., Moulin, L.S.: Confidence intervals for neural network based short-term load forecasting. IEEE Trans. Power Syst. **15**(4), 1191–1196 (2000)
65. Mastorakis, N.E., Tsekouras, G.J.: Short term load forecasting in Greek intercontinental power system using ANN: the confidence interval. Advanced aspects of theoretical electrical engineering. Sozopol, Bulgaria (2010)
66. Tsekouras, G.J., Mastorakis, N.E., Kanellos, F.D., Kontargyri, V.T., Tsirekis, C.D., Karanasiou, I.S., Elias, C.N., Salis, A.D., Kontaxis, P.A., Gialketsi, A.A.: Short term load forecasting in Greek interconnected power system using ANN: Confidence Interval using a novel re-sampling technique with corrective Factor. WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal Processing, Vouliagmeni, Athens, Greece (2010)
67. Khosravi, A., Nahavandi, S., Creighton, D.: Construction of optimal prediction intervals for load forecasting problems. IEEE Trans. Power Syst. **25**(3), 1496–1503 (2010)
68. Quan, H., Srinivasan, D., Khosravi, A.: Uncertainty handling using neural network-based prediction intervals for electrical load forecasting. Energy **73**, 916–925 (2014)
69. Hand, D., Manilla, H., Smyth, P.: Principles of data mining, 1st edn. The MIT Press, Cambridge (2001)
70. Fidalgo, J.N., Peças Lopes, J.A.: Load forecasting performance enhancement when facing anomalous events. IEEE Trans. Power Syst. **20**(1), 408–415 (2005)

Chapter 3
Analysis of Non-linear Vibrations of a Fractionally Damped Cylindrical Shell Under the Conditions of Combinational Internal Resonance

Yury Rossikhin and Marina Shitikova

Abstract Non-linear damped vibrations of a cylindrical shell subjected to the different conditions of the combinational internal resonance are investigated. Its viscous properties are described by Riemann-Liouville fractional derivative. The displacement functions are determined in terms of eigenfunctions of linear vibrations. The procedure resulting in decoupling linear parts of equations is proposed with the further utilization of the method of multiple time scales for solving nonlinear governing equations of motion, in so doing the amplitude functions are expanded into power series in terms of the small parameter and depend on different time scales. It is shown that the phenomenon of internal resonance can be very critical, since in a circular cylindrical shell the internal additive and difference combinational resonances are always present. The influence of viscosity on the energy exchange mechanism is analyzed. It is shown that each mode is characterized by its damping coefficient connected with the natural frequency by the exponential relationship with a negative fractional exponent.

Keywords Cylindrical shell • Free nonlinear damped vibrations • Combinational internal resonance • Fractional derivative • Method of multiple time scales

3.1 Introduction

Beginning with the paper by Witt and Gorelik [1], whose authors were among the first to show theoretically and experimentally the two-to-one internal resonance with the energy exchange from one subsystem to another using the simplest

Yu. Rossikhin • M. Shitikova (✉)
Research Center for Wave Dynamics of Solids and Structures
Voronezh State University of Architecture and Civil Engineering
20-letija Oktjabrja Str. 84, Voronezh 394006, Russia
e-mail: shitikova@vmail.ru; MVS@vgasu.vrn.ru

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_3

two-degree-of-freedom mechanical system as an example, interest of researchers to the problems of the internal resonance in mechanical systems does not relax. It will suffice to mention the state-of-the-art article [2] and the monograph [3], wherein the extensive review of literature in the field of internal resonances in different mechanical systems is presented. Different types of the internal resonance: one-to-one, two-to-one, as well as a variety of combinational resonances, when three and more natural modes interact, have been discussed. The enumerated internal resonances were investigated in mechanical systems with more than one degree-of-freedom, as well as in strings, beams, plates, and shells.

It has been emphasized by many researchers [4, 5] that the phenomenon of internal resonances can be very critical especially for circular cylindrical shells. Thus, the nonlinear vibrations of infinitely long circular cylindrical shells under the conditions of the two-to-one internal resonance were studied in [6] via the method of multiple time scales using the simple plane strain theory of shells. Parametrically excited vibrations of infinitely long cylindrical shells and nonlinear forced vibrations of a simply supported, circular cylindrical shell filled with an incompressible, inviscid, quiescent and dense fluid were investigated, respectively, in [4, 5, 7] using Donnell's nonlinear shallow-shell theory. The flexural deformation is usually expanded by using the linear shell eigenmodes, in so doing the flexural response involves several nodal diameters and one or two longitudinal half-waves. Internal resonances of different types have been analyzed in [8, 9].

In spite of the fact that many studies have been carried out on large amplitude vibrations of circular cylindrical shells and many different approaches to the problem have been used, we agree with Amabili et al. [5] that this research area is still far from being well understood. The extensive review of studies on shallow shells nonlinear vibrations could be found in the state-of-the-art articles by Kubenko and Koval'chuk [10], Amabili and Païdoussis [11], and Lee [12], as well as in recent papers [13, 14].

The problem of free, as well as forced nonlinear vibrations of cylindrical shells can be considered from different positions depending on the shell geometry, in so doing the nonlinear displacement field is approximated by a finite sum of global interpolating functions. However the choice of appropriate modal expansions is fundamental to guaranteeing accuracy of the results for large-amplitude vibrations [15]. Thus, for example, different expansions involving from 14 to 48 generalized coordinates, associated to natural modes of simply supported shells, have been discussed in [16].

A comparison of five shell theories for large-amplitude vibrations of circular cylindrical shells which are generally applied to geometrically non-linear problems that use only three variables, which are the three middle surface displacements, has been carried out in [16]. More complicated shell theories, suitable for moderately thick laminated shells exist, and they use five independent variables, three displacements and two rotations, [17, 18] or even six variables if thickness variation is taken into account [19].

In recent years much attention is given to damping features of mechanical systems subjected to the conditions of different internal resonances. Damping properties of non-linear systems are described mainly by the first-order time-derivative

of a generalized displacement [3]. However, as it has been shown by Rossikhin and Shitikova [20], who analyzed free damped vibrations of suspension combined system under the conditions of the one-to-one internal resonance, for good fit of the theoretical investigations with the experimental results it is better to describe the damping features of non-linear mechanical systems in terms of fractional time-derivatives of the generalized displacements [21]. The analysis of non-linear vibrations of a two-degree-of-freedom mechanical system, the damping features of which are described by a fractional derivative, has shown [22] that in the case when the system is under the conditions of the two-to-one or one-to-one internal resonance, viscosity may have a twofold effect on the system: a destabilizing influence producing unsteady energy exchange, and a stabilizing influence resulting in damping of the energy exchange mechanism. The same phenomenon was noted when considering non-linear vibrations of a fractionally damped plate under the conditions of the two-to-one [23] or one-to-one [24] internal resonance, as well as during non-linear vibrations of a fractionally damped cylindrical shell under the conditions of the internal resonance of the order of ε [25], where ε is a small parameter.

In the present paper, non-linear free damped vibrations of a thin cylindrical viscoelastic shell, the damping properties of which are described by the Riemann-Liouville fractional derivatives, are investigated. The dynamic behaviour of the shell is described by a set of three coupled non-linear differential equations with due account for the fact that the shell is being under the conditions of the internal combinational resonance resulting in the interaction of three modes corresponding to the mutually orthogonal displacements. The displacement functions are determined in terms of eigenfunctions of linear vibrations. The procedure resulting in decoupling linear parts of equations is proposed with the further utilization of the method of multiple scales for solving nonlinear governing equations of motion, in so doing the amplitude functions are expanded into power series in terms of the small parameter and depend on different time scales. It is shown that the phenomenon of combinational internal resonance, resulting in coupling of three modes of vibrations, can be very critical, since in a circular cylindrical shell internal additive and difference combinational resonances are always present.

3.2 Problem Formulation and Governing Equations

Let us consider the dynamic behaviour of a free supported non-linear elastic circular cylindrical shell of radius R and length l, vibrations of which in the cylindrical system of coordinates is described by the Donnell–Mushtari–Vlasov equations with respect to the three displacements [26]:

$$\frac{\partial^2 u}{\partial x^2} + \frac{1-\nu}{2}\frac{1}{R^2}\frac{\partial^2 u}{\partial \varphi^2} + \frac{1+\nu}{2}\frac{1}{R}\frac{\partial^2 v}{\partial x \partial \varphi} - \nu\frac{1}{R}\frac{\partial w}{\partial x} + \frac{\partial w}{\partial x}\frac{\partial^2 w}{\partial x^2}$$

$$+\frac{1+\nu}{2}\,\frac{1}{R^2}\,\frac{\partial w}{\partial\varphi}\,\frac{\partial^2 w}{\partial x\partial\varphi}+\frac{1-\nu}{2}\,\frac{1}{R^2}\,\frac{\partial w}{\partial x}\,\frac{\partial^2 w}{\partial\varphi^2}=\frac{\rho\left(1-\nu^2\right)}{E}\,\frac{\partial^2 u}{\partial t^2},\tag{3.1}$$

$$\frac{1}{R^2}\,\frac{\partial^2 v}{\partial\varphi^2}+\frac{1-\nu}{2}\,\frac{\partial^2 v}{\partial x^2}+\frac{1+\nu}{2}\,\frac{1}{R}\,\frac{\partial^2 u}{\partial x\partial\varphi}-\frac{1}{R^2}\,\frac{\partial w}{\partial\varphi}+\frac{1}{R^3}\,\frac{\partial w}{\partial\varphi}\,\frac{\partial^2 w}{\partial\varphi^2}$$
$$+\frac{1+\nu}{2}\,\frac{1}{R}\,\frac{\partial w}{\partial x}\,\frac{\partial^2 w}{\partial x\partial\varphi}+\frac{1-\nu}{2}\,\frac{1}{R}\,\frac{\partial w}{\partial\varphi}\,\frac{\partial^2 w}{\partial x^2}=\frac{\rho\left(1-\nu^2\right)}{E}\,\frac{\partial^2 v}{\partial t^2},\tag{3.2}$$

$$\frac{h^2}{12}\nabla^4 w+\frac{1}{R^2}\,w-\nu\,\frac{1}{R}\,\frac{\partial u}{\partial x}-\frac{1}{R^2}\,\frac{\partial v}{\partial\varphi}-\frac{1}{2}\,\frac{\nu}{R}\left(\frac{\partial w}{\partial x}\right)^2-\frac{1}{2}\,\frac{1}{R^3}\left(\frac{\partial w}{\partial\varphi}\right)^2$$
$$-\frac{\partial}{\partial x}\left[\frac{\partial w}{\partial x}\left(\frac{\partial u}{\partial x}+\frac{\nu}{R}\,\frac{\partial v}{\partial\varphi}-\frac{\nu}{R}\,w\right)+\frac{1-\nu}{2}\,\frac{1}{R}\,\frac{\partial w}{\partial\varphi}\left(\frac{1}{R}\,\frac{\partial u}{\partial\varphi}+\frac{\partial v}{\partial x}\right)\right]$$
$$-\frac{1}{R}\,\frac{\partial}{\partial\varphi}\left[\frac{1}{R}\,\frac{\partial w}{\partial\varphi}\left(\nu\,\frac{\partial u}{\partial x}+\frac{1}{R}\,\frac{\partial v}{\partial\varphi}-\frac{1}{R}\,w\right)+\frac{1-\nu}{2}\,\frac{\partial w}{\partial x}\left(\frac{1}{R}\,\frac{\partial u}{\partial\varphi}+\frac{\partial v}{\partial x}\right)\right]$$
$$=-\frac{\rho\left(1-\nu^2\right)}{E}\,\frac{\partial^2 w}{\partial t^2},\tag{3.3}$$

where the x-axis is directed along the axis of the cylinder, φ is the polar angle in the plane perpendicular to the x-axis, $u=u(x,\varphi,t)$, $v=v(x,\varphi,t)$, and $w=w(x,\varphi,t)$ are the displacements of points located in the shell's middle surface in three mutually orthogonal directions x,φ,r, r is the polar radius, h is the thickness, ρ is the density, E and ν are the elastic modulus and Poisson's ratio, respectively, t is the time, and

$$\nabla^4=\nabla^2\nabla^2=\frac{\partial^4}{\partial x^4}+2\,\frac{1}{R^2}\,\frac{\partial^4}{\partial x^2\partial\varphi^2}+\frac{1}{R^4}\,\frac{\partial^4}{\partial\varphi^4}\,.$$

The initial conditions

$$u|_{t=0}=v|_{t=0}=w|_{t=0}=0,\tag{3.4}$$

$$\dot{u}\Big|_{t=0}=\varepsilon V_1^0(x,\varphi),\quad \dot{v}\Big|_{t=0}=\varepsilon V_2^0(x,\varphi),\quad \dot{w}\Big|_{t=0}=\varepsilon V_3^0(x,\varphi),\tag{3.5}$$

where $V_\alpha^0(x,\varphi)$ are corresponding initial velocities and ε is a small value, should be added to Eqs. (3.1)–(3.3). Hereafter overdots denote time-derivatives.

The boundary conditions for the free supported shell (Navier conditions for the edges free supported in the x-direction) have the form [26]:

$$w\Big|_{x=0} = w\Big|_{x=l} = 0, \quad v\Big|_{x=0} = v\Big|_{x=l} = 0,$$
$$\frac{\partial^2 w}{\partial x^2}\Big|_{x=0} = \frac{\partial^2 w}{\partial x^2}\Big|_{x=l} = 0, \quad \frac{\partial u}{\partial x}\Big|_{x=0} = \frac{\partial u}{\partial x}\Big|_{x=l} = 0\,. \tag{3.6}$$

Let us rewrite Eqs. (3.1)–(3.3), the initial conditions (3.4) and (3.5), and the boundary conditions (3.6) in a dimensionless form introducing the following dimensionless values:

$$u^* = \frac{u}{l}, \quad v^* = \frac{v}{l}, \quad w^* = \frac{w}{l}, \quad x^* = \frac{x}{l}, \quad t^* = \frac{t}{l}\sqrt{\frac{E}{\rho\,(1-\nu^2)}}\,. \tag{3.7}$$

Omitting asterisks near the dimensionless values, rewriting Eqs. (3.1)–(3.6) in the dimensionless form with due account for (3.7), and considering that the shell vibrates in a viscoelastic medium yields [25]

$$u_{xx} + \frac{1-\nu}{2}\,\beta_1^2 u_{\varphi\varphi} + \frac{1+\nu}{2}\,\beta_1 v_{x\varphi} - \nu\beta_1 w_x + w_x\left(w_{xx} + \frac{1-\nu}{2}\,\beta_1^2 w_{\varphi\varphi}\right) + \frac{1+\nu}{2}\,\beta_1^2 w_\varphi w_{x\varphi} = \ddot{u} + æ_1\left(\frac{d}{dt}\right)^\gamma u\,, \tag{3.8}$$

$$\beta_1^2 v_{\varphi\varphi} + \frac{1-\nu}{2}\,v_{xx} + \frac{1+\nu}{2}\,\beta_1 u_{x\varphi} - \beta_1^2 w_\varphi + \beta_1 w_\varphi\left(\beta_1^2 w_{\varphi\varphi} + \frac{1-\nu}{2}\,w_{xx}\right) + \frac{1+\nu}{2}\,\beta_1 w_x w_{x\varphi} = \ddot{v} + æ_2\left(\frac{d}{dt}\right)^\gamma v, \tag{3.9}$$

$$\frac{\beta_2^2}{12}\left(w_{xxxx} + 2\beta_1^2 w_{xx\varphi\varphi} + \beta_1^4 w_{\varphi\varphi\varphi\varphi}\right) - \nu\beta_1 u_x - \beta_1^2 v_\varphi + \beta_1^2 w + \frac{1}{2}\,\nu\beta_1 (w_x)^2 + \frac{1}{2}\,\beta_1^3 (w_\varphi)^2 - w_x\left(u_{xx} + \frac{1-\nu}{2}\,\beta_1^2 u_{\varphi\varphi} + \frac{1+\nu}{2}\,\beta_1 v_{x\varphi}\right) - \beta_1 w_\varphi\left(\beta_1^2 v_{\varphi\varphi} + \frac{1-\nu}{2}\,v_{xx} + \frac{1+\nu}{2}\,\beta_1 u_{x\varphi}\right) - w_{xx}(u_x + \nu\beta_1 v_\varphi - \nu\beta_1 w) - \beta_1^2 w_{\varphi\varphi}(\nu u_x + \beta_1 v_\varphi - \beta_1 w) - (1-\nu)\beta_1 w_{x\varphi}(\beta_1 u_\varphi + v_x) = -\ddot{w} - æ_3\left(\frac{d}{dt}\right)^\gamma w. \tag{3.10}$$

Equations (3.8)–(3.10) are subjected to the initial

$$u|_{t=0} = v|_{t=0} = w|_{t=0} = 0, \tag{3.11}$$

$$\dot{u}\Big|_{t=0} = \varepsilon v_1^0(x,\varphi), \quad \dot{v}\Big|_{t=0} = \varepsilon v_2^0(x,\varphi), \quad \dot{w}\Big|_{t=0} = \varepsilon v_3^0(x,\varphi), \tag{3.12}$$

and boundary conditions

$$w|_{x=0} = w\Big|_{x=1} = 0, \quad v|_{x=0} = v\Big|_{x=1} = 0,$$

$$u_x\Big|_{x=0} = u_x\Big|_{x=1} = 0, \quad w_{xx}\Big|_{x=0} = w_{xx}\Big|_{x=1} = 0, \tag{3.13}$$

where $\beta_1 = l/R$ and $\beta_2 = h/l$ are the parameters defining the dimensions of the shell, lower indices label the derivatives with respect to the corresponding coordinates, $v_i^0(x,\varphi) = V_i^0(x,\varphi)\sqrt{\frac{\rho(1-\nu^2)}{E}}$ are dimensionless initial velocities, æ$_i$ $(i = 1,2,3)$ are damping coefficients.

It was shown in Samko et al. [27] (see Chap. 2, Paragraph 5, point 7^0) that the fractional order of the operator of differentiation $\left(\frac{d}{dt}\right)^\gamma$ is equal to Marsho fractional derivative, which, in its turn, equal to Riemann-Liouville derivative D_+^γ

$$\left(\frac{d}{dt}\right)^\gamma f = \frac{1}{\Gamma(-\gamma)}\int_0^\infty \frac{f(t-t') - f(t)}{t'^{1+\gamma}}\,dt' = D_+^\gamma f, \tag{3.14}$$

or with due account for equality $\gamma\Gamma(\gamma) = \Gamma(1+\gamma)$

$$\left(\frac{d}{dt}\right)^\gamma f = \frac{\gamma}{\Gamma(1-\gamma)}\int_0^\infty \frac{f(t) - f(t-t')}{t'^{1+\gamma}}\,dt' = D_+^\gamma f, \tag{3.15}$$

where

$$D_+^\gamma f = \frac{1}{\Gamma(1-\gamma)}\frac{d}{dt}\int_{-\infty}^t \frac{f(t')dt'}{(t-t')^\gamma} = \frac{1}{\Gamma(1-\gamma)}\frac{d}{dt}\int_0^\infty \frac{f(t-t')dt'}{t'^\gamma}. \tag{3.16}$$

Distinct to the traditional modeling the viscous resistance forces via first order time-derivatives [28], in the present research we adopt the fractional order time-derivatives D_+^γ, what, as it will be shown below, will allow us to obtain the damping coefficients dependent on the natural frequency of vibrations. It has been demonstrated in [20, 29] that such an approach for modeling the damped nonlinear vibrations of thin bodies provides the good agreement between the theoretical results and the experimental data through the appropriate choice of the fractional parameter (the order of the fractional derivative) and the viscosity coefficient.

It has been noted in [20, 23] that a fractional derivative is the immediate extension of an ordinary derivative. In fact, when $\delta \to 1$, $D_+^\gamma x$ tends to $\dot{x}$, i.e., at $\gamma \to 1$ the fractional derivative goes over into the ordinary derivative, and the mathematical model of the viscoelastic shell under consideration transforms into the Kelvin-Voigt model, wherein the elastic element behaves nonlinearly, but the viscous element behaves linearly. When $\gamma \to 0$, the fractional derivative $D_+^\gamma x$ tends to $x(t)$. To put it otherwise, the introduction of the new fractional parameter along with the parameters $æ_i$ allows one to change not only the magnitude of viscosity at the cost of an increase or decrease in the parameters $æ_i$, but also the character of viscosity at the sacrifice of variations in the fractional parameter.

Reference to (3.5) shows that free vibrations are excited by weak disturbance from the equilibrium position.

Since the set of Eqs. (3.8)–(3.10) admits the solution of the Navier type, then the displacements could be represented in the form

$$u(x,\varphi,t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} x_{1\,mn}(t)\eta_{1\,mn}(x,\varphi), \tag{3.17}$$

$$v(x,\varphi,t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} x_{2\,mn}(t)\eta_{2\,mn}(x,\varphi), \tag{3.18}$$

$$w(x,\varphi,t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} x_{3\,mn}(t)\eta_{3\,mn}(x,\varphi), \tag{3.19}$$

where m and n are integers, $x_{i\,mn}(t)$ are the generalized displacements, and $\eta_{i\,mn}(x,\varphi)$ $(i = 1,2,3)$ are the eigenfunctions satisfying the boundary conditions (3.13) at $0 \le x \le 1$ and $0 \le \varphi \le 2\pi$:

$$\eta_{i\,mn}(x,\varphi) = \begin{cases} \cos\pi mx \sin n\varphi \\ \sin\pi mx \cos n\varphi. \\ \sin\pi mx \sin n\varphi \end{cases} \tag{3.20}$$

Substituting (3.17)–(3.20) into Eqs. (3.8)–(3.10), multiplying (3.8), (3.9), and (3.10) by η_{1lk}, η_{2lk}, and η_{3lk}, respectively, integrating over x and φ, and using the orthogonality conditions for linear modes within the domains of $0 \le x \le 1$ and $0 \le \varphi \le 2\pi$, we are led to a coupled set of nonlinear ordinary differential equations of the second order in $x_{i\,mn}$ $(i = 1,2,3)$:

$$\ddot{x}_{i\,mn} + æ_i D^\gamma x_{i\,mn} + S_{ij}^{mn} x_{j\,mn} = -F_{i\,mn}, \tag{3.21}$$

where the summation is carried out over two repeated indices, the elements of the matrix S_{ij}^{mn} $(i,j = 1,2,3)$ are defined as follows:

$$S_{11}^{mn} = \left(\pi^2 m^2 + \frac{1-\nu}{2}\,\beta_1^2 n^2\right), \quad S_{12}^{mn} = S_{21}^{mn} = \frac{1+\nu}{2}\,\beta_1 \pi m n,$$

$$S_{13}^{mn} = S_{31}^{mn} = \nu \beta_1 \pi m, \qquad S_{23}^{mn} = S_{32}^{mn} = \beta_1^2 n\,, \tag{3.22}$$

$$S_{22}^{mn} = \left(\frac{1-\nu}{2}\,\pi^2 m^2 + \beta_1^2 n^2\right), \quad S_{33}^{mn} = \frac{\beta_2^2}{12}\,\left(\pi^2 m^2 + \beta_1^2 n^2\right)^2 + \beta_1^2\,.$$

The nonlinear parts $F_{i\ mn}$ of Eq. (3.21) have the form

$$F_{1mn} = 2 \sum_{m_1} \sum_{n_1} \sum_{m_2} \sum_{n_2} x_{3m_1n_1} x_{3m_2n_2} A_{mn}^{m_1n_1m_2n_2}, \tag{3.23}$$

$$F_{2mn} = \frac{2}{\pi} \sum_{m_1} \sum_{n_1} \sum_{m_2} \sum_{n_2} x_{3m_1n_1} x_{3m_2n_2} B_{mn}^{m_1n_1m_2n_2}, \tag{3.24}$$

$$\begin{aligned} F_{3mn} = -\frac{2}{\pi} \sum_{m_1} \sum_{n_1} \sum_{m_2} \sum_{n_2} \Big[& x_{3m_1n_1} x_{1m_2n_2} C_{mn}^{m_1n_1m_2n_2} \\ & + x_{3m_1n_1} x_{2m_2n_2} D_{mn}^{m_1n_1m_2n_2} + x_{3m_1n_1} x_{3m_2n_2} E_{mn}^{m_1n_1m_2n_2} \Big], \end{aligned} \tag{3.25}$$

where

$$\begin{aligned} A_{mn}^{m_1n_1m_2n_2} &= m_1 \left(\pi^2 m_2^2 + \frac{1-\nu}{2}\,\beta_1^2 n_2^2\right) a_{1mn}^{m_1n_1m_2n_2} \\ &\quad - \frac{1+\nu}{2}\,\beta_1^2 n_1 m_2 n_2 a_{2mn}^{m_1n_1m_2n_2}, \\ B_{mn}^{m_1n_1m_2n_2} &= \beta_1 n_1 \left(\beta_1^2 n_2^2 + \frac{1-\nu}{2}\,\pi^2 m_2^2\right) a_{3mn}^{m_1n_1m_2n_2} \\ &\quad - \frac{1+\nu}{2}\,\beta_1 \pi^2 m_1 m_2 n_2 a_{4mn}^{m_1n_1m_2n_2}, \\ C_{mn}^{m_1n_1m_2n_2} &= \pi m_2 \left(\pi^2 m_1^2 + \nu \beta_1^2 n_1^2\right) a_{5mn}^{m_1n_1m_2n_2} + (1-\nu)\beta_1^2 \pi m_1 n_1 n_2 a_{6mn}^{m_1n_1m_2n_2} \\ &\quad - \pi m_1 \left(\pi^2 m_2^2 + \frac{1-\nu}{2}\,\beta_1^2 n_2^2\right) a_{7mn}^{m_1n_1m_2n_2} \\ &\quad - \frac{1+\nu}{2}\,\beta_1^2 \pi n_1 m_2 n_2 a_{8mn}^{m_1n_1m_2n_2}, \\ D_{mn}^{m_1n_1m_2n_2} &= \beta_1 n_2 \left(\nu \pi^2 m_1^2 + \beta_1^2 n_1^2\right) a_{5mn}^{m_1n_1m_2n_2} + (1-\nu)\beta_1 \pi^2 m_1 n_1 m_2 a_{6mn}^{m_1n_1m_2n_2} \end{aligned}$$

$$
\begin{aligned}
&- \frac{1+\nu}{2}\,\beta_1\pi^2 m_1 m_2 n_2 a_{7mn}^{m_1 n_1 m_2 n_2} \\
&- \beta_1 n_1 \left(\frac{1-\nu}{2}\,\pi^2 m_2^2 + \beta_1^2 n_2^2\right) a_{8mn}^{m_1 n_1 m_2 n_2}, \\
E_{mn}^{m_1 n_1 m_2 n_2} &= \beta_1 \left(\nu\pi^2 m_2^2 + \beta_1^2 n_2^2\right) a_{5mn}^{m_1 n_1 m_2 n_2} \\
&- \frac{\beta_1}{2}\left(\nu\pi^2 m_1 m_2 a_{9mn}^{m_1 n_1 m_2 n_2} + \beta_1^2 n_1 n_2 a_{10mn}^{m_1 n_1 m_2 n_2}\right),
\end{aligned}
$$

and the coefficients $a_{k\,mn}^{m_1 n_1 m_2 n_2}$ $(k = 1, 2, \ldots, 10)$ depending on the combinations of sine and cosine functions entering into the eigenfunctions (3.20) are presented in Appendix A.

Since the matrix S_{ij}^{mn} is symmetric, then it has three real eigenvalues $\Omega_{i\,mn}$ $(i = 1, 2, 3)$ which are in correspondence with three mutually orthogonal eigenvectors

$$
\boldsymbol{L}_{mn}^{\mathrm{I}}\left\{L_{i\,mn}^{\mathrm{I}}\right\}, \quad \boldsymbol{L}_{mn}^{\mathrm{II}}\left\{L_{i\,mn}^{\mathrm{II}}\right\}, \quad \boldsymbol{L}_{mn}^{\mathrm{III}}\left\{L_{i\,mn}^{\mathrm{III}}\right\}. \tag{3.26}
$$

Thus, the matrix S_{ij}^{mn} and the generalized displacements $x_{i\,mn}$ could be expanded in terms of the vectors (3.26) as

$$
S_{ij}^{mn} = \Omega_{1\,mn}^2 L_{i\,mn}^{\mathrm{I}} L_{j\,mn}^{\mathrm{I}} + \Omega_{2\,mn}^2 L_{i\,mn}^{\mathrm{II}} L_{j\,mn}^{\mathrm{II}} + \Omega_{3\,mn}^2 L_{i\,mn}^{\mathrm{III}} L_{j\,mn}^{\mathrm{III}}, \tag{3.27}
$$

$$
x_{i\,mn} = X_{1\,mn} L_{i\,mn}^{\mathrm{I}} + X_{2\,mn} L_{i\,mn}^{\mathrm{II}} + X_{3\,mn} L_{i\,mn}^{\mathrm{III}}. \tag{3.28}
$$

Substituting (3.27) and (3.28) in Eq. (3.21) and then multiplying them successively by $L_{i\,mn}^{\mathrm{I}}$, $L_{i\,mn}^{\mathrm{II}}$, and $L_{i\,mn}^{\mathrm{III}}$ with due account for

$$
\begin{aligned}
L_{i\,mn}^{\mathrm{I}} L_{i\,mn}^{\mathrm{II}} &= L_{i\,mn}^{\mathrm{I}} L_{i\,mn}^{\mathrm{III}} = L_{i\,mn}^{\mathrm{II}} L_{i\,mn}^{\mathrm{III}} = 0, \\
L_{i\,mn}^{\mathrm{I}} L_{i\,mn}^{\mathrm{I}} &= L_{i\,mn}^{\mathrm{II}} L_{i\,mn}^{\mathrm{II}} = L_{i\,mn}^{\mathrm{III}} L_{i\,mn}^{\mathrm{III}} = 1,
\end{aligned} \tag{3.29}
$$

we obtain the following three equations:

$$
\ddot{X}_{1mn} + \text{æ}_1 D^{\gamma} X_{1mn} + \Omega_{1mn}^2 X_{1mn} = -\sum_{i=1}^{3} F_{i\,mn} L_{i\,mn}^{\mathrm{I}}, \tag{3.30}
$$

$$
\ddot{X}_{2mn} + \text{æ}_2 D^{\gamma} X_{2mn} + \Omega_{2mn}^2 X_{2mn} = -\sum_{i=1}^{3} F_{i\,mn} L_{i\,mn}^{\mathrm{II}}, \tag{3.31}
$$

$$
\ddot{X}_{3mn} + \text{æ}_3 D^{\gamma} X_{3mn} + \Omega_{3mn}^2 X_{3mn} = -\sum_{i=1}^{3} F_{i\,mn} L_{i\,mn}^{\mathrm{III}} \tag{3.32}
$$

in terms of three new generalized displacements $X_{1\,mn}$, $X_{2\,mn}$, and $X_{3\,mn}$

$$X_{1\,mn} = x_{i\,mn} L^{\mathrm{I}}_{i\,mn} = x_{1\,mn} L^{\mathrm{I}}_{1\,mn} + x_{2\,mn} L^{\mathrm{I}}_{2\,mn} + x_{3\,mn} L^{\mathrm{I}}_{3\,mn}, \tag{3.33}$$

$$X_{2\,mn} = x_{i\,mn} L^{\mathrm{II}}_{i\,mn} = x_{1\,mn} L^{\mathrm{II}}_{1\,mn} + x_{2\,mn} L^{\mathrm{II}}_{2\,mn} + x_{3\,mn} L^{\mathrm{II}}_{3\,mn}, \tag{3.34}$$

$$X_{3\,mn} = x_{i\,mn} L^{\mathrm{III}}_{i\,mn} = x_{1\,mn} L^{\mathrm{III}}_{1\,mn} + x_{2\,mn} L^{\mathrm{III}}_{2\,mn} + x_{3\,mn} L^{\mathrm{III}}_{3\,mn}\,. \tag{3.35}$$

It should be emphasized that left-hand side parts of (3.30)–(3.32) are linear and independent of each other, while Eqs. (3.30)–(3.32) are coupled only by non-linear terms in their right-hand sides.

In order to show the influence of the initial conditions (3.4) and (3.5) on the solution to be constructed, let us expand the desired functions $X_{i\,mn}$ $(i = 1, 2, 3)$ in a series in terms of the small parameter ε

$$X_{i\,mn} = \varepsilon X^0_{i\,mn} + \varepsilon^2 X^1_{i\,mn} + \ldots \quad (i = 1, 2, 3)\,. \tag{3.36}$$

Substituting (3.36) in the set of Eqs (3.30)–(3.32) and restricting ourselves by the terms of the order of ε, we are led to a linear homogeneous set of differential equations

$$\ddot{X}^0_{i\,mn} + \Omega^2_{i\,mn} X^0_{i\,mn} = 0 \quad (i = 1, 2, 3)\,. \tag{3.37}$$

The solution of (3.37) has the form

$$X^0_{i\,mn} = A_{i\,mn}(\varepsilon t) \exp(i\,\Omega_{i\,mn} t) + \bar{A}_{i\,mn}(\varepsilon t) \exp(-i\,\Omega_{i\,mn} t), \tag{3.38}$$

where $A_{i\,mn}(\varepsilon t)$ and $\bar{A}_{i\,mn}(\varepsilon t)$ $(i = 1, 2, 3)$ are complex conjugate functions to be found.

Representing functions $A_{i\,mn}(\varepsilon t)$ in the polar form

$$A_{i\,mn}(\varepsilon t) = a_{i\,mn}(\varepsilon t) e^{i\,\psi_{i\,mn}(\varepsilon t)} \quad (i = 1, 2, 3), \tag{3.39}$$

the solution (3.36) is reduced to

$$X_{i\,mn} = \varepsilon X^0_{i\,mn} = 2\varepsilon a_{i\,mn}(\varepsilon t) \cos\left[\Omega_{i\,mn} t + \psi_{i\,mn}(\varepsilon t)\right], \tag{3.40}$$

where $a_{i\,mn}(\varepsilon t)$ and $\psi_{i\,mn}(\varepsilon t)$ $(i = 1, 2, 3)$ are the amplitudes and phases of non-linear vibrations, respectively.

Differentiating (3.40) with respect to time t and ignoring the terms of the order of ε^2, we obtain

$$\dot{X}_{i\,mn} = -2\varepsilon a_{i\,mn}(\varepsilon t) \Omega_{i\,mn} \sin\left[\Omega_{i\,mn} t + \psi_{i\,mn}(\varepsilon t)\right]\,. \tag{3.41}$$

Now substituting (3.28) in (3.17)–(3.19) with due account for (3.40), we have

$$u(x,\varphi,t) = 2\varepsilon \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \{a_{1mn}(\varepsilon t)\cos[\Omega_{1mn}t + \psi_{1mn}(\varepsilon t)]\, L^{\mathrm{I}}_{1mn} + a_{2mn}(\varepsilon t)\cos[\Omega_{2mn}t + \psi_{2mn}(\varepsilon t)]\, L^{\mathrm{II}}_{1mn} + a_{3mn}(\varepsilon t)\cos[\Omega_{3mn}t + \psi_{3mn}(\varepsilon t)]\, L^{\mathrm{III}}_{1mn}\} \eta_{1mn}(x,\varphi), \quad (3.42)$$

$$v(x,\varphi,t) = 2\varepsilon \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \{a_{1mn}(\varepsilon t)\cos[\Omega_{1mn}t + \psi_{1mn}(\varepsilon t)]\, L^{\mathrm{I}}_{2mn} + a_{2mn}(\varepsilon t)\cos[\Omega_{2mn}t + \psi_{2mn}(\varepsilon t)]\, L^{\mathrm{II}}_{2mn} + a_{3mn}(\varepsilon t)\cos[\Omega_{3mn}t + \psi_{3mn}(\varepsilon t)]\, L^{\mathrm{III}}_{2mn}\} \eta_{2mn}(x,\varphi), \quad (3.43)$$

$$w(x,\varphi,t) = 2\varepsilon \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \{a_{1mn}(\varepsilon t)\cos[\Omega_{1mn}t + \psi_{1mn}(\varepsilon t)]\, L^{\mathrm{I}}_{3mn} + a_{2mn}(\varepsilon t)\cos[\Omega_{2mn}t + \psi_{2mn}(\varepsilon t)]\, L^{\mathrm{II}}_{3mn} + a_{3mn}(\varepsilon t)\cos[\Omega_{3mn}t + \psi_{3mn}(\varepsilon t)]\, L^{\mathrm{III}}_{3mn}\} \eta_{3mn}(x,\varphi)\,. \quad (3.44)$$

Using the initial conditions (3.11) and (3.12) with due account for relationships (3.41), from (3.42)–(3.44) at $t = 0$ we obtain an infinite set of algebraic equations for determining $a_{i\,mn}(0)$ and $\psi_{i\,mn}(0)$ $(i = 1, 2, 3)$

$$\begin{aligned} a_{1mn}(0)L^{\mathrm{I}}_{1mn}\cos\psi_{1mn}(0) + a_{2mn}(0)L^{\mathrm{II}}_{1mn}\cos\psi_{2mn}(0) \\ + a_{3mn}(0)L^{\mathrm{III}}_{1mn}\cos\psi_{3mn}(0) = 0, \\ a_{1mn}(0)L^{\mathrm{I}}_{2mn}\cos\psi_{1mn}(0) + a_{2mn}(0)L^{\mathrm{II}}_{2mn}\cos\psi_{2mn}(0) \\ + a_{3mn}(0)L^{\mathrm{III}}_{2mn}\cos\psi_{3mn}(0) = 0, \\ a_{1mn}(0)L^{\mathrm{I}}_{3mn}\cos\psi_{1mn}(0) + a_{2mn}(0)L^{\mathrm{II}}_{3mn}\cos\psi_{2mn}(0) \\ + a_{3mn}(0)L^{\mathrm{III}}_{3mn}\cos\psi_{3mn}(0) = 0, \end{aligned} \quad (3.45)$$

$$\begin{aligned} -a_{1mn}(0)\Omega_{1mn}L^{\mathrm{I}}_{1mn}\sin\psi_{1mn}(0) - a_{2mn}(0)\Omega_{2mn}L^{\mathrm{II}}_{1mn}\sin\psi_{2mn}(0) \\ - a_{3mn}(0)\Omega_{3mn}L^{\mathrm{III}}_{1mn}\sin\psi_{3mn}(0) = æ_{1mn}, \\ -a_{1mn}(0)\Omega_{1mn}L^{\mathrm{I}}_{2mn}\sin\psi_{1mn}(0) - a_{2mn}(0)\Omega_{2mn}L^{\mathrm{II}}_{2mn}\sin\psi_{2mn}(0) \\ - a_{3mn}(0)\Omega_{3mn}L^{\mathrm{III}}_{2mn}\sin\psi_{3mn}(0) = æ_{2mn}, \\ -a_{1mn}(0)\Omega_{1mn}L^{\mathrm{I}}_{3mn}\sin\psi_{1mn}(0) - a_{2mn}(0)\Omega_{2mn}L^{\mathrm{II}}_{3mn}\sin\psi_{2mn}(0) \\ - a_{3mn}(0)\Omega_{3mn}L^{\mathrm{III}}_{3mn}\sin\psi_{3mn}(0) = æ_{3mn}, \end{aligned} \quad (3.46)$$

where

$$\text{æ}_{1mn} = \frac{\int_0^1 \int_0^{2\pi} v_1^0(x,\varphi)\eta_{1mn}(x,\varphi)dxd\varphi}{\int_0^1 \int_0^{2\pi} \eta_{1mn}^2(x,\varphi)dxd\varphi},$$

$$\text{æ}_{2mn} = \frac{\int_0^1 \int_0^{2\pi} v_2^0(x,\varphi)\eta_{2mn}(x,\varphi)dxd\varphi}{\int_0^1 \int_0^{2\pi} \eta_{2mn}^2(x,\varphi)dxd\varphi},$$

$$\text{æ}_{3mn} = \frac{\int_0^1 \int_0^{2\pi} v_3^0(x,\varphi)\eta_{3mn}(x,\varphi)dxd\varphi}{\int_0^1 \int_0^{2\pi} \eta_{3mn}^2(x,\varphi)dxd\varphi}.$$

All subsequent approximations are determined from an inhomogeneous set of differential equations with known right parts. Since the general solution of such a system is the sum of two solutions, a particular solution of the inhomogeneous system and a general solution of the corresponding homogeneous system, then arbitrary constants could be chosen in such a way that the initial conditions of all subsequent approximations would be zero ones.

It is known [3, 30] that during nonstationary excitation of thin bodies not all possible modes of vibration would be excited. Moreover, the modes which are strongly coupled by any of the so-called internal resonance conditions are initiated and dominate in the process of vibration, resulting in the energy transfer from one subsystem to another between the coupled modes, in so doing the types of modes to be excited are dependent of the character of the external excitation.

Assume hereafter that the vibration process occurs in such a way that only three natural modes corresponding to the generalized displacements $X_{1s_1s_2}$, $X_{2k_1k_2}$, and $X_{3l_1l_2}$ are excited and dominate over other natural modes. In this case, the right parts of Eqs. (3.30)–(3.32) are significantly simplified, and equations of free vibrations (3.30)–(3.32) take the form

$$\begin{aligned}\ddot{X}_{1s_1s_2} &+ \text{æ}_1 D^\gamma X_{1s_1s_2} + \Omega_{1s_1s_2}^2 X_{1s_1s_2} + a_{11s_1s_2}^I X_{1s_1s_2}^2 \\ &+ a_{22s_1s_2}^I X_{2k_1k_2}^2 + a_{33s_1s_2}^I X_{3l_1l_2}^2 + a_{12s_1s_2}^I X_{1s_1s_2} X_{2k_1k_2} \\ &+ a_{13s_1s_2}^I X_{1s_1s_2} X_{3l_1l_2} + a_{23s_1s_2}^I X_{2k_1k_2} X_{3l_1l_2} = 0, \end{aligned} \tag{3.47}$$

$$\begin{aligned}\ddot{X}_{2k_1k_2} &+ \text{æ}_2 D^\gamma X_{2k_1k_2} + \Omega_{2k_1k_2}^2 X_{2k_1k_2} + a_{11k_1k_2}^{II} X_{1s_1s_2}^2 \\ &+ a_{22k_1k_2}^{II} X_{2k_1k_2}^2 + a_{33k_1k_2}^{II} X_{3l_1l_2}^2 + a_{12k_1k_2}^{II} X_{1s_1s_2} X_{2k_1k_2} \\ &+ a_{13k_1k_2}^{II} X_{1s_1s_2} X_{3l_1l_2} + a_{23k_1k_2}^{II} X_{2k_1k_2} X_{3l_1l_2} = 0, \end{aligned} \tag{3.48}$$

$$\begin{aligned}\ddot{X}_{3l_1l_2} &+ \text{æ}_3 D^\gamma X_{3l_1l_2} + \Omega_{3l_1l_2}^2 X_{3l_1l_2} + a_{11l_1l_2}^{III} X_{1s_1s_2}^2 \\ &+ a_{22l_1l_2}^{III} X_{2k_1k_2}^2 + a_{33l_1l_2}^{III} X_{3l_1l_2}^2 + a_{12l_1l_2}^{III} X_{1s_1s_2} X_{2k_1k_2} \\ &+ a_{13l_1l_2}^{III} X_{1s_1s_2} X_{3l_1l_2} + a_{23l_1l_2}^{III} X_{2k_1k_2} X_{3l_1l_2} = 0, \end{aligned} \tag{3.49}$$

where

$$a^{\alpha}_{11mn} = \frac{2}{\pi}\left\{\left(L^{\mathrm{I}}_{3s_1s_2}\right)^2\left(L^{\alpha}_{1mn}A^{s_1s_2s_1s_2}_{mn} + L^{\alpha}_{2mn}B^{s_1s_2s_1s_2}_{mn} + L^{\alpha}_{3mn}E^{s_1s_2s_1s_2}_{mn}\right)\right.$$
$$\left. + \left(L^{I}_{1s_1s_2}C^{s_1s_2s_1s_2}_{mn} + L^{I}_{2s_1s_2}D^{s_1s_2s_1s_2}_{mn}\right)L^{\mathrm{I}}_{3s_1s_2}L^{\alpha}_{3mn}\right\}, \tag{3.50}$$

$$a^{\alpha}_{22mn} = \frac{2}{\pi}\left\{\left(L^{\mathrm{II}}_{3k_1k_2}\right)^2\left(L^{\alpha}_{1mn}A^{k_1k_2k_1k_2}_{mn} + L^{\alpha}_{2mn}B^{k_1k_2k_1k_2}_{mn} + L^{\alpha}_{3mn}E^{k_1k_2k_1k_2}_{mn}\right)\right.$$
$$\left. + \left(L^{II}_{1k_1k_2}C^{k_1k_2k_1k_2}_{mn} + L^{II}_{2k_1k_2}D^{k_1k_2k_1k_2}_{mn}\right)L^{\mathrm{II}}_{3k_1k_2}L^{\alpha}_{3mn}\right\}, \tag{3.51}$$

$$a^{\alpha}_{33mn} = \frac{2}{\pi}\left\{\left(L^{\mathrm{III}}_{3l_1l_2}\right)^2\left(L^{\alpha}_{1mn}A^{l_1l_2l_1l_2}_{mn} + L^{\alpha}_{2mn}B^{l_1l_2l_1l_2}_{mn} + L^{\alpha}_{3mn}E^{l_1l_2l_1l_2}_{mn}\right)\right.$$
$$\left. + \left(L^{III}_{1l_1l_2}C^{l_1l_2l_1l_2}_{mn} + L^{III}_{2l_1l_2}D^{l_1l_2l_1l_2}_{mn}\right)L^{\mathrm{III}}_{3l_1l_2}L^{\alpha}_{3mn}\right\}, \tag{3.52}$$

$$a^{\alpha}_{13mn} = \frac{2}{\pi}\left\{L^{\mathrm{I}}_{3s_1s_2}L^{\mathrm{III}}_{3l_1l_2}\left[\left(A^{s_1s_2l_1l_2}_{mn} + A^{l_1l_2s_1s_2}_{mn}\right)L^{\alpha}_{1mn}\right.\right.$$
$$+ \left(B^{s_1s_2l_1l_2}_{mn} + B^{l_1l_2s_1s_2}_{mn}\right)L^{\alpha}_{2mn} + \left(E^{s_1s_2l_1l_2}_{mn} + E^{l_1l_2s_1s_2}_{mn}\right)L^{\alpha}_{3mn}\Big]$$
$$+ \Big[L^{\mathrm{I}}_{3s_1s_2}L^{\mathrm{III}}_{1l_1l_2}C^{s_1s_2l_1l_2}_{mn} + L^{\mathrm{III}}_{3l_1l_2}L^{\mathrm{I}}_{1s_1s_2}C^{l_1l_2s_1s_2}_{mn}$$
$$\left.\left. + L^{\mathrm{I}}_{3s_1s_2}L^{\mathrm{III}}_{2l_1l_2}D^{s_1s_2l_1l_2}_{mn} + L^{\mathrm{III}}_{3l_1l_2}L^{\mathrm{I}}_{2s_1s_2}D^{l_1l_2s_1s_2}_{mn}\Big]L^{\alpha}_{3mn}\right\}, \tag{3.53}$$

$$a^{\alpha}_{23mn} = \frac{2}{\pi}\left\{L^{\mathrm{II}}_{3k_1k_2}L^{\mathrm{III}}_{3l_1l_2}\left[\left(A^{k_1k_2l_1l_2}_{mn} + A^{l_1l_2k_1k_2}_{mn}\right)L^{\alpha}_{1mn}\right.\right.$$
$$+ \left(B^{k_1k_2l_1l_2}_{mn} + B^{l_1l_2k_1k_2}_{mn}\right)L^{\alpha}_{2mn} + \left(E^{k_1k_2l_1l_2}_{mn} + E^{l_1l_2k_1k_2}_{mn}\right)L^{\alpha}_{3mn}\Big]$$
$$+ \Big[L^{\mathrm{II}}_{3k_1k_2}L^{\mathrm{III}}_{1l_1l_2}C^{k_1k_2l_1l_2}_{mn} + L^{\mathrm{III}}_{3l_1l_2}L^{\mathrm{II}}_{1k_1k_2}C^{l_1l_2k_1k_2}_{mn}$$
$$\left.\left. + L^{\mathrm{II}}_{3k_1k_2}L^{\mathrm{III}}_{2l_1l_2}D^{k_1k_2l_1l_2}_{mn} + L^{\mathrm{III}}_{3l_1l_2}L^{\mathrm{II}}_{2k_1k_2}D^{l_1l_2k_1k_2}_{mn}\Big]L^{\alpha}_{3mn}\right\}, \tag{3.54}$$

$$a^{\alpha}_{12mn} = \frac{2}{\pi}\left\{L^{\mathrm{I}}_{3s_1s_2}L^{\mathrm{II}}_{3k_1k_2}\left[\left(A^{s_1s_2k_1k_2}_{mn} + A^{k_1k_2s_1s_2}_{mn}\right)L^{\alpha}_{1mn}\right.\right.$$
$$+ \left(B^{s_1s_2k_1k_2}_{mn} + B^{k_1k_2s_1s_2}_{mn}\right)L^{\alpha}_{2mn} + \left(E^{s_1s_2k_1k_2}_{mn} + E^{k_1k_2s_1s_2}_{mn}\right)L^{\alpha}_{3mn}\Big]$$
$$+ \Big[L^{\mathrm{I}}_{3s_1s_2}L^{\mathrm{II}}_{1k_1k_2}C^{s_1s_2k_1k_2}_{mn} + L^{\mathrm{II}}_{3k_1k_2}L^{\mathrm{I}}_{1s_1s_2}C^{k_1k_2s_1s_2}_{mn}$$
$$\left.\left. + L^{\mathrm{I}}_{3s_1s_2}L^{\mathrm{II}}_{2k_1k_2}D^{s_1s_2k_1k_2}_{mn} + L^{\mathrm{II}}_{3k_1k_2}L^{\mathrm{I}}_{2s_1s_2}D^{k_1k_2s_1s_2}_{mn}\Big]L^{\alpha}_{3mn}\right\}. \tag{3.55}$$

From relationships (3.50)–(3.55) we could calculate all coefficients entering in Eq. (3.47) at $\alpha = I$, $m = s_1$ and $n = s_2$, in Eq. (3.48) at $\alpha = II$, $m = k_1$ and $n = k_2$, and in Eq. (3.49) at $\alpha = III$, $m = l_1$ and $n = l_2$.

Omitting hereafter the subindices s_1s_2, k_1k_2, and l_1l_2 for ease of presentation, Eqs. (3.47)–(3.49) could be rewritten as

$$\ddot{X}_1 + æ_1 D^\gamma X_1 + \Omega_1^2 X_1 + a_{11}^I X_1^2 + a_{22}^I X_2^2 + a_{33}^I X_3^2 + a_{12}^I X_1 X_2 + a_{13}^I X_1 X_3 + a_{23}^I X_2 X_3 = 0, \tag{3.56}$$

$$\ddot{X}_2 + æ_2 D^\gamma X_2 + \Omega_2^2 X_2 + a_{11}^{II} X_1^2 + a_{22}^{II} X_2^2 + a_{33}^{II} X_3^2 + a_{12}^{II} X_1 X_2 + a_{13}^{II} X_1 X_3 + a_{23}^{II} X_2 X_3 = 0, \tag{3.57}$$

$$\ddot{X}_3 + æ_3 D^\gamma X_3 + \Omega_3^2 X_3 + a_{11}^{III} X_1^2 + a_{22}^{III} X_2^2 + a_{33}^{III} X_3^2 + a_{12}^{III} X_1 X_2 + a_{13}^{III} X_1 X_3 + a_{23}^{III} X_2 X_3 = 0\,. \tag{3.58}$$

3.3 Method of Solution

An approximate solution of Eqs. (3.56)–(3.58) for small but finite amplitudes weakly varying with time can be represented by a third-order uniform expansion in terms of different time scales in the following form [31]:

$$X_i = \varepsilon X_{i1}(T_0, T_1, T_2 \ldots) + \varepsilon^2 X_{i2}(T_0, T_1, T_2 \ldots) + \varepsilon^3 X_{i3}(T_0, T_1, T_2 \ldots) + \ldots, \tag{3.59}$$

where $i = 1, 2, 3$, ε is a small dimensionless parameter of the same order of magnitude as the amplitudes, $T_n = \varepsilon^n t$ are new independent variables, among them: $T_0 = t$ is a fast scale characterizing motions with the natural frequencies, and $T_1 = \varepsilon t$ and $T_2 = \varepsilon^2 t$ are slow scales characterizing the modulation of the amplitudes and phases of the modes with nonlinearity.

Recall that the first and the second time-derivatives are defined, respectively, as follows

$$\frac{d}{dt} = D_0 + \varepsilon D_1 + \varepsilon^2 D_2 + \ldots,$$

$$\frac{d^2}{dt^2} = D_0^2 + 2\varepsilon D_0 D_1 + \varepsilon^2 \left(D_1^2 + 2 D_0 D_2\right) + \ldots, \tag{3.60}$$

while the fractional-order time-derivative could be represented following Rossikhin and Shitikova [20] as

$$\left(\frac{d}{dt}\right)^\gamma = \left(D_0 + \varepsilon D_1 + \varepsilon^2 D_2 + \ldots\right)^\gamma \tag{3.61}$$

$$= D_0^\gamma + \varepsilon \gamma D_0^{\gamma-1} D_1 + \frac{1}{2}\,\varepsilon^2 \gamma \left[(\gamma - 1) D_0^{\gamma-2} D_1^2 + 2 D_0^{\gamma-1} D_2\right] + \ldots,$$

where $D_n = \partial / \partial T_n$, and D_0^γ, $D_0^{\gamma-1}$, $D_0^{\gamma-2}$, ... are the Riemann-Liouville fractional derivatives in time t defined in (3.16).

Considering that the viscosity is small, i.e.,

$$æ_i = \varepsilon^k \mu_i \tau_i^\gamma \ ,$$

where τ_i is the relaxation time of the i-th generalized displacement, μ_i is a finite value, and the choice of k depends on the order of smallness of the viscosity coefficients $æ_i$, substituting (3.59)–(3.61) in Eqs. (3.56)–(3.58), after equating the coefficients at like powers of ε to zero, we are led to a set of recurrence equations to various orders:

to order ε

$$D_0^2 X_{11} + \Omega_1^2 X_{11} = 0, \tag{3.62}$$

$$D_0^2 X_{21} + \Omega_2^2 X_{21} = 0, \tag{3.63}$$

$$D_0^2 X_{31} + \Omega_3^2 X_{31} = 0; \tag{3.64}$$

to order ε^2

$$\begin{aligned} D_0^2 X_{12} + \Omega_1^2 X_{12} &= -2D_0 D_1 X_{11} - a_{11}^I X_{11}^2 - a_{22}^I X_{21}^2 - a_{33}^I X_{31}^2 \\ &- a_{12}^I X_{11} X_{21} - a_{13}^I X_{11} X_{31} - a_{23}^I X_{21} X_{31} - \mu_1 (2-k) \tau_1^\gamma D_0^\gamma X_{11}, \end{aligned} \tag{3.65}$$

$$\begin{aligned} D_0^2 X_{22} + \Omega_2^2 X_{22} &= -2D_0 D_1 X_{21} - a_{11}^{II} X_{11}^2 - a_{22}^{II} X_{21}^2 - a_{33}^{II} X_{31}^2 \\ &- a_{12}^{II} X_{11} X_{21} - a_{13}^{II} X_{11} X_{31} - a_{23}^{II} X_{21} X_{31} - \mu_2 (2-k) \tau_2^\gamma D_0^\gamma X_{21}, \end{aligned} \tag{3.66}$$

$$\begin{aligned} D_0^2 X_{32} + \Omega_3^2 X_{32} &= -2D_0 D_1 X_{31} - a_{11}^{III} X_{11}^2 - a_{22}^{III} X_{21}^2 - a_{33}^{III} X_{31}^2 - a_{12}^{III} X_{11} X_{21} \\ &- a_{13}^{III} X_{11} X_{31} - a_{23}^{III} X_{21} X_{31} - \mu_3 (2-k) \tau_3^\gamma D_0^\gamma X_{31} \ , \end{aligned} \tag{3.67}$$

to order ε^3

$$\begin{aligned} D_0^2 X_{13} + \Omega_1^2 X_{13} &= -2D_0 D_1 X_{12} - \left(D_1^2 + 2D_0 D_2\right) X_{11} \\ &- 2a_{11}^I X_{11} X_{12} - 2a_{22}^I X_{21} X_{22} - 2a_{33}^I X_{31} X_{32} \\ &- a_{12}^I \left(X_{11} X_{22} + X_{12} X_{21}\right) - a_{13}^I \left(X_{11} X_{32} + X_{12} X_{31}\right) \\ &- a_{23}^I \left(X_{21} X_{32} + X_{22} X_{31}\right) - \mu_1 (2-k) \tau_1^\gamma D_0^\gamma X_{12} \\ &- \mu_1 \gamma (2-k) \tau_1^\gamma D_0^{\gamma-1} D_1 X_{11} - \mu_1 (k-1) \tau_1^\gamma D_0^\gamma X_{11}, \end{aligned} \tag{3.68}$$

$$\begin{aligned} D_0^2 X_{23} + \Omega_2^2 X_{23} &= -2D_0 D_1 X_{22} - \left(D_1^2 + 2D_0 D_2\right) X_{21} \\ &- 2a_{11}^{II} X_{11} X_{12} - 2a_{22}^{II} X_{21} X_{22} - 2a_{33}^{II} X_{31} X_{32} \\ &- a_{12}^{II} \left(X_{11} X_{22} + X_{12} X_{21}\right) - a_{13}^{II} \left(X_{11} X_{32} + X_{12} X_{31}\right) \\ &- a_{23}^{II} \left(X_{21} X_{32} + X_{22} X_{31}\right) - \mu_2 (2-k) \tau_2^\gamma D_0^\gamma X_{22} \\ &- \mu_2 \gamma (2-k) \tau_2^\gamma D_0^{\gamma-1} D_1 X_{21} - \mu_2 (k-1) \tau_2^\gamma D_0^\gamma X_{21}, \end{aligned} \tag{3.69}$$

$$
\begin{aligned}
D_0^2 X_{33} + \Omega_3^2 X_{33} = & -2D_0 D_1 X_{32} - \left(D_1^2 + 2D_0 D_2\right) X_{31} \\
& - 2a_{11}^{III} X_{11} X_{12} - 2a_{22}^{III} X_{21} X_{22} - 2a_{33}^{III} X_{31} X_{32} \\
& - a_{12}^{III} \left(X_{11} X_{22} + X_{12} X_{21}\right) - a_{13}^{III} \left(X_{11} X_{32} + X_{12} X_{31}\right) \\
& - a_{23}^{III} \left(X_{21} X_{32} + X_{22} X_{31}\right) - \mu_3 (2-k) \tau_3^{\gamma} D_0^{\gamma} X_{32} \\
& - \mu_3 \gamma (2-k) \tau_3^{\gamma} D_0^{\gamma-1} D_1 X_{31} - \mu_3 (k-1) \tau_3^{\gamma} D_0^{\gamma} X_{31} \,. \qquad (3.70)
\end{aligned}
$$

In order to construct the uniformly valid solution, it is necessary on each step to use the solution from the preceding step and to eliminate secular terms during integration [31].

We shall seek the solution of Eqs. (3.62)–(3.64) in the form

$$
X_{j1} = A_j(T_1, T_2) \exp(i \Omega_j T_0) + \bar{A}_j(T_1, T_2) \exp(-i \Omega_j T_0), \qquad (3.71)
$$

where $A_j(T_1, T_2)$ $(j = 1, 2, 3)$ are unknown complex functions, and $\bar{A}_j(T_1, T_2)$ are the complex conjugates of $A_j(T_1, T_2)$.

To solve the sets of Eqs. (3.65)–(3.67) and (3.68)–(3.70), it is necessary to specify the action of the fractional derivative D_0^{γ} on the functions X_{j1}, i.e., to calculate $D_0^{\gamma} e^{i \Omega_j t}$. It has been shown in [32] that

$$
D_0^{\gamma} e^{i \Omega_j t} = (i \Omega_j)^{\gamma} e^{i \Omega_j t}. \qquad (3.72)
$$

Note that since the process of vibrations starts at $t = 0$, then the fractional derivative should be defined on the segment $[0, t]$, i.e.,

$$
D_0^{\gamma} x(t) = \frac{1}{\Gamma(1-\gamma)} \frac{d}{dt} \int_0^t \frac{x(s) ds}{(t-s)^{\gamma}}. \qquad (3.73)
$$

Then instead of formula (3.72), the application of the fractional derivative (3.73) on the exponent results in the following relationship [32]:

$$
D_0^{\gamma} e^{i \Omega_j t} = (i \Omega_j)^{\gamma} e^{i \Omega_j t} + \frac{\sin \pi \gamma}{\pi} \int_0^{\infty} \frac{u^{\gamma}}{u + i \Omega_j} e^{-ut} du. \qquad (3.74)
$$

However, the second term of (3.74), as it has been proved in [33], does not influence the solution constructed via the method of multiple time scales restricted to the first- and second-order approximations.

In other words, even the utilization of exact formula (3.74) in the problem under consideration produces completely equivalent results given by the approximate formula (3.72) if the solution is constructed via the method of multiple time scales within the considered orders of approximation. Thus, in further analysis we will utilize formula (3.72).

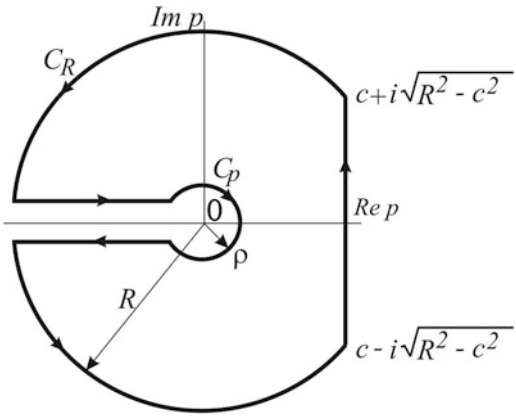


Fig. 3.1 Contour of integration

It has been shown by Rossikhin and Shitikova [32] that formula (3.74) could be obtained from the Mellin-Fourier formula for the function $D_0^{\gamma} e^{i\Omega_j t}$. Really,

$$D_0^{\gamma} e^{i\Omega_j t} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{p^{\gamma} e^{pt} dp}{p - i\Omega_j}. \tag{3.75}$$

To calculate the integral in the right-hand side of (3.75), one can use the contour of integration L shown in Fig. 3.1. Applying the theorem of residues for the integral over the contour L, we have

$$\frac{1}{2\pi i} \int_L \frac{p^{\gamma} e^{pt} dp}{p - i\Omega_j} = (i\Omega_j)^{\gamma} e^{i\Omega_j t}. \tag{3.76}$$

Writing the contour integral in terms of the sum of integrals along the vertical segment of the straight line, along the arcs of the circumferences with radii C_R and C_ρ, and along the branches of the cut of the negative real semi-axis, then tending R to ∞ and ρ to 0, and considering Jordan lemma, we arrive at the relationship (3.74).

Reference to Fig. 3.1 shows that the first term in (3.74) is defined by the residues at the point $p = i\Omega_j$, while the second term of (3.74) is governed by the integration along the cut of the negative semi-axis.

Note also that it has been shown in [33] that the utilization of the fractional derivatives due to the Caputo definition and Riemann-Liouville definition together with the method of multiple time scales produces the completely equivalent solution within the limits of the leading and second-order approximations.

For further analysis it is also a need to specify the order of weak damping.

3.3.1 Viscosity of the Order of ε

Let us first consider the case of viscosity of the order of ε. Then at $k = 1$ Eqs. (3.65)–(3.67) are reduced to

$$D_0^2 X_{12} + \Omega_1^2 X_{12} = -2D_0 D_1 X_{11} - a_{11}^I X_{11}^2 - a_{22}^I X_{21}^2 - a_{33}^I X_{31}^2 - a_{12}^I X_{11} X_{21} - a_{13}^I X_{11} X_{31} - a_{23}^I X_{21} X_{31} - \mu_1 \tau_1^\gamma D_0^\gamma X_{11}, \quad (3.77)$$

$$D_0^2 X_{22} + \Omega_2^2 X_{22} = -2D_0 D_1 X_{21} - a_{11}^{II} X_{11}^2 - a_{22}^{II} X_{21}^2 - a_{33}^{II} X_{31}^2 - a_{12}^{II} X_{11} X_{21} - a_{13}^{II} X_{11} X_{31} - a_{23}^{II} X_{21} X_{31} - \mu_2 \tau_2^\gamma D_0^\gamma X_{21}, \quad (3.78)$$

$$D_0^2 X_{32} + \Omega_3^2 X_{32} = -2D_0 D_1 X_{31} - a_{11}^{III} X_{11}^2 - a_{22}^{III} X_{21}^2 - a_{33}^{III} X_{31}^2 - a_{12}^{III} X_{11} X_{21} - a_{13}^{III} X_{11} X_{31} - a_{23}^{III} X_{21} X_{31} - \mu_3 \tau_3^\gamma D_0^\gamma X_{31}\,. \quad (3.79)$$

Substituting (3.71) in the right-hand side of Eqs. (3.77)–(3.79) with due account for (3.72) yields

$$\begin{aligned} D_0^2 X_{12} + \Omega_1^2 X_{12} = & -2i\Omega_1 D_1 A_1(T_1)\exp(i\Omega_1 T_0) - a_{11}^I \left[A_1^2 \exp(2i\Omega_1 T_0) + A_1 \bar{A}_1\right] \\ & - a_{22}^I \left[A_2^2 \exp(2i\Omega_2 T_0) + A_2 \bar{A}_2\right] - a_{33}^I \left[A_3^2 \exp(2i\Omega_3 T_0) + A_3 \bar{A}_3\right] \\ & - a_{12}^I \left\{A_1 A_2 \exp\left[i(\Omega_1 + \Omega_2)T_0\right] + A_1 \bar{A}_2 \exp\left[i(\Omega_1 - \Omega_2)T_0\right]\right\} \\ & - a_{13}^I \left\{A_1 A_3 \exp\left[i(\Omega_1 + \Omega_3)T_0\right] + A_1 \bar{A}_3 \exp\left[i(\Omega_1 - \Omega_3)T_0\right]\right\} \\ & - a_{23}^I \left\{A_2 A_3 \exp\left[i(\Omega_2 + \Omega_3)T_0\right] + A_2 \bar{A}_3 \exp\left[i(\Omega_2 - \Omega_3)T_0\right]\right\} \\ & - \mu_1 \tau_1^\gamma A_1 (i\Omega_1)^\gamma \exp(i\Omega_1 T_0) + cc, \end{aligned} \quad (3.80)$$

$$\begin{aligned} D_0^2 X_{22} + \Omega_2^2 X_{22} = & -2i\Omega_2 D_1 A_2(T_1)\exp(i\Omega_2 T_0) - a_{11}^{II} \left[A_1^2 \exp(2i\Omega_1 T_0) + A_1 \bar{A}_1\right] \\ & - a_{22}^{II} \left[A_2^2 \exp(2i\Omega_2 T_0) + A_2 \bar{A}_2\right] - a_{33}^{II} \left[A_3^2 \exp(2i\Omega_3 T_0) + A_3 \bar{A}_3\right] \\ & - a_{12}^{II} \left\{A_1 A_2 \exp\left[i(\Omega_1 + \Omega_2)T_0\right] + A_1 \bar{A}_2 \exp\left[i(\Omega_1 - \Omega_2)T_0\right]\right\} \\ & - a_{13}^{II} \left\{A_1 A_3 \exp\left[i(\Omega_1 + \Omega_3)T_0\right] + A_1 \bar{A}_3 \exp\left[i(\Omega_1 - \Omega_3)T_0\right]\right\} \\ & - a_{23}^{II} \left\{A_2 A_3 \exp\left[i(\Omega_2 + \Omega_3)T_0\right] + A_2 \bar{A}_3 \exp\left[i(\Omega_2 - \Omega_3)T_0\right]\right\} \\ & - \mu_2 \tau_2^\gamma A_2 (i\Omega_2)^\gamma \exp(i\Omega_2 T_0) + cc, \end{aligned} \quad (3.81)$$

$$\begin{aligned} D_0^2 X_{32} + \Omega_3^2 X_{32} = & -2i\Omega_3 D_1 A_3(T_1)\exp(i\Omega_3 T_0) - a_{11}^{III} \left[A_1^2 \exp(2i\Omega_1 T_0) + A_1 \bar{A}_1\right] \\ & - a_{22}^{III} \left[A_2^2 \exp(2i\Omega_2 T_0) + A_2 \bar{A}_2\right] - a_{33}^{III} \left[A_3^2 \exp(2i\Omega_3 T_0) + A_3 \bar{A}_3\right] \\ & - a_{12}^{III} \left\{A_1 A_2 \exp\left[i(\Omega_1 + \Omega_2)T_0\right] + A_1 \bar{A}_2 \exp\left[i(\Omega_1 - \Omega_2)T_0\right]\right\} \end{aligned}$$

$$- a_{13}^{III}\left\{A_1 A_3 \exp\left[i(\Omega_1+\Omega_3)T_0\right] + A_1\bar{A}_3 \exp\left[i(\Omega_1-\Omega_3)T_0\right]\right\}$$
$$- a_{23}^{III}\left\{A_2 A_3 \exp\left[i(\Omega_2+\Omega_3)T_0\right] + A_2\bar{A}_3 \exp\left[i(\Omega_2-\Omega_3)T_0\right]\right\}$$
$$- \mu_3\tau_3^{\gamma} A_3\,(i\Omega_3)^{\gamma}\exp(i\Omega_3 T_0) + cc\,, \tag{3.82}$$

where cc is the complex conjugate part to the preceding terms.

Reference to Eqs. (3.80)–(3.82) shows that the following types of the internal resonance could occur on this step [25]:

1. the two-to-one internal resonance, when one natural frequency is twice the other natural frequency,

$$\Omega_1 = 2\Omega_2 \quad (\Omega_3 \neq \Omega_1, \quad \Omega_3 \neq 2\Omega_2), \quad \text{or} \quad \Omega_2 = 2\Omega_1, \tag{3.83}$$
$$\Omega_1 = 2\Omega_3 \quad (\Omega_2 \neq \Omega_1, \quad \Omega_2 \neq 2\Omega_3), \quad \text{or} \quad \Omega_3 = 2\Omega_1, \tag{3.84}$$
$$\Omega_2 = 2\Omega_3 \quad (\Omega_1 \neq \Omega_2, \quad \Omega_1 \neq 2\Omega_3), \quad \text{or} \quad \Omega_3 = 2\Omega_2; \tag{3.85}$$

2. the one-to-one-to-two or one-to-two-to-two internal resonance, i.e.,

$$\Omega_1 = \Omega_2 = 2\Omega_3, \quad \text{or} \quad 1:1:2, \tag{3.86}$$
$$\Omega_1 = 2\Omega_2 = \Omega_3, \quad \text{or} \quad 1:2:1, \tag{3.87}$$
$$2\Omega_1 = \Omega_2 = \Omega_3, \quad \text{or} \quad 2:1:1, \tag{3.88}$$
$$\Omega_1 = 2\Omega_2 = 2\Omega_3, \quad \text{or} \quad 1:2:2, \tag{3.89}$$
$$2\Omega_1 = \Omega_2 = 2\Omega_3, \quad \text{or} \quad 2:1:2, \tag{3.90}$$
$$2\Omega_1 = 2\Omega_2 = \Omega_3, \quad \text{or} \quad 2:2:1; \tag{3.91}$$

3. the combinational resonance of the additive-difference type of the first order, i.e.,

$$\Omega_1 = \Omega_2 + \Omega_3, \quad \text{or} \quad \Omega_2 = \Omega_1 - \Omega_3, \quad \text{or} \quad \Omega_3 = \Omega_1 - \Omega_2, \tag{3.92}$$
$$\Omega_1 = \Omega_2 - \Omega_3, \quad \text{or} \quad \Omega_2 = \Omega_1 + \Omega_3, \quad \text{or} \quad \Omega_3 = \Omega_2 - \Omega_1, \tag{3.93}$$
$$\Omega_1 = \Omega_3 - \Omega_2, \quad \text{or} \quad \Omega_2 = \Omega_3 - \Omega_1, \quad \text{or} \quad \Omega_3 = \Omega_1 + \Omega_2. \tag{3.94}$$

3.3.2 *Viscosity of the Order of ε^2*

Now let us consider vibrations of a nonlinear shell putting $k = 2$ in its equations of motion (3.65)–(3.67)

$$D_0^2 X_{j2} + \Omega_j^2 X_{j2} = -2D_0 D_1 X_{j1} - \sum_{k=1}^{3} a_{kk}^J X_{k1}^2$$
$$- a_{12}^J X_{11} X_{21} - a_{13}^J X_{11} X_{31} - a_{23}^J X_{21} X_{31}, \tag{3.95}$$

where the upper index J takes on the values of I, II, III for $j = 1, 2, 3$, respectively.

Substituting (3.71) in the right-hand side of Eqs. (3.95) yields

$$D_0^2 X_{j2} + \Omega_j^2 X_{j2} = -2i\Omega_j D_1 A_j(T_1)\exp(i\Omega_j T_0) - \sum_{k=1}^{3} a_{kk}^J \left[A_k^2 \exp(2i\Omega_k T_0) + A_k \bar{A}_k\right]$$
$$- a_{12}^J \left\{A_1 A_2 \exp[i(\Omega_1+\Omega_2)T_0] + A_1 \bar{A}_2 \exp[i(\Omega_1-\Omega_2)T_0]\right\}$$
$$- a_{13}^J \left\{A_1 A_3 \exp[i(\Omega_1+\Omega_3)T_0] + A_1 \bar{A}_3 \exp[i(\Omega_1-\Omega_3)T_0]\right\}$$
$$- a_{23}^J \left\{A_2 A_3 \exp[i(\Omega_2+\Omega_3)T_0] + A_2 \bar{A}_3 \exp[i(\Omega_2-\Omega_3)T_0]\right\} + cc\,. \qquad (3.96)$$

Assume that the natural frequencies Ω_j $(j = 1, 2, 3)$ possess such magnitudes that any of the combinations from (3.83) to (3.94) could not occur. Then the functions $\exp(\pm i\Omega_j T_0)$ entering into the right-hand side of Eqs. (3.96) produce secular terms.

To eliminate circular terms in Eqs. (3.96), it is necessary to vanish to zero the coefficients standing at $\exp\left(\pm i\Omega_j T_0\right)$, i.e.,

$$D_1 A_j(T_1, T_2) = 0, \qquad (3.97)$$

whence it follows that A_j are T_1-independent.

Considering (3.97), solution of Eqs. (3.96) at $j = 1, 2, 3$ has the form

$$X_{12} = k_1 A_1^2 \exp(2i\Omega_1 T_0) + k_2 A_1 \bar{A}_1 + k_3 A_2^2 \exp(2i\Omega_2 T_0) + k_4 A_2 \bar{A}_2$$
$$+ k_5 A_3^2 \exp(2i\Omega_3 T_0) + k_6 A_3 \bar{A}_3 + k_7 A_1 A_2 \exp[i(\Omega_1+\Omega_2)T_0]$$
$$+ k_8 A_1 \bar{A}_2 \exp[i(\Omega_1-\Omega_2)T_0] + k_9 A_1 A_3 \exp[i(\Omega_1+\Omega_3)T_0]$$
$$+ k_{10} A_1 \bar{A}_3 \exp[i(\Omega_1-\Omega_3)T_0] + k_{11} A_2 A_3 \exp[i(\Omega_2+\Omega_3)T_0]$$
$$+ k_{12} A_2 \bar{A}_3 \exp[i(\Omega_2-\Omega_3)T_0] + F_1(T_2)\exp(i\Omega_1 T_0) + cc\,, \qquad (3.98)$$

$$X_{22} = g_1 A_1^2 \exp(2i\Omega_1 T_0) + g_2 A_1 \bar{A}_1 + g_3 A_2^2 \exp(2i\Omega_2 T_0) + g_4 A_2 \bar{A}_2$$
$$+ g_5 A_3^2 \exp(2i\Omega_3 T_0) + g_6 A_3 \bar{A}_3 + g_7 A_1 A_2 \exp[i(\Omega_1+\Omega_2)T_0]$$
$$+ g_8 A_1 \bar{A}_2 \exp[i(\Omega_1-\Omega_2)T_0] + g_9 A_1 A_3 \exp[i(\Omega_1+\Omega_3)T_0]$$
$$+ g_{10} A_1 \bar{A}_3 \exp[i(\Omega_1-\Omega_3)T_0] + g_{11} A_2 A_3 \exp[i(\Omega_2+\Omega_3)T_0]$$
$$+ g_{12} A_2 \bar{A}_3 \exp[i(\Omega_2-\Omega_3)T_0] + F_2(T_2)\exp(i\Omega_2 T_0) + cc, \qquad (3.99)$$

$$X_{32} = q_1 A_1^2 \exp(2i\Omega_1 T_0) + q_2 A_1 \bar{A}_1 + q_3 A_2^2 \exp(2i\Omega_2 T_0) + q_4 A_2 \bar{A}_2$$
$$+ q_5 A_3^2 \exp(2i\Omega_3 T_0) + q_6 A_3 \bar{A}_3 + q_7 A_1 A_2 \exp[i(\Omega_1+\Omega_2)T_0]$$
$$+ q_8 A_1 \bar{A}_2 \exp[i(\Omega_1-\Omega_2)T_0] + q_9 A_1 A_3 \exp[i(\Omega_1+\Omega_3)T_0]$$

$$+ q_{10} A_1 \bar{A}_3 \exp\left[i(\Omega_1 - \Omega_3)T_0\right] + q_{11} A_2 A_3 \exp\left[i(\Omega_2 + \Omega_3)T_0\right]$$
$$+ q_{12} A_2 \bar{A}_3 \exp\left[i(\Omega_2 - \Omega_3)T_0\right] + F_3(T_2)\exp(i\Omega_3 T_0) + cc, \qquad (3.100)$$

where $F_j(T_2)$ $(j = 1, 2, 3)$ are new functions to be determined, and coefficients k_i, g_i, and q_i $(i = 1, 2, \ldots, 12)$ are presented in Appendix B.

Taking (3.97) into account, Eqs. (3.68)–(3.70) at $k = 2$ are reduced to the following set of equations:

$$D_0^2 X_{j3} + \Omega_j^2 X_{j3} = -2D_0 D_2 X_{j1} - 2a_{11}^J X_{11} X_{12} - 2a_{22}^J X_{21} X_{22} - 2a_{33}^J X_{31} X_{32}$$
$$- a_{12}^J (X_{11} X_{22} + X_{12} X_{21}) - a_{13}^J (X_{11} X_{32} + X_{12} X_{31})$$
$$- a_{23}^J (X_{21} X_{32} + X_{22} X_{31}) - \mu_j \tau_j^\gamma D_0^\gamma X_{j1} \quad (j = 1, 2, 3)\,. \qquad (3.101)$$

Substituting (3.71) and (3.98)–(3.100) with due account for (3.97) in Eqs. (3.101), we have

$$D_0^2 X_{j3} + \Omega_j^2 X_{j3} = -\left[2i\Omega_j D_2 A_j + \mu_j \tau_j^\gamma \left(i\Omega_j\right)^\gamma A_j\right]\exp(i\Omega_j T_0)$$
$$- \left(d_1^J A_1^2 \bar{A}_1 + d_2^J A_1 A_2 \bar{A}_2 + d_3^J A_1 A_3 \bar{A}_3\right)\exp(i\Omega_1 T_0)$$
$$- \left(d_4^J A_2^2 \bar{A}_2 + d_5^J A_2 A_1 \bar{A}_1 + d_6^J A_2 A_3 \bar{A}_3\right)\exp(i\Omega_2 T_0)$$
$$- \left(d_7^J A_3^2 \bar{A}_3 + d_8^J A_3 A_1 \bar{A}_1 + d_9^J A_3 A_2 \bar{A}_2\right)\exp(i\Omega_3 T_0)$$
$$- d_{10}^J A_1^3 \exp(3i\Omega_1 T_0) - d_{11}^J A_2^3 \exp(3i\Omega_2 T_0) - d_{12}^J A_3^3 \exp(3i\Omega_3 T_0)$$
$$- a_{12}^J (A_1 F_2 + A_2 F_1)\exp\left[i(\Omega_1 + \Omega_2)T_0\right]$$
$$- a_{12}^J \left(A_1 \bar{F}_2 + \bar{A}_2 F_1\right)\exp\left[i(\Omega_1 - \Omega_2)T_0\right]$$
$$- a_{13}^J (A_1 F_3 + A_3 F_1)\exp\left[i(\Omega_1 + \Omega_3)T_0\right]$$
$$- a_{13}^J \left(A_1 \bar{F}_3 + \bar{A}_3 F_1\right)\exp\left[i(\Omega_1 - \Omega_3)T_0\right]$$
$$- a_{23}^J (A_2 F_3 + A_3 F_2)\exp\left[i(\Omega_2 + \Omega_3)T_0\right]$$
$$- a_{23}^J \left(A_2 \bar{F}_3 + \bar{A}_3 F_2\right)\exp\left[i(\Omega_2 - \Omega_3)T_0\right]$$
$$- e_1^J A_1 A_2 A_3 \exp\left[i(\Omega_1 + \Omega_2 + \Omega_3)T_0\right] - e_2^J A_1 A_2 \bar{A}_3 \exp\left[i(\Omega_1 + \Omega_2 - \Omega_3)T_0\right]$$
$$- e_3^J A_1 \bar{A}_2 A_3 \exp\left[i(\Omega_1 - \Omega_2 + \Omega_3)T_0\right] - e_4^J \bar{A}_1 A_2 A_3 \exp\left[i(\Omega_2 - \Omega_1 + \Omega_3)T_0\right]$$
$$- c_1^J A_1 A_2^2 \exp\left[i(\Omega_1 + 2\Omega_2)T_0\right] - c_2^J A_1 A_3^2 \exp\left[i(\Omega_1 + 2\Omega_3)T_0\right]$$
$$- c_3^J A_1^2 A_2 \exp\left[i(2\Omega_1 + \Omega_2)T_0\right] - c_4^J A_1^2 A_3 \exp\left[i(2\Omega_1 + \Omega_3)T_0\right]$$
$$- c_5^J A_2 A_3^2 \exp\left[i(\Omega_2 + 2\Omega_3)T_0\right] - c_6^J A_2^2 A_3 \exp\left[i(2\Omega_2 + \Omega_3)T_0\right]$$
$$- c_7^J A_1^2 \bar{A}_2 \exp\left[i(2\Omega_1 - \Omega_2)T_0\right] - c_8^J A_1^2 \bar{A}_3 \exp\left[i(2\Omega_1 - \Omega_3)T_0\right]$$
$$- c_9^J \bar{A}_1 A_2^2 \exp\left[i(2\Omega_2 - \Omega_1)T_0\right] - c_{10}^J \bar{A}_1 A_3^2 \exp\left[i(2\Omega_3 - \Omega_1)T_0\right]$$

$$- c_{11}^J A_2^2 \bar{A}_3 \exp\left[i(2\Omega_2 - \Omega_3)T_0\right] - c_{12}^J A_3^2 \bar{A}_2 \exp\left[i(2\Omega_3 - \Omega_2)T_0\right]$$

$$- \sum_{k=1}^{3} 2a_{kk}^J A_k \left[F_k \exp(2i\,\Omega_k T_0) + \bar{F}_k\right] + cc\,, \tag{3.102}$$

where all coefficients d_i^J, c_i^J $(i = 1, 2, \ldots, 12)$, and e_k^J $(k = 1, 2, 3, 4)$ are presented in Appendix B.

Reference to Eqs. (3.102) shows that the following types of the internal resonance could occur on this step:

1. the one-to-one internal resonance, when one natural frequency is nearly equal to the other natural frequency,

$$\Omega_1 = \Omega_2 \quad (\Omega_3 \neq \Omega_1, \quad \Omega_3 \neq \Omega_2), \tag{3.103}$$

$$\Omega_1 = \Omega_3 \quad (\Omega_2 \neq \Omega_1, \quad \Omega_2 \neq \Omega_3), \tag{3.104}$$

$$\Omega_2 = \Omega_3 \quad (\Omega_1 \neq \Omega_2, \quad \Omega_1 \neq \Omega_3); \tag{3.105}$$

2. the one-to-one-to-one internal resonance

$$\Omega_1 = \Omega_2 = \Omega_3\,; \tag{3.106}$$

3. the three-to-one internal resonance, when one natural frequency is three times larger than the other natural frequency,

$$\Omega_1 = 3\Omega_2 \ \ (\Omega_3 \neq \Omega_1, \ \ \Omega_3 \neq 3\Omega_2), \ \ \text{or} \ \ \Omega_2 = 3\Omega_1, \tag{3.107}$$

$$\Omega_1 = 3\Omega_3 \ \ (\Omega_2 \neq \Omega_1, \ \ \Omega_2 \neq 3\Omega_3), \ \ \text{or} \ \ \Omega_3 = 3\Omega_1, \tag{3.108}$$

$$\Omega_2 = 3\Omega_3 \ \ (\Omega_1 \neq \Omega_2, \ \ \Omega_1 \neq 3\Omega_2), \ \ \text{or} \ \ \Omega_3 = 3\Omega_2\,; \tag{3.109}$$

4. the one-to-one-to-three or one-to-three-to-three internal resonance, i.e.,

$$\Omega_1 = 3\Omega_2 = \Omega_3, \ \ \text{or} \ \ 1:3:1, \tag{3.110}$$

$$\Omega_1 = \Omega_2 = 3\Omega_3, \ \ \text{or} \ \ 1:1:3, \tag{3.111}$$

$$3\Omega_1 = \Omega_2 = \Omega_3, \ \ \text{or} \ \ 3:1:1, \tag{3.112}$$

$$\Omega_1 = 3\Omega_2 = 3\Omega_3, \ \ \text{or} \ \ 1:3:3, \tag{3.113}$$

$$3\Omega_1 = \Omega_2 = 3\Omega_3, \ \ \text{or} \ \ 3:1:3, \tag{3.114}$$

$$3\Omega_1 = 3\Omega_2 = \Omega_3, \ \ \text{or} \ \ 3:3:1; \tag{3.115}$$

5. the combinational resonance of the additive-difference type of the second order, i.e.,

$$\Omega_1 = \Omega_2 + 2\Omega_3, \ \text{or} \ \Omega_2 = \Omega_1 - 2\Omega_3, \ \text{or} \ 2\Omega_3 = \Omega_1 - \Omega_2, \tag{3.116}$$

$$\Omega_1 = 2\Omega_2 + \Omega_3, \ \text{or} \ 2\Omega_2 = \Omega_1 - \Omega_3, \ \text{or} \ \Omega_3 = \Omega_1 - 2\Omega_2, \tag{3.117}$$

$$\Omega_1 = 2\Omega_3 - \Omega_2, \text{ or } \Omega_2 = 2\Omega_3 - \Omega_1, \text{ or } 2\Omega_3 = \Omega_1 + \Omega_2\,, \quad (3.118)$$

$$\Omega_1 = \Omega_2 - 2\Omega_3, \text{ or } \Omega_2 = 2\Omega_3 + \Omega_1, \text{ or } 2\Omega_3 = \Omega_2 - \Omega_1, \quad (3.119)$$

$$\Omega_1 = 2\Omega_2 - \Omega_3, \text{ or } 2\Omega_2 = \Omega_1 + \Omega_3, \text{ or } \Omega_3 = 2\Omega_2 - \Omega_1, \quad (3.120)$$

$$\Omega_1 = \Omega_3 - 2\Omega_2, \text{ or } 2\Omega_2 = \Omega_3 - \Omega_1, \text{ or } \Omega_3 = 2\Omega_2 + \Omega_1, \quad (3.121)$$

$$2\Omega_1 = \Omega_2 + \Omega_3, \text{ or } \Omega_2 = 2\Omega_1 - \Omega_3, \text{ or } \Omega_3 = 2\Omega_1 - \Omega_2, \quad (3.122)$$

$$2\Omega_1 = \Omega_2 - \Omega_3, \text{ or } \Omega_2 = 2\Omega_1 + \Omega_3, \text{ or } \Omega_3 = \Omega_2 - 2\Omega_1, \quad (3.123)$$

$$2\Omega_1 = \Omega_3 - \Omega_2, \text{ or } \Omega_2 = \Omega_3 - 2\Omega_1, \text{ or } \Omega_3 = 2\Omega_1 + \Omega_2\,. \quad (3.124)$$

The different cases of the three-to-one, one-to-one-to-three and one-to-three-to-three internal resonances (3.107)–(3.115) have been recently analyzed in detail in [34]. Further we will restrict ourselves by considering the cases of the combinational resonances characterized by coupling of three interacting modes of vibrations.

3.4 Governing Nonlinear Differential Equations Describing Amplitude-Phase Modulation for Different Types of the Combinational Internal Resonance

To deduce the nonlinear differential equations describing the modulation of amplitudes and phases of the cylindrical shell under consideration, we should consider separately each type of the combinational internal resonance which could occur with due account for weak damping of the order ε or ε^2.

3.4.1 Combinational Additive-Difference Internal Resonance of the First Order

Now let us consider the case of the combinational additive-difference internal resonance of the first order, i.e., when the viscosity has the order of ε. As an example, we will analyze the case (3.92), when $\Omega_1 = \Omega_2 + \Omega_3$. Other possible cases, (3.93) and (3.94), could be treated similarly.

Eliminating secular terms in Eqs. (3.80)–(3.82) for the case under consideration, we obtain the following solvability equations:

$$2i\,\Omega_1 D_1 A_1(T_1) + \mu_1\,(i\,\Omega_1\tau_1)^{\gamma}\,A_1 + a^{I}_{23}A_2A_3 = 0, \quad (3.125)$$

$$2i\,\Omega_2 D_1 A_2(T_1) + \mu_2\,(i\,\Omega_2\tau_2)^{\gamma}\,A_2 + a^{II}_{13}A_1\bar{A}_3 = 0, \quad (3.126)$$

$$2i\,\Omega_3 D_1 A_3(T_1) + \mu_3\,(i\,\Omega_3\tau_3)^{\gamma}\,A_3 + a^{III}_{12}A_1\bar{A}_2 = 0\,. \quad (3.127)$$

Let us multiply Eqs. (3.125)–(3.127), respectively, by $\bar{A}_1$, $\bar{A}_2$, and $\bar{A}_3$ and find their complex conjugates. Adding every pair of the mutually adjoint equations with each other and subtracting one from another, and considering that

$$A_i = a_{ie}^{i\varphi_i} \quad (i = 1, 2, 3), \tag{3.128}$$

as a result we have

$$\left(a_1^2\right)^{\cdot} + s_1 a_1^2 = \Omega_1^{-1} a_{23}^I a_1 a_2 a_3 \sin\delta, \tag{3.129}$$

$$\left(a_2^2\right)^{\cdot} + s_2 a_2^2 = -\Omega_2^{-1} a_{13}^{II} a_1 a_2 a_3 \sin\delta, \tag{3.130}$$

$$\left(a_3^2\right)^{\cdot} + s_3 a_3^2 = -\Omega_3^{-1} a_{12}^{III} a_1 a_2 a_3 \sin\delta, \tag{3.131}$$

$$\dot{\varphi}_1 - \frac{1}{2}\sigma_1 - \frac{1}{2}\frac{a_{23}^I}{\Omega_1}\frac{a_2 a_3}{a_1}\cos\delta = 0, \tag{3.132}$$

$$\dot{\varphi}_2 - \frac{1}{2}\sigma_2 - \frac{1}{2}\frac{a_{13}^{II}}{\Omega_2}\frac{a_1 a_3}{a_2}\cos\delta = 0, \tag{3.133}$$

$$\dot{\varphi}_3 - \frac{1}{2}\sigma_3 - \frac{1}{2}\frac{a_{12}^{III}}{\Omega_3}\frac{a_1 a_2}{a_3}\cos\delta = 0, \tag{3.134}$$

where the phase difference has the form $\delta = \varphi_1 - (\varphi_2 + \varphi_3)$.

Introducing new functions $\xi_1(T_1)$, $\xi_2(T_1)$, and $\xi_3(T_1)$, such that

$$a_1^2 = \frac{a_{23}^I}{\Omega_1}\xi_1 \exp(-s_1 T_1), \quad a_2^2 = \frac{a_{13}^{II}}{\Omega_2}\xi_2 \exp(-s_2 T_1), \quad a_3^2 = \frac{a_{12}^{III}}{\Omega_3}\xi_3 \exp(-s_3 T_1), \tag{3.135}$$

and substituting (3.135) in the left-hand sides of (3.129)–(3.131) yield

$$\dot{\xi}_1 \exp(-s_1 T_1) = a_1 a_2 a_3 \sin\delta, \tag{3.136}$$

$$\dot{\xi}_2 \exp(-s_2 T_1) = -a_1 a_2 a_3 \sin\delta, \tag{3.137}$$

$$\dot{\xi}_3 \exp(-s_3 T_1) = -a_1 a_2 a_3 \sin\delta, \tag{3.138}$$

the summation of which results in the following:

$$2\dot{\xi}_1 \exp(-s_1 T_1) + \dot{\xi}_2 \exp(-s_2 T_1) + \dot{\xi}_3 \exp(-s_3 T_1) = 0\,. \tag{3.139}$$

From Eqs. (3.136)–(3.138) we could find that

$$\frac{a_2 a_3}{a_1} = \frac{\dot{\xi}_1}{\xi_1}\frac{\Omega_1}{a_{23}^I}\frac{1}{\sin\delta}, \tag{3.140}$$

$$\frac{a_1 a_3}{a_2} = -\frac{\dot{\xi}_2}{\xi_2}\frac{\Omega_2}{a_{13}^{II}}\frac{1}{\sin\delta}, \tag{3.141}$$

$$\frac{a_1 a_2}{a_3} = -\frac{\dot{\xi}_3}{\xi_3}\frac{\Omega_3}{a_{12}^{III}}\frac{1}{\sin\delta}, \tag{3.142}$$

while Eqs. (3.132)–(3.134) could be reduced to

$$\dot{\delta} = \Sigma + \frac{1}{2}\frac{a_{23}^{I}}{\Omega_1}\frac{a_2 a_3}{a_1}\cos\delta - \frac{1}{2}\frac{a_{13}^{II}}{\Omega_2}\frac{a_1 a_3}{a_2}\cos\delta - \frac{1}{2}\frac{a_{12}^{III}}{\Omega_3}\frac{a_1 a_2}{a_3}\cos\delta, \tag{3.143}$$

where $\Sigma = \frac{1}{2}(\sigma_1 - \sigma_2 - \sigma_3)$.

Substituting (3.140)–(3.142) in (3.143) yields

$$\dot{\delta} - \Sigma = \frac{1}{2}\left(\frac{\dot{\xi}_1}{\xi_1} + \frac{\dot{\xi}_2}{\xi_2} + \frac{\dot{\xi}_3}{\xi_3}\right)\cot\delta\ . \tag{3.144}$$

The first integral of (3.144) could be written in the form

$$G_0 \exp\left(-\Sigma\int_0^{T_1}\tan\delta dT_1\right) = \sqrt{\xi_1}\sqrt{\xi_2}\sqrt{\xi_3}\cos\delta, \tag{3.145}$$

where G_0 is a constant of integration to be found from the initial conditions

$$\xi_1\Big|_{T_1=0} = \xi_{1\,0}, \quad \xi_2\Big|_{T_1=0} = \xi_{2\,0}, \quad \xi_3\Big|_{T_1=0} = \xi_{3\,0}, \quad \delta\Big|_{T_1=0} = \delta_0\ . \tag{3.146}$$

From the other hand, substituting (3.135) in (3.143) and (3.136) yields

$$\dot{\delta} = \Sigma + \frac{1}{2}b\left(\frac{\sqrt{\xi_2}\sqrt{\xi_3}}{\sqrt{\xi_1}}e^{\frac{1}{2}(s_1-s_2-s_3)T_1} - \frac{\sqrt{\xi_1}\sqrt{\xi_3}}{\sqrt{\xi_2}}e^{\frac{1}{2}(s_2-s_1-s_3)T_1}\right.$$
$$\left. - \frac{\sqrt{\xi_1}\sqrt{\xi_2}}{\sqrt{\xi_3}}e^{\frac{1}{2}(s_3-s_1-s_2)T_1}\right)\cos\delta, \tag{3.147}$$

$$\dot{\xi}_1 = b\sqrt{\xi_1}\sqrt{\xi_2}\sqrt{\xi_3}\,e^{\frac{1}{2}(s_1-s_2-s_3)T_1}\sin\delta, \tag{3.148}$$

where

$$b = \sqrt{\frac{a_{23}^{I}}{\Omega_1}\frac{a_{13}^{II}}{\Omega_2}\frac{a_{12}^{III}}{\Omega_3}}\ .$$

The nonlinear set of Eqs. (3.139), (3.145), (3.147), and (3.148) with the initial conditions (3.146) completely describe the vibrational process of the mechanical system being investigated under the condition of the internal combinational resonance of the first order and could be solved numerically.

3.4.1.1 Particular Case

In the particular case at $\Sigma = 0$ and $s_1 = s_2 = s_3 = s$, Eq. (3.139) has the form

$$2\dot{\xi}_1 + \dot{\xi}_2 + \dot{\xi}_3 = 0, \tag{3.149}$$

which could be integrated as

$$2\xi_1 + \xi_2 + \xi_3 = E_0, \tag{3.150}$$

resulting in the following law of energy dissipation:

$$E = 2\frac{\Omega_1}{a_{23}^{I}}a_1^2 + \frac{\Omega_2}{a_{13}^{II}}a_2^2 + \frac{\Omega_3}{a_{12}^{III}}a_3^2 = E_0 e^{-sT_1}, \tag{3.151}$$

where E_0 is the initial energy of the system.

Introducing a new variable ξ such that

$$E_0\dot{\xi} = b\sqrt{\xi_1\xi_2\xi_3}\exp(-\frac{1}{2}sT_1)\sin\delta, \tag{3.152}$$

Eqs. (3.136)–(3.138) could be reduced to

$$\dot{\xi}_1 = \dot{\xi}E_0, \tag{3.153}$$

$$\dot{\xi}_2 = -\dot{\xi}E_0, \tag{3.154}$$

$$\dot{\xi}_3 = -\dot{\xi}E_0, \tag{3.155}$$

the integration of which yields

$$\xi_1 = E_0(c_1 + \xi), \quad \xi_2 = E_0(c_2 - \xi), \quad \xi_3 = E_0(c_3 - \xi), \tag{3.156}$$

where c_i $(i = 1, 2, 3)$ are constants of integration.

Note that Eq. (3.149) is fulfilled automatically under the substitution of (3.153)–(3.155) in it, while the substitution of (3.156) in (3.150) results in the relationship between the constants of integration

$$2c_1 + c_2 + c_3 = 1\ . \tag{3.157}$$

Considering (3.156), Eqs. (3.145), (3.147), and (3.148) take, respectively, the form

$$G(\xi, \delta) = \sqrt{(c_1 + \xi)(c_2 - \xi)(c_3 - \xi)}\cos\delta = G_0(\xi_0, \delta_0)\ , \tag{3.158}$$

$$\dot{\delta} = \frac{1}{2}\, b\sqrt{E_0}\left[\frac{(c_2-\xi)(c_3-\xi)-(c_1+\xi)(c_3-\xi)-(c_1+\xi)(c_2-\xi)}{\sqrt{(c_1+\xi)(c_2-\xi)(c_3-\xi)}}\right] e^{-\frac{1}{2}sT_1}\cos\delta\,, \tag{3.159}$$

$$\dot{\xi} = b\sqrt{E_0}\sqrt{(c_1+\xi)(c_2-\xi)(c_3-\xi)}\; e^{-\frac{1}{2}sT_1}\sin\delta\,. \tag{3.160}$$

The second first integral (3.158) defines the stream function $G(\xi,\delta)$ such that

$$v_\xi = \dot{\xi} = -b\sqrt{E_0}\,\frac{\partial G}{\partial \delta}\; e^{-\frac{1}{2}sT_1}, \quad v_\delta = \dot{\delta} = b\sqrt{E_0}\,\frac{\partial G}{\partial \xi}\; e^{-\frac{1}{2}sT_1}\,, \tag{3.161}$$

which describes steady-state vibrations of an elastic shell attenuating with time.

Eliminating the variable δ from (3.158) and (3.160) and integrating over T_1, we have

$$\int_{\xi_0}^{\xi} \frac{d\xi}{\sqrt{(c_1+\xi)(c_2-\xi)(c_3-\xi)-G_0^2}} = \frac{2b\sqrt{E_0}}{s}\left(1-e^{-\frac{1}{2}sT_1}\right)\,. \tag{3.162}$$

The integral in the left-hand side of Eq. (3.162) can be transformed to an incomplete elliptic integral of the first kind and can be easily calculated using special tables [35]. So the solution of (3.162) allows one to find the value $\xi(T_1)$, and thus, to solve the problem under consideration.

3.4.2 *Combinational Additive-Difference Internal Resonance of the Second Order*

Now let us consider the case of the combinational additive-difference internal resonance of the second order, i.e., when the viscosity has the order of ε^2. As an example, we will analyze the case (3.116), when $\Omega_1 = \Omega_2 + 2\Omega_3$. Other possible cases, (3.117)–(3.124), could be treated similarly.

Eliminating secular terms in Eqs. (3.102), we obtain the following solvability equations:

$$\begin{aligned} 2i\,\Omega_1 D_2 A_1(T_2) &+ \mu_1\tau_1^\gamma A_1\,(i\,\Omega_1)^\gamma + d_1^I A_1^2\bar{A}_1 + d_2^I A_1 A_2\bar{A}_2 + d_3^I A_1 A_3\bar{A}_3 \\ &+ c_5^I A_2 A_3^2 = 0\,, \end{aligned} \tag{3.163}$$

$$\begin{aligned} 2i\,\Omega_2 D_2 A_2(T_2) &+ \mu_2\tau_2^\gamma A_2\,(i\,\Omega_2)^\gamma + d_4^{II} A_2^2\bar{A}_2 + d_5^{II} A_2 A_1\bar{A}_1 + d_6^{II} A_2 A_3\bar{A}_3 \\ &+ c_{10}^{II} A_1\bar{A}_3^2 = 0, \end{aligned} \tag{3.164}$$

$$\begin{aligned} 2i\,\Omega_3 D_2 A_3(T_2) &+ \mu_3\tau_3^\gamma A_3\,(i\,\Omega_3)^\gamma + d_7^{III} A_3^2\bar{A}_3 + d_8^{III} A_3 A_1\bar{A}_1 + d_9^{III} A_3 A_2\bar{A}_2 \\ &+ e_4^{III} A_1\bar{A}_2\bar{A}_3 = 0\,. \end{aligned} \tag{3.165}$$

Applying to (3.163)–(3.165) the same procedure as it was done above in the case of the combinational internal resonance of the first kind, we obtain a set of six nonlinear equations describing modulations in amplitudes and phases

$$\left(a_1^2\right)^{\cdot} + s_1 a_1^2 = -\Omega_1^{-1} c_5^I a_1 a_2 a_3^2 \sin\delta, \tag{3.166}$$

$$\left(a_2^2\right)^{\cdot} + s_2 a_2^2 = \Omega_2^{-1} c_{10}^{II} a_1 a_2 a_3^2 \sin\delta, \tag{3.167}$$

$$\left(a_3^2\right)^{\cdot} + s_3 a_3^2 = \Omega_3^{-1} e_4^{III} a_1 a_2 a_3^2 \sin\delta, \tag{3.168}$$

$$\dot{\varphi}_1 - \frac{1}{2}\sigma_1 - \frac{1}{2}\frac{d_1^I}{\Omega_1} a_1^2 - \frac{1}{2}\frac{d_2^I}{\Omega_1} a_2^2 - \frac{1}{2}\frac{d_3^I}{\Omega_1} a_3^2 - \frac{1}{2}\frac{c_5^I}{\Omega_1}\frac{a_2 a_3^2}{a_1}\cos\delta = 0, \tag{3.169}$$

$$\dot{\varphi}_2 - \frac{1}{2}\sigma_2 - \frac{1}{2}\frac{d_5^{II}}{\Omega_2} a_1^2 - \frac{1}{2}\frac{d_4^{II}}{\Omega_2} a_2^2 - \frac{1}{2}\frac{d_6^I}{\Omega_2} a_3^2 - \frac{1}{2}\frac{c_{10}^{II}}{\Omega_2}\frac{a_1 a_3^2}{a_2}\cos\delta = 0, \tag{3.170}$$

$$\dot{\varphi}_3 - \frac{1}{2}\sigma_3 - \frac{1}{2}\frac{d_8^{III}}{\Omega_3} a_1^2 - \frac{1}{2}\frac{d_9^{III}}{\Omega_3} a_2^2 - \frac{1}{2}\frac{d_7^{III}}{\Omega_3} a_3^2 - \frac{1}{2}\frac{e_4^{III}}{\Omega_3} a_1 a_2 \cos\delta = 0, \tag{3.171}$$

where the phase difference has the form $\delta = \varphi_2 + 2\varphi_3 - \varphi_1$, and an overdot denotes the differentiation with respect to T_2.

Introducing new functions $\xi_1(T_2)$, $\xi_2(T_2)$, and $\xi_3(T_2)$, such that

$$a_1^2 = \frac{c_5^I}{\Omega_1}\xi_1 \exp(-s_1 T_2), \quad a_2^2 = \frac{c_{10}^{II}}{\Omega_2}\xi_2 \exp(-s_2 T_2), \quad a_3^2 = \frac{e_4^{III}}{\Omega_3}\xi_3 \exp(-s_3 T_2), \tag{3.172}$$

and substituting (3.172) in (3.166)–(3.168) yield

$$\dot{\xi}_1 e^{-s_1 T_2} = -a_1 a_2 a_3^2 \sin\delta, \tag{3.173}$$

$$\dot{\xi}_2 e^{-s_2 T_2} = a_1 a_2 a_3^2 \sin\delta, \tag{3.174}$$

$$\dot{\xi}_3 e^{-s_3 T_2} = a_1 a_2 a_3^2 \sin\delta, \tag{3.175}$$

the summation of which gives

$$2\dot{\xi}_1 e^{-s_1 T_2} + \dot{\xi}_2 e^{-s_2 T_2} + \dot{\xi}_3 e^{-s_3 T_2} = 0\,. \tag{3.176}$$

Equations (3.173)–(3.175) could be rewritten as

$$\frac{a_2 a_3^2}{a_1} = -\frac{\dot{\xi}_1 e^{-s_1 T_2}}{a_1^2}\frac{1}{\sin\delta} = -\frac{\Omega_1}{c_5^I}\frac{\dot{\xi}_1}{\xi_1}\frac{1}{\sin\delta}, \tag{3.177}$$

$$\frac{a_1 a_3^2}{a_2} = \frac{\dot{\xi}_2 e^{-s_2 T_2}}{a_2^2} \frac{1}{\sin\delta} = \frac{\Omega_2}{c_{10}^{II}} \frac{\dot{\xi}_2}{\xi_2} \frac{1}{\sin\delta}, \tag{3.178}$$

$$a_1 a_2 = \frac{\dot{\xi}_3 e^{-s_3 T_2}}{a_3^2} \frac{1}{\sin\delta} = \frac{\Omega_3}{e_4^{III}} \frac{\dot{\xi}_3}{\xi_3} \frac{1}{\sin\delta}. \tag{3.179}$$

Equations (3.169)–(3.171) could be reduced to one following equation:

$$\begin{aligned}\dot{\delta} = \Sigma &+ \left(\frac{1}{2}\frac{d_5^{II}}{\Omega_2} + \frac{d_8^{III}}{\Omega_3} - \frac{1}{2}\frac{d_1^{I}}{\Omega_1}\right) a_1^2 + \left(\frac{1}{2}\frac{d_4^{II}}{\Omega_2} + \frac{d_9^{III}}{\Omega_3} - \frac{1}{2}\frac{d_2^{I}}{\Omega_1}\right) a_2^2 \\ &+ \left(\frac{1}{2}\frac{d_6^{II}}{\Omega_2} + \frac{d_7^{III}}{\Omega_3} - \frac{1}{2}\frac{d_3^{I}}{\Omega_1}\right) a_3^2 \\ &+ \left(\frac{1}{2}\frac{c_{10}^{II}}{\Omega_2}\frac{a_1 a_3^2}{a_2} + \frac{e_{10}^{III}}{\Omega_3} a_1 a_2 - \frac{1}{2}\frac{c_5^{I}}{\Omega_1}\frac{a_2 a_3^2}{a_1}\right)\cos\delta,\end{aligned} \tag{3.180}$$

where $\Sigma = \frac{1}{2}\sigma_2 + \sigma_3 - \frac{1}{2}\sigma_1$.

Considering (3.172) and (3.177)–(3.179), Eq. (3.180) could be reduced to

$$\dot{\delta} - \Sigma = \kappa_4 \xi_1 e^{-s_1 T_2} + \kappa_5 \xi_2 e^{-s_2 T_2} + \kappa_6 e^{-s_3 T_2} + \left(\frac{1}{2}\frac{\dot{\xi}_1}{\xi_1} + \frac{1}{2}\frac{\dot{\xi}_2}{\xi_2} + \frac{\dot{\xi}_3}{\xi_3}\right)\cot\delta, \tag{3.181}$$

where

$$\kappa_4 = \left(\frac{1}{2}\frac{d_5^{II}}{\Omega_2} + \frac{d_8^{III}}{\Omega_3} - \frac{1}{2}\frac{d_1^{I}}{\Omega_1}\right)\frac{c_5^{I}}{\Omega_1},$$

$$\kappa_5 = \left(\frac{1}{2}\frac{d_4^{II}}{\Omega_2} + \frac{d_9^{III}}{\Omega_3} - \frac{1}{2}\frac{d_2^{I}}{\Omega_1}\right)\frac{c_{10}^{II}}{\Omega_2},$$

$$\kappa_6 = \left(\frac{1}{2}\frac{d_6^{II}}{\Omega_2} + \frac{d_7^{III}}{\Omega_3} - \frac{1}{2}\frac{d_3^{I}}{\Omega_1}\right)\frac{e_4^{II}}{\Omega_3}.$$

It should be noted that the equation defining the stream function could be obtained from (3.181), if we neglect the first three terms in (3.181) which decay rather rapidly as compared with its last term.

Equations (3.166) and (3.180) could be rewritten in another form if we substitute (3.172) in all its terms

$$\dot{\xi}_1 = -b\xi_1^{1/2}\xi_2^{1/2}\xi_3 e^{(1/2\,s_1 - 1/2\,s_2 - s_3)T_2}\sin\delta, \tag{3.182}$$

$$\dot{\delta} - \Sigma = \kappa_4 \xi_1 e^{-s_1 T_2} + \kappa_5 \xi_2 e^{-s_2 T_2} + \kappa_6 \xi_3 e^{-s_3 T_2}$$
$$+ b \left[\xi_1^{1/2} \xi_2^{1/2} e^{-\frac{1}{2}(s_2+s_1)T_2} - \frac{1}{2} \xi_1^{-1/2} \xi_2^{1/2} \xi_3 e^{(1/2\, s_1 - 1/2\, s_2 - s_3)T_2} \right.$$
$$\left. + \frac{1}{2} \xi_1^{1/2} \xi_2^{-1/2} \xi_3 e^{(-1/2\, s_1 + 1/2\, s_2 - s_3)T_2} \right] \cos\delta \,, \tag{3.183}$$

where

$$b = \frac{e_4^{II}}{\Omega_3} \left(\frac{c_5^I}{\Omega_1} \right)^{1/2} \left(\frac{c_{10}^{II}}{\Omega_2} \right)^{1/2} .$$

The nonlinear set of Eqs. (3.176) and (3.181)–(3.183) with the initial conditions (3.146) completely describe the vibrational process of the mechanical system being investigated under the condition of the combinational internal resonance of the second order and could be solved numerically.

3.4.2.1 Particular Case

In the particular case at $\Sigma = 0$ and $s_1 = s_2 = s_3 = s$, Eq. (3.176) has the form

$$2\dot{\xi}_1 + \dot{\xi}_2 + \dot{\xi}_3 = 0 \,, \tag{3.184}$$

whence it follows that

$$2\xi_1 + \xi_2 + \xi_3 = E_0 \,, \tag{3.185}$$

and the energy of the system decays according the following law:

$$E = E_0 e^{-sT_2} \,. \tag{3.186}$$

Introducing a new variable ξ such that

$$\dot{\xi} E_0 = b \sqrt{\xi_1 \xi_2} \xi_3 \exp(-sT_2) \sin\delta, \tag{3.187}$$

Eqs. (3.173)–(3.175) could be reduced to

$$\dot{\xi}_1 = -\dot{\xi} E_0, \tag{3.188}$$

$$\dot{\xi}_2 = \dot{\xi} E_0, \tag{3.189}$$

$$\dot{\xi}_3 = \dot{\xi} E_0, \tag{3.190}$$

the integration of which yields

$$\xi_1 = E_0(c_1 - \xi)\,, \quad \xi_2 = E_0(c_2 + \xi)\,, \quad \xi_3 = E_0(c_3 + \xi)\,, \tag{3.191}$$

where c_i $(i = 1, 2, 3)$ are constants of integration.

Note that Eq. (3.184) is fulfilled automatically under the substitution of (3.188)–(3.190) in it, while the substitution of (3.191) in (3.185) results in the relationship between the constants of integration

$$2c_1 + c_2 + c_3 = 1\,. \tag{3.192}$$

Considering (3.191), Eqs. (3.182) and (3.183), wherein three terms rapidly decaying with time have been neglected, are reduced to the following:

$$\dot{\xi} = bE_0\sqrt{(c_1 - \xi)(c_2 + \xi)}(c_3 + \xi)e^{-sT_2}\sin\delta\,, \tag{3.193}$$

$$\dot{\delta} = bE_0e^{-sT_2}\cos\delta\left[\frac{(c_1-\xi)(c_3+\xi)+2(c_1-\xi)(c_2+\xi)-(c_2+\xi)(c_3+\xi)}{2\sqrt{(c_1 - \xi)(c_2 + \xi)}}\right], \tag{3.194}$$

while Eq. (3.181), wherein three terms rapidly decaying with time have been neglected, takes the form

$$\dot{\delta}\tan\delta \approx \frac{1}{2}\frac{\dot{\xi}_1}{\xi_1} + \frac{1}{2}\frac{\dot{\xi}_2}{\xi_2} + \frac{\dot{\xi}_3}{\xi_3}\,, \tag{3.195}$$

and it could be integrated, resulting in its first integral

$$G(\xi, \delta) = \sqrt{(c_1 - \xi)(c_2 + \xi)}(c_3 + \xi)\cos\delta = G_0(\xi_0, \delta_0)\,. \tag{3.196}$$

The second first integral (3.196) defines the stream function $G(\xi, \delta)$ such that

$$v_\xi = \dot{\xi} = -bE_0\frac{\partial G}{\partial\delta}e^{-sT_2}, \quad v_\delta = \dot{\delta} = bE_0\frac{\partial G}{\partial\xi}e^{-sT_2}\,, \tag{3.197}$$

which describes steady-state vibrations of an elastic shell decaying with time.

Eliminating the variable δ from (3.193) and (3.196) and integrating over T_2, we have

$$\int_{\xi_0}^{\xi}\frac{d\xi}{\sqrt{(c_1 - \xi)(c_2 + \xi)(c_3 + \xi)^2 - G_0^2}} = \frac{bE_0}{s}\left(1 - e^{-sT_2}\right)\,. \tag{3.198}$$

The solution of (3.198) allows one to find the value $\xi(T_2)$, and thus, to solve the problem under consideration.

If we represent $\dot{\delta}$ considering (3.193) as

$$\dot{\delta} = \frac{d\delta}{d\xi}\dot{\xi} = \frac{d\delta}{d\xi} bE_0 \sqrt{(c_1-\xi)(c_2+\xi)}(c_3+\xi)\, e^{-sT_2} \sin\delta\,, \tag{3.199}$$

and retain all terms in (3.183) with due account for (3.191), then (3.183) could be rewritten as follows

$$\frac{d\cos\delta}{d\xi} + \frac{1}{2}\left(\frac{1}{c_2+\xi} + \frac{2}{c_3+\xi} - \frac{1}{c_1-\xi}\right)\cos\delta = -\frac{1}{b}\left[\kappa_4 \frac{\sqrt{c_1-\xi}}{\sqrt{c_2+\xi}(c_3+\xi)} + \kappa_5 \frac{\sqrt{c_2+\xi}}{\sqrt{c_1-\xi}(c_3+\xi)} + \kappa_6 \frac{1}{\sqrt{(c_1-\xi)(c_2+\xi)}}\right]. \tag{3.200}$$

Integrating (3.200) yields

$$G(\xi,\delta) = \sqrt{(c_1-\xi)(c_2+\xi)}(c_3+\xi)\cos\delta + \frac{1}{2b}\Big[-\kappa_4(c_1-\xi)^2 + \kappa_5(c_2+\xi)^2 + \kappa_6(c_3+\xi)^2\Big] = G_0(\xi_0,\delta_0)\,. \tag{3.201}$$

Note that relationships (3.197) are valid in this case as well, while the integral (3.198) takes the form

$$\int_{\xi_0}^{\xi} \frac{d\xi}{\sqrt{(c_1-\xi)(c_2+\xi)(c_3+\xi)^2 - C^2(\xi)}} = \frac{bE_0}{s}\left(1-e^{-sT_2}\right), \tag{3.202}$$

where

$$C(\xi) = G_0 - \frac{1}{2b}\left[-\kappa_4(c_1-\xi)^2 + \kappa_5(c_2+\xi)^2 + \kappa_6(c_3+\xi)^2\right].$$

3.5 Numerical Investigations

As examples, let us carry out the qualitative analysis of the cases of combinational internal resonances of the first order (3.92) and the second order (3.116).

3.5.1 Combinational Internal Resonance of the First Order $\Omega_1 = \Omega_2 + \Omega_3$ (3.92)

For this case, the stream-function $G(\xi, \delta)$ is defined by relationship (3.158), and the phase portrait to be constructed according to (3.158) depends essentially on the magnitudes of the coefficients c_i $(i = 1, 2, 3)$. Let us carry out the phenomenological analysis of the phase portraits constructing them at different magnitudes of the system parameters.

3.5.1.1 The Case (3.92) at $c_1 = \frac{1}{2}$ and $c_2 = c_3 = 0$

Let us first consider the case (3.92) when $c_1 = \frac{1}{2}$ and $c_2 = c_3 = 0$. The stream-lines of the phase fluid in the phase plane $\xi - \delta$ for this particular case are presented in Fig. 3.2.

In this case, the stream-function is defined as

$$G(\xi, \delta) = \xi \sqrt{\frac{1}{2} + \xi}\ \cos\delta = G(\xi_0, \delta_0),$$

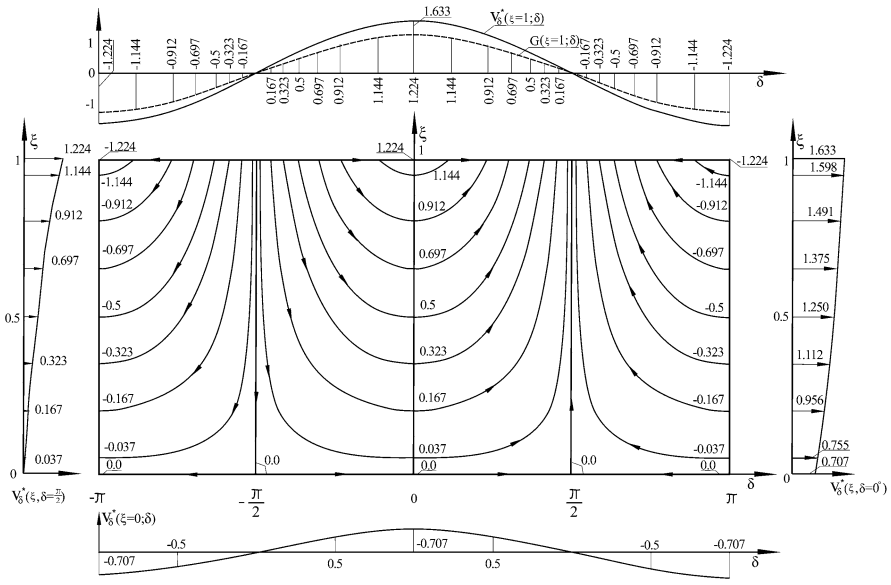


Fig. 3.2 Phase portrait for the case of the combinational resonance of the first type: $c_1 = \frac{1}{2}$, $c_2 = c_3 = 0$

and its magnitudes are indicated by digits near the curves which correspond to the stream-lines; the flow direction of the phase fluid elements are shown by arrows on the stream-lines.

Reference to Fig. 3.2 shows that the phase fluid flows within the circulation zones, which tend to be located around the perimeter of the rectangles bounded by the lines $\xi = 0$, $\xi = 1$, and $\delta = \pm(\pi/2) \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$. As this takes place, the flow in each such rectangle becomes isolated. On three sides of the rectangle, namely: $\xi = 0$ and $\delta = \pm(\pi/2) \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$, $G = 0$ and inside each rectangle the value G preserves its sign. Along the side $\xi = 1$ the stream function G changes periodically attaining its extreme magnitudes at the points with the coordinates $\xi = 1, \delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ (Fig. 3.2).

All stream-lines inside the rectangle are nonclosed, in so doing their initial and terminal points locate on the line $\xi = 1$. The distribution of the velocities of the phase fluid points is shown in Fig. 3.2 along the lines $\xi = 0$, $\xi = 1$, $\delta = 0$, and $\delta = \frac{\pi}{2}$, wherein $v^*_{\delta \text{ or } \xi} = \frac{v_{\delta \text{ or } \xi}}{b\sqrt{E_0}}$.

Along the lines $\delta = \pm(\pi/2) \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$ in the presence of conventional viscosity the solution could be written for the amplitude modulated regimes decaying with time

$$\ln \frac{\sqrt{1+2\xi}-1}{\sqrt{1+2\xi}+1}\Big|_{\xi_0}^{\xi} = \pm\frac{\sqrt{2}}{s}\, b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right),$$
$$\delta(T_1) = \delta_0 = \pm\frac{\pi}{2} \pm 2\pi n, \quad n = 0, 1, 2, \ldots \tag{3.203}$$

Along the lines $\xi = 0$ and $\xi = 1$ the phase modulated regimes decaying with time are realized

$$\xi(T_1) = \xi_0 = 0$$
$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = \frac{\sqrt{2}}{s}\, b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right), \tag{3.204}$$

and

$$\xi(T_1) = \xi_0 = 1$$
$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = \frac{4\sqrt{6}}{3s}\, b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right). \tag{3.205}$$

3.5.1.2 The Case (3.92) at $c_1 = c_2 = c_3 = \frac{1}{4}$

In this case, the stream-function is defined as

$$G(\xi, \delta) = \left(\frac{1}{4} - \xi\right)\sqrt{\frac{1}{4} + \xi}\,\cos\delta = G(\xi_0, \delta_0),$$

and Fig. 3.3 shows the streamlines of the phase fluid in the phase plane.

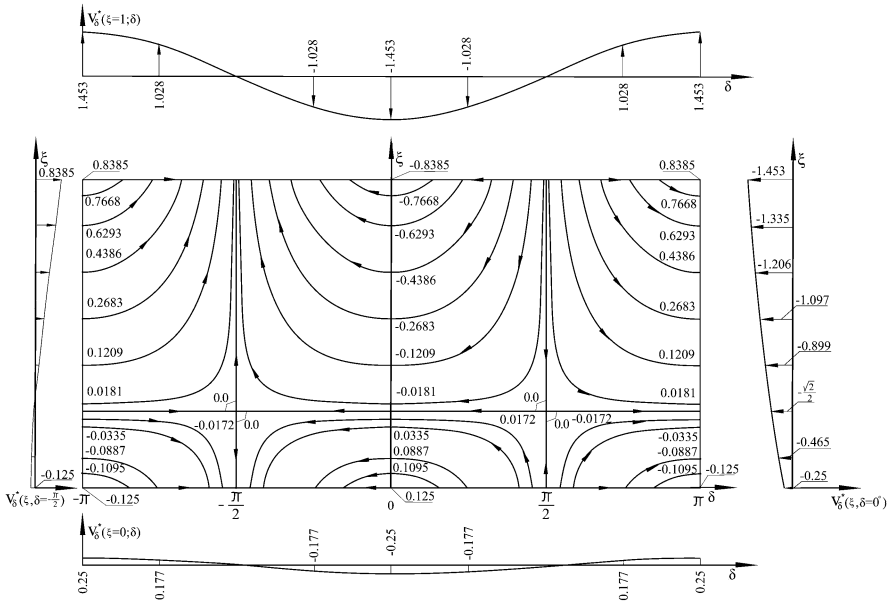


Fig. 3.3 Phase portrait for the case of the combinational resonance of the first type: $c_1 = c_2 = c_3 = \frac{1}{4}$

Reference to Fig. 3.3 shows that the infinite channel ($-\infty < \delta < \infty$) bounded by the lines $\xi = 0$ and $\xi = 1$ is divided into a set of rectangles by the lines $\xi = \frac{1}{4}$ and $\delta = \pm\frac{\pi}{2} \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$. Within each rectangle, the value of G preserves its sign and all stream-lines are nonclosed with initial and terminal points locating on the boundary lines $\xi = 0$ and $\xi = 1$, along which the stream-function changes periodically attaining its extreme magnitudes at the points with the coordinates $\xi = 1$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ and $\xi = 0$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ (Fig. 3.3).

Along all boundaries of rectangles the analytical solutions could be found. Thus, three phase modulated regimes are the following:

$$\xi(T_1) = \xi_0 = 0$$
$$\ln\tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = -\frac{1}{2s}\, b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right), \tag{3.206}$$

$$\xi(T_1) = \xi_0 = \frac{1}{4}$$
$$\ln\tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = -\frac{\sqrt{2}}{s}\, b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right), \tag{3.207}$$

and

$$\xi(T_1) = \xi_0 = 1$$

$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = \frac{13\sqrt{5}}{10s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right). \tag{3.208}$$

Along the lines $\delta = \pm(\pi/2) \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$ the amplitude modulated regimes decaying with time are realized

$$\ln \frac{\sqrt{1+4\xi} - \sqrt{2}}{\sqrt{1+4\xi} + \sqrt{2}}\Big|_{\xi_0}^{\xi} = \pm\frac{\sqrt{2}}{s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right),$$

$$\delta(T_1) = \delta_0 = \pm\frac{\pi}{2} \pm 2\pi n, \quad n = 0, 1, 2, \ldots \tag{3.209}$$

3.5.1.3 The Case (3.92) at $c_1 = 0$, and $c_2 = c_3 = \frac{1}{2}$

In this case, the stream-function is defined as

$$G(\xi, \delta) = \left(\frac{1}{2} - \xi\right)\sqrt{\xi}\,\cos\delta = G(\xi_0, \delta_0),$$

and Fig. 3.4 shows the streamlines of the phase fluid in the phase plane.

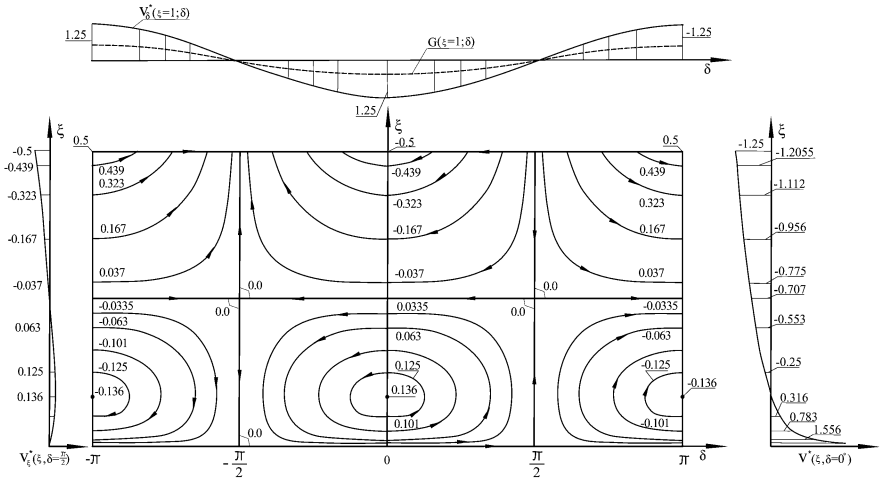


Fig. 3.4 Phase portrait for the case of the combinational resonance of the first type: $c_1 = 0$, $c_2 = c_3 = \frac{1}{2}$

Reference to Fig. 3.4 shows that the infinite channel ($-\infty < \delta < \infty$) bounded by the lines $\xi = 0$ and $\xi = 1$ is divided into a set of equal rectangles by the lines $\xi = \frac{1}{2}$ and $\delta = \pm\frac{\pi}{2} \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$. Within each rectangle, the value of G preserves its sign. As this takes place, in the upper rectangles all stream-lines are nonclosed with initial and terminal points locating on the boundary line $\xi = 1$, along which the stream-function changes periodically attaining its extreme magnitudes at the points with the coordinates $\xi = 1$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ (Fig. 3.3), while along the line $\xi = 0$ the magnitude of the stream function is constant and equal to $G = 0$, and in the bottom rectangles all stream-lines are closed. The function G attains its extreme magnitudes at the points with the coordinates $\xi = 1$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ and $\xi = \frac{1}{6}$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ within the upper and bottom rectangles, respectively, in so doing the points with the coordinates $\xi = \xi_0 = \frac{1}{6}$, $\delta = \delta_0 = \pm\pi n$ are centers corresponding to stationary motions.

The distribution of the velocity along the boundary line is shown in Fig. 3.4 as well, whence it follows that the velocity distribution along the vertical lines $\delta = \pm\pi n/2$ $(n = 0, 1, 2, \ldots)$ has the aperiodic character, while in the vicinity of the line $\xi = 1$ it possesses the periodic character. The transition of fluid elements from the points $\xi = 0$, $\delta = -\pi/2 \pm 2\pi n$ to the points $\xi = 0$, $\delta = \pi/2 \pm 2\pi n$ $(n = 0, 1, 2 \ldots)$ proceeds instantly, because according to the distribution of the phase velocity along the section $\delta = 0$ (see Fig. 3.4) the magnitude of $\mathbf{v}$ tends to infinity as $\xi \to 0$.

Along the boundaries of rectangles the analytical solutions could be found. Thus, two phase modulated regimes are the following:

$$\xi(T_1) = \xi_0 = 1$$
$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = -\frac{5}{2s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right), \tag{3.210}$$

and

$$\xi(T_1) = \xi_0 = \frac{1}{2}$$
$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = -\frac{\sqrt{2}}{s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right). \tag{3.211}$$

Along the lines $\delta = \pm(\pi/2) \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$ the amplitude modulated regimes decaying with time are realized

$$\ln \frac{\sqrt{2\xi} - 1}{\sqrt{2\xi} + 1}\Big|_{\xi_0}^{\xi} = \pm\frac{1}{s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right),$$
$$\delta(T_1) = \delta_0 = \pm\frac{\pi}{2} \pm 2\pi n, \quad n = 0, 1, 2, \ldots \tag{3.212}$$

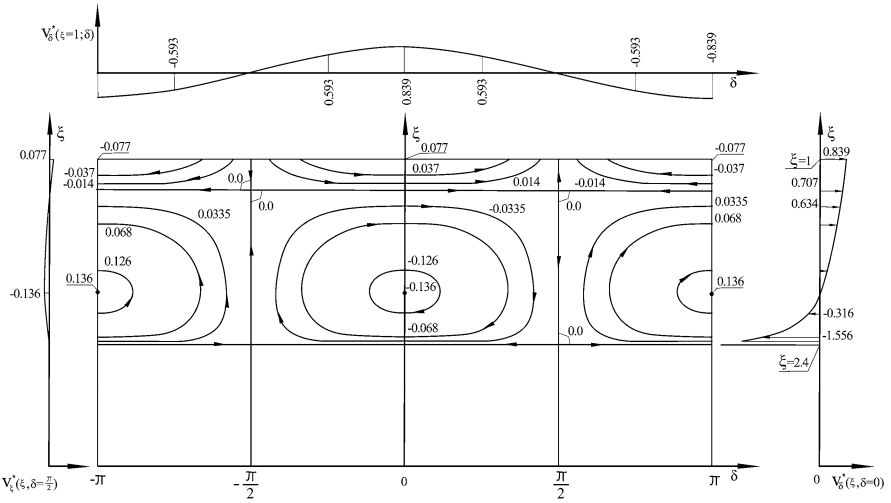


Fig. 3.5 Phase portrait for the case of the combinational resonance of the first type: $c_1 = -0.4$, $c_2 = c_3 = 0.9$

3.5.1.4 The Case (3.92) at $c_1 = -0.4$, and $c_2 = c_3 = 0.9$

In this case, the stream-function is defined as

$$G(\xi, \delta) = (0.9 - \xi)\sqrt{\xi - 0.4}\,\cos\delta = G(\xi_0, \delta_0),$$

and Fig. 3.5 shows the streamlines of the phase fluid in the phase plane.

Reference to Fig. 3.5 shows that, as distinct to the previous cases, the infinite channel ($-\infty < \delta < \infty$) is narrowed and is bounded by the lines $\xi = 0.4$ and $\xi = 1$. It is divided into a set of rectangles by the lines $\xi = 0.9$ and $\delta = \pm\frac{\pi}{2} \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$. Within each rectangle, the value of G preserves its sign. As this takes place, in the upper rectangles all stream-lines are nonclosed with initial and terminal points locating on the boundary line $\xi = 1$, along which the stream-function changes periodically attaining its extreme magnitudes at the points with the coordinates $\xi = 1$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ (Fig. 3.3), while along the line $\xi = 0.4$ the magnitude of the stream function is constant and equal to $G = 0$, and in the bottom rectangles all stream-lines are closed. The function G attains its extreme magnitudes at the points with the coordinates $\xi = 1$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ and $\xi = 0.567$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ within the upper and bottom rectangles, respectively, in so doing the points with the coordinates $\xi = \xi_0 = 0.567$, $\delta = \delta_0 = \pm\pi n$ are centers corresponding to stationary motions.

The distribution of the velocity along the boundary line is shown in Fig. 3.5 as well, whence it follows that the velocity distribution along the vertical lines $\delta = \pm\pi n/2$ $(n = 0, 1, 2, \ldots)$ has the aperiodic character, while in the vicinity of the line $\xi = 1$ it possesses the periodic character. The transition of fluid elements from

the points $\xi = 0, \quad \delta = \pi/2 \pm 2\pi n$ to the points $\xi = 0, \quad \delta = -\pi/2 \pm 2\pi n$ $(n = 0, 1, 2 \ldots)$ proceeds instantly, because according to the distribution of the phase velocity along the section $\delta = 0$ (see Fig. 3.5) the magnitude of $\mathbf{v}$ tends to infinity as $\xi \to 0$.

Along the boundaries of rectangles the analytical solutions could be found. Thus, two phase modulated regimes are the following:

$$\xi(T_1) = \xi_0 = 1$$
$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = -\frac{11\sqrt{15}}{15s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right), \tag{3.213}$$

and

$$\xi(T_1) = \xi_0 = 0.9$$
$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = -\frac{19\sqrt{2}}{10s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right). \tag{3.214}$$

Along the lines $\delta = \pm(\pi/2) \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$ the amplitude modulated regimes decaying with time are realized

$$\ln \frac{\sqrt{2\xi - 0.8} - 1}{\sqrt{2\xi - 0.8} + 1}\Big|_{\xi_0}^{\xi} = \pm\frac{\sqrt{2}}{s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right),$$
$$\delta(T_1) = \delta_0 = \pm\frac{\pi}{2} \pm 2\pi n, \quad n = 0, 1, 2, \ldots \tag{3.215}$$

3.5.1.5 The Case (3.92) at $c_1 = 0.1$, $c_2 = 0.3$, and $c_3 = 0.5$

In this case, the stream-function is defined as

$$G(\xi, \delta) = \sqrt{(0.1 + \xi)(0.3 - \xi)(0.5 - \xi)}\ \cos\delta = G(\xi_0, \delta_0),$$

and Fig. 3.6 shows the streamlines of the phase fluid in the phase plane.

Reference to Fig. 3.6 shows that in this case the phase fluid flows in two separate channels bounded by the lines $\xi = 0$, $\xi = 0.3$ and $\xi = 0.5$, $\xi = 1$. The distribution of the velocity along the boundary lines reveals the fact that in the vicinity of the external lines $\xi = 0$ and $\xi = 1$ it possesses the periodic character. However, in the vicinity of the internal boundaries $\xi = 0.3$ and $\xi = 0.5$ the magnitude of $\mathbf{v}$ tends to infinity, that is why the transition of fluid elements from the points $\xi = 0.5$, $\delta = -\pi/2 \pm 2\pi n$ to the points $\xi = 0.5, \quad \delta = \pi/2 \pm 2\pi n$ $(n = 0, 1, 2 \ldots)$ and from the points $\xi = 0.3, \quad \delta = \pi/2 \pm 2\pi n$ to the points $\xi = 0.3, \quad \delta = -\pi/2 \pm 2\pi n$ $(n = 0, 1, 2 \ldots)$ proceeds instantly.

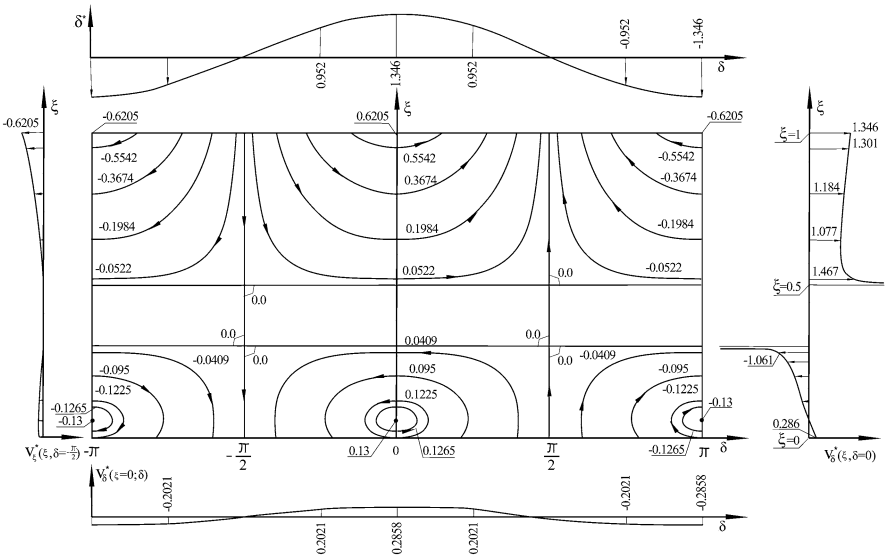


Fig. 3.6 Phase portrait for the case of the combinational resonance of the first type: $c_1 = 0.1$, $c_2 = 0.3$, and $c_3 = 0.5$

Within each rectangle, the value of G preserves its sign. The function G attains its extreme magnitudes at the points with the coordinates $\xi = 1$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ and $\xi = 0.0567$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ within the upper and bottom rectangles, respectively, in so doing the points with the coordinates $\xi = \xi_0 = 0.0567$, $\delta = \delta_0 = \pm\pi n$ are centers corresponding to stationary motions.

As this takes place, in the upper rectangles all stream-lines are nonclosed with initial and terminal points locating on the upper external boundary line $\xi = 1$, along which the stream-function changes periodically attaining its extreme magnitudes at the points with the coordinates $\xi = 1$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ (Fig. 3.6), while along the lines $\xi = 0.5$ and $\xi = 0.3$ the magnitude of the stream function is constant and equal to $G = 0$. In the bottom rectangles there are closed stream-lines surrounded the center-like points, as well as nonclosed stream lines with initial and terminal points locating on the bottom external boundary line $\xi = 0$, along which the stream-function changes periodically attaining its extreme magnitudes at the points with the coordinates $\xi = 0$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$.

Along the external boundaries of rectangles the analytical solutions could be found corresponding to two phase modulated regimes:

$$\xi(T_1) = \xi_0 = 0$$
$$\ln\tan\left(\frac{1}{2}\,\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = \frac{0.5715}{s}\,b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}\,sT_1}\right), \qquad (3.216)$$

and

$$\xi(T_1) = \xi_0 = 1$$
$$\ln \tan\left(\frac{1}{2}\delta + \frac{\pi}{4}\right)\Big|_{\delta_0}^{\delta} = \frac{2.6914}{s} b\sqrt{E_0}\left(1 - e^{-\frac{1}{2}sT_1}\right). \tag{3.217}$$

3.5.2 *Combinational Internal Resonance of the Second Order $\Omega_1 = \Omega_2 + 2\Omega_3$ (3.116)*

For this case, the stream-function $G(\xi, \delta)$ is defined by relationship (3.196), and the phase portrait to be constructed according to (3.196) depends essentially on the magnitudes of the coefficients c_i $(i = 1, 2, 3)$.

3.5.2.1 The Case (3.92) at $c_1 = \frac{1}{2}$, $c_2 = c_3 = 0$

In this case, the stream-function $G(\xi, \delta)$ takes the form

$$G(\xi, \delta) = \xi\sqrt{\xi(\frac{1}{2} - \xi)}\cos\delta = G(\xi_0, \delta_0),$$

and Fig. 3.7 shows the stream lines of the phase fluid in the phase plane.

Reference to Fig. 3.7 shows that phase fluid flows within the infinite channel $(-\infty < \delta < \infty)$ bounded by the bottom boundary line $\xi = 0$ and the upper boundary line $\xi = 0.5$. This channel is divided by vertical lines $\delta = \pm\frac{\pi}{2} \pm 2\pi n$ $(n = 0, 1, 2, \ldots)$ into rectangles, along all side of which the stream function G is equal to zero.

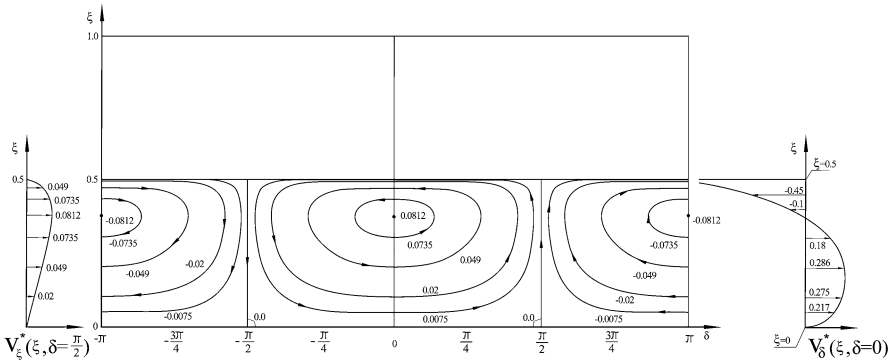


Fig. 3.7 Phase portrait for the case of the combinational resonance of the second type: $c_1 = \frac{1}{2}$, $c_2 = c_3 = 0$

Within each rectangle, the value of G preserves its sign. The function G attains its extreme magnitudes at the center-like points with the coordinates $\xi = \frac{3}{8}$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$. The stream lines with $0.0075 < G < G_{\max} = 0.0812$ are closed corresponding to periodic change of amplitudes and phase difference. The stream line with $G = 0.0075$ is the first nonclosed line with the initial point $\xi = 0.5$, $\delta = -\frac{\pi}{4}$ and the terminal point $\xi = 0.5$, $\delta = \frac{\pi}{4}$. All stream lines with $0 < G < 0.0075$ are nonclosed with the initial and terminal points located on the line $\xi = 0.5$.

The distribution of velocities along the lines $\delta = \pm 2\pi n$ and $\delta = \frac{\pi}{2} \pm 2\pi n$ show that along the bottom boundary line $\xi = 0$ the velocity is equal to zero, while along the upper boundary line $\xi = 0.5$ within the segments $\frac{\pi}{4} \pm 2\pi n < \delta < \frac{\pi}{2} \pm 2\pi n$ and $\frac{\pi}{2} \pm 2\pi n < \delta < \frac{3\pi}{2} \pm 2\pi n$ velocity changes from zero to $-\infty$ and $+\infty$, respectively. Transition of the phase fluid particles from the points $\xi = 0.5$, $\delta = \frac{\pi}{4} \pm 2\pi n$ to the points $\xi = 0.5$, $\delta = -\frac{\pi}{4} \pm 2\pi n$ proceeds instantly.

Along the vertical lines $\delta = \pm\frac{\pi}{2}$ the analytical solution could be written in the following form:

$$\frac{4}{\xi - 0.5}\sqrt{(1-\xi)(\xi - 0.5)}\Big|_{\xi_0}^{\xi} = \mp\frac{bE_0}{s}\left(1 - e^{-sT_2}\right),$$

$$\delta(T_2) = \delta_0 = \pm\frac{\pi}{2} \pm 2\pi n, \quad n = 0, 1, 2, \ldots \tag{3.218}$$

3.5.2.2 The Case (3.92) at $c_1 = 0.8$, $c_2 = c_3 = -0.3$

In this case, the stream-function $G(\xi, \delta)$ takes the form

$$G(\xi, \delta) = (\xi - 0.3)\sqrt{(0.8 - \xi)(\xi - 0.3)}\cos\delta = G(\xi_0, \delta_0),$$

and Fig. 3.8 shows the stream lines of the phase fluid in the phase plane.

Comparison of Figs. 3.7 and 3.8 shows that the channel with the phase fluid moves up, in so doing the bottom and upper boundaries $\xi = 0$ and $\xi = 0.5$ in Fig. 3.7 go over to the bottom and upper boundaries $\xi = 0.3$ and $\xi = 0.8$ in Fig. 3.8, respectively. As this takes place, the stationary center points with the coordinates $\xi = \frac{3}{8}$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ in Fig. 3.7 move to the points with the coordinates $\xi = 0.675$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ in Fig. 3.8.

All other reasoning presented for the previous case is valid for the case under consideration.

3.5.2.3 The Case (3.92) at $c_1 = 1$, $c_2 = c_3 = -0.5$

In this case, the stream-function $G(\xi, \delta)$ takes the form

$$G(\xi, \delta) = (\xi - 0.5)\sqrt{(1 - \xi)(\xi - 0.5)}\cos\delta = G(\xi_0, \delta_0),$$

and Fig. 3.9 shows the stream lines of the phase fluid in the phase plane.

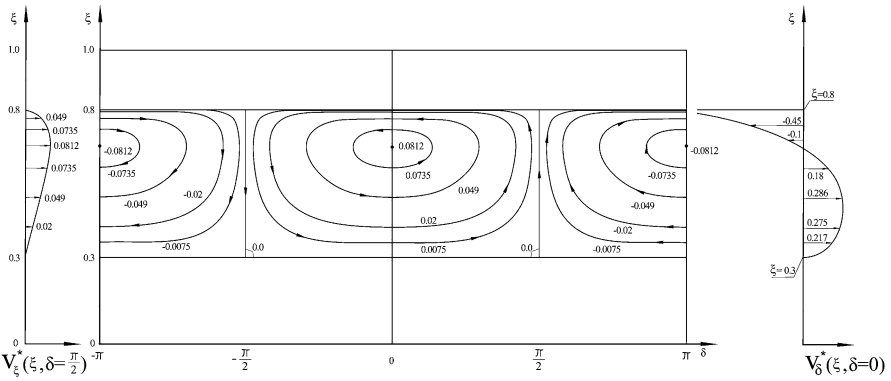


Fig. 3.8 Phase portrait for the case of the combinational resonance of the second type: $c_1 = 0.8$, $c_2 = c_3 = -0.3$

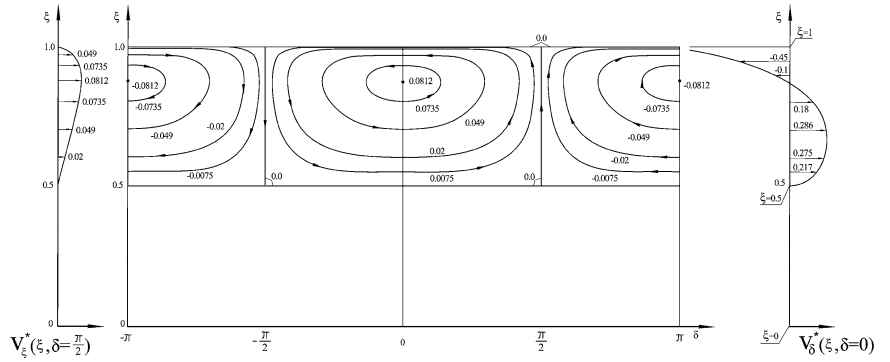


Fig. 3.9 Phase portrait for the case of the combinational resonance of the second type: $c_1 = 1$, $c_2 = c_3 = -0.5$

Comparison of Figs. 3.7, 3.8, and 3.9 shows that the channel with the phase fluid moves up once more, in so doing the bottom and upper boundaries now take place at $\xi = 0.5$ and $\xi = 1$ in Fig. 3.9, in so doing the stationary center points have the coordinates $\xi = 0.875$, $\delta = \pm\pi n$ $(n = 0, 1, 2, \ldots)$ in Fig. 3.9.

All other reasoning presented for the last two previous cases is valid for the case under consideration.

3.6 Conclusion

Free damped vibrations of a non-linear cylindrical shell in a fractional derivative viscoelastic medium have been investigated. Damped vibrations occur under the different conditions of the internal resonance: two-to-one, one-to-one, one-to-one-to-two, one-to-two-to-two, three-to-one, one-to-one-to-three, one-to-three-to-three,

and combinational resonances of the additive and difference types, in so doing the type of the resonance depends on the order of smallness of the fractional derivative entering in the equations of motion of the shell.

Combinational internal resonances resulting in coupling of three modes have been investigated in detail. The nonlinear sets of resolving equations in terms of amplitudes and phase differences have been obtained. For both types of the combinational internal resonances there exist such particular cases when it is possible to obtain two first integrals, namely: the energy integral and the stream-function, what allows one to reduce the problem to the calculation of elliptic integrals.

The proposed analytical approach for investigating the damped vibrations of the nonlinear cylindrical shell subjected to the conditions of the internal resonance has been possible owing to the suggested procedure resulting in decoupling linear parts of equations with the further utilization of the method of multiple scales for solving nonlinear governing equations of motion. It is shown that the phenomenon of internal resonance can be very critical, since in the circular cylindrical shell the internal resonance is always present.

The internal resonances belong to the resonances of the constructive type, despite of external resonances which could be eliminated by changing the frequency of external loads. Since all of internal resonances depend on the geometrical dimensions of the shell under consideration and its mechanical characteristics, that is why such resonances could not be ignored and eliminated for a particularly designed shell.

Internal resonances occurring in non-linear systems could be dangerous, since they could result in the energy transfer from one subsystem to another and reverse. Thus, in the problem considered in this paper it has been shown that the energy exchange could occur between three subsystems at a time: normal vibrations of the shell, its torsional vibrations and shear vibrations along the shell axis. Such an energy exchange, if it takes place for a rather long time, could result in crack formation in the shell, and finally to its failure. It is convenient to trace the energy exchange on the phase portraits, wherein closed phase trajectories are often changed by non-closed trajectories. The lines separating the closed lines from non-closed ones are called separatrices, along which irreversible energy interchange takes place. Along the closed trajectories the energetic state of the subsystem repeats periodically in time, while along the non-closed lines the functions change aperiodically.

The influence of viscosity on the energy exchange mechanism has been analyzed. It has been shown that each mode is characterized by its damping coefficient connected with the natural frequency by the exponential relationship with a negative fractional exponent. It has been revealed that during free vibrations of the shell under consideration with due account for the conditions of the combinational internal resonance three regimes could be observed: stationary (absence of damping at $\gamma=0$ and $\mu = 0$), quasi-stationary (damping is defined by an ordinary derivative at $\gamma = 1$), and transient (damping is defined by a fractional derivative at $0 < \gamma < 1$). Phenomenological analysis of nonlinear vibrations of a conventionally damped

cylindrical shell with the combinational internal resonance has been carried out by constructing the phase portraits of the stream-function at different magnitudes of the coefficients depending on the shell parameters.

Acknowledgements This research was made possible by the Grant No. 7.22.2014/K as a Government task from the Ministry of Education and Science of the Russian Federation.

Appendix A

$$a_{1mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \sin \pi m_1 x \sin n_1 \varphi \cos \pi m_2 x \sin n_2 \varphi \cos \pi m x \sin n\varphi \; dx d\varphi,$$

$$a_{2mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \cos \pi m_1 x \cos n_1 \varphi \sin \pi m_2 x \cos n_2 \varphi \cos \pi m x \sin n\varphi \; dx d\varphi,$$

$$a_{3mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \sin \pi m_1 x \cos n_1 \varphi \sin \pi m_2 x \sin n_2 \varphi \sin \pi m x \cos n\varphi \; dx d\varphi,$$

$$a_{4mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \cos \pi m_1 x \sin n_1 \varphi \cos \pi m_2 x \cos n_2 \varphi \sin \pi m x \cos n\varphi \; dx d\varphi,$$

$$a_{5mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \sin \pi m_1 x \sin n_1 \varphi \sin \pi m_2 x \sin n_2 \varphi \sin \pi m x \sin n\varphi \; dx d\varphi,$$

$$a_{6mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \cos \pi m_1 x \cos n_1 \varphi \cos \pi m_2 x \cos n_2 \varphi \sin \pi m x \sin n\varphi \; dx d\varphi,$$

$$a_{7mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \cos \pi m_1 x \sin n_1 \varphi \cos \pi m_2 x \sin n_2 \varphi \sin \pi m x \sin n\varphi \; dx d\varphi,$$

$$a_{8mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \sin \pi m_1 x \cos n_1 \varphi \sin \pi m_2 x \cos n_2 \varphi \sin \pi m x \sin n\varphi \; dx d\varphi,$$

$$a_{9mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \cos \pi m_1 x \sin n_1 \varphi \cos \pi m_2 x \sin n_2 \varphi \sin \pi m x \sin n\varphi \; dx d\varphi,$$

$$a_{10mn}^{m_1n_1m_2n_2} = \int_0^1 \int_0^{2\pi} \sin \pi m_1 x \cos n_1 \varphi \sin \pi m_2 x \cos n_2 \varphi \sin \pi m x \sin n\varphi \; dx d\varphi \; .$$

The above integrals could be easily calculated using the following formulas [36]:

$$\sin\alpha \sin\beta \sin\gamma = \frac{1}{4}\Big[\sin(\alpha+\beta-\gamma) + \sin(\beta+\gamma-\alpha) + \sin(\gamma+\alpha-\beta) - \sin(\alpha+\beta+\gamma)\Big],$$

$$\sin\alpha \cos\beta \cos\gamma = \frac{1}{4}\Big[\sin(\alpha+\beta-\gamma) - \sin(\beta+\gamma-\alpha) + \sin(\gamma+\alpha-\beta) - \sin(\alpha+\beta+\gamma)\Big],$$

$$\sin\alpha\sin\beta\cos\gamma = \frac{1}{4}\Big[-\cos(\alpha+\beta-\gamma)+\cos(\beta+\gamma-\alpha)+\cos(\gamma+\alpha-\beta)+\cos(\alpha+\beta+\gamma)\Big],$$

$$\cos\alpha\cos\beta\cos\gamma = \frac{1}{4}\Big[\cos(\alpha+\beta-\gamma)+\cos(\beta+\gamma-\alpha)+\cos(\gamma+\alpha-\beta)+\cos(\alpha+\beta+\gamma)\Big].$$

Appendix B

$$k_1 = \frac{a_{11}^I}{3\Omega_1^2},\quad k_2 = -\frac{a_{11}^I}{\Omega_1^2},\quad k_3 = \frac{a_{22}^I}{4\Omega_2^2-\Omega_1^2},\quad k_4 = -\frac{a_{22}^I}{\Omega_1^2},\quad k_5 = \frac{a_{33}^I}{4\Omega_3^2-\Omega_1^2},$$

$$k_6 = -\frac{a_{33}^I}{\Omega_1^2},\quad k_7 = \frac{a_{12}^I}{(\Omega_1+\Omega_2)^2-\Omega_1^2},\quad k_8 = \frac{a_{12}^I}{(\Omega_1-\Omega_2)^2-\Omega_1^2},$$

$$k_9 = \frac{a_{13}^I}{(\Omega_1+\Omega_3)^2-\Omega_1^2},\quad k_{10} = \frac{a_{13}^I}{(\Omega_1-\Omega_3)^2-\Omega_1^2},$$

$$k_{11} = \frac{a_{23}^I}{(\Omega_2+\Omega_3)^2-\Omega_1^2},\quad k_{12} = \frac{a_{23}^I}{(\Omega_2-\Omega_3)^2-\Omega_1^2};$$

$$g_1 = \frac{a_{11}^{II}}{4\Omega_1^2-\Omega_2^2},\quad g_2 = -\frac{a_{11}^{II}}{\Omega_2^2},\quad g_3 = \frac{a_{22}^{II}}{3\Omega_2^2},\quad g_4 = -\frac{a_{22}^{II}}{\Omega_2^2},\quad g_5 = \frac{a_{33}^{II}}{4\Omega_3^2-\Omega_2^2},$$

$$g_6 = -\frac{a_{33}^{II}}{\Omega_2^2},\quad g_7 = \frac{a_{12}^{II}}{(\Omega_1+\Omega_2)^2-\Omega_2^2},\quad g_8 = \frac{a_{12}^{II}}{(\Omega_1-\Omega_2)^2-\Omega_2^2},$$

$$g_9 = \frac{a_{13}^{II}}{(\Omega_1+\Omega_3)^2-\Omega_2^2},\quad g_{10} = \frac{a_{13}^{II}}{(\Omega_1-\Omega_3)^2-\Omega_2^2},$$

$$g_{11} = \frac{a_{23}^{II}}{(\Omega_2+\Omega_3)^2-\Omega_2^2},\quad g_{12} = \frac{a_{23}^{II}}{(\Omega_2-\Omega_3)^2-\Omega_2^2};$$

$$q_1 = \frac{a_{11}^{III}}{4\Omega_3^2-\Omega_1^2},\quad q_2 = -\frac{a_{11}^{III}}{\Omega_3^2},\quad q_3 = \frac{a_{22}^{III}}{4\Omega_2^2-\Omega_3^2},\quad q_4 = -\frac{a_{22}^{III}}{\Omega_3^2},$$

$$q_5 = \frac{a_{33}^{III}}{3\Omega_3^2},\quad q_6 = -\frac{a_{33}^{III}}{\Omega_3^2},\quad q_7 = \frac{a_{12}^{III}}{(\Omega_1+\Omega_2)^2-\Omega_3^2},\quad q_8 = \frac{a_{12}^{III}}{(\Omega_1-\Omega_2)^2-\Omega_3^2},$$

$$q_9 = \frac{a_{13}^{III}}{(\Omega_1 + \Omega_3)^2 - \Omega_3^2}, \quad q_{10} = \frac{a_{13}^{III}}{(\Omega_1 - \Omega_3)^2 - \Omega_3^2},$$

$$q_{11} = \frac{a_{23}^{III}}{(\Omega_2 + \Omega_3)^2 - \Omega_3^2}, \quad q_{12} = \frac{a_{23}^{III}}{(\Omega_2 - \Omega_3)^2 - \Omega_3^2};$$

$$d_1^J = 2a_{11}^J(k_1 + 2k_2) + a_{12}^J(g_1 + 2g_2) + a_{13}^J(q_1 + 2q_2),$$

$$d_2^J = 4a_{11}^J k_4 + 2a_{22}^J(g_7 + g_8) + a_{12}^J(2g_4 + k_7 + k_8) + 2a_{13}^J q_4 + a_{23}^J(q_7 + q_8),$$

$$d_3^J = 4a_{11}^J k_6 + 2a_{33}^J(q_9 + q_{10}) + 2a_{12}^J g_6 + a_{13}^J(2q_6 + k_9 + k_{10}) + a_{23}^J(g_9 + g_{10}),$$

$$d_4^J = 2a_{22}^J(g_3 + 2g_4) + a_{12}^J(k_3 + 2k_4) + a_{23}^J(q_3 + 2q_4),$$

$$d_5^J = 2a_{11}^J(k_7 + k_8) + 4a_{22}^J g_2 + a_{12}^J(2k_2 + g_7 + g_8) + a_{12}^J(q_7 + q_8) + 2a_{23}^J q_2,$$

$$d_6^J = 4a_{22}^J g_6 + 2a_{33}^J(q_{11} + q_{12}) + 2a_{12}^J k_6 + a_{13}^J(k_{11} + k_{12}) + a_{23}^J(2q_6 + g_{11} + g_{12}),$$

$$d_7^J = 2a_{33}^J(q_5 + 2q_6) + a_{13}^J(k_5 + 2k_6) + a_{23}^J(g_5 + 2g_6),$$

$$d_8^J = 2a_{11}^J(k_9 + k_{10}) + 4a_{33}^J q_2 + a_{12}^J(g_9 + g_{10}) + a_{13}^J(2k_2 + q_9 + q_{10}) + 2a_{23}^J g_2,$$

$$d_9^J = 2a_{22}^J(g_{11} + g_{12}) + 4a_{33}^J q_4 + a_{12}^J(k_{11} + k_{12}) + 2a_{13}^J k_4 + a_{23}^J(2g_4 + q_{11} + q_{12}),$$

$$d_{10}^J = 2a_{11}^J k_1 + a_{12}^J g_1 + a_{13}^J q_1, \quad d_{11}^J = 2a_{22}^J g_3 + a_{12}^J k_3 + a_{23}^J q_3,$$

$$d_{12}^J = 2a_{33}^J q_5 + a_{13}^J k_5 + a_{23}^J g_5;$$

$$e_1^J = 2a_{11}^J k_{11} + 2a_{22}^J g_9 + 2a_{33}^J q_7 + a_{12}^J(g_{11} + k_9) + a_{13}^J(q_{11} + k_7) + a_{23}^J(q_9 + g_7),$$

$$e_2^J = 2a_{11}^J k_{12} + 2a_{22}^J g_{10} + 2a_{33}^J q_7 + a_{12}^J(g_{12} + k_{10}) + a_{13}^J(q_{12} + k_7) + a_{23}^J(q_{10} + g_7),$$

$$e_3^J = 2a_{11}^J k_{12} + 2a_{22}^J g_9 + 2a_{33}^J q_8 + a_{12}^J(g_{12} + k_9) + a_{13}^J(q_{12} + k_8) + a_{23}^J(q_9 + g_8),$$

$$e_4^J = 2a_{11}^J k_{11} + 2a_{22}^J g_{10} + 2a_{33}^J q_8 + a_{12}^J(g_{11} + k_{10}) + a_{13}^J(q_{11} + k_8) + a_{23}^J(q_9 + g_8);$$

$$c_1^J = 2a_{11}^J k_3 + 2a_{22}^J g_7 + a_{12}^J(g_3 + k_7) + a_{13}^J q_3 + a_{23}^J q_7,$$

$$c_2^J = 2a_{11}^J k_5 + 2a_{33}^J q_9 + a_{12}^J g_5 + a_{13}^J(q_5 + k_9) + a_{23}^J g_9,$$

$$c_3^J = 2a_{11}^J k_7 + 2a_{22}^J g_1 + a_{12}^J(g_7 + k_1) + a_{13}^J q_7 + a_{23}^J q_1,$$

$$c_4^J = 2a_{11}^J k_9 + 2a_{33}^J q_1 + a_{12}^J g_9 + a_{13}^J(q_9 + k_1) + a_{23}^J g_1,$$

$$c_5^J = 2a_{22}^J g_5 + 2a_{33}^J q_{11} + a_{12}^J k_5 + a_{13}^J k_{11} + a_{23}^J (q_5 + g_{11}),$$

$$c_6^J = 2a_{22}^J g_{11} + 2a_{33}^J q_3 + a_{12}^J k_{11} + a_{13}^J k_3 + a_{23}^J (q_{11} + g_3),$$

$$c_7^J = 2a_{11}^J k_8 + 2a_{22}^J g_1 + a_{12}^J (g_8 + k_1) + a_{13}^J q_8 + a_{23}^J q_1,$$

$$c_8^J = 2a_{11}^J k_{10} + 2a_{33}^J q_1 + a_{12}^J g_{10} + a_{13}^J (q_{10} + k_1) + a_{23}^J g_1,$$

$$c_9^J = 2a_{11}^J k_3 + 2a_{22}^J g_8 + a_{12}^J (g_3 + k_8) + a_{13}^J q_3 + a_{23}^J q_8,$$

$$c_{10}^J = 2a_{11}^J k_5 + 2a_{33}^J q_{10} + a_{12}^J g_5 + a_{13}^J (q_5 + k_{10}) + a_{23}^J g_{10},$$

$$c_{11}^J = 2a_{22}^J g_{12} + 2a_{33}^J q_3 + a_{12}^J k_{12} + a_{13}^J k_3 + a_{23}^J (q_{12} + g_3),$$

$$c_{12}^J = 2a_{22}^J g_5 + 2a_{33}^J q_{12} + a_{12}^J k_5 + a_{13}^J k_{12} + a_{23}^J (q_5 + g_{12}).$$

References

1. Witt, A.A., Gorelik, G.A.: Vibrations of an elastic pendulum as an example of vibrations of two parametrically coupled linear systems (in Russian). J. Tech. Phys. **2–3**, 294–307 (1933)
2. Nayfeh, A.H., Balachandran, S.: Modal interactions in dynamical and structural systems. Appl. Mech. Rev. **42**, S175–S201 (1989)
3. Nayfeh, A.H.: Nonlinear Interaction: Analytical, Computational, and Experimental Methods. Wiley, New York (2000)
4. Popov, A.A., Thompson, J.M.T., McRobie, F.A.: Low dimensional models of shell vibrations. Parametrically excited vibrations of cylindrical shells. J. Sound Vib. **209**, 163–186 (1998)
5. Amabili, M., Pellicano, F., Vakakis, A.F.: Nonlinear vibrations and multiple resonances of fluid-filled circular shells. Part 1: Equations of motion and numerical results. ASME J. Vib. Acoust. **122**, 346–354 (2000)
6. Nayfeh, A.H., Raouf, R.A.: Non-linear oscillation of circular cylindrical shells. Int. J. Solids Struct. **23**, 1625–1638 (1987)
7. McRobie, F.A., Popov, A.A., Thompson, J.M.T.: Auto-parametric resonance in cylindrical shells using geometric averaging. J. Sound Vib. **227**, 65–84 (1999)
8. Avramov, K.V.: Nonlinear forced vibrations of a cylindrical shell with two internal resonances. Int. Appl. Mech. **42**(2), 169–175 (2006)
9. Popov, A.A.: Auto-parametric resonance in thin cylindrical shells using the slow fluctuation method. Thin-Walled Struct. **42**, 475–495 (2004)
10. Kubenko, V.D., Koval'chuk, P.S.: Nonlinear problems of the vibration of thin shell (review). Int. Appl. Mech. **34**(8), 703–728 (1998)
11. Amabili, M., Païdoussis, M.P.: Review of studies on geometrically nonlinear vibrations and dynamics of circular cylindrical shells and panels, with and without fluid-structure interaction. ASME Appl. Mech. Rev. **56**, 349–381 (2003)
12. Lee, Y.-S. : Review on the cylindrical shell research (in Korean). Trans. KSME **33**, 1–26 (2009)
13. Breslavsky, I.D., Avramov, K.V.: Nonlinear modes of cylindrical panels with complex boundaries. R-function method. Meccanica **46**, 817–832 (2011)
14. Avramov, K.V., Mikhin, Yu.V., Kurilov, E.: Asymptotic analysis of nonlinear dynamics of simply supported cylindrical shells. Nonlinear Dyn. **47**, 331–352 (2012)

15. Goncalves, P.B., del Prado, Z.J.G.N.: Low-dimensional Galerkin models for nonlinear vibration and instability analysis of cylindrical shells. Nonlinear Dyn. **41**, 129–145 (2005)
16. Amabili, M.: A comparison of shell theories for large-amplitude vibrations of circular cylindrical shells: Lagrangian approach. J. Sound Vib. **264**, 1091–1125 (2003)
17. Amabili, M., Reddy, J.N.: A new non-linear higher-order shear deformation theory for large-amplitude vibrations of laminated doubly curved shells. Int. J. Non-Linear Mech. **45**, 409–418 (2010)
18. Amabili, M.: Internal resonances in non-linear vibrations of a laminated circular cylindrical shell. Nonlinear Dyn. **69**, 755–770 (2012)
19. Amabili, M.: Reduced-order models for nonlinear vibrations, based on natural modes: The case of the circular cylindrical shell. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **371**(1993), art. no. 20120474 (2013)
20. Rossikhin, Yu.A., Shitikova, M.V.: Application of fractional calculus for analysis of nonlinear damped vibrations of suspension bridges. ASCE J. Eng. Mech. **124**, 1029–1036 (1998)
21. Rossikhin, Yu.A., Shitikova, M.V.: Application of fractional calculus for dynamic problems of solid mechanics: Novel trends and recent results. Appl. Mech. Rev. **63**(1), art. no. 010801-1–52 (2010)
22. Rossikhin, Yu.A., Shitikova, M.V.: Analysis of nonlinear vibrations of a two-degree-of-freedom mechanical system with damping modelled by a fractional desrivative. J. Eng. Math. **37**, 343–362 (2000)
23. Rossikhin, Yu.A., Shitikova, M.V.: Free damped non-linear vibrations of a viscoelastic plate under the two-to-one internal resonance. Mater. Sci. Forum **440–441**, 29–36 (2003)
24. Rossikhin, Yu.A., Shitikova, M.V.: Analysis of free non-linear vibrations of a viscoelastic plate under the conditions of different internal resonances. Int. J. Non-Linear Mech. **41**, 313–325 (2006)
25. Rossikhin, Yu.A., Shitikova, M.V.: A new approach for studying nonlinear dynamic response of a thin fractionally damped cylindrical shell with internal resonances of the order of ε. In: Altenbach, H., Mikhasev, G.I. (eds.) Shell and Membrane Theories in Mechanics and Biology: From Macro- to Nanoscale Structures. Advanced Structured Materials, vol. 45, pp. 301–321. Springer, Heidelberg (2015)
26. Volmir, A.S.: Nonlinear Dynamics of Plates and Shells (in Russian). Nauka, Moscow (1972)
27. Samko, S.G., Kilbas, A.A., Marichev, O.I.: Fractional Integrals and Derivatives. Theory and Applications (in Russian). Nauka i Tekhnika, Minsk (1988) (Engl. transl. by Gordon and Breach Science Publ., Amsterdam 1993)
28. Clough, R.W., Penzien, J.: Dynamics of Structures. McGraw-Hill, New York (1975)
29. Rossikhin, Yu.A., Shitikova, M.V.: Nonlinear free damped vibrations of suspension bridges with uncertain fractional damping. J Eur des Syst Automatises **42**(6–8), 879–894 (2008)
30. Emama, S.A., Nayfeh, A.H.: Non-linear response of buckled beams to 1:1 and 3:1 internal resonances. Int. J. Non-Linear Mech. **52**, 12–25 (2013)
31. Nayfeh, A.H.: Perturbation Methods. Wiley, New York (1973)
32. Rossikhin, Yu.A., Shitikova, M.V., Shcheglova, T.A.: Forced vibrations of a nonlinear oscillator with weak fractional damping. J. Mech. Mat. Struct. **4**(9), 1619–1636 (2009)
33. Rossikhin, Yu.A., Shitikova, M.V.: On fallacies in the decision between the Caputo and Riemann-Liouville fractional derivatives for the analysis of the dynamic response of a nonlinear viscoelastic oscillator. Mech. Res. Commun. **45**, 22–27 (2012)
34. Rossikhin, Yu.A., Shitikova, M.V.: Nonlinear dynamic response of a fractionally damped cylindrical shell with a three-to-one internal resonance. Appl. Math. Comput. **257**, 498–525 (2015). doi:10.106/j.amc.2015.01.018
35. Abramowitz, M., Stegun I. (eds.): Handbook of Mathematical Functions With Formulas, Graphs, and Tables, vol. 55. National Bureau of Standards USA, Applied Mathematical Services, Washington, DC (1964)
36. Bronshtein, I.N., Semendyayev, K.A.: Handbook of Mathematics, 3rd edn. Springer, New York (1985)

Chapter 4
Schwartz-Christoffel Panel Method Improvements and Applications

Etsuo Morishita

Abstract Schwartz-Christoffel panel method is improved and applied to general and unique airfoils. A potential flow around a two-dimensional circular cylinder can be transformed to that of a two-dimensional airfoil by the Schwartz-Christoffel conformal mapping. Multiple straight panels approximate a two-dimensional airfoil. This method is particularly effective for very thin airfoils. The conventional panel method would suffer for these extremely thin airfoil problems. First, the method is improved and tested against a circular cylinder and the analytical Joukowski airfoil. Several real airfoils are studied together with the unique polygonal airfoil sections for the propeller of Mars exploration and the dragonfly airfoil. It is shown that the method gives satisfactory results for all cases.

Keywords Potential flow • Conformal mapping • Airfoil • Panel method

4.1 Introduction

Conformal mapping in the aerodynamics is a classical subject and the most famous one is the Joukowski transformation. Although the transformation is theoretically and mathematically beautiful, the method and the airfoil both are hardly used in the real engineering world. One difficulty lies in the fact that the trailing edge is very thin. Actually, it has zero thickness. The fact also imposes a numerical difficulty for the conventional panel methods. No one ever sees a numerical solution for very thin Joukowski airfoils by the panel methods in the textbooks. This might be due to the interaction of the singularities arrayed on the airfoil surface particularly around the trailing edge.

The present author found that the Schwartz-Christoffel transformation [1] can be applied to the two-dimensional airfoil [2]. A two-dimensional airfoil is approximated as a polygon. Several numerical results are reported for flows around two-dimensional cross sections including regular polygons and two-dimensional

E. Morishita (✉)
Department of Advanced Interdisciplinary Sciences, Graduate School of Engineering, Utsunomiya University, 7-1-2, Yoto, Utsunomiya, Tochigi 321-8585, Japan
e-mail: tmorisi@cc.utsunomiya-u.ac.jp

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_4

airfoils [3, 4]. It was found that the mid-shoulder velocity around some of the regular polygons, i.e. flat plate, square, hexagon, octagon etc. follow very beautiful formulae during the course of development [4]. The solution procedure was also improved from the original paper to the second one to simplify the calculation [2, 4].

In this paper, we would like to extend the application of the Schwartz-Christoffel method to various general and unique airfoils and show that the method is very effective for thin and polygonal airfoils for which the conventional panel methods might face difficulty [5, 6].

4.2 Schwartz-Christoffel Panel Method

The Schwartz Christoffel transformation from a circle to a paneled airfoil can be obtained from Eq. (4.1) [7].

$$\frac{d\zeta}{dz} = A \cdot \prod_{j=1}^{N} \left(1 - \frac{z_j}{z}\right)^{\mu_j} = A \cdot \left(1 - \frac{e^{i\theta_1}}{z}\right)^{\mu_1} \cdot \left(1 - \frac{e^{i\theta_2}}{z}\right)^{\mu_2} \cdot \ldots \cdot \left(1 - \frac{e^{i\theta_N}}{z}\right)^{\mu_N}. \tag{4.1}$$

where A is a complex constant $\left(A \equiv Ke^{i\kappa}\right)$, N is the number of apexes, z is the complex coordinate of an original plane, $z_j \left(= e^{i\theta_j},\ \theta_1 < \theta_2 < .. < \theta_j .. < \theta_N\right)$ is the j-th point on a unit circle, μ_j is the outer angular ratio to π at the j-th apex of a polygon and $\zeta \equiv \xi + i\eta$ is the transformed coordinate. A schematic view is shown in Fig. 4.1. The uniform velocity is U_∞ at the angle of attack α_z in the z plane and V_∞ at the angle of attack α in the ζ plane, respectively. The stagnation points are $z_1 = 1$, and the trailing edge, respectively.

The following equations must be satisfied in Eq. (4.1).

$$\sum_{j=1}^{N} \mu_j = 2. \tag{4.2}$$

$$\sum_{j=1}^{N} \mu_j z_j = 0. \tag{4.3}$$

where Eq. (4.2) represents the fact that the sum of the angular rotation at each j-th apex of a polygon is equal to 2π, and Eq. (4.3) is a necessary condition to avoid singularity by Eq. (4.1) and to form a closed polygon.

The complex potential is given by w in Fig. 4.1a. In Fig. 4.1b,

$$\frac{dw}{d\zeta}\Big|_{|\zeta|\to\infty} = \frac{\left(\frac{dw}{dz}\right)_\infty}{\left(\frac{d\zeta}{dz}\right)_\infty} = \frac{U_\infty e^{-i\alpha_z}}{A} = \left(\frac{U_\infty}{K}\right) e^{-i(\alpha_z+\kappa)} \equiv V_\infty e^{-i\alpha}. \tag{4.4}$$

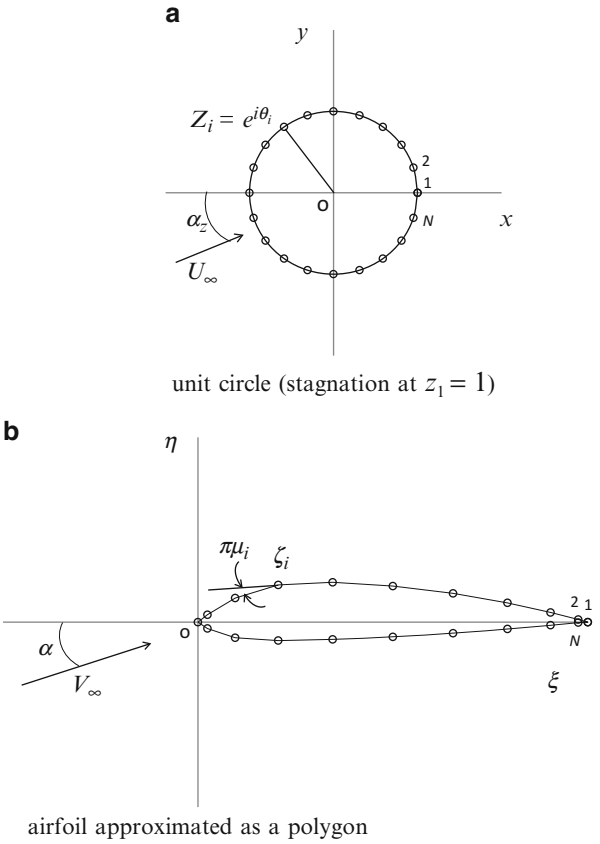


Fig. 4.1 Schwartz-Christoffel transformation from a unit circle (**a**) to an airfoil (**b**)

Therefore, the uniform velocity V_∞ and the angle of attack α in the transformed plane become respectively as follows where K and κ are both real constants.

$$V_\infty = \frac{U_\infty}{K} \tag{4.5}$$

$$\alpha = \alpha_z + \kappa \tag{4.6}$$

Equation (4.1) becomes on the unit circle $z = e^{i\theta}$ as follows.

$$d\zeta = K \cdot 2^2 \cdot e^{i \cdot \left[\frac{3}{2}\pi + \frac{1}{2} \sum_{j=1}^{N} \mu_j \cdot \theta_j + \kappa \right]} \cdot \prod_{j=1}^{N} \left(\sin \frac{\theta_j - \theta}{2} \right)^{\mu_j} \cdot d\theta. \tag{4.7}$$

From Eq. (4.7), for the panel i

$$\zeta_{i+1} - \zeta_i = \left[K \cdot 2^2 \cdot \int_{\theta_i}^{\theta_{i+1}} \prod_{j=1}^{N} \left| \sin \frac{\theta_j - \theta}{2} \right|^{\mu_j} \cdot d\theta \right] \cdot e^{i \left[\frac{3}{2}\pi + \frac{1}{2} \sum_{j=1}^{N} \mu_j \cdot \theta_j + \kappa + \pi \sum_{j=1}^{i} \mu_j \right]} \equiv l_i \cdot e^{i\delta_i}. \tag{4.8}$$

where

$$l_i \equiv K \cdot 2^2 \cdot \int_{\theta_i}^{\theta_{i+1}} \prod_{j=1}^{N} \left| \sin \frac{\theta_j - \theta}{2} \right|^{\mu_j} \cdot d\theta \quad (1 \le i \le N). \tag{4.9}$$

and

$$\delta_i \equiv \frac{3}{2}\pi + \frac{1}{2} \sum_{j=1}^{N} \mu_j \cdot \theta_j + \kappa + \pi \sum_{j=1}^{i} \mu_j \quad (1 \le i \le N). \tag{4.10}$$

The panel length is l_i and the panel inclination is δ_i for the i -th panel, respectively.

We have to find the proper angle θ_j $(1 \le j \le N)$ on the unit circle so that the panel length l_i becomes exactly the same value as that of the given airfoil. The initial angle θ_1 on the unit circle can be chosen arbitrary and we can set $\theta_1 = 0$. Note that $\theta_{N+1} = 2\pi$. Therefore for a given airfoil, the governing equation of the Schwartz-Christoffel panel method becomes Eqs. (4.9), i.e.

$$K \cdot 2^2 \cdot \int_{\theta_i}^{\theta_{i+1}} \prod_{j=1}^{N} \left| \sin \frac{\theta_j - \theta}{2} \right|^{\mu_j} \cdot d\theta = l_i \quad (1 \le i \le N). \tag{4.11}$$

We have $N-1$ unknown angles θ_j $(2 \le j \le N)$ and one unknown real constant K here, and therefore we can determine these N unknowns from Eq. (4.11). By adding the length of each panel, we can get the airfoil perimeter P as follows.

$$\sum_{i=1}^{N} l_i = P. \tag{4.12}$$

Equation (4.12) can replace one of the Eq. (4.11), for example $i = N$, and is used to determine the real constant K. When Eqs. (4.11) and (4.12) are satisfied, the polygon is the same as the given paneled airfoil and it is expected that Eq. (4.3) is satisfied

automatically. Therefore, Eq. (4.3) was not used in the present study to simplify the solution process, although it was used in the original analysis [2]. Equation (4.12) is newly introduced instead of Eq. (4.3).

The governing equations are non-linear and iterations are necessary to determine $\theta_i \ (2 \leq i \leq N)$, K and κ.

A possible solution procedure is as follows. The angle of attack α in the ζ -plane is first given. A given airfoil chord locates on the ξ -axis. The initial guess for $\theta_i \ (2 \leq i \leq N)$ may be given on the unit circle as follows.

$$\theta_i = \frac{2\pi}{N} \cdot (i-1) \quad (2 \leq i \leq N)\,. \tag{4.13}$$

The first panel ($i = 1$) inclination measured counterclockwise from the ξ axis is a given constant $\delta = \delta_1$ and the geometrical condition becomes

$$\tan(\pi - \delta_1) = \frac{\eta_2 - \eta_1}{\xi_1 - \xi_2}. \tag{4.14}$$

where the right-hand side of Eq. (4.14) is calculated from the given airfoil panel coordinates and

$$\delta_1 = \pi - \tan^{-1}\frac{\eta_2 - \eta_1}{\xi_1 - \xi_2}. \tag{4.15}$$

From Eq. (4.10) for the panel $i = 1$, δ_1 satisfies

$$\delta_1 = \frac{3}{2}\pi + \frac{1}{2}\sum_{j=1}^{N}\mu_j \cdot \theta_j + \kappa + \pi\mu_1. \tag{4.16}$$

From Eq. (4.16), we can determine κ with δ_1 of Eq. (4.15) as follows.

$$\kappa = \delta_1 - \left[\frac{3}{2}\pi + \pi\mu_1 + \frac{1}{2}\sum_{j=1}^{N}\mu_j \cdot \theta_j\right]. \tag{4.17}$$

Once κ is determined, each panel inclination $\delta_i \ (2 \leq i \leq N)$ is also calculated.

We can evaluate each tentative panel length l_i' relative to K from Eq. (4.11) although we do not know the specific value of the real constant K at this stage.

$$\frac{l_i'}{K} = 2^2 \cdot \int_{\theta_i}^{\theta_{i+1}} \prod_{j=1}^{N} \left|\sin\frac{\theta_j - \theta}{2}\right|^{\mu_j} \cdot d\theta \quad [\equiv I(\theta_i, \theta_{i+1})] \quad (1 \leq i \leq N)\,. \tag{4.18}$$

where I is an integrated constant obtained numerically.

For the given airfoil, the exact total panel length, i.e. the perimeter is given by Eq. (4.12). On the other hand, the perimeter P' during the iteration is given by

$$P' = \sum_{j=1}^{N} l'_j = K \sum_{j=1}^{N} I\left(\theta_j, \theta_{j+1}\right) = K \cdot I\left(0, 2\pi\right). \tag{4.19}$$

The geometric requirement is $P' = P$, and therefore K is determined as follows.

$$K = \frac{P}{I\left(0, 2\pi\right)}. \tag{4.20}$$

The above procedure is the first iteration (see [1–3] in Fig. 4.2).

We then update the angles $\theta_i \quad (2 \leq i \leq N)$. A simple update method can be as follows.

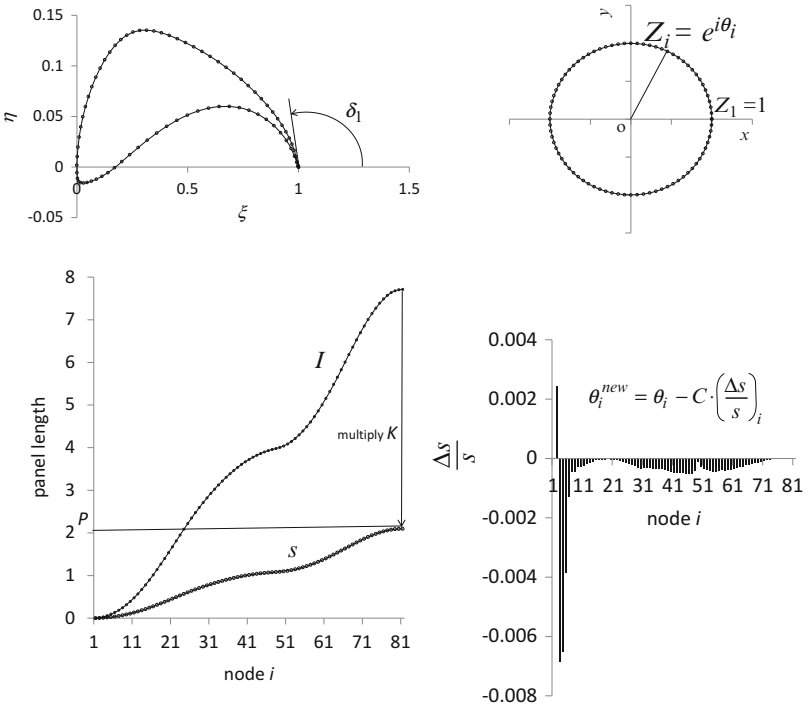


Fig. 4.2 Schwartz-Christoffel panel method iteration procedure. [1] airfoil N panels given. [2] assume $\theta_i = \frac{2\pi}{N} \cdot (i-1) \quad (2 \leq i \leq N)$ and determine κ from δ_1. [3] K determined from integral I and airfoil perimeter P. [4] update $\theta_i \quad (2 \leq i \leq N)$ from excess panel length $(\Delta s/s)_i$ and obtain κ. [5] repeat [3] and [4] to achieve $|\Delta s/s|_i < \varepsilon$ and restore original airfoil

$$\theta_i^{new} = \theta_i - C \cdot \frac{\left(\sum_{j=1}^{i-1} l'_j - \sum_{j=1}^{i-1} l_j \right)}{\sum_{j=1}^{i-1} l_j} \equiv \theta_i - C \cdot \left(\frac{s' - s}{s} \right)_i \quad (2 \le i \le N) . \quad (4.21)$$

where C is a positive constant that may be determined empirically, and s is the length measured counterclockwise along the airfoil surface from the trailing edge. The above equation means that when the total panel length up to i -th panel during the iteration $s'_i \equiv \sum_{j=1}^{i-1} l'_j$ exceeds that of the given one $s_i \equiv \sum_{j=1}^{i-1} l_j$, the corresponding θ_i is reduced and vice versa. Convergence is reached when

$$\left| \frac{s' - s}{s} \right|_i \equiv \left| \frac{\Delta s}{s} \right|_i < \varepsilon. \quad (4.22)$$

where ε is a small positive constant ($\varepsilon << 1$) and we finally determine θ_i $(2 \le i \le N)$, K and κ. The angle α_z is determined from Eq. (4.6).

The panel mid complex conjugate velocity is given by

$$\frac{\left(\frac{dw}{d\zeta} \right)}{V_\infty} |_i = \frac{\sin(\theta_{i\ \mathrm{mid}} - \alpha_z) + \sin \alpha_z}{2 \cdot \prod_{j=1}^{N} \left| \sin \frac{\theta_j - \theta_{i\ \mathrm{mid}}}{2} \right|^{\mu_j}} \cdot e^{-i\beta_i}. \quad (4.23)$$

where the numerator is derived from the surface velocity along the unit circle, and

$$\theta_{i\ \mathrm{mid}} = \frac{\theta_i + \theta_{i+1}}{2}. \quad (4.24)$$

$$\beta_i = -\pi\mu_1 - \pi + \delta_1 + \pi \sum_{j=1}^{i} \mu_j . \quad (4.25)$$

We can show that the lift coefficient C_l is directly obtained from K as follows.

$$C_l = 4 \cdot \left(\frac{K}{c} \right) \cdot 2\pi \sin \alpha_z. \quad (4.26)$$

where c is the chord length. Note that the location on each panel for Eq. (4.24) is very close to the middle of the panel, but not the same point.

The average velocity $\overline{V}_i$ of the i -th panel can be calculated as follows.

$$\frac{dw}{d\zeta} \cdot d\zeta = (u - iv) \cdot (d\xi + i d\eta) = u \cdot d\xi + v \cdot d\eta + i\,(u d\eta - v d\xi)$$
$$= \mathbf{V} \cdot d\mathbf{l} = V dl. \tag{4.27}$$

where u and v are the velocity components and $V = \sqrt{u^2 + v^2}$. When we integrate Eq. (4.27) counterclockwise along the panel surface, we get

$$-\overline{V}_i \cdot l = -\int_{\theta_i}^{\theta_{i+1}} V dl = -V_\infty \int_{\theta_i}^{\theta_{i+1}} \frac{dw}{d\zeta} \cdot d\zeta = 2K V_\infty [\cos(\theta - \alpha_z) - \theta \cdot \sin\alpha_z]_{\theta_i}^{\theta_{i+1}}$$

Therefore, the average velocity of the panel i becomes

$$\frac{\overline{V}_i}{V_\infty} = -\frac{K}{l_i} \cdot 2 \cdot [\cos(\theta_{i+1} - \alpha_z) - \cos(\theta_i - \alpha_z) - (\theta_{i+1} - \theta_i) \cdot \sin\alpha_z]. \tag{4.28}$$

This average velocity $\overline{V}_i$ has nearly the same value as that of the panel mid Eq. (4.23) although it is not necessarily the same.

The moment coefficient around the leading edge and/or the origin C_{m0} can be derived as follows.

$$dC_{m0} = -C_p \eta d\eta - C_p \xi d\xi = \left[\left(\frac{V}{V_\infty}\right)^2 - 1\right](\xi d\xi + \eta d\eta). \tag{4.29}$$

where C_p is the pressure coefficient and the coordinates are normalized by the chord length c. When the panel average velocity in Eq. (4.28) is used, the moment coefficient becomes (excluding -1 term)

$$dC_{m0} \approx \left(\frac{\overline{V}}{V_\infty}\right)^2 (\xi d\xi + \eta d\eta) = \left(\frac{\overline{V}}{V_\infty}\right)^2 \frac{1}{2} d\left(\xi^2 + \eta^2\right). \tag{4.30}$$

So the panel moment contribution can be approximated as follows.

$$\Delta C_{m0_i} \approx \frac{1}{2}\left(\frac{\overline{V}_i}{V_\infty}\right)^2 \left[|\zeta_{i+1}|^2 - |\zeta_i|^2\right]. \tag{4.31}$$

4.3 Method Validation

4.3.1 Regular Polygons and Circular Cylinder

The Schwartz-Christoffel transformation can generate regular polygons from a unit circle [3, 4, 8]. The complex conjugate velocity is obtained for the regular polygons with n apexes and the rear stagnation point on the ξ axis [4]. For a n-th regular polygon for $\theta_1 = 0$ ($\kappa = 0, \alpha = \alpha_z$)

$$\frac{\left(\frac{dw}{d\zeta}\right)}{V_\infty} = 2^{1-\frac{2}{n}} \cdot i^{1-\frac{2}{n}} \cdot \frac{\sin(\theta-\alpha)+\sin\alpha}{\left(\sin\frac{n}{2}\theta\right)^{\frac{2}{n}}}\Big|_{\alpha=0} = \frac{V}{V_\infty}\cdot e^{-i\beta_j} \quad . \tag{4.32}$$

where for the panel j, $\frac{2\pi}{n}\cdot(j-1) \le \theta < \frac{2\pi}{n}\cdot j \qquad (1 \le j \le n)$ and $\alpha = 0$

$$\frac{V_j}{V_\infty} = 2^{1-\frac{2}{n}} \cdot \frac{\sin\theta}{\left|\sin\frac{n}{2}\theta\right|^{\frac{2}{n}}}, \quad \frac{\overline{V}_j}{V_\infty} = 2^{1-\frac{2}{n}} \cdot \sin\left[\frac{2\pi}{n}\cdot\left(j-\frac{1}{2}\right)\right] \cdot \frac{2\cdot\sin\left(\frac{\pi}{n}\right)}{\int_0^{\frac{2\pi}{n}} \left(\sin\frac{n}{2}\theta\right)^{\frac{2}{n}}\cdot d\theta},$$

$$\beta_j = \left(\frac{2}{n}j - \frac{1}{n} - \frac{1}{2}\right)\cdot\pi$$

For a n-th regular polygon for $\theta_1 = -\pi/n$ ($\kappa = 0, \alpha = \alpha_z$),

$$\frac{\left(\frac{dw}{d\zeta}\right)}{V_\infty} = 2^{1-\frac{2}{n}} \cdot i \cdot \frac{\sin(\theta-\alpha)+\sin\alpha}{\left(\cos\frac{n}{2}\theta\right)^{\frac{2}{n}}}\Big|_{\alpha=0} = \frac{V}{V_\infty}\cdot e^{-i\beta_j} \tag{4.33}$$

where for the panel j, $-\frac{\pi}{n} + \frac{2\pi}{n}\cdot(j-1) \le \theta < -\frac{\pi}{n} + \frac{2\pi}{n}\cdot j \qquad (1 \le j \le n)$ and $\alpha = 0$

$$\frac{V_j}{V_\infty} = 2^{1-\frac{2}{n}} \cdot \frac{\sin\theta}{\left|\cos\frac{n}{2}\theta\right|^{\frac{2}{n}}}, \quad \frac{\overline{V}_j}{V_\infty} = 2^{1-\frac{2}{n}} \cdot \sin\left[\frac{2\pi}{n}\cdot(j-1)\right] \cdot \frac{2\cdot\sin\left(\frac{\pi}{n}\right)}{\int_{-\frac{\pi}{n}}^{+\frac{\pi}{n}} \cos\left(\frac{n}{2}\theta\right)^{\frac{2}{n}} d\theta},$$

$$\beta_j = \left(\frac{2}{n}j - \frac{2}{n} - \frac{1}{2}\right)\cdot\pi$$

From Eqs. (4.32) and (4.33), for the n -th even regular polygons, the magnitude of the shoulder mid velocity V_{s_mid} at $\alpha = 0$ and $\theta = \pi/2$ becomes

$$\frac{V_{s\_mid}}{V_\infty} = 2^{1-\frac{2}{n}} = 2^{\frac{m-1}{m}} \ (n = 2, 4, \ldots, 2m, ..) \tag{4.34}$$

where m is an integer. Equation (4.34) is shown in Fig. 4.3.

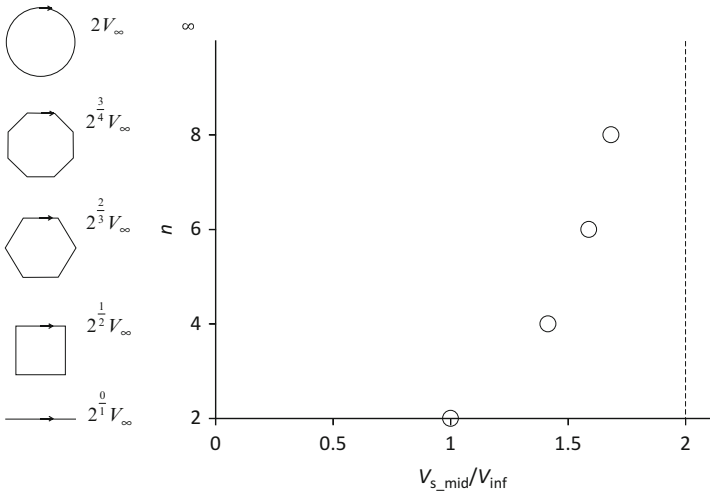


Fig. 4.3 Shoulder mid velocity of n -th even regular polygons

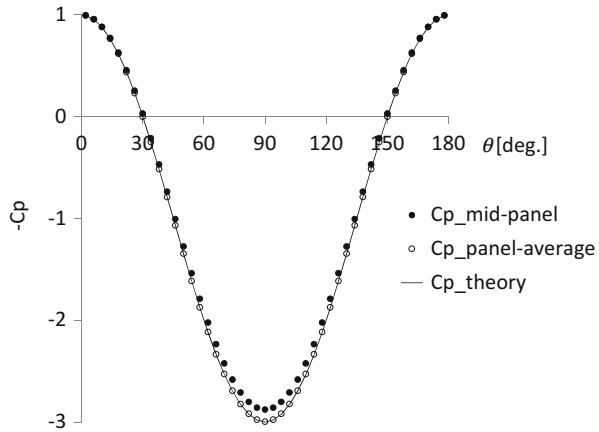


Fig. 4.4 Circular cylinder approximated by a regular polygon ($n = 90$)

Figure 4.4 shows the surface pressure distribution around a regular polygon $n = 90$ obtained from Eq. (4.32) at $\alpha = 0$. This is a model for a circular cylinder. Pressure is calculated from both V_j / V_∞ and $\overline{V}_j / V_\infty$ in Eq. (4.32). The shoulder mid velocity becomes approximately 1.969 for $n = 90$ and it is smaller than the panel average velocity 1.999. The equations for V_j / V_∞ and $\overline{V}_j / V_\infty$ both approach to the theory $2 \sin\theta$ for $n \to \infty$. The panel average velocity is almost identical with that of the theory in spite of the finite panels $n = 90$.Note that the solution itself is an exact analytical one for the regular polygon, although it is an approximate model for a circular cylinder.

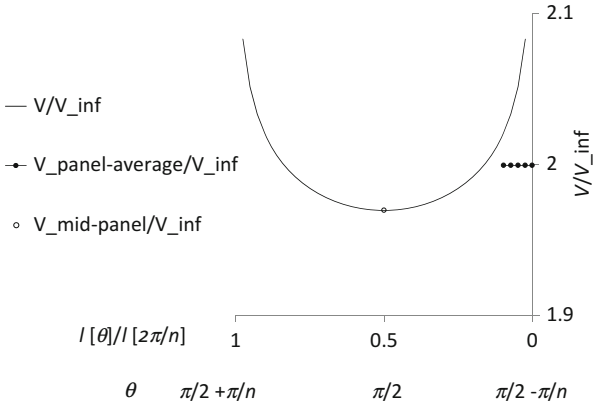


Fig. 4.5 Regular polygon shoulder velocity $n = 90, \quad \frac{\pi}{2} - \frac{\pi}{n} \le \theta \le \frac{\pi}{2} + \frac{\pi}{n}$

Figure 4.5 shows the surface velocity $V/V_\infty \left(= V_j/V_\infty \big|_{j=23}\right)$ on the shoulder panel for $n = 90, \quad 88^\circ \le \theta \le 90^\circ$. The panel mid velocity is the minimum 1.969 and the velocity goes to infinity at the both panel ends. The panel average velocity 1.999 is very close to the theoretical value 2. The lateral coordinate in Fig. 4.5 is given by

$$\frac{l(\theta)}{l\left(\frac{2\pi}{n}\right)} = \frac{\int_0^{\theta} \left(\sin\frac{n}{2}\theta\right)^{\frac{2}{n}} d\theta}{\int_0^{\frac{2\pi}{n}} \left(\sin\frac{n}{2}\theta\right)^{\frac{2}{n}} \cdot d\theta} \tag{4.35}$$

where $l(\theta)$ is the panel length as a function of the angle θ and

$$l\left(\frac{2\pi}{n}\right) = 2\sin\left(\frac{\pi}{n}\right) \tag{4.36}$$

when a unit circle circumscribes the regular polygon. It is interesting note that $l(\theta)$ is effectively a linear function of the angle θ for the large value of n.

4.3.2 Joukowski Airfoil

The Joukowski airfoil is used to test the validity of the present method. This is to prove that the technique is useful for a very thin trailing edge. Figure 4.6a, b show

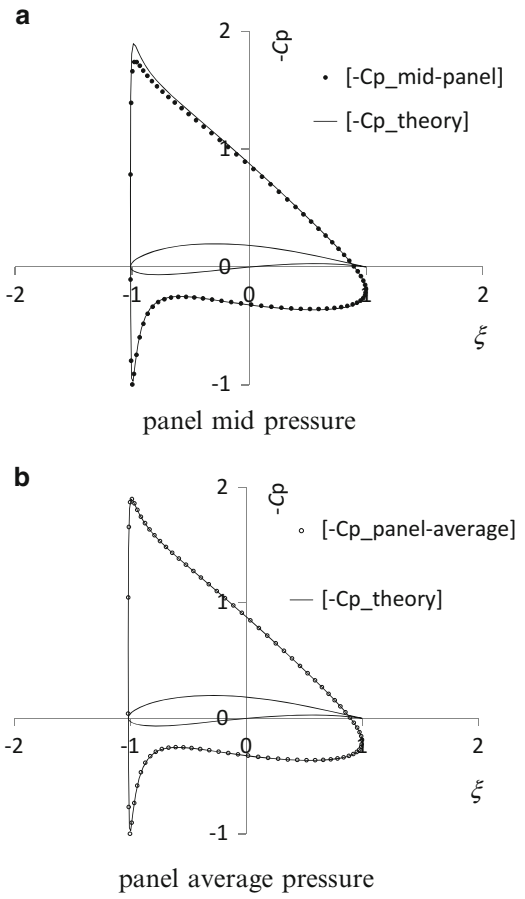


Fig. 4.6 Joukowski airfoil $-C_p$ at $\alpha = 5°$ $z_0 = (-0.1, 0.1)$

the results where z_0 is the center of a base circle that crosses the real axis at $(1, 0)$. The theoretical lift coefficient is $C_l \approx 1.22$ at $\alpha = 5°$ and the present calculation gives $C_l \approx 1.22$. The moment coefficient around the origin becomes $C_{m0} \approx 0.163$ which is numerically obtained from the theoretical pressure distribution, while for the present method $C_{m0} \approx 0.161$. The peak pressure is slightly lower for the panel mid value of the present method (Fig. 4.6a). This is due to the finite numbers of panels. The same is true for a circular cylinder which is approximated as a polygon as mentioned above. However, the panel average pressure is almost identical with that of the conformal mapping theory (Fig. 4.6b).

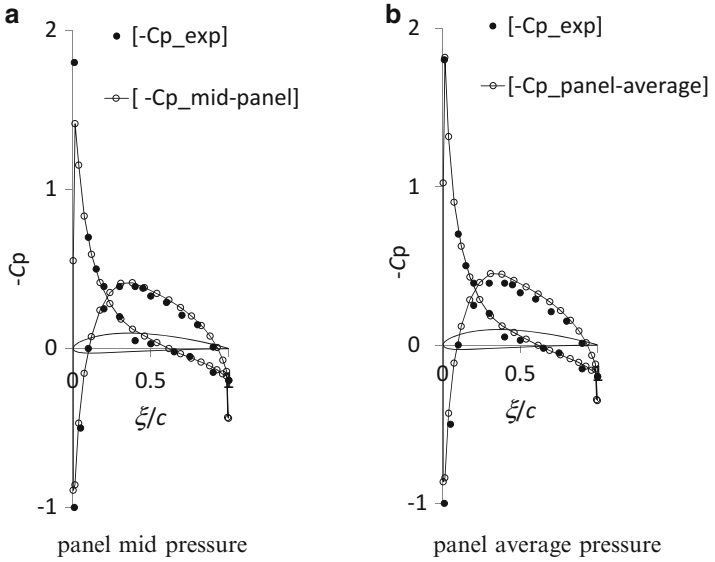


Fig. 4.7 NACA4412 airfoil ($N = 40$ panels) [9]

4.3.3 NACA4412 Airfoil

Figure 4.7a, b show the comparison between the present Schwartz-Christoffel panel method with $N = 40$ and NACA experiment at $\alpha = -4°$ [9]. The calculations are satisfactory in spite of the potential flow assumption. The panel average pressure well simulates the peak value.

4.4 Application to Unique Airfoils

4.4.1 FX63-137, M06-13-128 and S1223 Airfoils

Selig and Guglielmo [10] study FX63-137, M06-13-128 and S1223 airfoils and the present method is applied to these airfoils by using the given coordinates [11].

Figure 4.8 shows the magnitude of the airfoil panel average surface velocity for FX63-137 ($N = 96$) and M06-13-128 ($N = 63$). These results are identical with that of Selig and Guglielmo [10].

A S1223 Gurney flap of 1 % chord is modeled as a single panel with two surfaces in Fig. 4.9 (panel average velocity, $N = 80 + 2$). The stagnation is on the tip of the flap. It is visible that the Schwartz-Christoffel transformation restores the original airfoil coordinates very accurately, for example $\varepsilon \approx 10^{-7}$, after proper iterations. The Gurney flap increases the rear loading in the calculation and this is observed

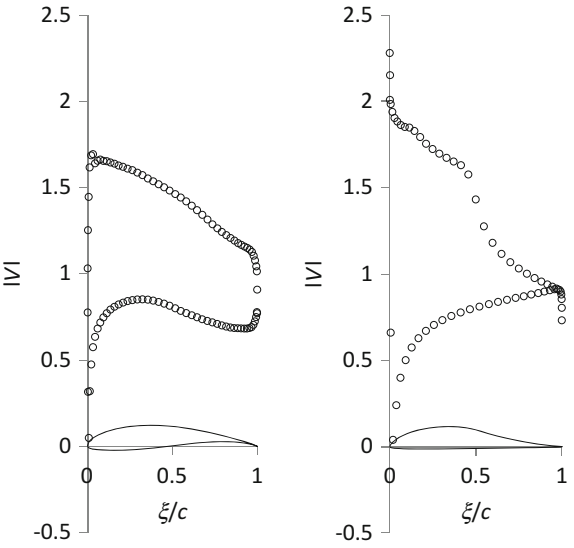
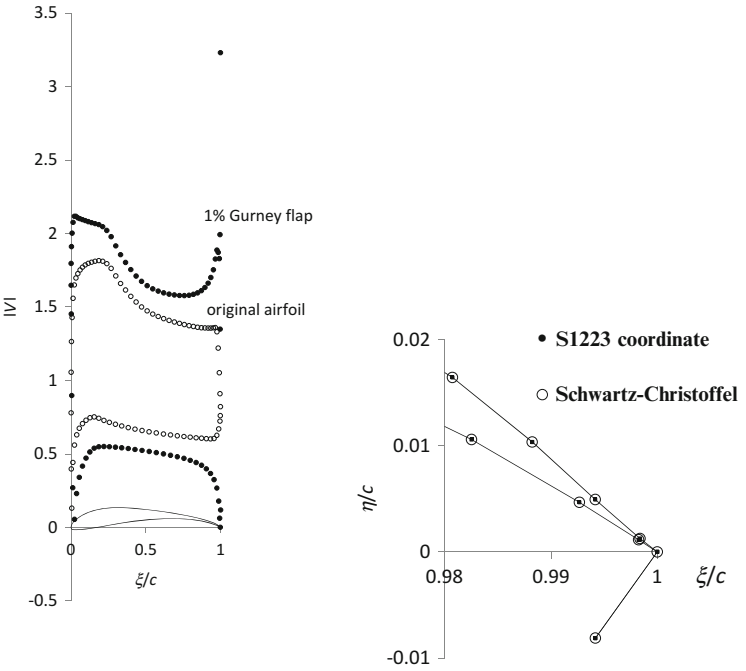


Fig. 4.8 FX63-137 (*left*) and M06-13-128 (*right*) airfoils at $C_l \approx 1.5$ (cf. [10])



Airfoil surface velocity at $C_l \approx 1.95 \circ / 2.99 \bullet$ 1% chord Gurney flap

Fig. 4.9 S1223 airfoil and Gurney flap (cf. [10])

in the experiment [2, 12]. A very sharp negative pressure is observed on the airfoil upper surface just prior to the Gurney flap. The potential flow calculation gives much higher lift increase than that of the experimental results; $\Delta C_l \approx 1$ by the present calculation vs. $\Delta C_l \approx 0.2$ by the experiment at $Re \approx 2 \times 10^5$ [10], both with the Gurney flap.

4.4.2 *Propeller Airfoil for Mars Exploration Airplane*

Figure 4.10 shows possible propeller airfoils for Mars exploration airplane, i.e. for low Reynolds number flows proposed and studied by Yonezawa et al. [13]. They are named airfoil A, B, C and D, respectively. The maximum thickness is 5 % chord for A, B and C airfoils, and 5.3 % for D airfoil at 30 % chord for all airfoils. Airfoil A is a triangle. Airfoil B is a polygon with four apexes and has trailing edge thickness of 2.5 % chord. Airfoil C is also a polygon with four apexes and has the same thickness of 5 % chord up to the trailing edge after the maximum thickness is reached. Airfoil D has thickness of 2.44 % chord at 30 % chord, the half of the lower surface consists of two straight lines, and the latter half of the lower surface is a circular arc of radius 50 % chord. Details are given in [13].

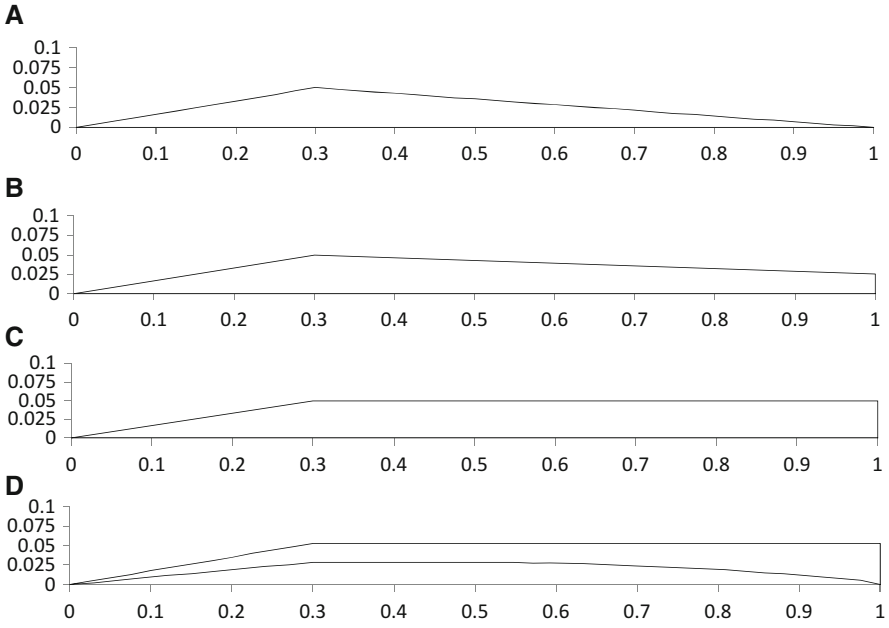


Fig. 4.10 Propeller airfoils for Mars exploration (**A**, **B**, **C** and **D** from *top*) [13]

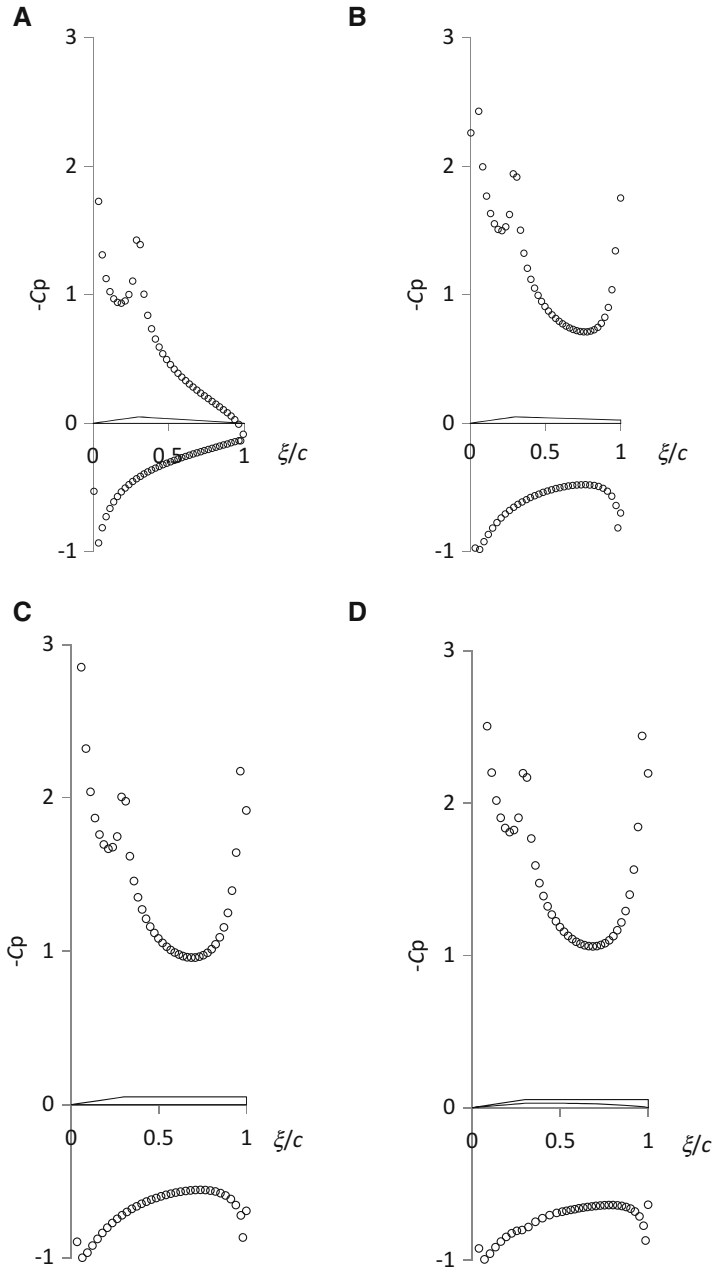


Fig. 4.11 (**A**) (*upper left*), (**B**) (*upper right*), (**C**) (*lower left*) and (**D**) (*lower right*) airfoils at $\alpha = 8^{\circ}$

Figure 4.11 shows the present results for the research propeller airfoils for Mars exploration (panel mid pressure). It is assumed that the rear stagnation point locates at (1, 0) even for the finite trailing edge thickness. Calculated theoretical lift coefficients are $C_l \approx 1.07$ for A, $C_l \approx 0.96$ for B, $C_l \approx 1.15$ for C and $C_l \approx 1.26$ for D, respectively. Airfoil D gives the highest lift. The experimental results are $C_l \approx 0.74$ at $Re \approx 3300$ for A and $C_l \approx 0.86$ at $Re \approx 4000$ for D at the same $\alpha = 8°$. The experimental results show that not only the lift coefficient but also the lift drag ratio of airfoil D are better than those of airfoil A. The potential calculation for airfoil D predicts higher pressure along the lower surface and lower pressure along the upper surface both towards the trailing edge. This is primarily because the Kutta condition is given at the lower edge of the airfoil and this has a similar effect as that of the Gurney flap.

4.4.3 *Dragonfly Airfoil*

Dragonfly has a unique airfoil section. Figure 4.12 shows an idealized dragonfly airfoil [14, 15]. The airfoil coordinates are derived from a patent figure [15]. The thickness of the airfoil is zero and has several peaks and valleys. The round symbols show the nodal points for the calculation. There are 90 panels with 45 panels each on the upper and lower surfaces, respectively. It is argued that the airfoil is effective in the low Reynolds number region.

Figure 4.13 shows the panel mid pressure distribution on the airfoil surface at the angle of attack 5°. Positive loading is observed for the convex while negative loading is visible for the concave. The pressure distribution looks very irregular.

Although the potential analysis is not necessarily important for the dragonfly airfoils, it is worthwhile to apply the present method to them because it gives the theoretical solution otherwise impossible. In the real viscous flow, it is said that vortices occupy the valleys and the airfoil is effectively streamlined. The up-to-date modern viscous and high-speed fluid dynamic calculations and applications are underway to solve the real life problems more accurately. See [16–18].

The calculated lift coefficient becomes $C_l \approx 0.37$ and the center of the pressure locates around at $x_{cp}/c \approx 0.14$ at $\alpha = 5°$. The lift coefficient is smaller than that of a flat plate and this is possibly due to the negative loading on the concaves. The

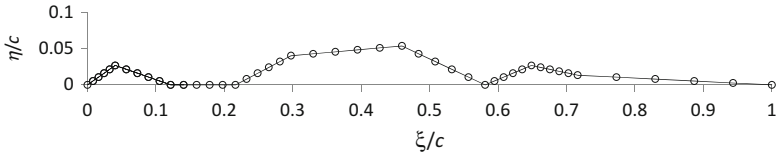


Fig. 4.12 Dragonfly airfoil and nodes [14, 15]

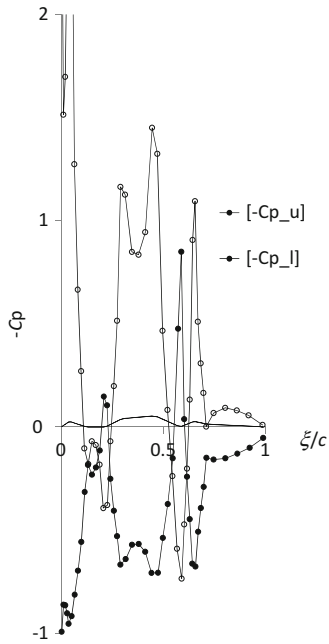


Fig. 4.13 Dragonfly airfoil pressure distribution at $\alpha = 5°$

concave corners must be the stagnation points and the convex peaks have infinite velocities although the lines in Fig. 4.13 are continuously drawn.

4.5 Concluding Remarks

We apply the Schwartz-Christoffel transformation to general and unique airfoils.

The solution procedure of the Schwartz-Christoffel panel method is improved and a simpler numerical procedure is proposed. First, a given airfoil is approximated as a polygon, and the each panel lengths are calculated together with the outer angles of the polygon apexes. A unit circle is deformed to a polygon, i.e. to an airfoil by assuming the angular position of the apexes on a unit circle by the Schwartz-Christoffel transformation. With the numerical integration, the each panel length of a polygonal airfoil is calculated. Then the corresponding angle on the unit circle is reduced if the panel length exceeds the desired given length and vice versa. After iterations, the original polygonal airfoil geometry is restored.

The method is tested for a circular cylinder and the Joukowski airfoil, and the results are quite satisfactory. It is also examined that the theoretical results for regular polygons give the exact difference between the continuous and polygonal geometries. For a convex surface, the panel method always underestimates the magnitude of the pressure; $-C_{\mathrm{p}}$.is smaller than the theoretical value for the upper

suction surface. However, the panel average pressure is very close to the theoretical value even for a convex surface.

The Schwartz-Christoffel panel method is the applied to several unique airfoils FX63-137, M06-13-128 and S1223. The present solutions give the same values as those by Selig and Guglielmo. The Gurney flap is tested numerically for S1223 and the rear loading increases theoretically although the potential calculation gives much higher lift than that of the experiment. The propeller airfoils for Mars exploration are also studied and it is found that the thick trailing edge may have a favourable effect which is consistent with the experiments. The dragonfly airfoil is also calculated although it is used in the low Reynolds number regions.

In spite of the fact that the present method involves the numerical integration of the panel length, the Schwartz-Christoffel panel method gives the theoretical results for polygons, i.e. approximated airfoils because it is based on the conformal mapping. The method is particularly useful for thin airfoils which are challenging for the conventional panel methods.

References

1. Shivamoggi, B.K.: Theoretical fluid dynamics, Wiley, 53(1998)
2. Morishita, E.: Schwartz-Christoffel panel method. Trans. Japan Soc. Aero. Space. Sci. **47**, 153–157 (2004)
3. Tian, Z.W., Wu, Z.N.: A study of two-dimensional flow past regular polygons via conformal mapping. J. Fluid Mech. 628, 121–154 (2009)
4. Morishita, E, Schwartz-Christoffel.: Transformation applied to polygons and airfoils, The 2014 International Conference Applied Mathematics, Computational Science & Engineering (AMCSE 2014), Varna, Bulgaria, September 13–15, (2014)
5. Moran, J.: An introduction to theoretical and computational aerodynamics. Wiley, 286 (1984)
6. Tanaka, S., Murata, S., Kurata, K.: Computation of the potential flow through cascades using the conformal mapping and the singularity method, JSME Int. J. Series II, 34, 423–430 (1991)
7. Moriya, T.: Introduction to aerodynamics (in Japanese), Baifu-kan, 139 (1977)
8. Imai, I.: Fluid dynamics and complex analysis (in Japanese), Nihon Hyoron-sha, 174–175, Tokyo (1981)
9. Pinkerton, R.M.: Calculated and measured pressure distribution over the mid-span section of the N.A.C.A. 4412 airfoil, NACA-TR-563, 367–380 (1937)
10. Selig, M.S., Guglielmo, J.J.: High-lift low Reynolds number airfoil design. J Aircraft **34**, 72–79 (1997)
11. http://m-selig.ae.illinois.edu/ads/coord_database.html
12. Li, Y., Wang, J., Zhang, P.: Effects of gurney flap on a NACA0012 airfoil. Flow Turbul. Combust. **68**, 27–39 (2002)
13. Yonezawa, K., Goto, Y., Sunada, S., Hayashida, T., Suwa, T., Sakai, N., Nagai, H., Asai, K., Tsujimoto, Y.: An investigation of airfoils for development of a propeller of mars exploration airplane (in Japanese). J Japan Soc. Aero. Space. Sci. **62**, 24–30 (2014)
14. Okamoto, M., Yasuda, K., Azuma, A.: Aerodynamic characteristics of the wing and body of a dragonfly. J. Exp. Biol. **199**, 281–294 (1996)
15. Ikeda, K.: http://www.google.com/patents/WO2014030465A1?cl=en (2014)
16. Mainum, A., Jamei, S., Priyanto, A. and Azwadi, N.: Aerodynamics characteristics of WIG catamaran vehicle during ground effect. WSEAS Trans. Fluid Mech. 3(5), 196–205 (2010)

17. Bernardini, C., Carnevate, M., Salvadori, S., Martelli, F.: On the assessment of an unstructured finite-volume DES/LES solver for turbomachinery applications, WSEAS Trans. Fluid Mech. 3(6), 160–173 (2011)
18. Maciel, E.S.D.G.,: A review of some numerical methods to the Euler Equations in two-dimensions, WSEAS Trans. Fluid Mech. 3(7), 81–95 (2012)

Chapter 5
Mining Latent Attributes in Neighborhood for Recommender Systems

Na Chang and Takao Terano

Abstract Neighborhood-based collaborative filtering (CF) algorithms have been extensively studied and discussed. In the traditional way of these methods, user-based CF predicts a target user's preference for an item based on the integrated preference of the user's neighbors for the item, and item-based CF is based on the integrated preference of the user's preference for the item's neighbors. Both the two ways underestimate the effect of structure of the target user or item's neighbors. That is, for instance, these neighbors may form two distinct groups: some neighbors like the target item or give high ratings; on the other hand, some neighbors dislike the target item or give low ratings. The difference between the two groups may influence user's choice. As an extension of neighborhood-based collaborative filtering, this paper focuses on the analysis of such structure by mining latent attributes of users or items' neighborhood, and corresponding correlations with users' preference by several popular data mining techniques. Mining latent attributes and experiment evaluation were conducted on MovieLens dataset. The experimental results reveal that the proposed method can improve the performance of pure user-based and item-based collaborative filtering algorithm.

5.1 Introduction

With the development of Internet and E-commerce, recommender systems have already been adopted by most web-sites, and play important roles in helping users find the most interesting products, such as books, articles, webpages, movies, and valuable information [1]. As one of the most successful approaches in building recommender systems, neighborhood-based CF methods are centered on computing the relationships between users or, alternatively, between items [2]. User-based CF approach evaluates the preference of a user to an item based on ratings of similar users; on the other hand, an item-based approach evaluates the preference of a user to an item based on ratings of similar items by the same user [3].

N. Chang (✉) • T. Terano
Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259, Nagatsuta-cho, Midori-ku, Yokohama 226-8503, Japan
e-mail: changna@trn.dis.titech.ac.jp; terano@dis.titech.ac.jp

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_5

As the definition above, user-based CF method predicts a user's preference for an item based on the integrated preference of the neighbors' preference, ignoring the fact that these neighbors may have opposite preference, i.e. the neighbors form two groups based on their preference. One group contains users who like the target item or give high rating, and we call it like-mind group. On the other hand, the other group contains users who dislike the target item or give low rating, and we call it dislike-mind group. For instance, the effect of the size of each group is underestimated by traditional user-based CF method. If the number of like-mind group is greater than that of dislike-mind group, the target user may tend to take the integrated preference of like-mind group which contains majority users. These factors, which are not considered in tradition user-based CF method, may actually play important roles in the prediction of target user's preference.

Thus, in this paper, we focus on the study of latent attributes of user or items' neighborhood. That is, mining latent attributes in users neighborhood (user-MLAN) and mining latent attributes in items neighborhood (item-MLAN). Specifically, taking user-MLAN for example, firstly we try to discover several useful and meaningful latent attributes, such as group size difference, average similarity difference, and standard deviation difference in user-MLAN systems. Then, corresponding correlations between users' preference and these attributes by several popular data mining techniques are suggested. Finally, we use real data set to evaluate the proposed approach and show the performance.

The rest of this paper is organized as follows: Sect. 5.2 give the related work and background of this paper. Section 5.3 shows the discovery of mining latent factors. Section 5.4 introduces the strategy of the proposed approach and several popular data mining techniques could be used. Section 5.5 presents the experiment evaluation and according results. Section 5.6 gives the final remarks.

5.2 Related Work

User-based CF method is a basic approach and has been used in many recommender systems. Since the basic intuition of user-based CF method is that predictions for a user is based on the preference patterns of other user who have similar interests. Therefore, the first step of user-based CF is to find most similar neighbors through measuring the similarity of users. Users similarity can be measured by Pearson correlation or other correlated similarities [4, 5]. For example, Pearson correlation between user u and v is as follows:

$$s_{uv} = \frac{\sum_{i \in I} (r_{ui} - \overline{r}_u)(r_{vi} - \overline{r}_v)}{\sqrt{\sum_{i \in I} (r_{ui} - \overline{r}_u)^2} \sqrt{\sum_{i \in I} (r_{vi} - \overline{r}_v)^2}} \tag{5.1}$$

where $\overline{r}_u$ and $\overline{r}_v$ mean the average rating of user u and v. The second step is to obtain preference model. The basic model is as follows:

$$\widehat{r}_{ui} = \overline{r}_u + \frac{\sum_{v \in U} (r_{vi} - \overline{r}_v) \cdot s_{uv}}{\sum_{v \in U} s_{uv}} \tag{5.2}$$

As a traditional and basic method, user-based CF can be simply extended by incorporating data mining and machine learning techniques. For instance, reference [6] proposed a novel framework which combines user-based CF and clustering. Clusters generated from the training data provide the basis for data smoothing and neighbor selection. Reference [7] incorporated Tag Transfer Learning in user-based CF by transferring tag topics. Reference [8] proposed to combine user-based CF and latent factor models together, and take into account the number of neighbors in the prediction model.

Most of the current research on user-based CF mainly focuses on the explicit attributes, such as user-item-rating matrix, the number of users, the number of items, similarity, average rating of a user and so on. There may be latent attributes which can influence the preference prediction implicitly. Under this guideline, we attempt to mine several useful latent attributes in user-based CF method.

5.3 Mining Latent Attributes

5.3.1 Problem Definition

Consider that there are m users and n items in a recommender system. We can define the set of users U as the set of integers $\{1, 2 \ldots m\}$ and the set of items I as the set of integers $\{1, 2 \ldots n\}$. The ratings of users for items r_{ui} are stored in a $m \times n$ rating matrix $R = \{r_{ui} | 1 \leq u \leq m; 1 \leq i \leq n, 0 \leq r_{ui} \leq r\}$. In addition to specific ratings, we also use binary measurement, 1 and -1, indicating user's preference, i.e.

$$o_{ui} = \begin{cases} 1, & if \ r_{ui} \geq \lambda \\ -1, & if \ 0 < r_{ui} < \lambda \end{cases} \tag{5.3}$$

where $0 < \lambda \leq r$, and r denotes the maximum ratings. In the definition, $o_{ui} = 1$ indicates user u like item i, while $o_{ui} = -1$ indicates user u dislike item i.

In the case of user-MLAN model, we define a target user u's neighbors who have rated item i as $N(u; i)$. As discussed in Sect. 5.1, the target user's neighbors can be distinguished into two groups: like-mind group $N^+(u; i)$ including users who like the target item, and dislike-mind group $N^-(u; i)$ including users who dislike the target item, and the size of each group as $\left|N^+(u; i)\right|$ and $|N^-(u; i)|$. This paper argues that the difference between like-mind group and dislike-mind group may influence the target user's preference. For instance, if there are 20 neighbors, who like the target item, and 5 neighbors, who dislike the target item, among a target user u's neighbors, u may be likely to take the majority users' opinion. This example

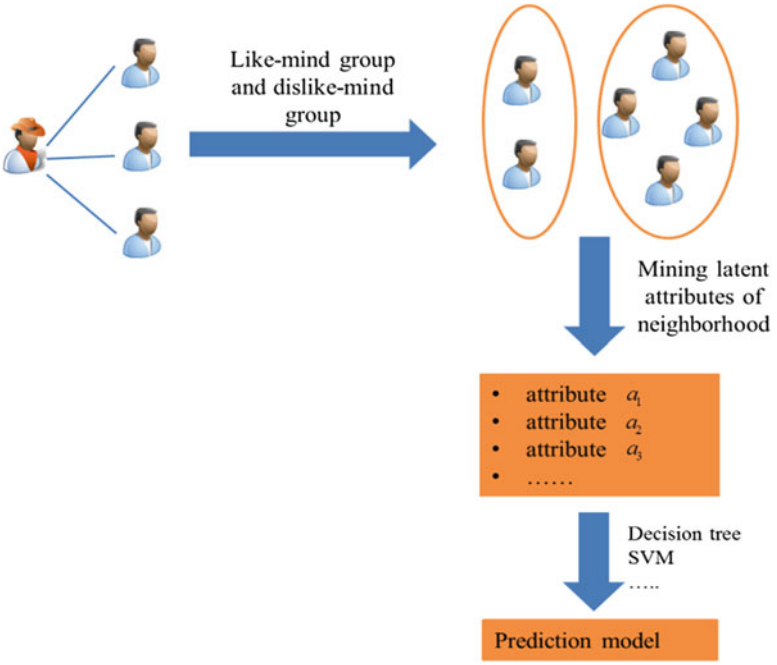


Fig. 5.1 Framework of user-MLAN

shows that the important effect of group size. There may be other latent attributes in the neighborhood, which can decide the target user's preference and choice.

Thus, the aim of the proposed user-MLAN algorithm is to find useful and meaningful latent attributes between the like-mind group and dislike-mind group, and predict whether the target user will like take the integrated preference of the like-mind group or dislike the target item. The framework of user-MLAN is as Fig. 5.1 shows.

Based on this idea, we assume that there is a set of latent attributes $A = \{a^1, a^2, \cdots, a^l\}$ for arbitrary users' neighbors related to item i, and then the prediction becomes a classification problem, i.e. each observed user u's latent attributes of neighborhood consists of a pair: a attributes vector $\mathbf{a}_{ui} = \{a^1_{ui}, a^2_{ui}, \cdots a^l_{ui}\}$, where $\mathbf{a}_{ui} \in R^l$, and the associated label y_{ui},where $y_{ui} \in \{1, -1\}$. The classification task is to learn the mapping $\boldsymbol{\alpha}_{ui} \mapsto y_{ui}$. In the next section, we define several related latent attributes of user neighborhood.

In the case of item-MLAN model, we define a target item i's neighbors which have been rated by the target user u as $N(i;u)$. Like user-MLAN model, the item's neighbors can be distinguished into like-mind group $N^+(i;u)$ including items which are liked by the target user u and dislike-mind group $N^-(i;u)$ including items which are disliked by the target user u. Also the preference prediction for item-MLAN can be seen as a classification problem, by switching the roles of users and items.

5.3.2 *Latent Attributes in User-MLAN Model*

In reference [9], the authors list 24 potential attributes of user-item-rating matrix, such as number of ratings, degree of agreement with others, average of user's rating, standard deviation in user rating and so on. But clearly these attributes are just the statistical features of the matrix. In this paper, we attempt to discovery several latent attributes of the neighborhood which have correlations with user's preference.

In this paper, we use Movielens[1] data set to discover useful and meaningful attributes for user-MLAN. The data set contains 100,000 records (user-item-rating) with rating scale 1–5 and consists of 943 users and 1,682 movies. The basic attribute selection rule is based on [10], which indicates that the attributes should be highly correlated with the class (preference) and uncorrelated with each other. After several trials, we have the following latent attributes in users' preference prediction:

- Group size difference (GSD): the difference of size of like-mind group and dislike-mind group. This attribute measures the difference of two groups in terms of neighbor number. The computation is as Eq. (5.4) shows. If GSD > 0, it indicates that the like-mind group has more users who like the target item than that of dislike-mind group, and vice versa.

$$GSD = \frac{\left|N^+(u;i)\right| - \left|N^-(u;i)\right|}{\left|N^+(u;i)\right| + \left|N^-(u;i)\right|} \tag{5.4}$$

- Average similarity difference (ASD): the difference of average similarity between like-mind group and dislike-mind group with a target user. This attribute measures difference of two groups in terms of the average similarity. The computation is as Eq. (5.5) shows. If ASD > 0 indicates that the like-mind group is much more similar to the target user than that of dislike-mind group on average, and vice versa.

$$ASD = \frac{\sum_{j \in N^+(i;u)} s_{ij}}{N^+(i;u)} - \frac{\sum_{j \in N^-(i;u)} s_{ij}}{N^-(i;u)} \tag{5.5}$$

- Standard Deviation Difference (SDD): the difference of standard deviation between like-mind group and dislike-mind group. This attribute measures difference of ratings in two groups in terms of standard deviation. SDD > 0 indicates that the ratings in like-mind group are closer than that of dislike-mind group, and vice versa. In the following computation for SDD, $\overline{r_i^+}$ and $\overline{r_i^-}$ indicate the average rating of like-mind and dislike-mind group for item i respectively.

[1] https://movielens.org/

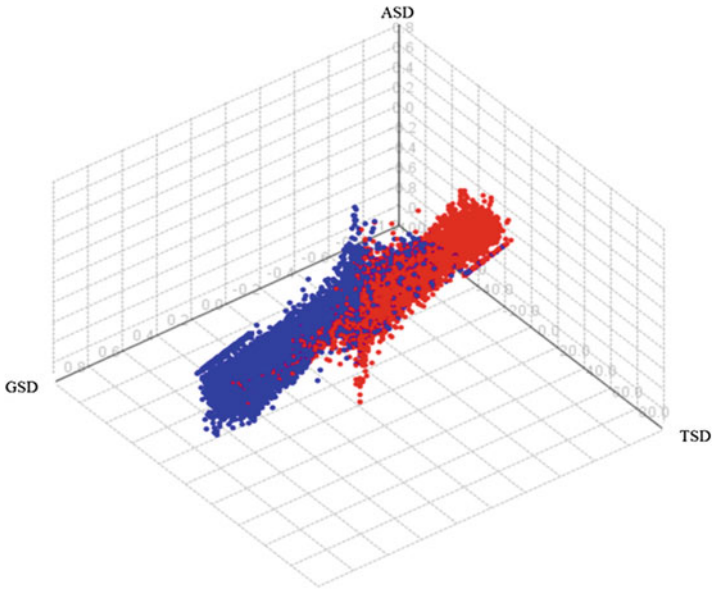


Fig. 5.2 Correlations of proposed latent attributes and preferences

$$SDD = \sqrt{\frac{\sum_{v \in N^+(u;i)} \left(r_{vi} - \overline{r}_i^+\right)^2}{|N^+(u;i)|}} - \sqrt{\frac{\sum_{v \in N^-(u;i)} \left(r_{vi} - \overline{r}_i^-\right)^2}{|N^-(u;i)|}} \tag{5.6}$$

The correlation of these attributes and users' preference tendency is as Fig. 5.2 shows. The blue points mean users would like to take the opinion of like-mind group, and the red points mean users would like to take the opinion of dislike-mind group. From this figure, we can see that the two classes do not have distinct boundaries, but still are separated into two parts. This indicates that there are correlations between these latent attributes and preferences.

Based on these findings, we can see that users' preference is related to these three attributes: group size difference, total similarity difference and average similarity difference. And we can have the following conclusions:

- Users' preference may be related to the structure of neighborhood;
- Users' preference is related to neighbors' group size difference, total similarity difference and average similarity difference.
- Given threshold for those attributes, the boundary of like-mind group and dislike-mind group is clear.

5.3.3 Latent Attributes in Item-MLAN Model

We use the same data sets as user-MLAN model to conduct experiments, and after several trials, we have the following latent attributes in items' neighborhood:

- Group Size Difference (GSD): the difference of size of persuasive group and supportive group. This attribute measures the difference of two groups in terms of neighbors' number. If GSD > 0, it indicates that the persuasive group contains more items who are liked by the target user than that of supportive group, and vice versa.

$$GSD = \frac{\left|N^+(u;i)\right| - \left|N^-(u;i)\right|}{\left|N^+(u;i)\right| + \left|N^-(u;i)\right|} \tag{5.7}$$

- Average Similarity Difference (ASD): the difference of average similarity of like-mind group and dislike-mind group to target item. This attribute measures the difference of closeness between two groups and the target item. If ASD > 0, the like-mind group is much closer to the target item than that of dislike-mind group, and vice versa.

$$ASD = \frac{\sum_{j \in N^+(i;u)} s_{ij}}{N^+(i;u)} - \frac{\sum_{j \in N^-(i;u)} s_{ij}}{N^-(i;u)} \tag{5.8}$$

- Average Popularity Difference (APD): the difference of average popularity between like-mind group and dislike-mind group. This attribute measures difference of two groups in terms of the average popularity. APD > 0 indicates that the like-mind group is much more popular than that of dislike-mind group on average, and vice versa. In the following computation for APD, $|Y_j|$ denotes the ratings received by item j.

$$APD = \frac{\sum_{j \in N^+(i;u)} \left|Y_j\right|}{\left|N^+(i;u)\right|} - \frac{\sum_{j \in N^-(i;u)} \left|Y_j\right|}{\left|N^-(i;u)\right|} \tag{5.9}$$

The correlation of these attributes and users' preference tendency is as Fig. 5.3 shows. The blue points mean users would like to take the opinion of like-mind group, and the red points mean users would like to take the opinion of dislike-mind group. From this figure, we can see that the two classes do not have distinct boundaries, but still are separated into two parts. This indicates that there are correlations between these latent attributes and preferences.

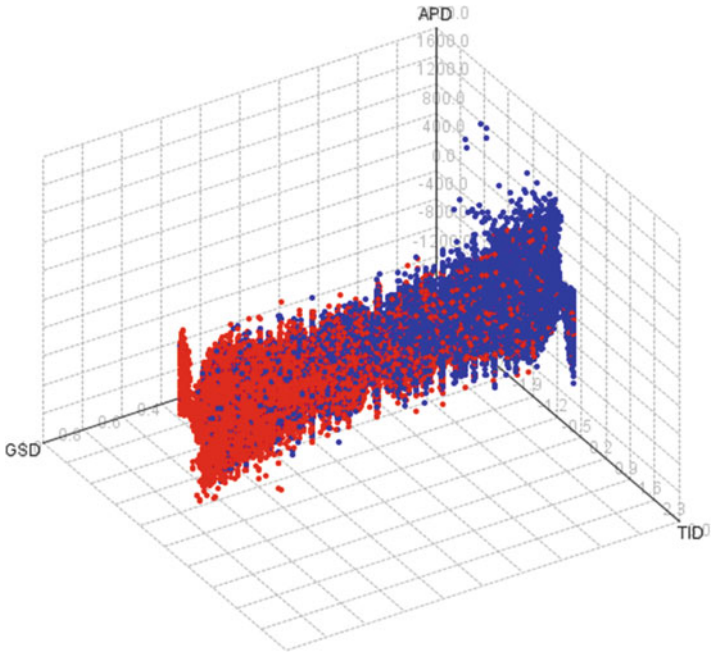


Fig. 5.3 Correlations of proposed latent attributes and preferences

5.4 Applying in Recommendations

In this section, we use four data mining and machine learning techniques, i.e. Decision Tree, Naïve Bayesian, Neural Network and Support Vector Machine, to learn models respectively for the prediction of users' preference based on the three latent attributes discussed in Sect. 5.4.

- Decision Tree: Decision trees [11, 12] are classifiers on a target attribute (or class) in the form of a tree structure. Each node corresponds to one of these attributes, and edges correspond to the possible values of the attributes. Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree.
- Naive Bayesian: a Bayesian classifier is a probabilistic framework which is based on the definition of conditional probability and the Bayes theorem [13]. In simple terms, a naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable.
- Neural Networks: An Artificial Neural Network (ANN) [14] is an assembly of inter-connected nodes and weighted links that is inspired in the architecture of the biological brain. Nodes in an ANN are called neurons as an analogy with

biological neurons. These simple functional units are composed into networks that have the ability to learn a classification problem after they are trained with sufficient data [15, 16].

- Support Vector Machine: the goal of a SVM classifier [17] is to find a linear hyperplane (decision boundary) that separates the data in such a way that the margin is maximized [18]. SVM can be used for both regression and classification problems. For classification problems, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

5.5 Experiments with MovieLens Dataset

In this section, we use the data set from MovieLens to verify our research (see details in Sect. 5.3). Firstly, we investigate the attribute correlations for user-MLAN and item-MLAN model. The correlations are computed by Pearson coefficient. The experiment results are as Tables 5.1 and 5.2 shows.

From the tables, we can see that the correlations between attributes are relatively independent to each other; meanwhile, the correlations between attributes and class are dependent.

Then, we use fivefold cross validation by randomly selecting 80 % of the data set as training data and 20 % of it as test data to measure the performance accuracy [19]. The computation of accuracy is as Fig. 5.4 show.

Table 5.1 The attribute correlations for user-MLAN

	GSD	TSD	ASD	Preference
GSD	1.000	0.045	0.024	0.429
TSD		1.000	0.008	0.114
ASD			1.000	0.115

Table 5.2 The attribute correlation for item-MLAN

	GSD	ASD	APD	Preference
GSD	1.000	0.012	0.028	0.641
ASD		1.000	0.014	0.312
APD			1.000	0.145

Fig. 5.4 Computation of accuracy

	Condition	
Test outcome	True positive (tp)	False positive (fp)
	False negative (fn)	True negative (tn)

$$\text{Accuracy} = \frac{\#tp + \#tn}{\#tp + \#fp + \#fn + \#tn}$$

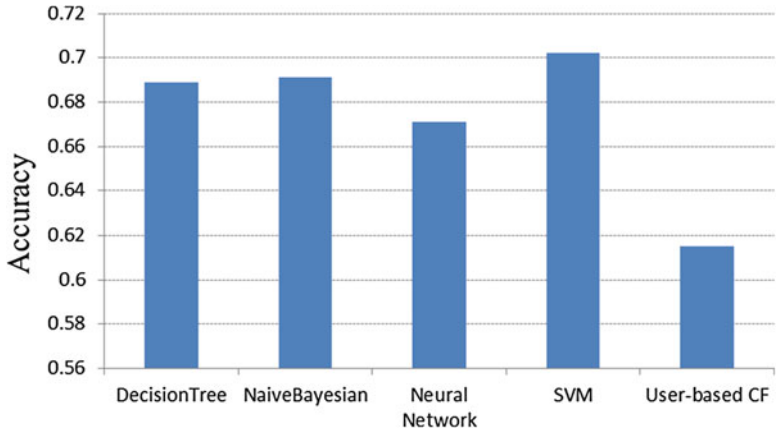


Fig. 5.5 Experiment results for user-MLAN

In the experiment, we set the threshold for GSD, TSD, ASD is 0.1, 0, 0 in user-MLAN model. Figure 5.5 shows the experimental result. From the figure, it is clear that proposed method, which learning users' preference by mining latent attribute of neighborhood, outperforms pure user-based CF method in terms of accuracy. This verified that the proposed approach can give better performance than traditional user-based collaborative filtering algorithm and the three latent attributes have correlations with users' preference. Moreover, among the four machine learning techniques which are used to learn preference model based on latent attributes of neighborhood, we can see that SVM gives best performance, Neural networks shows the worst performance in terms of accuracy.

In case of item-MLAN model, we set the threshold for GSD, TSD, APD is 0.2, 0, 0. Figure 5.6 shows the experiment result. Similarly, we can see that the proposed item-MLAN can give better performance that tradition item-based CF method. And SVM gives best performance, neural networks shows the worst performance in terms of accuracy.

Furthermore, we can also compare the performance of user-MLAN and item-MLAN by both Figs. 5.5 and 5.6. Clearly, item-MLAN outperforms user-MLAN. This indicates that item-MLAN gives better performance than user-MLAN.

5.6 Conclusions

In this paper, we introduced a new idea for neighborhood-based CF approach by mining latent attributes of user or item's neighborhood. Specifically, we firstly introduced the importance of latent attributes in neighborhood-based CF systems, then conducted experiment to discover several meaningful latent attributes: group size difference, average similarity difference, standard deviation difference for

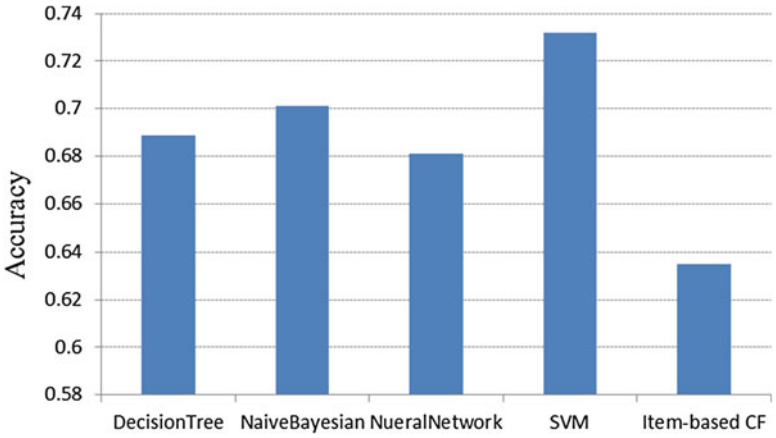


Fig. 5.6 Experiment results for item-MLAN

user-MLAN, and group size difference, average similarity difference, average popularity difference for item-MLAN. In addition, we proposed that the prediction model can be learned by Decision Tree, Naïve Bayesian, Neural Networks and Support Vector machine. Finally, we verified that the proposed approach working better in terms of accuracy on MovieLens dataset.

References

1. Melville, P., Vikas, S.: Recommender systems. J. Encycl. Mach. Learn. 829–838 (2010)
2. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2000)
3. Xu, J., Johnson-Wahrmann K., Li, S.: The development, status and trends of recommender systems: a comprehensive and critical literature review. In: Proceedings of International Conference Mathematics and Computers in Science and Industry. 117–122 (2014)
4. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. J. Adv. Artif. Intell. (2009)
5. Kularbphettong, K., Somngam, S., Tongsiri, C., Roonrakwit, P.: A Recommender System using Collaborative Filtering and K-mean based on Android Application. In: Proceedings of International Conference Applied Mathematics, Computational Science and Engineering. 161–166 (2014)
6. Xue, G., Lin, C., Yang,Q., Xi, W., Zeng, H., Yu, Y.: Scalable collaborative filtering using cluster-based smoothing. SIGIR'05 August. 15–19 (2005)
7. Wang, W., Chen, Z., Liu, J., Qi Q., Zhao, Z.: User-based Collaborative filtering on cross domain by Tag transfer learning. In: ACM KDD'18 August 12–16 (2012)
8. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Proceedings of the 14th ACM SIGKDD Conference. 426–434 (2008)

9. Morid, M.A., Shajari, M., Golpayegni, A.H.: Who are the most influential users in a recommender system? In: Proceedings of the 13th International Conference on Electronic Commerce, ACM, New York (2012)
10. Hall, M.A.: Correlation-based feature selection for machine learning. Doctoral dissertation, The University of Waikato (1999)
11. Quinlan, J.R.: Induction of decision trees. J. Mach. Learn. **1**(1), 81–106 (1986)
12. Rokach, L.: Data mining with decision trees: theory and applications. J. World Scientific. 69 (2008)
13. Miyahara, K., Pazzani, M.J.: Collaborative filtering with the simple bayesian classifier. In: Pacific Rim International Conference on Artificial Intelligence (2000)
14. Zurada, J.: Introduction to artificial neural systems. West Publishing, St. Paul (1992)
15. Berka, T. Behrendt, W., Gams, E.: A trail based internet-domain recommender system using artificial neural networks. In: Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (2002)
16. Kongsakun, K., Fung, C.C.: Neural Network Modeling for an Intelligent Recommendation System Supporting SRM for Universities in Thailand. WSEAS Trans. Comput. **11**(2), 34–44 (2012)
17. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other Kernel-based learning methods. Cambridge University Press, New York (2000)
18. Kang, H. Yoo, S.: SVM and collaborative filtering-based prediction of user preference for digital fashion recommendation systems. IEICE Trans. Inf. Syst. (2007)
19. Shani, G., Gunawardana, A.: Evaluating recommendation systems. Recommender systems handbook, pp. 257–297. Springer, New York (2011)

Chapter 6
An Assessment of the Effect of Varying Popov's Parameter on the Region of Robust Absolute Stability of Nonlinear Impulsive Control Systems with Parametric Uncertainty

Tseligorov Nikolai, Tseligorova Elena, and Mafura Gabriel

Abstract This paper focuses on finding and the assessment of the variation of the region of robust absolute stability of an impulse control system, with monotonous nonlinearities, as Popov's parameter varies. A mathematical model of a nonlinear impulsive control system (NICS) is considered. The criterion for absolute stability on the equilibrium position for NICS, with monotonous nonlinearities, can be expressed as a polynomial expression. The robust stability of NICS is tested using Kharitonov's theorem and a modified root locus method for interval transfer functions. A graphical illustration of roots of the characteristic equation, which have been gotten from the interval transfer function, on the complex plane is used in the assessment of the stability of control system. To evaluate the effect of Popov's parameter, a specially written program complex Stability is used. An illustrative example is given to demonstrate the effect of varying Popov's parameter on the region of absolute robust stability.

Keywords Absolute robust stability • Nonlinear impulsive system • Transfer function • Perturbed polynomial • Popov's parameter • Monotonous nonlinearity • Root locus

T. Nikolai
Rostov on Don's affilliate of Russian Customs Academy
Avenue Budenovsky D.20, 344000, Rostov on Don, Russia
e-mail: nzelig@rambler.ru

T. Elena
Don State Technical University
Gagarin's square D.1, 344000, Rostov on Don, Russia
e-mail: celena@yandex.ru

M. Gabriel (✉)
LLC Rostovgiproshaht
Street Krasnoarmeiskaya D.157, 344000, Rostov on Don, Russia
e-mail: mafurag@hotmail.com

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_6

6.1 Introduction

Today, the process of designing any technically complex system is accompanied by the creation of its mathematical model. The mathematical model allows the control engineer to get information about the stability of the plant [1, 2]. Mathematical models, which are based on the application of the criteria of absolute stability to the system under design, focused mainly on control systems with standard nonlinear characteristics, with theoretical forms of nonlinear characteristics, which differed considerably from actual real world systems. To get results which are closer to the real world conditions of exploitation, it is necessary to take into account the parametric uncertainty of the control system in question.

In practice, control systems used for various purposes will face various problems due to uncertainty. Uncertainty can be defined as the incompleteness or inaccuracy of information regarding a given control system during its mathematical modelling. Uncertainty can be caused by both internal or external factors. Uncertainty can affect the accuracy of a control system and sometimes even cause the control system to loose stability.

In control theory several types of uncertainties are identified: parametrical uncertainty [3–5], non-parametric uncertainty, non-stationary uncertainty, nonlinear uncertainty [5] and other types of uncertainties. The use of algebraic [5, 6] and graphical [5, 7, 8] methods, for testing various control systems, has become very popular. The main advantage of such methods is that they make it possible to apply interval methods. Among other beneficial features, interval methods do not depend on the knowledge of the probability characteristics of uncertainty or the concrete values of the parameters of the plant to be controlled since this values fall in a known interval. In response to the challenges caused by uncertainty, the programs and mathematical apparatus, for investigating such control systems, have been further advanced and developed. This has also led to the use of computer aided algebra systems (CAS) for manipulating symbolic data [6, 10]. In order to investigate the stability of NICS, with parametric uncertainty, a specially written program complex Stability [11] is used. This program complex can graphically illustrate the region of stability on the complex plane.

6.2 Statement of Problem

The Academic Ya.Z. Tsypkin proposed the criteria for absolute stability of nonlinear control systems (NICS) with monotonous characteristics in the form of the following inequalities [1]

$$Re[1 + q(1 - e^{-j\bar{\omega}})]W(j\bar{\omega}, 0) + k^{-1} > 0. \tag{6.1}$$

This inequality must be satisfied for all frequencies $\bar{\omega}$ within the interval $[0, \pi]$ for real V.M. Popov's parameters $q \geq 0$. The nonlinear elements (NE) $\phi(\sigma)$ fulfill the following condition:

$$0 \leq \frac{\phi(\sigma)}{\sigma} \leq k, \phi(0) = 0. \tag{6.2}$$

The above criterion (6.2) can be interpreted geometrically as plotting a modified amplitude-phase characteristics graph $\tilde{W}(j\bar{\omega}, 0)$ with Popov's line plotted on the $\tilde{W}(j\bar{\omega}, 0)$ plane [1]

$$U^*(j\bar{\omega}, 0) + qV^*(j\bar{\omega}, 0) + \frac{1}{k} \geq 0,$$

$$\text{where, } U^*(j\bar{\omega}, 0) = ReW^*(j\bar{\omega}, 0),$$

$$U^*(j\bar{\omega}, 0) = -Re[e^{-j\bar{\omega}} W * (j\bar{\omega}, 0)] + ReW * (j\bar{\omega}, 0).$$

The plotted line crosses through the point $-1/k$, on the real axis and at an angle of $arctg1/q$. A graphical illustration of the criterion of absolute stability is given above in Fig. 6.1.

However to the best of our knowledge, there is no literature which graphically illustrates the region of absolute robust stability and the effect of varying Popov's parameter on the size and shape of the region of stability.

It is necessary to evaluate the change in the region of absolute robust stability and define the effect Popov's parameter on the region of absolute robust stability of a control system under evaluation.

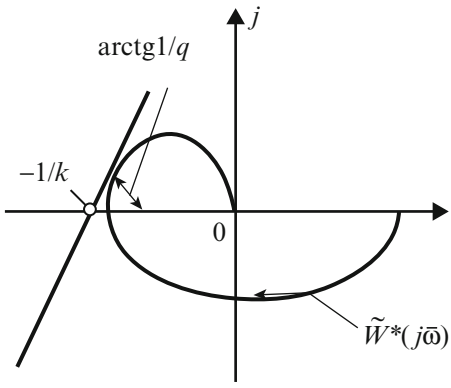


Fig. 6.1 Graphical interpretation of Tsypkin's criterion with Popov's line

6.3 Solution of Problem

6.3.1 Mathematical Model of Problem

Using w-transform, the test for absolute stability of NICS is reduced to testing if the resultant characteristic polynomial is Hurwitz [3]. Criterion (6.1) takes the following form in w-plane

$$Re[(1+q\frac{2w}{1+w})W(w)]+k^{-1}>0,$$

$$\text{or} \tag{6.3}$$

$$Re[(1+q\frac{2j\nu}{1+j\nu})W(j\nu)]+k^{-1}>0\forall\nu\in[0,\infty].$$

Where $w=j\nu, \nu=tg\frac{\omega T_0}{2}$—relative pseudo frequency, T_0—sampling interval.

If the transfer function's frequency response characteristics are presented as shown below

$$W(j\nu)=\frac{\alpha_1(\nu)+j\beta_1(\nu)}{\alpha_1(\nu)+j\beta_1(\nu).} \tag{6.4}$$

then after substituting (4) in (3) we get the following polynomial expression which corresponds to the criterion (3) [5].

$$k(\alpha_1(\nu)\alpha_2(\nu)+\beta_1(\nu)\beta_2(\nu))(1+\nu^2)+2q[(\alpha_1(\nu)\alpha_2(\nu)+\beta_1(\nu)\beta_2(\nu))\nu^2+$$
$$(\alpha_1(\nu)\beta_2(\nu)-\alpha_2(\nu)\beta_1(\nu))\nu]+[\alpha_2^2(\nu)+\beta_2^2(\nu)](1+(\nu)^2)=0. \tag{6.5}$$

Using criterial expression (6.5), we can easily get the polynomial expression in symbolic form for transfer functions of various orders. For a transfer function of the first order i.e. n = 1:

$$W(j\nu)=\frac{j\alpha_1(\nu)+\alpha_0}{j\beta_1(\nu)+\beta_0}$$

The characteristic expression can be written as follows:

$$P(x)|_{x=\nu^2}=kq(b_1(2a_1x^2+2a_0x)-2b_0a_1x+2a_0b_0x)$$
$$+k(a_1b_1(x^2+x)+a_0b_0(x+1))+b_1^2(x^2+x)+b_0^2(x+1). \tag{6.6}$$

For a transfer function of the second order i.e. n = 2:

$$W(j\nu)=\frac{-a_2v^2+ia_1v+a_0}{-b_2v^2+ib_1v+b_0}.$$

The characteristic expression in symbolic form can be written as follows:

$$\begin{aligned} P(x)|_{x=v^2} = k(b_2(a_2(x^3+x^2)+a_0(-x^2-x))+a_1b_1(x^2+x)+b_0a_2 \\ (-x^2-x)+a_0b_0(x+1))+kq(b_2,(2a_2x^3+2a_1x^2-2a_0x^2)+ \\ a_2(-2b_1x^2-2b_0x^2)+b_1(2a_1x^2+2a_0x)-2b_0a_1x+2a_0b_0x) \\ +b_2^2(x^3+x^2)+b_1^2(x^2+x)+b_0b_2(-2x^2-2x)+b_0^2(x+1). \end{aligned} \tag{6.7}$$

The above polynomial expressions (6.6) and (6.7) allow us to get the characteristic polynomial for testing the absolute stability without any intermediate calculations. Only the known values of coefficients in the numerator and denominator polynomials of the transfer functions are needed. The code fragment below shows an Excel macros, written in visual basic, which calculates the characteristic polynomial for transfer function of the third degree i.e. n = 3.

```
Sub processcoeff()
Dim a1, b1, a0, b0, a2, b2, a3, b3, k,
Dim coeff_x_4, coeff_x_3, coeff_x_2, coeff_x_1, coeff_x_0

a0 = Range("B1").Value
a1 = Range("B2").Value
a2 = Range("B3").Value
a3 = Range("B4").Value
b0 = Range("E1").Value
b1 = Range("E2").Value
b2 = Range("E3").Value
b3 = Range("E4").Value
k = Range("B6").Value
q = Range("D6").Value

coeff_x_4 = (2 * a3 * b3 * k * q + a3 * b3 * k + b3 ^ 2)
coeff_x_3 = 2 * a2 * b3 * k * q - 2 * a1 * b3 * k * q - 2 * b2 *
a3 * k * q - 2 * b1 * a3 * k * q + 2 * a2 * b2 * k * q + a3 * b3 * k
- a1 * b3 * k - b1 * a3 * k + a2 * b2 * k + b3 ^ 2 - 2 * b1 * b3 + b2 ^ 2
coeff_x_2 = -2 * a0 * b3 * k * q + 2 * b0 * a3 * k * q + 2 * a1 * b2
* k * q - 2 * b1 * a2 * k * q - 2 * b0 * a2 * k * q + 2 * a1 * b1 * k
* q - 2 * a0 * b1 * k * q - a1 * b3 * k - b1 * a3 * k + a2 * b2  * k
- a0 * b2 * k - b0 * a2 * k + a1 * b1 * k - 2 * b1 * b3 + b2 ^ 2 - 2
* b0 * b2 + b1 ^ 2
coeff_x_1 = 2 * a0 * b1 * k * q - 2 * b0 * a1 * k * q + 2 * a0 * b0 *
k * q - a0 * b2 * k - b0 * a2  * k + a1 * b1 * k + a0 * b0 * k - 2 *
b0 * b2 + b1 ^ 2 + b0 ^ 2
coeff_x_0 = a0 * b0 * k + b0 ^ 2
Range("A8").Value = coeff_x_4
Range("C8").Value = coeff_x_3
Range("E8").Value = coeff_x_2
Range("G8").Value = coeff_x_1
Range("I8").Value = coeff_x_0
Dim f_2kq, f_k
 f_2kq = 2 * a3 * b3 & "x ^ 4 + (" & 2 * (a2 * b3 - a1 * b3 - b2 * a3 -
b1 * a3 + a2  * b2) & ")x ^ 3 + (" & 2 * (-a0 * b3 + b0 * a3 + a1 * b2
- b1 * a2 - b0 * a2 + a1         * b1 - a0 * b1) & ")x ^ 2 + (" & 2 *
(a0 * b1 - b0 * a1 + a0 * b0) & ")x"
f_k = a3 * b3 & "x ^ 4 + (" & (a3 * b3 - a1 * b3 - b1 * a3 + a2 * b2)
& ")x ^ 3 + (" & (-a1 * b3 - b1 * a3 + a2 * b2 - a0 * b2 - b0 * a2 +
a1 * b1)  & ")x ^ 2 + (" & (-a0 * b2 - b0 * a2 + a1 * b1 + a0 * b0)
& ")x  + (" & a0 * b0 & ")"
f_x_0 = b3 ^ 2 & "x ^ 4 + (" & (b3 ^ 2 - 2 * b1 * b3 + b2 ^ 2) & ")x ^ 3
 + (" & (-2 * b1   * b3 + b2 ^ 2 - 2 * b0 * b2 + b1 ^ 2) & ")x ^ 2 +
(" & (-2 * b0 * b2 + b1 ^ 2 + b0  ^ 2) & ")x + (" & b0 ^ 2 & ")"
```

```
Range("B10").Value = f_k
Range("B12").Value = f_2kq
Range("B14").Value = f_x_0
End Sub
```

The characteristic polynomial, that has been gotten after substituting concrete coefficient values from the transfer function, can now be tested for Hurwitz stability.

6.3.2 Testing Hurwitz Stability of the Characteristic Equation

There are two groups of methods used for testing if a given characteristic polynomial is Hurwitz. This groups are the exact methods (analytical) and approximate (graphical-analytical methods) methods.

6.3.2.1 Analytical Methods

Analytical methods include the Hurwitz, Jury and Sturm methods.

According to the Hurwitz method, for a given characteristic polynomial $P(x)$, the following conditions have to be met [13]:

1. The number of times the sign of the coefficients of $P(x)$ changes should be even i.e. the first and last coefficient of $P(x)$ should be positive
2. The number of times the sign changes in $V(\bullet) - V(1, \nabla_2, \nabla_4, \cdots, \nabla_{2p})$ should be equal to n.

The determinant ∇_{2p} is constructed from the coefficients of $P(x)$ as follows:

$$\underset{2p}{\nabla} = \begin{vmatrix} a_0 & a_1 & a_2 & \cdots & a_{2p-1} \\ 0 & b_1 & b_2 & \cdots & b_{2p-1} \\ 0 & a_0 & a_1 & \cdots & a_{2p-2} \\ 0 & 0 & b_1 & \cdots & b_{2p-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{vmatrix} \quad p = 1, 2, \cdots, 2n$$

Where b_i—coefficients of the first derivative of $P(x)$.

Jury method, for testing if the characteristic equation is stable, involves forming a matrix Δ_{2n-1} of dimension (2n-1)*(2n-1). This matrix is constructed as follows [14]:

$$\Delta_{2n-1} = \begin{bmatrix} 1 & -a''_{n-1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & -a''_{n-1} & -a'_{n-2} & -a''_{n-3} & -a'_{n-4} & -a''_{n-5} & -a'_{n-6} & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & -a''_{n-1} & -a'_{n-2} & -a''_{n-3} & -a'_{n-4} & -a''_{n-5} & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & -a''_{n-1} & -a'_{n-2} & -a''_{n-3} & -a'_{n-4} & \cdots & 0 & * \\ 0 & 0 & \cdots & 0 & 0 & 0 & -a''_{n-1} & -a'_{n-2} & -a''_{n-3} & -a'_{n-4} & \cdots & 0 & \circ \\ 0 & 0 & \cdots & 0 & 0 & a'_{n-1} & -a''_{n-2} & -a'_{n-3} & a''_{n-4} & a'_{n-5} & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & a'_{n-1} & -a''_{n-2} & -a'_{n-3} & a''_{n-4} & a'_{n-5} & -a''_{n-6} & \cdots & 0 & 0 \\ 0 & 0 & \cdots & a'_{n-1} & -a''_{n-2} & -a'_{n-3} & -a''_{n-4} & -a'_{n-5} & -a''_{n-6} & -a'_{n-7} & \cdots & 0 & 0 \\ 0 & -a''_{n-1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 \\ a'_{n-1} & -a''_{n-1} & & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 \end{bmatrix},$$

$* = (-1)^{(n+1)/2} a_0''$, when n is odd,$(-1)^{n/2} a_0'$,when n is even,
$\circ = (-1)^{(n+1)/2} a_0''$, when n is odd,$(-1)^{n/2} a_0'$,when n is even.

Sturm's method involves the construction of the sturm chain and finding the difference in the number times the sign changes when $x = 0$ and $x = \infty$ [15].

6.3.2.2 Graphical Analytical Methods

This methods involve finding the roots of the characteristic equations. The most widely used method is the root locus method. The root locus method involves finding the roots of the characteristic equation and drawing them on the complex plane.

In classical root locus method, the characteristic equation takes the following form

$$A(x) + hB(x) = 0,$$

where $A(x)$, $B(x)$—polynomials, $h = ck$—variable parameter.

We propose the following root locus equation with variable parameters [17]

$$A(x) + h_1 B(x) + h_2 C(x) = 0. \tag{6.8}$$

Polynomial expressions (6.6) and (6.7) can be expressed in form (6.8) where $h_1 = k$ and $h_2 = qk$.

In order to investigate the robust stability of given NICS we turn to Kharitonov's polynomials.

6.3.3 Mathematical Model for Testing Robust Stability of NICS

The robust stability of a nonlinear impulsive control system (NICS) can be tested by applying the strong Kharitonov's theorem [18].

Since the interval values of the given system are known, it is possible to get the four kharitonov polynomials from the criterial expression (6.5). If the four kharitonov polynomials are Hurwitz polynomials then the system in question is robust stable.

If we use the modified root locus equation (6.8) with several variable parameters, we get the combined locus of roots in the complex plane. For robust stability the roots should not fall on the real positive axis of the complex plane.

6.4 Illustrative Example

Consider a NICS, with the following transfer function with perturbed coefficients

$$W(w) = \frac{(0.12\ldots0.18)w^3 + (0.41\ldots0.59)w^2 - (0.22\ldots0.54)w + (0.09\ldots0.15)}{(0.55\ldots1.99)w^3 + (2.21\ldots3.22)w^2 + (0.45\ldots1.21)w + (0.11\ldots0.16)}. \quad (6.9)$$

The nonlinear elements characteristics lie within the interval 0; 1.5.

The values of V.M. Popov's parameter fall between q = 0.1 and q = 14. Criterion expression for the transfer function (6.9) in symbolic form is written as follows

$$\begin{aligned} P(x)|_{x=\nu^2} = {} & k(b_3(a_3(x^4 + x^3) + a_1(-x^3 - x^2)) + b_2(a_2(x^3 + x^2) \\ & + a_0(-x^2 - x)) + b_1a_3(-x^3 - x^2) + a_1b_1(x^2 + x) \\ & + b_0a_2(-x^2 - x) + a_0b_0(x + 1)) \\ & + kq(b_3(2a_3x^4 + 2a_2x^3 - 2a_1x^3 - 2a_0x^2) \\ & + a_3(-2b_2x^3 - 2b_1x^3 + 2b_0x^2) + b_2(2a_2x^3 + 2a_1x^2 - 2a_0x^2) \\ & + a_2(-2b_1x^2 - 2b_0x^2) + b_1(2a_1x^2 + 2a_0x) - 2b_0a_1x + 2a_0b_0x) \\ & + b_3^2(x^4 + x^3) + b_2^2(x^3 + x^2) + b_1b_3(-2x^3 - 2x^2) + b_1^2(x^2 + x) \\ & + b_0b_2(-2x^2 - 2x) + b_0^2(x + 1). \end{aligned} \quad (6.10)$$

The variation of numeric coefficient values of criterial Eq. (6.6) as the value of parameter k changes is illustrated on the graphs below.

The variation of coefficient values of the free polynomial, as parameter k of the characteristic equation changes, in the criterial Eq. (6.6) is illustrated below.

The calculated coefficient values are typed into the form, as shown in Fig. 6.2, of the program complex Stability. The program complex then plots a modified root

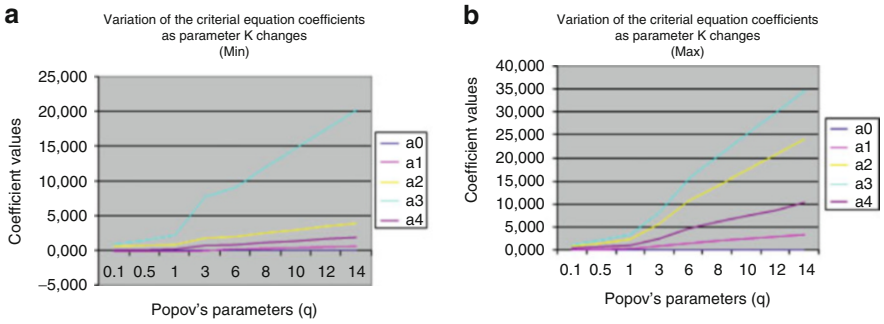


Fig. 6.2 The variation of polynomial coefficients as parameter k of the characteristic equation changes. (**a**) Min values of the transfer function; (**b**) max values of the transfer function

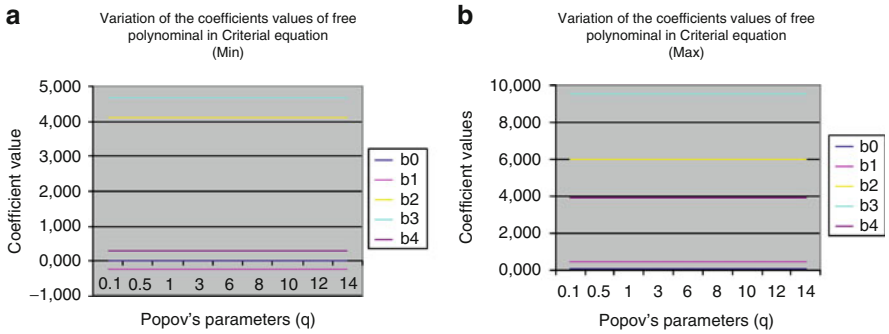


Fig. 6.3 Variation of coefficient values of the free polynomial in the characteristic equation. (**a**) Min values of the transfer function; (**b**) max values of the transfer function

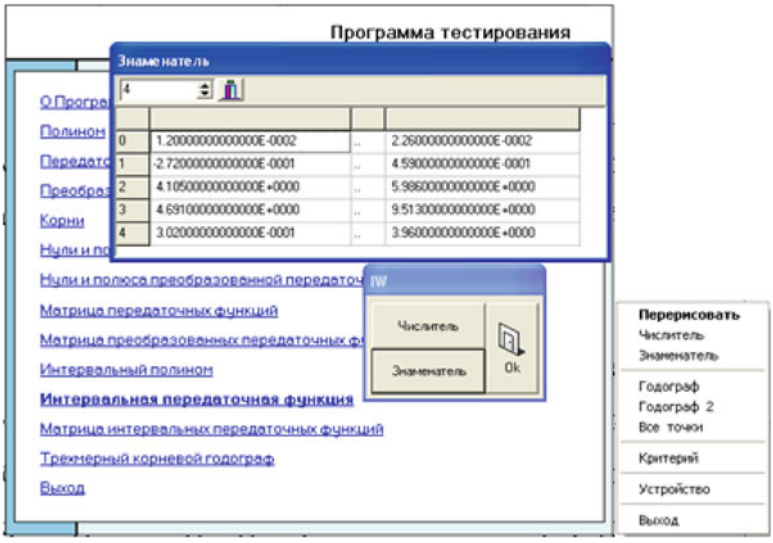


Fig. 6.4 User interface of program complex Stability

locus diagram, which takes into account the perturbed nature of the coefficients, for concrete values of q (Fig. 6.3).

Screenshots of the regions of robust stabilities for several values of q. As shown in Fig. 6.4.

6.5 Conclusion

The proposed approach enables one to get the symbolic coefficient expression of the criteria, for a nonlinear impulsive control system with monotonous nonlinearity and transfer function of given degree, in analytical form and the resultant characteristic

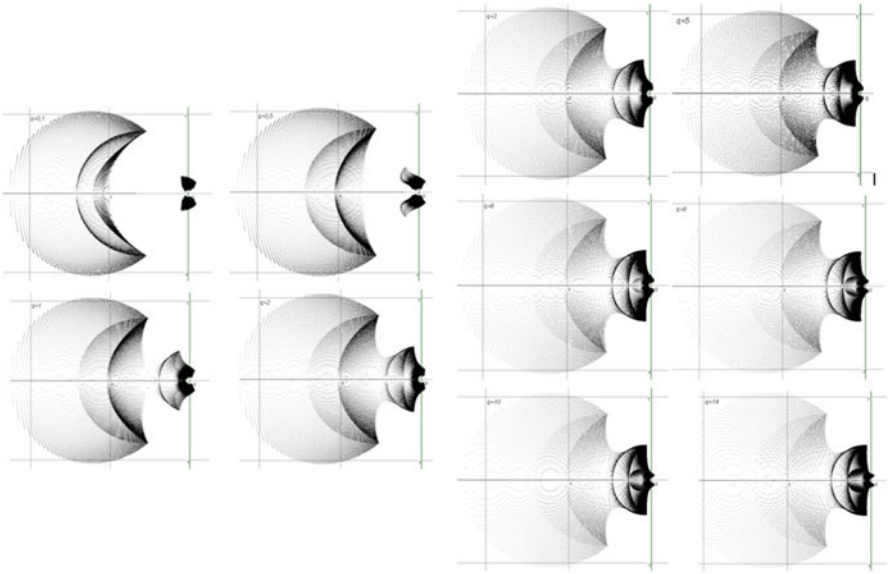


Fig. 6.5 Variation of coefficient values of the free polynomial in the characteristic equation

expression. The resultant characteristic expression, with concrete numeric values from the transfer function, can be used to plot a modified root locus plot, which takes into account the perturbed nature of coefficients. The resultant root locus diagram shows that the region of stability increases while the value of Popov's parameter increases from 0.1 to 14 (Fig. 6.5).

References

1. Arvanitis, K.G., Soldatos, A.G., Boglou, A.K., Bekiaris-Liberis, N.K.: New simple controller rules for integrating and stable or unstable first order plus dead-time processes. In: Recent Advances in Systems-13th WSEAS International Conference on SYSTEMS, Rhodes, pp. 328–337 (2009)
2. Leonov, G.A., Kuznetsov, N.V., Seledzhi, S.M., Shumakov, M.M.: Stabilization of unstable control system via design of delayed feedback. In: Recent Researches in Applied and Computational Mathematics-EUROPMENT/WSEAS International Conference on Applied and Computational Mathematics (ICACM '11), Lanzarote, pp. 18–25 (2011)
3. Fuh, C.C., Tuhg, P.C.: Robust stability bounds for Lur'e systems with parametric uncertainty. J. Mar. Sci. Technol. **7**(2), 73–78 (1999)
4. Haddad, W.M., Collins, E.G., Jr., Bernstein, D.S.: Robust stability analysis using the small gain, circle positivity, and Popov theorems: A comparative study. IEEE Trans. Control Syst. Technol. **1**, 290–293 (1993)
5. Matusu, R.: Robust stability analysis of discrete-time systems with parametric uncertainty: A graphical approach. Int. J. Math Models Methods Appl. Sci. **8**, 95–102 (2014)

6. Prokop, R., Volkova, N., Prokopova, Z.: Tracking and disturbance attenuation for unstable systems: Algebraic approach. In: Recent Researches in Automatic Control - 13th WSEAS International Conference on Automatic Control, Modeling and Simulation (ACMOS '11), Lanzarote, pp. 57–62 (2011)
7. Gazdoš, F., Dostal, P., Marholt, J.: Robust control of unstable systems: Algebraic approach using sensitivity functions. Int. J. Math. Models Methods Appl. Sci. **5**(7), 1189–1196 (2011)
8. Matušu, R., Prokop, R.: Graphical Analysis of Robust Stability for Polynomials with Uncertain Coefficients in Matlab Environment. In: Proceeding of the 16th WSEAS International Conference on Systems, Kos (2012)
9. Matusu, R., Prokop, R.: Graphical approach to robust stability analysis for discrete-time interval polynomials. In: Proceeding of the 15th WSEAS International Conference on Mathematical and Computational Methods in Science and Engineering, Kuala Lumpur (2013)
10. Tseligorov, N.A., Tseligorova, E.N., Mafura, G.M.: Using information technology for computer modeling of nonlinear monotonous impulse control system with uncertainties. In: Proceeding of the Computer Modeling and Simulation, Saint Petersburg, pp. 75–80 (2014)
11. Tseligorov, N.A., Leonov, M.V., Tseligorova, E.N.: Modeling complex "Sustainanability" for the study of robust absolute stability of nonlinear impulsive control systems. In: Proceeding of the Computer Modeling 2012, Saint Petersburg, pp. 9–14 (2012)
12. Tspkin, Ya.Z., Popkov, Yu.S.: Theory of Nonlinear Impulsive Systems. Nauka, Moscow (1973)
13. Serkov, V.I., Tseligorov, N.A.: The analysis of absolute stability of nonlinear impulsive automated systems by analytical methods. Autom. Telemech. **9**, 60–65 (1975)
14. Jury, E.I.: Inners and Stability of Dynamic Systems. Wiley, London (1974)
15. Gantmacher, F.R.: The Theory of Matrices. Taylor & Francis, London (1964)
16. Rimsky, G.V.: Foundations of the general theory of the root locus for automatic control systems. Science and Technology, Minsk (1972)
17. Tseligorov, N.A., Tseligorova, E.N.: Using a modified root locus method to study the robust absolute stability of a multivariate control system. System identification and control problem. In: Proceedings of the IV International Conference SICPRO '07, 29 January–1 February 2007. IPU RAN, 1 DVD Disk, Number: 13034 (2007)
18. Kharitonov, V.L.: Asymptotic stability of an equilibrium position of a family of systems of linear differential equations. Differentsial'nye Uravneniya **14**, 2086–2088 (1978)

Chapter 7
Analytical Modeling of the Viscoelastic Behavior of Periodontal Ligament with Using Rabotnov's Fractional Exponential Function

Sergei Bosiakov and Sergei Rogosin

Abstract The mathematical modeling of a stress-strain state of the viscoelastic periodontal membrane is carried out. Internal and external surfaces of the periodontal ligament are described by a symmetrical two-sheeted hyperboloid. Tooth root is assumed to be a rigid body. Displacements of points on the internal surface of the periodontal ligament coincide with the displacements of the corresponding points of the external surface of the tooth root. The relationships between the displacements and strains for periodontal ligaments are formulated as an assumption that the periodontal tissue approaches to incompressible materials. Viscoelastic constitutive law with a fractional exponential kernel for periodontal ligament was used. The equations of motion for the periodontal ligament relative to translational displacements and rotation angles of its points are derived. In the particular case the vertical translational motion of the tooth root, as well as corresponding displacements are analyzed. Constants of the fractional viscoelastic function were assessed on the basis of the experimental data about the behavior of the periodontal ligament. The obtained results can be used to determine a load for orthodontic tooth movement corresponding to the optimal stresses, as well as to simulate bone remodeling on the basis of changes of stresses and strains in the periodontal ligament during orthodontic movement.

Keywords Periodontal ligament • Tooth root • Viscoelastic model • Fractional exponential function • Translational displacement

S. Bosiakov (✉)
Department of Theoretical and Applied Mechanics, Belarusian State University, Nezavisimosti Avenue 4, 220030 Minsk, Belarus
e-mail: bosiakov@bsu.by

S. Rogosin
Institute of Mathematics, Physics and Computer Sciences, Aberystwyth University, Penglais, Aberystwyth Ceredigion SY23 3BZ, UK

Department of Economics, Belarusian State University, Nezavisimosti Avenue 4, 220030 Minsk, Belarus
e-mail: rogosinsv@gmail.com

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_7

7.1 Introduction

Periodontal ligament is a thin membrane that holds the root of the tooth in the alveolar bone. It reduces and distributes the occlusal load on the tooth by means of collagen fibers. In normal conditions there is no contact between the tooth root and the bone tissue. The load acting on the tooth, is transmitted to the alveolar bone by the periodontal ligament strain. Periodontium can be loaded by long-term (orthodontic) forces or by short-term (occlusal) load. It is occurred an orthodontic movement of teeth as a result of the biological response of bone alveolar process [1, 2].

Linearly elastic (bilinear elastic), viscoelastic, hyperelastic and biphasic (multiphasic) models are used to predict the behavior of the periodontal ligament under the various loading conditions. Overview of the specific applications of different models is given in [3]. The main drawback of the periodontal ligament simulation on the base of medium with complex properties is the lack of accurate quantitative data of mechanical parameters. For the viscoelastic models it is compensated by existence of known values of the relaxation times and elasticity moduli [4–6], and the experimental data determining the viscoelastic properties [7–12]. In [13], it is shown that all the tissues involved in the reconstruction of bone tissue, demonstrate viscoelastic properties which do not depend on the applied forces. Experimental determination and modeling of periodontium properties is discussed in [14].

Several viscoelastic models of the periodontal ligament behavior, based on the laws of Maxwell, Voigt, Kelvin-Voigt [3] have been proposed. Such material possesses a rheological behaviour. Rheology as a branch of science is concerned with extending continuum mechanics to characterize flow of materials, that exhibits a combination of elastic, viscous and plastic behaviour by properly combining elasticity and (Newtonian) fluid mechanics. In particular, the materials studied in the framework of the rheological investigations could have a memory (so called hereditary materials). For modeling of this effect the fractional approaches are used, e.g., [15–17]. The history of fractional modeling in rheology is presented in [18] (see also [19] and references therein). The fractional viscoelastic model is very natural for the study of periodontal membrane. In addition, the fractional models (models with fractional derivatives) are successfully used to solve different problems of mechanics [20–22].

Rabotnov [23] presented a general theory of hereditary solid mechanics using integral equations, and Koeller [24] reviewed the use of integral equations for viscoelasticity and interjects fractional calculus into Rabotnov's theory by the introduction of the spring-pot, which he used to generalize the classical models. Rossikhin [25] summarized Rabotnov's theory (see also [26]). Rabotnov's fractional exponential function is related to the well known Mittag-Leffler function [27]. By using this relation it can be shown the equivalence of Rabotnov's theory to Torvik and Bagley's theory based on the fractional polynomial constitutive equation.

The aim of this work is to formulate equations of motion of the periodontal ligament. We use an approach based on viscoelastic model similar to Rabotnov's model. It allows us to determine the translational displacement and rotation angles

of the periodontium under the action of a concentrated load. In the particular case the vertical motion of the tooth root, as well as arising from this stresses and displacements were analyzed. The experimental data on the behavior of the periodontal ligament allows us to estimate corresponding constants of the fractional viscoelastic function.

7.2 Equations of Movements for Viscoelastic Periodontal Ligament

7.2.1 Geometrical Form of Tooth Root and Periodontal Ligament

The external surface of the tooth root (supposed to be an absolutely rigid body) and the adjacent inner surface of the periodontal ligament are modeling by a two-sheeted hyperboloid

$$F(x,y,z) = y - \frac{h}{\sqrt{1-p^2}-p} \times \\ \times\left(\sqrt{(1-e^2)\left(\frac{x}{b}\right)^2 + \left(\frac{z}{b}\right)^2 + p^2} - p\right) = 0, \tag{7.1}$$

where h is the height of alveolar crest; $e = \sqrt{1-(b/a)^2}$ is the eccentricity of the ellipse in cross-section of tooth in alveolar crest; a and b are the axes of this ellipse; p is the parameter of rounding of the tooth root.

The internal surface of the periodontal ligament adjacent to the dental alveoli bone is shifted along the normal to the surface of the tooth root in $\delta > 0$. Its equation is as follows:

$$F_1(x,y,z) = y + n_y\delta - \frac{h}{\sqrt{1-p^2}-p} \times \\ \times\left(\sqrt{(1-e^2)\left(\frac{1}{b}(x+n_x\delta)\right)^2 + \left(\frac{1}{b}(z+n_z\delta)\right)^2 + p^2} - p\right) = 0,$$

where n_x, n_y, and n_z are components of the unit normal vector to the surface of the first hyperboloid. The components of the normal vector are determined from (7.1):

$$n_x = -\frac{1}{\Delta}\frac{h(1-e^2)x}{bAB}, n_y = \frac{1}{\Delta}, n_z = \frac{1}{\Delta}\frac{hz}{bAB}, A = \sqrt{1+p^2} - p, \\ B = \sqrt{(1-e^2)x^2 + z^2 + (bp)^2}, \Delta = \sqrt{1 + \frac{h^2((1-e^2)x^2+z^2)}{b^2AB}}. \tag{7.2}$$

Under the action on a tooth a concentrated force, the points of the periodontal ligament contiguous to surface of the tooth root (7.1) begin to move equally to the root itself. The external surface of the periodontal ligament is fixed. There is no significant difference between the model considering the fixing of the outer surface of the periodontal ligament in the alveolar bone or its rigid fixing. Therefore, for calculating the initial movement of the teeth in the periodontal ligament, both the teeth and the alveolar bone could be considered as solids [28].

7.2.2 *Relations for Strains and Displacements*

It is supposed that the periodontal ligament consists of a material with Poisson's ratio equal to 0.49. This means it behaves almost as fluid, i.e. the periodontal tissue begins to flow around the surface of the root of the tooth when the root is displaced to the wall of the dental alveolus [29]. Hence the strains and relative shears associated with the normal vector, generatrix and guide to the external surface of the tooth root could be represented in the coordinate system as follows [29, 30]:

$$\varepsilon_{nn} = -\frac{u_n}{\delta}, \varepsilon_{tt} = \varepsilon_{\theta\theta} = 0, \gamma_{n\theta} = -\frac{u_\theta}{\delta}, \gamma_{nt} = -\frac{u_t}{\delta}, \gamma_{t\theta} = 0, \tag{7.3}$$

where u_n, u_t and u_θ are displacements of the periodontium points in the direction of the $\boldsymbol{n}$ (the normal vector to the root surface), $\boldsymbol{t}$ is the generatrix vector to the root surface, $\boldsymbol{\theta}$ is the tangential vector to the root surface, and δ is the width of the periodontal ligament in the normal direction. The normal vector, generatrix and guide to the root surface of the tooth, as well as their geometrical dimensions are shown in Fig. 7.1.

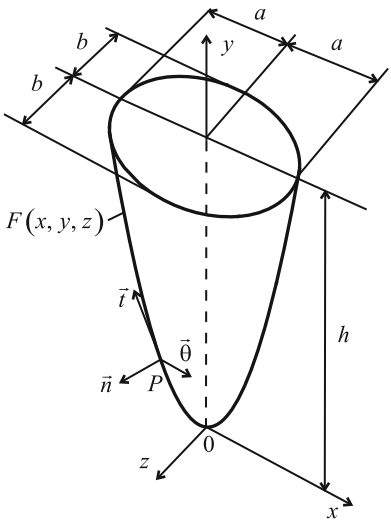


Fig. 7.1 Root of tooth in two-sheeted hyperboloid form: $\boldsymbol{n}$ is the normal, $\boldsymbol{t}$ is the generatrix, and $\boldsymbol{\theta}$ is the guide to the surface of the hyperboloid at the point P

The strains and relative shears can be expressed in the coordinate system (x, y, z) by the components of the strain tensor in the coordinate system (n, t, θ) [30]:

$$\begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{12} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{13} & \varepsilon_{23} & \varepsilon_{33} \end{pmatrix} = T_2 \cdot T_1 \cdot \begin{pmatrix} \varepsilon_{nn} & \varepsilon_{tn} & \varepsilon_{\theta n} \\ \varepsilon_{tn} & 0 & 0 \\ \varepsilon_{\theta n} & 0 & 0 \end{pmatrix} \cdot T_1^T \cdot T_2^T, \quad \varepsilon_{tn} = \frac{1}{2}\gamma_{tn}, \varepsilon_{\theta n} = \frac{1}{2}\gamma_{\theta n}, 1 \equiv x, 2 \equiv y, 3 \equiv z. \tag{7.4}$$

The components of the vectors (u_n, u_t, u_θ) and (u_x, u_y, u_z) satisfy the following equation:

$$\begin{pmatrix} u_n \\ u_t \\ u_\theta \end{pmatrix} = T_1^T \cdot T_2^T \cdot \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix}, T_1 = \begin{pmatrix} \sin(\alpha) & \cos(\alpha) & 0 \\ -\cos(\alpha) & \sin(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad T_2 = \begin{pmatrix} H & 0 & -G \\ 0 & 1 & 0 \\ G & 0 & H \end{pmatrix}, H = \frac{x(1-e^2)}{\sqrt{x^2(1-e^2)^2 + z^2}}, G = \frac{z}{\sqrt{x^2 + z^2}}, \tag{7.5}$$

where T_1 is the rotation matrix relative to the guide $\boldsymbol{\theta}$; T_2 is the rotation matrix relative to the z-axis on angle φ; T_1^T, T_2^T are the transpose matrixes T_1 and T_2, respectively. The angle α between generatrix to the root surface and xz-plane is given by the formula

$$\tan(\alpha) = \frac{h\sqrt{(1-e^2)^2x^2 + z^2}}{b(\sqrt{1+p^2} - p)\sqrt{(bp)^2 + (1-e^2)x^2 + z^2}}.$$

Any displacements of the tooth root can be described by a combination of the translational displacements u_{0x}, u_{0y}, u_{0z} and the angles of rotation $\theta_x, \theta_y, \theta_z$ relative of the coordinate axes. Since the thickness of periodontal is small, the rotation angles are small too. Therefore, we can use the following linearized expressions [30]:

$$u_x = u_{0x} + z\theta_y - y\theta_z, u_y = u_{0y} - z\theta_x + x\theta_z, u_z = u_{0z} + y\theta_x - x\theta_y. \tag{7.6}$$

Equations (7.2)–(7.6) allow us to express the strains and relative shifts via translational displacements and rotation angles in the coordinate system (x, y, z).

7.2.3 Constitutive Equation

Taking into account the viscoelastic properties of the periodontal ligament, the components of the stress tensor of the periodontal ligament are represented in the following form:

$$\sigma_{ij} = \frac{E_\infty}{(1-2\nu)(1+\nu)}\left\{(1-2\nu)\varepsilon_{ij} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right)\varepsilon_{ij}(t-\tau)d\tau + \right.$$

$$\left. + \nu\left(\sum_{k=1}^{3}\varepsilon_{kk} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right)\sum_{k=1}^{3}\varepsilon_{kk}(t-\tau)d\tau\right)\right\},$$

$$i, j = 1, 2, 3; 1, 2, 3 \text{ are identical } x, y, z, \text{respectively} \tag{7.7}$$

where τ_s is the relaxation time, $\nu_\varepsilon = \frac{E_\infty - E_0}{E_\infty}$, E_0 and E_∞ are, respectively, the relaxed (prolonged modulus of elasticity, or the rubbery modulus) and nonrelaxed (instantaneous modulus of elasticity, or the glassy modulus) magnitudes of the elastic modulus [25], and $\mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right)$ is Rabotnov's "fractional exponential function", which describes the relaxation of volume and shear stresses. It was introduced by Rabotnov in the form [23, 31]

$$\mathcal{E}_\gamma\left(-\frac{t}{\tau_\varepsilon}\right) = \frac{t^{\gamma-1}}{\tau_\varepsilon^\gamma}\sum_{n=0}^{\infty}(-1)^n\frac{(t/\tau_\varepsilon)^{\gamma n}}{\Gamma[\gamma(n+1)]},$$

where $0 < \gamma < 1$ is a fractional parameter. Note that Rabotnov's function is a special case of the classical Mittag-Leffler function highly used in fractional models (see [19, 27]).

7.2.4 Equations of Motion

To find the translational displacements and the rotation angles, we use the conditions of the dynamic equilibrium of the tooth root:

$$\iint_F (\boldsymbol{n}\cdot\sigma)dF + M\frac{d^2\boldsymbol{u_0}}{dt^2} - \boldsymbol{f} = 0, \quad \iint_F \boldsymbol{r}\times(\boldsymbol{n}\cdot\sigma)dF + J\frac{d^2\boldsymbol{\theta}}{dt^2} - \boldsymbol{m} = 0, \tag{7.8}$$

where $\boldsymbol{m} = (m_x, m_y, m_z)$ is the principal moment of external forces, $\boldsymbol{f} = (f_x, f_y, f_z)$ is the principal vector of external forces, $\boldsymbol{r}$ is the radius-vector, $\boldsymbol{n} = (n_x, n_y, n_z)$ is the unit normal vector to the surface (7.1), σ is the stress tensor, M is the mass of the tooth root (7.1), J is the axial moment of inertia of the tooth root, $\boldsymbol{u}_0 = (u_{0x}, u_{0y}, u_{0z})$ is the vector of translational displacements of the tooth root

along the coordinate axes, and $\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)$ is the vector of rotation angles of the tooth root with respect to the coordinate axes.

Taking into account relations (7.2) and (7.7) one can reduce equations of motion (7.8) after the transformations to the following form

$$
\begin{aligned}
& c_x(u_{0x} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma u_{0x}(t-\tau)d\tau) + \\
& + c_{\theta xy}(\theta_z - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma \theta_z(t-\tau)d\tau) + M\frac{d^2u_{0x}}{dt^2} = f_x, \\
& c_y(u_{0y} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma u_{0y}(t-\tau)d\tau) + M\frac{d^2u_{0y}}{dt^2} = f_y, \\
& c_z(u_{0z} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma u_{0z}(t-\tau)d\tau) + \\
& + c_{\theta yz}(\theta_x - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma \theta_x(t-\tau)d\tau) + M\frac{d^2u_{0z}}{dt^2} = f_z, \\
& c_{\theta z}(u_{0z} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma u_{0z}(t-\tau)d\tau) + \mu_x(\theta_x - \\
& - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma \theta_x(t-\tau)d\tau) + J_x\frac{d^2\theta_x}{dt^2} = y_f f_z - z_f f_y, \\
& \mu_y(\theta_y - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma \theta_y(t-\tau)d\tau) + J_y\frac{d^2\theta_y}{dt^2} = z_f f_x - x_f f_z, \\
& c_{\theta x}(u_{0x} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma u_{0x}(t-\tau)d\tau) + \mu_z(\theta_z - \\
& - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma \theta_z(t-\tau)d\tau) + J_z\frac{d^2\theta_z}{dt^2} = x_f f_y - y_f f_x, \\
& \mathcal{E}_\gamma \equiv \mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right),
\end{aligned}
\tag{7.9}
$$

where c_x, c_y, and c_z are, respectively, the stiffness coefficients of the periodontal ligament at the tooth root translation along the co-ordinate axes; $c_{\theta xy}$ and $c_{\theta yz}$ are the static moments of stiffness; $c_{\theta x}$ and $c_{\theta z}$ are the stiffness coefficients of the periodontal ligament at the tooth root rotations relative to the x-axis and z-axis, under the force acting along these coordinate axis; μ_x, μ_y, and μ_z are the stiffness coefficients of the periodontal ligament at the tooth root rotations relative to the axes x, y and z, respectively; and x_f, y_f and z_f are the coordinates of the point where the load is applied. The coefficients of system (7.9) are defined as follows

$$c_x = E_\infty \iint\limits_F (ABb(2\nu-1)\cos(\alpha) + \\ +h(2Hx(1-e^2)(\nu-1) + Gz(2\nu-1))\sin(\alpha))\frac{dF}{C},$$

$$c_y = E_\infty \iint\limits_F (ABb(1-2\nu)\cos(\alpha) - \\ -h(Hx(1-e^2)(1-2\nu) + 2Gz(1-\nu))\sin(\alpha))\frac{dF}{C},$$

$$c_z = E_\infty \iint\limits_F (ABb(1-2\nu)\cos(\alpha) - \\ -h(Hx(1-e^2)(1-2\nu) + 2Gz(1-\nu))\sin(\alpha))\frac{dF}{C},$$

$$c_{\theta x} = -E_\infty \iint\limits_F ((1-2\nu)((1-e^2)hx^2 + ABby)\times \\ \times\cos(\alpha) + (hy(2Hx(1-e^2)(1-\nu) + \\ +Gz(1-2\nu)) + 2ABHb\nu x)\sin(\alpha))\frac{dF}{C},$$

$$c_{\theta z} = E_\infty \iint\limits_F ((1-2\nu)(hz^2 + ABby)\cos(\alpha) + \\ +(hy(Hx(1-e^2)(1-2\nu) + 2Gz(1-\nu)) + 2ABGb\nu z)\sin(\alpha))\frac{dF}{C},$$

$$c_{\theta xy} = E_\infty \iint\limits_F ((2B(1-e^2)h\nu x^2 - Ab(1-2\nu)yB^2)\cos(\alpha) -$$
$$-(Bhy(2Hx(1-e^2)(1-\nu) + Gz(1-2\nu)) +$$
$$+AB^2bhx(1-2\nu))\sin(\alpha))\frac{dF}{C},$$

$$c_{\theta yz} = E_\infty \iint\limits_F ((ABby(1-2\nu) + 2hz^2\nu)\cos(\alpha) +$$
$$+(ABGbz(1-2\nu) + hy(Hx(1-e^2)(1-2\nu) + 2Gz(1-\nu)))\sin(\alpha))\frac{dF}{C},$$

$$\mu_x = E_\infty \iint\limits_F ((hyz^2 + ABb((1-2\nu)y^2 + 2z^2(1+\nu)))\cos(\alpha) +$$
$$+(ABGbyz + h(Hx(1-e^2)(1-2\nu)(y^2+z^2) +$$
$$+Gz(2y^2(1-\nu) + z^2(1-2\nu))))\cos(\alpha))\frac{dF}{C},$$

$$\mu_y = E_\infty \iint\limits_F (ABb(x^2+z^2)(1-2\nu)\cos(\alpha) -$$
$$-h(Gz(x^2(1+e^2-2\nu) + z^2(1-2\nu)) + BHx(x^2(1-e^2)(1-2\nu) +$$
$$+z^2(1-2e^2(1-\nu)-2\nu)))\sin(\alpha))\frac{dF}{C},$$

$$\mu_z = E_\infty \iint\limits_F (hx^2y(1-e^2) + ABb(2x^2(1-\nu) +$$
$$+y^2(1-2\nu))\cos(\alpha) + (Hx(ABby + h(1-e^2)(x^2(1-2\nu) +$$
$$+2y^2(1-\nu))) + Ghz(x^2+y^2))\sin(\alpha))\frac{dF}{C},$$
$$C = 2ABb\delta(1+\nu)(2\nu-1).$$

Note that the expressions for the coefficients $c_{\theta yz}$, $c_{\theta z}$ and $c_{\theta xy}$, $c_{\theta x}$ are not the same, while from conservatism of the system (7.9) it follows that $c_{\theta yz} = c_{\theta z}$ and $c_{\theta xy} = c_{\theta x}$. The ratios for the corresponding constants can be reduced to one, after the substitution of all intermediate values. However, in this case, the expressions are rather huge, therefore they have to be written in a more compact form. Calculation of the coefficients of the system (7.9) shows that equalities $c_{\theta yz} = c_{\theta z}$ and $c_{\theta xy} = c_{\theta x}$ are really satisfied. To be specific these coefficients are calculated for the tooth

Table 7.1 Stiffnesses of periodontal ligament

c_x (MN/m)	c_y (MN/m)	c_z (MN/m)	μ_x (N · m)	μ_y (N · m)	μ_z (N · m)
4.52307	1.05533	6.28229	553.949	4.99209	429.647

root with geometrical dimensions $h = 13.0\,\text{mm}$, $b = 3.9\,\text{mm}$, $p = 0.4$ and $e = 0.6$. Elastic properties of the periodontal ligament are assigned by constants $E = E_\infty = 680\,\text{kPa}$ and $\nu = 0.49$ [32]. Thickness δ of the periodontal ligament is 0.229 mm [29]. At this case, $c_{\theta xy} = c_{\theta x} = -410178.8\,\text{N}$ and $c_{\theta yz} = c_{\theta z} = 54796.7\,\text{N}$. The stiffnesses are given in Table 7.1.

The stiffnesses of the periodontal ligament and the moments of stiffnesses depend on the geometrical shape of the tooth root, Poisson's ratio and the relaxed and nonrelaxed elastic moduli of periodontal tissue and are time-independent. Therefore, the stiffnesses and the moments of stiffnesses could be eliminated from the integrals in Eq. (7.9).

7.3 Vertical Displacement of Tooth Root

7.3.1 Approximate Solution

To find the material constants and the relaxation time, the experimental data on the stress-strain state of the periodontal ligament can be used and, in particular, the dependence of the periodontal points displacements on time. Typically, such data are obtained for the translational motion of the tooth root in the vertical and horizontal directions under the action of a load, which takes on discrete values with time, or which is changed with a predetermined frequency.

During the motion of the tooth root along the y-axis, the corresponding extrusion (or intrusion), the translational displacement along the x- and z-axes, as well as the angles of rotation are equal to zero, i.e., $u_{0x} = u_{0z} = 0$, and $\theta_x = \theta_y = \theta_z = 0$. Load is acting only in the y-axis direction. In this case, one obtains from (7.9)

$$c_y(u_{0y} - v_\varepsilon \int_0^t \mathcal{E}_\gamma u_{0y}(t-\tau)d\tau) + M\frac{d^2u_{0y}}{dt^2} = f_y. \tag{7.10}$$

We assume that the duration of the load action on the tooth root is large enough (from 0 to 300 s [5, 33]), and the mass of the tooth root is small ($m \sim 10^{-3}\,\text{kg}$). Therefore one can neglect the inertial term in Eq. (7.10)

$$c_y(u_{0y} - v_\varepsilon \int_0^t \mathcal{E}_\gamma u_{0y}(t-\tau)d\tau) = f_y.$$

According to [34] the solution of this equation can be written as

$$u_{0y}(t) = \frac{f_y}{c_y}\left(1 + \nu_\sigma \frac{t^\gamma}{\tau_\sigma} \sum_{n=0}^{\infty} \frac{(-1)^n (\frac{t}{\tau_\sigma})^{\gamma n}}{\Gamma[\gamma(n+1)]}\right), \tag{7.11}$$

where $\nu_\sigma = \frac{E_\infty - E_0}{E_0}$, τ_σ is the retardation time. Solution (7.11) corresponds to the initial conditions $u_{0y}(t)|_{t=0} = \frac{f_y}{c_y}$ and $\frac{du_{0y}(t)}{dt}|_{t=0} = \frac{d^2u_{0y}(t)}{dt^2}|_{t=0} = 0$.

7.3.2 *Effect of Fractional Parameter*

In the function $u_{0y}(t)$ the stiffness c_y is known (see Table 7.1), the load f_y must be specified. Retardation time τ_σ, parameter ν_σ and fractional parameter γ are unknown. The magnitudes of these constants are estimated by using the models of the tooth movement versus time in viscoelastic periodontal ligament that are analyzed in [5, 33]. Displacement of the tooth root in the viscoelastic periodontal ligament over time is determined for continuous load that changes from 0 to 15 N [5], as well as for the discrete change of the load from 0.5 to 3.0 N with step of 0.5 N [33]. Time is changed from 0 to 300 s [5] and from 0 to 1,200 s [33]. In our case, the calculation of displacements is provided in the time interval from 0 to 300 s. The transition phase is 20–25 s [5, 33]. Figure 7.2 shows effect of the fractional parameter on the displacement versus the time. The tooth crown is loaded by constant compressive force of 2 N, the retardation time and the parameter are equal to 550 s and $1.3 \cdot 10^3$, respectively.

Figure 7.2 shows that the increase of the fractional parameter leads to an increase of the transition phase and maximum displacement of the tooth root (parameter ν_σ and retardation time τ_σ have constant values).

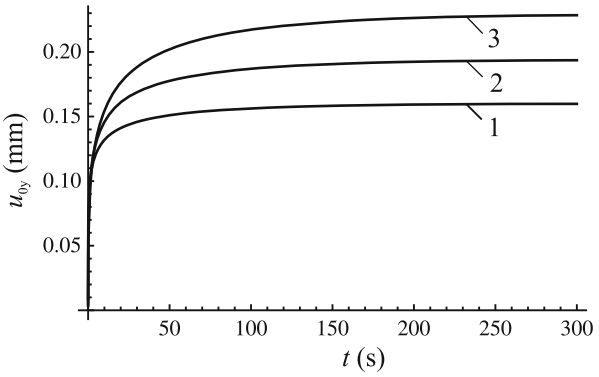


Fig. 7.2 Displacement vs time: 1—$\gamma = 0.25$, 2—$\gamma = 0.30$, 3—$\gamma = 0.35$

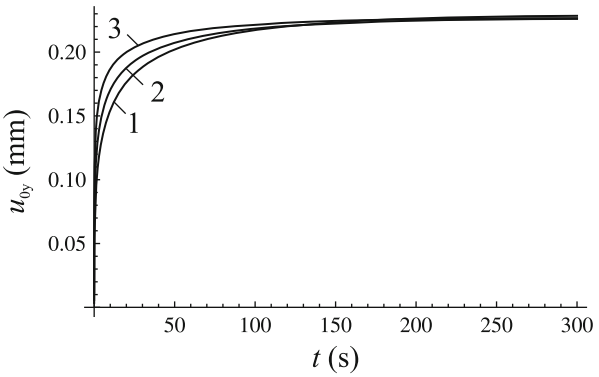


Fig. 7.3 Displacement vs time: 1—$\gamma = 0.25$ and $\nu_\sigma = 1850$, 2—$\gamma = 0.30$ and $\nu_\sigma = 1520$, 3—$\gamma = 0.35$ and $\nu_\sigma = 1300$

The change of the parameter ν_σ for different fractional parameter can give the same maximum displacements of the tooth root in the periodontium. Figure 7.3 shows the displacements versus time of the for load of 2 N and retardation time of 550 s. From Figs. 7.2 and 7.3 it follows that the simultaneous change of the fractional parameter γ and parameter ν_σ allows us to specify a necessary transitional phase and the maximum displacement of the tooth root in periodontal ligament. This can be done for any load. The magnitude of the maximum displacement can be specified by changing the parameters ν_σ, which can be a function of the load.

7.3.3 *Effect of the Number of Terms in the Series*

The number of terms extracting from the series in the approximate solution (7.11) substantially affects the displacement of the tooth root over time. Especially significant is their impact for small values of fractional parameter. In particular, the ineffective terms in the series are those with $n \geq 20$, for $\gamma = 0.5$ we have $n \geq 10$, and for $\gamma = 0.75$ we have $n \geq 3$.

7.3.4 *Effect of Inertial Term*

Now we carry out estimates of the inertial term $M\frac{d^2u_{0y}(t)}{dt^2}$ taking into account the solution (7.11) for the retardation time between 350 and 550 s, the fractional parameter from 0.25 to 0.90, and parameter ν_σ from $1.3 \cdot 10^3$ to $1.8 \cdot 10^3$. Mass of the tooth $1 \cdot 10^{-3}$ kg, the geometric dimensions of the tooth root and the elastic properties of the periodontal membranes were previous. The calculation indicated

that on the time interval from 0 to 300 s the inertial member has the order of $10^{-11}\ \frac{m}{s^2}$ to $10^{-10}\ \frac{m}{s^2}$. Thus, the solution of (7.11) can be used to sufficiently good approximation of the vertical movement of the tooth root.

7.4 Conclusions

To generalize viscoelastic models corresponding to different types of the stress-strain state of the periodontal ligament under the action of concentrated forces and moments, the equations of motion of the tooth root involving the fractional exponential function are suggested. The advantage of this model is in the use of the fractional parameter γ and the parameter ν_σ for the description of the various pathological processes and age-related changes in the periodontium. Fractional parameter make possible to take into account the different behavior of the periodontal tissue during the action of different short-term and long-term loads.

Constants of the fractional viscoelastic function determining on the base of the experimental data on the behavior of the periodontal ligament during vertical displacement of the tooth root were assessed. To clarify the role of the material constants, the results of experiments on the cyclic loading of the tooth [10, 11], as well as the experimental data on the root displacement in a horizontal plane [12] are used.

The obtained results can be applied to determine magnitude of a load for orthodontic tooth movement corresponding to the optimal stresses, as well as to simulate bone remodeling on the basis of changes of stresses and strains in the periodontal ligament during orthodontic movement.

Acknowledgements The research is supported by the FP7 IRSES Marie Curie grant TAMER No 610547. The authors are thankful to professor Francesco Mainardi and to professor Ivan Argatov for valuable discussions of the results of the paper.

References

1. Masella, R.S., Meister, M.: Current concepts in the biology of orthodontic tooth movement. Am. J. Orthod. Dentofac. Orthop. **129**, 458–468 (2006)
2. Wise, G.E., King, G.J.: Mechanisms of tooth eruption and orthodontic tooth movement. J. Dent. Res. **87**, 414–434 (2008)
3. Fill, T.S., Toogood, R.W., Major, P.W., Carey, J.P.: Analytically determined mechanical properties of, and models for the periodontal ligament: Critical review of literature. J. Biomech. **45**, 9–16 (2012)
4. Komatsu, K.: Mechanical strength and viscoelastic response of the periodontal ligament in relation to structure. J. Dent. Biomech. **1**, 1–18 (2010)
5. Qian, L., Todo, M., Morita, Y., Matsushita, Y., Koyano, K.: Deformation analysis of the periodontium considering the viscoelasticity of the periodontal ligament. Dent. Mater. **25**, 1285–1292 (2009)

6. Wood, S.A., Strait, D.S., Dumont, E.R., Ross, C.F., Grosse, I.R.: The effects of modeling simplifications on craniofacial finite element models: The alveoli (tooth sockets) and periodontal ligaments. J. Biomech. **44**, 1831–1838 (2011)
7. Ferrari, M., Sorrentino, R., Zarone, F., Apicella, D., Aversa, R., Apicella, A.: Non-linear viscoelastic finite element analysis of the effect of the length of glass fiber posts on the biomechanical behaviour of directly restored incisors and surrounding alveolar bone. Dent. Mater. J. **27**, 485–498 (2008)
8. Natali, A.N., Pavan, P.G., Scarpa, C.: Numerical analysis of tooth mobility: Formulation of a non-linear constitutive law for the periodontal ligament. Dent. Mater. **20**, 623–629 (2004)
9. Toms, S.R., Eberhardt, A.W.: A nonlinear finite element analysis of the periodontal ligament under orthodontic tooth loading. Am. J. Orthod. Dentofac. Orthop. **123**, 657–665 (2003)
10. Bergomi, M., Cugnoni, J., Galli, M., Botsis, J., Belser, U.C., Wiskott, H.W.A.: Hydromechanical coupling in the periodontal ligament: A porohyperelastic finite element model. J. Biomech. **44**, 34–38 (2011)
11. Naveh, G.R.S., Chattah, N.L.-T., Zaslansky, P., Shahar, R., Weiner, S.: Tooth-PDL-bone complex: Response to compressive loads encountered during mastication - a review. Arch. Oral Biol. **57**, 1575–1584 (2012)
12. Yoshida, N., Koga, Y., Peng, Ch.-L., Tanaka, E., Kobayashi, K.: In vivo measurement of the elastic modulus of the human periodontal ligament. Med. Eng. Phys. **23**, 567–572 (2001)
13. Cronau, M., Ihlow, D., Kubein-Meesenburg, D., Fanghanel, J., Dathe, H., Nagerl, H.: Biomechanical features of the periodontium: An experimental pilot study in vivo. Am. J. Orthod. Dentofac. Orthop. **129**, 599.e13–599.e21 (2006)
14. Fill, T.S., Carey, J.P., Toogood, R.W., Major, P.W.: Experimentally determined mechanical properties of, and models for, the periodontal ligament: Critical review of current literature. J. Dent. Biomech. **2**, 1–11 (2011)
15. Uchaikin, V.: Fractional Derivatives for Physicists and Engineers, vols. I–II. Springer/Higher Education Press, Berlin/Beijing (2013)
16. Koeller, R.C.: A theory relating creep and relaxation for linear materials with memory. J. Appl. Mech. **77**, 031008-1–031008-9 (2010)
17. West, B.J., Bologna, M., Grigolini, P.: Physics of Fractal Operators. Springer, NewYork (2003)
18. Rogosin, S., Mainardi, F.: George William Scott Blair - the pioneer of factional calculus in rheology. Commun. Appl. Ind. Math. **6**(1), e481 (2014)
19. Mainardi, F.: Fractional Calculus and Waves in Linear Viscoelasticity. Imperial College Press/World Scientific, London/Singapore (2010)
20. Rossikhin, Yu.A., Shitikova, M.V.: Nonlinear dynamic response of a fractionally damped suspension bridge subjected to small external force. Int. J. Mech. **7**, 155–163 (2013)
21. Rossikhin, Yu.A., Shitikova, M.V., Popov, I.I.: Dynamic response of a hereditarily elastic beam with Rabotnov's kernel impacted by an elastic rod. In: Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences, pp. 25–31. Saint Petersburg State Polytechnic University, Saint-Petersburg (2014)
22. Sibatov, R.T., Svetukhin, V.V., Uchaikin, V.V., Morozova, E. V.: Fractional model of electron diffusion in dye-sensitized nanocrystalline solar cells. In: Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences, pp. 118–121. Saint Petersburg State Polytechnic University, Saint-Petersburg (2014)
23. Rabotnov, Yu.N.: Elements of Hereditary Solid Mechanics. Mir Publishers, Moscow (1980)
24. Koeller, R.C.: Application of fractional calculus to the theory of viscoelasticity. J. Appl. Mech. **51**, 299–307 (1984)
25. Rossikhin, Yu.A., Shitikova, M.V.: Centennial jubilee of academician Rabotnov and contemporary handling of his fractional operator. Fract. Calc. Appl. Anal. **17**, 674–683 (2014)
26. Rossikhin, Yu.A., Shitikova, M.V.: Two approaches for studying the impact response of viscoelastic engineering systems: An overview. Comput. Math. Appl. **66**, 755–773 (2013)
27. Gorenflo, R., Kilbas, A., Mainardi, F., Rogosin, S.: Mittag-Leffler Functions, Related Topics and Applications. Springer, New York (2014)

28. Hohmann, A., Kober, C., Young, Ph., Dorow, Ch., Geiger, M., Boryor, A., Sander, F.M., Sander, Ch., Sander, F.G.: Influence of different modeling strategies for the periodontal ligament on finite element simulation results. Am. J. Orthod. Dentofac. Orthop. **139**, 775–783 (2011)
29. Provatidis, C.G.: An analytical model for stress analysis of a tooth in translation. Int. J. Eng. Sci. **39**, 1361–1381 (2001)
30. Van Schepdael, A., Geris, L., Van der Sloten, J.: Analytical determination of stress patterns in the periodontal ligament during orthodontic tooth movement. Med. Eng. Phys. **35**, 403–410 (2013)
31. Rabotnov, Yu.N.: Equilibrium of an elastic medium with after-effect. Fract. Calc. Appl. Anal. **17**, 684–696 (2014)
32. Tanne, K., Nagataki, T., Innoue, Y., Sakuda, M., Burstone, C.J.: Patterns of initial tooth displacement associated with various root lengths and alveolar bone heights. Am. J. Orthod. Dentofac Orthop. **100**, 66–71 (1991)
33. Slomka, N., Vardimon, A.D., Gefen, A., Pilo, R., Bourauel, C., Brosh, T.: Time-related PDL: Viscoelastic response during initial orthodontic tooth movement of a tooth with functioning interproximal contact – a mathematical model. J. Biomech. **41**, 1871–1877 (2008)
34. Rossikhin, Yu.A.: Reflections on two parallel ways in the progress of fractional calculus in mechanics of solids. Appl. Mech. Rev. **63**, 010701-1–010701-12 (2010)

Chapter 8
Simulation of Stiff Hybrid Systems with One-Sided Events and Nonsmooth Boundaries

Yury V. Shornikov, Maria S. Nasyrova, and Dmitry N. Dostovalov

Abstract Different classes of modal behavior of hybrid systems (HS) are considered. Architecture of instrumental environment is designed in accordance with CSSL standard. Library of original numerical solvers, embedded in simulation environment, is presented. Theorem about the choice of the integration step considering the HS event function dynamic has been formulated and proved. Algorithm of accurate HS event detection with implicit continuous behavior models is designed. Examples of specification and analysis of different hybrid systems models is given.

Keywords Computer aided analysis • Software architecture • Numerical simulation • Differential equations • Event detection and circuit simulation

8.1 Introduction

Hybrid systems (HS) theory is a modern and versatile apparatus for mathematical description of the complex dynamic processes in systems with different physical nature. Such systems are characterized by the composition of the continuous and discrete behaviors. Earlier the ISMA instrumental environment [1, 2] examined models and methods of HS analysis, continuous modes of which are described by the Cauchy problem with constraints. In this paper the extension of class of systems by models unresolved with respect to the derivative is proposed. Numerical analysis of the new class of problems requires using a specific integration and HS event detection algorithms. The described algorithms are implemented in the ISMA and successfully tested.

Yu.V. Shornikov (✉) • D.N. Dostovalov
Novosibirsk State Technical University, Novosibirsk, Russia

Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia
e-mail: shornikov@inbox.ru; dostovalov.dmitr@mail.ru

M.S. Nasyrova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: maria_myssak@mail.ru

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_8

8.2 Class of Systems

There are many systems (mechanical, electrical, chemical, biological, etc.), the behavior of which can be conveniently described as a sequential change of continuous modes. These systems are referred to as hybrid or event-continuous and described by system of differential Eq. (8.1) or differential-algebraic Eq. (8.2) equations with the constraints.

$$\begin{array}{l} y' = f\left(y,\ \ y\left(t-\tau\right),t\right), \\ pr:\ \ g\left(y,\ \ t\right) < 0, \\ t \in \left[t_0,\ \ t_k\right], y\left(t_0\right) = y_0, \\ y \in R^{N_y}, t \in R, \\ f: R^{N_y} \times R \to R^{N_y}, \\ g:\ \ R^{N_y} \times R \to R^S. \end{array} \tag{8.1}$$

$$\begin{array}{l} y' = f\left(x,\ \ y,\ \ t\right), x = \varphi\left(x,\ \ y,\ \ t\right), \\ pr:\ \ g\left(x,\ \ y,\ \ t\right) < 0, \\ t \in \left[t_0,\ \ t_k\right], x\left(t_0\right) = x_0, y\left(t_0\right) = y_0, \\ x \in R^{N_x}, y \in R^{N_y}, t \in R, \\ f:\ \ R^{N_x} \times R^{N_y} \times R \to R^{N_y}, \\ \varphi:\ \ R^{N_x} \times R^{N_y} \times R \to R^{Nx}, \\ g:\ \ R^{N_x} \times R^{N_y} \times R \to R^S. \end{array} \tag{8.2}$$

The vector-function $g(x, y, t)$ is referred to as event function or guard. A predicate pr determines the conditions of existence in the corresponding mode or state. Inequality $g\left(x, y, t\right) < 0$ means that the phase trajectory in the current mode should not cross the border $g\left(x, y, t\right) = 0$. Events occurring in violation of this condition and leading to transition into another mode without crossing the border are referred to as one-sided. Many practical problems are characterized by stiff modes, and the surface of boundary $g\left(x, y, t\right) = 0$ has sharp angles or solution has several roots at the boundary [2]. Numerical analysis of such models by traditional methods is difficult or impossible, as it gives incorrect results. Therefore it is necessary to use special methods to detect events accurately.

Computer analysis of these systems is typically performed in simulation tools, best of which are Charon (USA), AnyLogic (Russia), Scicos (France), MVS (Russia), Hybrid Toolbox and HyVisual (USA), DYMOLA (Sweden) and etc.

In the simulation of electrical circuits, processes of chemical kinetics, electromechanical processes and many other applications a necessity arises to numerically analyze HS, modes of which are given by stiff implicit systems of high-dimensional differential equations with strict one-sided constraint:

$$F\left(x,\ \ x',\ \ t\right) = 0, pr:\ \ g\left(x,\ \ t\right) < 0, t \in \left[t_0,\ \ t_k\right], x\left(t_0\right) = x_0, \tag{8.3}$$

where $x \in R^N$ is the vector of state variables, $t \in R$ is the independent variable, $F : R^N \times R^N \times R \to R^N$ is the vector-function of HS continuous behavior, $g : R^N \times R \to R$ is the event function, x_0 are the initial conditions.

The problem Eq. (8.2) is usually stiff that leads to the application of implicit numerical formulas required Jacobi matrix inversion. Due to the ease of implementation and good accuracy and stability properties Rosenbrock type methods [3, 4] are widely used in solving stiff problems.

8.3 Architecture of Instrumental Environment

Development of simulation languages, simulators, simulation systems, etc. is essentially influenced by the CSSL (continuous system simulation language) Standard 1968 [5]. Although forty years old, the structures defined in CSSL Standard are used up to now. The early CSSS standard determined basic necessary features for a simulator, the late developments to implicit systems fixed extended features for simulation systems—both referred as classical CSSL features. In 1968, the CSSL standard set first challenges for features of simulation systems, defining necessary basic features for simulators and a certain structure for simulators.

The CSSL standard also defines segments for discrete actions, first mainly used for modelling discrete control. So-called DISCRETE regions or sections manage the communication between discrete and continuous world and compute the discrete model parts. For incorporating discrete actions, the simulation engine must interrupt the ODE solver and handle the event. For generality, efficient implementations set up and handle event lists, representing the time instants of discrete actions and the calculations associated with the action, where in-between consecutive discrete actions the ODE solver is to be called. In order to incorporate DAEs and discrete elements, the simulator's translator must now extract from the model description the dynamic differential equations (derivative), the dynamic algebraic equations (algebraic), and the events with static algebraic equations and event time. In principle, initial equations, parameter equations and terminal equations (initial, terminal) are special cases of events at time t = 0 and terminal time. Some simulators make use of a modified structure, which puts all discrete actions into one event module, where CASE—constructs distinguish between the different events.

Simulation environment of complex dynamical and hybrid systems called ISMA (translated from Russian "Instrumental Facilities of Machine Analysis") is developed at the department of Automated control systems of Novosibirsk state technical university (Russia) [6].

Specification of hybrid systems is carried out using graphical and symbolic languages that are the system content of instrumental environment. Analytical content is provided by numerical methods and algorithms for computer analysis corresponding to the chosen class of systems and methods for solving these models. ISMA environment is developed subject to simplicity of description of dynamical

and hybrid models in the language that is maximally close to the object language. Main features of ISMA are the following:

- composition of hybrid models is carried out in visual structural-textual form;
- structural form of model description corresponds to the classical description of systems by block diagrams and includes all necessary components such as integrators, accumulators, amplifiers, signal sources, nonlinear elements and others;
- language of symbolic specification is approached maximal to the language of mathematical formulas;
- special module for specification of problems of chemical kinetics in the language of chemical reactions which automatically translates them into a system of differential equations;
- a variety of traditional and original numerical methods included methods that are intended for the analysis of ODE systems of medium and high stiffness;
- computer simulation in real time;
- graphic interpreter called GRIN provides a wide range of tools for analysis and visualization of simulation results such as scaling, tracing, optimization, displaying in the logarithmic scale and phase plane;
- extension of system functionality by adding new typical components and numerical methods.

Architecture of ISMA software package (Fig. 8.1) is designed [7] in accordance with CSSL to unify existing mathematical program software for analysis of problems in various object domains: chemical kinetics, automation, electricity, etc.

8.4 Library of Numerical Methods

Peculiarities of numerical analysis are defined by the configuration and implementation of the solver in the scheme interpreter. Solver is configured to numerical analysis not only of smooth dynamical systems but also systems with ordinary discontinuity and stiff systems [2, 8]. For the analysis of the stiff modes new original m-phasic methods of p-order (Table 8.1), developed by the authors, are included in the solver library.

Libraries of standard blocks and numerical methods are implemented as independent application modules that are loaded at run time.

This approach allows to allocate in the application programming interface (API) a set of functions and classes required for the implementation of element libraries and numerical methods. Using the API any user with basic knowledge of object-oriented programming able to develop and built in the system new typical elements and numerical methods without recompiling the entire system.

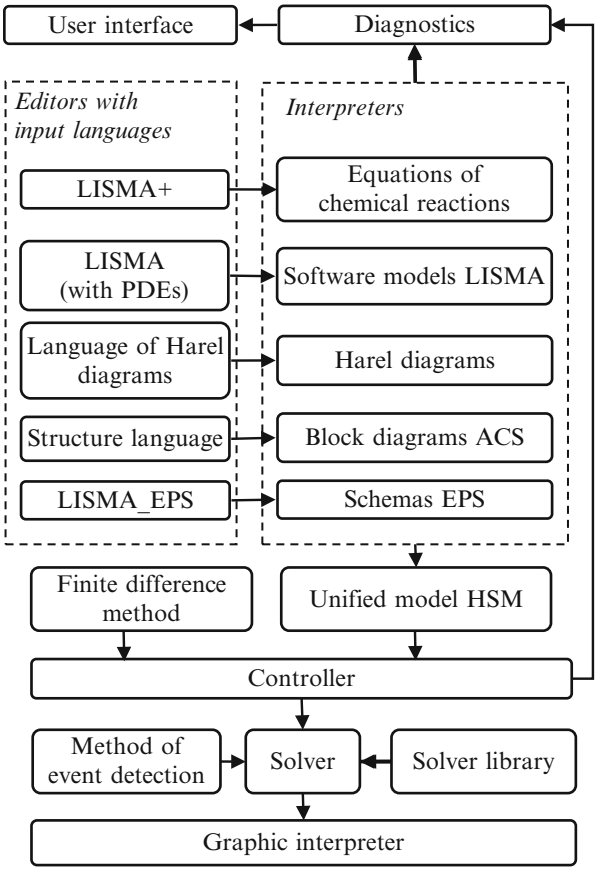


Fig. 8.1 ISMA architecture

8.5 Event Detection in Hybrid Systems

The correct analysis of hybrid models is significantly depends on the accuracy of detection of the change of the local states of the HS.

Therefore, the numerical analysis is necessary to control not only the accuracy and stability of the calculation, but also the dynamics of the event-function. The degree of approximation by the time the event occurred is defined by the behavior of event driven function.

Analyze the behavior of the event function $g(y,\ t)$ from Eq. (8.1). Let the method of the form $y_{n+1} = y_n + h_n\varphi_n$, where function φ_n is calculated in point t_n, is used for calculations.

Then the event-function $g(y,\ t)$ at point t_{n+1} has a form $g_{n+1} = g\ (y_n + h_n\varphi_n,\ t_n + h_n)$.

Table 8.1 Library of numerical methods

Method (p, m)	Comments
DISPF (5, 6)	Stability control, systems of medium and low stiffness
RADAU5 (3, 3)	Stiff systems
DISPF1_RADAU	Adaptive method DISPF in combination with RADAU5 with stiffness control, essentially stiff systems
DP78ST (8, 13)	Stability control, variable order and step, systems of medium stiffness and high precision
RKF78ST (7, 13)	Stability control, variable order and step, systems of medium stiffness and high precision
RK2ST (2, 2), RK3ST (2, 3)	Explicit methods with stability control for analysis of non-stiff systems
DISPS1	Algorithm of variable order with adaptive stability region
MK22 (2, 2), MK21 (2, 2)	Freezing of Jacobean matrix, stiff systems
MK11F	Algorithm of analysis of implicit problems

Decomposing the g_{n+1} in a Taylor series and taking into account the linearity of g_{n+1}, we obtain the dependence of g_{n+1} of the projected step h_n:

$$g_{n+1} = g_n + h_n \left(\frac{\partial g_n}{\partial y} \cdot \varphi_n + \frac{\partial g_n}{\partial t} \right). \tag{8.4}$$

Theorem 1. The choice of the step according to the formula

$$h_n = (\gamma - 1)\, g_n / \left(\frac{\partial g_n}{\partial y} \cdot \varphi_n + \frac{\partial g_n}{\partial t} \right), \gamma \in (0, 1)\,, \tag{8.5}$$

provides the event-dynamics behavior as a stable linear system, the solution of which is asymptotically approaching to the surface $g\,(x,\ t) = 0$.

Proof. Substituting Eqs. (8.5) in (8.4), we have $g_{n+1} = \gamma g_n$, $n = 0, 1, 2, \ldots$ Converting recurrently this expression we get $g_{n+1} = \gamma^{n+1} g_0$. Given that $\gamma < 1$, then $g_n \to \infty$ takes place when $n \to \infty$. In addition, condition $\gamma > 0$ implies that function g_n does not change sign. Therefore, when $g_0 < 0$, $g_n < 0$ will be valid for all n. Then the guard condition will never cross the potentially dangerous area $g\,(y_n,\ t_n) = 0$, which completes the proof.

8.5.1 Control of Event Function in the Integration Algorithm

Let the solution y_n at the point t_n is calculated with the step h_n. In addition, the new accuracy step h_{n+1}^{ac} is computed. Then the approximate solution at the point t_{n+1} is calculated as follows

Step 1. Calculate the functions

$$g_n = g\,(y_n,\ t_n)\,, \frac{\partial g_n}{\partial y} = \frac{\partial g\,(y_n,\ t_n)}{\partial y}, \frac{\partial g_n}{\partial t} = \frac{\partial g\,(y_n,\ t_n)}{\partial t}.$$

Step 2. Calculate $g'_n = \frac{\partial g_n}{\partial y}\varphi_n + \frac{\partial g_n}{\partial t}$, where $\varphi_n = f_n$.
Step 3. If $g'_n < 0$, then $h_{n+1} = h^{ac}_{n+1}$ and go to the Step 6.
Step 4. Calculate the new "Event" step h^{ev}_{n+1} by the formula

$$h^{ev}_{n+1} = (\gamma - 1)\,\frac{g_n}{g'_n}.$$

Step 5. Calculate the new step h_{n+1} by the formula
Step 6. Go to the next integration step.

In the Step 3, unlike the previously presented algorithm [9], we determine the direction of event-function change. Near the boundary regime denominator Eq. (8.4) will be positive, and away from the boundary $g\,(y,\ t) = 0$ it becomes negative. Then, defining the direction of event-function change, we do not impose any further restrictions on the integration step if the event-function is removed from the state boundary.

8.5.2 Test

To illustrate the event-detection algorithm we consider a hybrid system of two oscillating masses on springs [9], shown in Fig. 8.2.

The system can be in one of two local states: when masses move separately or together. Mathematical model is not presented here because of the proximity to description of the computer model. A computer model of system in the ISMA shown in Fig. 8.3.

Qualitative simulation results are obtained with enabled event-detection algorithm (Fig. 8.4). Traditional analysis of the system without using the event-detection algorithm does not allow to obtain valid results as shown in (Fig. 8.5).

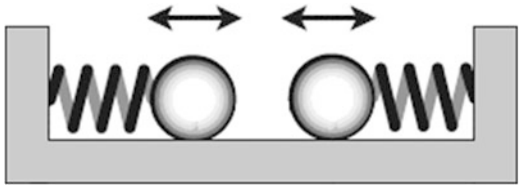


Fig. 8.2 The system of two oscillating masses

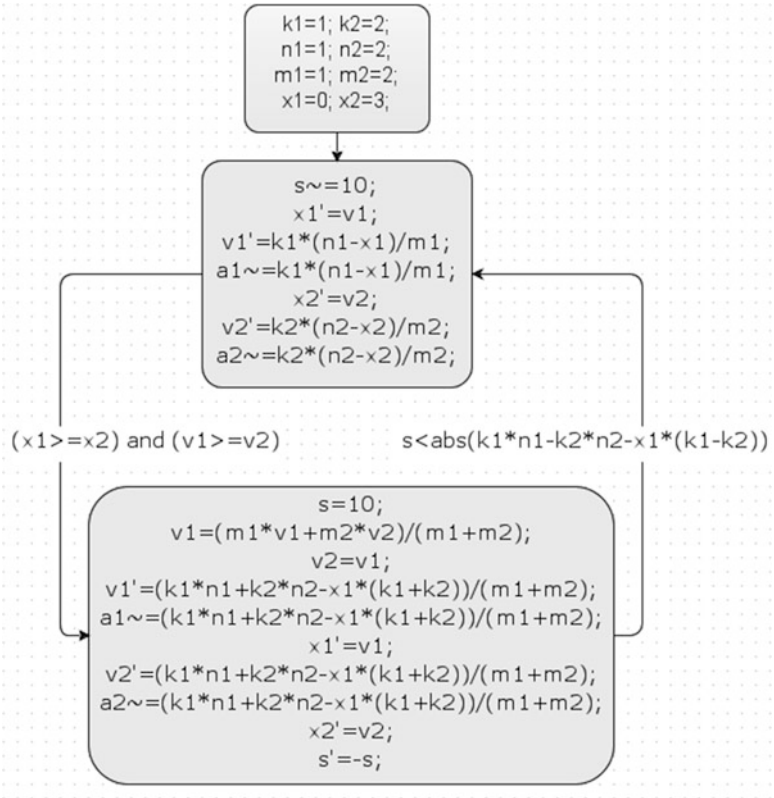


Fig. 8.3 A computer model of the system in ISMA instrumental environment

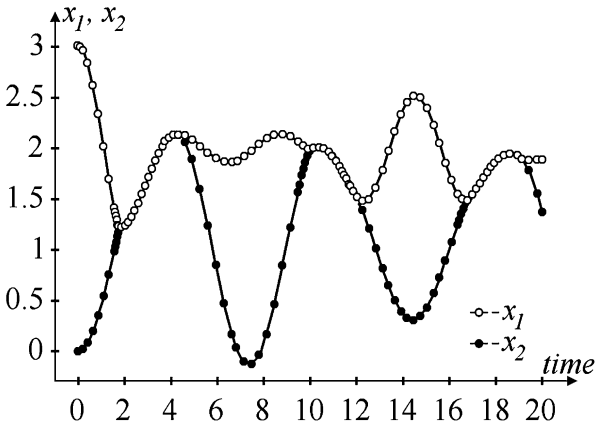


Fig. 8.4 Calculation results (using the event-detection algorithm)

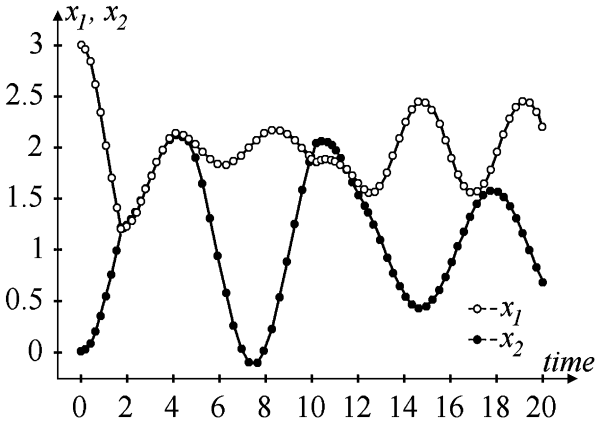


Fig. 8.5 Calculation results (excluding the event-function dynamics)

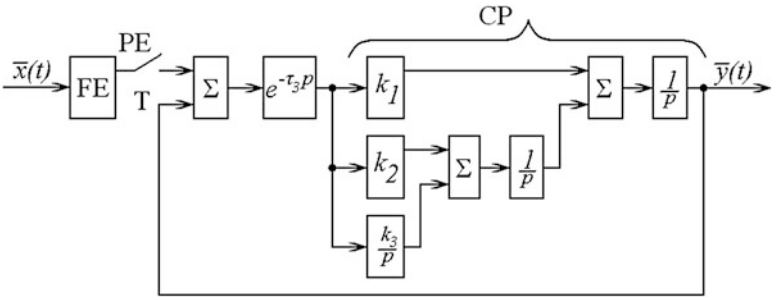


Fig. 8.6 Block diagram of autotracking system with delays

8.6 Analysis of Autotracking System

Block diagram of equivalent pulse system [10] given on the Fig. 8.6 and taking delays into account corresponds to the autotracking system of ballistic, space and aerodynamic objects.

The model of a discrete part of the system consists of a simple pulse element (PE) and the forming element (FE). The delay is taken into account by the dynamic unit with the transfer function $W(p) = e^{-\tau_3 p}$. Model of extrapolation and smoothing filter is represented by integrators constituting a continuous part (CP) of system model.

Switching to one integration step occurs at the moments of time $t = nT$, $n = 1, 2, 3\ldots$ with period T. Input action of second order is represented by parabola [10] as $\alpha(t) = -4 + 15t - 5t^2$ and implemented on the integrators with the phase variables x_6, x_7, x_8. Extended scheme of the state variables of the whole system, edited in the ISMA with the specified control logic is shown on the Fig. 8.7.

As opposed to the autotracking system structure presented in [10] the delay in ISMA is implemented by a typical unit DELAY which uses Pade approximation of the specified order depending on the accuracy.

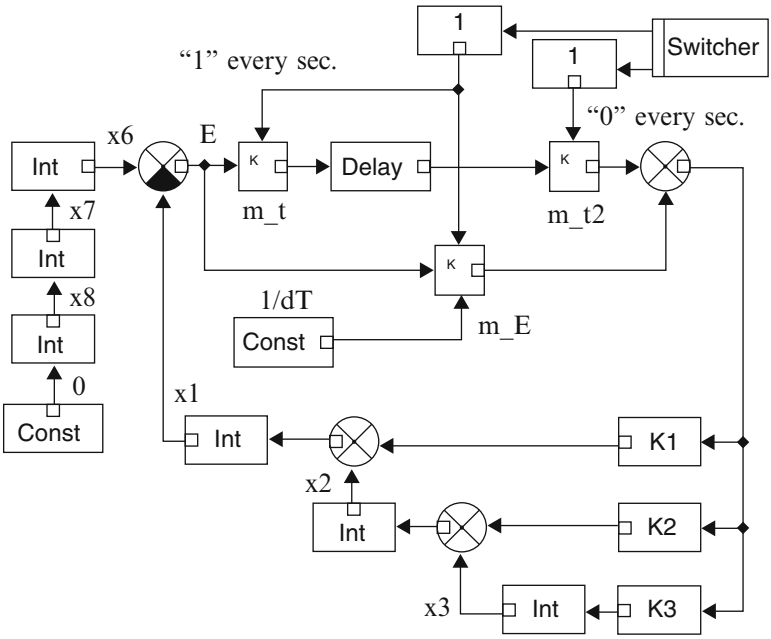


Fig. 8.7 Visual program model in ISMA environment

Switching logic in ISMA is provided by the Switcher macros with two control outputs with values of 0 or 1. The first output equals to 1 before the switching time and to 0 at the time of switching. The second output is the inverter of the first.

Simulation results are shown in the Fig. 8.8.

The traditional approach of pulse system simulation (for example, [11, 12]) has the significant disadvantage that the accepted in the control systems form of mathematical model representation in block diagrams becomes the art of structure composition for the applied user in a specific simulation environment. If the logic of continuous part control is complex the discrete part can be represented by a set of discrete control systems (structure in ISMA is shown on the Fig. 8.7, structure in SAU is in [10]) combining both continuous and discrete model part in a block diagram. Such structured visual models edited in different simulation environments becomes harder to read and sometimes available only to the author of a particular structure composition for the corresponding simulation environment.

Hybrid models [2] lack these disadvantages. The appearance of a new apparatus for specification and research of complex dynamical systems is caused by complexity of discrete control laws of continuous processes. Therefore traditional structural methods of presenting both continuous and discrete part in the same structure become insufficient.

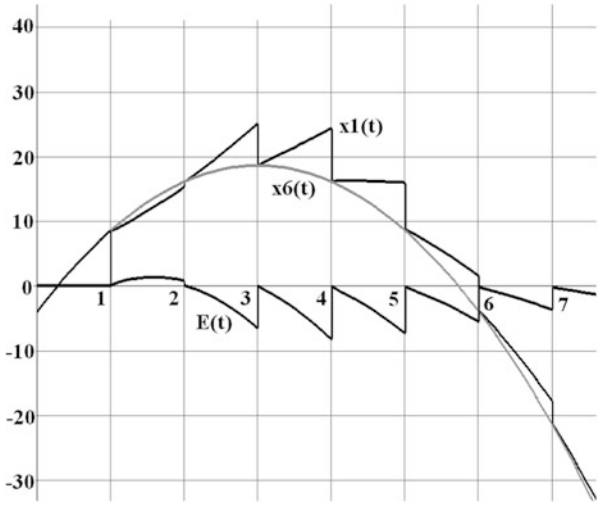


Fig. 8.8 State variables of autotracking system: x_6 is the measured trajectory, x_1 is the trajectory of tracking and E is the function of tracking errors

8.7 Model of Ring Modulator

As an example of differential-algebraic equations with the constraints Eq. (8.2) we selected the model of ring modulator [13] the scheme of which is shown in Fig. 8.9.

When receiving at input the low-frequency signal U_{in1} and high-frequency signal U_{in2}, the ring modulator generates mixed signal U_2 at output.

Type of problem depends on parameter C_s. If $C_s \neq 0$, we have a Cauchy problem for stiff system with fifteen ordinary differential equations. If $C_s = 0$, then we have a differential-algebraic system of index 2 consisting of eleven differential and four algebraic equations. In calculations we used the following parameters [13]: $C = 1.6 \cdot 10^{-8}$, $C_s = 2 \cdot 10^{-12}$, $C_p = 10^{-8}$, $L_h = 4.45$, $L_{s1} = 0.002$, $L_{s2} = 5 \cdot 10^{-4}$ $L_{s3} = 5 \cdot 10^{-4}$, $\gamma = 40.67286402 \cdot 10^{-9}$, $R = 25000$, $R_p = 50$, $R_{g1} = 36.3$, $R_{g2} = 17.3$, $R_{g3} = 17.3$, $R_i = 50$, $R_c = 600$, $\delta = 17.7493332$.

The computer model of the system, written down in language LISMA [14], is:

```
C1=1.6e-8;
Cs1=2e-12;
Cp1=1e-8;

Pi=3.14159265358979323846264338 3;

Lh1=4.45;
Ls11=2e-3;
Ls21=5e-4;
Ls31=5e-4;
```

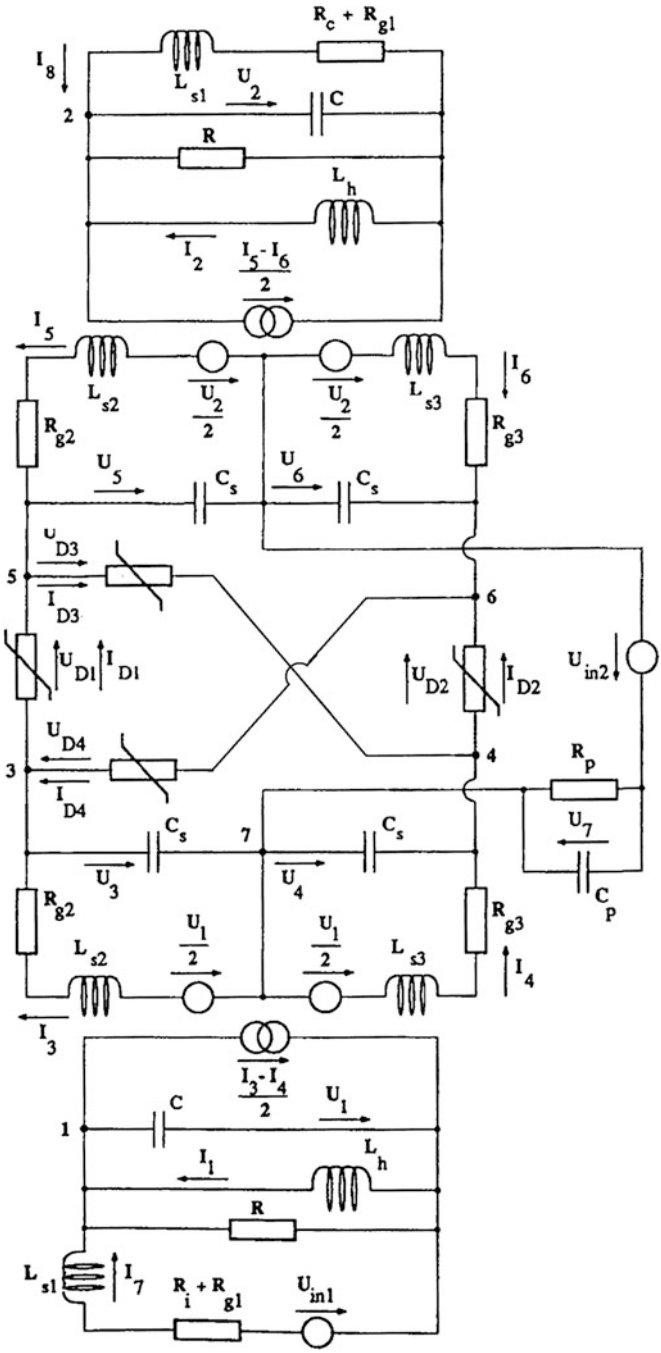


Fig. 8.9 Schematic diagram of ring modulator

```
gamma=40.67286402e-9;
Rg1=36.3;
Rg2=17.3;
Rg3=17.3;
Ri=50;
Rc=600;
d=17.7493332;

Rr1=25e+3;
Rp1=50;
Uin1~=0.5*sin(2000*Pi*TIME);
Uin2~=2.0*sin(20000*Pi*TIME);

Ud1~=y3-y5-y7-Uin2;
Ud2~=-y4+y6-y7-Uin2;

Ud3~=y4+y5+y7+Uin2;
Ud4~=-y3-y6+y7+Uin2;

Qd1~=gamma*(exp(d*Ud1)-1);
Qd2~=gamma*(exp(d*Ud2)-1);
Qd3~=gamma*(exp(d*Ud3)-1);
Qd4~=gamma*(exp(d*Ud4)-1);

y1'=(y8-0.5*y10+0.5*y11+y14-y1/Rr1)/C1;
y2'=(y9-0.5*y12+0.5*y13+y15-y2/Rr1)/C1;
y3'=(y10-Qd1+Qd4)/Cs1;
y4'=(-y11+Qd2-Qd3)/Cs1;
y5'=(y12+Qd1-Qd3)/Cs1;
y6'=(-y13-Qd2+Qd4)/Cs1;
y7'=(-y7/Rp1+Qd1+Qd2-Qd3-Qd4)/Cp1;
y8'=-y1/Lh1;
y9'=-y2/Lh1;
y10'=(0.5*y1-y3-Rg2*y10)/Ls21;
y11'=(-0.5*y1+y4-Rg3*y11)/Ls31;
y12'=(0.5*y2-y5-Rg2*y12)/Ls21;
y13'=(-0.5*y2+y6-Rg3*y13)/Ls31;
y14'=(-y1+Uin1-(Ri+Rg1)*y14)/Ls11;
y15'=(-y2-(Rc+Rg1)*y15)/Ls11;
```

The results of the computer model analysis by ISMA instrumental environment [6] are presented in Fig. 8.10.

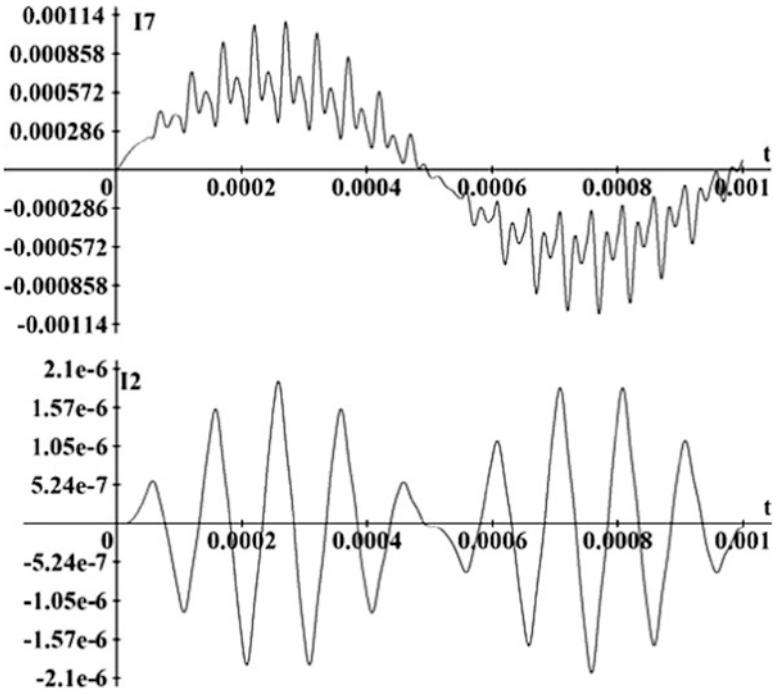


Fig. 8.10 Results of modeling in ISMA

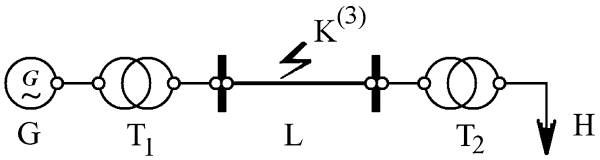


Fig. 8.11 Schematic diagram of the electrical network

8.8 Simulation of Fault in an Electrical Network

As an illustration of the system class Eq. (8.3) we analyze the model of three-phase short circuit. Schematic diagram of the electrical power system (EPS) built in graphics editor of the ISMA instrumental environment is shown in Fig. 8.11.

Considered scheme consists of generator G, transformers T_1, T_2, line L and load H. In the equivalent circuit in Fig. 8.12 capacitive conductivity of the line and transformer non-load loses are not taken into account and the load is taken into account by approximately active and inductive reactance.

Transient is initiated by the contact closure K. In this case previously established mode of power system is changed to the new mode corresponding fault and another system configuration. Thus, the model is a two-mode hybrid system (HS) [2].

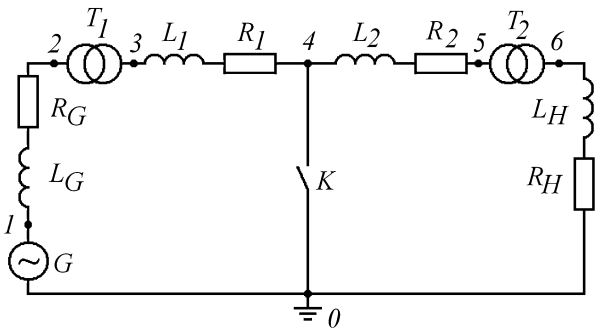


Fig. 8.12 Schematic diagram of the electrical network

Fig. 8.13 Behavior map

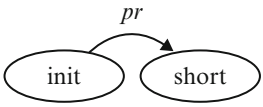
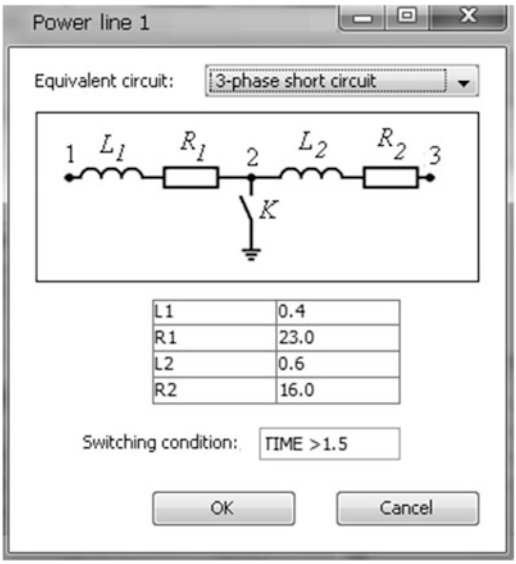


Fig. 8.14 Configuring the parameters of the equivalent circuit



The discrete behavior of the hybrid system is illustrated by the state chart [15] shown in Fig. 8.13. State init corresponds to the functioning of EPS before the fault. Switching to state short corresponded to the fault condition occurs when a logical predicate pr is carried out.

In the graphics editor of schematic diagrams of EPS hybrid behavior is specified in the configuration editor window for the equivalent circuit of a transmission line *L* as shown in Fig. 8.14.

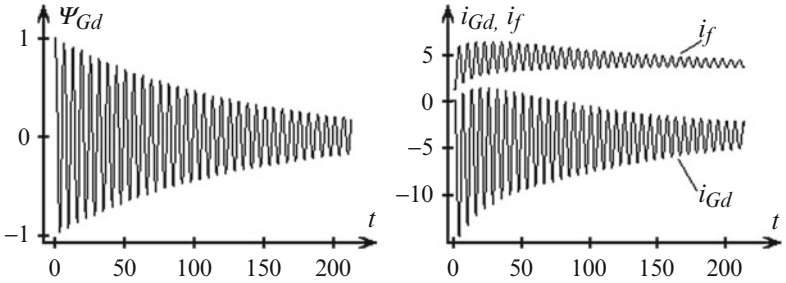


Fig. 8.15 Configuring the parameters of the equivalent circuit

The mathematical model is composed by Park-Gorev equations in rotating coordinate system (d, q) associated with the generator rotor G. Let the axis q is ahead of the axis d. Obtain an implicit system of equations for the generator G:

$$\begin{aligned}
&u_{1d}\cos(\theta - t) + u_{1q}\sin(\theta - t) + ri_{Gd} + \\
&+ L_d\frac{di_{Gd}}{dt} - L_{ad}\frac{di_f}{dt} - L_{ad}\frac{di_g}{dt} - \left(L_q i_{Gq} - L_{aq} i_h\right)\omega = 0, \\
&- u_{1d}\sin(\theta - t) + u_{1q}\cos(\theta - t) + ri_{Gq} + \\
&+ L_q\frac{di_{Gq}}{dt} - L_{aq}\frac{di_h}{dt} + \left[L_d i_{Gd} - L_{ad}\left(i_f + i_g\right)\right]\omega = 0, \\
&- u_f + r_f i_f + L_f\frac{di_f}{dt} + L_{ad}\frac{di_g}{dt} - L_{ad}\frac{di_{Gd}}{dt} = 0, \\
&r_g i_g + L_g\frac{di_g}{dt} + L_{ad}\frac{di_f}{dt} - L_{ad}\frac{di_{Gd}}{dt} = 0, \\
&r_h i_h + L_h\frac{di_h}{dt} - L_{aq}\frac{di_{Gq}}{dt} = 0, \\
&\frac{d\omega}{dt} = \frac{T_\partial + \left[\left(L_d - L_q\right)i_{Gq} + L_{aq} i_h\right]i_{Gd} - L_{ad} i_{Gq}\left(i_f + i_g\right)}{T_J} \\
&\frac{d\theta}{dt} = \omega.
\end{aligned} \tag{8.6}$$

Here the index f refers to the excitation winding and indices g and h refers to the longitudinal and transverse damper contours respectively. Equations for other areas of the equivalent circuit are conventional and therefore not given.

When an event corresponded to the fault occurs in HS, the voltage in point 4 is equated to zero $u_{4d} = u_{4q} = 0$. In this case in the equivalent circuit two independent contours are formed. The equations for sections of the contours remain the same. Plots of some state variables obtained in ISMA are shown in Fig. 8.15. Calculation results correspond to theoretical statements and coincide with results obtained in MATLAB.

8.9 Conclusions

In this paper the new class of hybrid systems within the ISMA instrumental environment, the modal behavior of which is defined by a system of ODE with delays, differential-algebraic and implicit systems with nonlinear differential

equations are introduced. Architecture of instrumental environment is designed in accordance with CSSL standard adapt to the current configuration of computing technology and the designated problem classes. The new original method of switching point's localization is proposed. The algorithm easily complements the existing numerical solvers based on explicit and semi-explicit schemes including the proposed algorithm of implicit problem's analysis. Models of HS are presented and studied in ISMA.

Acknowledgments This work was supported by grant 14-01-00047-a from the Russian Foundation for Basic Research, RAS Presidium project № 15.4 "Mathematical modeling, analysis and optimization of hybrid systems". Yu.V. Shornikov is with the Design Technological Institute of Digital Techniques Siberian Branch of Russian Academy of Science, Novosibirsk, Russia (e-mail: shornikov@inbox.ru). M.S. Myssak, D.N. Dostovalov is with the Department of Automated Control Systems, Novosibirsk State Technical University, Novosibirsk, Russia (e-mails: maria.myssak@gmail.com, dostovalov.dmitr@gmail.com).

References

1. Shornikov, Yu.V., Tomilov, I.N., Dostovalov, D.N., Denisov, M.S.: Numerical modeling of dynamic processes in electric power systems as a strategic management tool. Scientific Bulletin of the NSTU **4**, 129–134 (2011)
2. Novikov, E.A., Shornikov, Yu.V.: Computer simulation of stiff hybrid systems: monograph. Publishing house of NSTU, Novosibirsk (2012)
3. Rosenbrock, H.H.: Some general implicit processes for the numerical solution of differential equations. The Computer Journal **5**, 329–330 (1963)
4. Shornikov, Yu.V., Dostovalov, D.N., Myssak, M.S.: Simulation of hybrid systems with implicitly specified modal behavior in the ISMA environment. Humanities and Science University Journal **5**, 175–182 (2013)
5. Breitenecker, F., Popper, N.: Classification and evaluation of features in advanced simulators. In: MATHMOD, Vienna (2009)
6. Shornikov, Yu.V., Druzhinin, V.S., Makarov, N.A., Omelchenko, K.V., Tomilov, I.N.: Official Registration License for Computers 2005610126. Rospatent, Moscow (2005)
7. Shornikov, Yu.V., Myssak, M.S., Dostovalov, D.N., Bessonov, A.V.: Using ISMA simulation environment for numerical solution of hybrid systems with PDE. Computer Modeling and Simulation **4**, 101–108 (2014)
8. Dostovalov, D.N.: Computer simulation and algorithms of numerical analysis of hybrid systems. Control System and Information Technologies **53**(3.1), 128–133 (2013)
9. Shornikov, Yu.V., Dostovalov, D.N., Myssak, M.S., et al.: Specification and analysis of discrete behavior of hybrid systems in the workbench ISMA. Open Journal of Applied Sciences. **3**(2b), 51–55 (2013)
10. Arseniev, G.N.: Synthesis and Analysis of Autotracking Systems with Delays in Radio-Electronic Facilities. Information-Measuring Systems **3**(2), 25–31 (2005)
11. Oltean, V.E., Dobrescu, R., Popescu, D., Nicolae, M.: On a modal approach for oscillations damping in affine and piecewise affine systems. International Journal of Systems Applications, Engineering & Development **6**(1), 1–8 (2012)
12. Popescu, M.-C.O.S., Mastorakis, N.E.: Testing and Simulation of a Motor Vehicle Suspension. International Journal of Systems Applications, Engineering & Development **3**(2), 74–83 (2009)

13. Kampowski, W., Rentrop, P., Schmidt, W.: Classication and Numerical Simulation of Electric Circuits. Surveys on Mathematics for Industry **2**(1), 23–65 (1992)
14. Shornikov, Yu.V., Tomilov, I.N.: The program of language processor from language LISMA. Official registration license for computers № 2007611024. Rospatent, Moscow (2007)
15. De Leeuw, B., Hoogewijs, A.: Statechart normalizations. In: WSEAS Trans. Inf. Sci. Appl. **7**(11), 1358–1367 (2010)

Chapter 9
New Methods of Complex Systems Inspection: Comparison of the ADC Device in Different Operating Modes

Raoul R. Nigmatullin, Yury K. Evdokimov, Evgeny S. Denisov, and Wei Zhang

Abstract The authors suggest a general concept for quantitative inspection of complex systems (when the "best fit" model is absent) data in one unified scheme with the help of the sequence of ranged amplitudes (SRA). Moreover, the "up" and "down" branches of SRA distribution can replace a conventional histogram (having uncontrollable errors) and can be expressed in terms of the fitting parameters that are associated with a combination of power-law functions. As an example of a complex system we considered an analog-to-digital convertor having 16 channels. For four compared different operating modes of this device the calculated SRAs have different behavior and the significant quantitative parameters found enable to differentiate all these regimes from each other. We hope that this new approach will find a proper place in analysis of different complex systems and in different engineering applications, where the urgent necessity in quantitative comparison of complex systems without model exists.

Keywords Intermediate model • Data/signal processing • Measurements with/without memory • Sequence of the ranged amplitudes

R.R. Nigmatullin • Y.K. Evdokimov • E.S. Denisov (✉)
Radioelectronics and Information & Measuring Techniques Department, Kazan National Research Technical University named after A.N. Tupolev–KAI
10 Karl Marx str., 420111 Kazan, Tatarstan, Russia
e-mail: genia-denisov@yandex.ru; renigmat@gmail.com

W. Zhang
Department of Electronic Engineering, College of Information Science and Technology, Jinan University
Shi-Pai, Guangzhou, Guangdong, China

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_9

9.1 Introduction

The section of experimental physics associated with treatment of different data is considered as well-developed and to suggest some new and general ideas or principles that can touch the grounds of this section it seems rather difficult. One can numerate a lot of excellent monographs written by prominent scientists (mathematicians, experimentalists, specialists in statistics of different kind and etc.) [1–10] then we can add here a lot of journals that publish the peer reviewed information in this area and also the specialized conferences that every year evaluates new achievements in this important area. The fresh information related to recent achievement in the fractal signal processing is collected in books [11–14]. This information "explosion" creates a "rigid" trend and it definitely increases the limits of applicability of many methods that are developed in the area as processing/treatment of different random sequences and signals. The chaotic and random phenomena are originated from variety of reasons and their specificity dictates different methods for their quantitative description. The first author (RRN) of this paper also tried to develop different methods that proved their efficiency in solution of many complex situations [15–20], where the conventional methods did not work properly.

All data can be divided into two large classes: reproducible and unreproducible data, accordingly. In the first case an experimentalist enables to reproduce relatively stable conditions of his experiment and can measure the response of the system (object) studied again in the same period of time with some accuracy. For the second type of data (economical, meteorological, geological, biological, medical, etc.) the reproducibility of the same external conditions become impossible and many special methods for analysis of different time series were suggested [11–23]. But the needs of the science associated with analysis of different complex systems require the methods that should be error controllable (at least in any treatment/processing procedure applied to analysis of different complex data) and they should contain a minimal number of the fitting parameters for the functions that, in turn, follow from some general principles. These principles as self-similarity [24–26], quasi-periodicity [27–29] and others should be free from the specific models applied to analysis of different complex systems.

In this paper we want to apply for quantitative analysis of reproducible random data the sequences of the ranged amplitudes (SRA). They are obtained easily if the amplitudes of the random sequence analyzed are located in the descending order, i.e. $y_1 > y_2 > \ldots > y_N$. Here and below the index $j = 1, 2, \ldots, N$ numerates the number of the measured points. The division of the SRA forms a distribution of positive/negative amplitudes relatively the mean value. We want to show its universal behavior and the calculated SRAs can be used as a tool of replacement the corresponding histograms having uncontrollable errors in their constructions.

The content of the paper is organized as follows. In Sect. 9.2 we formulate a new approach and give the general algorithm in construction of the SRAs that allows treating all reproducible data in the unified scheme. Section 9.3 contains

the description of experiment and confirms the applicability of this algorithm to analysis of the ADC data in its different operating modes. Section 9.4, as usual, contains the basic results and outlines the perspectives of the future research in different engineering area.

9.2 General Description of Data in the Frame of SRA Approach

Let us remind some important points that are necessary for understanding of the new concept. Under ideal experiment we understand the measured response from the object studied (during the period of time T) that is reproduced in each measurement with the same accuracy. If $Pr(x)$ is chosen as the response (measured) function then from the mathematical point of view it implies that the following relationship is satisfied

$$y_m \cong Pr(x + m \cdot T_x) = Pr(x + (m-1) \cdot T_x), m = 1, 2, \ldots, M. \tag{9.1}$$

Here x is the external (control) variable, T_x is a "period" of experiment expressed in terms of the control variable x. If $x = t$ coincides with temporal variable then $T_x = T$ coincides with conventional definition of a period. The solution of this functional equation is well-known and (in case of discrete distribution of the given data points $x = x_j$, $j = 1, 2, \ldots, N$) it coincides with the segment of the Fourier series. It can be written as

$$Pr(x) = A_0 + \sum_{k=1}^{K \gg 1} \left[Ac_k \cos\left(2\pi k \frac{x}{T_x}\right) + As_k \sin\left(2\pi k \frac{x}{T_x}\right) \right]. \tag{9.2}$$

We deliberately show only the segment of the Fourier series because in reality all data points are always discrete and the number of "modes" k (coinciding with the coefficients of the Fourier decomposition) is limited. We define here and below by the capital letter K the finite mode. This final mode K is chosen from the requirement that it is sufficient to fit experimental data by expression (9.2) with the given (or acceptable) accuracy. As we will see below the value of K can be calculated from the expression (8) for the relative error located in the given interval [1–10 %]. But in reality the condition of an "ideal" experiment cannot be realized. In reality (as it has been confirmed on many available data) we have the following functional equation

$$F(x + m \cdot T_x) = a_m F(x) + b_m, m = 1, 2, \ldots, M,$$

$$F(x) = \frac{1}{M} \sum_{m=1}^{M} F(x + m \cdot T_x), \tag{9.3}$$

where the parameters a_m and b_m coincide with some real constants and depend on the number of the current measurement m. The letter M defines the total number of measurements. Here we suppose that the initial measurement coincides with the mean measurement $F(x)$. The general solution of the functional equation (9.3) has been considered in the first time in paper [33] but in the present paper we do not consider this solution and concentrate on the distribution of the slopes that follows from (9.3). The quasi-periodic solutions associated with Prony decomposition and their confirmations on available data were considered in papers [27–29]. This distribution of slopes is calculated easily and has a form

$$\begin{aligned} SI_m &= slope\,(F(x), F(x + mT_x)) \equiv a_m, \\ b_m &= intercept\,(F(x), F(x + mT_x))\,. \end{aligned} \tag{9.4}$$

The distribution of intercepts b_m in this case is not important for further purposes and can be omitted. This distribution allows to calculate the desired SRA when all slopes are located in the descending order ($Sl_1 > Sl_2 > \ldots > Sl_M$). For this SRA we obtain the desired distribution of the measurements that characterizes the quality of the measurements performed. We want to stress here that this distribution has a principle differences from the conventional histograms. These differences are the following: (a) the desired SRA does not require an arbitrary selection of the histogram column while in construction of histograms the selection of the proper width of the column plays an important role. In the case of the calculation of the desired SRA the "width" of the column is set up by the accuracy of the significant digits used in calculation of the data analyzed; (b) The lengths of the "up" and "down" branches of the SRA relatively mean value (equaled 1) cannot coincide with each other near the mean value and has a specific jump. These two peculiarities that present in the behavior of the SRA allow considering these two ("up" and "down") branches separately. One of the authors (RRN) has proved [30, 31] that the branches can be identified by two power-law function

$$Y(x) = A_0 + A_1 x^{\nu_1} + A_2 x^{\nu_2}, x \equiv m \in [1, M], \tag{9.5}$$

with the help of the eigen-coordinates method (ECs) [32, 33]. This method allows to reduce the calculation of the nonlinear power-law exponents ($\nu_{1,2}$) to the LLSM with the help of the basic linear relationship (BLR). The idea of calculation of the BLR by power-law functions and eigen-coordinates for the function (9.5) was considered in paper [32]. In Figs. 9.1, 9.2, 9.3, 9.4 9.5, and 9.6 we explain the successive stages in construction of the desired SRA for the first channel of the ADC and their fit to expression (9.5). We want to stress here that attentive analysis allows selecting the important quantitative parameter that can characterize the quality of any measurement. Really, for an "ideal" experiment all slopes should be equaled to the unit value. For the "accurate" device the maximal value of the range between "up" (defined as $1 + \Delta_{up}$) and "dn" ($1 - \Delta_{dn}$) branches of the slopes distribution

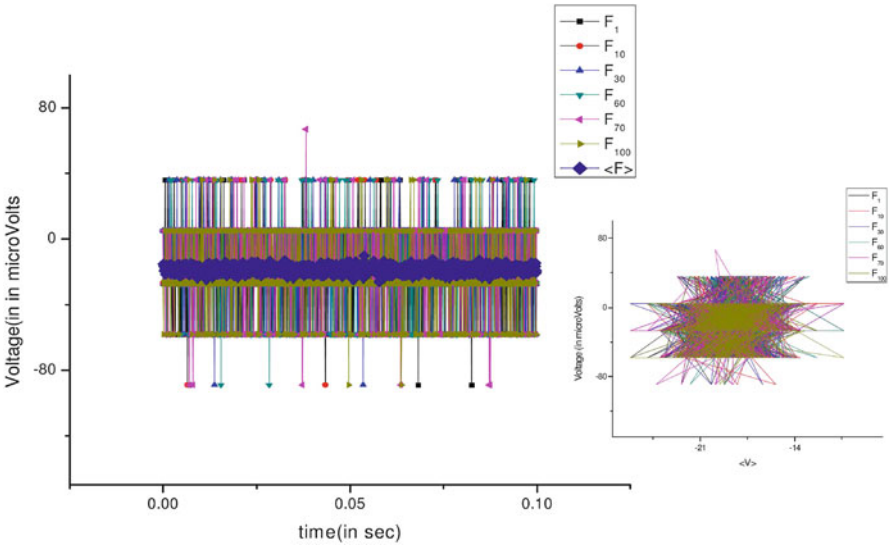


Fig. 9.1 The original response (Voltage) (in the microvolts units) with respect to time (in s) registered for the first channel. The high-frequency fluctuations destroy the memory that exists between successive measurements. The small figure on the *right* clearly demonstrates the absence of strong correlations (memory) between successive measurements

$$Rs = \max(\text{“}up\text{”}) - \min(\text{“}dn\text{”}) = \Delta_{up} + \Delta_{dn} \equiv \Delta_{R_s}, \tag{9.6}$$

should accept minimal values ($0.01 < \Delta_{R_s} < 0.1$) while for other devices this quantitative measure should be used for comparison of different devices with each other or/and the operating modes of the same device in different external conditions. In addition, one can add five fitting parameters (for each branch) that follow from the fit to expression (9.5). Besides, the SRA that follows from consideration of different slopes one can consider as an alternative distribution that follows from analysis of distribution of the differences between maximal and minimal values of the response function obtained in each measurement. This random function is defined as

$$Rmx_m = \max(F(x + mT_x)) - \min(F(x + mT_x)) \equiv \Delta_{Rmx}. \tag{9.7}$$

The SRA calculated for this function and its range (defined quantitatively as Δ_{Rmx}) are considered as an additional distribution. For the second channel (defined as number 5) we show only their two distributions (slopes and extreme values). They are shown in Figs. 9.7 and 9.8, accordingly.

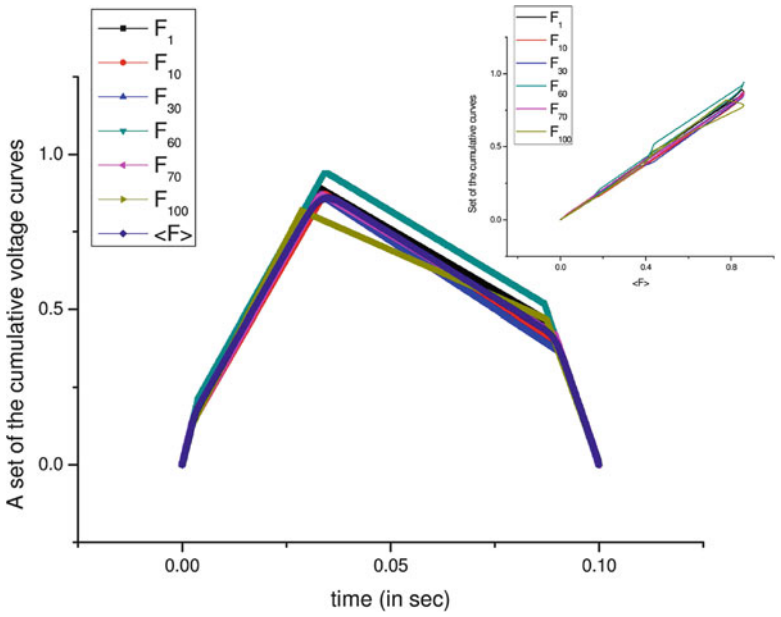


Fig. 9.2 After formation of the SRA for the original measurements (shown on the previous Fig. 9.1) and their subsequent integration we obtain the *cumulative curves* depicted on the central figure. Being plotted with respect to mean measurement we obtain the *strongly-correlated curves* shown on the small figure above. The high-frequency fluctuations are eliminated and the *curves* obtained make a set close to the segments of *straight lines*. For this set one can calculate the distribution of the slopes with respect to mean measurement

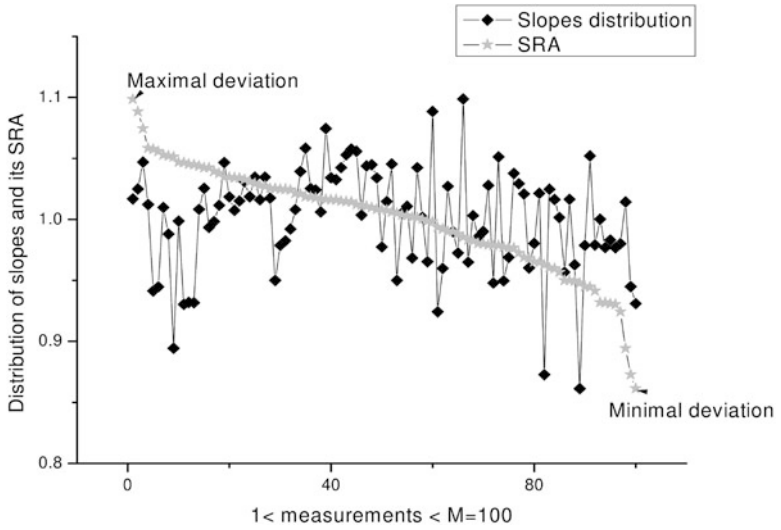


Fig. 9.3 Distribution of the desired slopes that are calculated with the help of expression (9.4) for all realized measurements ($M = 100$). The sequence of the ranged amplitudes is shown by the *grey stars*. The mean value of this distribution is equaled to one

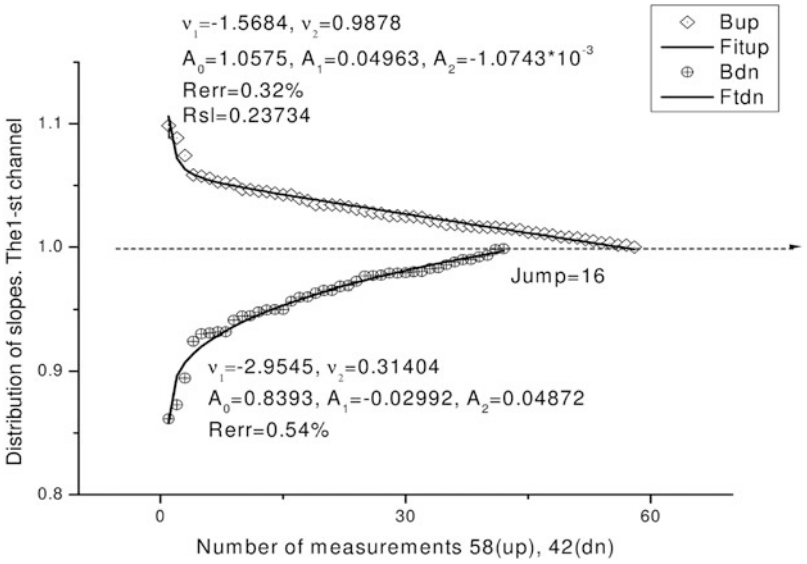


Fig. 9.4 Here we show the distribution of the slopes with respect to its mean value (=1) and fit the "up" and "dn" branches to the power-law function (9.5). All necessary fitting parameters are shown correspondingly in the *up* and *down parts* of this figure. The maximal value of the range for this distribution is equaled to $Rsl = 0.23734$. The value of the jump between branches is equaled to $58(up) - 42(dn) = 16 points$. In complete analogy with two previous figures one can consider the distribution of the ranges associated with (9.7)

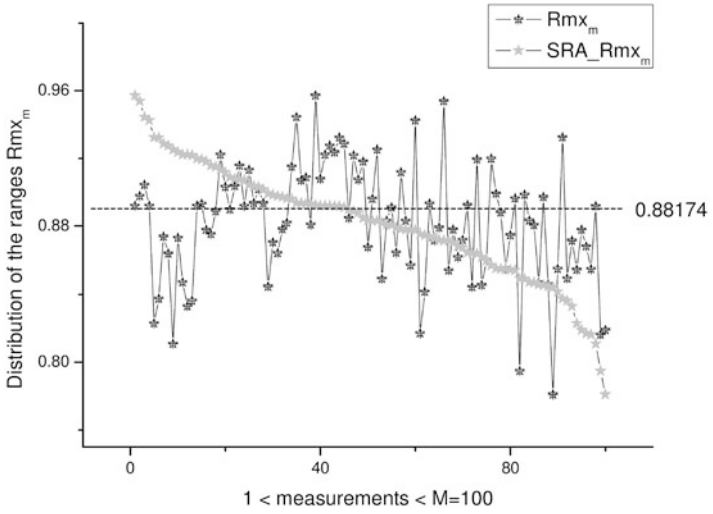


Fig. 9.5 Here we demonstrate the distribution of the extreme values calculated in accordance with expression (9.7) for the cumulative curves (see Fig. 9.2). The calculated SRA is marked by *grey stars*. The mean value of this distribution is defined as (=0.88174). After these preliminary calculations one can evaluate the corresponding distribution similar to Fig. 9.4

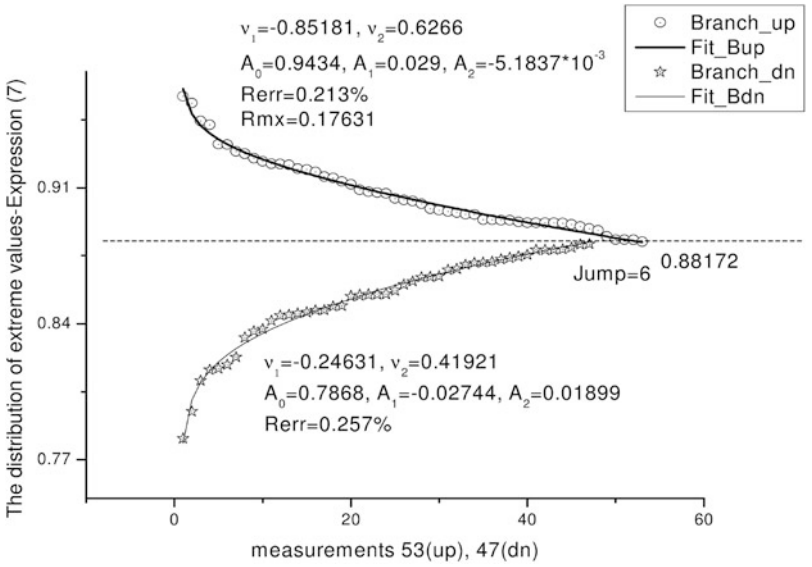


Fig. 9.6 Here we show the distribution of the extreme values for the first channel (expression (9.7)) with respect to its mean value (=0.88172) and fit the "up" and "dn" branches to the power-law function (9.5). All necessary fitting parameters are shown correspondingly in the *up* and *down parts* of this figure. The maximal value of the range for this distribution is equaled to $Rmx = 0.17631$. The value of the jump between branches is equaled to $53(up) - 47(dn) = 6 points$

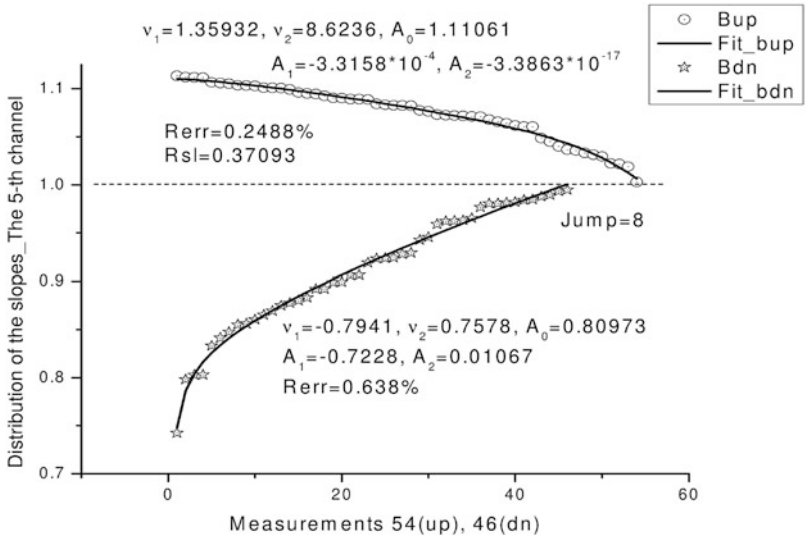


Fig. 9.7 Here we show the distribution of the slopes for the fifth channel fit the "up" and "dn" branches to the power-law function (9.5). All necessary fitting parameters are shown correspondingly in the *up* and *down parts* of this figure. The maximal value of the range for this distribution is equaled to $Rsl = 0.37093$. The value of the jump between branches is equaled to $54(up) - 46(dn) = 8 points$

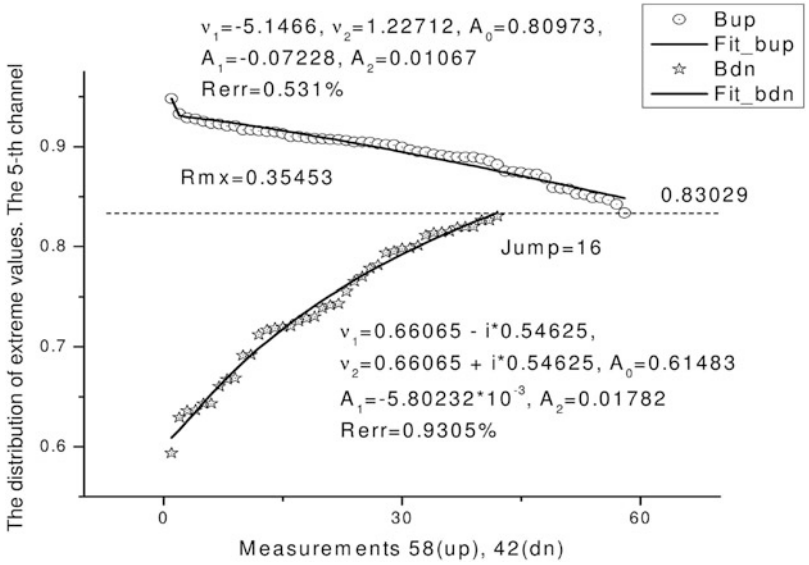


Fig. 9.8 Here we show the distribution of the extreme values for the fifth channel (expression (9.7)) with respect to its mean value (=0.83029) and fit the "up" and "dn" branches to the power-law function (9.7). All necessary fitting parameters are shown correspondingly in the *up* and *down parts* of this figure. The maximal value of the range for this distribution is equaled to $Rmx = 0.35453$. The value of the jump between branches is equaled to $58(up) - 42(dn) = 16 points$. It is interesting to note that for the down branch the power-law exponents accept the complex-conjugated values

9.3 Experimental Part: Data for the ADC in Different Regimes and Their Analysis

9.3.1 The Description of the Measurement Procedure

For the testing of the proposed algorithms the experiment consisting in quantitative comparison of intrinsic electrical fluctuations of the chosen ADC has been carried out. The selection of ADC is evoked by the fact that the ADC is one of the main parts for the increasing part of modern measurement systems. In other words, it allows supposing that the results of this experiment can be interesting for scientists and engineers specialized in different fields and we want to prove that the proposed approach can be applied in different applications.

Therefore, one of the typical ADC, namely, ADC of NI PXIe-6368 Simultaneous X Series Data Acquisition [34] produced by the National Instrument, Inc. has been used as a device under test for implementation of the proposed approach. NI PXIe-6368 comprises 16 differential simultaneous analog inputs with 16-bit resolution. Maximum sampling frequency is limited by the value of 2 MHz per channel.

Traditionally, it is assumed that all channels of ADC are approximately equivalent to each other. However, it is clear that quality of each specific channel that can be caused for example by imbalance of input cascades, by imperfections of soldered joints, and by other random and uncontrollable reasons. Within these frameworks the vicinity of parameters of the channels can be used as ADC quality criteria. However traditional approaches instead of the proposed one do not allow estimating quantitatively the differences between fluctuation characteristics of the channels compared.

Thus, the main purpose of the experimental investigation is to apply the proposed approach to compare the different ADC channels operating under different operation modes. For this purpose the schemes for registration of intrinsic voltage fluctuations of ADC channels for two different ADC setups are shown in Fig. 9.9, accordingly.

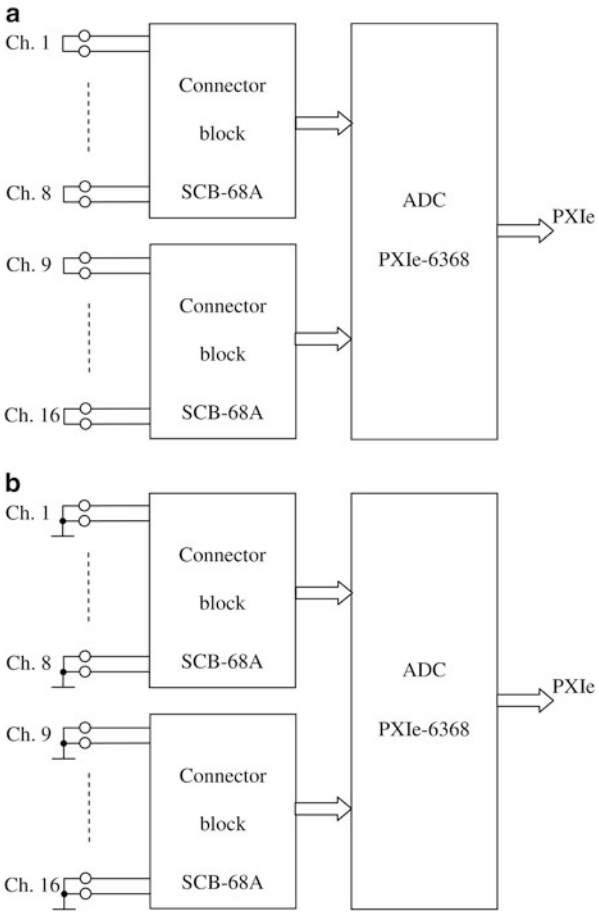


Fig. 9.9 The experimental setup: (**a**) the first configuration short circuit of ADC differential input terminals without grounding, and (**b**) the second configuration short circuit of ADC differential input terminals with grounding

It can be seen that both configurations use SCB-68A connector block to implement one of two input configurations: (a) short circuit of differential input terminals without grounding Fig. 9.9a and short circuit of differential input terminals with grounding (Fig. 9.9b). These two configurations differ by the operation modes of input cascades of the ADC. Based on the general principles of ADC operation it can be preliminary expected that the second configuration should provide more stable ADC fluctuation characteristics. For example, the first configuration (a) can be more sensitive to external interferences in comparison with the case (b). However, it should be mentioned that the experimental installation has been covered by effective Faraday cage to reduce random interferences.

Here it should be noted that ADC has been investigated at 0 V input voltage produced by shot circuit of input terminals to exclude influence of any additional elements on the results of analysis. However, it cannot be considered as a limitation of the proposed approach; the suggested setup can be used obviously for other input voltages and different types of signals.

During experiment the ADC under test has operated with sampling frequency 1 MHz according to the following protocol: (1) ADC iteratively acquires array of 100,000 samples for each of 16 channels during 0.1s and transmits each acquired array of samples into LabVIEW based program; (2) The program collects the acquired arrays by the following way: add one acquired array of samples in predetermined time interval T to the file stored on hard disk (other arrays are ignored to reduce volume of the stored data); (3) The measurement is stopped when N arrays stored, there $N = 100$ is determined as the fixed value. This measurement procedure is shown schematically in Fig. 9.10. The arrays of samples within the stored file have been decimated with the factor 100 to reduce the volume of stored data and decrease computational costs.

It has been used two types of measurements on the basis of the mentioned protocol: short-time ($T = 1$ min) and long-time ($T = 50$ min) for each of two considered ADC configurations.

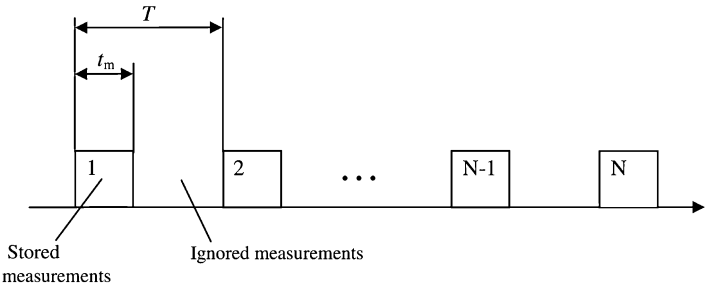


Fig. 9.10 Time diagram of the measuring procedure: (1) number of measurements (N); (2) period between measurements (T); (3) measurement duration $t_m = f_s \cdot N_s$, where f_s is the chosen sampling frequency, N_s is number of samples corresponding to a single measurement

Hence, during experiment it has been obtained information about four different regimes of the 16-bit 16 channel ADC of NI PXIe-6368 operated with 1 MHz sampling frequency:

1. short-time operation of 16 channels of ADC in the first configuration: 16 differential channels of ADC are closed without grounding; the stored data comprises 16 files, each file includes 100 iterations of the array containing 1,000 time-counts obtained as result of decimation of measured samples by the factor 100; time between iteration is 1 min (total duration of exposition is 100 min). This set of measurements is defined as: mode 1–16ch_wgr_shtm;
2. short-time operation of 16 channels of ADC in the second configuration: similar to regime 1 except that the second configuration is used, i.e. the input terminals are grounded that makes the effect of the interference induction negligible (total duration of exposition is 100 min). This set of measurements is defined as: mode 2–16ch_gr_shtm;
3. long-time operation of 16 channels of ADC in the first configuration: similar to regime 1 except that time between iteration is 50 min (total duration of exposition is about 83 h). This set of measurements is defined as: mode 3–16ch_wgr_lgtm;
4. long-time operation of 16 channels of ADC in the second configuration: similar to regime 2 except that time between iteration is 50 min (total duration of exposition is about 83 h). This set of measurements is defined as: mode 4–6ch_gr_lgtm.

9.3.2 The Description of the Treatment Procedure

Based on the basic approach described in Sect. 9.2 we have a possibility to compare four operating modes described in the previous section. For quantitative comparison of these ADC modes one can suggest the following algorithm. It comprises the following five basic steps.

S1. The construction of the distribution of slopes in accordance with expression (9.4) for the given number of measurements M. For practical calculations it is necessary to have sufficient number of measurements M (for practical calculations we choose the value $M = 100$). As an example we show the distribution of measurements for the mode 1 file—16ch_wgr_shtm.
S2. The construction of the SRA corresponding to the ordered measurements. This set of measurements is important because it shows the range of the measurements. The curves corresponding to four mode files for the channel 1 are shown in Fig. 9.11.
S3. The ordered set of measurements helps to form the desired branches (“up” and “dn”) bisected with respect to its mean value. For the case of distribution of slopes the corresponding mean value = 1. These branches are shown for modes 1, 2 and 3, 4 in Figs. 9.12 and 9.13, correspondingly. The most important quantitative parameter is defined by expression (9.6). This parameter

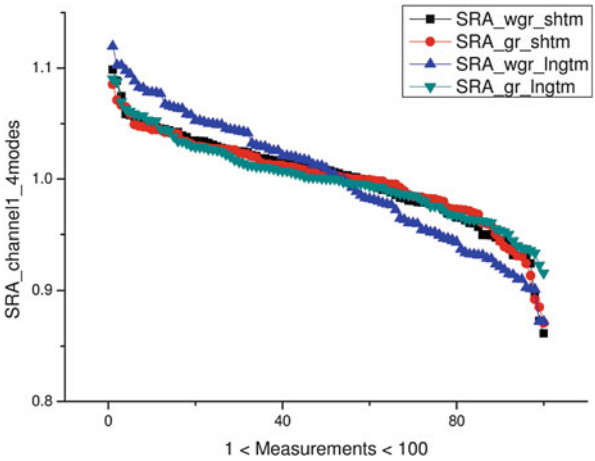


Fig. 9.11 Here we show the distributions of measurements for the channel 1 corresponding to four operating modes: mode 1: without grounding with short time exposition, mode 2: grounding with short time exposition, mode 3: without grounding with long time exposition, mode 4: grounding with long time exposition

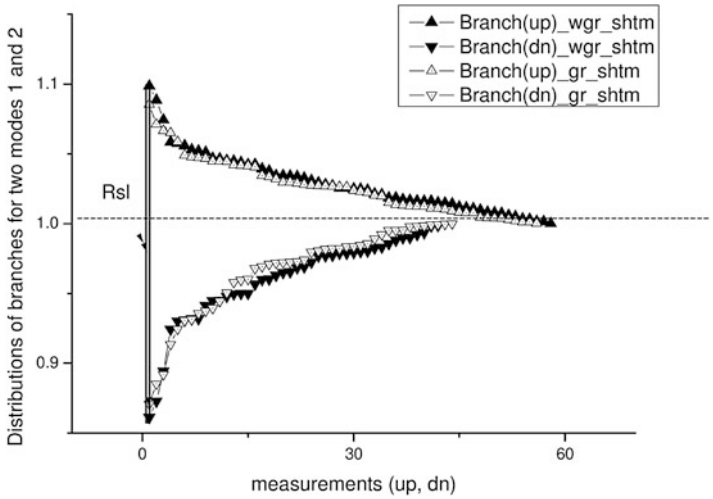


Fig. 9.12 The distribution of "up" and "dn" branches for the first two modes. The most important parameter that should be taken into account is related to the range of the distribution that is defined as the difference between maximal and minimal values

can characterize the quality of the channel for four different operating modes analyzed. We show the variation of this important parameter on Fig. 9.14. The minimal value of this parameter is associated with the quality of the channel analyzed. If this parameter accepts large value it signifies that the quality of the work of the considered channel is low. From analysis of Fig. 9.14 one can notice that the "grounding" procedure plays a key role. For the same channel this procedure decreases the range and increases the quality of the work at the given operating conditions.

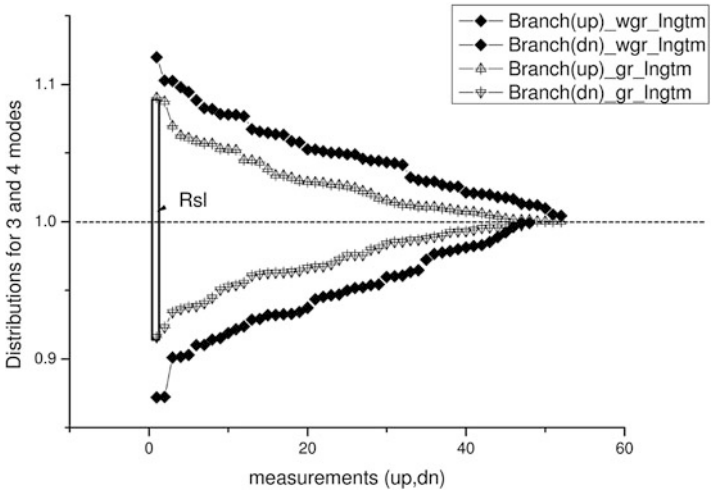


Fig. 9.13 The distribution of "up" and "dn" branches for the last two modes (3 and 4). The most important parameter that should be taken into account is related to the range of the distribution that is defined as the difference between maximal and minimal values. It is shown by *arrow*

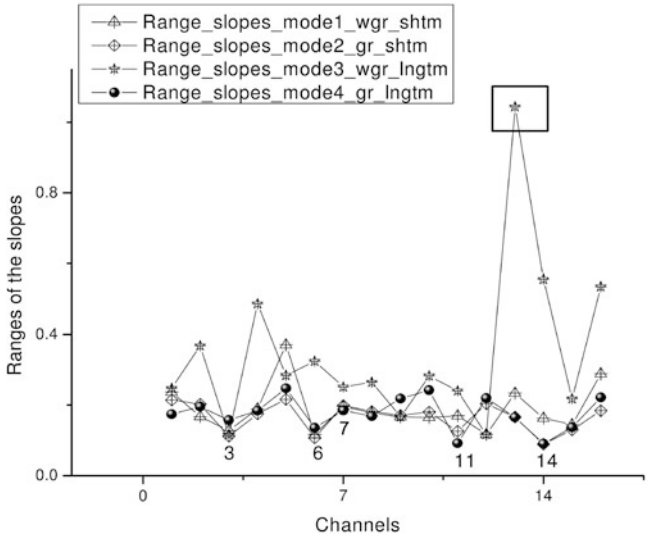


Fig. 9.14 This figure represents the basic result of this research—the distributions of the ranges (calculated, in turn, from the distribution of the slopes) for each channel. One can select the most qualitative channels—3, 6, 7, 11, 14. One can notice also that the grounding modes (2, 4) decrease the range and thereby increase the "quality" of the channel. Long-time "outlier" without grounding (mode 3) is the "worst" mode for ADC work

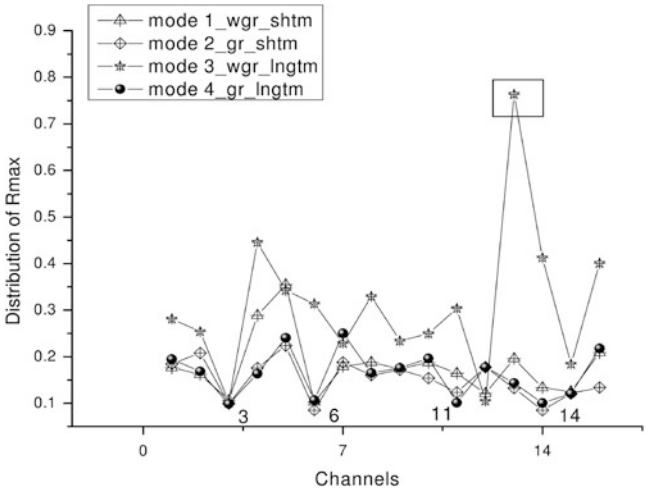


Fig. 9.15 This figure demonstrates the distribution of the ranges Δ_{Rmx} corresponding to the maximal values. These values are defined by expression (9.7). These plots, in fact, repeat the general tendencies that were marked earlier in Fig. 9.14. The grounding procedure increases the quality of a mode irrespective to the time of duration. Long-time exposition has an opposite tendency. The most unstable channel 13 is detected by both distributions

S4. Then each branch ("up" and "dn") can be fitted to the power-law function (9.5) that gives in addition 5 (for each branch) $\times$ 2 = 10 fitting parameters in total. But in this paper we skip this step in order not to overload the content by the large numbers of figures and tables.

S5. If necessary one can repeat these steps for another distribution related to distribution of ranges. This distribution is defined by expression (9.7). In this paper the results of the detailed analysis of the second distribution are demonstrated also. See Fig. 9.15. It is interesting to note that comparison of the Figs. 9.14 and 9.15 leads to the same conclusions: (a) the grounding procedure is important and has a tendency to decrease the value of the range in both cases; (b) the long-time exposition has an opposite tendency and promotes to increasing the value of the corresponding range; (c) the grounding allows to decrease the negative tendency of the long-time exposition. For both case the same "good" channels are detected (3, 6, 11, 14).

9.4 Results and Discussions

The suggested algorithm allows diagnosing a possible disalignment (malfunction) of the technical state of the system considered relatively other states of the same system or relatively another system working in the same working mode.

The possibilities of the algorithm were demonstrated on the working modes of the multichannel ADC. The selection of ADC as an example was stipulated by the following reason. The ADC is a typical device for the most systems of measurements, inspection and control and serves, in the same time, as a source of preliminary information that is provides an accuracy, reliability and trustworthiness of the whole measurement system under study.

In the high-accuracy ADCs with high digit capacity the influence of numerous non-stationary random factors (external and internal ones) leads to systematic dissipation ("diffusion") of errors touching presumably the lower order bits. In multichannel ADCs these factors act with different intensities ("weights") on input channels that disturb the information equivalence of measurement channels. In the result of these random disturbances the data obtained from different ADCs have different credibility and errors.

So, it becomes important to realize the self-inspection regime of the ADC mode when its "metrological" state is under the constant inspection during the monitoring period. In the frame of the suggested algorithm the realization of semi-inspection regime becomes possible in the "rest" interval of times exceeding the intervals of measurements. For example, if the step of quantization equals 10^{-3} s then at operation speed of the ADC equaled 10^{-5} s one can obtain 100 counts that are quite sufficient for self-inspection purposes. These diagnostic results can be used for self-calibration or self-correction of the ADC data obtained. In the simplest case the self-correction procedure includes the elimination of parts of data obtained from "marginal" channels which are out from admissible interval or switching over on part of "normal" channels that are still in admissible interval for the given mode of the ADC analyzed.

The high sensitivity of the proposed method allows implementing corresponding self-inspecting procedures for ADCs. This conclusion is confirmed, for example, by the fact that it provides possibilities to distinguish fluctuation of short and long time operation modes (see, for example, Figs. 9.11, 9.12, and 9.13) while from the point of view of traditional approach based on power estimation they are quite similar and difficult for distinguishing.

In conclusion we would like to stress two basic points. With development of high-intellectual measurement systems, including in itself diagnostics and control systems it becomes important to develop new methods and algorithms that allow to organize the self-calibration, self-diagnosis and self-testing procedures during the whole monitoring mode of the device (ADC in our case) considered. In this paper we suggest new method (algorithm) that allows organizing the self-inspection procedure of the technical system under consideration. It allows detecting possible deviations, misalignment (malfunction) that can appear under the influence of external/internal uncontrollable factors. The algorithm is based on fundamental property of symmetry of the SRA that bisected relatively its mean value [33]. The "ideal" SRA of the slopes, for example, bisected with respect to the unit value with minimal range (defined by relationship (9.6)) can correspond to an ideal "device". The deviations from the ideal symmetry and simple quantitative value expressed in the form of the corresponding range can serve as a measure of deviation of

the objects that are diagnosed or under testing procedure. In addition one can add the fitting parameters of the two-power law function (9.6) that fits "up" and "dn" branches of the desired SRA distribution. From our point of view the following advantages of the suggested algorithm should be emphasized:

– It does not need in any a priori information;
– It has invariant properties to the type of statistics of the signal and noises (internal or external) analyzed;
– It has a simple structure based on small computing resources that allows to realize it in the form of simple embedded device or in the form of small program working in background mode;
– It has high operating speed and sensitivity to possible deviations from the required/standard metrological values and characteristics of the "normal" technical state.

Acknowledgements This paper is stimulated by the R&D project realized in the frame of the JNU-KNRTU(KAI) Joint Laboratory of "Fractal Radio-electronics and Fractal Signal Processing".

References

1. Rabiner, L.R., Gold, B.: Theory and Application of Digital Signal Processing. Prentice-Hall, Englewood Cliffs (1975)
2. Singleton, Jr., Royce, A., Straits, B.C., Straits, M.M.: Approaches to Social Research. Oxford University Press, Oxford (1993)
3. Mendel, J.M.: Lessons in Estimation Theory for Signal Processing, Communications, and Control. Pearson Education, Upper Saddle River (1995)
4. Hagan, M.T., Demuth, H.B., Beale M.H.: Neural Network Design. PWS Publishing, Boston (1996)
5. Ifeachor, E.C., Jervis, B.W.: Digital Signal Processing: A Practical Approach. Pearson Education, Harlow (2002)
6. Montgomery, D.C., Jennings, C.L., Kulahci, M.: Introduction to Time Series Analysis and Forecasting. Wiley, Hoboken (2011)
7. Bendat, J.S., Piersol, A.G.: Random Data: Analysis and Measurement Procedures. Wiley, New York (2011)
8. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data analysis. CRC Press, Boca Raton (2013)
9. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control. Wiley, Hoboken (2008)
10. Chatfield, C. (ed.): The Analysis of Time Series: An Introduction. CRC Press, New York (2003)
11. Sheng, H., Chen, Y., Qui, T.: Fractal Processes and Fractional-Order Signal Processing. Techniques and Applications. Springer, Berlin (2012)
12. Baleanu, D., Guvench, Z.B., Tenreiro Machado, J.A.: New trends in Nanotechnology and Fractional Calculus Applications. Springer, Heidelberg (2010)
13. Baleanu, D., Tenreiro Machado, J.A., Luo, A.C.J.: Fractional Dynamics and Control. Springer, New York (2012)
14. Luo, A.C.J., Tenreiro Machado, J.A., Baleanu, D.: Dynamical Systems and Methods. Springer, New York (2012)

15. Ciurea, M.L., Lazanu, S., Stavaracher, I., Lepadatu, A.-M., Iancu, V., Mitroi, M.R., Nigmatullin, R.R., Baleanu, C.M.: Stress-induced traps in multilayered structures. J. Appl. Phys. **169**, 013717 (2011)
16. Nigmatullin, R.R., Baleanu, D., Dinch, E., Ustundag, Z., Solak, A.O., Kargin, R.V.: Analysis of a nanofilm of the mercaptophenyl diazonium modified gold electrode within new statistical parameters. J. Comput. Theor. Nanosci. **7**(3), 562–570 (2010)
17. Nigmatullin, R.R.: New noninvasive methods for 'reading' of random sequences and their applications in nanothechnology. In: Baleanu, D., Guvench, Z.B., Tenreiro Machado, J.A. (eds.) New Trends in Nanotechnology and Fractional Calculus Applications, pp. 43–56. Springer, Heidelberg (2010)
18. Nigmatullin, R.R.: Universal distribution function for the strongly-correlated fluctuations: General way for description of different random sequences. Commun. Nonlinear Sci. Numer. Simul. **15**, 637–647 (2010)
19. Nigmatullin, R.R.: The statistics of the fractional moments: Is there any chance to "read quantitatively" any randomness? Signal Process. **86**, 2529–2547 (2006)
20. Nigmatullin, R.R., Ionescu, C., Baleanu, D.: NIMRAD: Novel technique for respiratory data treatment. Signal Image Video Process. (2012). doi:10.1007/s11760-012-0386-1
21. Nigmatullin, R.R., Striccoli, D., Zhang, W.: General theory for reproducible data processing: Apparatus function and reduction to an "ideal" experiment. In: Books of Abstracts, 2014 International Conference on Mathematics Models and Methods in Applied Sciences (MMMAS 2014), pp. 303–305. S-Petersburg State Politechnical University, Saint-Petersburg (2014)
22. El-Bakry, H.M., Mastorakis, N.: A new fast forecasting technique using high speed neural networks. In: Proceedings of 8th WSEAS International Conference on Signal, Speech and Image Processing (SSIP '08), pp. 116–138. Santander, Cantabria (2008)
23. Kan, B., Yazici, B.: Comparison of the results of factorial experiments, fractional factorial experiments, regression trees and mars for fuel consumption data. WSEAS Trans. Math. **9**(2), 110–119 (2010)
24. Nigmatullin, R.R., Tenreiro Machado, J., Menezes, R.: Self-similarity principle: The reduced description of randomness. Cent. Eur. J. Phys. **11**(6), 724–739 (2013)
25. Nigmatullin, R.R., Baleanu, D.: The derivation of the generalized functional equations describing self-similar processes. Fract. Calc. Appl. Anal. **15**(4), 718–740 (2012)
26. Nigmatullin, R.R., Baleanu, D.: New relationships connecting a class of fractal objects and fractional integrals in space. Fract. Calc. Appl. Anal. **16**(4), 911–936 (2013)
27. Nigmatullin, R.R., Khamzin, A.A., Machado, J.T.: Detection of quasi-periodic processes in complex systems: How do we quantitatively describe their properties? Phys. Scr. (2014). doi:10.1088/0031-8949/89/01/015201
28. Nigmatullin, R.R., Osokin, S.I., Baleanu, D., Al-Amri, S., Azam, A., Memic, A.: The first observation of memory effects in the infraRed (FT-IR) measurements: Do successive measurements remember each other? PLoS One (2014). doi:10.1371/journal.pone.0094305
29. Nigmatullin, R., Rakhmatullin, R.: Detection of quasi-periodic processes in repeated measurements: New approach for the fitting and clusterization of different data. Commun. Nonlinear Sci. Numer. Simul. **19**(12), 4080–4093 (2014)
30. Nigmatullin, R.R.: Eigen-coordinates: New method of analytical functions identification in experimental measurements. Appl. Magn. Reson. **14**, 601–633 (1998)
31. Nigmatullin, R.R.: Recognition of nonextensive statistical distributions by the eigencoordinates method. Phys. A **285**, 547–565 (2000)
32. Nigmatullin, R.R., Bras, A.R., Coreia, N.T.: Evidences of the fractional kinetics in temperature region: Evolution of extreme points in ibuprofen. Commun. Nonlinear Sci. Numer. Simul. **15**, 2942–2966 (2010)
33. Nigmatullin, R.R.: Strongly correlated variables and existence of a universal distribution function for relative fluctuations. Phys. Wave Phenom. **16**(2), 119–145 (2008)
34. NI 6366/6368 Specifications. http://www.ni.com/pdf/manuals/370084d.pdf (2013). Accessed 10 Sep 2014

Chapter 10
Maximum Principle for Delayed Stochastic Switching System with Constraints

Charkaz Aghayeva

Abstract This paper is devoted to the stochastic optimal control problem of switching systems with constraints. Dynamic of the system is described by the collection of delayed stochastic differential equations which initial conditions depend on its previous state. The restriction on the system is defined by the functional constraints on the end of each interval. Maximum principle for stochastic control problems of delayed switching system is established. Afterwards, using Ekeland's Variational Principle the necessary condition of optimality for optimal control problem with constraints is obtained.

Keywords Stochastic control system • Differential equation with delay • Switching system • Switching law • Optimal control problem • Maximum principle

10.1 Introduction

Noise and time delay are associated with many real phenomena, and often they are sources of complex behaviors. Stochastic differential equations have the benefit in description of the natural systems, which in one or another degree are subjected to the influence of the random noises. Systems with stochastic uncertainties have provided a lot of interest for problems of nuclear fission, communication systems, self-oscillating systems and etc., where the influences of random disturbances cannot be ignored [1, 2].

Many real stochastic process cannot be considered as Markov process, because their future behavior obviously depends not only on their present, but also on their previous states. The differential equations with time delay can be used to model processes with a memory, when the behaviour of the system depends on values of

C. Aghayeva (✉)
Industrial Engineering Department, Anadolu University, Eskisehir, Turkey

Institute of Control Systems of ANAS, Baku, Azerbaijan
e-mail: c_aghayeva@anadolu.edu.tr; cherkez.agayeva@gmail.com

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_10

the process of the past [3, 4]. Optimization problems for delayed stochastic control systems have attracted a lot of interest [5–12].

Switching systems consist of several subsystems and a switching law indicating the active subsystem at each time instantly. For general theory of stochastic switching systems it is referred to [13]. Theoretical results and applications of stochastic switching systems were developed in [14–18]. Deterministic and stochastic optimal control problems of switching systems, described by differential equations with delay, are actual at present [19–21].

This article is concerned with optimal control problem of stochastic delayed switching system with constraints. The rest of paper is organized as follows. The next section formulates the main problem, presents some concepts and assumptions.

The necessary condition of optimality for stochastic switching systems with delay in case when endpoint constraints are imposed is obtained in Sect. 10.3. In Sect. 10.4, using Ekeland's Variational Principle investigated problem is convert into the sequence of unconstrained systems. A maximum principle and transversality conditions are established for transformed problem. Finally, taking the limit the necessary condition of optimality in the case with endpoint constraints is achieved. The paper is concluded in Sect. 10.5 with some possible developments and enlargements.

10.2 Problem Statement and Assumptions

First, we introduce notations are used throughout this paper. Let N be some positive constant, R^n denotes the n-dimensional real vector space, $|.|$ denotes the Euclidean norm and $\langle \cdot, \cdot \rangle$ denotes scalar product in R^n. E represents expectation and $\overline{1,r}$ denotes the set of integer numbers $1,\ldots,r$. Assume that $w_t^1, w_t^2, \ldots, w_t^r$ are independent Wiener processes, which generate filtrations $F_t^l = \overline{\sigma}(w_q^l, t_{l-1} \leq t \leq t_l)$, $l = 1,\ldots,r$. Let $\left(\Omega, F^l, P\right)$, $l = 1,\ldots,r$ be a probability spaces with corresponding filtrations $\left\{ F_t^l, t \in [t_{l-1}, t_l] \right\}$. $L^2_{F^l}(a,b;R^n)$ denotes the space of all predictable processes $x_t(\omega)$ such that: $E\int_a^b |x_t(\omega)|^2 dt < +\infty$. $R^{m\times n}$ is the space of all linear transformations from R^m to R^n. Let, $O_l \subset R^{n_l}$, $Q_l \subset R^{m_l}$ be open sets, $T = [0,T]$ be a finite interval and $0 = t_0 < t_1 < \cdots < t_r = T$. Following notation is used unless specified otherwise: $\mathbf{t} = (t_0, t_1, \ldots, t_r)$, $\mathbf{u} = \left(u^1, u^2, \ldots, u^r\right)$, $\mathbf{x} = \left(x^1, x^2, \ldots, x^r\right)$.

Consider the following stochastic control system with delay:

$$dx_t^l = g^l\left(x_t^l, x_{t-h}^l, u_t^l, t\right)dt + f^l\left(x_t, x_{t-h}, t\right)dw_t^l \quad t \in (t_{l-1}, t_l] \tag{10.1}$$

$$x_t^{l+1} = \mathrm{K}^{l+1}(t), \quad t \in [t_l - h, t_l), \quad l = 0, 1, \ldots, r-1, \tag{10.2}$$

$$x_{t_l}^{l+1} = \Phi^{l+1}\left(x_{t_l}^{l}, t_l\right), \quad l = 0, \ldots, r-1, \quad x_{t_0}^{1} = x_0, \tag{10.3}$$

$$u_t^l \in U_\partial^l \equiv \left\{u^l(\cdot,\cdot) \in L_F^2\left(t_{l-1}, t_l; R^m\right) \middle| u^l(t,\cdot) \in U^l \subset R^m\right\} \tag{10.4}$$

where $U^l,\ l = 1, \ldots, r$ are non-empty bounded sets. Let $\Lambda_l,\ l = 1, \ldots, r$ be the set of piecewise continuous functions $\mathrm{K}^l(\cdot),\ l = 1, \ldots, r : [t_{l-1} - h, t_{l-1}) \to N_l \subset O_l$ and $h \geq 0$.

The problem is concluded to find the control $u^1, u^2, \ldots, u^r$ and the switching law $t_1, t_2, \ldots, t_r$ which minimize the cost functional:

$$J(u) = E\left[\varphi^r\left(x_{t_r}^r\right) + \sum_{l=1}^{r} \int_{t_{l-1}}^{t_l} p^l\left(x_t^l, u_t^l, t\right) dt\right] \tag{10.5}$$

which is determined on the decisions of the system (10.1)–(10.3), which are generated by all admissible controls $U = U^1 \times U^2 \times \cdots \times U^r$ at conditions:

$$Eq^l\left(x_{t_l}^l\right) \in G^l, \quad l = \overline{1, r} \tag{10.6}$$

$G^1, \ldots, G^r$ are a closed convex sets in $R^{k_1}, R^{k_2}, \ldots, R^{k_r}$ respectively.

Consider the sets $A_i = \mathrm{T}^{i+1} \times \prod_{j=1}^{i} O_j \times \prod_{j=1}^{i} \Lambda_j \times \prod_{j=1}^{i} U^j$ with the elements $\pi^i = \left(t_0, \ldots, t_i, x_{t_1}^1, \ldots, x_{t_i}^i, K_1, \ldots, K_i, u^1, \ldots, u^i\right)$.

Definition 1. The set of functions $\left\{x_t^l = x^l\left(t, \pi^l\right)\right\}, t \in [t_{l-1} - h, t_l], l = 1, \ldots r$ is said to be a solution of the equation with variable structure which corresponds to an element $\pi^r \in \mathrm{A}_r$, if the function $x_t^l \in O_l$ on the interval $[t_l - h, t_l]$ satisfies the conditions (10.2) and (10.3), while on the interval $[t_{l-1}, t_l]$ it is absolutely continuous with probability 1 and satisfies the Eq. (10.1) almost everywhere.

Definition 2. The element $\pi^r \in A_r$ is said to be admissible if the pairs $\left(x_t^l, u_t^l\right),\ t \in [t_{l-1} - h, t_l],\ l = 1, \ldots, r$ are the solutions of system (10.1)–(10.4) and satisfied the conditions (10.6). A_r^0 denotes the set of admissible elements.

Definition 3. The element $\tilde{\pi}^r \in A_r^0$, is said to be an optimal solution of problem (10.1)–(10.6) if there exist admissible controls $\tilde{u}_t^l,\ t \in [t_{l-1}, t_l], l = 1, \ldots r$ and corresponding solutions $\left\{\tilde{x}_t^l, t \in [t_{l-1} - h, t_l], l = 1, \ldots, r\right\}$ of system (10.1)–(10.4) with constraints (10.6), and pairs $\left(\tilde{x}_t^l, \tilde{u}_t^l\right),\ l = 1, \ldots, r$ minimize the functional (10.5).

Assume that the following requirements are satisfied:

I. Functions $g^l, f^l, p^l, l = 1, \ldots, r$ and their derivatives are continuous in (x, y, u, t): $g^l(x, y, u, t) : O_l \times O_l \times Q_l \times T \to R^{n_l}$, $f^l(x, y, t) : O_l \times O_l \times T \to R^{n_l \times n_l}$, $p^l(x, u, t) : O_l \times Q_l \times T \to R$.
II. When (t, u) are fixed, functions $g^l, f^l, p^l,\ l = \overline{1, r}$ hold the conditions:

$$(1+|x|+|y|)^{-1}\left(\left|g^l(x,y,u,t)\right|+\left|g_x^l(x,y,u,t)\right|+\left|g_y^l(x,y,u,t)\right|\right.$$
$$+\left|f^l(x,y,t)\right|+\left|f_x^l(x,y,t)\right|+\left|f_y^l(x,y,t)\right|$$
$$\left.+\left|p^l(x,u,t)\right|+\left|p_x^l(x,u,t)\right|\right)\le N.$$

III. Function $\varphi^r(x): R^{n_r}\to R$ is continuously differentiable:

$$|\varphi^r(x)|+|\varphi_x^r(x)|\le N\,(1+|x|)$$

IV. Functions $\Phi^l(x,t): O_{l-1}\times \mathrm{T}\to O_l,\ l=1,\dots,r-1$ are continuously differentiable in respect to (x,t): $\left|\Phi^l(x,t)\right|+\left|\Phi_x^l(x,t)\right|\le N\,(1+|x|)$.

V. Functions $q^l(x): R^{k_l}\to R,\ l=\overline{1,r}$ are continuously differentiable in respect to (x,t): $\left|q^l(x)\right|+\left|q_x^l(x)\right|\le N\,(1+|x|)$.

10.3 Maximum Principle for Delayed Systems

Applying the similar technique as in [19] following result that is a necessary condition of optimality for problem (10.1)–(10.5) is obtained.

Theorem 1. Suppose that $\pi^r=\left(t_0,\dots,t_r,x_t^1,\dots,x_t^r,K_1,\dots,K_r,u^1,\dots,u^r\right)$ is an optimal solution of problem (10.1)–(10.5) and random processes $\left(\psi_t^l,\beta_t^l\right)\in L^2_{F^l}\left(t_{l-1},t_l;R^{n_l}\right)\times L^2_{F^l}\left(t_{l-1},t_l;R^{n_l x n_l}\right)$ are the solutions of the following adjoint equations:

$$\begin{cases} d\psi_t^l=-\left[H_x^l\left(\psi_t^l,x_t^l,y_t^l,u_t^l,t\right)+H_y^l\left(\psi_{t+h}^l,x_{t+h}^l x_t^l,u_t^l,t\right)\right]dt+\beta_t^l dw_t^l,\\ \qquad t_{l-1}\le t<t_l-h,\\ d\psi_t^l=-H_x^l\left(\psi_t^l,x_t^l,y_t^l,u_t^l,t\right)dt+\beta_t^l dw_t^l,\quad t_{l-1}-h_l\le t<t_l,\\ \psi_{t_l}^l=\psi_{t_l}^{l+1}\Phi_x^l\left(x_{t_l}^l,t_l\right),\quad l=1,\dots,r-1,\\ \psi_{t_r}^r=-\varphi^r\left(x_{t_r}^r\right). \end{cases} \tag{10.7}$$

Then,

(a) a.c. for $\forall\tilde{u}^l\in U^l,\ l=1,\dots r$, a.e. in $[t_{l-1},t_l]$ the maximum principle hold:

$$H^l\left(\psi_\theta^l,x_\theta^l,y_\theta^l,\tilde{u}^l,\theta\right)-H^l\left(\psi_\theta^l,x_\theta^l,y_\theta^l,u_\theta^l,\theta\right)\le 0 \tag{10.8}$$

(b) Following transversality conditions hold for each $l = \overline{0,r}$:

$$\begin{aligned} & b_l \psi^{l+1*}_{t_l+h} \Phi^{l+1}_t \left(x^l_{t_l}, t_l\right) + b_l \psi^{l+1*}_{t_l} g^{l+1} \left(x^{l+1}_{t_l}, y^{l+1}_{t_l}, u^{l+1}_{t_l}, t_l\right) \\ & \quad - b_l \psi^{l+1*}_{t_l+h} g^{l+1}(x^{l+1}_{t_l+h}, K^{l+1}(t_l), u^{l+1}_{t_l+h}, t_l + h) \\ & \quad - a_l \psi^{l*}_{t_l} g^l \left(x^l_{t_l}, y^l_{t_l}, u^l_{t_l}, t_l\right) = 0 \end{aligned} \tag{10.9}$$

Here, $H^l(\psi_t, x_t, y_t, u_t, t) = \psi^*_t g^l(x_t, y_t, u_t, t) + \beta^*_t f^l(x_t, y_t, t) - p^l(x_t, u_t, t)$, $t \in [t_{l-1}, t_l]$

$y^l_t = x^l_{t-t}, a_0 = 0, a_1 = \cdots = a_r = 1$ and $b_0 = \cdots = b_{r-1} = 1,\ b_r = 0.$

Proof. Let $\overline{u}^l_t = u^l_t + \Delta\overline{u}^l_t \ \ l = \overline{1,r}$ be some admissible controls and $\overline{x}^l_t = x^l_t + \Delta\overline{x}^l_t \ \ l = \overline{1,r}$ be corresponding trajectories of system (10.1)–(10.3) and $0 = t_0 < t_1 < \cdots < t_r \le T$ be switching sequence. Then for some sequence of $0 = \overline{t}_0 < \overline{t}_1 < \cdots < \overline{t}_r \le T$

$$\begin{cases} d\Delta\overline{x}^l_t = \left\{\Delta_{\overline{u}^l} g^l \left(x^l_t, y^l_t, u^l_t, t\right) + g^l_x \left(x^l_t, y^l_t, u^l_t, t\right) \Delta\overline{x}^l_t + g^l_y \left(x^l_t, y^l_t, u^l_t, t\right) \Delta\overline{y}^l_t\right\} dt \\ \qquad\qquad + \left\{f^l_x \left(x^l_t, y^l_t, t\right) \Delta\overline{x}^l_t + f^l_y \left(x^l_t, y^l_t, t\right) \Delta\overline{y}^l_t\right\} dw^l_t + \eta^1_t, \quad t \in (t_{l-1}, t_l], \\ \Delta\overline{x}^l_t = 0,\ t \in [t_{l-1} - h, t_l),\ l = \overline{1,r}, \Delta\overline{x}^1_{t_0} = 0, \\ \Delta\overline{x}^l_{t_{l-1}} = \Phi^{l-1}\left(\overline{x}^{l-1}_{t_{l-1}}, \overline{t}_{l-1}\right) - \Phi^{l-1}\left(x^{l-1}_{t_{l-1}}, t_{l-1}\right), l = \overline{2,r}, \end{cases} \tag{10.10}$$

where $\Delta_{\overline{u}^l} g(x_t, x_{t-h}, u_t, t) = g\left(x_t, x_{t-h}, , \overline{u}^l_t, t\right) - g\left(x_t, x_{t-h}, , u^l_t, t\right)$,

$$\begin{aligned} \eta^1_t = & \int_0^1 \left[g^{l*}_x (x^l_t + \mu\Delta\overline{x}^l_t, \overline{y}^l_t, \overline{u}^l_t, t) - g^{l*}_x \left(x^l_t, y^l_t u^l_t, t\right)\right] \Delta\overline{x}^l_t d\mu dt \\ & + \int_0^1 \left[g^{l*}_y (x^l_t, y^l_t + \mu\Delta\overline{y}^l_t, \overline{u}^l_t, t) - g^{l*}_y \left(x^l_t, y^l_t u^l_t, t\right)\right] \Delta\overline{y}^l_t d\mu dt \\ & + \int_0^1 \left[f^{l*}_x (x^l_t + \mu\Delta\overline{x}^l_t, \overline{y}^l_t, t) - f^{l*}_x \left(x^l_t, y^l_t, t\right)\right] \Delta\overline{x}^l_t d\mu dw^l_t \\ & + \int_0^1 \left[f^{l*}_y \left(x^l_t, y^l_t + \mu\Delta\overline{y}^l_t, t\right) - f^{l*}_y (x^l_t, y^l_t, t)\right] \Delta\overline{x}^l_t d\mu dw^l_t. \end{aligned}$$

According to Ito's formula [2] the following has been yielded:

$$
\begin{aligned}
d\left(\psi_t^{l*}\cdot\Delta\overline{x}_t^l\Delta\overline{t}_l\right) &= \psi_t^{l*}\cdot\Delta\overline{x}_t^l dt_l + d\psi_t^{l*}\Delta\overline{x}_t^l\Delta\overline{t}_l + \psi_t^{l*}d\Delta\overline{x}_t^l\Delta\overline{t}_l \\
&+ \left\{\beta_t^{l*}\left[f_x^l\left(x_t^l,y_t^l,t\right)\Delta\overline{x}_t^l + f_y^l\left(x_t^l,y_t^l,t\right)\Delta\overline{y}_t^l\right]\Delta\overline{t}_l\right. \\
&+ \beta_t^{l*}\int_0^1\left[f_x^l\left(x_t^l+\mu\Delta\overline{x}_t^l,\overline{y}_t^l,t\right) - f_x^l\left(x_t^l,y_t^l,t\right)\right]\Delta\overline{x}_t^l\Delta\overline{t}_l d\mu \\
&+ \left.\beta_t^{l*}\int_0^1\left[f_y^l\left(x_t^l,y_t^l+\mu\Delta\overline{y}_t^l,t\right) - f_y^l\left(x_t^l,y_t^l,t\right)\Delta\overline{y}_t^l\Delta\overline{t}_l d\mu\right\}dt\right.
\end{aligned}
$$

The stochastic processes ψ_t^l, at the points $t_1, t_2, \ldots, t_r$ can be defined as follows:

$$
\psi_{t_l}^l = \psi_{t_l}^{l+1}\Phi_x^l\left(x_t^l,t_l\right), \quad l=\overline{1,r-1} \quad \text{and} \quad \psi_{t_r}^r = -\varphi_x^r\left(x_{t_r}^r\right) \tag{10.11}
$$

Taking into consideration (Eqs. 10.9–10.11) the expression of increment of a cost functional (Eq. 10.5) along the admissible control looks like:

$$
\begin{aligned}
\Delta J(u) = -\sum_{l=1}^{r} E\int_{t_{-1}}^{t_l}\Big[&\psi_t^{l*}\Delta_{\overline{u}^l}g^l\left(x_t^l,y_t^l,u_t^l,t\right) + \psi_t^{l*}g_x^l\left(x_t^l,y_t^l,u_t^l,t\right)\Delta\overline{x}_t^l \\
&+ \psi_t^{l*}g_y^l\left(x_t^l,y_t^l,u_t^l,t\right)\Delta\overline{y}_t^l + \beta_t^{l*}f_x^l\left(x_t^l,y_t^l,t\right)\Delta\overline{x}_t^l \\
&+ \beta_t^{l*}f_y^l\left(x_t^l,y_t^l,t\right)\Delta\overline{y}_t^l - \Delta_{\overline{u}^l}p^l\left(x_t^l,u_t^l,t\right) - p_x^l\left(x_t^l,u_t^l,t\right)\Delta\overline{x}_t^l\Big]\Delta\overline{t}_l dt \\
&+ \sum_{l=1}^{r-1}\psi_{t_l}^{l+1}\Phi_t\left(x_{t_l}^l,t_l\right)\Delta\overline{t}_l + \sum_{l=1}^{r}\eta_{t_{l-1}}^{t_l}
\end{aligned} \tag{10.12}
$$

$$\eta_{t_{l-1}}^{t_l} = -E\int_0^1 (1-\mu)\left[\varphi_x^{r*}\left(x_{t_r}^r + \mu\Delta\overline{x}_{t_r}^r\right) - \varphi_x^{r*}\left(x_{t_r}^r\right)\right]\Delta\overline{x}_{t_r}^r d\mu$$

$$- E\int_{t_{l-1}}^{t_l}\left\{\int_0^1 (1-\mu)\left[p_x^{l*}\left(x_t^l + \mu\Delta\overline{x}_t^l, u_t^l, t\right) - p_x^{l*}\left(x_t^l, u_t^l, t\right)\right]\Delta\overline{x}_t^l d\mu dt\right\}$$

$$+ E\int_{t_{l-1}}^{t_l}\int_0^1 (1-\mu)\,\psi_t^{l*}\left[g_x^l\left(\overline{x}_t^l, \overline{y}_t^l, u_t^l, t\right) - g_x^l\left(x_t^l, \overline{y}_t^l, u_t^l, t\right)\right]\Delta\overline{x}_t^l \Delta t_l d\mu dt$$

$$+ E\int_{l-1}^{t_l}\int_0^1 (1-\mu)\,\psi_t^{l*}\left[g_y^l\left(x_t^l, \overline{y}_t^l, u_t^l, t\right) - g_y^l\left(x_t^l, y_t^l, u_t^l, t\right)\right]\Delta\overline{y}_t^l \Delta t_l d\mu dt$$

$$+ E\int_{t_{l-1}}^{t_l}\int_0^1 (1-\mu)\,\beta_t^{l*}\left[f_x^l\left(x_t^l + \mu\Delta\overline{x}_t^l, \overline{y}_t^l, t\right) - f_x^l\left(x_t^l, y_t^l, t\right)\right]\Delta\overline{x}_t^l \Delta t_l d\mu dt$$

$$+ E\int_{t_{l-1}}^{t_l}\int_0^1 (1-\mu)\,\beta_t^{l*}\left[f_y^l\left(x_t^l, y_t^l + \mu\Delta\overline{y}_t^l, t\right) - f_y^l\left(x_t^l, y_t^l, t\right)\right]\Delta\overline{y}_t^l \Delta t_l d\mu dt$$

$$- E\int_0^1 (1-\mu)\,\psi_{t_l}^{l+1*}\left[\Phi_x^l\left(x_{t_l}^l + \mu\Delta\overline{x}_{t_l}^l, t_l\right) - \Phi_x^l\left(x_{t_l}^l, t_l\right)\right]\Delta\overline{x}_{t_l}^l \Delta t_l d\mu \tag{10.13}$$

According to a necessary condition for an optimal solution, we obtain that, the coefficients of the independent increments $\Delta\overline{x}_t^l$, $\Delta\overline{y}_t^l$ $\Delta\overline{t}_l$ equal zero. By assumption IV and using the expression (10.10) from the identity (10.12), we obtain that Eq. (10.9) is true.

According to Eqs. (10.9) and (10.11), through the simple transformations, expression (10.12) may be written as:

$$\Delta J(u) = -\sum_{l=1}^{r} E\int_{t_{l-1}}^{t_l}\left[\Delta_{\overline{u}^l} H^l\left(\psi_t^l, x_t^l, u_t^l, t\right) + \Delta_{\overline{u}^l} H_{x^l}^l\left(\psi_t^l, x_t^l, y_t^l, u_t^l, t\right)\Delta\overline{x}_t^l\right.$$

$$\left. + \Delta_{\overline{u}^l} H_y^l\left(\psi_t^l, x_t^l, y_t^l, u_t^l, t\right)\Delta\overline{y}_t^l\right]\Delta\overline{t}_l dt + \sum_{l=1}^{r}\eta_{t_{l-1}}^{t_l} \tag{10.14}$$

Now, consider the following spike variation:

$$\Delta u_t^l = \Delta u_{t,\varepsilon^l}^{\theta^l} = \begin{cases} 0, & t \notin [\theta_l, \theta_l + \varepsilon_l), \ \varepsilon_l > 0, \theta_l \in [t_{l-1}, t_l) \\ \tilde{u}^l - u_t^l, & t \in [\theta_l, \theta_l + \varepsilon_l), \ \tilde{u}^l \in L^2\left(\Omega, F^{\theta_l}, P; R^m\right) \end{cases}$$

where ε_l are sufficiently small numbers. Then the expression (10.14) takes the form of:

$$\Delta_\theta J(u) = -\sum_{l=1}^{r} E \int_{\theta_l}^{\theta_l+\varepsilon_l} \left[\Delta_{\tilde{u}^l} H^l\left(\psi_t^l, x_t^l, u_t^l, t\right) + \Delta_{\tilde{u}^l} H_x^l\left(\psi_t^l, x_t^l, y_t^l, u_t^l, t\right) \Delta \overline{x}_t^l \right.$$

$$\left. + \Delta_{\tilde{u}^l}^l H_y^l\left(\psi_t^l, x_t^l, y_t^l, u_t^l, t\right) \Delta \overline{y}_t^l\right] \Delta \bar{t}_l dt + \sum_{l=1}^{r} \eta_{\theta_l}^{\theta_l+\varepsilon_l} \tag{10.15}$$

Following lemma will be used in estimating for increment (Eq. 10.15).

Lemma 1 [19]. Assume that the conditions I–II are fulfilled, $x_{t,\varepsilon_l}^{\theta_l}$ are trajectories of system (10.1)–(10.3), corresponding to controls $u_{t,\varepsilon_l}^{\theta_l} = u_t^l + \Delta u_{t,\varepsilon_l}^{\theta_l}$ respectively. If $\varepsilon_l \to 0$, $l = \overline{1,r}$. Then for all $t \in [t_{l-1}, t_l)$ there occurs

$$E\left|\frac{x_{t,\varepsilon_l}^{\theta_l} - x_t^l}{\varepsilon_l}\right|^2 \leq N.$$

Proof. Let us denote the following: $\tilde{x}_{t,\varepsilon_l}^l = \left(x_{t,\varepsilon_l}^{\theta_l} - x_t^l\right)\varepsilon_l^{-1}$, $\tilde{y}_{t,\varepsilon_l}^l = \tilde{x}_{t-h,\varepsilon_l}^l = \left(x_{t-h,\varepsilon_l}^{\theta_l} - x_{t-h}^l\right)\varepsilon_l^{-1}$

It is clear that $\forall t \in [t_{l-1}, \theta_l)$ $\tilde{x}_{t,\varepsilon_l}^l = 0$, $l = \overline{1,r}$. Then for $\forall t \in [\theta_l, \theta_l + \varepsilon_l)$

$$d\tilde{x}_{t,\varepsilon_l}^l = \varepsilon_l^{-1}\left[g^l\left(x_{t,\varepsilon_l}^{\theta_l}, y_{t,\varepsilon_l}^{\theta_l}, \tilde{u}^l, t\right) - g^l\left(x_t^l, x_t^l, u_t^l, t\right)\right] dt$$

$$+ \varepsilon_l^{-1}\left[f^l\left(x_{t,\varepsilon_l}^{\theta_l}, y_{t,\varepsilon_l}^{\theta_l}, t\right) - f^l\left(x_t^l, y_t^l, t\right)\right] dw_t^l,$$

$$\tilde{x}_{\theta_l,\varepsilon_l}^l = -\left(g^l\left(x_{\theta_l}^l, y_{\theta_l}^l, \tilde{u}^l, \theta_l\right) - g\left(x_{\theta_l}^l, y_{\theta_l}^l, u_{\theta_l}^l, \theta_l\right)\right),$$

or

$$\begin{aligned}\tilde{x}^l_{\theta_l+\varepsilon_l,\varepsilon_l} &= \varepsilon_l^{-1}\int_{\theta_l}^{\theta_l+\varepsilon_l}\left[g^l\left(x^{\theta_l}_{t,\varepsilon_l}, y^{\theta_l}_{t,\varepsilon_l}, u^l, s\right) - g^l\left(x^l_s, y^l_s, u^l_s, s\right)\right]ds\\ &+ \int_{\theta_l}^{\theta_l+\varepsilon_l}\left[g^l\left(x^l_{\theta_l}, y^l_{\theta_l}, u^l_{\theta_l}, \theta_l\right) - g^l\left(x^l_s, y^l_s, u^l_s, s\right)\right]ds\\ &+ \int_{\theta_l}^{\theta_l+\varepsilon_l}\left[f^l\left(x^{\theta_l}_{t,\varepsilon_l}, y^{\theta_l}_{t,\varepsilon_l}, s\right) - f^l\left(x^l_s, y^l_s, s\right)\right]dw^l_s\\ &+ \int_{\theta_l}^{\theta v+\varepsilon_l}\left[g^l\left(x^l_s, y^l_s, \tilde{u}^l, s\right) - g^l\left(x^l_{\theta_l}, y^l_{\theta_l}, \tilde{u}^l, \theta_l\right)\right]ds.\end{aligned}$$

Therefore, using the Gronwall's inequality due to the conditions I–II the following is achieved:

$$\begin{aligned}E\left|\tilde{x}^l_{\theta_l+\varepsilon_l,\varepsilon_l}\right|^2 \le N\Bigg[&E\sup_{\theta_l\le t\le\theta_l+\varepsilon_l}\left|x^{\theta_l}_{t,\varepsilon_l} - x^l_t\right|^2 + E\sup_{\theta_l\le t\le\theta_l+\varepsilon_l}\left|x^l_t - x^l_{\theta_l}\right|^2\\ &+ E\sup_{\theta_l\le t\le\theta_l+\varepsilon_l}\left|y^{\theta_l}_{t,\varepsilon_l} - y^l_t\right|^2 + E\sup_{\theta_l\le t\le\theta_l+\varepsilon_l}\left|y^l_t - y^l_{\theta_l}\right|^2\\ &+ \sup_{\theta_l\le t\le\theta_l+\varepsilon_l}\varepsilon_l^2 E\left|g^l\left(x^l_t, y^l_t, \tilde{u}^l, t\right) - g^l\left(x^l_{\theta_l}, y^l_{\theta_l}, \tilde{u}^l, \theta_l\right)\right|^2\\ &+ \varepsilon_l^{-1}E\int_{\theta_l}^{\theta_l+\varepsilon_l}\left|f^l\left(x^l_t, y^l_t, t\right) - f^l\left(x^l_{\theta_l}, y^l_{\theta_l}, \theta_l\right)\right|^2 dt\\ &+ \varepsilon_l^{-1}E\int_{\theta_l}^{\theta_l+\varepsilon_l}\left|g^l\left(x^l_t, y^l_t, u^l_t, t\right) - g^l\left(x^l_{\theta_l}, y^l_{\theta_l}, u^l_{\theta_l}, \theta_l\right)\right|^2 dt\Bigg]\end{aligned}$$

Hence: $\forall t \in [\theta_l, \theta_l+\varepsilon_l)$, $E\left|\tilde{x}^l_{t+\varepsilon_l,\varepsilon_l}\right|^2 \le N$ if $\varepsilon_l \to 0$. Further for $\forall t \in [\theta_l+\varepsilon_l, t_l]$:

$$\begin{aligned}d\tilde{x}^l_{t,\varepsilon} &= \left[g^l\left(x^{\theta_l}_{t,\varepsilon_l}, y^{\theta_l}_{t,\varepsilon_l}, u^l_t, t\right) - g^l\left(x^l_t, y^l_t, u^l_t, t\right)\right]dt\\ &+ \left[f^l\left(x^{\theta_l}_{t,\varepsilon_l}, y^{\theta_l}_{t,\varepsilon_l}, t\right) - f\left(x^l_t, y^l_t, t\right)\right]dw^l_t.\end{aligned}$$

Consequently there occurs:

$$d\tilde{x}^l_{t,\varepsilon_l} = \int_0^1 g^l_x\left(x^l_t + \mu\varepsilon_l\tilde{x}^l_{t,\varepsilon_l}, y^l_t, u^l_t, t\right)\tilde{x}^l_{t,\varepsilon_l}d\mu dt$$

$$+ \int_0^1 g^l_y\left(x^l_t, y^l_t + \mu\varepsilon_l\tilde{y}^l_{t,\varepsilon_l}, u^l_t, t\right)\tilde{y}^l_{t,\varepsilon_l}d\mu dt$$

$$+ \int_0^1 f^l_x\left(x^l_t + \mu\varepsilon_l\tilde{x}^l_{t,\varepsilon_l}, y^l_t, t\right)\tilde{x}^l_{t,\varepsilon_l}d\mu dt$$

$$+ \int_0^1 f^l_y\left(x^l_t, y^l_t + \mu\varepsilon_l\tilde{y}^l_{t,\varepsilon_l}, t\right)\tilde{x}^l_{t,\varepsilon_l}d\mu dt$$

$$+ \tilde{x}^l_{\theta_l+\varepsilon_l,\varepsilon_l} = g\left(x^l_{\theta_l+\varepsilon}, y^l_{\theta_l+\varepsilon}, \tilde{u}^l, \theta_l\right) - g^l\left(x^l_{\theta_l+\varepsilon}, y^l_{\theta_l+\varepsilon}, u^l_{\theta_l+\varepsilon}, \theta_l\right).$$

Hence: $E\left|\tilde{x}^l_{t_l,\varepsilon_l}\right|^2 \le N$, for $\forall t \in [\theta_l + \varepsilon_l, t_l]$, if $\varepsilon_l \to 0$. Thus: $\sup\limits_{t_{l-1}\le t\le t_l} E|\tilde{x}_{t,\varepsilon_l}|^2 \le N$. Lemma 1 is proved.

From the expression (10.13), due to Lemma 1 the following estimation is obtained: $\eta^{\theta_l+\varepsilon_l}_{\theta_l} = o(\varepsilon_l)$. Then according to optimality of controls u^l_t, $l = \overline{1,r}$ from Eq. (10.15) for each l it follows that:

$$\Delta_{\theta^l} J(u) = -E\varepsilon_l\left\lfloor\psi^{l*}_{\theta_l}\Delta_{\tilde{u}^l}g^l\left(x^l_{\theta_l}, u^l_{\theta_l}, \theta_l\right) - \Delta_{\tilde{u}^l}p^l\left(x^l_{\theta_l}, u^l_{\theta_l}, \theta_l\right)\right\rfloor\Delta\bar{t}_l + o(\varepsilon_l) \ge 0$$

Hence, due to sufficient smallness of ε_l it follows that Eq. (10.8) is fulfilled. Theorem 1 is proved.

10.4 Switching System with Constraints

Further, by applying Theorem 1 and Ekeland's Variational Principle [22] it is obtained the necessary condition of optimality for stochastic control problem of switching systems with delay (Eqs. 10.1–10.6).

Theorem 2. Suppose that $\pi^r = \left(t_0, \ldots, t_r, x^1_t, x^2_t, \ldots x^r_t, K_1, \ldots, K_r, u^1, u^2, \ldots, u^r\right)$ is an optimal solution of problem (10.1)–(10.6) and random processes $\left(\psi^l_t, \beta^l_t\right) \in L^2_{F^l}(t_{l-1}, t_l; R^{n_l}) \times L^2_{F^l}(t_{l-1}, t_l; R^{n_l x n_l})$ are the solutions of the following adjoint equations:

$$\begin{cases} d\psi_t^l = -\left[H_x^l\left(\psi_t^l, x_t^l, y_t^l, u_t^l, t\right) + H_y^l\left(\psi_{t+h}^l, x_{t+h}^l x_t^l, u_t^l, t\right)\right] dt + \beta_t^l dw_t^l, \\ \qquad\qquad t_{l-1} \le t < t_l - h, \\ d\psi_t^l = -H_x^l\left(\psi_t^l, x_t^l, y_t^l, u_t^l, t\right) dt + \beta_t^l dw_t^l, \quad t_{l-1} - h \le t < t_l \\ \psi_{t_l}^l = \psi_{t_l}^{l+1}\Phi_x^l\left(x_{t_l}^l, t_l\right) - \lambda_l q_x^l\left(x_{t_l}^l\right), \quad l = \overline{1, r-1}, \\ \psi_{t_r}^r = -\lambda_0\varphi_x^r\left(x_{t_r}^r\right) - \lambda_r q_x^r\left(x_{t_r}^r\right). \end{cases} \tag{10.16}$$

Then, maximum principle (Eq. (10.8)) and transversality conditions (10.9) hold.

Proof. For any natural j let's introduce the approximating functional:

$$I_j(\mathbf{u}) = S_j^l\left(E\varphi^r\left(x_{t_r}^r\right) + E\sum_{l=1}^{r}\int_{t_{l-1}}^{t_l} p^l\left(x_t^l, u_t^l, t\right) dt, Eq^l\left(x_{t_l}^l\right)\right)$$

$$= \min_{(c_j, y_l)\in\varepsilon}\sqrt{\left|c_j - 1/j - E\left[\varphi_{t_r}^r\left(x_{t_r}^r\right) + \sum_{l=1}^{r} M_l\right]\right|^2 + \sum_{l=1}^{r}\left|y_l - Eq^l\left(x_{t_l}^l\right)\right|^2},$$

where $\varepsilon = \left\{c : c \le J^0, y_l \in G^l\right\}$, $M_l = \int_{t_{l-1}}^{t_l} p\left(x_t^l, u_t^l, t\right) dt$ and J^0 is minimal value of the functional in the problem (10.1)–(10.5). Let $\mathbf{V} \equiv \left(V^1, \ldots, V^r\right)$, here $V^k \equiv \left(U^k, d\right)$ be space of controls obtained by means of the following metric: $d\left(u^k, v^k\right) = (l \otimes P)\left\{(t, \omega) \in [t_{k-1}, t_k] \times \Omega : v_t^k \ne u_t^k\right\}$.

It is easy to prove the following fact:

Lemma 2. Assume that $u_t^{l,n}$, $l = \overline{1, r}$ be the sequence of admissible controls from V^l, and $x_t^{l,n}$ be the sequence of corresponding trajectories of the system (10.1)–(10.3). If the following condition is met: $d\left(u_t^{l,n}, u_t^l\right) \to 0$. Then

$$\lim_{n\to\infty}\left\{\sup_{t_{l-1}\le t\le t_l} E\left|x_t^{l,n} - x_t^l\right|^2\right\} = 0$$

where x_t^l is a trajectory corresponding to an admissible controls u_t^l, $l = 1, \ldots, r$.

According to Ekeland's variational principle, there are controls such as; $u_t^{l,j}$: $d\left(u_t^{l,j}, u_t^l\right) \le \sqrt{\varepsilon_j^l}$ and for $\forall u_t^l \in V^l$ the following is achieved:

$$I_j\left(\mathbf{u}^j\right) \le I_j(\mathbf{u}) + \sum_{l=1}^{r}\sqrt{\varepsilon_j^l} d\left(u^{l,j}, u^l\right), \quad \varepsilon_j^l = \frac{1}{j}.$$

This inequality means that $(t_0, t_1, \ldots, t_r, x_t^{1,j}, \ldots, x_t^{r,j}, K_1, \ldots, K_r, u_t^{1,j}, \ldots, u_t^{r,j})$ is a solution of the following problem:

$$\begin{cases} J_j(\mathbf{u}) = I_j\left(\mathbf{u}^j\right) + \sum_{l=1}^{r} \sqrt{\varepsilon_j^l}\, E \int_{t_{l-1}}^{t_l} \delta\left(u_t^l, u_t^{l,j}\right) dt \to \min \\ dx_t^l = g^l\left(x_t^l, y_t^l, u_t^l, t\right) dt + f^l\left(x_t^l, y_t^l, t\right) dw_t, \quad t \in (t_{l-1}, t_l] \\ x_t^{1+1} = \mathrm{K}^{l+1}(t), \quad t \in [t_l - h, t_l), \ l = 0, 1, \ldots, r-1, \\ x_{t_l}^{l+1} = \Phi^l\left(x_{t_l}^l, t_l\right) \quad l = 1, \ldots, r, \ x_{t_0}^1 = x_0, \ u_t^l \in U_\partial^l \end{cases} \tag{10.17}$$

Function $\delta(u, v)$ is determined in the following way: $\delta(u, v) = \begin{cases} 0, u = v \\ 1, u \neq v. \end{cases}$

Then according to the Theorem 1, it is obtained as follows:

1. There exist the random processes $\psi_t^{l,j} \in L_{F^l}^2(t_{l-1}, t_l; R^{n_l})$, $\beta_t^{l,j} \in L_{F^l}^2(t_{l-1}, t_l; R^{n_l \times n_l})$ which are solutions of the following system in $t \in [t_{l-1}, t_l - h)$:

$$\begin{cases} d\psi_t^{l,j} = -H_x^l\left(\psi_t^{l,j}, x_t^{l,j}, y_t^{l,j}, u_t^{l,j}, t\right) dt \\ \qquad\qquad -H_y^l\left(\psi_{t+h}^{l,j}, x_{t+h}^{l,j}, y_{t+h}^{l,j}, u_{t+h}^{l,j}, t+h\right) dt + \beta_t^{l,j}\, dw_t, \\ d\psi_t^{l,j} = -H_x^l\left(\psi_t^{l,j}, x_t^{l,j}, y_t^{l,j}, u_t^{l,j}, t\right) dt + \beta_t^{l,j}\, dw_t, \quad t \in [t_{l-1} - h, t_l), \\ \psi_{t_l}^{l,j} = \psi_{t_l}^{l+1,j}\, \Phi_x^l\left(x_{t_l}^{l,j}, t_l\right) - \lambda_l^j q_x^l\left(x_{t_l}^{l,j}\right), \quad l = 1, \ldots, r-1 \\ \psi_{t_r}^r = -\lambda_0^j \varphi_x^r\left(x_{t_r}^{r,j}\right) - \lambda_r^j q_x^r\left(x_{t_r}^{r,j}\right). \end{cases} \tag{10.18}$$

where non-zero $\left(\lambda_0^j, \lambda_1^j, \ldots, \lambda_r^j\right) \in R^{r+1}$ meet the following requirement:

$$\lambda_l^j = \left[-y + E q^l\left(x_{t_l}^{l,j}\right)\right] / J_j^0,$$

$$\lambda_0^j = \left(-c_l + 1/j + E\varphi^r\left(x_{t_r}^{r,j}\right) + E\sum_{l=1}^{r} \int_{t_{l-1}}^{t_l} p^l\left(x_t^{l,j}, u_t^{l,j}, t\right) dt\right) / J_j^0$$

here $J_j^0 = \sqrt{\left|c_j - 1/j - E\left[\varphi_{t_r}^r\left(x_{t_r}^r\right) + \sum_{l=1}^{r} M_l\right]\right|^2 + \sum_{l=1}^{r} \left|y - E q^l\left(x_{t_l}^l\right)\right|^2}$,

2. Almost certainly for any $\tilde{u}^l \in U^l$ and a.e. $t \in [t_{l-1}, t_l]$ is satisfied:

$$H^l\left(\psi_t^{l,j}, x_t^{l,j}, y_t^{l,j}, \tilde{u}_t^l, t\right) - H^l\left(\psi_t^{l,j}, x_t^{l,j}, y_t^{l,j}, u_t^{l,j}, t\right) \leq 0 \tag{10.19}$$

3. The following transversality conditions hold:

$$\begin{aligned} &- a_l \psi_{t_l}^{l,j} g^l\left(x_{t_l}^{l,j}, y_{t_l}^{l,j}, u_{t_l}^{l,j}, t_l\right) + b_l \psi_{t_l}^{l+1,j} g^{l+1}\left(x_{t_l}^{l,j}, y_{t_l}^{l,j}, u_{t_l}^{l,j}, t_l\right) \\ &+ b_l \psi_{t_l+h}^{l+1,j} g^{l+1}\left(x_{t_l}^{l,j}, K^{l,j}(t_l), u_{t_l}^{l,j}, t_l\right) + b_l \Phi_t^l\left(x_{t_l}^{l,j}, t_l\right) = 0, \quad l = \overline{0,r} \end{aligned} \tag{10.20}$$

Since the following exists $\left|\left(\lambda_0^j, \lambda_1^j, \ldots, \lambda_r^j\right)\right| = 1$, then according to conditions I–IV it is implied that $\left(\lambda_0^j, \lambda_1^j, \ldots, \lambda_r^j\right) \to (\lambda_0, \lambda_1, \ldots, \lambda_r)$ if $j \to \infty$.

Let us introduce the following result which will be needed in the future.

Lemma 3. Let $\psi_{t_l}^l$ be a solution of system (10.16), and $\psi_{t_l}^{l,j}$ be a solution of system (10.18). Then $E\int_{t_{l-1}}^{t_l} \left|\psi_t^{l,j} - \psi_t^l\right|^2 dt + E\int_{t_{l-1}}^{t_l} \left|\beta_t^{l,j} - \beta_t^l\right|^2 dt \to 0$, if $d\left(u_t^{l,j}, u_t^l\right) \to 0, \quad j \to \infty$.

Proof. It is clear that $\forall t \in [t_{l-1}, t_l]$, $l = 1, \ldots, r-1$:

$$\begin{aligned} d\left(\psi_t^{l,j} - \psi_t^l\right) = &-\left[H_x^l\left(\psi_t^{l,j}, x_t^{l,j}, y_t^{l,j}, u_t^{l,j}, t\right) - H_x^l\left(\psi_t^l, x_t^l, y_t^l, u_t^l, t\right)\right] dt \\ &+ \left(\beta_t^{l,j} - \beta_t^l\right) dw_t. \end{aligned}$$

According to Ito formula, for $\forall s \in [t_l - h, t_l]$ it is satisfied:

$$\begin{aligned} E\left|\psi_{t_l}^{l,j} - \psi_{t_l}^l\right|^2 - E\left|\psi_s^{l,j} - \psi_s^l\right|^2 = 2E\int_s^{t_r} &\left[\psi_t^{l,j} - \psi_t^l\right]\Big[\left(g_x^{l*}\left(x_t^{l,j}, y_t^{l,j}, u_t^{l,j}, t\right)\right. \\ &- g_x^{l*}\left(x_t^l, y_t^l, u_t^l, t\right)\Big)\psi_t^{l,j} + g_x^{l*}\left(x_t^l, y_t^l, u_t^l, t\right)\left(\psi_t^{l,j} - \psi_t^l\right) \\ &+ \left(f_x^{l*}\left(x_t^{l,j}, y_t^{l,j}, t\right) - f_x^{l*}\left(x_t^l, y_t^l, t\right)\right)\beta_t^{l,j} - p^l\left(x_t^{l,j}, u_t^{l,j}, t\right) \\ &+ p_x^l\left(x_t^l, u_t^l, t\right)\Big]dt + E\int_s^{t_r} \left|\beta_t^{l,j} - \beta_t^l\right|^2 dt. \end{aligned}$$

Due to assumptions I–IV and using simple transformations, the following is obtained:

$$E\int_s^{t_l}\left|\beta_t^{l,j}-\beta_t^l\right|^2dt+E\left|\psi_s^{l,j}-\psi_s^l\right|^2\le EN\int_s^{t_l}\left|\psi_t^{l,j}-\psi_t^l\right|^2dt$$

$$+EN\varepsilon\int_s^{t_r}\left|\beta_t^{l,j}-\beta_t^l\right|^2dt+E\left|\psi_{t_l}^{l,j}-\psi_{t_l}{}^l\right|^2.$$

Consequently, according to Gronwall inequality [10] it suggests that:

$$E\left|\psi_s^{l,j}-\psi_s^l\right|^2\le De^{N(t_r-s)}\quad\text{a.e. in }[t_l-h,t_l]\tag{10.21}$$

where constant D is determined in the way below: $D=E\left|\psi_{t_l}^{l,j}-\psi_{t_l}^l\right|^2$. According to Eqs. (10.16) and (10.18), it is obtained that: $\psi_{t_l}^{l,j}\to\psi_{t_l}^l$, which leads to $D\to 0$ if $j\to\infty$. Hence, from Eq. (10.21) it follows: $\psi_s^{l,j}\to\psi_s^l$ in $L_{F^l}^2(t_l-h,t_l;R^{n_l})$ and $\beta_s^{l,j}\to\beta_s^l$ in $L_{F^l}^2(t_l-h,t_l;R^{n_l\times n_l})$. Then, $\forall t\in[t_{l-1},t_l-h)$, $l=1,\ldots,r$ from the expression:

$$\begin{aligned}d\left(\psi_t^{l,j}-\psi_t^l\right)=&-\left[H_x^l\left(\psi_t^{l,j},x_t^{l,j},y_t^{l,j},u_t^{l,j},t\right)-H_x^l\left(\psi_t^l,x_t^l,y_t^l,u_t^l,t\right)\right]dt\\&-\Big[H_y^l\left(\psi_{t+h}^{l,j},x_{t+h}^{l,j},y_{t+h}^{l,j},u_{t+h}^{l,j},t+h\right)\\&-H_y^l\left(\psi_{t+h}^l,x_{t+h}^l,y_{t+h}^l,u_{t+h}^l,t+h\right)\Big]dt+\left(\beta_t^{l,j}-\beta_t^l\right)dw_t\end{aligned}$$

using simple transformations, in view of assumptions I–IV the following is obtained:

$$E\int_s^{t_l-h}\left|\beta_t^{l,j}-\beta_t^l\right|^2dt+E\left|\psi_s^{l,j}-\psi_s^l\right|^2\le EN\int_s^{t_l-h}\left|\psi_t^{l,j}-\psi_t^l\right|^2dt$$

$$+EN\varepsilon\int_s^{t_l-h}\left|\beta_t^{l,j}-\beta_t^l\right|^2dt+E\left|\psi_{t_l-h}^{l,j}-\psi_{t_l-h}^l\right|^2.$$

Hence, according to Gronwall inequality, the following result is achieved:

$$E\left|\psi_s^{l,j}-\psi_s^l\right|^2\le De^{N(t_l-s)}\quad\text{a.e. in }[t_{l-1},t_l-h)$$

where constant D is determined as follows: $D = E\left|\psi^{l,j}_{t_l-h} - \psi^{l}_{t_l-h}\right|^2$, which leads to $D \to 0$ if $j \to \infty$.It is inferred that $\psi^{l,j}_s \to \psi^l_s$ in $L^2_{F^l}(t_{l-1}, t_l; R^{n_l})$ and $\beta^{l,j}_s \to \beta^l_s$ in $L^2_{F^l}(t_{l-1}, t_l; R^{n_l \times n_l})$. Lemma 2 is proved.

It follows from Lemma 2 and Lemma 3 that it can be proceeded to the limit in system (10.18) and the fulfillments of Eq. (10.16) are obtained. Following the similar scheme by taking limit in Eqs. (10.19) and (10.20) it is proved that Eqs. (10.8) and (10.9) are true. Theorem 2 is proved.

10.5 Conclusion

This work deals with description the natural phenomena with memory and investigation of optimal control problems of such systems. Necessary conditions satisfied by an optimal solution, play an important role for analysis of control problems. It is well known that every optimal solution satisfies the maximum principle. In this paper a maximum principle for stochastic optimal control problem of switching systems with delay on state is obtained. The results can be used in various optimal control problems of biological, physics, economic systems and a lot of life science, financial market applications. The necessary conditions developed in this manuscript can be viewed as a stochastic analogues of the problems formulated in [21, 23–25]. Withal, Theorem 1 and Theorem 2 is a natural improving of the results given in [20, 26–28].

References

1. Gikhman, I.I., Skorokhod, A.V.: Stochastic Differential Equations. Springer, Berlin (1972)
2. Mao, X.: Stochastic Differential Equations and Their Applications. Horwood Publication House, Chichester (1997)
3. Chojnowska-Michalik, A.: Representation theorem for general stochastic delay equations. Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys. **26**(7), 635–642 (1978)
4. Kolmanovsky, V.B., Myshkis, A.D.: Applied Theory of Functional Differential Equations. Kluwer, Dordrecht (1992)
5. Agayeva, C.A., Allahverdiyeva, J.J.: On one stochastic optimal control problem with variable delays, Journal of Theory of stochastic processes, Kiev **13**(29), 3–11 (2007)
6. El-Bakry, H.M., Mastorakis, N.: Fast packet detection by using high speed time delay neural networks. In: Chen, S., Guan, Q. (eds.) Proceedings of the 10th WSEAS International Conference on Multimedia Systems & Signal Processing, pp. 222–227 (2010)
7. Chernousko, F.L., Ananievski, I.M., Reshmin, S.A.: Control of Nonlinear Dynamical Systems: Methods and Applications (Communication and Control Engineering). Springer, Berlin (2008)
8. Elsanosi, I., Øksendal, B., Sulem, A.: Some solvable stochastic control problems with delay. Stoch. Stoch. Rep. **71**(1–2), 69–89 (2000)
9. Federico, S., Golds, B., Gozzi, F.: HJB equations for the optimal control of differential equations with delays and state constraints, II: optimal feedbacks and approximations. SIAM J. Control Optim. **49**, 2378–2414 (2011)

10. Fleming, W.H., Rishel, R.W.: Deterministic and Stochastic Optimal Control. Springer, New York (1975)
11. Larssen, B.: Dynamic programming in stochastic control of systems with delay. Stoch. Stoch. Rep. **74**(3–4), 651–673 (2002)
12. Vinter, R.B., Kwong, R.H.: The infinite time quadratic control problem for linear systems with state and control delays: an evolution equation approach. SIAM J. Control Optim. **19**(1), 139–153 (1981)
13. Boukas, E.-K.: Stochastic Switching Systems: Analysis and Design. Birkhauer, Boston (2006)
14. Avezedo, N., Pinherio, D., Weber, G.W.: Dynamic programing for a Markov-switching jump diffusion. J. Comput. Appl. Math. **267**, 1–19 (2014)
15. Shen, H., Xu, S., Song, X., Luo, J.: Delay-dependent robust stabilization for uncertain stochastic switching systems with distributed delays. Asian J. Control **5**(11), 527–535 (2009)
16. Aghayeva, C.A., Abushov, Q.: The maximum principle for the nonlinear stochastic optimal control problem of switching systems. J. Glob. Optim. **56**(2), 341–352 (2013)
17. Aghayeva, Ch.A., Abushov, Q.: Stochastic optimal control problem for switching system with controlled diffusion coefficients. In: Ao, S.I., Gelman, L., Hukins, D.W.L. (eds.) Book Series: Lecture Notes in Engineering and Computer Science, vol. 1, pp. 202–207 (2013)
18. Hall, E., Hanagud, S.: Control of nonlinear structural dynamic systems—chaotic vibrations. J. Guid. Control Dyn. **16**(3), 470–476 (1993)
19. Aghayeva, C.A.: Stochastic optimal control problem of switching systems with lag. Trans. ANAS Math. Mech. Ser. Phys.-Tech. Math. Sci. **31**(3), 68–73 (2011)
20. Aghayeva, Ch.A.: Necessary condition of optimality for stochastic switching systems with delay. In: Senichenkov, Y., Korablev, V., et al. (eds.) Proceedings of International Conference MMAS'14, pp. 54–58 (2014).
21. Kharatatishvili, G., Tadumadze, T.: The problem of optimal control for nonlinear systems with variable structure, delays and piecewise continuous prehistory. Mem. Diff. Equat. Math. Phys. **11**, 67–88 (1997)
22. Ekeland, I.: On the variational principle. J. Math. Anal. Appl. **47**, 324–353 (1974)
23. Capuzzo, D.I., Evans, L.C.: Optimal switching for ordinary differential equations. SIAM J. Control Optim. **22**(1), 143–161 (1984)
24. Bengea, S.C., Raymond, A.C.: Optimal control of switching systems. Automatica **41**, 11–27 (2005)
25. Seidmann, T.I.: Optimal control for switching systems. In: Proceedings of the 21st Annual Conference on Informations Science and Systems, pp. 485–489 (1987)
26. Agayeva, Ch., Abushov, Q.: Necessary condition of optimality for stochastic control systems with variable structure. In: Sakalauskas, L., Weber, G., Zavadskas, E. (eds.) Proceedings of EurOPT 2008, pp. 77–81 (2008)
27. Abushov, Q., Aghayeva, C.: Stochastic maximum principle for the nonlinear optimal control problem of switching systems. J. Comput. Appl. Math. **259**, 371–376 (2014)
28. Aghayeva, Ch., Abushov, Q.: Stochastic maximum principle for switching systems. In: AidaZade, K. (eds.) 4th International Conference PCI, vol. 3, pp. 198–201 (2012)

Chapter 11
Computer Simulation of Emission and Absorption Spectra for LH2 Ring

Pavel Heřman and David Zapletal

Abstract Computer simulation of absorption and steady state fluorescence spectra for molecular system is presented. We focus on the B850 ring from peripheral cyclic antenna unit LH2 of the bacterial photosystem from purple bacteria. Uncorrelated static disorder in radial positions of molecules on the ring is taking into account in our simulations. We consider also influence of dynamic disorder, interaction with phonon bath, in Markovian approximation. Spectral responses are calculated by the cumulant-expansion method of Mukamel et al. Procedure in Fortran was created for calculation of single ring spectra within full Hamiltonian model. These new results are compared with our previous ones (within the nearest neighbour approximation model) that were obtained by software package Mathematica.

11.1 Introduction

The first (light) stage of photosynthesis consists of very effective processes. Solar photon is absorbed by a complex system of membrane-associated pigment-proteins (light-harvesting (LH) antenna) and absorbed energy is transferred to a reaction center (RC), where it is converted into a chemical energy [1].

Our interest is focused on antenna systems from purple bacteria, that are formed by ring antenna complexes LH1, LH2, LH3, and LH4. These systems are relatively simple and symmetric and their geometric structures are known in great details from X-ray crystallography. General organization of above mentioned light-harvesting complexes is the same: identical subunits are repeated cyclically in such a way that a ring-shaped structure is formed. However the symmetries of these rings are different.

P. Heřman (✉)
Faculty of Science, Department of Physics, University of Hradec Králové, Rokitanského 62, Hradec Králové 500 03, Czech Republic
e-mail: pavel.herman@uhk.cz

D. Zapletal
Faculty of Economics and Administration, Institute of Mathematics and Quantitative Methods, University of Pardubice, Studentská 95, Pardubice 532 10, Czech Republic
e-mail: david.zapletal@upce.cz

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_11

Crystal structure of LH2 complex contained in purple bacterium *Rhodopseudomonas acidophila* was first described in high resolution by McDermott et al. [2], then further e.g. by Papiz et al. [3]. The bacteriochlorophyll (BChl) molecules are organized in two concentric rings. One ring features a group of nine well-separated BChl molecules (B800) with absorption band at about 800 nm. The other ring consists of 18 closely packed BChl molecules (B850) absorbing around 850 nm. LH2 complexes from other purple bacteria have analogous ring structure.

Some bacteria contain also other types of complexes such as the B800–820 LH3 complex (*Rhodopseudomonas acidophila* strain 7050) or the LH4 complex (*Rhodopseudomonas palustris*). LH3 complex like LH2 one is usually nonameric but LH4 one is octameric. While the B850 dipole moments in LH2 ring have tangential arrangement, in the LH4 ring they are oriented more radially. Mutual interactions of the nearest neighbour BChls in LH4 are approximately two times smaller in comparison with LH2 and have opposite sign. The other difference is the presence of an additional BChl ring in LH4 complex [4]. Different arrangements manifest themselves in different optical properties.

Despite intensive study of bacterial antenna systems, e.g. [2–4], the precise role of the protein moiety for governing the dynamics of the excited states is still under debate. At room temperature the solvent and protein environment fluctuates with characteristic time scales ranging from femtoseconds to nanoseconds. The simplest approach is to substitute fast fluctuations by dynamic disorder and slow fluctuations by static disorder.

In our previous papers we presented results of our simulations doing within the nearest neighbour approximation model. In several steps we extended the former investigations of static disorder effect on the anisotropy of fluorescence made by Kumble and Hochstrasser [5] and Nagarajan et al. [6–8] for LH2 ring. After studying the influence of diagonal dynamic disorder for simple systems (dimer, trimer) [9–11], we added this effect into our model of LH2 ring by using a quantum master equation in Markovian and non-Markovian limits [12–16]. We also studied influence of four types of uncorrelated static disorder (Gaussian disorder in local excitation energies, Gaussian disorder in transfer integrals, Gaussian disorder in radial positions of BChls on the ring and Gaussian disorder in angular positions of BChls on the ring) [17–21]. Influence of correlated static disorder, namely an elliptical deformation of the ring, was also taken into account [14, 21]. We also investigated the time dependence of fluorescence anisotropy for the LH4 ring with different types of uncorrelated static disorder [15, 19, 20].

Recently we have focused on the modeling of absorption and steady state fluorescence spectra. Our results for LH2 and LH4 rings within the nearest neighbour approximation model have been presented in [22–27]. The results within full Hamiltonian model have been published in [28–30].

Main goal of our present paper is the comparison of the results for B850 ring from LH2 complex calculated within full Hamiltonian model with our previous results calculated within the nearest neighbour approximation model [23, 24]. The rest of the paper is organized as follows. Section 11.2 introduces the ring model with the static disorder and dynamic disorder (interaction with phonon bath) and

the cumulant expansion method, which is used for the calculation of spectral responses of the system with exciton-phonon coupling. Computational point of view is mentioned in Sect. 11.3. Results of our simulations and used units and parameters could be found in Sect. 11.4, some conclusions are drawn in Sect. 11.5.

11.2 Physical Model

Because of strong interaction between bacteriochlorophyll molecules our theoretical approach considers an extended Frenkel exciton model. We assume that only one exciton is present on the ring after an impulsive excitation. Hamiltonian of an exciton on the ideal ring coupled to a bath of harmonic oscillators reads

$$H^0 = H^0_{\mathrm{ex}} + H_{\mathrm{ph}} + H_{\mathrm{ex-ph}}. \tag{11.1}$$

Here the first term,

$$H^0_{\mathrm{ex}} = \sum\nolimits_{m,n(m\neq n)} J_{mn} a^+_m a_n, \tag{11.2}$$

corresponds to an exciton, e.g. the system without any disorder. The operator a^+_m (a_n) creates (annihilates) an exciton at site m, J_{mn} (for $m \neq n$) is the so-called transfer integral between sites m and n. The second term in Eq. (11.1),

$$H_{\mathrm{ph}} = \sum\nolimits_q \hbar\omega_q b^+_q b_q, \tag{11.3}$$

represents phonon bath in harmonic approximation (phonon creation and annihilation operators are denoted by b^+_q and b_q, respectively). Last term in Eq. (11.1),

$$H_{\mathrm{ex-ph}} = \frac{1}{\sqrt{N}} \sum\nolimits_m \sum\nolimits_q G^m_q \hbar\omega_q a^+_m a_m \left(b^+_q + b_q\right), \tag{11.4}$$

describes exciton-phonon interaction which is assumed to be site-diagonal and linear in bath coordinates (the term G^m_q denotes exciton-phonon coupling constant).

Inside one ring the pure exciton Hamiltonian H^0_{ex} can be diagonalized using the wave vector representation with corresponding delocalized "Bloch" states α and energies E_α. Considering homogeneous case with only the nearest neighbour transfer matrix elements

$$J_{mn} = J_0 \left(\delta_{m,n+1} + \delta_{m,n-1}\right) \tag{11.5}$$

and using Fourier transformed excitonic operators (Bloch representation)

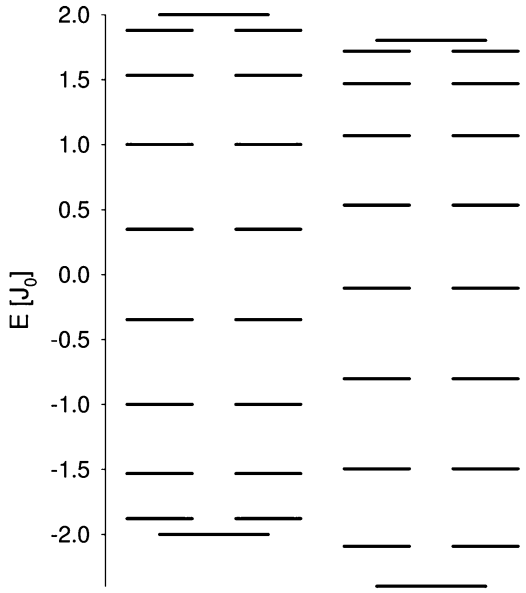


Fig. 11.1 Energetic band structure of B850 ring from LH2 (*left column*—the nearest neighbour approximation model, *right column*—full Hamiltonian model)

$$a_\alpha = \sum\nolimits_n a_n \mathrm{e}^{\mathrm{i}\alpha n}, \quad \alpha = \frac{2\pi}{N} l, \quad l = 0, \pm 1, \ldots, \pm\frac{N}{2}, \tag{11.6}$$

the simplest exciton Hamiltonian in α-representation reads

$$H_{\mathrm{ex}}^0 = \sum\nolimits_\alpha E_\alpha a_\alpha^+ a_\alpha, \quad E_\alpha = -2J_0 \cos\alpha, \tag{11.7}$$

(see Fig. 11.1—left column). In case of the full Hamiltonian model (dipole-dipole approximation), energetic band structure slightly differs (Fig. 11.1—right column). Differences of energies in lower part of the band are larger and in upper part of the band are smaller in comparison with the nearest neighbour approximation model.

Influence of uncorrelated static disorder is modeled by the fluctuations δr_n of radial positions of bacteriochlorophylls on the ring with Gaussian distribution and standard deviation Δ_r. The Hamiltonian H_s of the uncorrelated static disorder adds to the Hamiltonian H_{ex}^0 of the ideal ring.

The cumulant-expansion method of Mukamel et al. [31, 32] is used for the calculation of spectral responses of the system with exciton-phonon coupling. Absorption $OD(\omega)$ and steady-state fluorescence $FL(\omega)$ spectrum can be expressed as

$$OD\,(\omega) = \omega \sum\nolimits_\alpha d_\alpha^2 \; Re \int_0^\infty dt \; \mathrm{e}^{\mathrm{i}(\omega-\omega_\alpha)t - g_{\alpha\alpha\alpha\alpha}(t) - R_{\alpha\alpha\alpha\alpha}t}, \tag{11.8}$$

$$FL(\omega) = \omega \sum\nolimits_{\alpha} P_{\alpha} d_{\alpha}^{2}\, Re \int_{0}^{\infty} dt\; e^{\mathrm{i}(\omega-\omega_{\alpha})t + \mathrm{i}\lambda_{\alpha\alpha\alpha\alpha} t - g^{*}_{\alpha\alpha\alpha\alpha}(t) - R_{\alpha\alpha\alpha\alpha} t}. \tag{11.9}$$

Here $\boldsymbol{d}_{\boldsymbol{\alpha}} = \sum_{n} c_{n}^{\alpha} \boldsymbol{d}_{\boldsymbol{n}}$ is the transition dipole moment of eigenstate α, c_{n}^{α} are the expansion coefficients of the eigenstate α in site representation and P_{α} is the steady state population of the eigenstate α. The inverse lifetime $R_{\alpha\alpha\alpha\alpha}$ of exciton state α [33] is given by the elements of Redfield tensor $R_{\alpha\beta\gamma\delta}$ [34]. It is a sum of the relaxation rates between exciton states,

$$R_{\alpha\alpha\alpha\alpha} = -\sum\nolimits_{\beta\neq\alpha} R_{\beta\beta\alpha\alpha}. \tag{11.10}$$

The g-function and λ-values in Eqs. (11.8) and (11.9) are given by

$$g_{\alpha\beta\gamma\delta} = -\int_{-\infty}^{\infty} \frac{d\omega}{2\pi\omega^{2}} C_{\alpha\beta\gamma\delta}(\omega) \left[\mathrm{cotgh} \frac{\omega}{2k_{\mathrm{B}}T} (\cos\omega t - 1) - \mathrm{i}(\sin\omega t - \omega t) \right], \tag{11.11}$$

$$\lambda_{\alpha\beta\gamma\delta} = -\lim_{t\to\infty} \frac{d}{dt} \mathrm{Im}\left\{ g_{\alpha\beta\gamma\delta}(t) \right\} = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi\omega} C_{\alpha\beta\gamma\delta}(\omega). \tag{11.12}$$

The matrix of spectral densities $C_{\alpha\beta\gamma\delta}(\omega)$ in the eigenstate (exciton) representation reflects one-exciton states coupling to the manifold of nuclear modes. In what follows only a diagonal exciton phonon interaction in site representation is used (see Eq. (11.4)), i.e., only fluctuations of the pigment site energies are assumed and the restriction to the completely uncorrelated dynamic disorder is applied. In such case each site (i.e. each chromophore) has its own bath completely uncoupled from the baths of the other sites. Furthermore it is assumed that these baths have identical properties [13, 35, 36]

$$C_{mnm'n'}(\omega) = \delta_{mn}\delta_{mm'}\delta_{nn'} C(\omega). \tag{11.13}$$

After transformation to the exciton representation we have

$$C_{\alpha\beta\gamma\delta}(\omega) = \sum\nolimits_{n} c_{n}^{\alpha} c_{n}^{\beta} c_{n}^{\gamma} c_{n}^{\delta} C(\omega). \tag{11.14}$$

Various models of spectral density of the bath are used in literature [33, 37, 38]. In our present investigation we have used the model of May and Kűhn [37]

$$C(\omega) = \Theta(\omega)\, j_{0} \frac{\omega^{2}}{2\omega_{c}^{3}} \mathrm{e}^{-\omega/\omega_{c}} \tag{11.15}$$

which has its maximum at $2\omega_{c}$.

11.3 Computational Point of View

To obtain absorption and steady state fluorescence spectra, it is necessary to calculate single ring $OD(\omega)$ and $FL(\omega)$ spectra for large number of different static disorder realizations created by random number generator. Finally, these results have to be averaged over all realizations of static disorder.

For our previous calculations of absorption and steady state fluorescence spectra (with Gaussian uncorrelated static disorder in local excitation energies δE_n and transfer integrals δJ_{mn} taking into account) software package Mathematica [39] was used. Standard numerical integration method used in Mathematica proved to be unsuitable in case of full Hamiltonian model and static disorder δr_n in radial positions of molecules. It was not possible to achieve satisfactory convergence by above mentioned integration method from Mathematica. This is the reason a procedure in Fortran was created for present calculations.

Integrated functions are oscillating and damped (see Eqs. 11.8 and 11.9) and function $\mathrm{Re}\, g_{\alpha\alpha\alpha\alpha}(t)$ is non-negative. Therefore absolute values of integrated functions (for individual α) satisfy inequalities

$$\left|\mathrm{Re}\left\{\mathrm{e}^{\mathrm{i}(\omega-\omega_\alpha)t-g_{\alpha\alpha\alpha\alpha}(t)-R_{\alpha\alpha\alpha\alpha}t}\right\}\right| \leq \mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t}, \tag{11.16}$$

$$\left|\mathrm{Re}\left\{\mathrm{e}^{\mathrm{i}(\omega-\omega_\alpha)t+i\lambda_{\alpha\alpha\alpha\alpha}t-g^*_{\alpha\alpha\alpha\alpha}(t)-R_{\alpha\alpha\alpha\alpha}t}\right\}\right| \leq \mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t}. \tag{11.17}$$

The whole $OD(\omega)$ and $FL(\omega)$ then satisfy

$$OD\,(\omega) \leq \omega \sum\nolimits_\alpha d_\alpha^2 \int_0^\infty \mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t}, \tag{11.18}$$

$$FL\,(\omega) \leq \omega \sum\nolimits_\alpha P_\alpha d_\alpha^2 \int_0^\infty \mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t} \leq \omega \sum\nolimits_\alpha d_\alpha^2 \int_0^\infty \mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t}. \tag{11.19}$$

Predetermined accuracy could be achieved by integration over finite time interval $t \in \langle 0, t_0\rangle$ (instead of $\langle 0, \infty\rangle$). If

$$t_0 \geq \max\{t_\alpha\}, \quad \alpha = 1, \ldots, 18, \tag{11.20}$$

where t_α satisfy condition (Q is arbitrary real positive number)

$$d_\alpha^2 \left[\int_0^\infty dt\ \mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t} - \int_0^{t_\alpha} dt\ \mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t}\right] = d_\alpha^2 \frac{\mathrm{e}^{-R_{\alpha\alpha\alpha\alpha}t_\alpha}}{R_{\alpha\alpha\alpha\alpha}} \leq \frac{Q}{18\omega}, \tag{11.21}$$

i.e.

$$t_\alpha \geq \frac{1}{R_{\alpha\alpha\alpha\alpha}} \ln \frac{18\omega d_\alpha^2}{QR_{\alpha\alpha\alpha\alpha}}, \tag{11.22}$$

then deviations of $OD(\omega)$ and $FL(\omega)$ from precise values are not larger then Q. $OD(\omega)$ and $FL(\omega)$ are therefore integrated as sums of contributions from individual cycles of oscillation. These contributions are added until upper limit of integration exceeds t_0.

11.4 Results

Above mentioned uncorrelated static disorder in radial positions of molecules on the ring has been taken into account in our simulations simultaneously with dynamic disorder in Markovian approximation. Dimensionless energies normalized to the transfer integral $J_{12} = J_0$ in B850 ring from LH2 complex have been used. Estimation of J_0 varies in literature between 250 and 400 cm^{-1}.

All our simulations of LH2 spectra have been done with the same values of J_0 and unperturbed transition energy ΔE_0 from the ground state, that we found for LH2 ring in case of the nearest neighbour approximation model ($J_0 = 400$ cm^{-1}, $\Delta E_0 = 12,300$ cm^{-1}) [23, 24].

Contrary to Novoderezhkin et al. [33], different model of spectral density (the model of May and Kűhn [37]) has been used. In agreement with our previous results [16, 17] we have used $j_0 = 0.4\ J_0$ and $\omega_c = 0.212\ J_0$ (see Eq. 11.15). The strength of uncorrelated static disorder has been taken in agreement with [18]. That is why six strengths $\Delta_r = 0.01, 0.015, 0.02, 0.025, 0.03, 0.06\ r_0$ are used in our simulations.

Resulting steady state fluorescence spectra $FL(\omega)$ for B850 ring from LH2 complex averaged over 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring at low temperature ($kT = 0.1\ J_0$) for both models (full Hamiltonian model and the nearest neighbour approximation one) can be seen in Fig. 11.2. The same, but for room temperature ($kT = 0.5\ J_0$), is shown in Fig. 11.3. Figure 11.4 shows calculated absorption spectra $OD(\omega)$ at low temperature ($kT = 0.1\ J_0$) for both models. The same, but for room temperature ($kT = 0.5\ J_0$), is drawn in Fig. 11.5.

For clarification of the spectral line splitting appearance for low temperature ($kT = 0.5\ J_0$) in case of full Hamiltonian model (see Fig. 11.2), the distribution of the quantity $P_\alpha d_\alpha^2$ (see Eq. 11.9) has been investigated. Here P_α is the steady state population of the eigenstate α and d_α^2 is the dipole strength of eigenstate α. Distributions of this quantity as a function of wavelength λ for low temperature ($kT = 0.1\ J_0$) and 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules are presented in Fig. 11.6 (full Hamiltonian model) and in Fig. 11.7 (the nearest neighbour approximation model). The same, but for room temperature ($kT = 0.5\ J_0$) can be seen in Fig. 11.8 (full Hamiltonian model) and in Fig. 11.9 (the nearest neighbour approximation model).

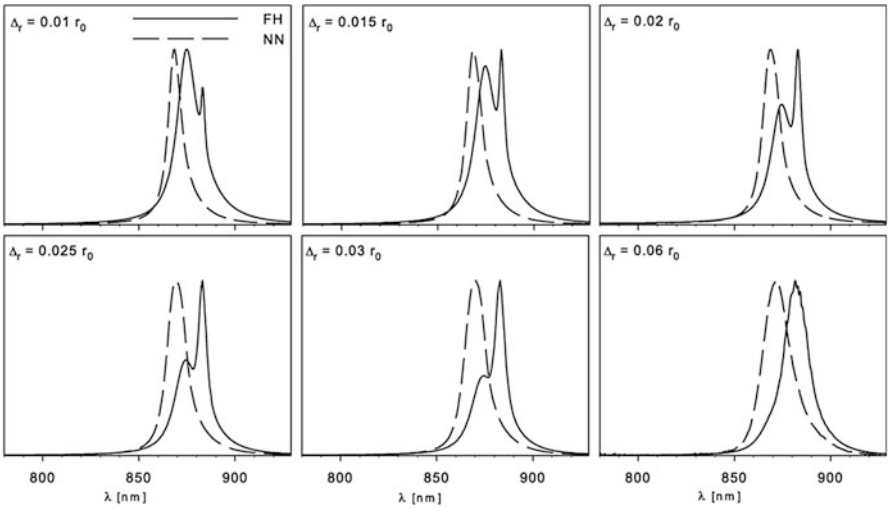


Fig. 11.2 Calculated $FL(\omega)$ spectra averaged over 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n for low temperature $kT = 0.1\ J_0$ (six strengths $\Delta_r = 0.01,\ 0.015,\ 0.02,\ 0.025,\ 0.03,\ 0.06\ r_0$, full Hamiltonian model (FH)—*solid line*, the nearest neighbour approximation model (NN)—*dashed line*)

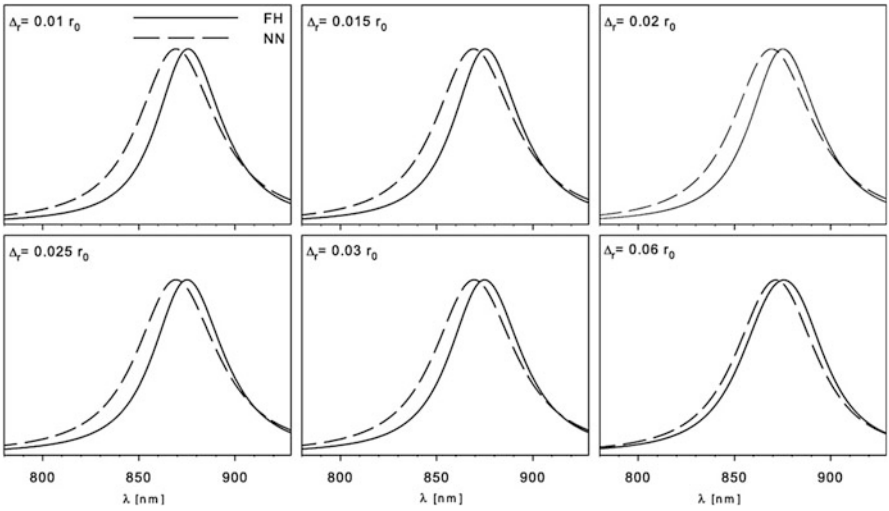


Fig. 11.3 Calculated $FL(\omega)$ spectra averaged over 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n for room temperature $kT = 0.5\ J_0$ (six strengths $\Delta_r = 0.01,\ 0.015,\ 0.02,\ 0.025,\ 0.03,\ 0.06\ r_0$, full Hamiltonian model (FH)—*solid line*, the nearest neighbour approximation model (NN)—*dashed line*)

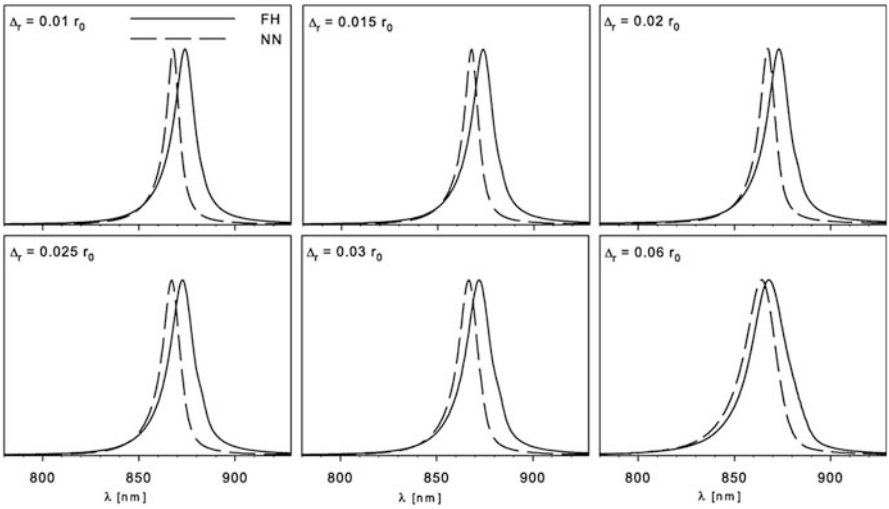


Fig. 11.4 Calculated $OD(\omega)$ spectra averaged over 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n for low temperature $kT = 0.1\ J_0$ (six strengths $\Delta_r = 0.01,\ 0.015,\ 0.02,\ 0.025,\ 0.03,\ 0.06\ r_0$, full Hamiltonian model (FH)—*solid line*, the nearest neighbour approximation model (NN)—*dashed line*)

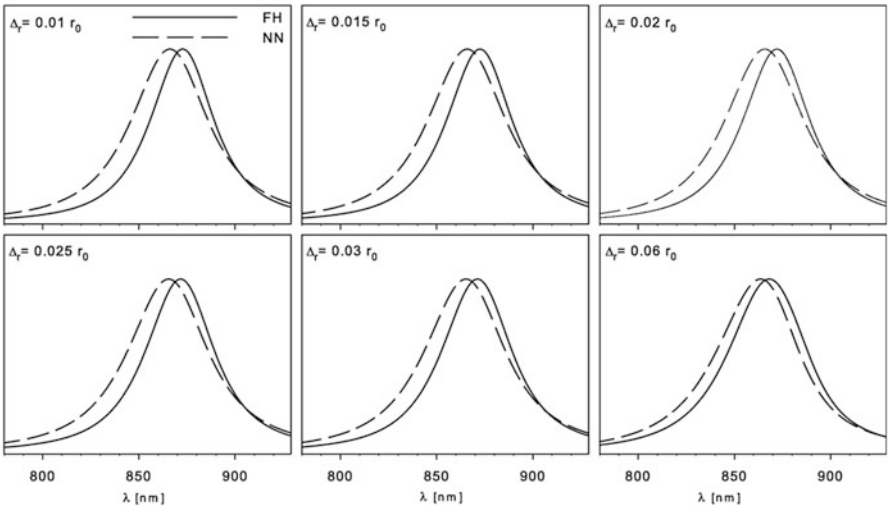


Fig. 11.5 Calculated $OD(\omega)$ spectra averaged over 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n for room temperature $kT = 0.5\ J_0$ (six strengths $\Delta_r = 0.01, 0.015, 0.02, 0.025, 0.03, 0.06\ r_0$, full Hamiltonian model (FH)—*solid line*, the nearest neighbour approximation model (NN)—*dashed line*)

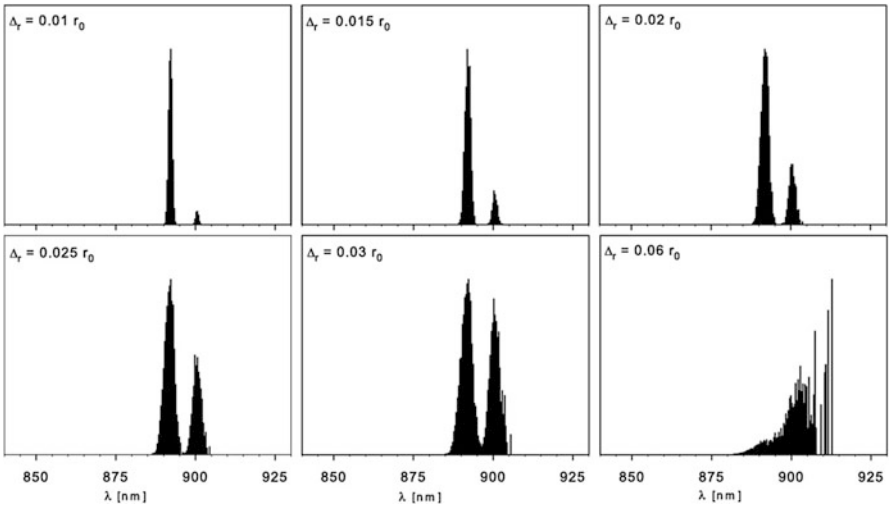


Fig. 11.6 The distribution of the quantity $P_\alpha d_\alpha^2$ as a function of wavelength λ for low temperature $kT = 0.1\ J_0$ and 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n—full Hamiltonian model

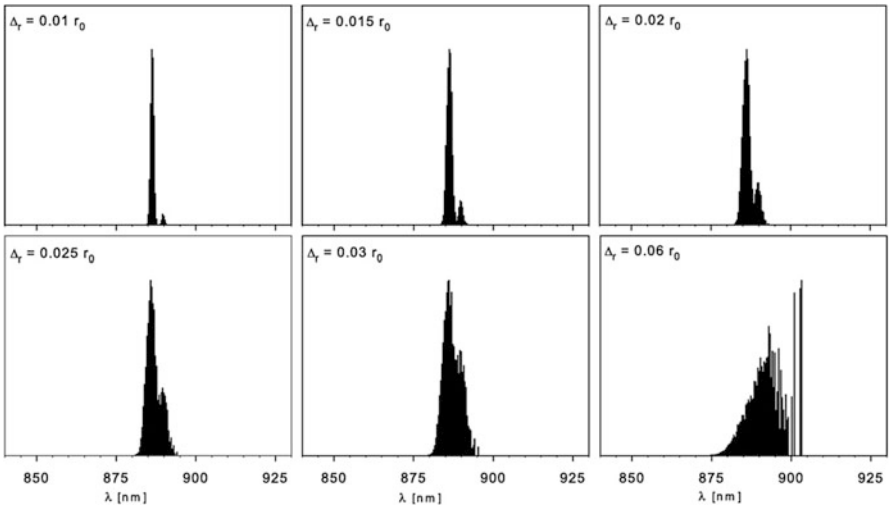


Fig. 11.7 The distribution of the quantity $P_\alpha d_\alpha^2$ as a function of wavelength λ for low temperature $kT = 0.1\ J_0$ and 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n—the nearest neighbour approximation model

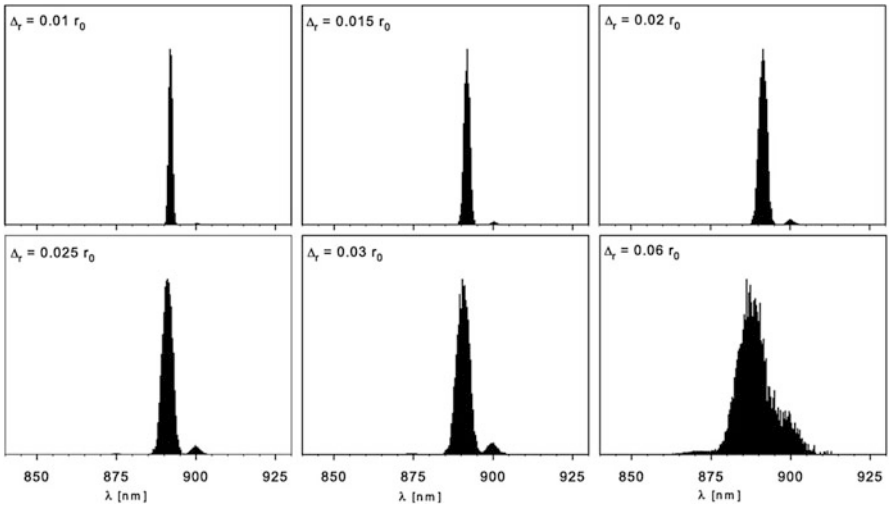


Fig. 11.8 The distribution of the quantity $P_\alpha d_\alpha^2$ as a function of wavelength λ for room temperature $kT = 0.5\ J_0$ and 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n—full Hamiltonian model

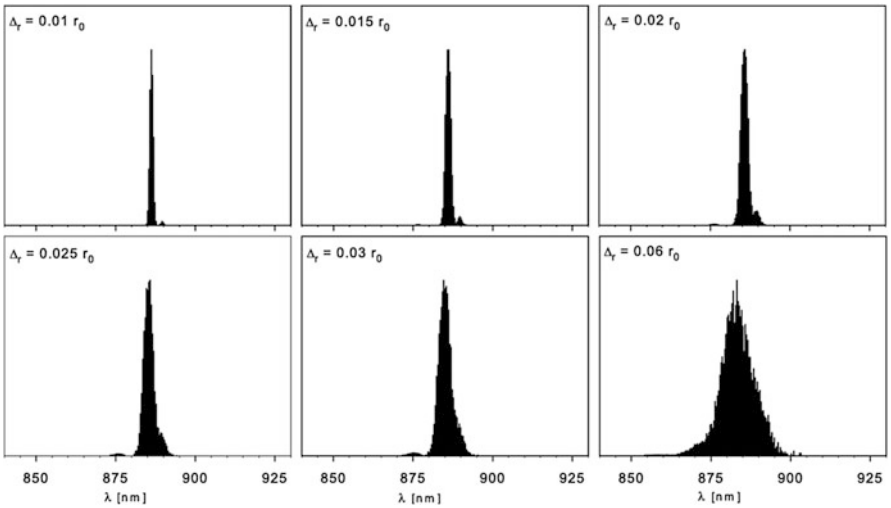


Fig. 11.9 The distribution of the quantity $P_\alpha d_\alpha^2$ as a function of wavelength λ for room temperature $kT = 0.5\ J_0$ and 2,000 realizations of Gaussian uncorrelated static disorder in radial positions of molecules on the ring δr_n—the nearest neighbour approximation model

11.5 Conclusions

Software package Mathematica was found by us very useful for the simulations of the molecular ring spectra in case of static disorder in local excitation energies and transfer integrals. But standard numerical integration method from Mathematica does not provide satisfactory results in case of the static disorder in radial positions of molecules on the ring and full Hamiltonian model. This problem has been solved by use of our procedure created in Fortran.

We compare our new simulated steady state fluorescence and absorption spectra for B850 ring from LH2 complex (full Hamiltonian model, static disorder in radial positions of molecules on the ring) from two aspects. At first the comparison with the spectra in case of the nearest neighbour approximation model (the same type of static disorder) is done (see Figs. 11.2, 11.3, 11.4 and 11.5). Then we compare simulated $FL(\omega)$ and $OD(\omega)$ spectra (full Hamiltonian model, static disorder in radial positions of molecules on the ring) with our previous results (full Hamiltonian model, static disorder in local excitation energies [28] and in transfer integrals [40]). Following conclusions can be made.

General difference in fluorescence and absorption spectra is a shift of the spectral line peak position to higher wavelength for full Hamiltonian model in comparison with the nearest neighbour approximation model. Absorption spectral lines in case of full Hamiltonian model are wider (in particular on the right hand side of spectral profile) in comparison with the nearest neighbour approximation model. For growing strength Δ_r of static disorder, the absorption spectral peak positions move to lower wavelength (especially for full Hamiltonian model). On the other hand, any substantial shift is not visible for the fluorescence spectral line in case of the nearest neighbour approximation model.

The most essential difference in case of low temperature ($kT = 0.1\ J_0$) is fluorescence spectral line splitting ($\Delta_r \in (0.01\ r_0, 0.03\ r_0)$) for full Hamiltonian model. It is caused by different energetic band structure of full Hamiltonian model in comparison with the nearest neighbour approximation model (see Fig. 11.1). In case of the nearest neighbour approximation model any fluorescence spectral line splitting is not visible. In case of room temperature ($kT = 0.5\ J_0$), full Hamiltonian model does not give substantially different fluorescence spectral lines in comparison with the nearest neighbour approximation model and no splitting appears. The reason is dependence of steady state populations P_α on temperature. Also spectral line widening due to dynamic disorder hides differences between both models. This conclusion is supported by different distributions of the quantity $P_\alpha d_\alpha^2$ (see Eq. 11.9) that can be seen in Figs. 11.6, 11.7, 11.8 and 11.9.

As concerns the comparison of the case with static disorder in radial positions of molecules on the ring with other types of static disorder, following conclusion can be done. Fluorescence spectral line splitting is also visible in case of static disorder in local excitation energies [28] and in transfer integrals [40] (again for full Hamiltonian model and low temperature $kT = 0.1\ J_0$). Comparable splitting can be seen for the strengths $\Delta = 0.1\ J_0$ (static disorder in local excitation energies [28]), $\Delta_J = 0.05\ J_0$ (static disorder in transfer integrals [40]) and $\Delta_r = 0.015\ r_0$ (static disorder in radial positions of molecules on the ring).

Acknowledgments This work was supported in part by the Faculty of Science, University of Hradec Králové—specific research project no. 2106/2014.

References

1. van Grondelle, R., Novoderezhkin, V.I.: Energy transfer in photosynthesis: experimental insights and quantitative models. Phys. Chem. Chem. Phys. **8**, 793–807 (2003)
2. McDermott, G., Prince, S.M., Freer, A.A., Hawthornthwaite-Lawiess, A.M., Papiz, M.Z., Cogdell, R.J., Isaacs, N.: Crystal structure of an integral membrane light harvesting complex from photosynthetic bacteria. Nature **374**, 517–521 (1995)
3. Papiz, M.Z., Prince, S.M., Howard, T., Cogdell, R.J., Isaacs, N.W.; The structure and thermal motion of the B800-850 LH2 complex from Rps. acidophila at 2.0 Å over-circle resolution and 100 K: new structural features and functionally relevant motions. J. Mol. Biol. **326**, 1523–1538 (2003)
4. de Ruijter, W., Oellerich, S., Segura, J.-M., Lawless, A., Papiz, M., Aartsma, T.: Observation of the energy level structure of the low-light adapted B800 LH4 complex by single-molecule spectroscopy. Biophys. J. **87**(5), 3413–3420 (2004)
5. Kumble, R., Hochstrasser, R.: Disorder-induced exciton scattering in the light-harvesting systems of purple bacteria: influence on the anisotropy of emission and band → band transitions. J. Chem. Phys. **109**, 855–865 (1998)
6. Nagarajan, V., Alden, R., Williams, J., Parson, W.: Ultrafast exciton relaxation in the B850 antenna complex of Rhodobacter sphaeroides. Proc. Natl. Acad. Sci. U. S. A. **93**(24), 13774–13779 (1996)
7. Nagarajan, V., Johnson, E.T., Williams, J.C., Parson, W.W.: Femtosecond pump-probe spectroscopy of the B850 antenna complex of Rhodobacter sphaeroides at room temperature. J. Phys. Chem. B **103**, 2297–2309 (1999)
8. Nagarajan, V., Parson, W.W.: Femtosecond fluorescence depletion anisotropy: application to the B850 antenna complex of Rhodobacter sphaeroides. J. Phys. Chem. B **104**, 4010–4013 (2000)
9. Čápek, V., Barvík, I., Heřman, P.: Towards proper parametrization in the exciton transfer and relaxation problem: dimer. Chem. Phys. **270**, 141–156 (2001)
10. Heřman, P., Barvík, I.: Towards proper parametrization in the exciton transfer and relaxation problem II. Trimer. Chem. Phys. **274**, 199–217 (2001)
11. Heřman, P., Barvík, I., Urbanec, M.: Energy relaxation and transfer in excitonic trimer. J. Lumin. **108**, 85–89 (2004)
12. Heřman, P., Kleinekathöfer, U., Barvík, I., Schreiber, M.: Exciton scattering in light-harvesting systems of purple bacteria. J. Lumin. **94–95**, 447–450 (2001)
13. He Heřman, P., Kleinekathöfer, U., Barvík, I., Schreiber, M.: Influence of static and dynamic disorder on the anisotropy of emission in the ring antenna subunits of purple bacteria photosynthetic systems. Chem. Phys. **275**, 1–13 (2002)
14. Heřman, P., Barvík, I.: Non-Markovian effects in the anisotropy of emission in the ring antenna subunits of purple bacteria photosynthetic systems. Czech. J. Phys. **53**, 579–605 (2003)
15. Heřman, P., Barvík, I.: Temperature dependence of the anisotropy of fluorescence in ring molecular systems. J. Lumin. **122–123**, 558–561 (2007)
16. Heřman, P., Zapletal, D., Barvík, I.: Lost of coherence due to disorder in molecular rings. Phys. Stat. Sol. C **6**, 89–92 (2009)
17. Heřman, P., Barvík, I.: Coherence effects in ring molecular systems. Phys. Stat. Sol. C **3**, 3408–3413 (2006)
18. Heřman, P., Barvík, I., Zapletal, D.: Energetic disorder and exciton states of individual molecular rings. J. Lumin. **119–120**, 496–503 (2006)

19. Heřman, P., Zapletal, D., Barvík, I.: The anisotropy of fluorescence in ring units III: tangential versus radial dipole arrangement. J. Lumin. **128**, 768–770 (2008)
20. Heřman, P., Barvík, I., Zapletal, D.: Computer simulation of the anisotropy of fluorescence in ring molecular systems: tangential vs. radial dipole arrangement. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) Computational science—ICCS 2008. LNCS, vol. 5101, pp. 661–670. Springer, Heidelberg (2008)
21. Heřman, P., Zapletal, D., Barvík, I.: Computer simulation of the anisotropy of fluorescence in ring molecular systems: influence of disorder and ellipticity. In: Proceedings of IEEE 12th International Conference on Computational Science and Engineering, pp. 437–442. IEEE Computer Society (2009)
22. Heřman, P., Zapletal, D., Šlégr, J.: Comparison of emission spectra of single LH2 complex for different types of disorder. Phys. Procedia **13**, 14–17 (2011)
23. Heřman, P., Zapletal, D., Horák, M.: Computer simulation of steady state emission and absorption spectra for molecular ring. In: ADVCOMP2011—The Fifth International Conference on Advanced Engineering Computing and Applications in Sciences, pp. 759–762. IARIA (2011)
24. Zapletal, D., Heřman, P.: Simulation of molecular ring emission spectra: localization of exciton states and dynamics. Int. J. Math. Comp. Sim. **6**, 144–152 (2012)
25. Horák, M., Heřman, P., Zapletal, D.: Simulation of molecular ring emission spectra-LH4 complex: localization of exciton states and dynamics. Int. J. Math. Comp. Sim. **7**(1), 85–93 (2013)
26. Heřman, P., Zapletal, D.: Intermolecular coupling fluctuation effect on absorption and emission spectra for LH4 ring. Int. J. Math. Comp. Sim. **7**(3), 249–257 (2013)
27. Horák, M., Heřman, P., Zapletal, D.: Modeling of emission spectra for molecular rings—LH2 and LH4 complexes. Phys. Procedia **44**, 10–18 (2013)
28. Heřman, P., Zapletal, D., Horák, M.: Emission spectra of LH2 complex: full hamiltonian model. Eur. Phys. J. B **86**, Art. number 215 (2013)
29. Heřman, P., Zapletal, D.: Emission spectra of LH4 complex: full Hamiltonian model. Int. J. Math. Comp. Sim. **7**(6), 249–257 (2013)
30. Heřman, P., Zapletal, D.: Simulation of emission spectra for LH4 ring: intermolecular coupling fluctuation effect. Int. J. Math. Comp. Sim. **8**, 73–81 (2014)
31. Mukamel, S.: Principles of nonlinear optical spectroscopy. Oxford University Press, New York (1995)
32. Zhang, W., Chernyak, V., Mukamel, S.: Exciton-migration and three-pulse femtosecond optical spectroscopies of photosynthetic antenna complexes. J. Chem. Phys. **108**(18), 7763–7774 (1998)
33. Novoderezhkin, V.I., Rutkauskas, D., van Grondelle, R.: Dynamics of the emission spectrum of a single LH2 complex: interplay of slow and fast nuclear motions. Biophys. J. **90**, 2890–2902 (2006)
34. Redfield, A.G.: The theory of relaxation processes. Adv. Magn. Reson. **1**, 1–32 (1965)
35. Rutkauskas, D., Novoderezhkin, V., Cogdel, R., van Grondelle, R.: Fluorescence spectral fluctuations of single LH2 complexes from Rhodopseudomonas acidophila strain 10050. Biochemistry **43**(15), 4431–4438 (2004)
36. Rutkauskas, D., Novoderezhkin, V., Cogdel, R., van Grondelle, R.: Fluorescence spectroscopy of conformational changes of single LH2 complexes. Biophys. J. **88**(1), 422–435 (2005)
37. May, V., Kűhn, O.: Charge and energy transfer in molecular systems. Wiley, Berlin (2000)
38. Zerlauskiene, O., Trinkunas, G., Gall, A., Robert, B., Urboniene, V., Valkunas, L.: Static and dynamic protein impact on electronic properties of light-harvesting complex LH2. J. Phys. Chem. B **112**, 15883–15892 (2008)
39. Wolfram, S.: The Mathematica Book, 5th edn. Wolfram Media, Champaign (2003)
40. Zapletal, D., Heřman, P.: Photosynthetic complex LH2—absorption and steady state fluorescence spectra. Energy **77**, 212–219 (2014)

Chapter 12
On the Throughput of the Scheduler for Virtualization of Links

Andrzej Chydzinski

Abstract We deal with the scheduler for virtualization of links. The scheduler switches the service of the physical link between virtual links in constant time intervals, thus providing the isolation of the performance between virtual links. Most important characteristics of this scheduler are the throughput and delay of created virtual links. In this paper we demonstrate how the throughput of a virtual link can be controlled either by the virtual link buffer or the virtual link work phase.

Keywords Scheduler • Queue • Virtualization of links • Buffer size • Work phase

12.1 Introduction

Network virtualization enables coexistence of multiple architectures on common hardware, thus making possible creation of a new, versatile networking paradigm [1, 2]. It has been widely discussed on Future Internet forums like FIA, ETSI and ITU-T and included to proposed Future Internet architectures in several large networking projects, including FIA MANA [3], AKARI (http://akari-project.nict.go.jp/eng/index2.htm), GENI (http://www.geni.net/) and IIP (http://iip.net.pl/).

Network virtualization techniques differ from each other either in the layer of virtualization (which layer is virtualized in the stack), the components of the infrastructure which are virtualized (nodes, links, other resources), the underlying networking technology or in the devices used for virtualization purposes.

In this paper we deal with a link virtualization algorithm, proposed in [4], in which all virtual links are given a constant time of the physical link in a cyclic manner. This algorithm has been used in the IIP System [5], which was built within the IIP project (http://iip.net.pl/), using devices described in [6]. The algorithm has been also analyzed in [7, 8].

A. Chydzinski (✉)
Institute of Informatics, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
e-mail: andrzej.chydzinski@polsl.pl

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_12

In order to make the algorithm more useful in engineering, we have to be able to programm the throughput of each virtual link via the system parameters. In this paper we discuss how the throughput of a virtual link can be programmed using the buffer size of the virtual link or, alternatively, the duration of the service phase of this link in the physical link cycle.

In particular, in Sect. 12.2 the queueing model of the link virtualization scheme is recalled. The analytical solutions of this model are summarized in Sect. 12.3. Then it is demonstrated, how the throughput of a virtual link can be programmed using the buffer size (Sect. 12.4), or the duration of the service phase of this link (Sect. 12.5). The paper is concluded in Sect. 12.7.

12.2 The Queueing Model

Assume there are N independent Poisson streams of packets (customers) arriving to a single physical link (service station). Each arrival stream is characterized by the packet arrival rate, λ_i, and the packet length distribution, D_i. Moreover, a separate buffer for packets is assigned to each arrival stream. The size of the i-th buffer is equal to b_i packets. In these buffers the packets from the arrival streams are queued, waiting for service (transmission through the physical link). If upon a packet arrival the buffer is full, the arriving packet is dropped.

There is also a physical link of capacity C bits/s. All the queues are serviced by the physical link in a cyclic way. Namely, to each queue a predefined constant work time is assigned (work phase); the first one is serviced for W_1 seconds, the second one is serviced for W_2 seconds and so on. After the last, N-th queue, the cycle is repeated.

Obviously, the service time of a single packet (its transmission time) depends on the packet length and equals d_j/C for a packet of length d_j. The following convention is adopted: if the end of phase W_i occurs during the service of a packet, the packet remains in the buffer and its service is repeated from the beginning in the next cycle. This discipline is called the *preemptive repeat-identical (PRI) discipline* (see [9]).

Within each queue the packets are serviced according to the natural discipline, i.e. First-In-First-Out (FIFO).

12.3 Analytical Results

The queueing model presented in the previous section has been recently solved in [8].

In this section we will recall the most important results of [8] without the proofs.

As the scheduler assures the performance isolation between the virtual links, each virtual link can be treated as a separate system with vacations (more on the theory of

such systems can be found in e.g. [10–16], while on their applications in [17–19]). To avoid additional indexes, the results for one virtual will be presented. Of course, the same formulas apply to all virtual links.

One particular virtual link is characterized by the following parameters:

- the length of the work phase (in seconds), denoted as W,
- the length of the vacation phase (in seconds), denoted as V (we have $V = \sum_{i \neq j} W_i$),
- the capacity of the physical link (in bits/s), denoted as C,
- the buffer size (in packets), denoted as b,
- the rate of the Poisson process (in packets), denoted as λ,
- the distribution of packet lengths, denoted as D.

The distribution of packet lengths is characterized by M pairs:

$$D : (d_1, p_1), \ldots, (d_M, p_M), \qquad \sum_{i=1}^{M} p_i = 1, \tag{12.1}$$

where d_i is a packet length, whilst p_i is the probability of length d_i. The mean packet length equals

$$\overline{d} = \sum_{i=1}^{M} d_i \, p_i .$$

It is assumed that the service phase is no shorter than the transmission time of the largest possible packet, i.e.

$$\max\{d_i / C : i = 1, \ldots, M\} \leq W.$$

The offered load of the virtual link is defined as:

$$\rho = \lambda \, \overline{d} \, \frac{V + W}{CW}.$$

By $X(t)$ the queue length at time t will be denoted.

12.3.1 Queue Length

In [8] is was shown that the stationary distribution of the queue length,

$$\lim_{t \to \infty} \mathbf{P}\{X(t) = m\},$$

does not exit. Instead, we can compute the distribution of the queue length at the beginning of the vacation phase, i.e.:

$$\overline{q}_n = \lim_{k\to\infty} \mathbf{P}\{X(\alpha_k) = n\}, \quad n = 0, \ldots, b, \tag{12.2}$$

where

$$\alpha_k = (k-1)(V+W), \quad k \geq 1.$$

Set

$$X_k = X(\alpha_k), \quad S_k = S(\alpha_k), \quad k \geq 1,$$

where $S(t)$ is the number of the length of the packet serviced at time t. Namely, S_k denotes the number of the length of the packet that could not be serviced in the k-th work phase, due to forthcoming end of the phase. (We adopt the convention that $S_k = 0$ means that after the k-th work phase there was no packets in the buffer.)

The pair (X_k, S_k) is a two-dimensional Markov chain in space:

$$\Omega = \{(m, j) : m = 1, \ldots, b, \;\; j = 1, \ldots, M\} \cup (0, 0).$$

In [8] it was proven that the transition probabilities for (X_k, S_k) are the following:

$$Q_{n,i,m,j} = \begin{cases} \sum_{l=0}^{b} U_{0,0,l}\, \Theta_l(W, m, j), \\ \qquad\qquad \text{if} \;\; n = 0, i = 0, \\ \sum_{l=n}^{b} U_{n,i,l}\, \Theta_{l-1}(W - d_i/C, m, j), \\ \qquad\qquad \text{if} \;\; n > 0, i > 0, \end{cases} \tag{12.3}$$

where

$$Q_{n,i,m,j} = \mathbf{P}\{X_{k+1} = m, S_{k+1} = j \,|\, X_k = n, S_k = i\},$$

$$U_{n,i,l} = \begin{cases} \frac{e^{-\lambda V}(\lambda V)^l}{l!}, \quad \text{if} \;\; n = 0, i = 0, n \leq l < b, \\ \sum_{j=b}^{\infty} \frac{e^{-\lambda V}(\lambda V)^j}{j!}, \;\; \text{if} \;\; n = 0, i = 0, l = b, \\ \frac{e^{-\lambda(V+d_i/C)}[\lambda(V+d_i/C)]^{l-n}}{(l-n)!}, \\ \qquad\qquad \text{if} \;\; n > 0, i > 0, n \leq l < b, \\ \sum_{j=b-n}^{\infty} \frac{e^{-\lambda(V+d_i/C)}[\lambda(V+d_i/C)]^j}{j!}, \\ \qquad\qquad \text{if} \;\; n > 0, i > 0, l = b, \\ 0, \text{otherwise}. \end{cases} \tag{12.4}$$

and $\Theta_n(t, m, j)$ is the probability that in the classic $M/G/1/b$ queueing model (without vacations) at time t the queue length is m and a packet of length d_j is being

transmitted. $\Theta_n(t, m, j)$ can be calculated using the Laplace transform method—detailed calculation can be found in [8]. (For more examples of the method used, the reader is referred to [20–22].)

Using (12.3) we can calculate the stationary distribution of the chain (X_k, S_k), i.e.:

$$q_{n,i} = \lim_{k\to\infty} \mathbf{P}\{X_k = n, S_k = i\}, \quad (n, i) \in \Omega, \tag{12.5}$$

by exploiting the set of equations in the form:

$$\sum_{n=1}^{b}\sum_{i=1}^{M} q_{n,i} Q_{n,i,m,j} + q_{0,0} Q_{0,0,m,j} = q_{m,j}, \tag{12.6}$$

for $1 \le m \le b$, $1 \le j \le M$, and

$$\sum_{n=1}^{b}\sum_{i=1}^{M} q_{n,i} + q_{0,0} = 1. \tag{12.7}$$

By means of distribution $q_{n,i}$ we then get the main result:

$$\overline{q}_n = \begin{cases} q_{0,0}, & \text{if} \quad n = 0, \\ \sum_{i=1}^{M} q_{n,i}, & \text{if} \quad 0 < n \le b, \end{cases} \tag{12.8}$$

with the mean:

$$\mathbf{E}(\overline{q}) = \sum_{n=0}^{b} n\overline{q}_n. \tag{12.9}$$

12.3.2 Throughput

Firstly, we define the loss ratio, *LR*, of a virtual link as the long-run fraction of packets lost due to the buffer overflow. Then the throughput, γ, of the link can be defined as the percentage of the input traffic that is carried by the virtual link, namely:

$$\gamma = (1 - LR) \cdot 100\,\%.$$

As the probability that an arriving packet is dropped does not depend on the length of this packet, the loss ratio calculated for packets is equal to the loss ratio calculated

for bytes. The same applies to throughput—it does not matter whether it is measured in packets or bytes.

The formula for the loss ratio of a virtual link was proven in [8]. It has the following form:

$$LR = \frac{1}{\lambda(V+W)}\Big[\sum_{n=1}^{b}\sum_{i=1}^{M}\sum_{j=0}^{\infty}\sum_{m=0}^{b} q_{n,i}\,Y_{n,i,m}(j)\big(j+\Delta_m(W-d_i/C)\big) + \sum_{j=0}^{\infty}\sum_{m=0}^{b} q_{0,0}\,Y_{0,0,m}(j)\big(j+\Delta_m(W)\big)\Big], \tag{12.10}$$

where

$$Y_{n,i,m}(j) = \begin{cases} \frac{e^{-\lambda V}(\lambda V)^m}{m!}, & \text{if } n=0, i=0, 0\le m\le b, j=0,\\ \frac{e^{-\lambda V}(\lambda V)^{b+j}}{(b+j)!}, & \text{if } n=0, i=0, m=b, j>0,\\ \frac{e^{-\lambda(V+d_i/C)}[\lambda(V+d_i/C)]^{m-n+1}}{(m-n+1)!}, & \text{if } n>0, i>0, n-1\le m\le b-1, j=0,\\ \frac{e^{-\lambda(V+d_i/C)}[\lambda(V+d_i/C)]^{b-n+j}}{(b-n+j)!}, & \text{if } n>0, i>0, m=b-1, j>0,\\ 0, & \text{otherwise.}\end{cases} \tag{12.11}$$

and $\Delta_n(t)$ is the average number of losses in interval $(0,t]$ in the classic $M/G/1/b$ queueing model (without vacations). The formula for $\Delta_n(t)$ with the proof can be found in [8].

12.4 Controlling the Throughput Using the Buffer Size

In [8] several examples of calculations of the queue length distributions and loss ratios for predefined virtual link parameters can be found. In this paper, the problem is reversed: we search for buffers or work phases that provide some predefined loss ratios (throughputs).

The simplest way to solve this problem is probing the space of buffer sizes or work phases (or both) for a satisfactory solution.

In this section we will demonstrate this using the buffer sizes.

For example, consider the following situation with three virtual links created on a physical link (see also [8]): the work phases of all virtual link are the same ($W_1 = W_2 = W_3$), the physical link capacity is equal to

$$C = 1\,Gb/s.$$

Moreover, the packet length distributions are different in each arrival stream. To link 1 only 256-bytes-long packets arrive, to link 2 only 512-bytes-long packets

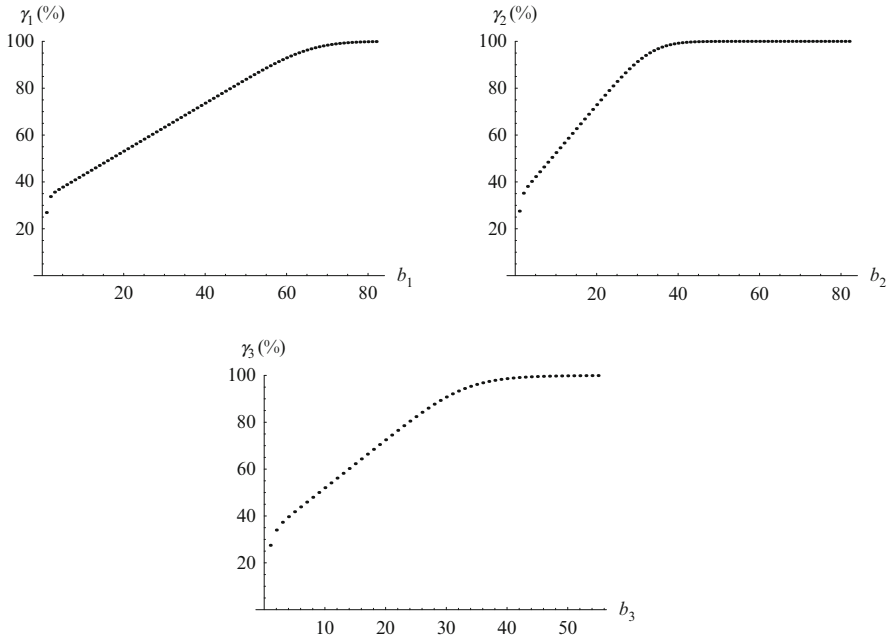


Fig. 12.1 The throughput of links 1, 2 and 3 versus the buffer size. $\rho = 0.80$, $W_1 = W_2 = W_3 = 250\,\mu s$

arrive. Finally, to link 3 the packets of length 40 bytes, 512 bytes and 1,500 bytes arrive, following the distribution:

$$D: \quad d_1 = 40B, \;\; d_2 = 512B, \;\; d_3 = 1,500B,$$
$$p_1 = 0.494, \;\; p_2 = 0.270, \;\; p_3 = 0.236. \tag{12.12}$$

Therefore, the mean packet length of the third link is

$$\overline{d} = 512B.$$

We assume that the work phases are $W_1 = W_2 = W_3 = 250\,\mu s$.

In Fig. 12.1, the dependence of the throughput on the buffer size for links 1, 2 and 3, is presented assuming the load offered to each virtual link of 0.80 (equivalent to average input bitrate of 266.66 Mb/s on each link). As we can see, all the curves approach 100 % for the buffer sizes over some threshold (e.g. about 70 in the case of link 1). This is a consequence of a moderate load and its not a case for higher load values.

As the curves are strictly increasing, we can check the buffer sizes for a minimal buffer size which provides a predefined throughput.

Table 12.1 Buffer sizes guaranteeing the predefined (or higher) throughput

	Predefined throughput					
Link	99.9 %	99.8 %	99.5 %	99 %	98 %	95 %
Link 1, 256B	82	80	76	73	69	63
Link 2, 512B	46	44	42	40	37	33
Link 3, distr. D	54	50	45	42	39	34

$\rho = 0.80$, $W_1 = W_2 = W_3 = 250\,\mu s$

Table 12.2 The throughput and the average queue length for three virtual links

Link	Throughput	Avg. queue length	Stddev. queue length
Link 1, 256B	99.12 %	0.315	0.573
Link 2, 512B	99.23 %	0.353	0.683
Link 3, dist. D	99.10 %	1.174	2.663

$\rho = 0.80$, $b_1 = 73$, $b_2 = 40$, $b_3 = 42$, $W_1 = W_2 = W_3 = 250\,\mu s$

For instance, let us assume that we are searching for the buffer sizes which guarantee the throughputs of 99.9, 99.8, 99.5, 99, 98 and 95 %.

Checking the resulting throughputs for buffer sizes in interval [30,90] we obtain the results gathered in Table 12.1. Namely, Table 12.1 presents the minimal buffer sizes that guarantee the predefined throughput. For instance, in order to have the throughput no smaller than 99 % in all three virtual links, we have to provide the following buffers: $b_1 \geq 73$, $b_2 \geq 40$, $b_3 \geq 42$.

Detailed results for these minimal buffer sizes are presented in Table 12.2. The exact throughputs are slightly higher than 99 %, which is to be expected—the buffer sizes smaller by 1 give values slightly smaller than 99 %.

12.5 Controlling the Throughput Using the Work Phase Duration

Similarly, we can obtain a predefined throughput manipulating the work phase. However, this is different than manipulating the buffer size, because changing the work phase of one virtual link we change also the whole cycle, thus influencing the performance of other virtual links in an uncontrolled way. (This does not happen when we change the buffer size.)

Moreover, the dependence of the throughput on the work phase duration can be very irregular. For instance, in Figs. 12.2, 12.3 and 12.4, the dependence of the throughput of virtual links 1, 2 and 3 on the work phase for $\rho = 1$ is depicted, respectively. Even in the case of simple distribution of the packet size, like in Figs. 12.2 and 12.3, the curve is highly variable and non-monotonic. It has an "increasing" part for short work phases, a "flat" part, and a "decreasing"

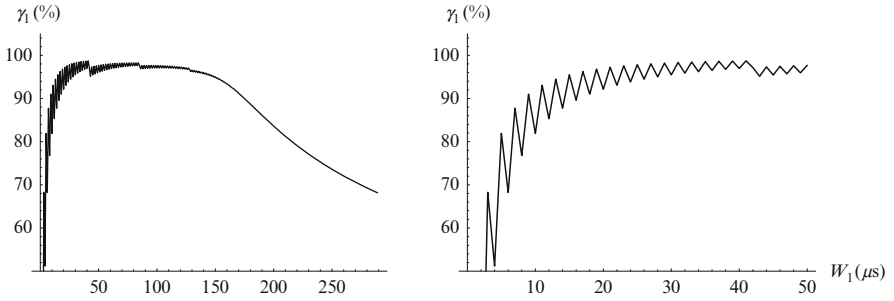


Fig. 12.2 The throughput of link 1 versus its work phase duration (in microseconds). $\rho = 1$, $b_1 = 50$, $W_1 = W_2 = W_3$

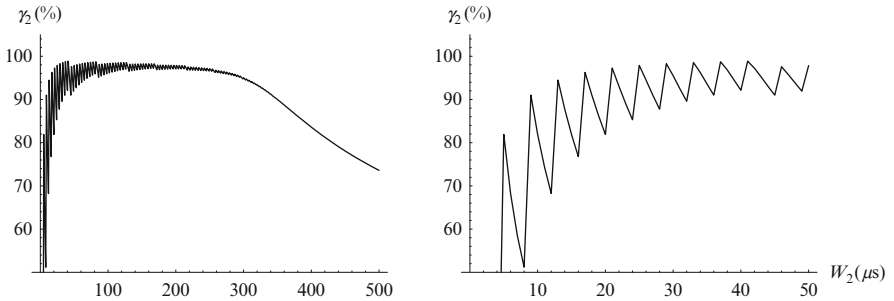


Fig. 12.3 The throughput of link 2 versus its work phase duration (in microseconds). $\rho = 1$, $b_2 = 50$, $W_1 = W_2 = W_3$

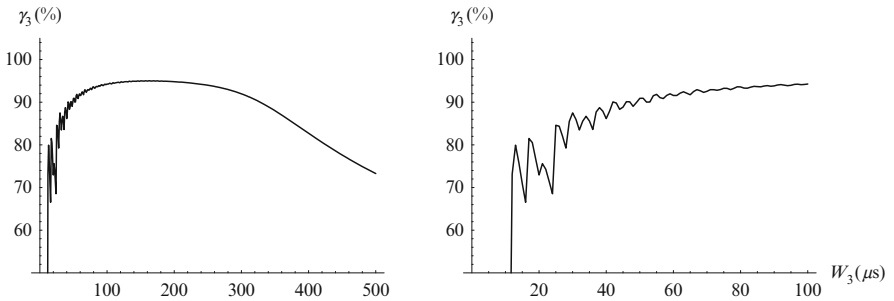


Fig. 12.4 The throughput of link 3 versus its work phase duration (in microseconds). $\rho = 1$, $b_1 = 50$, $W_1 = W_2 = W_3$

part for long work phases. Moreover, the general shape is disturbed by a high-scale sawtooth-like irregularities, a middle-scale irregularities (see Fig. 12.2 around buffers sizes 45 and 90). This gets even more complicated, when the packet sizes are distributed, as in Fig. 12.4—now the sawtooth have irregular heights. Moreover, the maximum possible throughput is below 100 % and barely reaches 95 %. This is connected with the the bandwidth loss at the end of the work phase—large packets are blocked because there is not enough time to transmit them within the current work phase and the physical link is idle.

Nevertheless, obtaining a needed throughput by manipulating the phases is possible, but more tricky than in the previous section. It is also easier for lower load values.

To demonstrate this, let us assume again that we are searching for the work phases which guarantee the throughputs of 99.9, 99.8, 99.5, 99, 98 and 95 % for load of 0.80.

We do not manipulate the buffer sizes now—they are set as follows. For the link operating on 256-bytes-long packets the buffer is set to 100 packets, while the other buffers are set to 50 packets. In this way the buffer sizes are similar, if measured in bytes.

Probing the throughputs for the work phases in range [220 μs, 420 μs] we can obtain Table 12.3, in which the work phases that provide the predefined throughputs are presented. For instance, if we want to have the throughput of link 1 equal to 99.5 %, we have to set $W = 341\,\mu s$ and $V = 682\,\mu s$, i.e. $W_1 = W_2 = W_3 = 341\,\mu s$.

Unfortunately, due to the reasons explained above, such parameterization gives unspecified throughputs of other virtual links. Therefore, each entry in Table 12.3 should be treated as a separate parameterization of the scheduler, which guarantees the throughput of one virtual link only.

However, using the structure of Table 12.3, we can manipulate (up to some extend) the length of the work phases in such a way, that all three throughputs are over the assumed threshold. For instance, from Table 12.3 we may expect that for $W_1 = W_2 = W_3 = 293\,\mu s$, not only the third throughput is 99.5 %, but also the first and the second are over 99.5 %. Detailed results for three virtual links working together, which are presented in Table 12.4, confirm this supposition.

Table 12.3 The work phase duration, W, guaranteeing the predefined (or higher) throughput (μs)

	Predefined throughput					
Link type	99.9 %	99.8 %	99.5 %	99 %	98 %	95 %
256B packets, $b = 100$	316	325	341	355	373	405
512B packets, $b = 50$	280	292	313	331	353	393
D distributed packets, $b = 50$	221	254	293	319	346	389

$\rho = 0.80$, $V = 2W$

Table 12.4 The throughput and the average queue length for three virtual links

Link	Throughput	Avg. queue length	Stddev. queue length
Link 1, $256B$	99.985 %	0.327	0.629
Link 2, $512B$	99.79 %	0.387	0.803
Link 3, dist. D	99.5 %	1.132	2.703

$\rho = 0.80$, $b_1 = 100$, $b_2 = 50$, $b_3 = 50$, $W_1 = W_2 = W_3 = 293\,\mu s$

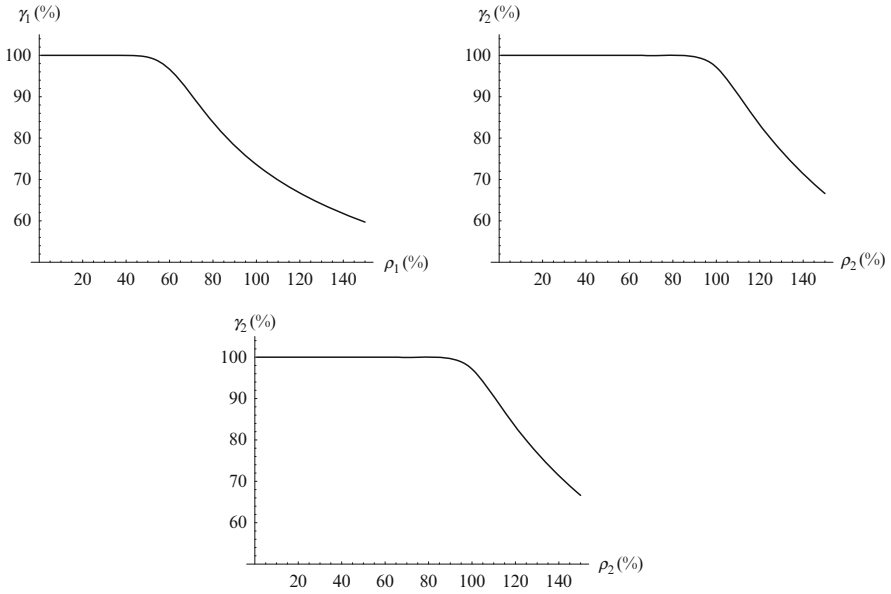


Fig. 12.5 The throughput of links 1, 2 and 3 versus the offered load (in percent). $b_1 = b_2 = b_3 = 50$, $W_1 = W_2 = W_3 = 250\,\mu s$

It must be stressed that the dependence of the throughput on the phase duration may be highly variable and non-monotonic, especially for short phases. In other words, a longer work phase does not necessary mean a higher throughput. Thus each parameterization, that we expect to produce satisfactory results, should be carefully checked, as it was done in Table 12.4.

12.6 Throughput Versus the Offered Load

Finally, it is worth recalling that the throughput of a virtual link depends on the load offered to this link. The dependence of the throughput on the offered load for links 1, 2 and 3 and $W_1 = W_2 = W_3 = 250\,\mu s$ is depicted in Fig. 12.5.

Table 12.5 The throughput and the average queue size for three virtual links and $\rho = 0.80$

Link	Throughput	Avg. queue length	Stddev. queue length
Link 1, $256B$	99.85 %	1.280	2.083
Link 2, $512B$	99.77 %	2.965	3.554
Link 3, distr. D	95.69 %	6.614	5.687

$b_1 = b_2 = b_3 = 20$, $W_1 = W_2 = W_3 = 32\,\mu$s

Table 12.6 The throughput and the average queue size for three virtual links and $\rho = 0.50$

Link	Throughput	Avg. queue length	Stddev. queue length
Link 1, $256B$	99.99996 %	0.194	0.462
Link 2, $512B$	>99.99999 %	0.273	0.625
Link 3, D	99.99509 %	0.778	1.629

$b_1 = b_2 = b_3 = 20$, $W_1 = W_2 = W_3 = 32\,\mu$s

In some cases, where the buffers are small and the work phases are short, we may use the load to control the throughput. This is demonstrated in the following example. Assume that the buffer sizes are $b_1 = b_2 = b_3 = 20$ and the work phases are $W_1 = W_2 = W_3 = 32\,\mu$s. The results for $\rho = 0.80$ are presented in Table 12.5. As we can see, the throughput of the scheduler is quite low–it is below 96 % in link 3. Now assume that and the buffer sizes and the work phases cannot be manipulated for some reasons. Therefore the only way to improve the throughput is to decrease the load of virtual links. In Table 12.6 the sample results (for the load reduced to 0.50) are presented. As we can see, the throughput is at least 99.99509 % now.

12.7 Conclusions

In the paper we dealt with a scheduler for creating several virtual links on a physical link by assigning constant service time for each virtual link in a cyclic manner. It was demonstrated how the throughput of a virtual link can be set either using the buffer size of the virtual link or the length of the work phase of this link.

Acknowledgements This is an extended version of the paper [23] presented during International Conference on Applications of Computer Engineering in Lisbon, October 2014.

References

1. Mosharaf Kabir Chowdhury, N.M., Boutaba, R.: A survey of network virtualization. Comput. Netw. **54**(5), 862–876 (2010)
2. Anderson, T., Peterson, L., Shenker, S., Turner, J.: Overcoming the internet impasse through virtualization. Computer **38**(4), 34–41 (2005)
3. Galis, A., et al.: Management and Service-Aware Networking Architectures (MANA) for Future Internet. System Functions, Capabilities and Requirements, Position Paper, Version V6.0, 3 May 2009
4. Burakowski, W., Tarasiuk, H., Beben, A., Goralski, W., Wisniewski, P.: Ideal device supporting virtualization of network infrastructure in system IIP (in Polish). In: Proceedings of KSTiT '11, Lodz, 14–16 September 2011, pp. 818–823
5. Burakowski, W., Tarasiuk, H., Beben, A., Danilewicz, G.: Virtualized network infrastructure supporting co-existence of parallel internets. In: Proceedings of SNPD '12, Kyoto, 8–10 August 2012, pp. 679–684
6. Chydzinski, A., Rawski, M., Wisniewski, P., Adamczyk, B., Olszewski, I., Szotkowski, P., Chrost, L., Tomaszewicz, P., Parniewicz, D.: Virtualization devices for prototyping of Future Internet. In: Proceedings of SNPD '12, Kyoto, 8–10 August 2012, pp. 672–678
7. Sosnowski, M., Burakowski, W.: Analysis of the system with vacations under Poissonian input stream and constant service times. In: Proceedings of Polish Teletraffic Symposium, Zakopane, 6–7 December 2012, pp. 9–13
8. Chydzinski, A., Adamczyk, B.: Analysis of a scheduler for virtualization of links with performance isolation. Appl. Math. Inf. Sci. **8**(6), 2653–2666 (2014)
9. Katayama, T.: Waiting time analysis for a queueing system with time-limited service and exponential timer. Nav. Res. Logist. **48**(7), 638–651 (2001)
10. Takagi, H.: Queueing Analysis - Vacation and Priority Systems. North-Holland, Amsterdam (1991)
11. Tian, N., Zhang, Z.G.: Vacation Queueing Models - Theory and Applications. Springer, New York (2006)
12. Doshi, B.T.: Queueing systems with vacation: A survey. Queueing Syst. **1**, 29–66 (1986)
13. Ke, J.C., Wu, C.H., Zhang, Z.G.: Recent developments in vacations queueing models: A short survey. Int. J. Oper. Res. **7**(4), 3–8 (2010)
14. Hur, S., Ahn, S.: Batch arrival queues with vacations and server setup. Appl. Math. Model. **29**(12), 1164–1181 (2005)
15. Gupta, U.C., Sikdar, K.: Computing queue length distributions in MAP/G/1/N queue under single and multiple vacation. Appl. Math. Comput. **174**, 1498–1525 (2006)
16. Wu, J., Liu, Z., Peng, Y.: On the BMAP/G/1 G-queues with second optional service and multiple vacations. Appl. Math. Model. **33**(12), 4314–4325 (2009)
17. Chung, S.-P., Chen, V.: Performance of power efficient wake-up mechanisms for mobile multimedia communication with bursty traffic. In: Proceedings of the 5th WSEAS International Conference on Data Networks, Communications & Computers, Bucharest, pp. 51–56 (2006)
18. Ho, J.-H.: A carrier fragmentation aware CSMA/ID MAC protocol for IP over WDM ring networks. WSEAS Trans. Commun. **4**(9), 271–280 (2010)
19. Liu, D., Xu, G., Mastorakis, N.E.: Reliability analysis of a deteriorating system with delayed vacation of repairman. WSEAS Trans. Syst. **12**(10), 413–424 (2011)
20. Chydzinski, A.: Duration of the buffer overflow period in a batch arrival queue. Perform. Eval. **63**(4–5), 493–508 (2006)
21. Chydzinski, A.: Transient analysis of the MMPP/G/1/K queue. Telecommun. Syst. **32**(4), 247–262 (2006)
22. Chydzinski, A., Chrost, L.: Analysis of AQM queues with queue-size based packet dropping. Int. J. Appl. Math. Comput. Sci. **21**(3), 567–577 (2011)
23. Chydzinski, A.: Controlling the throughput of virtual links with performance isolation. In: Proceedings on International Conference on Applications of Computer Engineering (ACE '14), Lisbon, October 2014, pp. 13–18

Chapter 13
A Simulation Study on Generalized Pareto Mixture Model

Mustafa Cavus, Ahmet Sezer, and Berna Yazici

Abstract The Generalized Pareto Distribution is commonly used for extreme value problems. Especially, the values which exceed the finite threshold, is the focus in extreme value problems like in insurance sector. The Generalized Pareto Distribution is well approach for modeling the samples which include these extreme values. In the real life, samples are heterogeneous. In such cases, the mixture models are better way for modeling the data. In this study, we generate random samples from the Generalized Pareto Mixture Distribution for modeling of heterogeneous data. For this purpose, we use two different Generalized Pareto Distribution as components of the Generalized Pareto Mixture Distribution. For generating random samples, The Inverse Transformation Method is used in the simulation study. The parameters of the mixture models are shape, scale and location are fixed. After generating random samples, Chi-Square Goodness-of-Fit Test is used for checking whether the generated samples are distributed based on the Generalized Pareto Distribution. R-Statistical Programming Language is used in simulation study.

Keywords The Generalized Pareto Mixture Distribution • Mixture models • The Inverse Transformation Method • Chi-Square Goodness-of-Fit Test • Generating random samples • Pareto Distribution

13.1 Motivation

The aim of this study is that generating random samples from Generalized Pareto Mixture Models safely. Because, in the applications of extreme value problems Generalized Pareto Distribution is used commonly. These applications usually intent with heterogeneous data. For modeling the heterogeneous data, mixture models is very important solution. Also constructing the mixture models can be handle safely. For this aim, in this study is focused on the generating random samples from Generalized Pareto Mixture Models for modeling the heterogeneous data.

M. Cavus (✉) • A. Sezer • B. Yazici
Department of Statistics, Anadolu University, Eskisehir 26470, Turkey
e-mail: mustafacavus@anadolu.edu.tr

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_13

The inverse transform method is chosen for the application part of the study. Lastly the important point of this study is that the show the goodness-of-fit test for generated random samples.

13.2 Introduction

The Generalized Pareto Distribution (GPD) was introduced by Pickands [1]. Then further studied was worked by Davison [2], Castillo [3, 4]. Mierlus-Mazilu studied on generalized pareto distribution, especially on generating random samples from it by the inverse transformation method [5]. However, he did not use any goodness-of-fit test for the generated random samples.

The generalized pareto distribution is very important tool for modeling of economical, financial and insurance data [9−11]. The companies are especially work on these areas want to plan their financial parameters. For instance, the financial crisis or same cases which are unexpected events can be damage to the company. Thus, to isolate the damaged case companies should model the extreme events with the statistical distributions. By this way, companies can predict the extreme and damaged events for their financial balance.

In this study, we focused on how to generate random samples from the generalized pareto mixture model and especially we test whether they fit well the generalized pareto mixture model by chi-squared goodness-of-fit test. The generalized pareto mixture model is important tool when the data has heterogeneous distribution. In real life, the risks in the financial sector can be heterogeneous dispersion so the generalized pareto mixture model is used on these areas.

By now, Chi-Square goodness-of-fit test was not used in studies which are related generating random samples from mixture models. For example, Beirlant used the Jackson statistics as a goodness-of-fit test for the generated random samples from Pareto-type behavior [6]. In the study, it is claimed that the log-transformed Pareto random variables are exponentially distributed and Jackson statistics, originally proposed as a goodness-of-fit statistics for testing exponentially.

13.3 Generalized Pareto Mixture Model

13.3.1 Pareto Distribution

Pareto Distribution is generally used for modeling of the data which is consists of income. It was proposed by an Italian economist and sociologist Vilfredo Frederico Damaso Pareto (1897). This distribution is constructed on Pareto Principle. Based on Pareto Principle, a large portion of wealth of many societies is owned by a smaller

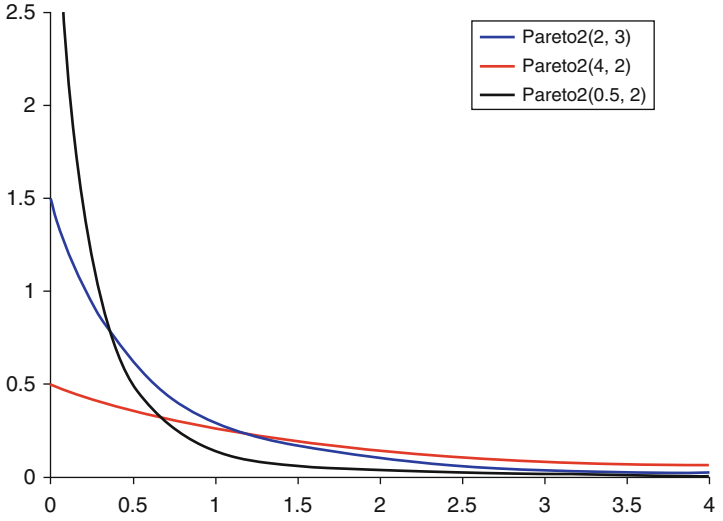


Fig. 13.1 Probability density function of pareto distribution with different parameters

percentage of the people in that society. This principle is explained more simply as 80-20 rule which says that 20 % of the population owns 80 % of the wealth [7].

Pareto Distribution is also useful for modeling of finance, actuary and economics. It can be used many situations in which an equilibrium is found in the distribution of the small to the large.

The probability density function of Pareto Distribution:

$$f(x) = \frac{\alpha\beta^{\alpha}}{x^{a+1}}, \quad 0 < \beta \leq x, \ \ \alpha > 0 \tag{13.1}$$

The Pareto Distribution is characterized by α is shape parameter which is positive and β is scale parameter which measures the heaviness in the upper tail as can be seen Fig. 13.1. Mostlyused notation of Pareto Distribution is P(β, α) with shape parameter β and scale parameter α.

13.3.2 Generalized Pareto Distribution

Generalized Pareto Distribution is one of the most preferred method for modeling of Extreme Value problems. Extreme Value Theory interests in the values of the function which exceed the finite threshold. Especially in actuarial sector, the payments of companies consist regular policies which are in expected limits. If the unexpected events are occurred, the payments of companies can be rise up

extremely. Thus the insurance companies need to model of rare events which are very important on payments.

Let X is a random variable of F distribution. The interested values are exceed the finite threshold of this distribution. Generalized Pareto Distribution applies in modeling of this values which are named rare events. These values constructs of the right tail of the distribution line.

The Generalized Pareto Distribution was introduced by Pickands [1] and has been further studied by Davison [2], Castillo [3, 4]. It has a wide application area in economics, insurance and finance.

The probability density function of Generalized Pareto Distribution:

$$p(x) = \frac{1}{\beta}\left(1 + \frac{\alpha\,(x-\mu)}{\beta}\right)^{\frac{-1}{\alpha}-1} \tag{13.2}$$

The Generalized Pareto Distribution is characterized by α is shape parameter which domain is negative infinity to positive infinity. β is scale parameter which measures the heaviness in the upper tail. μ is location parameter can be explains as the threshold.

13.3.3 Mixture Models

In statistics, there are many available method for modeling the data. These methods are named as statistical distributions. Statistical distributions are classifying to discrete probability distribution and continuous probability distributions. Data analysis used these statistical distributions which is appropriate to data's behavior. All analysis want to describe the data is concerned optimally. Occasionally, some data's can be distributed in different clusters that as seen on the distribution graph of the data. This difference also can be named as heterogeneity on the distribution. In such cases there is a new approach of modeling these data is known as the mixture models in literature. The Mixture Models provide a natural representation of heterogeneity in a number of clusters.

Mixture Models provide a method of describing more complex probability distributions which is mentioned earlier as clusters, by combining several probability distributions. The combining probability distributions can be same distributions with different parameters and also different distributions. These distributions are combined with mixture weights. The notation of the probability distribution of the mixture models:

$$P\left(x\middle|\theta_1,\theta_2,\ldots,\theta_k\right) = P\left(x\middle|\theta_i\right) = \sum_{i=1}^{k} w_i\, P\left(x_i \;\middle|\; \theta_i\right) \tag{13.3}$$

The Mixture Models are weighted with w_i parameters are defined the weights of each component distributions. θ_i is defined as the parameter of the component distribution. In the formula, there is only a parameter, it can be differ according to the component distributions.

The distribution can be constructed by combining of two or more probability density function. Defining of the component distributions is related to data analyst and his own experience. Plot of the data is very useful tool for defining of the component distribution.

The advantages of mixture models can be explained with four main titles. First, component distributions can be multimodal. Second, the mixture model cover the data well to standard models. Third, it includes well-studied statistical inference techniques available. Last one is flexibility in choosing the component distributions.

13.3.4 Generalized Pareto Mixture Model

Generalized Pareto Mixture Model (GPMM) is a parametric probability density function represented as a weighted sum of Generalized Pareto component densities. It is a weighted sum of k component Generalized Pareto densities as given by the equation:

$$P\left(x \,\middle|\, \mu_1, \mu_2, \ldots, \mu_k; \alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k\right) = P\left(x \,\middle|\, \mu_i, \alpha_i, \beta_k\right) = \sum_{i=1}^{k} w_i \, P\left(x_i \middle| \mu_i, \alpha_i, \beta_i\right) \tag{13.4}$$

where x is a data vector, $w_i = 1, 2, \ldots, k$, are the mixture weights and $P\left(x_i \middle| \mu_i, \alpha_i, \beta_k\right)$, i = 1, 2, ... , k are the component Generalized Pareto densities. Each component density is a Generalized Pareto probability density function of the form:

$$P\left(x_i \middle| \mu_i, \alpha_i, \beta_i\right) = \frac{1}{\beta_i}\left(1 + \frac{\alpha_i \left(x - \mu_i\right)}{\beta_i}\right)^{\frac{-1}{\alpha_i} - 1} \tag{13.5}$$

where μ_i are location parameters, α_i are shape parameter and β_i are scale parameter of the Generalized Pareto probability density function. The mixture weights satisfy the constraint that $\sum_{i=1}^{k} w_i = 1$.

13.4 Simulation Study

13.4.1 Inverse Transformation Method

Generating random samples is one of the most important subject in simulation studies. Researchers often refer generating random samples for their studies. Simulation study is not a proof for the problems but it can be preferable way for demonstrating some facts.

There are several methods are used for generating random samples. Purpose of applying these methods is to generate random samples from an arbitrary distribution. Most of them are quite complex and requiring computer support. Some of them can be easier for researchers. One of the simple methods is the inverse transformation.

The Inverse Transformation Method applies with the uniform distribution. After the calculation of cumulative probability function of intended distribution, random samples can be generated with the inverse transformation of random samples which are generated from uniform distribution.

Let X be a random variable with cumulative distribution function is F. F is a nondecreasing function, the inverse function F^{-1} may be defined as:

$$F^{-1} = inf\,\{x : F(x) \geq y\}, \quad 0 \leq y \leq 1 \tag{13.6}$$

Let $u \sim Uniform\,(0, 1)$. The cumulative distribution function of the inverse transform $F^{-1}(u)$ is given by

$$P\left(F^{-1}(u) \leq x\right) = P\,(u \leq F(x)) = F(x) \tag{13.7}$$

Thus, to generate a random variable X with cumulative distribution F, draw $u \sim Uniform\,(0, 1)$ and set $X = F^{-1}(u)$. This leads to the general method for generating random samples from an arbitrary cumulative probability distribution F [8].

Algorithm:

1. Generate $u \sim Uniform\,(0, 1)$.
2. Set $X = F^{-1}(u)$.

Requirements for applying the inverse transformation method:

1. The cumulative probability function of intended distribution F must be nondecreasing.
2. The inverse of the cumulative probability function of intended distribution F^{-1} must be found analytically.

13.4.2 Chi-Square Goodness-of-Fit Test

In the simulation study, generated random samples must be check by a goodness-of-fit test be certain of belong the indented distribution. For this purpose, chi-square goodness-of-fit test is used in R.

Chi-square goodness-of-fit test calculate a test statistic from the differences between observed frequencies and expected frequencies of theoretical distribution. The cumulative probability function of the theoretical distribution is used for calculating the expected frequencies. The test statistics is calculated below:

$$\chi^2 = \sum_{i=1}^{k} \frac{(Obs.F_{.i} - Exp.F_{.i})^2}{Exp.F_{.i}} \tag{13.8}$$

The result is calculated from the formula is named as chi-square test statistics. This is compared with chi-square table value and then the conclusion is defined about the hypotheses.

In this study, every sample is divided ten groups. Then the observed frequencies of these groups are compared with expected frequencies. Ten group is used because it is optimal for many sample size. It is decided after the control of the simulation study.

13.4.3 Numerical Results

In the application part of this study, the random samples are generated from GPMM by the inverse transformation method. For this, R is used. In R code, the three-parameters of the distribution are fixed when the random samples are generated. The mixture weights are used as 0.3 and 0.7.

Algorithm:

1. Generate

$$u \sim Uniform\,(0, 1)$$

2. If $u \leq w_1$ then set $X = F_1^{-1}(u)$
   Else (or $u > w_1$) then set $X = F_2^{-1}(u)$.

According to the application:

1. Generate $u \sim Uniform\,(0, 1)$
2. If $u \leq 0.3$ (or $u > 0.7$) then set $X = F_1^{-1}(u) = \mu_1 + \frac{\beta_1\left[(1-u)^{-\alpha_1}-1\right]}{\alpha_1}$
   Else $u > 0.7$ (or $u > 0.3$) then set $X = F_2^{-1}(u) = \mu_2 + \frac{\beta_2\left[(1-u)^{-\alpha_2}-1\right]}{\alpha_2}$.

By these fixed parameters and values, the algorithm is repeated 1,000 times in R with different sample sizes and the appropriate samples which have appropriate distribution are detected according to the level of significance 0.05. The success ratio of the results are showed as (successful trials/all trials) in Table 13.1. Each cell shows the success rate of the generated random samples which are passed the goodness-of-fit test successfully.

As you seen on the table of the results, if the sample size is 100, the generated random samples distribute generalized pareto with three-parameters appropriately more than 90 % success with significance level is 0.05. If the sample size is increased, the success of the generator is be higher.

After the seeing accordance of the generator we can generate the random samples from the GPMM. According to the algorithm, the generalized pareto distributions which are shown in Table 13.1 third and fourth distributions are mixtured with mixture weights are 0.3 and 0.7 is used for generating random samples. Then generated random samples are shown in Fig. 13.2. In Fig. 13.3, first and second distributions are mixtured with mixture weights are 0.3 and 0.7 is used for generating random samples and the result is shown.

13.4.4 Performance of the Generator

In Section 4.3 of the simulation study, we can use the inverse transform method for generating random samples. Also, we can generate random samples with different parameters. In this section, we examine the effect of changes in parameters on the success rate of the generator. For this, the success rates values are calculated with different parameters. Results are compared with the different combinations.

Table 13.1 The results of chi-square test

Shape Scale Location	n = 10	n = 20	n = 50	n = 100
5 2 1	0.673	0.785	0.845	0.9
3 1.3 0.5	0.771	0.849	0.889	0.91
14 4 2	0.299	0.619	0.779	0.842
4 5 0.9	0.683	0.813	0.872	0.906

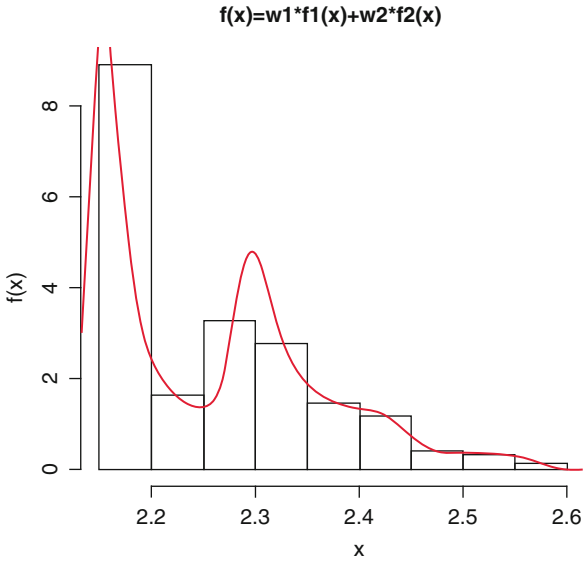


Fig. 13.2 Distribution function of mixtured GPD 3 and 4 in Table 1

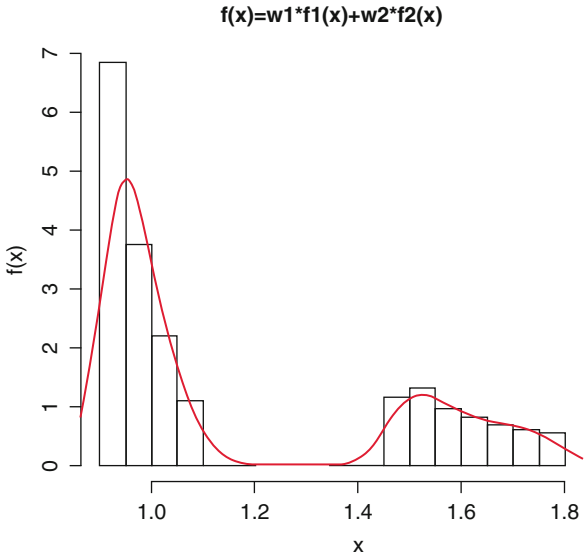


Fig. 13.3 Distribution function of mixtured GPD 1 and 2 in Table 1

In Table 13.2, there are the success rates of the different parameters combinations in the narrow range. According to result, we can say that shape parameter has an effect on the success rates of generator in the narrow range.

In Table 13.3, same procedure is followed in Table 13.2 but in the wide range. Parameters are changed in the wide range. This is shown that the shape parameter is most effect parameters on the success rate of generator. If the shape parameter is increased in the wide range, the success rate of the generator is decreasing. But it is

Table 13.2 Success rates of the generator when changes in parameters in narrow range

Parameters	$\alpha = 1$ $\beta = 1$ $\mu = 1$	$\alpha = 2$ $\beta = 2$ $\mu = 1$	$\alpha = 2$ $\beta = 2$ $\mu = 1$	$\alpha = 1$ $\beta = 1$ $\mu = 2$	$\alpha = 2$ $\beta = 2$ $\mu = 2$
Success rate	0.902	0.91	0.91	0.902	0.91
Parameters	$\alpha = 3$ $\beta = 2$ $\mu = 2$	$\alpha = 3$ $\beta = 3$ $\mu = 2$	$\alpha = 3$ $\beta = 3$ $\mu = 3$	$\alpha = 4$ $\beta = 3$ $\mu = 3$	$\alpha = 4$ $\beta = 4$ $\mu = 3$
Success rate	0.91	0.91	0.91	0.906	0.906

Table 13.3 Success rates of the generator when changes in parameters in wide range

Parameters	$\alpha = 9$ $\beta = 1$ $\mu = 1$	$\alpha = 15$ $\beta = 1$ $\mu = 1$	$\alpha = 9$ $\beta = 7$ $\mu = 1$	$\alpha = 15$ $\beta = 7$ $\mu = 1$	$\alpha = 9$ $\beta = 1$ $\mu = 6$
Success rate	0.874	0.832	0.874	0.832	0.874
Parameters	$\alpha = 1$ $\beta = 7$ $\mu = 1$	$\alpha = 1$ $\beta = 12$ $\mu = 1$	$\alpha = 9$ $\beta = 7$ $\mu = 6$	$\alpha = 9$ $\beta = 12$ $\mu = 1$	$\alpha = 15$ $\beta = 7$ $\mu = 1$
Success rate	0.902	0.902	0.874	0.874	0.832
Parameters	$\alpha = 1$ $\beta = 1$ $\mu = 6$	$\alpha = 1$ $\beta = 1$ $\mu = 10$	$\alpha = 9$ $\beta = 1$ $\mu = 6$	$\alpha = 9$ $\beta = 7$ $\mu = 6$	$\alpha = 9$ $\beta = 7$ $\mu = 10$
Success rate	0.902	0.902	0.874	0.874	0.874

just about 0.03 %. Changes in the scale parameter have no important effect on the success rate of the generator. Mathematically, it is same for location parameter. As a result, shape parameter has an important effect on the success rate of the generator.

13.5 Conclusions

Researchers are faced with homogeneous data in their studies. They can model these data with a known distribution. In the growing research work area, the modeling might be easy like this. There are many data which are heterogeneous in many areas. In these cases, the mixture models can be more appropriate for modeling the data.

In this study, the GPMM is constructed with finite mixture weights with two components. After this, the random samples are generated from this mixture model. The results are tested with chi-squared goodness-of-fit test, and the success rate of the inverse transformation method is shown in Table 13.1 with significance level is

0.05. In this test, the components of mixture model are tested separately. After the seeing the success of the method, the generated random samples' graphs are drawn in Figs. 13.1 and 13.2.

As a result, the inverse transformation method is useful way for generating random samples from the generalized pareto distribution. After researchers obtain the success of the generator, the mixture model is used for generating random samples with mixture weights. We can say that the random samples from GPMM can be safely generated by the inverse transformation method.

References

1. Pickands, J.: Statistical inference using extreme order statistics. Ann. Stat. **3**, 119–131 (1975)
2. Davison, A.C.: Modeling excesses over high threshold with an application. In: de Oliveira, T. (ed.) Statistical Extremes and Applications, pp. 416–482. Riedel, Dordrecht (1984)
3. Castillo, E., Hadi, A.S.: Fitting the generalized pareto distribution to data. JASA **92**, 1609–1620 (1997)
4. Castillo J, Daoudib J.: Estimation of the generalized Pareto distribution. Stat. Probab. Lett. **79**(5):684–688 (2008)
5. Mierlus-Mazilu, I.: On generalized pareto distributions. Rom. J. Econ. Forecast. **8**, 107–117 (2010)
6. Beirlant, J., de Wet, T., Goegebeur, Y.: A goodness-of-fit statistic for Pareto-type behavior. J. Comput. Appl. Math. **186**, 99–116 (2006)
7. Raja, T.A., Mir, A.H.: On fitting of generalized pareto distribution. Global J. Human Soc. Sci. Econ. **13** (2013)
8. Kroese, D.P., Taimre, T., Botev, Z.I.: Handbook of Monte Carlo Methods. Wiley, New York, 8–11 (2011)
9. Pocatilu, P., Alecu, F., Vetrici, M.: Measuring the efficiency of cloud computing for E-learning systems. WSEAS Trans. Comput. **9**, 42–51 (2010)
10. Farnoosh, R., Zarpak, B.: Image restoration with Gaussian mixture models. WSEAS Trans. Math. **4**(3), 773–777 (2004)
11. Deguenon J., Barbulescu A.: GPD models for extreme rainfall in Dobrudja. Comput. Eng. Syst. Appl. **2**, 131–136 (2011)

Chapter 14
Lecture Notes in Computer Science: Statistical Causality and Local Solutions of the Stochastic Differential Equations Driven with Semimartingales

Ljiljana Petrović and Dragana Valjarević

Abstract The paper considers a statistical concept of causality in continuous time between filtered probability spaces which is based on Granger's definition of causality. Then, the given causality concept is connected with a local weak solutions of the stochastic differential equations driven with semimartingales. Also, we establish connection between the local solution and the local weak solution.

Keywords Filtration • Causality • Local weak solution

14.1 Introduction

The concept of local weak solutions were investigated by many scientists. In this paper we consider a local weak solutions of the stochastic differential equations driven with semimartingales.

In Sect. 14.2 we give some definitions and basic properties of the causality concept (see [1–5]). The given causality concept is closely connected to the orthogonality of local martingales (see [6]) and with stable subspaces of H^p which contains right continuous modifications of martingales (see [7]). Also, the preservation of the martingale representation property, if the information σ-algebra decreases is shown to be strongly connected to the concept of causality (see [6]).

L. Petrović (✉) • D. Valjarević
Faculty of Economics, Department of Mathematics and Statistics,
University of Belgrade, Kamenička 6, 11000 Beglrade, Serbia

Science Faculty, Department of Mathematics, University of Kosovska Mitrovica,
Lole Ribara br. 29, 38240 Kosovska Mitrovica, Serbia
e-mail: petrovl@ekof.bg.ac.rs; dragana.valjarevic@pr.ac.rs

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_14

In Sect. 14.3 we give a definition of local weak solution (in terms of causality) for stochastic differential equation (introduced in [8])

$$\begin{cases} dX_t = u_t(X)dZ_t \\ X_0 = x_0 \end{cases} \tag{14.1}$$

where Z_t is a semimartingale (with $Z_0 = 0$), $u_t(X)$ is $(\mathcal{F}_t^{Z,X})$-predictable process. Definitions of local strong and local weak solutions, using the standard methods, are given in [9]. This method involves enlarging of the probability space and showing that a weak solution exists on the enlarged space, with an argument that the new semimartingale is in a reasonable sense "the same" as the original semimartingale. In this paper we give a definition of local weak solutions using the concept of causality, according to Definition of weak solution for the Eq. (14.1) given in [2, 4, 10].

The concept of local weak solutions for equations driven by process of Brownian motion, their existence, uniqueness and connection with strong solutions, have been studied in [11, 12]. In [11] those results are connected with a model for the stochastic evolution of forward rate curves.

14.2 Concept of Statistical Causality

The study of Granger-causality has been mostly concerned with time series. But, many of the systems to which it is natural to apply tests of causality, take place in continuous time, so we will consider continuous time processes.

A probabilistic model for a time-dependent system is described by $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where $(\Omega, \mathcal{F}, P)$ is a probability space and $\{\mathcal{F}_t,\ t \in I\}$ is a "framework" filtration, i.e. $(\mathcal{F}_t)$ are all events in the model up to and including time t and $(\mathcal{F}_t)$ is a subset of $\mathcal{F}$. We suppose that the filtration $(\mathcal{F}_t)$ satisfy the "usual conditions", which means that $\{\mathcal{F}_t, t \in I\}$ is right continuous and each $(\mathcal{F}_t)$ is complete.

An analogous notation will be used for filtrations $\mathbf{H} = \{\mathcal{H}_t\}$, $\mathbf{G} = \{\mathcal{G}_t\}$.

A family of σ-algebras induced by a stochastic process $X = \{X_t, t \in I\}$ is given by $\mathbf{F}^X = \{\mathcal{F}_t^X, t \in I\}$, where $(\mathcal{F}_t^X) = \sigma\{X_u, u \in I, u < t\}$, being the smallest σ-algebra with respect to which the random variables $X_u, u \leqslant t$ are measurable. The process X_t is $(\mathcal{F}_t)$-adapted if $(\mathcal{F}_t^X) \subseteq (\mathcal{F}_t)$ for each t.

The intuitively plausible notion of causality for families of Hilbert spaces is given in [13, 14] and generalized in [3]. It is natural to introduce the following definition of causality between filtrations.

Definition 1 (Compare with [3, 14]). It is said that $\mathbf{G}$ causes $\mathbf{H}$ within $\mathbf{F}$ relative to P (and written as $\mathbf{H} |< \mathbf{G}; \mathbf{F}; P)$ if $(\mathcal{H}_\infty) \subseteq (\mathcal{F}_\infty)$, $\mathbf{G} \subseteq \mathbf{F}$ and if $(\mathcal{H}_\infty)$ is conditionally independent of $(\mathcal{F}_t)$ given $(\mathcal{G}_t)$ for each t. If there is no doubt about P, we omit "relative to P".

If $\mathbf{G}$ and $\mathbf{F}$ are such that $\mathbf{G} \mid\!\prec \mathbf{G};\mathbf{F}$, we shall say that $\mathbf{G}$ is its own cause within $\mathbf{F}$ (compare with [15]).

The assertion $\mathbf{G} \mid\!\prec \mathbf{G};\mathbf{F};\ P$ implies that $\mathcal{G}_t = \mathcal{F}_t \cap \mathcal{G}_\infty$ for every $t \geqslant 0$. Also, $(\mathcal{G}_t)$ is a filtration generated by the continuous martingales of the form $M_t = P(A \mid \mathcal{F}_t)$, $A \in \mathcal{G}_\infty$ (see [16]).

This definition can be applied to stochastic process: it will be said that stochastic processes are in a certain relationship if and only if the corresponding induced filtrations are in this relationship. For example, $(\mathcal{F}_t)$-adapted stochastic process X_t is its own cause if $\mathbf{F}^X = (\mathcal{F}_t^X)$ is its own cause within $\mathbf{F} = (\mathcal{F}_t)$ i.e. if

$$\mathbf{F}^X \mid\!\prec \mathbf{F}^X;\mathbf{F};\ P.$$

Let T be a stopping time. With stopping times can be defined stopped variables, stopped processes and the stopped σ-algebras.

Definition 2 ([9]). Let X be a stochastic process and let T be a stopping time.

1) By a stopped or a truncated process we mean the process

$$\mathbf{X^T} = \{X_{t\wedge T} \mid t \in \mathbf{R}_+\}.$$

2) The random variable $X_T(\omega) = X(T(\omega), \omega)$ is called a stopped variable.
3) The stopped σ-algebra $\mathcal{F}_T$ is the set of events $A \in \mathcal{F}_\infty$ for which

$$\mathcal{F}_T = \sigma\{A \in \mathcal{F}_\infty : A \cap \{T \leqslant t\} \in \mathcal{F}_t, \forall t \geqslant 0\}.$$

We give now the characterization of causality using stopping times—a class of random variables that plays the essential role in the Theory of Martingales (see [17]). More precisely, we define causality using σ-fields associated to stopping times.

Theorem 1 ([16, 18]). *Let $X = \{X_t\}$ is $(\mathcal{F}_t)$-adapted stochastic process and $\mathbf{G} \subseteq \mathbf{F}$. The filtration $\mathbf{G}$ causes $\mathbf{F}^X$ within $\mathbf{F}$, i.e. $\mathbf{F}^X \mid\!\prec \mathbf{G};\mathbf{F};\ P$ if and only if one of the following properties is satisfied:*

(i) $(\mathcal{F}_\infty^X)$ is conditionally independent of $(\mathcal{F}_t)$ given $(\mathcal{G}_t)$ for each t, i.e. $\mathcal{F}_\infty^X \perp \mathcal{F}_t | \mathcal{G}_t$ for each t.
(ii) For any stopping time S relative to filtration $\mathbf{G}$, $\mathcal{F}_\infty^X \perp \mathcal{F}_S | \mathcal{G}_S$.
(iii) For any stopping time T relative to filtration $\mathbf{F}^X$ and any stopping time S relative to filtration $\mathbf{G}$, $\mathcal{F}_T^X \perp \mathcal{F}_S | \mathcal{G}_S$.
(iv) There exists an increasing sequence of stopping times $\{T_n\}$ (depending on X), such that $\lim_n T_n = \infty$ a.s.and that for each n holds $\mathcal{F}_{T_n}^X \perp \mathcal{H}_T \mid \mathcal{G}_T$.

Theorem 2 ([9]). *Let T be a stopping time. Then*

$$\int_0^{t\wedge T} H_s(X)dZ_s = \int_0^t H_s(X)I\{s \leqslant T\}dZ_s = \int_0^t H_s(X)dZ_{s\wedge T}.$$

14.3 Local Solution and Local Weak Solution

In this section we introduce the notion of local solution and local weak solution. As a start we give three basic canonical spaces, as introduced in [8, 19].

14.3.1 The Canonical Space of Driving Processes

Let $(\dot{\Omega}, \dot{\mathcal{F}}, \dot{\mathcal{F}}_t, \dot{P})$ be a filtered probability space on which is defined the driving process $\dot{Z}$ (with $\dot{Z}_0 = 0$. Then $\dot{\mathcal{F}} = \sigma\{\dot{Z}_s, s \geqslant 0\}$ and $(\dot{\mathcal{F}}_t) = \bigcap_{s>t} \sigma\{\dot{Z}_r, r \leqslant s\} \Rightarrow (\dot{\mathcal{F}}_t) = (\mathcal{F}_t^{\dot{Z}}) \Rightarrow \dot{\mathbf{F}} = \mathbf{F}^{\dot{Z}}$ (where $\dot{\mathbf{F}} = \{\dot{\mathcal{F}}\}_{t\geqslant 0}$). $\dot{T}$ is a stopping time relative to $(\mathcal{F}_t^{\dot{Z}})$.

14.3.2 The Canonical Space of Solutions

Let $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathcal{F}}_t, \hat{P})$ be a filtered probability space on which is defined the solution process $\hat{X}$. Then we have $\hat{\mathcal{F}} = \sigma\{\hat{X}_s, s \geqslant 0\}$ and $(\hat{\mathcal{F}}_t) = \bigcap_{s>t} \sigma\{\hat{X}_r, r \leqslant 0\}$, that is $(\hat{\mathcal{F}}_t) = (\mathcal{F}_t^{\hat{X}}) \Rightarrow \hat{\mathbf{F}} = \mathbf{F}^{\hat{X}}$ (where $\hat{\mathbf{F}} = \{\hat{\mathcal{F}}_t\}_{t\geqslant 0}$).

14.3.3 The Joint Canonical Space

We consider the product space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where is:

$$\Omega = \dot{\Omega} \times \hat{\Omega} \ ; \ \mathcal{F} = \dot{\mathcal{F}} \otimes \hat{\mathcal{F}} \ ; \ \mathcal{F}_t = \bigcap_{s>t} (\dot{\mathcal{F}}_s \otimes \hat{\mathcal{F}}_s)$$

and $\mathbf{F} = \{\mathcal{F}_t\}_{t\geqslant 0}$. Let

$$\dot{\varphi} : \Omega \to \dot{\Omega} \ \text{ with } \ \dot{\varphi}(\dot{\omega}, \hat{\omega}) = \dot{\omega},$$
$$\hat{\varphi} : \Omega \to \hat{\Omega} \ \text{ with } \ \hat{\varphi}(\dot{\omega}, \hat{\omega}) = \hat{\omega},$$

be the projection mappings. We denote by Z and X processes on Ω such that

$$\dot{\varphi}(Z) = \dot{Z} \ ; \ \hat{\varphi}(X) = \hat{X}$$

and for stopping time T is $\dot{\varphi}(T) = \dot{T}$. We also, have

$$\dot{\varphi}(\mathbf{F}) = \dot{\mathbf{F}} = \mathbf{F}^Z \quad \text{and} \quad \hat{\varphi}(\mathbf{F}) = \hat{\mathbf{F}} = \mathbf{F}^X.$$

The coefficient $u_t(X)$ which is $(\mathcal{F}_t)$-predictable and bounded process on Ω is defined on this space.

We associate to semimartingale Z a triplet (A, C, ν) which is called a local characteristics of Z. The concept is carried over from [19, 20].

Let us consider a stochastic differential equation (as introduced in [8]):

$$\begin{cases} dX_t = u_t(X)dZ_t \\ X_0 = x_0 \end{cases} \tag{14.2}$$

This equation is interesting mostly because solutions of this equation are semimartingales.

Notion of local strong solution and local weak solution of the Eq. (14.2) is introduced in [9, 21].

Due to introduced canonical space of the driving process and considering the Definition 3.5 in [19], we introduced next definition of the local strong solution.

Definition 3 (Compare with [8, 19]). $(\dot{X}, \dot{T})$ is a local solution on the driving system $(\dot{\Omega}, \dot{\mathcal{F}}, \dot{\mathcal{F}}_t, \dot{P}, \dot{Z})$ if $\dot{T}$ is $(\dot{\mathcal{F}}_t)$-stopping time and $\dot{X}$ is a right continuous and left hand limited $\dot{\mathbf{F}}$-adapted process which satisfy the equation

$$X_{t\wedge T} = x_0 + \int_0^{t\wedge T} u_s(X)dZ_s \tag{14.3}$$

up to $\dot{P}$-evanescent set.

By $\dot{X}$ we denote the mapping $\dot{\Omega} \to \hat{\Omega}$ defined by $\dot{X}(\dot{\omega}) = \hat{\omega}$, and $u_t(X)$ is $(\mathcal{F}_t)$-predictable process.

Next Definition gives a local weak solutions for Eq. (14.2) on a joint canonical space, due to Definition 1.7 in [8].

Definition 4 (Compare with [8, 19]). A local weak solution of Eq. (14.2) is a probability measure P on $(\Omega, \mathcal{F})$ which satisfies:

1) $P \circ \dot{\varphi}^{-1} = \dot{P}$,
2) Z is a semimartingale on $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ keeping all the local characteristics,
3) T is $(\mathcal{F}_t^{Z,X})$-stopping time,
4) X is adapted process and satisfy the Eq. (14.3).

On the joint canonical space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ we define T as a stopping time relative to $(\mathcal{F}_t^{Z,X})$, or $T : \Omega \to [0, +\infty]$ is a random time and it is a stopping time if

$$\{(\dot{\omega}, \hat{\omega}); T(\dot{\omega}, \hat{\omega}) \leqslant t\} \in \mathcal{F}_t^{Z,X}, \quad t \in [0, +\infty].$$

Using the projective mapping $\dot{\varphi} : \Omega \to \dot{\Omega}$, we have that $\dot{\varphi}(T) = \dot{T}, \dot{T} : \dot{\Omega} \to [0, +\infty]$ and $\dot{\varphi}(T) = \dot{\varphi}(T(\dot{\omega}, \hat{\omega})) = \dot{T}(\dot{\omega})$ or

$$\{\omega; T(\omega) \leqslant t\} \in \mathcal{F}_t^{Z,X} \iff \{(\dot{\omega}, \hat{\omega}); T(\dot{\omega}, \hat{\omega}) \leqslant t\} \in \mathcal{F}_t^{Z,X}$$
$$\iff \dot{\varphi}\left(\{(\dot{\omega}, \hat{\omega}); T(\dot{\omega}, \hat{\omega}) \leqslant t\}\right) \in \dot{\varphi}\left(\mathcal{F}_t^{Z,X}\right)$$
$$\iff \{\dot{\omega}; \dot{T}(\dot{\omega}) \leqslant t\} \in \dot{\mathcal{F}}_t = \mathcal{F}_t^{\dot{Z}}.$$

In other words $\dot{\varphi}(T) = \dot{T}$ is $(\mathcal{F}_t^Z)$-stopping time.

Using the fact that T is $(\mathcal{F}_t^{Z,X})$-stopping time, by Proposition 4.6 in [19] and Theorem 1, the conditions (2) and (3) from Definition 4, can be replaced with

$$\mathbf{F}^Z \models \mathbf{F}^Z; \mathbf{F}^T; P \quad \text{or}$$
$$\forall A \in (\mathcal{F}_\infty^Z) \qquad P(A \mid \mathcal{F}_t^Z) = P(A \mid \mathcal{F}_{t\wedge T}).$$

Now, based on definition of regular weak solution (introduced in [4, 10]) and using the connection between the solution measure and conditional expectation (see Proposition 4.6 in [19]) we can give the definition of local weak solution using the concept of causality.

Definition 5. An object $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, Z_t, T)$ is a local weak solution of Eq. (14.2) if

1) $\mu_Z(A) = P\{Z \in A\}$ coincides with predetermined measure on the function space where Z takes values;
2) $\mathbf{F}^Z \models \mathbf{F}^Z; \mathbf{F}^T; P$ or
$$\forall A \in (\mathcal{F}_\infty^Z) \qquad P(A \mid \mathcal{F}_t^Z) = P(A \mid \mathcal{F}_{t\wedge T}),$$
3) T is $(\mathcal{F}_t^{Z,X})$-stopping time (called the lifetime of X);
4) X is adapted and satisfy the Eq. (14.3).

If $T = \infty$ we just refer to $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, Z_t)$ as weak solution of the Eq. (14.2) (defined in [4, 10]).

The following theorem gives the connection between regular weak solution and weak local solution.

Theorem 3. *If T is $(\mathcal{F}_t^{Z,X})$-stopping time then the following two assertions are equivalent:*

a) $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, Z_t)$ is regular weak solution of Eq. (14.2);
b) $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, Z_t, T)$ is local weak solution of Eq. (14.2).

Proof. Using the definition of regular weak solution (see Definition in [4, 10]) we see that the difference is only in the fourth condition, because we already have assumption that T is $(\mathcal{F}_t^{Z,X})$-stopping time.

Regular weak solution of Eq. (14.2) must satisfy the equation:

$$X_t = x_0 + \int_0^t u_s(X) dZ_s \tag{14.4}$$

but local weak solution must satisfy the equation

$$X_{t\wedge T} = x_0 + \int_0^{t\wedge T} u_s(X)dZ_s.$$

If $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, Z_t)$ is a regular weak solution then for $t \leqslant T$ local weak solution satisfy the Eq. (14.4), but for $T \leqslant t$ local weak solution must satisfy the equation

$$X_T = x_0 + \int_0^T u_s(X)dZ_s.$$

According to Theorem 2 it follows that

$$X_T = x_0 + \int_0^T u_s(X)dZ_s = x_0 + \int_0^t u_s(X)I\{s \leqslant T\}dZ_s = x_0 + \int_0^t u_s(X)dZ_{s\wedge T}.$$

The same equality holds for $T \leqslant t$ the previous equation is obviously satisfied because $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, Z_t)$ is a regular weak solution. So, $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, Z_t, T)$ is a local weak solution of the Eq. (14.2).

Conversely, it is obviously satisfied if we set $T = \infty$.

14.4 Some Practical Applications

Causality is a topic which nowadays receives much attention. Many scientist in statistics, social science, computer science (especially those in artificial intelligence), econometrica, medicine and philosophy are investigating questions like "what would have happened if" and "what would happen if".

Causality is, in any case, a prediction property and the central question is: is it possible to reduce available information in order to predict a given filtration?

The study of Granger-causality has been mainly preoccupied with time series (see seminal paper Econometrica, 1969) and was first extended to a Markov chains, and later to more general stochastic processes.

We shall instead concentrate on continuous time processes. Many of systems to which it is natural to apply tests of causality, take place in continuous time. For example, this is generally the case within economy.

Continuous-time concepts of causality become more and more frequent in econometric practice. Let us mention some important fields of applications. In labor economics, duration models, Markovian and, more generally, counting processes appear to be powerful tools describing individual mobilities between participation states, or to analyze cohort data in demographics. At the same time, modern finance theory uses extensively diffusion processes.

Also, the continuous time concepts of causality are relatively easy to deal with in the context of processes admitting a (Gaussian) continuous time invertible moving average (CIMA) representation, which are particular cases of the MAR processes (i.e., processes with moving average representation).

Models based od stochastic differential equations are frequently used in many areas, such as climate science, pollution, traffic monitoring and ecology [22, 23].

Acknowledgement The work is supported by the Serbian Ministry of Science and Technology (Grants 044006 and 179005).

References

1. Granger, C.W.J.: Investigation causal relations by econometric models and cross spectral methods. Econometrica **37**, 424–438 (1969)
2. Mykland, P.A.: Statistical Causality. Statistical Report No. 14, University of Bergen (1986)
3. Petrović, L.: Causality and Markovian representations. Stat. Probab. Lett. **29**, 223–227 (1996)
4. Petrović, L., Stanojević, D.: Statistical causality, extremal measures and weak solutions of stochastical differential equations with driving semimartingales. J. Math. Model. Algorithm. **9**, 113–128 (2010)
5. Petrović, L.: Statistical causality and stochastic dynamic systems. Int. J. Appl. Math. Inf. **5**(3), 153–156 (2011) [ISSN 2074–1278]
6. Valjarević, D., Petrović, L.: Statistical causality and orthogonality of local martingales. Stat. Probab. Lett. **82**, 1326–1330 (2012)
7. Petrović, L., Valjarević, D.: Statistical causality and stable subspaces of H^p. Bull. Aust. Math. Soc. **88**, 17–25 (2013)
8. Jacod, J., Memin, J.: Existence of weak solutions for stochastic differential equation with driving semimartingales. Stochastics **4**, 317–337 (1981)
9. Protter, P.: Stochastic Integration and Diferential Equations. Springer, Berlin (2004)
10. Petrović, L., Stanojević, D.: Some models of causality and weak solutions of stochastic differential equations with driving semimartingales. Fac. Univ. Ser. **20**, 103–112 (2005)
11. Filipović, D.: Invariant manifolds for weak solutions to stochastic equations. Probab. Theory Relat. Fields **118**(3), 323–341 (2000)
12. Skorokhod, A.: On stochastic differential equations in a configuration space. Georegian Math. J. **8**(2), 389–400 (2001)
13. Gill, J.B., Petrović, L.: Causality and stochastic dynamic systems. SIAM J. Appl. Math. **47**, 1361–1366 (1987)
14. Petrović, L.: Causality and stochastic realization problem. Publ. Inst. Math. (Beograd) **45**(59), 203–212 (1989)
15. Mykland, P.A.: Stable subspaces over regular solutions of martingale problems. Statistical report No. 15, University of Bergen (1986)
16. Bremaud, P., Yor, M.: Changes of filtrations and of probability measures. Z. Wahrscheinlichkeitstheorie Verw. Gebiete **45**, 269–295 (1978)
17. Jacod, J.: Calcul Stochastique et Problemes de Martingales. Lecture Notes. Springer, Berlin (1979)
18. Petrović, L., Dimitrijević, S.: Invariance of statistical causality under convergence. Stat. Probab. Lett. **81**, 1445–1448 (2011)
19. Jacod, J.: Weak and strong solutions of stochastic differential equations. Stochastics **3**, 171–191 (1980)
20. Jacod, J., Shiryaev, A.N.: Limit Theorems for Stochastic Processes. Springer, Berlin (1994)

21. Kurtz, T., Protter, P., Weak convergence of stochastic integrals and differential equations, Lecture Notes in Mathematics: Probabilistics Models for Nonlinear Partial Differential Equations, vol. 1627, 1–41, Springer, Berlin (1996)
22. Guarnaccia, C.: Analysis of traffic noise in a road intersection configuration. WSEAS Trans. Syst. **9**(8), 865–874 (2010) [ISSN: 1109–2777]
23. Guarnaccia, C., Lenza, T.L.L., Mastorakis, N., Quartieri, J.: Traffic noise predictive models comparison with experimental data. In: Proceedings of the 4th WSEAS International Conference on Urban Planning and Transportation (UPT '11), Corfu Island, 14–16 July 2011, pp. 365–371

Chapter 15
A Mathematical Model to Optimize Transport Cost and Inventory Level in a Single Level Logistic Network

Laila Kechmane, Benayad Nsiri, and Azeddine Baalal

Abstract This paper proposes a mathematical model that minimizes transportation costs and optimizes distribution organization in a single level logistic network. The objective is to allocate customers to distribution centers and vehicles to travels in order to cut down the traveled distances, while observing the storage capacities of vehicles and distribution centers and covering the customers' needs. We propose a mixed integer programming formula that can be solved using Lingo 14.0. A digital example will be given in the end to illustrate the practicability of the model.

Keywords Distribution organization • Mixed integer programming • Single level logistic network • Transportation costs

15.1 Introduction

Supply chain is the succession of processes transforming raw materials into finished products delivered to the final customers, it consists of activities of supply, manufacturing, storage, distribution and sale. Transport is one of the main functions of supply chain which is present in several levels; to connect suppliers to factories, factories to distribution centers and the latter to customers. Connecting these various points via means of transport is what we call a logistic distribution network (see Fig. 15.1).

A logistic network consists of one or several levels; a single level logistic network is a network where there is a single intermediary between factories and customers, for example: distribution centers, whereas a multi level logistic network consists of several intermediaries between factories and customers, for example: distribution centers, centers of transfer, hubs, etc.

L. Kechmane (✉) • B. Nsiri • A. Baalal
Faculty of Sciences Casablanca, MACS Laboratory, Department of Mathematics and Computing, University Hassan II, Km 8 Eljadida Road, P.B. 5366, Maarif 20100, Morocco
e-mail: kechmanelaila@gmail.com; benayad.nsiri@telecom-bretagne.eu; abaalal@gmail.com

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_15

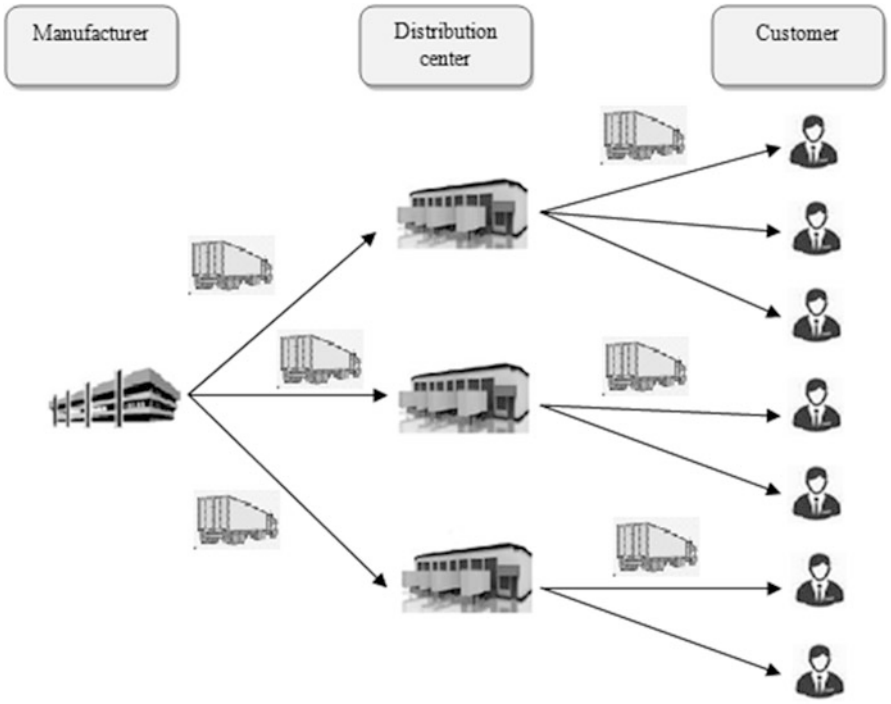


Fig. 15.1 Transport role in distribution network

Companies that manage distribution of their goods seek to cut down transport costs and avoid stock shortages at their distribution centers. To ensure a certain quality of customer service, companies have to manage the allocation of customers to distribution centers in a way that minimizes the costs of transport and considers the storage capacity of the vehicles as well as of the various distribution network's nodes.

The problem related to transport and distribution of goods ranges from vehicle routing problems such as the TSP (Travelling Salesman Problem) formulated by the mathematicians WR Hamilton and Thomas Kirkman in 1800 [1, 2], to problems that consider the interactions between the different activities like production, warehousing and transport to minimize the total costs [3] and problems of constructing the whole networks and thus to factories setting up and allocation of various nodes to customers [4, 5]. Vehicle Routing Problem has been an active area of research; Traveling Salesman Problem (TSP) focuses on finding the optimal route to visit a given number of cities while minimizing transportation cost [6, 7], the Vehicle routing problem (VRP), which is an extension of the TSP, was formulated in 1959 by Dantzig and Ramser [8], according to Laporte [9], this problem aims at building the optimal tours of pickup or delivery, from one or several warehouses

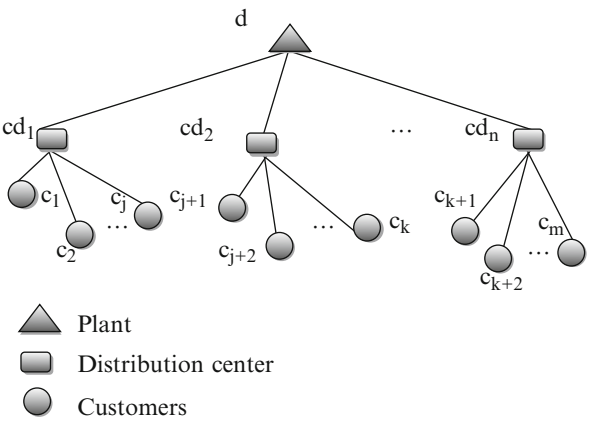


Fig. 15.2 A single level logistic network

towards a number of customers or cities that are geographically scattered, while respecting certain constraints. There exist four variants of the VRP [10]: VRP with Time Windows (VRPTW) [11], VRP with Pickup and Delivery (VRPPD) [12], the capacitated VRP (CVRP) [13] and the VRP with Backhauls (VRPB) [14].

The first algorithm to solve the VRP problem, was proposed by Clarke and Wright in 1964 [15], and since then, several methods were proposed and which are either exact methods that allow to find an optimal solution, or approximate methods that allow to obtain a solution to the problem but which is not optimal [16].

Since its introduction, the formulation of several models aiming at the optimization of the transport costs has been based on the VRP. Likewise, the proposed mathematical model is based on the VPR and addresses the minimization of transport costs as well as those of storage, both being parts of logistic distribution.

In the following section, we will begin by presenting our mathematical model, then, we will apply it to a real case and solve it by the Lingo 14.0 software to test its reliability.

We consider a logistic network consisted of one plant, n distribution centers and m customers (see Fig. 15.2). The first objective is to allocate customers to distribution centers and vehicles to travels so as to minimize the distances to travel, and ultimately the transport costs. The second objective is to minimize the storage costs at the distribution centers by minimizing the stored quantities while respecting the daily quantities that can be delivered to every center and satisfy the customers' needs (see Fig. 15.3).

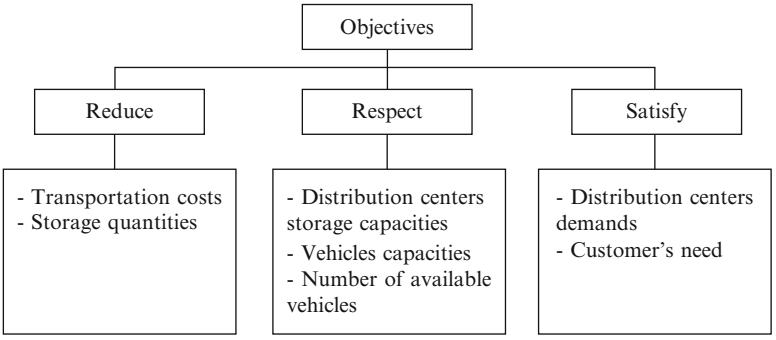


Fig. 15.3 Objectives of the model

15.2 Mathematical Formulation

Sets

I: Collection of distribution centers $i \in I, i = 1, 2, \ldots, n$;
M: Set of customers j, $j \in M, j = 1, 2 \ldots m$;
T: Set of periods t, $t \in T, t = 1, 2 \ldots t'$;
C_h: Set of vehicles vh with a capacity h;
C: Set of C_h;
Parameters

d_{ip}: Distance between plant p and center i
d_{ij}: Distance between center i and customer j
c_{vh}: Transportation cost per km for a vehicle vh
c_{si}: Unit cost of storage per day at distribution center i
n_h: Number of vehicles of category C_h
cap_{vh}: Vehicle vh capacity
b_i^t: Center i demand on day t
c_i: Storage capacity at center i
bes_j^t: Customer j demand on day t
stk_i^t: Available stock at center i on day t

Decision Variables

x_{ip}^t: Quantity to deliver from the plant p to center i on day t
y_{ij}^t: Quantity to deliver from center i to customer j on day t

$$l_{vh}^{it} = \begin{cases} 1 \text{ if vehicle vh visits center i on day t} \\ 0 \text{ else} \end{cases}$$

$$l_{vh}^{ijt} = \begin{cases} 1 \text{ if vehicle vh travels from center i to customer j on day t} \\ 0 \text{ else} \end{cases}$$

We assume all parameters are nonnegative.
Objective function:

$$\begin{aligned} Min & \sum_{v\in Ch} \sum_{t=1}^{t'} \sum_{i\in I} l_{vh}^{it} x_{ip}^{t} d_{ip} c_{vh} + \sum_{v\in Ch} \sum_{t=1}^{t'} \sum_{i\in I} \sum_{j\in M} l_{vh}^{ijt} y_{ip}^{t} d_{ij} c_{vh} \\ & + \sum_{t=1}^{t'} \sum_{i\in I} c_{si} \left(x_{ip}^{t} - b_i^t + stk_i^t \right). \end{aligned} \tag{15.1}$$

Subject to

$$x_{ip}^{t} - b_i^t \le c_i \quad \forall i \in I \quad \forall t \in T. \tag{15.2}$$

$$b_i^t \ge x_{ip}^{t} + stk_i^t \quad \forall i \in I \quad \forall t \in T. \tag{15.3}$$

$$stk_i^t = stk_i^{t-1} + x_{ip}^{t} - b_i^t \quad \forall i \in I \quad \forall t \in T. \tag{15.4}$$

$$\sum_{i\in I} y_{ij}^{t} = bes_j^t \quad \forall i \in I \quad \forall j \in M \quad \forall t \in T. \tag{15.5}$$

$$\sum_{vh\in C_h} l_{vh}^{it} \le n_h \quad \forall i \in I \quad \forall t \in T \quad \forall vh \in C_h. \tag{15.6}$$

$$l_{vh}^{it} x_{ip}^{t} \le cap_{vh} \qquad \forall i \in I \qquad \forall t \in T \qquad \forall vh \in C_h. \tag{15.7}$$

$$l_{vh}^{ijt} y_{ij}^{t} \le cap_{vh} \quad \forall i \in I \quad \forall j \in M \quad \forall t \in T \quad \forall vh \in C_h. \tag{15.8}$$

$$l_{vh}^{it}, l_{vh}^{ijt} \in \{0,1\} \quad \forall i \in I \quad \forall j \in M \quad \forall t \in T \quad \forall vh \in C_h. \tag{15.9}$$

The objective function (15.1) expresses the cost to be minimized and which is the sum of:

- Travelling costs from the plant to distribution centers;
- Travelling costs from centers to the customers;
- Storage costs at the distribution centers.

Constraint (15.2) assures the respect of the storage capacity of every distribution center.

Constraint (15.3) assures that the daily need for every distribution center is satisfied.

Constraint (15.4) calculates the quantity available in every distribution center.

Constraint (15.5) assures that the daily need of every customer is satisfied.

Constraint (15.6) assures the respect of the number of vehicles available in each category.

Constraints (15.7) and (15.8) assure the respect of each vehicle capacity.

15.3 Illustrative Example

To illustrate our model, we apply it to a network consisted of a single plant, 4 distribution centers and 13 customers. Table 15.1 includes storage parameters. There are two categories of vehicles for travels linking plant to centers and two other categories for travels linking centers to customers. Table 15.2 represents the characteristics of different vehicles, we note that when sending a vehicle to a center, it is completely filled, even if the sent quantity exceeds the center need, what explains the existence of stock.

Tables 15.3 and 15.4 represent respectively distances between plant and various distribution centers, and distances between the latter and customers. Tables 15.5 and 15.6 represent respectively the daily needs of distribution centers and of customers over a period of 4 days.

Table 15.1 Parameters values

Parameter	Value
c_{si}	0.20
c_i	500

Table 15.2 Characteristics of each type of vehicle vh

vh	A	B	C	D
c_{vh}	0.21	0.21	0.20	0.20
n_h	2	4	8	10
cap_{vh}	1,000	600	350	200

Table 15.3 Distances between the plant and distribution centers

	cd_2	cd_3	cd_4	cd_1
Plant	511	0	291	369

Table 15.4 Distances between distribution centers and customers

	cd_1	cd_2	cd_3	cd_4
c_1	419	99	390	469
c_2	172	351	642	721
c_3	651	133	166	237
c_4	719	259	204	93
c_5	303	241	485	611
c_6	746	23	60	267
c_7	614	93	198	281
c_8	772	314	29	422
c_9	439	72	361	433
c_{10}	735	217	82	221
c_{11}	87	483	773	818
c_{12}	907	411	120	423
c_{13}	910	390	286	60

Table 15.5 Daily distribution centers' need during period T

	j_1	j_2	j_3	j_4
cd_1	822	832	840	838
cd_2	793	1,007	985	1,015
cd_3	508	531	516	513
cd_4	476	472	460	481

Table 15.6 Daily customers' need during period T

	j_1	j_2	j_3	j_4
c_1	300	311	298	288
c_2	302	298	288	278
c_3	200	188	185	186
c_4	150	147	156	148
c_5	340	345	329	330
c_6	188	201	210	198
c_7	347	300	321	311
c_8	150	165	140	160
c_9	250	248	211	200
c_{10}	139	128	148	144
c_{11}	180	189	223	230
c_{12}	170	165	166	155
c_{13}	187	197	156	189

Table 15.7 Optimal quantities to be sent to the distribution centers during period T

	j_1	j_2	j_3	j_4
cd_1	1,000	1,000	1,000	1,000
cd_2	1,000	1,000	1,000	1,000
cd_3	600	600	600	600
cd_4	600	600	600	600

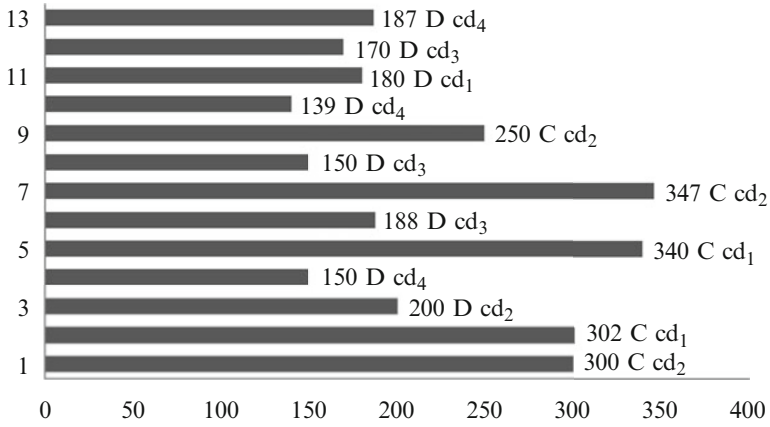


Fig. 15.4 Affectations on day J_1

15.3.1 Discussion

We solve this problem using a Mixed Integer Linear Programming solver LINGO 14.0 [17] on an Acer Aspire ONE D255 1.00 GHz machine, running Windows 7 Starter Edition. Results are obtained in 0.56 s, and the objective value is 901032.1.

Table 15.7 represents the optimal quantities to be sent to distribution centers during period T and which meet their needs. Figures 15.4, 15.5, 15.6, and 15.7 represent each the affectation of customers to distribution centers, optimal quantities to be sent on every day of period T and which category of vehicle to use.

We notice that obtained results respect the various constraints of our example, which are the storage capacity of distribution centers and the needs of the final customers. According to these results, we can easily deduct the optimal affectation of customers to distribution centers, which is, in this example represented in Fig. 15.8.

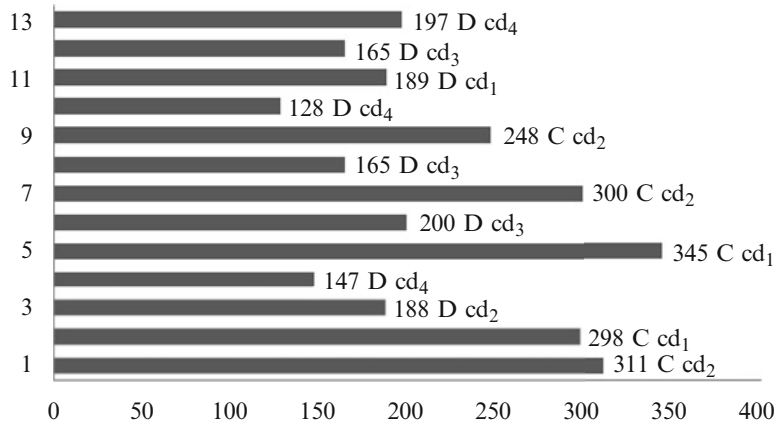


Fig. 15.5 Affectations on day J_2

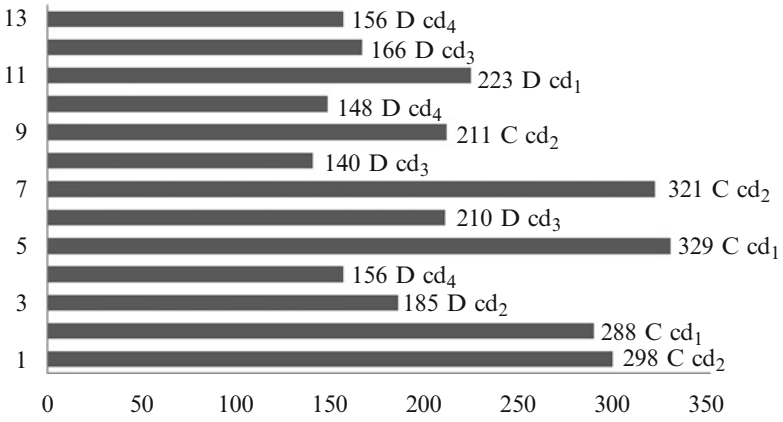


Fig. 15.6 Affectations on day J_3

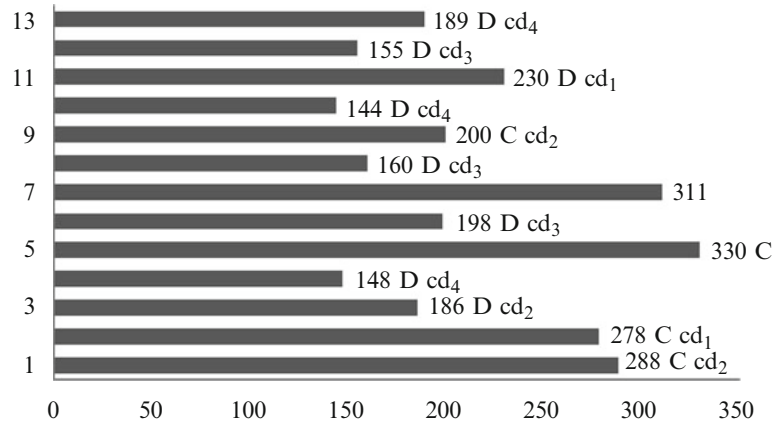


Fig. 15.7 Affectations on day J_4

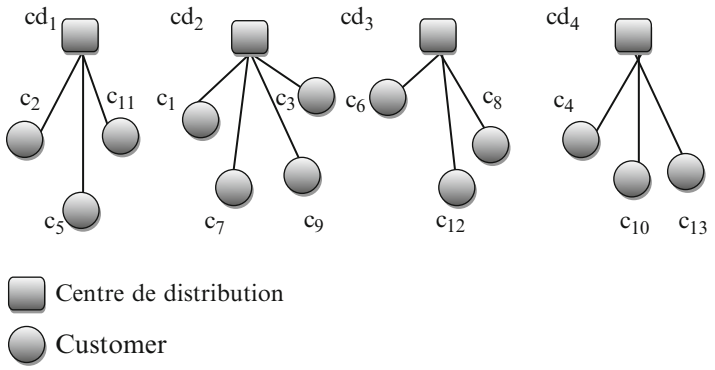


Fig. 15.8 Affectation of customers to distribution centers

15.4 Conclusion and Perspectives

The number of scientific publications handling transport problems continues to increase, so proving the importance of this function of supply chain. In this paper, we investigate the optimization of the distribution problem, the objective is to minimize both the traveled distances and the storage level, and allocate vehicles to travels. We relied on the vehicle routing problem VRP to develop our mathematical formula.

In this work, a single level logistic network is considered to apply our model. As perspective, we can consider a multi level logistic network, the model can be easily developed and applied in that case.

References

1. Davendra, D.: Traveling Salesman Problem, Theory and Applications. InTech, Hyderabad (2010)
2. Exnar, F., Machac, O.: The travelling salesman problem and its application in logistic practice. WSEAS Trans. Bus. Econ. **8**(4), 163–173 (2011)
3. Fahimnia, B., Luong, L., Marian, R.: Optimisation/simulation modeling of the integrated production-distribution plan: an innovative survey. WSEAS Trans. Bus. Econ. **5**(3), 44–57 (2008)
4. Chopra, S.: Designing the distribution network in a supply chain. Transp. Res. Part E Logist. Transp. Rev. **39**, 123–140 (2003)
5. Guerra, L., Murino, E., Romano, E.: The location-routing problem: an innovative approach. In: 6th WSEAS Transactions on System Science and Simulation in Engineering, Venice, Italy, 21–23 November 2007
6. Laporte, G.: The traveling salesman problem: an overview of exact and approximate algorithms. Eur. J. Oper. Res. **59**, 231–247 (1992)
7. Dantzig, G.B., Fulkerson, D.R., Johnson, S.M.: The solution of a large-scale traveling salesman problem. Oper. Res. **2**, 393–410 (1954)

8. Dantzig, G.B., Ramser, J.H.: The truck dispatching problem. Manag. Sci. **6**, 80–91 (1959)
9. Laporte, G.: The vehicle routing problem: an overview of exact and approximate algorithms. Eur. J. Oper. Res. **59**, 345–358 (1992)
10. Toth, P., Vigo, D.: The Vehicle Routing Problem. SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics, Philadelphia (2001)
11. Solomon, M.: Algorithms for the vehicle routing problem with time windows. Transp. Sci. **29**(2), 156–166 (1995)
12. Parragh, S.N., Doerner, K.F., Hartl, R.F.: A survey on pickup and delivery problems, part I: transportation between customers and depot. J. Betriebswirt. **58**, 21–51 (2008)
13. Ralphs, T.K., Kopman, L., Pulleyblank, W.R., Trotter Jr., L.E.: On the capacitated vehicle routing problem. Math. Program. **94**(2–3), 343–359 (2003)
14. Toth, P., Vigo, D.: An exact algorithm for the vehicle routing problem with backhauls. Transp. Sci. **31**(4), 372–385 (1997)
15. Clarke, G., Wright, J.W.: Scheduling of vehicles from a central depot to a number of delivery points. Oper. Res. **12**, 568–581 (1964)
16. Laporte, G.: What you should know about the vehicle routing problem. Nav. Res. Logist. **54**(8), 811–819 (2007)
17. LINGO: The Modeling Language and Optimizer. LINDO Systems Inc., Chicago (2013)

Chapter 16
Cost Optimization and High Available Heterogeneous Series-Parallel Redundant System Design Using Genetic Algorithms

Walid Chaaban, Michael Schwarz, and Josef Börcsök

Abstract Heterogeneous redundant series-parallel systems allow the mixing of components within the same subsystem. This diversity feature may improve the overall characteristics of the system compared with the homogeneous case in term of less susceptibility against so called common-cause failures and reduced cost. That means they guarantee longer availability and are quite suitable for systems that are designed to perform continuous processes. But the main challenging task is to determine the optimal design that corresponds to the minimal investment costs and satisfies the predefined constraints. This kind of combinatorial optimization tasks is perfectly solved using heuristic methods, since those approaches showed stability, powerfulness, and computing effectiveness in solving such matters. This task is more complex than the homogeneous case since the search space is getting larger due to the fact that every component available and that can be deployed in a subsystem has to be taken into account. This fact leads definitely to greater chromosome length and makes the search more time consuming. The algorithm has been implemented in Matlab and three different existing models (Levitin, Lisnianski, and Ouzineb) have been considered for a comparison with the homogeneous case and for validation purposes.

Keywords Common cause failure (CCF) • Genetic algorithms • Heterogeneous series-parallel systems • Redundancy allocation problem (RAP) • Universal moment generating function (UMGF)

W. Chaaban (✉) • M. Schwarz • J. Börcsök
Department of Computer Architecture and System Programming, University of Kassel, Willhelmshöher Allee. 71, Kassel 34121, Germany
e-mail: walid.chaaban@uni-kassel.de; m.schwarz@uni-kassel.de; j.boercsoek@uni-kassel.de

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_16

16.1 Introduction

Heuristic search methods represent powerful and effective means in solving combinatorial optimization problems since they do not require any additional information compared with the classical optimization methods and they accelerate the search towards objective or convergence through their parallel performed exploration and exploitation of the search space.

One well known combinatorial optimization task solved using such heuristic approaches is the Redundancy Allocation Problem (RAP) also referred to as Redundancy Optimization Problem (ROP) which consists of determining best series-parallel system designs in terms of redundancy depth on different subsystems level corresponding to minimal investment costs and that satisfies at the same time the predefined constraints and system design requirement specifications like availability, weight, volume and etc.

The Redundancy Allocation Problem (RAP) or Redundancy Optimization Problem (ROP) [1] is a single objective optimization and can often be encountered in many applications areas of the safety engineering world like electrical power systems and in the consumer electronic industry where system designs are mostly assembled using standard certified component types with different characteristics, e.g., reliability, availability, nominal performance, cost, etc. This matter has been intensively studied over last two decades and has been classified as a complex nonlinear integer programming combinatorial problem, where deterministic or conventional mathematical optimization approaches become ineffective by means of computational effort and quality of solution [1].

Using heuristic and metaheuristic search methods, e.g. Genetic Algorithms (GAs), Tabu Search (TS), Simulated Annealing, etc., in solving such kind of combinatorial optimization problems aims to determine an optimal or near optimal, also called pseudo-optimal solution, to the proposed RAP, i.e. to find the best or at least one acceptable solution that satisfies the constraint(s). However, these approaches have shown instead how powerful and effective they are in finding high qualitative solutions for the addressed kind of problems, especially when the search space corresponding to the problem becomes too large and where conventional classical optimization methods become ineffective and useless. This kind of problems was first introduced by Ushakov [1] and has been further analyzed by Levitin and Lisnianski et al. [2–4], Ouzineb [5–7] and many others.

This paper deals with the cost optimization of heterogeneous structured series assembled systems, where mixing of components or usage of non-identical components within the same subsystem is allowed. This feature, compared with the homogeneous case, includes more complexity to the task because the corresponding search or solution space becomes larger, since every component available on the market has to be taken into account. It represents a single objective optimization (cost function) subject to one constraint which is the availability of the system.

The remainder of the paper is organized as follows. Section 16.2 gives a short introduction into heterogeneous series-parallel multi-states configurations and

a brief overview on the advantages obtained through mixing of components in addition to a short comparison with homogeneous systems. Section 16.3 shortly discusses genetic algorithms and its different operators while chromosomal encoding and random generating of solution candidates are implemented in Sect. 16.4. In Sect. 16.5 a detailed formulation of the optimization problem, which is solved using heuristic traditional GA genetic techniques, is presented. Section 16.6 deals with the Universal Moment Generating Function (UMGF), also called the *Ushakov*-transform, which represents the function used for the determination and evaluation of the availability of the different redundant structures. Section 16.7 reports different numerical and experimental results, evaluations and graphical representations obtained by the implemented GA algorithm for different analyzed models which will be compared with previously published evaluations in terms of efficiency, solution quality and accuracy in addition to algorithm computation speed and convergence time. Finally concluding remarks are resumed in Sect. 16.8.

16.2 Homogeneous vs. Heterogeneous Series-Parallel Configurations

In order to improve or to increase the system's reliability and provide longer operation time, safety system designers may introduce different parallel technologies into a system also called redundancy [8]. Including homogeneous components redundancy is a great and effective technique to achieve a desired level of reliability in binary state systems or to increase the availability of multi-state systems. Reliability analysis have shown that the availability of homogeneous redundant structures or systems is extremely affected by common cause failure (CCF) that cannot be ignored since the CCF is the simultaneous failure of all components of the same type due to a common cause (CC), which leads in homogeneous redundant structures definitely to the failure of complete subsystems consisting of identical components causing herewith a total system failure. Common cause events may arise from environmental loads (humidity, temperature, vibration, shock, etc.), errors in maintenance and system design flaws [9]. In order to partly overcome this kind of facing problems and avoid total system failure subject to CCF heterogeneous redundancy is used.

The main concept of heterogeneous or non-homogeneous structures consists of the mixture of non-identical components within the same subsystem. That means that all non-identical components with the same functionality available on the market and which can be deployed in a redundant manner within the same subsystem have to be taken into account in this case, the fact that would definitely enlarge the size of the search space of feasible solutions and increase the exploration and hence the convergence time towards acceptable solutions.

The main advantages and benefits of components mixing lie in the improvement of the availability of the whole system and reducing the effect of common cause failure in addition of introducing flexibility and diversification into redundant system design through the allowed multiple component choice.

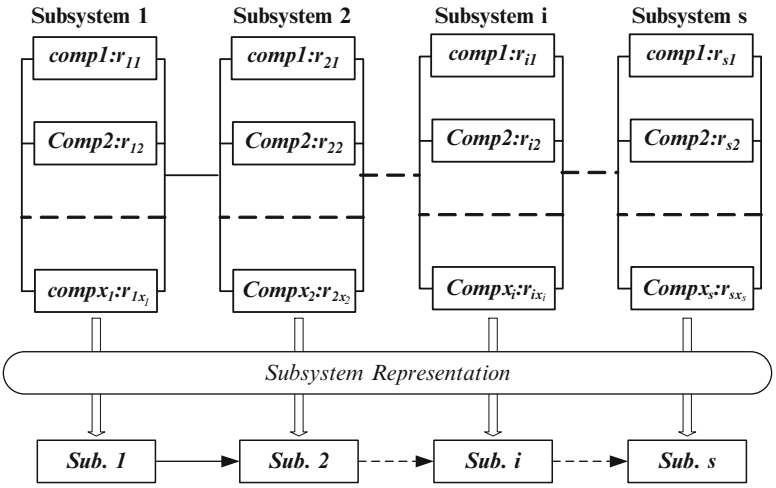


Fig. 16.1 Heterogeneous series-parallel configuration consisting of s-nodes or subsystems

Figure 16.1 represents a heterogeneous series-parallel multi-state system consisting of s subsystems, which are connected serially.

In Fig. 16.1 r_{ij} represents the reliability of a component of version j within the subsystem i. In the case of a homogeneous configuration the reliabilities of all components within the same subsystem are the same since subsystems consists of identical components, i.e. all r_{ij}'s are equal for the same i, which must not be the case in heterogeneous systems since they mostly deal with non-identical components on the same stage.

For a brief explanation, in addition to a short mathematical computation that shows why series-parallel configurations are more suitable and studied than parallel-series configurations, the reader should refer to [1, 10]. This is due to the fact that the overall reliability or availability of a system in a series-parallel configuration is better than the corresponding parallel-series configuration using the same set of components.

16.3 Genetic Algorithms

Genetic algorithms (GAs) are biologically inspired metaheuristic search and optimization routines that mimic the act of self-evolution concept of natural species that has been first laid by Charles Darwin. Nowadays, they are frequently used in many engineering and mathematical fields like optimization, self-adaptiveness, artificial intelligence, machine learning, etc. As computational efforts and speeds have been increasingly improved over the last decade, GAs have been expanded to cover a wide variety of applications including numerical and combinatorial

optimization tasks in engineering like the one discussed in the recent work. For further fundamentals and detailed information on genetic algorithms, the reader should refer to [11, 12].

GAs represent iterative self-adaptive stochastic techniques based on the concept of randomness. They mimic the process of the natural evolution of species. GAs have become very popular and widely used over the last decade and are very well suited as universal or common techniques for solving combinatorial optimization problems, e.g. multimodal functions (many peaks and local optima), the very well-known TSP (Travelling Salesman Problem) and redundancy allocation problems like the one discussed in this paper and many other matters [13].

GAs differ from normal optimization and search methods in four fundamental different ways [12]:

- The first difference is that GAs require an encoding of the parameter set in so called solution candidates, also called chromosomes according to the biological genetic terminology.
- Another difference is that GA starts the search from a start (initial) population and not from a single point like classical deterministic algorithms.
- GAs use information provided directly by the objective function and do not require any additional information or auxiliary knowledge like derivatives, gradients, etc.
- GAs apply probabilistic transition rules or operators (crossover, mutation) and not deterministic ones.

As mentioned before the search procedure starts from a random generated population of chromosomes that are encoded according to the addressed problem (binary, integer, decimal, etc.) conducting herewith a simultaneous search in many areas of the feasible solution space at once. The encoding of solutions constitutes the most difficult and challenging task of GAs and the evolving procedure from one population to the next is referred to as generation.

After each generation the new generated solutions are decoded and evaluated in terms of fitness with the help of the objective function. The fitness value of a chromosome represents a measure for its quality (fitness) and represents the decision maker of the selection operator since the probability of selection of chromosomes is proportional to its corresponding fitness value. A general overview of the genetic cycle is given in Fig. 16.2.

The genetic run process terminates when at least one of its predefined termination criterions is met, e.g., when the predefined maximum number of generations or repetitions N_{rep} or a specific number of successive runs without any solution's improvement is reached, or for example when the current solution satisfies the predefined requirement specifications or constraints.

Three main operators, also called genetic operators, will be executed during one genetic cycle, hence the selection, crossover or recombination and the mutation operator. These operators are shortly discussed in the following subsections.

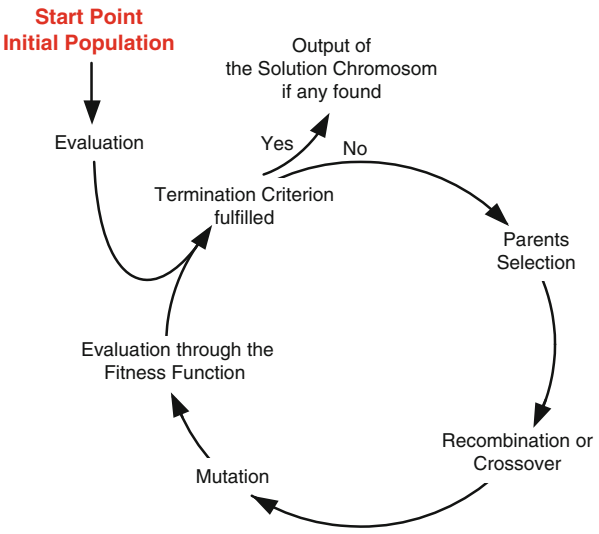


Fig. 16.2 General overview of the genetic process

16.3.1 Selection or Reproduction: Operator

Outgoing from a start or an initial population of different solution candidates the selection operator is used to randomly select or choose individuals or chromosomes which will reproduce and help building the next population during the genetic cycle. This operator represents an artificial version of natural selection, the Darwinian principle of the survival of the fittest, which drives the evolution towards optimization. Since the selection probability is proportional to relative fitness, chromosomes with higher fitness have better chance or higher probability to survive into next population, while chromosomes with bad or lower fitness will die off. This phenomenon will improve the population's average fitness from one population to the next.

There are many selection methods, some of them are listed in the following [14]

- Roulette Wheel selection,
- Tournament selection,
- Rank selection,
- etc.

16.3.2 Recombination or Crossover: Operator

Whereas the selection operator determines which chromosomes of the recent population are going to reproduce, the crossover operator performs jumps between

the different solution subspaces enabling the exploration of new areas of the solution space and avoiding herewith premature convergence in addition to the exchange of some basic characteristics or genetic materials and inheriting these properties to the offsprings which will join next populations. The crossover occurs with a predefined crossover rate or probability of crossover p_c. There are many crossover techniques used in genetic algorithms [1, 14, 15] like the one-point crossover, two-point crossover, uniform and half uniform crossover and many other crossover techniques.

In the following the one-point crossover operator is shortly discussed. For this purpose two parent chromosomes *Parent1* and *Parent2* are selected. Afterwards a pseudorandom real number is generated in [0; 1]. If and only if the generated number is less or equal p_c both parents will undergo crossover; otherwise they will have to be recopied in the population and wait to undergo the next operator, hence the mutation operator.

In case 2 chromosomes undergo crossover, a crossover point or position depending on the length of the chromosome is randomly selected on both selected parent chromosomes. This step is done by a pseudorandom integer number generator that generates numbers in the interval $[1;l_c - 1]$, where l_c is the length of the vector representing the chromosome. All data beyond that point in either chromosome will be swapped between the two parent organisms which will result in two new individuals called offsprings or children chromosomes. The one-point crossover technique is depicted in Fig. 16.3.

Figure 16.4 shows a pseudo code which resumes the crossover procedure.

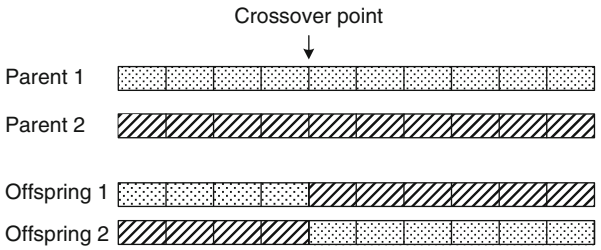


Fig. 16.3 One-point crossover technique

```
% Crossover(Parent1,Parent2,pc)
if (rand(1)<=pc)
     CrossPos=randi(lc);
     [child1,child2]=crossover(Parent1,Parent2,CrossPos)
end
```

Fig. 16.4 Pseudocode—one-point crossover

```
Chromosom before mutation: 1 0 0 0 1 0 1 0 1 1 1 1 1

Chromosom after mutation : 1 0 1 0 1 0 1 0 1 1 0 1 1
```

Fig. 16.5 Mutation of a binary encoded chromosome

Fig. 16.6 Pseudocode—mutation operator

```
% Mutation(Chromosom,p_m)
for i=1:length(chromosome)
    if (rand(1)<=p_m)
      chromosome(i)=mutate(chromosome(i),i)
    end
end
```

16.3.3 Mutation Operator

After crossovering parent chromosomes the resulting offsprings or children undergo mutation with a low mutation rate or probability p_m. The mutation operator introduces diversity into the GA algorithm and inserts small disturbance into the properties (genes) of the proposed solutions avoiding herewith premature convergence into local maxima. Mutation also helps recovering loss that might have been caused by crossover. After the mutation process has been accomplished, the new resulting mutated chromosomes constitute the new next population. The mutation of a binary encoded chromosome consists of inverting the randomly selected bit or position like depicted in Fig. 16.5, while a pseudo code which represents the mutation routine is shown in Fig. 16.6.

16.4 Chromosomal Encoding and Random Generating of Solution Candidates

Genetic algorithms are population based combinatorial optimization approaches where populations consist of a predefined number of solution candidates, also called chromosomes or individuals. These candidates represent vectors of encoded information, referred to as genetic materials, which will be decoded using the fitness or objective function in order to find the optimal (minimum or maximum) or near optimal solution of the addressed problem, whereas the recent approaches for solving the RAP problem are based on the Universal Moment Generating Function (UMGF) for estimating the availability of multi-state systems [2, 5–7, 16]. The encoding of chromosomes represents one of the major challenges faced in the context of genetic computing.

With regard to such problems that are dealt with in this paper, i.e., in the case of heterogeneous redundant structures the chromosome length corresponding to a

system is definitely longer than chromosome corresponding to the homogeneous case since each component version or type available on the market that may be deployed in a subsystem has to be taken into account, whereas in homogeneous systems, the chromosome length is equal to twice the number of subsystems, since only one component type or version is allowed on each stage. The chromosomes are integer encoded and each element x_{ij} of the chromosome vector corresponds to the number of components of version j used in subsystem i. The chromosome dimension or length l_c is given therefore through:

$$l_c = \sum_{i=1}^{s} J_i . \tag{16.1}$$

where s is the total number of subsystems or stages and J_i the total number of component versions available on the market that can be used on stage or in subsystem i. For more clarification and to gain a deeper insight it is important to mention at this point that two components of different versions or non-identical components connected in parallel are supposed to perform the same task or function. The difference lies in the technical data (reliability/availability, nominal performance and etc.) in addition to purchasing costs.

For example let us consider a system consisting totally of four subsystems with the following version vector $m = [4\ 6\ 8\ 5]$ representing the number of components available on the market for each subsystem. The chromosome length results in this case as the sum of all elements of the version vector and would give according to Eq. (16.1) a total chromosome length of $4 + 6 + 8 + 5 = 23$ and the chromosome encoding or vector X will look like in the following:

$$X = \left([x_{11} \ldots x_{14}], [x_{21} \ldots x_{26}], [x_{31} \ldots x_{38}], [x_{41} \ldots x_{45}] \right). \tag{16.2}$$

where x_{ij} denotes as mentioned previously the number of components of type j used in subsystem i.

For generating chromosomes or solution candidates according to the addressed optimization problem, discussed in this paper, a pseudo random number generator is used. Since events should happen at random but some events or numbers within the chromosomal encoding should have a higher probability of occurrence or happening than others, e.g. zeros which means that no components of this kind are used, a weighted pseudo random number generator is used.

As mentioned above, the step of generating numbers with predefined probabilities of occurrence in addition to the limitation of the maximum number of totally used components within each subsystem would reduce the search hypervolume. This would increase the computation speed drastically towards convergence and make the algorithm more efficient and time consuming.

16.5 Cost Optimization and Redundancy Allocation: Task Formulation

The cost optimization problem of heterogeneous series-parallel redundant systems discussed in this work deals with the determination of optimal redundant designs and the level of redundancy (redundancy allocation) to use in each subsystem that corresponds to the minimal total purchasing costs of the system and which satisfies at the same time the predefined availability constraints. This kind of optimization gives a rise to safety vs. economics conflicts resumed in the following two points [17]:

Choice of components: choosing high reliable components guarantees high system availability but may be largely non-economic due to high purchase prices; whereas choosing less reliable components for lower costs on one hand may decrease the availability of the system and increase drastically the accident costs on the other hand.

Choice of redundancy configuration: choosing highly redundant configurations increases definitely the reliability and availability of the system and is accompanied at the same time with higher purchase costs caused by additional equipment units required to improve individual subsystems reliabilities.

The previously described aspects of safety system design call for compromise choices which optimize system operation in view of recommended safety and longer operation time or budget constraint. As mentioned before this paper deals with customizing a GA for budgetary optimization and redundancy determination of multi state systems under a given availability constraint. This problem is considered as a single objective optimization and can be mathematically formulated as to minimize the cost function $C_{sys}(X)$ (objective function) of the whole system given by [1, 5–7, 16]:

$$C_{sys}(X) = \sum_{i=1}^{s} \sum_{j=1}^{m_i} c_{ij} x_{ij}. \tag{16.3}$$

where c_{ij} being the cost of component of type j in subsystem i and x_{ij} the number of components of type j used in subsystem i. m_i is the number of component choices available on the market which may be deployed in subsystem i. The (cost) objective function represented in Eq. (16.3) results over the sum of the purchasing costs of all components used in system that should meet at the same time the system specified availability constraint, which implies that the total availability of the system $A_{sys}(X)$ must satisfy a minimum level of availability required A_0 (inequality or availability constraints)

$$A_{sys}(X) \geq A_0. \tag{16.4}$$

Based on the UMGF or the *Ushakov*-transform, the total availability of the system $A_{sys}(X)$ is estimated as a function of system structure, performance and availability characteristics of its constituting components.

For a detailed overview of the UMGF in computing the availability of series-parallel systems the reader should refer to [2, 3, 5–7, 10, 17, 18].

16.6 Universal Moment Generating Function

The UMGF, also referred to as *Ushakov*-transform according to *I. Ushakov* (mid 1980s) or *u*-function, is a polynomial representation of the different states corresponding to a component or a system. For a great understanding of the mathematical fundamentals of the UMGF the reader should refer to [17]. For example, the *u*-transform $u_j(z)$ of a component or a random variable j having M different discrete states is given by

$$u_j(z) = \sum_{m=1}^{M} p_m z^{W_m}. \tag{16.5}$$

In Eq. (16.5) p_m is the probability that the nominal performance of the component or value of the discrete random variable j at state m equal to W_m.

Since in the recent study it is assumed that the used components are binary state, i.e., have two particular states (perfect working or complete failing), the *Ushakov*-polynomial representation of a binary state component j is given by

$$u_j(z) = \sum_{m=1}^{2} p_m z^{W_m} = \left(1 - A_j\right) z^0 + A_j z^{W_j} = \left(1 - A_j\right) + A_j z^{W_j}. \tag{16.6}$$

In Eq. (16.6) A_j represents the probability that the component is available (perfect functioning) and delivers a nominal performance of W_j ($\Pr[W_m = W_j] = A_j$) while $(1 - A_j)$ represents the probability of unavailability or failing (system not available, i.e., $\Pr[W_m = 0] = 1 - A_j$). The 0 power factor in the failing state results from the absence of delivered performance in this state.

In order to determine the *u*-function of an entire series-parallel system for availability computation purposes, two different basic operators will be respectively applied [5, 6, 10, 17, 18]. These two operators also called composition operators applied respectively [17] will be implemented separately in the following subsections.

16.6.1 Γ-Operator: Ushakov Transform of Parallel Configurations

The Γ-composition operator is used to determine the u-function of parallel systems. Suppose a system consisting of x_i parallel connected components, the corresponding u-function is given by the following equation

$$u_{parallel}(z) = \Gamma\left(u_1(z), u_2(z), \ldots, u_{x_i}(z)\right). \tag{16.7}$$

where the total performance or the structure function $f(W_1, W_2, \ldots, W_{xi})$ is given by the sum of the performances or capacities of the individual components or elements as described in the following equation

$$f\left(W_1, W_2, \ldots, W_{x_i}\right) = \sum_{i=1}^{xi} W_i. \tag{16.8}$$

For a pair of parallel connected components with the corresponding u-functions $u_1(z)$ and $u_2(z)$ given according to Eq. (16.5) the resulting u-function $u_{parallel}(z)$ of the entire parallel system is given by

$$\begin{aligned} u_{parallel}(z) &= \Gamma\left(U_1(z), U_2(z)\right) \\ &= \sum_{i=1}^{n} P_{1i} z^{W_{1i}} \sum_{j=1}^{m} P_{2j} z^{W_{2j}} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} P_{1i} P_{2j} z^{W_{1i}+W_{2j}}. \end{aligned} \tag{16.9}$$

In Eq. (16.9), n and m represent the number of states of the components 1 and 2. W_{1i} and W_{2j} are respectively the nominal performances of the components 1 and 2 at states i and j, which occur with the respective probabilities P_{1i} and P_{2j} ($i = 1 \ldots n$ and $j = 1 \ldots m$). All this means that the Γ-operator is nothing else than the polynomial product of the individual u-functions corresponding to all parallel connected components and therefore Eq. (16.9) can be represented as

$$u_{parallel}(z) = \prod_{e=1}^{x_i} u_e(z). \tag{16.10}$$

For a subsystem i within a series-parallel configuration consisting of x_i different parallel connected binary state components, whose UMGF representation is given by Eq. (16.6), the u-function according to Eq. (16.10) is written in the form

$$u_{parallel}(z) = \prod_{e=1}^{x_i} \left[\left(1 - A_{ij}\right) + A_{ij} z^{W_{ij}}\right]. \tag{16.11}$$

j represents the index corresponding to the version of the component used in case of non-homogeneity. Suppose that all parallel connected components are identical, Eq. (16.11) is rewritten as

$$u_{parallel}(z) = \left[\left(1 - A_{ij}\right) + A_{ij} z^{W_{ij}}\right]^{x_i}. \tag{16.12}$$

Using the binomial theorem the power representation of Eq. (16.12) can be expanded in a sum of the form

$$\begin{aligned} u_{parallel}(z) &= \sum_{k=0}^{x_i} \binom{x_i}{k} \left(A_{ij} z^{W_{ij}}\right)^k \left(1 - A_{ij}\right)^{x_i - k} \\ &= \sum_{k=0}^{x_i} \binom{x_i}{k} A_{ij}^k \left(1 - A_{ij}\right)^{x_i - k} z^{k W_{ij}} \\ &= \sum_{k=0}^{x_i} \alpha_{ik} z^{k W_{ij}} \end{aligned} \tag{16.13}$$

with

$$\alpha_{ik} = binm\left(k, A_{ij}, x_i\right) = \binom{x_i}{k} A_{ij}^k \left(1 - A_{ij}\right)^{x_i - k} \tag{16.14}$$

and where the binomial coefficients are given by

$$\binom{x_i}{k} = \frac{x_i!}{k!\,(x_i - k)!}. \tag{16.15}$$

That means that the u-function or the polynomial z-representation of each parallel subsystem consisting of x_i non-identical components (heterogeneous case) can be computed according to Eq. (16.11). If all x_i or a specific number x_n of components are identical, some simplification can be made using the binomial theorem according to Eq. (16.13).

16.6.2 *η-Operator: Ushakov Transform of Series Configurations*

In order to determine the u-function of a system consisting of s elements or components connected in series, the η-operator is applied according to

$$u_{series}(z) = \eta\left(u_1(z), u_2(z), \ldots, u_s(z)\right). \tag{16.16}$$

In the case of series connected components, the element with the minimal or least performance becomes the bottleneck of the system. This system is the decision maker about the total system performance or productivity. Therefore the structure or the performance function f is given by

$$f\ (W_1, W_2, \ldots, W_s) = \min\left(W_1, W_2, \ldots, W_s\right). \tag{16.17}$$

For a pair of components $u_1(z)$ and $u_2(z)$ connected in series and according to Eq. (16.17) the resulting u-function of the system $u_{series}(z)$ is given by

$$\begin{aligned} u_{series}(z) &= \eta\left(u_1(z), u_2(z)\right) \\ &= \eta\left(\sum_{i=1}^{n} P_{1i} z^{W_{1i}}, \sum_{j=1}^{m} P_{2j} z^{W_{2j}}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} P_{1i} P_{2j} z^{f\left(W_{1i}, W_{2j}\right)}. \end{aligned} \tag{16.18}$$

Replacing the structure function according through Eq. (16.17), the u-function of two series connected components will be given by

$$u_{series}(z) = \sum_{i=1}^{n}\sum_{j=1}^{m} P_{1i} P_{2j} z^{\min\left(W_{1i}, W_{2j}\right)}. \tag{16.19}$$

16.6.3 Ushakov Transform of Series-Parallel Systems: (Γ, η)

As mentioned previously, in order to determine the u-function of the entire series-parallel MSS both composition operators Γ and η have to be performed respectively like depicted in Fig. 16.7.

In Fig. 16.7 the individual s parallel systems consisting of non-identical binary state components will be replaced by single elements having multi-states u-functions ($u_1(z)$ … $u_s(z)$) determined by the Γ-composition operator according to Eq. (16.11). Afterwards the η-composition operator will be applied on the resulting system consisting of the s-multi-states components connected in series which leads to the u-function $u_{sys}(z)$ of the entire system that would be computed with the help of Eq. (16.16) like in the following

$$\begin{aligned} u_{sys}(z) &= \eta\left(u_1(z), u_2(z), \ldots, u_s(z)\right) \\ &= \eta\left(\sum_{k=0}^{x_1} \alpha_{1k} z^{kW_{1j}}, \sum_{k=0}^{x_2} \alpha_{2k} z^{kW_{2j}}, \ldots, \sum_{k=0}^{x_s} \alpha_{sk} z^{kW_{sj}}\right) \\ &= \sum_{m=0}^{M} \delta_m z^{W_m}. \end{aligned} \tag{16.20}$$

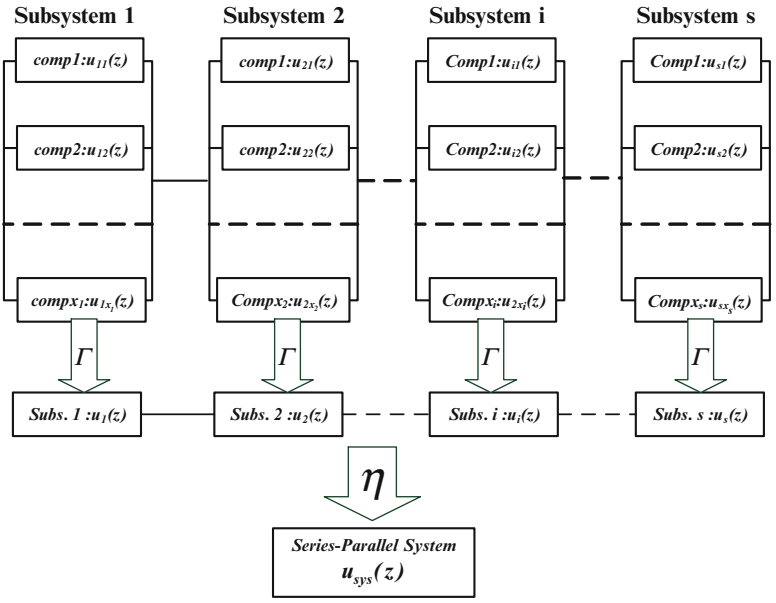


Fig. 16.7 Determination of the $u(z)$-function for the entire series-parallel system

δ_m and W_m are real numbers determined according to Eq. (16.19). The evaluation of the probability that the entire system satisfies a specific level of performance W_0 is given by the sum over all coefficients δ_m that correspond to a nominal performance W_m greater or equal W_0. The resulting sum represents the availability $A_{sys}(X)$ corresponding to the series-parallel system, whose design is represented by X and is given by

$$A_{sys}\,(W_0) = \Pr\,(W_m \geq W_0) = \sum_{W_m \geq W_0} \delta_m. \tag{16.21}$$

Given K different demand levels represented by W_0^k where $k = 1 \ldots K$, which should be satisfied over different operation intervals T_k the total availability $A_{sys}(X)$ of the system is obtained by the sum of the instantaneous availabilities corresponding to the different demand levels divided by the total operation or mission time and is given by

$$\begin{aligned} A_{sys}(X) &= \tfrac{1}{T}\sum_{k=1}^{K} \sum_{W_m \geq W_0^k} \delta_m T_k \\ &= \frac{1}{\sum_{k=1}^{K} T_k} \sum_{k=1}^{K} \sum_{W_m \geq W_0^k} \delta_m T_k. \end{aligned} \tag{16.22}$$

16.7 Tuning Parameters and Experimental Results: Validation

The simple genetic algorithm with some modifications in the context of its operators was used in this work. The algorithm has been implemented in Matlab which provides uniform pseudorandom number generators and powerful matrix and vector operations and allows great visualization and graphical representation. The three different models, which have been analyzed in the homogeneous case in a previous publication [10, 18], have been treated again in the heterogeneous case in order to show the effect of mixing of components on investment cost reduction and hence getting safer systems subject to CCFs for lower cost than the homogeneous case and for same given constraint factors. The used components are assumed to be binary state (perfect working or totally failing). The models and data, as mentioned previously, have been taken from [5–7], are listed below:

- Lev4_4_6_3
- Lev5_5_9_4
- Ouz6_4_11_4

For a brief understanding of the decoding of the denotation of the individual models it is referred to Ouzineb in [5–7]. The purchasing price, reliability and nominal performance capacity for the components corresponding to the upper listed systems are supposed to be known and can be retrieved from a list with technical data (excel sheets).

The algorithm starts by retrieving the data of the analyzed problem from the appropriate sheet and by random generating a so called initial population of size Pop_{size} that has been set to 100 chromosomes. The integer encoded chromosomes constituting the initial population have been generated in such a way that generated solutions that do not satisfy the given availability constraint are rejected and replaced by new acceptable ones in order to get a high qualitative start population. The constituting individuals or chromosomes have been created using a weighted pseudo random number generator which generates numbers between zero and the total number of components allowed in each subsystem, which have been set to ten. The probability of occurrence of 0 has been varied between 0.7 and 0.9 during the random chromosome generation process depending on the length of the chromosome corresponding to the analyzed problem. This limitation to the number of components allowed within a subsystem in addition to the rejection of non-feasible solutions. Furthermore a weighted generation of chromosome would limit the area that will be explored within the search space and should accelerate the search procedure towards convergence.

After evaluating and ranking populations (cost and availability estimation) chromosomes are selected to mate, recombine and finally mutate in order to build new offsprings that complete the next population of size Pop_{size}. This genetic procedure repeats until the predefined maximal number of generations N_{rep} is reached (Termination criterion). After completing each population through crossover and

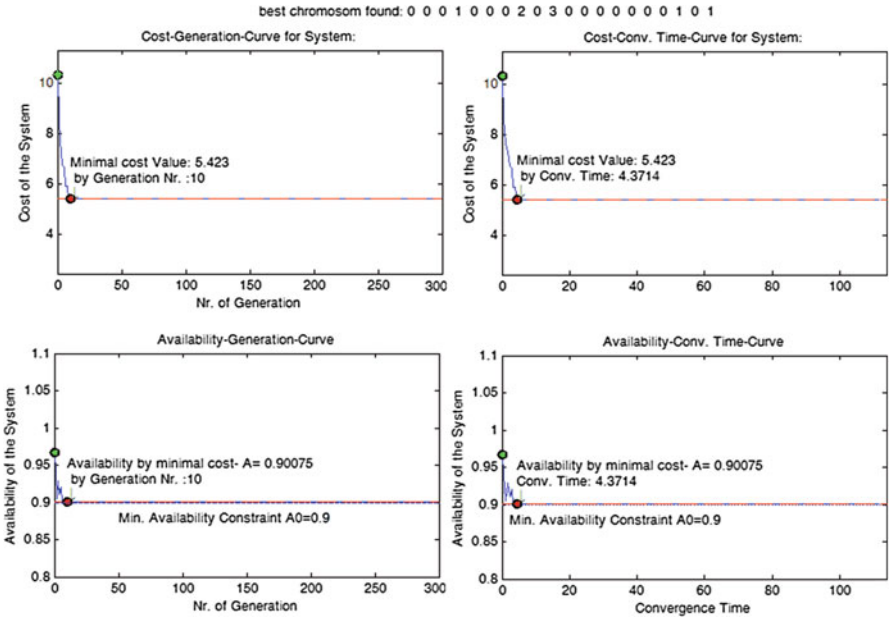


Fig. 16.8 Results of the genetic algorithm run of the heterogeneous case (Levitin—model containing four subsystems, availability constraint $A_0 = 0.900$ and 300 generations). The cost-generation and availability-generation curves dependencies are depicted on the left-hand side. The cost-time and availability-time dependencies are depicted on the right-hand side

mutation the population will be checked for multiplicity before evaluation by term of fitness function and new chromosomes would be generated to replace the chromosomes that appear more than once and have therefore been removed. This procedure of inserting new chromosomes to the population is compared to the act of inserting new genetic materials and may lead to new search areas that have not been explored or searched before and may accelerate the convergence speed. Figures 16.8 and 16.9 show the results of one run of the GA over the Lev4-(4/6)-3 model (heterogeneous case) data by predefined availability constraints of $A_0 = 0.900$ and $A_0 = 0.960$, and Fig. 16.10 shows the run of the GA over the Ouz6_(4/11)_4 by an availability constraint of $A_0 = 0.99$. The different plots show the evolution progress outgoing from the random initial population up to the predefined maximum number of generations. The best result (Cost—upper plots and Availability—lower plots), received after each genetic cycle, is depicted.

The time needed to find the best solution (convergence time) depends on the quality of the start population and on how the selected fittest chromosomes evolve throughout crossover and mutation.

On the left-hand side of Figs. 16.8, 16.9 and 16.10 (heterogeneous case) the best solution found (Top: cost value, bottom: availability value for found cost) during each generation is plotted against generation number whereas the same plots are

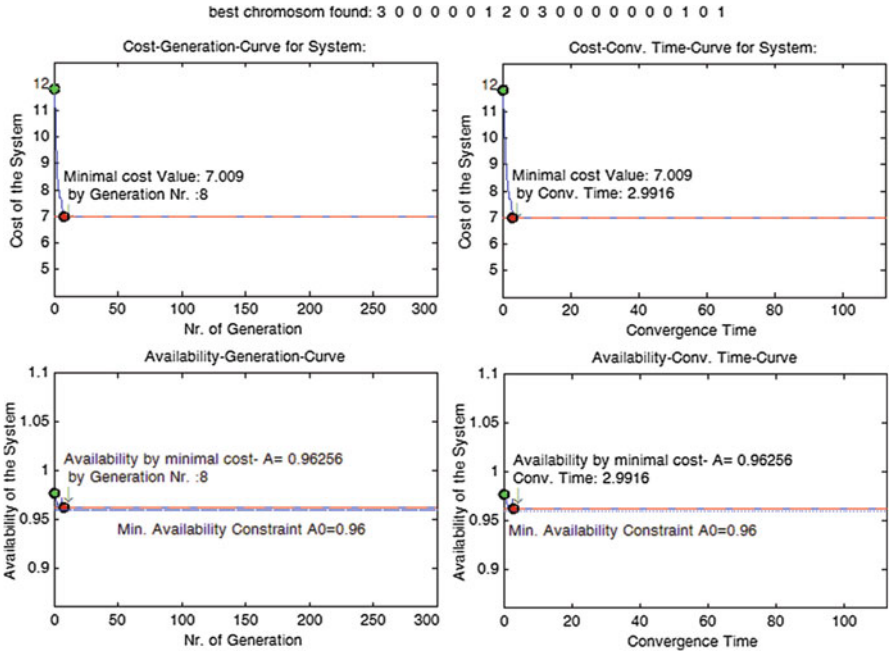


Fig. 16.9 Results of the genetic algorithm run of the heterogeneous case (Levitin—model containing four subsystems, availability constraint $A_0 = 0.960$ and 300 generations). The cost-generation and availability-generation curves dependencies are depicted on the left-hand side. The cost-time and availability-time dependencies are depicted on the right-hand side

represented on the right hand side against algorithm processing time. On the head of each plot the best chromosome or system design corresponding to the optimal (minimal) found cost subject to the given availability constraint is represented. In the context of the plots the generation number and convergence time are reported for which the best result has been identified.

Figures 16.11, 16.12 and 16.13 represent the homogeneous case of the heterogeneous problems analyzed successively in Figs. 16.8, 16.9, and 16.10. These figures have been included in order to show that through mixing of components lower or better system costs (Lev4-(4/6)-3—Model) can be reached in comparison to the homogeneous case subject to the same availability constraint. In the analyzed Ouz6_(4/11)_4 model no better cost results have been achieved but at least the same results as in the homogeneous case have been reached. One additional reason is to show that with the GA approach analyzed in this paper it was also possible to get the same results got with the hybridized GA + TS algorithm implemented in [5, 7]. This fact shows the effectiveness and accuracy of the GA approach discussed in this work since with the GA implemented in [6, 7] different results have been achieved.

The best test results got within 15 successive runs of the genetic algorithm over the different models mentioned previously are shown in Table 16.1. Computing and convergence time for best achieved results are also included in the same table and

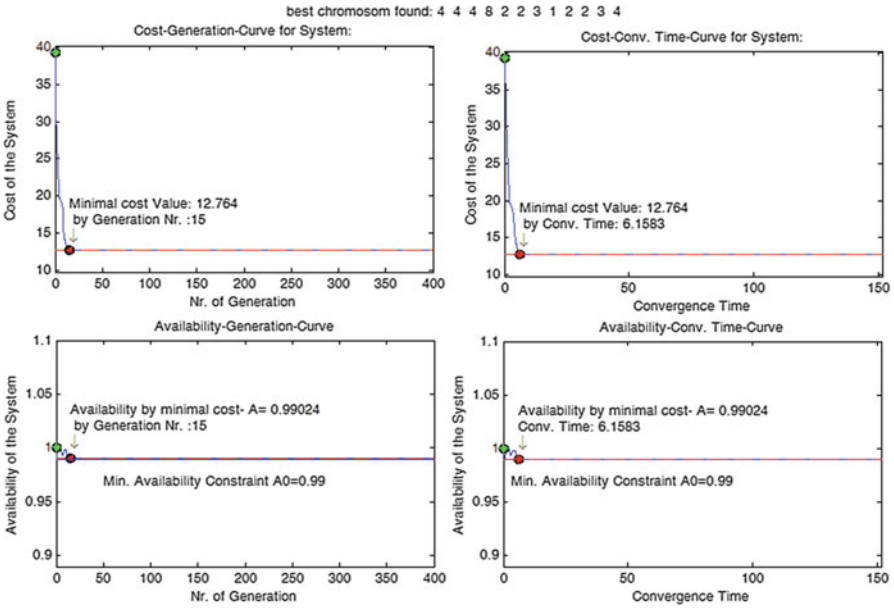


Fig. 16.10 Results of the genetic algorithm run of the heterogeneous case (Ouzineb—model containing six subsystems, availability constraint $A_0 = 0.990$ and 300 generations). The cost-generation and availability-generation curves dependencies are depicted on the left-hand side. The cost-time and availability-time dependencies are depicted on the right-hand side

serve to show how well the genetic algorithm customized for the heterogeneous case is performing in term of convergence speed, which can also be seen in the results depicted in Figs. 16.8, 16.9 and 16.10.

16.8 Conclusion and Future Works

Based on the facts and experimental results shown in the previous section and resumed in Table 16.1 for the different analyzed models, it can be recognized and concluded that the GA algorithm for heterogeneous series-parallel multi-states systems implemented in this work was performing in a great and efficient manner in term of convergence speed towards optimal results being expected and represented by Ouzineb in [5–7] and obtained by the hybrid GA + TS metaheuristic approach,

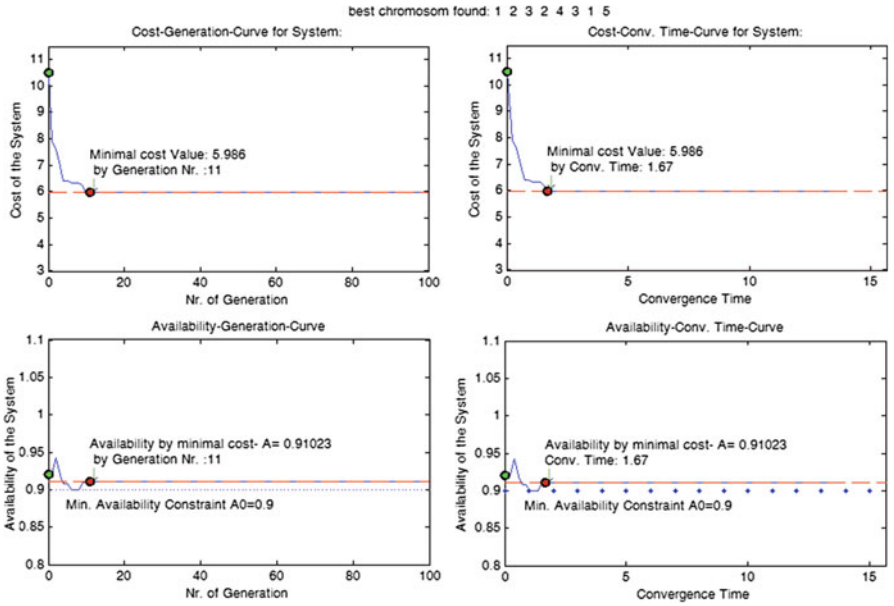


Fig. 16.11 Results of the genetic algorithm run of the homogeneous case for the same upper system (Levitin—model containing four subsystems) and subject to the same availability constraint $A_0 = 0.900$. As one can see in the heterogeneous case a better cost factor can be reached (5.423) than the homogeneous case (5.986) due to the fact that components have been mixed

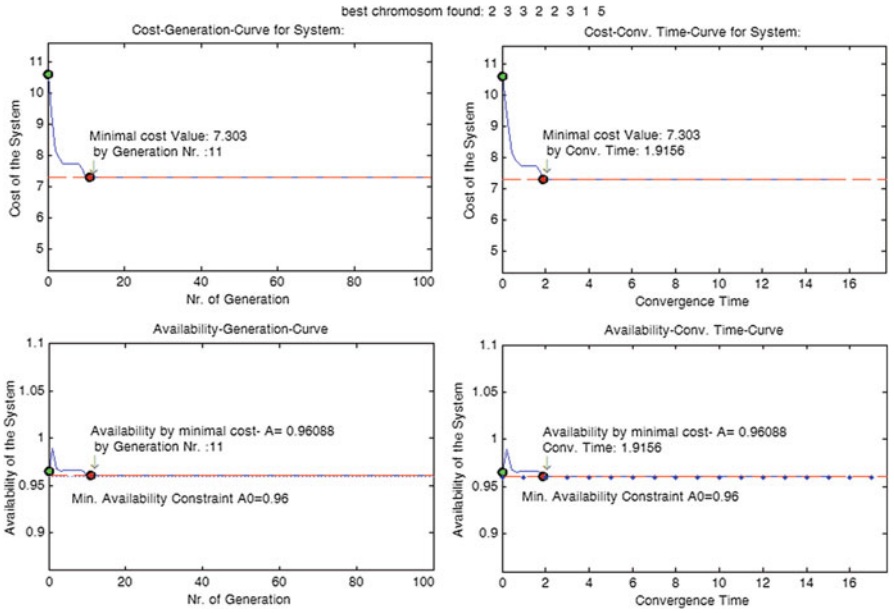


Fig. 16.12 Results of the genetic algorithm run of the homogeneous case for the same upper system (Levitin—model containing four subsystems) and subject to the same availability constraint $A0 = 0.960$. As one can see in the heterogeneous case a better cost factor can be reached (7.009) than the homogeneous case (7.303) due to the fact that components have been mixed

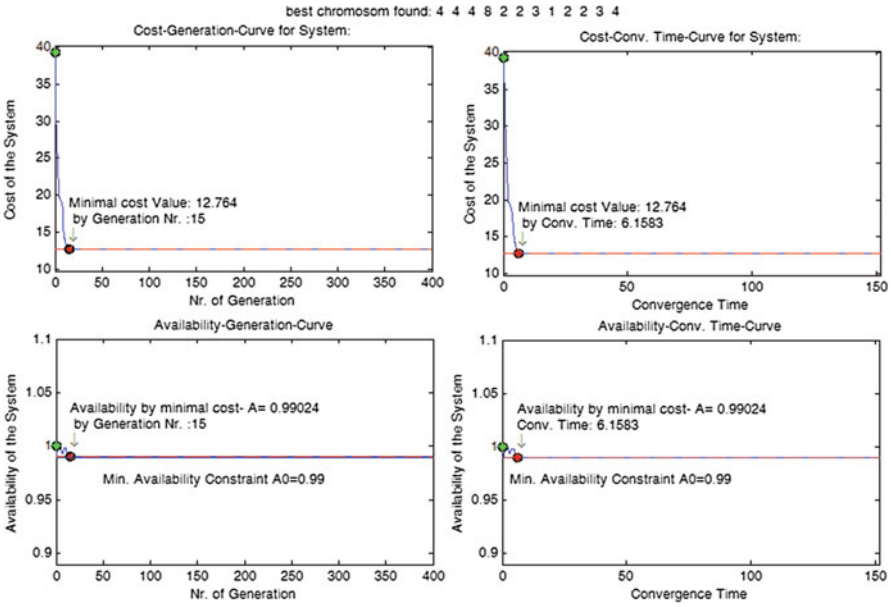


Fig. 16.13 Results of the genetic algorithm run of the homogeneous case for the same upper system (Ouzineb—model containing six subsystems) and subject to the same availability constraint $A0 = 0.990$. As one can see in both cases the same cost factor has been reached (12.764) but definitely for less computation time in the homogeneous case

in addition to the high accuracy in determining or finding the optimal solution (minimal system cost). And since genetic searching seems like searching for a small fish in a big ocean one small disadvantage or drawback is the standard one known in (heuristic) genetic approaches and that is resumed in the fact that the best optimal solution is not guaranteed or ensured in each run due to the limitation of the maximum number of iterations that may result, that some regions of the search or solution space that may include the optimal solution remains unexplored or out of reach.

The genetic approach implemented in this paper represents a very effective means in solving single objective constrained redundancy design problems like the complex heterogeneous one discussed in this work.

One of our future intentions is to tune genetic algorithms with local search algorithms targeting to increase the level of search accuracy. This kind of tuning is referred to in the literature as hybridization of genetic global searching algorithms.

Table 16.1 Computation results of the homogeneous and the heterogeneous case using the GA

Model name	Availability constraint A_0	Av. value $A(X,J)$ (hom.)	Av. value $A(X)$ (het.)	Cost $C(X,J)$ (*mln* \$) (hom.)	Cost $C(J(X))$ (*mln* \$) (het.)
lev4-(4/6)-3	0.900	0.9102	0.90075	5.986	5.423
	0.960	0.9609	0.96256	7.303	7.009
	0.990	0.9917	0.99148	8.328	8.180
lev5-(4/9)-4	0.975	0.9774	0.97615	16.450	12.855
	0.980	0.9808	0.98009	16.520	14.770
	0.990	0.9937	0.99211	17.050	15.870
ouz6-(4/11)-4	0.975	0.9790	0.9790	11.241	11.241
	0.980	0.9802	0.9802	11.369	11.369
	0.990	0.9902	0.9902	12.764	12.764

Model name	Best found chromosome homogeneous case [X,J]	Best found chromosome heterogeneous case J(X)	Conv. time (het.)
lev4-(4/6)-3	[1 2 3 2 4 3 1 5]	[4(1) 3(2) 1(3) 3(1),5(1)]	2.1992
	[2 3 3 2 2 3 1 5]	[1(3) 2(1),3(2) 1(3) 3(1),5(1)]	1.6700
	[3 3 3 5 1 3 1 2]	[1(3) 3(3) 1(3) 3(1),4(2)]	2.4501
lev5-(4/9)-4	[2 2 3 3 1 2 3 2 7 2]	[4(2),6(1) 5(6) 1(1),4(1) 7(3) 4(3)]	16.7701
	[2 6 3 3 1 2 5 2 7 2]	[4(2),6(1) 3(2) 2(1),3(2) 7(3) 3(2),4(1)]	14.0462
	[2 2 3 3 3 2 3 2 7 4]	[4(2),6(1) 3(2) 2(2),3(1) 7(3) 4(3)]	11.1849
ouz6-(4/11)-4	[4 4 5 7 2 1 3 1 2 2 3 4]	[3(4) 1(4) 2(5) 2(7) 3(2) 4(1)]	24.3273
	[4 5 5 8 2 1 3 1 2 2 3 4]	[3(4) 1(5) 2(5) 2(8) 3(2) 4(1)]	17.4778
	[4 4 4 8 2 2 3 1 2 2 3 4]	[3(4) 1(4) 2(4) 2(8) 3(2) 4(1)]	16.8972

It is important to mention at this point that the results got with the GA match either for homogeneous or for the non-homogeneous case the one got by Ouzineb using the hybrid GA + TS (Tabu Search) [5–7]. The best values of convergence time of the heterogeneous GA within 15 runs have been also included

References

1. Kuo, W., Rajendra Prasad, V., Tillman, F.A., Hwang, C.-L.: Optimal Reliability Design, Fundamentals and Applications. Cambridge University Press, Cambridge (2001)
2. Levitin, G., Lisnianski, A., Haim, H.B., Elmakis, D.: Genetic Algorithm and Universal Generating Function Technique for Solving Problems of Power System Reliability Optimization. The Israel Electric Corporation Ltd., Planning Development & Technology Division (2000)
3. Levitin, G., Lisnianski, A., Haim, H.B.: Redundancy optimization for series-parallel multi state systems. IEEE Trans. Reliab. **47**(2) (1998)
4. Lisnianski, A., Livitin, G., Haim, H.B., Elmakis, D.: Power system optimization subject to reliability constraints. Electr. Power Syst. Res. **39**, 145–152 (1996)
5. Ouzineb, M.: Heuristiques éfficaces pour l'optimisation de la performance des systèmes séries-parallèles. Département d'informatique et de recherche opérationnelle Faculté des arts et des sciences, Université de Montréal, 2009
6. Ouzineb, M., Nourelfath, M., Gendreau, M.: Tabu search for the redundancy allocation problem of homogenous series-parallel multi-state systems. Reliab. Eng. Syst. Saf. **93**, 1257–1272 (2008)
7. Ouzineb, M., Nourelfath, M., Gendreau, M.: A heuristic method for non-homogeneous redundancy optimization of series-parallel multi-state systems. J. Heuristics **17**(1), 1–22 (2009)
8. Yalaoui, A., Chu, C., Châtelet, E.: Reliability allocation problem in a series–parallel system. Reliab. Eng. Syst. Saf. **90**, 55–61 (2005)
9. Li, C.-y., Chen, X., Yi, X.-s., Tao, J.-y.: Heterogeneous redundancy optimization for multi-state series–parallel systems subject to common cause failures. Reliab. Eng. Syst. Saf. **95**, 202–207 (2010)
10. Chaaban, W., Schwarz, M., Börcsök, J.: Budgetary and redundancy optimisation of homogeneous series-parallel systems subject to availability constraints using Matlab implemented genetic computing. In: 24th IET Irish, Signals and Systems Conference (ISSC 2013)
11. Holland, J.: Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor (1975)
12. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, Reading (1989)
13. Tian, Z., Zuo, M.J., Huang, H.: Reliability-redundancy allocation for multi-state series-parallel systems. IEEE Trans. Reliab. **57**(2), 303–310 (2008)
14. Affenzeller, M., Winkler, S., Wagner, S., Beham, A.: Genetic Algorithms and Genetic Programming, Modern Concepts and Applications. CRC Press, Boca Raton (2009)
15. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs, 3rd revised and extended edition. Springer, Berlin (2011)
16. Tillman, F.A., Hwang, C.-L., Kuo, W.: Optimization techniques for system reliability with redundancy—a review. IEEE Trans. Reliab. **R-26**(3), 148–155 (1977)
17. Levitin, G.: The Universal Generating Function in Reliability Analysis and Optimization. Springer, London (2005)
18. Chaaban, W., Schwarz, M., Börcsök, J.: Cost and redundancy optimization of homogeneous series-parallel multi-state systems subject to availability constraints using a Matlab implemented genetic algorithm. In: Recent Advances in Circuits, Systems and Automatic Control, WSEAS 2013, Budapest, Hungary, 2013

Chapter 17
Random Hypernets in Reliability Analysis of Multilayer Networks

Alexey Rodionov and Olga Rodionova

Abstract The general approach to constructing structural models of non-stable multi-level networks is proposed. This approach is based on hypernets—relatively new mathematical object, which is successively used for modeling different multi-level networks in the Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia, for last 30 years. Hypernets allow standard description of neighboring levels interconnection in a mathematically correct way. Using this mathematical object allows easy modifications of data with model changing and/or development and efficiently organize data search for different computational or optimization algorithms. Optimization of mapping of secondary (logical) network onto structure of unreliable primary (physical) network is considered as example.

Keywords Multilevel networks • Modeling • Hypernets • Reliability analysis

17.1 Introduction

Reliability is one of major indexes of networks' and other complex systems' quality of service [1–3]. Random graphs and hypergraphs are usually used for the reliability analysis of unreliable networks [4–9] of different nature and destination (data transferring, public transport, pipeline systems, etc.). This model seems quite appropriate as a structural model for analysis of structural and functional reliability of information networks: failures of nodes or links are simulated by removal of correspondent vertexes or edges (arcs) of a random graph with given probabilities, or by reducing their throughput. At the same time, in many cases modeling network

A. Rodionov (✉)
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
Prospect Akademika Lavrentjeva 6, Novosibirsk 630090, Russia
e-mail: alrod@sscc.ru
http://www.sscc.ru/

O. Rodionova
Higher College of Informatics of the Novosibirsk State University,
35 Russkaya St., Novosibirsk 630058, Russia
e-mail: rolcon@mail.ru

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_17

by a random graph is not sufficient, which can be shown by the following example of two-level network. Let us have a cable network that is presented by a graph G_1, and let us have some data-transferring network realized inside it. This network can be presented by some graph G_2, that not necessarily coincides with G_1. We will name G_1 as primary network (PN) while G_2 we will name secondary network (SN). Laying of G_2 into G_1 may be done by different ways (if G_1 is not a tree). Thus, we have some mapping of links of *SN* onto paths constructed from links of *PN* (further, we will name these links as branches). Obviously, failure or change of throughput of a branch may lead to failure or change of throughput of several *SN*'s links or may not touch any of them at all. So, for analyzing multi-level networks more complicated models than random graphs are needed.

We can find different descriptions of such models: bigraphs [10], sandwiching graphs [11], graphs with different kinds of edges [12], descriptions on the application level [13–15], layered complex networks (LCN) [16], etc. All these models (may be excluding LCN) are not universal, but all of them take into account different connections between layers. Even LCN model consider mapping of neighbor layers only. Yet for more than 30 years, the hypernet model is successively used for modeling multi-layer embedded networks of different nature in several Russian, Kirghiz and Kazakh universities. Unfortunately, until now most of papers and books with description of the model and its applications are in Russian (main monograph is [17]), but there are some conference publications in English in which the hypernet models are used also [18–20]. Hypernets allow adequately describe multilevel networks with an arbitrary number of levels. In this paper, we discuss the simplest case of two-level networks and its usage for reliability analysis and optimization.

17.2 Random Hypernet Model Specification

General description of a hypernet model is given in [17]. Here concept of abstract hypernet is formally presented that describes multi-level network in general case; each layer is presented by a hypergraph. In partial case, hypergraphs retrograde to graphs and we have simple multi-level hypernet. For the purpose of this paper, the concept of two-level hypernet or simply hypernet is enough.

Definition: Hypernet $H = (X, V, R; P, F, W)$ consists of the following objects (see Fig. 17.1):

$X = (x_1, \ldots, x_n)$—the set of vertexes;
$V = (v_1, \ldots, v_m)$— the set of branches (edges of the graph of a primary network);
$R = (r_1, \ldots, r_g)$— the set of edges (edges of the graph of a secondary network);
$P : V \to X \times X$—the mapping that defines graph $PN = (X, V)$ named a primary network;
$W : R \to X \times X$—the mapping that defines graph $SN = (X, R)$ named a secondary network;

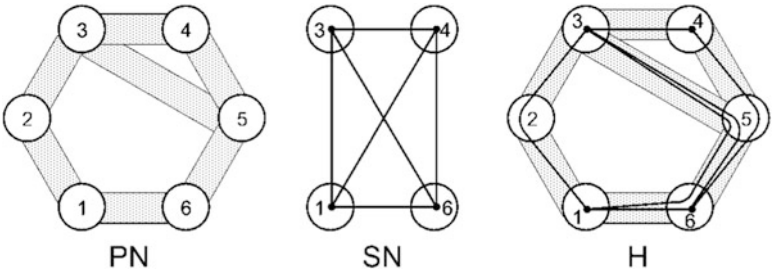
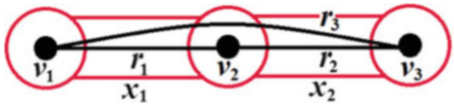


Fig. 17.1 Illustration to the hypernet definition: *PN* is the primary network; the shadowed vertexes $\{1, 3, 4, 6\}$ form the set of vertexes that belongs to *SN*; *SN* is the secondary network (that is a complete graph in our case); H is the hypernet (SN is mapped to the *PN*)

Fig. 17.2 Three-vertex simple hypernet



F—the mapping that assigns to each element $r \in R$ the set $F(r) \subseteq V$ of its branches (route in graph $PN = (X, V)$).

The incidence and adjacency in *PN* and *SN* are defined similar to those for graphs, while mapping F gives these concepts for a hypernet in a whole. F may be presented by special adjacency matrices that describe adjacency of edges (v-adjacency if edges are incidental to one vertex and x-adjacency if edges lays in one branch). If edge goes *through* a vertex without being incidental to it, then we have *weak* incidence of these vertex and edge.

Let us have a hypernet presented in the Fig. 17.2, here *PN* is the 3-vertex chain and *SN* is the 3-node complete graph. r_3 is incidental to v_1 and v_3 and weakly incidental to v_2. Embedding of *SN* into *PN* is presented by the following matrices of adjacency between branches and edges and vertexes and edges:

$$XR = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}; \qquad VR = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

In many cases several secondary networks are embedded in a single primary network, for example, one cable network may be used for public phone network, cable TV and Internet; several independent working groups can use the same LAN and so on. In other models it may be needed place several nodes of a *SN* into one node (vertex) of a *PN*. For modeling such situations the extension of hypernet named S-hypernet is proposed in [21]. S-hypernet allows embedding of several hypernets of one level into one hypernet of another level. In partial case several *SNs* presented by graphs may be embedded into single *PN*. The example of such embedding is shown in Fig. 17.3.

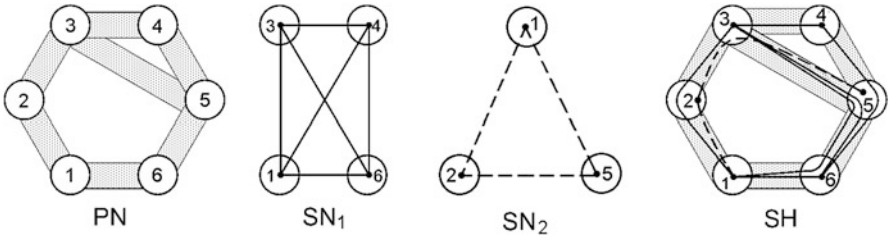


Fig. 17.3 Illustration to the S-hypernet definition: *PN* is the primary network; vertexes {1,3,4,6} form the set of vertexes that belongs to SN_1 and vertexes {1,2,5} form the set of vertexes that belongs to SN_2, both are complete graphs; *SH* is the S-hypernet (SNs are mapped onto the *PN*)

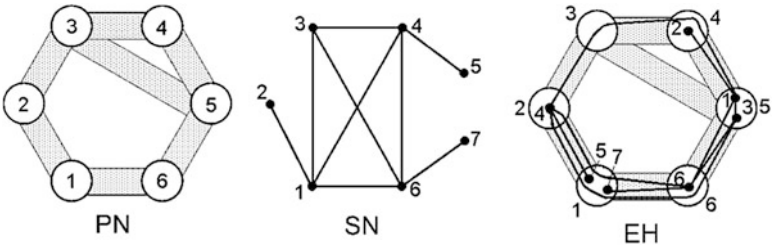


Fig. 17.4 Illustration to the E-hypernet definition

Another special case corresponds to different sets of vertexes of *PN* and *SN*. This situation is quite common and we name corresponding S-hypernet as E-hypernet (Enhanced hypernet). Dedication of this special class of S-hypernets allows decrease complexity of a model presentation while it allows adequately describe a lot of networks (for example, several switches of different networks may be installed in one control cabinet of a structural cable network, several bus stops may be located on different sides of a square, etc.). The example of E-hypernet is presented in Fig. 17.4.

Representation of an E-hypernet requires additional adjacency matrix between vertexes of *PN* and nodes of *SN*. Thus the hypernet in the Fig. 17.4 (in addition to presentation of its *PN* and *SN*) may be presented by the adjacency matrices

$$XR=\begin{pmatrix}0&0&0&0&1&0&1&1&0\\0&0&0&0&1&0&0&1&1\\0&0&1&0&0&0&0&0&0\\0&0&1&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\\1&0&1&0&0&0&0&0&0\\0&0&0&1&1&1&0&0&0\end{pmatrix};\ VR=\begin{pmatrix}0&0&0&0&1&0&1&1&1\\0&0&1&0&1&0&1&1&0\\0&0&1&0&0&0&0&0&0\\1&0&1&0&0&0&0&0&0\\1&1&1&1&1&1&0&0&0\\0&0&0&1&1&1&0&1&1\end{pmatrix};\ VN=\begin{pmatrix}0&0&0&0&1&0&1\\0&0&0&1&0&0&0\\0&0&0&0&0&0&0\\0&1&0&0&0&0&0\\1&0&1&0&0&0&0\\0&0&0&0&0&1&0\end{pmatrix}.$$

Here branches and edges are enumerated in lexicographical order, that is x_1 connects vertexes 1 and 2, x_2 connects vertexes 1 and 6, x_3 connects vertexes 2 and 3, and so on.

Random hypernet is random realization of feasible sextuple H. We consider the following cases:

1. Fixed *PN* with given probabilities of existence of v_i, that corresponds to unreliable physical network with reliable nodes and unreliable links; *SN* and F are fixed. This model suits for the case of long-term interconnections.
2. *PN* and *SN* are fixed, F is random (randomly chosen feasible mapping). This model better suits case of short-term interconnections.

First model allows analyze probabilistic connectivity of a logical network at possible failures of channels of a physical one, which may occur as because of natural reasons, as anthropogenic ones.

17.3 Solving Reliability Problems with Hypernets

There are possible different kinds of failures in a hypernet. They are presented in the Fig. 17.5.

If we consider a logical network as reliable one, then its failures is possible due to failures of a physical one. Thus for calculating its reliability we must search all possible destructions of a *PN* and consider corresponding states of a *SN*. For example, failure of the branch that connects vertexes 3 and 5 in the Fig. 17.4 does not influent on the state of *SN* at all, while destruction of the branch that connects vertexes 1 and 6 destroys three edges of *SN* (3–4, 4–6 and 6–7) making *SN* disconnected.

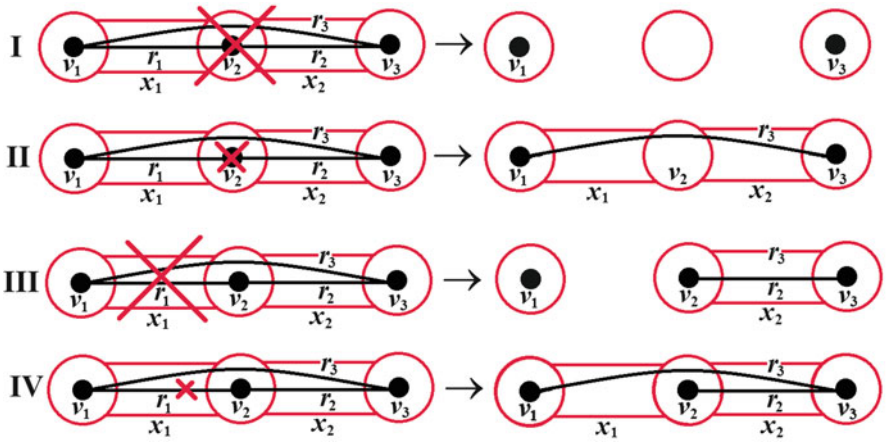


Fig. 17.5 Possible kinds of destructions in a hypernet

In [22] we discuss some improvements of well-known factoring method [23] for calculating reliability of a random graph. These improvements could be adopted for calculating reliability of *SN*. Factoring is executed by states of unreliable elements of *PN* (branches in our case) while connectivity of *SN* is checked. Thus, we have:

$$R(SN) = p_{ij} R(SN|v_{ij} \text{ works}) + (1 - p_{ij}) R(SN|v_{ij} \text{ fails}),$$

where p_{ij} is a reliability of a branch $v_i j$. When calculating first summand, nodes x_i and x_j are contracted into one new node, while when calculating second summand the branch v_{ij} is simply removed. On the other hand, in some cases the following direct equation may be more convenient:

$$R(SN) = \sum_{i=0}^{g} \sum_{j=1}^{C_g^i} A_{ij} I(SN_{ij}).$$

Here A_{ij} is a probability of an event corresponding to realization of j-th mode of removal exactly i branches from *PN*, SN_{ij}—corresponding to this event remaining part of *SN*, and $I(SN)$—indicator function, 1 if *SN* is connected and 0 otherwise. For highly reliable branches this equation may be used for lower approximation of *SN*'s reliability by stopping summation at some $Vg \leq g$. If reliabilities of all branches are equal to some p, we can obtain a reliability polynomial for *SN*:

$$R(SN, p) = \sum_{i=0}^{g} B_i \, p^{g-i} (1 - p)^i,$$

where B_i—number of connected realizations of *SN* when exactly i branches are removed from *PN*. Examination of this polynomial may help in choice of best mapping of *SN* onto unreliable *PN*.

Two possible mappings of a given *SN* onto cyclic *PN* are presented in the Fig. 17.6, E-hypernet is used as a model. Mappings differs in placement of *SN*'s nodes into nodes of *PN*, shortest paths realize edges. Reliability polynomials are easily obtained by exhaustive search of all possible destructions of the *PN*:

$$R(SN|H_1) = p^5 + 2p^4(1 - p); \quad R(SN|H_2) = p^5 + 3p^4(1 - p).$$

Comparison of polynomials shows that second mapping is better for all values of p (see Fig. 17.7).

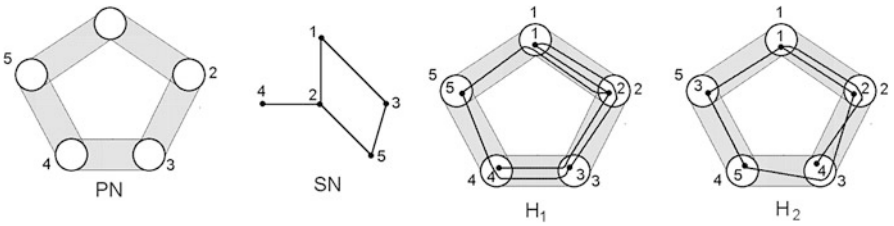


Fig. 17.6 Two possible mappings of *SN* onto *PN*

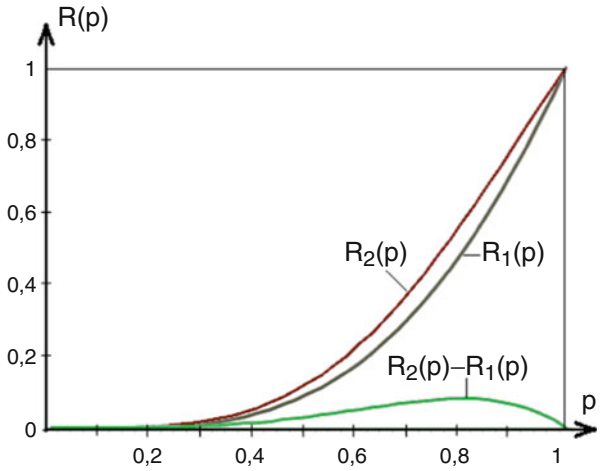


Fig. 17.7 Reliability polynomials for *SN* in case of different mappings

Second model (random mapping) is harder to analyze as it assumes averaging of reliability indices by all possible mappings and so requires exhaustive search for exact calculation or Monte-Carlo method for approximate one. In the last case, using Hypernet model allows unification of data presentation and storage. When modeling one must take into account that if there are some restrictions on *PN* or *SN*, then mapping is not always possible. For example, throughput limitations on branches of *PN* may not allow realize all edges of *SN*, or there may be limitation on a number of branches realizing an edge and so on.

17.4 Conclusion

In this paper, we discuss possible usage of the Hypernet model for analyzing multi-layer network reliability. Possible advantages of using this model are greater. First to all, it gives general means for describing multi-layer network structure, and then it allows designing algorithms that are more effective and solving problems that are

impossible or hard model by other means. Our further researches in this direction concerns constructing models of multi-service sensor networks and networks in which several networks of upper level may exist inside one network of lower one.

Acknowledgements This work is supported by the grant of the Program of basic researches of the Presidium of Russian Academy of Science.

References

1. Jain, M., Chand, S.: On connectivity of ad hoc network using fuzzy logic. In: Proceedings of the 2014 International Conference on Applied Mathematics and Computational Methods in Engineering II (AMCME '14) and the 2014 International Conference on Economics and Business Administration II (EBA '14), pp. 159–165 (2014)
2. Seytnazarov, S., Kim, Y.-T.: QoS-aware MPDU aggregation of IEEE 802.11n WLANs for VoIP services. In: Proceedings of the 2014 International Conference on Electronics and Communication Systems II (ECS '14) and the 2014 International Conference on Education and Educational Technologies II (EET '14), pp. 64–71 (2014)
3. Mosharraf, N., Khayyambashi, M.R.: Improving performance and reliability of adaptive fault tolerance structure in distributed real time systems. Comput. Simul. Mod. Sci. **3**, 133–143 (2010)
4. Waxman, B.M.: Routing of multipoint connections. IEEE J. Sel. A. Commun. **6**(9), 1617–1622 (2006)
5. Doar, M.: Multicast in the ATM environment. Ph.D. Thesis, Cambridge University, Computer Lab (1993)
6. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E.: Stochastic models for the Web graph. In: Proceedings 41st Annual Symposium on Foundations of Computer Science, pp. 57–65 (2000)
7. Albert, R., Barabasi A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
8. Yano, A., Wadayama, T.: Probabilistic analysis of the network reliability problem on a random graph ensemble. http://arxiv.org/pdf/1105.5903.pdf (2011) [arXiv:1105.5903v3]
9. Bobbio, A., Terruggia, R., Ciancamerla, E., Minichino, M.: Evaluating Network Reliability Versus Topology by Means of BDD Algorithms. PSAM-9, Hong Kong (2008)
10. Milner, R.: Bigraphs, a Tutorial, at http://www.cl.cam.ac.uk/users/rm135 (2005)
11. Kim, J.H., Vu, V.: Sandwiching random graphs. Adv. Math. **188**, 444–469 (2004)
12. Dijkstra, F., Andree, B., Koymans, K., van der Hama, J., Grosso, P., de Laat, C.: A multi-layer network model based on ITU-T G.805. Comput. Netw. **52**, 1927–1937 (2008)
13. He, F., Xin, C.: Cross-layer path computation for dynamic traffic grooming in mesh WDM optical networks. Technical Report #NSUCS-2004-009, Norfolk State University (2004)
14. Koster, A.M.C.A., Orlowski, S., Raack, C., Baier, G., Engel, T., Belotti, P.: Branch-and-cut techniques for solving realistic two-layer network design problems. In: Graphs and Algorithms in Communication Networks, pp. 95–118. Springer, Heidelberg (2009)
15. Chigan, C., Atkinson, G., Nagarajan, R.: On the modeling issue of joint cross-layer network protection/restoration. In: Proceedings of Advanced Simulation Technologies Conference (ASTC '04), pp. 57–62 (2004)
16. Kurant, M., Thiran, P.: Layered complex networks. Phys. Rev. Lett. **96**, 138701-1–138701-4 (2006)
17. Popkov, V.K.: Mathematical models of connectivity. Inst. Comp. Math. Math. Geophys. Novosibirsk (2006) (in Russian)

18. Popkov, V.K., Sokolova, O.D.: Application of hyperneet theory for the networks optimazation problems. In: 17th IMACS World Congress, July 2005, Paper T4-I-42-011 (2005)
19. Rodionov, A.S., Sokolova, O., Yurgenson, A., Choo, H.: On Optimal placement of the monitoring devices on channels of communication network. In: ICCSA 2009, Part II. Lecture Notes in Computer Science, vol. 5593, pp. 465–478 (2009)
20. Rodionov, A.S., Choo, H., Nechunaeva, K.A.: Framework for biologically inspired graph optimization. In: Proceedings of ICUIMC 2011, Seoul, Paper 2.5 (2011)
21. Popkov, V.K.: Using s-hypernet theory for modeling systems with network structure. Probl. Inf. **4**, 17–40 (2010) (in Russian)
22. Rodionova, O.K., Rodionov, A.S., Choo, H.: Network probabilistic connectivity: Exact calculation with use of chains. In: ICCSA-2004. Springer Lecture Notes in Computer Science, vol. 3046, pp. 315–324 (2004)
23. Satyanarayana, A., Chang, M.K.: Network reliability and the factoring theorem. Networks **13**, 107–120 (1983)

Chapter 18
Profiling Power Analysis Attack Based on Multi-layer Perceptron Network

Zdenek Martinasek, Lukas Malina, and Krisztina Trasy

Abstract In 2013, an innovative method of power analysis was presented in Martinasek and Zeman (Radioengineering **22**(2), IF 0.687, 2013) and Martinasek et al. (Smart Card Research and Advanced Applications. Lecture Notes in Computer Science. Springer International Publishing, New York, 2014). Realized experiments proved that the proposed method based on Multi-Layer Perceptron (MLP) can provide almost 100 % success rate. This description based on the first-order success rate is not appropriate enough. Moreover, the above mentioned works contain other lacks: the MLP has not been compared with other well-known attacks, an adversary uses too many points of power trace and a general description of the MLP method was not provided. In this paper, we eliminate these weaknesses by introducing the first fair comparison of power analysis attacks based on the MLP and templates. The comparison is accomplished by using the identical data sets, number of interesting points and guessing entropy as a metric. The first data set created contains the power traces of an unprotected AES implementation in order to classify the secret key stored. The second and third data sets were created independently from public available power traces corresponding to a masked AES implementation (DPA Contest v4). Secret offset is revealed depending on the number of interesting points and power traces in this experiment. Moreover, we create a general description of the MLP attack.

Keywords Power analysis • MLP • Machine learning • Template attack • Comparison

Z. Martinasek (✉) • L. Malina
Department of Telecommunications, Brno University of Technology
Technicka 12, 612 00 Brno, Czech Republic
e-mail: martinasek@feec.vutbr.cz

K. Trasy
Department of Garden and Landscape Architecture, Mendel University in Brno
Valticka 337, 691 44 Lednice, Czech Republic

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_18

18.1 Introduction

Power analysis (PA) measures and analyzes the power consumption of cryptographic devices depending on their activity. The goal of PA is to determine the sensitive information (mostly secret key stored) of cryptographic devices from the measured power consumption and to apply the obtained information in order to abuse the cryptographic device. This whole process is called power analysis attack. These types of attacks represent extremely effective and successful way of attacks on so far confidential cryptographic algorithms such as AES (Advanced Encryption Standard [1–4]), RSA (Rivest Shamir Adleman) [5, 6] and cryptographic devices such as smart cards [7–9]. Power analysis was introduced by Kocher in [10] and generally includes two basic methods: simple PA and differential PA. An adversary tries to determine the secret key directly from the traces measured in the simple power analysis (SPA). In the most extreme case, this means that the adversary attempts to reveal the key based on one single power trace. The goal of the differential power analysis (DPA) attacks is to reveal the secret key of the cryptographic device by using a large number of power traces that have been recorded while the device was encrypting or decrypting various input data.

To prevent power analysis attacks, one can implement some of the countermeasure techniques. The goal of every countermeasure is to create the power consumption of a cryptographic device independent of intermediate values that are processed. Generally countermeasure techniques are divided into two basic groups, hiding and masking. Masking randomizes each intermediate values by adding random values called masks. One of the widespread countermeasures represents Boolean masking [11, 12]. By contrast, hiding tries to break the link between the power consumption and the data values processed [10, 13–15]. A detailed description of power analysis including side-channel sources, testbeds, statistical and countermeasures is summarized in the book [9].

18.1.1 Related Work

A typical example of SPA is the attack on the implementation of the RSA asymmetric cryptographic algorithm, where the difference in power consumption between the operations of multiplication and squaring can be observed [5]. Template based attacks were introduced in [16] and can be considered as the strongest leakage analysis in an information theoretic sense. Practical aspects of template attacks (TA) have been discussed in [7, 17, 18]. The concept of the DPA attack was first described in [10] and the basic principle was introduced on a DES algorithm using the statistical method based on the Difference of Means. Nowadays, DPA based on correlation coefficient is one of the most widely used methods [19].

Application of neural networks in the field of power analysis was first published in [20]. Naturally, this work was followed by other authors, e.g. [21, 22], who

dealt with the classification of individual power prints. These works are mostly oriented towards reverse engineering. Yang et al. [23] proposed MLP in order to create a power consumption model of a cryptographic device in DPA based on correlation coefficient. In recent years, the cryptographic community has explored new approaches based on machine learning models. Lerman et al. [24, 25] compared a template attack (TA) with a binary machine learning approach based on non-parametric methods. Hospodar et al.[26, 27] analysed the Suprot Vector Machine (SVM) on a software implementation of a block cipher. Heuser et al. [28] created the general description of the SVM attack and compared this approach with the template attack. In 2013, Bartkewitz [29] applied a multi-class machine learning model that improves the attack success rate with respect to the binary approach. Recently, Lerman et al. [30] proposed a machine learning approach that takes into account the temporal dependencies between power values. This method improves the success rate of an attack in a low signal-to-noise ratio with respect to classification methods. Lerman et al. [31] presented a machine learning attack against a masking countermeasure, using the dataset of the DPA Contest v4. Interesting method of power analysis based on a multi-layer perceptron was first presented in [32]. In this work, the authors used a neural network directly for the classification of the AES secret key. In [33], this MLP approach was optimized by using the preprocessing of the power traces measured.

18.1.2 Contribution

In [32, 33], the authors used the first-order success rate for efficiency description of the proposed MLP power analysis method. This is not sufficiently reliable because this value can be deceiving [34]. According the framework, the guessing entropy represents an appropriate metric of two side analysis attack implementation [34]. The metric measures the average number of key candidates to test after the side-channel attack.

The other important fact is that both methods based on the MLP (original implementation and optimized one) have not been yet compared with other well-known approaches such as the template attack or the stochastic attack. In previous researches described in [32, 33], the adversary uses 1,200 interesting points to realize the attack. This large number of interesting points is not practically applicable to TA because of possible numerical problems connected with a covariance matrix.[1] In this paper, we introduce the first fair comparison of power analysis attacks based on the MLP and templates. For a first time, the comparison is accomplished by

[1]The size of the covariance matrix grows quadratically with the number of points in the trace, more information in [9].

using the identical data set including a number of interesting points. Moreover, we create a general description of the MLP aimed for byte classification including the structure, setting and training algorithm, because this information was also missing in previous research.

Our research was based on three datasets, the first dataset (DS1) contains the power traces of an unprotected AES implementation where we classify one byte of the secret key and whole secret key subsequently. We measured and prepared the first data set in our lobar using our testbed [35, 36]. The second (DS2) and the third datasets (DS3) were independently prepared from public datasets of power traces corresponding to the masked AES implementation (DPA Contest v4 [37]) where we classify the secret offset. In this experiment we compare the power analysis attacks efficiency depending on the number of interesting points and the number of power traces. We chose public available power traces to obtain independent datasets corresponding to the different cryptographic device. It is clear, that making fair comparison based only on own power traces is not trusted. The second data set was created by our research group and the third data set was created by Liran Lerman during preparation of the attack in [31].

18.2 General Description of the MLP

This section provides only a basic information about the neural networks that we used during the attack (the basic structure and the training algorithm of the MLP). We refer the work [38, 39] for more specific information. The main goal of this section is to show how to use MLP to realize the power analysis attack (analogy to attack based on templates).

Preliminaries: a learning set $\mathbf{Y}$ (sometimes denoted as a training set) and a test set $\mathbf{X}$ with n and m instance represents power traces measured in the context of the side channel analysis. Each instance $\mathbf{y}_i$ where $i = 1, \ldots, n$ and $\mathbf{x}_j$ where $j = 1, \ldots, m$ in the learning and training set contains one assignment (a class label which determines which class the concrete instant belongs to) and several attributes $\mathbf{y}_i = y_1, \ldots, y_N, \mathbf{x}_j = x_1, \ldots, x_N$ (features or observed variables). These attributes represent the points of power traces in time (samples). Mostly, one chooses only some interesting points that leak information [7, 9, 18, 40] which the attack is aimed on. The learning set is used in the profiling phase of the profiled attack and test set is used during the attack phase.

The basic element of an artificial neural network is a formal neuron, often called as a perceptron in the literature. The basic model of the neuron is shown on the left side in Fig. 18.1. The neuron contains x_i inputs that are multiplied by the weights w_i, where $i = 1$ to n. Input x_0 multiplied by the weight $w_0 = -\theta$ determines the threshold of the neuron (bias). During the training of the neuron, weights are updated to achieve a desired output value. Firstly, a post-synaptic potential is calculated. It is defined as the internal function of the neuron:

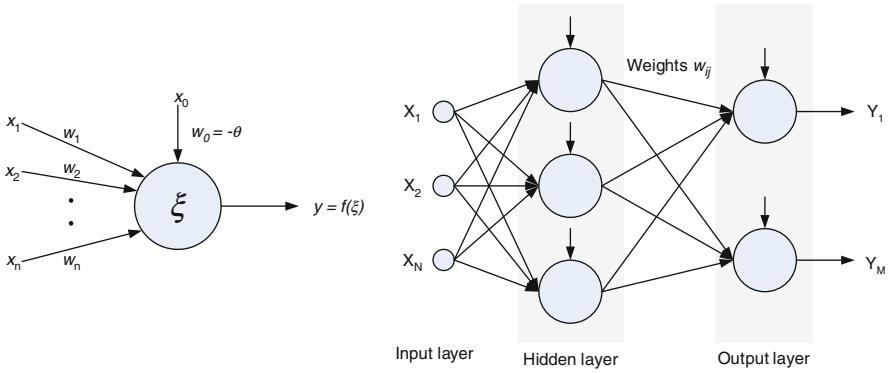


Fig. 18.1 The general structure of neural network

$$\xi = \sum_{i=1}^{n} x_i w_i - \theta. \tag{18.1}$$

Subsequently, the output value of the neuron is calculated as $y = f(\xi)$ where f represents a non-linear function, mostly a sigmoid. Naturally, one formal neuron is not able to solve complex problems, therefore we use neurons (perceptrons) connected into a network. The multilayer perceptron consists of two or more layers of neurons that are denoted as an output layer and a hidden layer. Each neuron in one layer is connected with a certain weight w_{ij} to every neuron in the following layer. Frequently, the input layer is not included when one is counting the number of layers because the input layer is not composed of neurons. We follow this notation in this article. An example of the two-layer neural network is shown in Fig. 18.1 (on the right side).

These networks are modifications of the standard linear perceptron and can distinguish data that are not linearly separable [39]. These networks are widely used for a pattern classification, recognition, prediction and approximation and utilize mostly a supervised learning method called backpropagation [41]. The backpropagation (BPG) algorithm is an iterative gradient learning algorithm which minimizes squares of a cost function using the adaptation of the synaptic weights. This method is described with the following steps (the following equations are valid for the two-layer neural network which is shown in Fig. 18.1):

- Step 1: Weights w_{ij} and thresholds θ of each neuron are initialized with random values.
- Step 2: An input vector $\mathbf{X} = [x_1, \ldots, x_N]^T$ and a desired output vector $\mathbf{D} = [d_1, \ldots, d_M]^T$ are applied to the neural network. In other words, one creates a training set containing pairs of $\mathbf{T} = \{[\mathbf{X_1}, \mathbf{D_1}], [\mathbf{X_2}, \mathbf{D_2}], \cdots, [\mathbf{X_n}, \mathbf{D_n}]\}$, where n denotes the number of training set patterns and the training set prepared is applied to the neural network. Provided, that NN represents an ordinary classifier

which classifies input data to the desired output groups, the **D** represents mostly a classification matrix where the desired outputs are labelled by value 1 and other outputs 0.

– Step 3: The current output of each neuron is calculated by the following equations:

$$y_k(t) = f_s(\sum_{k=1}^{N_1} w'_{jk}(t)x'_j(t) - \theta'_k), \tag{18.2}$$

$$x'_j(t) = f_s(\sum_{i=1}^{N} w_{ij}(t)x_i(t) - \theta_j), \tag{18.3}$$

where $1 \leq k \leq M$ denotes output layer and $1 \leq j \leq N_1$ hidden layer.

– Step 4: Weights and thresholds are applied according to the following equation:

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_j x_i. \tag{18.4}$$

Adaptation of weight values starts at the output neurons and proceeds recursively back to the input neurons. In this equation, w_{ij} denotes weights between the ith hidden or input neuron and the neuron jth at time t. Output of the ith neuron is denoted as x'_i, η represents the learning coefficient and δ_j is an error of neuron which is calculated as follows:

$$\delta_j = y_j(1-y_j)(d_j - y_j), \text{ (outputlayer)}, \tag{18.5}$$

$$\delta_j = x'_j(1-x'_j)(\sum_{1}^{M} \delta_k w_{jk}), \text{ (hiddenlayer)}, \tag{18.6}$$

where k represents all neurons in the output layer.

– Step 5: Steps from 3 to 5 are repeated until the error value is less than the predetermined value.

During the training of NN which is based on the BPG algorithm, some problems may occur. These problems are caused by inappropriate setting of training parameters or the improper initialization of weights and thresholds. These difficulties can be reduced by using a modification of the basic algorithm such as Back-Propagation with Momentum or Conjugate Gradient Backpropagation.

18.3 General Description of MLP Attack

In this section, we describe the general usage of the MLP in power analysis attack. Machine learning algorithms are mostly used in profiled attacks where an adversary needs a physical access to a pair of identical devices, which we call a profiling

device and a target device. Basically, these attacks consist of two phases. In the first phase, the adversary analyzes the profiling device and then, in the second phase, the adversary attacks the target device. Typical examples are template-based attacks [7, 9, 16]. By contrast, non-profiled attacks are one-phase attacks that perform the attack directly on the target device such as DPA based on the correlation coefficient [19].

18.3.1 Profiling Phase

In the attack based on the MLP, we assume that we can characterize the profiling device using a well trained neural network. We assume that desired value by adversary is the secret key stored in cryptographic device. This means that one can create and train a NN for a certain part of a cryptographic algorithm. We execute this sequence of instructions on the profiling device with the same data d and different key values k_j to record the power consumption. After measuring n power traces, it is possible to create the matrix $\mathbf{X}_n$ that contains power traces corresponding to a pair of (d, k_i). These pairs represent a training set $\mathbf{T}$ of the neural network. Input values are power traces measured and values of secret key k_i represent the desired output of the neural network. In this case, secret key values k_i can be easily represented using the $n \times 256$ classification matrix $\mathbf{D}$.

After the measurement phase, an adversary creates a neural network. The number of input neurons has to be equal to the numbers of chosen interesting points. We use only interesting points because of memory limitation and time-consuming training process (similar situation like in classical Template attack). Generally, the setting of the hidden layer depends on the problem to solve and the training set, therefore the adversary has to set the number of hidden layers and neurons experimentally. The output layer should contain the desired number of neurons corresponding to the aim of the attack (output byte of S-Box, byte of the secret key, Hamming weight etc.). In our example, the NN is aimed on byte classification, therefore the output layer contains 256 neurons. In the last step of the profiling phase, the adversary trains the neural network created by the prepared training set and the chosen training algorithm.

18.3.2 Attack Phase

During the attack phase, the adversary uses a well-trained NN together with a measured power trace from the target device (denoted as $\mathbf{t}$) to determine the secret key value. The adversary puts the $\mathbf{t} = [x_1, \ldots, x_N]^T$ as an input to NN and it classifies the output values using the calculation:

$$y_k = f_s\left(\sum_{k=1}^{N_1} w'_{jk} x'_j - \theta'_k\right),\ 1 \leq k \leq M, \tag{18.7}$$

where w_{ij} denotes weights between ith hidden neuron (or the input neuron) and the neuron jth and x'_i denotes the output of hidden neurons:

$$x'_j = f_s(\sum_{i=1}^{N} w_{ij}x_i - \theta_j),\ 1 \leq j \leq N_1. \tag{18.8}$$

The result of this classification is a vector $\mathbf{g} = [g_1, g_2, \ldots, g_M]$ which contains the probability value 0 to 1 for every output value. The probabilities show how well the measured trace $\mathbf{t}$ corresponds to the training patterns. Intuitively, the highest probability should indicate the correct training pattern in the training set $\mathbf{T}$ and because each training pattern $\mathbf{X_n}$ is associated with a desire value (in our case secret key), the adversary obtains the information about secret key stored in the target device.

18.4 Experiments Realized

The following text summarizes experiments realized including the information about the experimental setup and the attacks implementation. At first, we focus on attack implementation description, because these are identical for every experiment realized.

18.4.1 Template Attack Implementation

We implemented classical template attack, reduced template attack and effective template attack based on pooled covariance matrix [7]. Calculation of the probability density function was performed according the following Eq. (18.9):

$$p(\mathbf{t}; (\mathbf{m}, \mathbf{C})_{d_i,k_j}) = \frac{exp(-\frac{1}{2} \cdot (\mathbf{t} - \mathbf{m})' \cdot \mathbf{C}^{-1} \cdot (\mathbf{t} - \mathbf{m}))}{\sqrt{(2 \cdot \pi)^{NP} \cdot det(\mathbf{C})}} \tag{18.9}$$

where $(\mathbf{m}, \mathbf{C})$ represents templates prepared in profiling phase based on multivariate normal distribution that is fully defined by a mean vector and a covariance matrix. The power trace measured from the target device is denoted as $\mathbf{t}$ and NI is the number of interesting points. In the case of effective template attack, we calculated the pool covariance matrix as an average value of all covariance matrices and we calculate the Eq. (18.9) using this matrix. In following text, the classical, the reduced template attack and the template attack based on the pooled covariance matrix are denoted as T_{cls}, T_{red} and T_{pool} sequentially.

18.4.2 MLP Attack Implementation

We created and trained the neural network in Matlab using the Netlab neural network toolbox [38]. Ian Nabney and Christopher Bishop from Aston University in Birmingham are the authors of this toolbox and it is available for download. We created a typical two layer perception network and we used optimized learning based on the scaled conjugate gradient algorithm (see Sect. 18.2). A standard sigmoid was chosen as an activation function. The NNs created are shown in Figs. 18.2 and 18.3. The input layer contained five inputs corresponding with interesting points, hidden layer contained 1,000 neurons and output layer had 256 neurons and we used 200 training cycles for DS1. The input layer contained 48 and 50 inputs corresponding with interesting points, hidden layer contained 1,000 neurons and output layer had 16 neurons and we used 180 training cycles for DS1 and DS2. These implementations of NNs are denoted as NN_{org} and practically correspond to the original approach described in [32]. We created the other NNs according the optimization based on preprocessing of measured power traces [33]. These implementations are denoted as NN_{opt}.

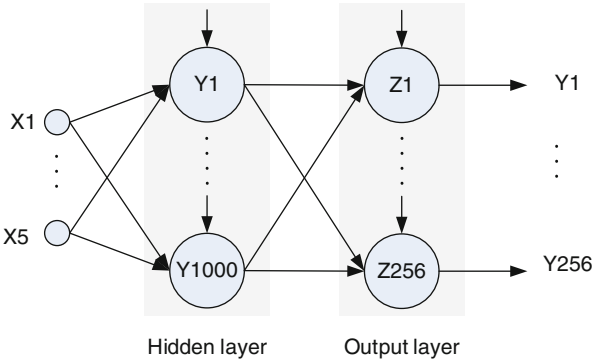


Fig. 18.2 Created NN for DS1

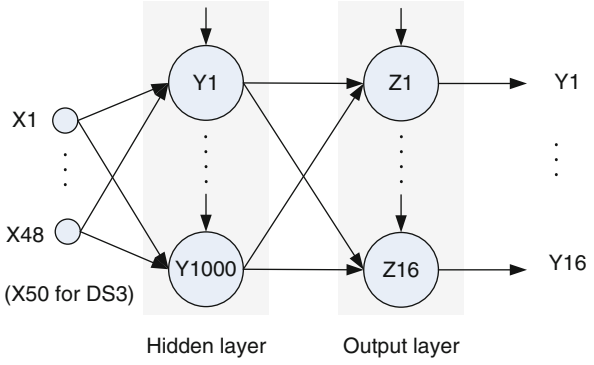


Fig. 18.3 Created NN for DS2 and DS3

18.4.3 First Experiment: Secret Key Revelation

The first data set is focused on secret key classification. DS1 is prepared from power traces of unprotected AES-128 implementation in our testbed. The cryptographic module was represented by the PIC 8-bit microcontroller, and for the power consumption measurement we used the CT-6 current probe and the Tektronix DPO-4032 digital oscilloscope. We used standard operating conditions with 5 V power supply. Stored power traces have 100,000 of samples and cover the `AddRoudnKey` and `SubBytes` operations in the initialization phase of the algorithm (see Fig. 18.4). We can denote stored secret key as $K_{sec} = \{k_1, k_2, \ldots, k_{16}\}$ where k_i represents individual bytes of the key. Because our implementation was realized in the assembly language and the executed instruction of examined operation (`AddRoudnKey` and `SubBytes`) were exactly the same for every key byte k_i, we assume that it is possible to use parts of power traces where first byte is processed to create a model (or templates) to determine the whole secret key byte by byte. In the first step, we determine the value of k_1 and in the second step byte k_2, and so on. The difference between these steps is in the division of the power traces measured into parts corresponding to the time intervals in which the cryptographic device works with the respective bytes of the secret key. The division of power traces is indicated in Fig. 18.4 by numbers. We verified this assumption experimentally and it is naturally conditioned by the excellent synchronization of measured power traces.

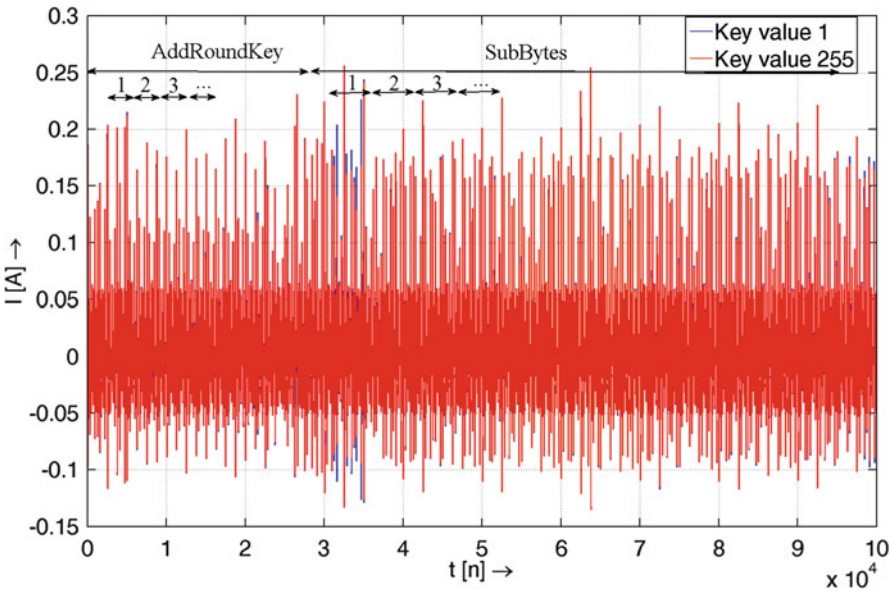


Fig. 18.4 Measured power traces for different first key value

We measured a set of 2,560 power traces where ten power traces were independently stored for each value of the first secret key byte. This number of power traces was chosen because we wanted to compare both implementation of the attacks (the MLP approach and the template attack) using the typical tenfold cross-validation. In data mining and machine learning, the tenfold cross-validation is the most common method of model verification. Cross-validation (CV) is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one is used for learning a model and the other one is used for the model validation. In typical cross-validation, the training and validation sets must cross-over in successive rounds that each data point has a chance of being validated against. Therefore, we used nine power traces in profiling phase of the individual attack and one power trace in attack phase in every step of validation.

We chose five interesting points according to the information provided in [18]. Our algorithm searched for the maximum differences of an average power consummation and power consumption corresponding to key value 1. The algorithm accepted only the maximums that had a distance of at least one clock cycle from each other. This restriction for having interesting points not too close from each other avoids numerical problems during the covariance matrix inverting. Measured power traces were properly synchronized and our device leaks Hamming weight (HW) of processed data. These facts confirm the plots shown in Figs. 18.5 and 18.6. Figure 18.5 shows the detail of power traces that correspond to MOV instruction where data values 0 to 255 were processed. Figure 18.6 shows plot of these measured power traces for one point $t = 4086$. Each of our chosen interesting points leaked HW of processed data. Same chosen points were used for the template creation and the neural network model. Consequently, our first dataset represents a

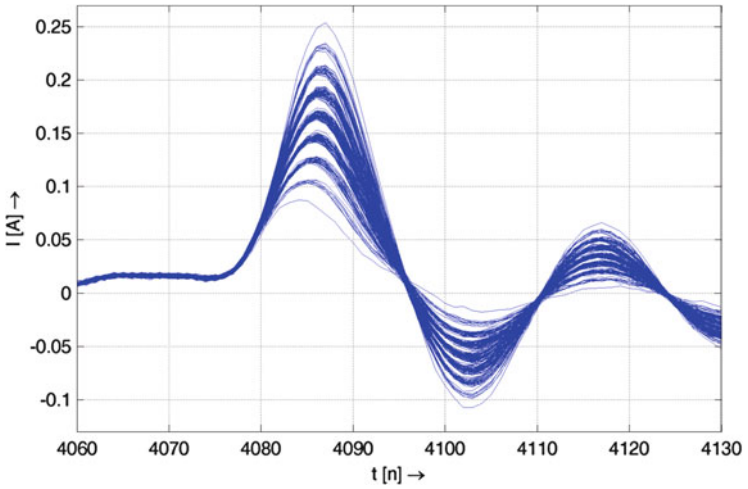


Fig. 18.5 Detail of measured 256 power traces

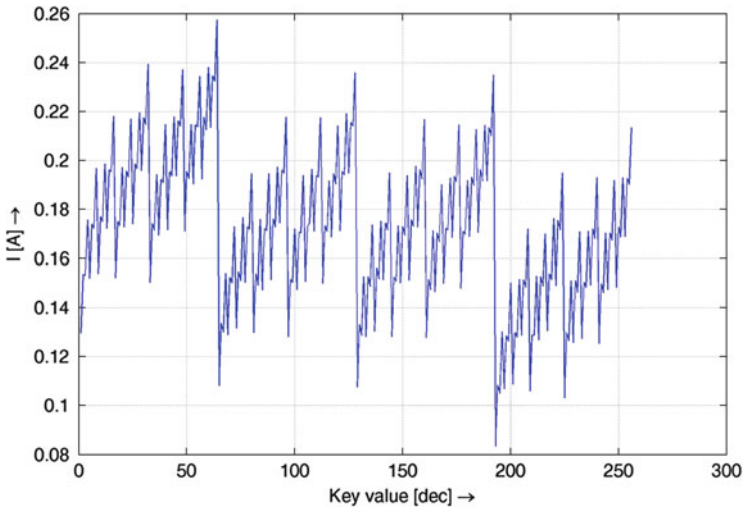


Fig. 18.6 Measured leaks of Hamming weight for point 4,086

matrix 2560×13 where each of the last eight vale is a label. Label is expressed by four columns where every row represents a class using the binary expression 00000000 to 11111111.

A well known fact is that noise always poses the problem during the power consumption measurement. During the preparation of DS1, we performed the experimental measurements of a test bed according to the information provided in [9] and we established that the noise level was distributed according to the normal distribution with the parameters $\mu = 0$ mA and $\sigma = 5$ mA. Every stored power trace was calculated as an average power trace from ten power traces measured using the digital oscilloscope to reduce electronic noise.

In our experiment, we use the guessing entropy to compare our implemented attacks. The guessing entropy is defined as follows: let $\mathbf{g} = [p_1, p_2, \ldots, p_N]$ contain the probability such as $p_1 \geq p_2, \geq \ldots, \geq p_N$ of all possible key candidates after N iterations of Eq. (18.9) or Eq. (18.7). Indices i correspond with the correct key in $\mathbf{g}$. After the realization of S experiments, one obtains a matrix $\mathbf{G} = [g_1, \ldots, g_S]$ and a corresponding vector $\mathbf{i} = [i_1, \ldots, i_S]$. Then the guessing entropy determines the average position of the correct key:

$$GE = \frac{1}{S} \sum_{x=1}^{S} i_x. \tag{18.10}$$

In other words, the guessing entropy describes the average number of guesses, required for recovering the secret key [28, 34].

In the first experiment, we determined the value of one byte of the secret key from one measured power trace. We tried this for all 256 power traces measured

Table 18.1 Guessing entropy for the individual byte determination

CV	NN_{org}	NN_{opt}	T_{cls}	T_{red}	T_{pol}
1	1.16	1.02	1.07	1.04	1.02
2	1.18	1.04	1.07	1.06	1.02
3	1.32	1.03	1.04	1.04	1.03
4	1.16	1.05	1.04	1.04	1.02
5	1.16	1.05	1.07	1.05	1.02
6	1.23	1.04	1.04	1.04	1.02
7	1.15	1.03	1.08	1.03	1.02
8	1.11	1.05	1.07	1.02	1.02
9	1.18	1.06	1.08	1.02	1.00
10	1.17	1.03	1.03	1.04	1.01
ϕ	1.18	1.04	1.06	1.04	1.02

Table 18.2 Guessing entropy for the whole secret key determination

CV	NN_{org}	NN_{opt}	T_{cls}	T_{red}	T_{pol}
1	4.00	2.00	4.00	4.00	2.00
2	24.00	4.00	4.00	1.00	2.00
3	32.00	2.00	4.00	4.00	8.00
4	24.00	8.00	2.00	4.00	4.00
5	4.00	2.00	4.00	4.00	4.00
6	30.00	4.00	16.00	4.00	4.00
7	8.00	2.00	4.00	2.00	2.00
8	16.00	6.00	8.00	2.00	2.00
9	32.00	2.00	4.00	4.00	2.00
10	2.00	1.00	2.00	1.00	1.00
ϕ	17.60	3.30	5.20	3.00	3.10

corresponding to every key values from 0 to 255. In other words, we determined the value of 256 individual bytes in every step of the cross-validation. After the realization, we calculated the GE according to the Eq. (18.10). Obtained results are summarized in Table 18.1, where ϕ denotes an average value calculated from every cross-validations realized. The template attack based on the pooled covariance matrix T_{pol} achieved the best result in one byte guessing but it is important that the classification based on NN was not much worse. The original implementation of the neural network NN_{org} was the worst of all implemented attacks and achieved $GE = 1.18$ in average. The optimized method achieved $GE = 1.04$ that was almost identical with template attacks.

In the following experiment, we determined the whole 128 bit secret key by using the 16 power traces measured. The secret key stored had the value $K = [29, 245, 48, 93, 215, 65, 139, 198, 5, 232, 81, 107, 173, 243, 24, 151]$. Obtained results are written in Table 18.2. The second experiment confirmed the previous results. The adversary needs about 18 guesses to determine the correct secret key after the side-channel attack based on the original implementation of neural network NN_{org}. The results of the optimized method were almost identical with template attacks.

Table 18.3 Obtained results

	NN_{org}	NN_{opt}	T_{cls}	T_{red}	T_{pol}
τ (ms)	1.59	1.11	174.89	149.85	221.66
m (kB)	1920.00	1920.00	4880.00	1448.00	4894.00

Potential adversary would need in average about four guesses to determine the secret key value after the side-channel attack. Our experiments confirm that success secret key revelation of MLP attack is comparable with template based attacks (identical number of interesting points, number of power traces and so on). The MLP approach is able to be trained only for a few interesting points of power traces. In order to complete the comparison of implemented attacks, Table 18.3 provides the information about the time complexity of attack phase τ and memory complexity m.

18.4.4 Second Experiment: Secret Offset Revelation 1

The second experiment is focused on the mask classification. We created DS2 which is based on electromagnetic traces that are freely available on the website of DPA Contest v4 [37]. The masked block-cipher AES-256 in encryption mode without any mode of operation is implemented on the target cryptographic device Atmel ATMega-163-based smart card. Implemented masking scheme is a variant of the Rotating Sbox Masking [12, 42]. According to authors, this masking scheme keeps performance and complexity close to the unprotected scheme and is resistant against several side-channel attacks. The 16 masks are public information that are incorporated in the computation of the algorithm. Offset value, which is drawn randomly at the beginning of computation, is a secret value. Mask values are rotating according to the offset value [12, 42]. Each stored trace has 435,002 samples associated to the same secret key and corresponds to the first and to the beginning of the second round of AES algorithm. For DS2, we chose only the points that are the most correlated with the secret offset value. We realized the classical CPA (differential power analysis based on the correlation coefficient [9]) for operation `Plaintext blinding` depending on the offset value in order to locate interesting points. In other words, we chose output of `Plaintext blinding` as the intermediate value of the CPA attack [12]. Result of the CPA analysis is shown in Fig. 18.7. Finally, we selected 3 points for every mask value, together 48 interesting points were chosen. Selected interesting points of individual masks were exact distributed with distance of 4,342 samples (Mask 0 $t = (5222, 6777, 8777)$, Mask 1 $t = (29564, 11119, 13119)$ and so on). In this experiment, we divided DS2 into a learning set of 1,000 traces and a testing (validation) set of 1,500 traces to measure the success rate. In other words, learning set represents a matrix 1000×52 where

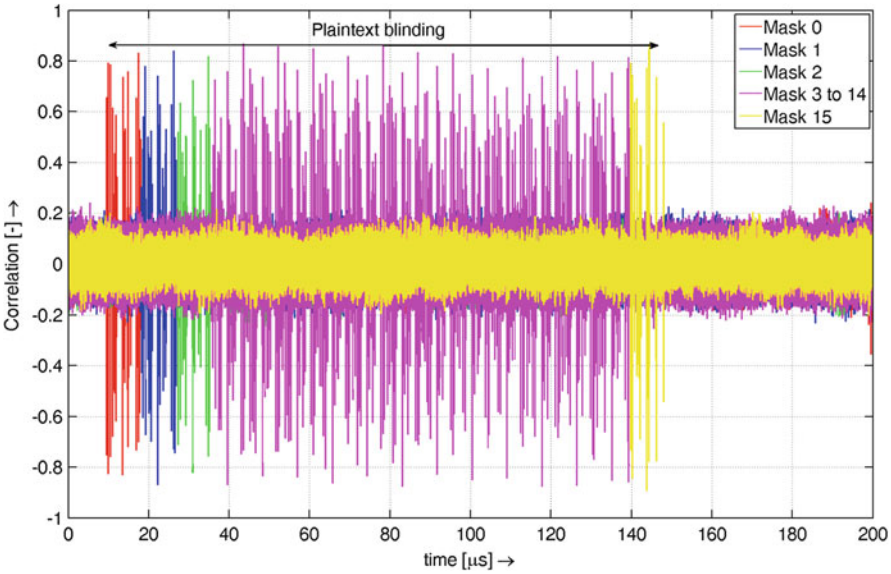


Fig. 18.7 Result of CPA for operation `Plaintext blinding`

each of the last four values is a label. In our case, the label values correspond with the offset value 0 to 15 (16 possible variants). Identically, the validation test was a matrix 1500×52.

In this experiment, we investigate a success rate of the masks revelation depending on the number of interesting points and the number of power traces. In comparison with the first experiment, we use first-order success rate as a metric because secret offset is revealing during this observation. Learning set contains $100, 250, 500$ and $1{,}000$ of power traces successively. Figures 18.8, 18.9, 18.10, and 18.11 report the success rate to predict the right offset value as a function of the number of interesting points selected for T_{cls}, T_{pool}, NN_D and NN. One can extract the following observations. First, as expected, the higher the number of traces in the learning set, the higher the accuracy. For example, maximal success rate achieved was 70 % and 99 % for learning set containing 100 and 1,000 power traces successively. Secondly, the number of selected points in each trace influences the success rate: the higher the number of interesting points, the higher the success rate of every attack implementations. The main finding is that the rise in success rate of the MLP attack occurs much earlier than for every TA attacks. We can observe success rate of 72 % for the MLP and of 7 % for TAs for 20 interesting points and 1,000 power traces.

It is remarkable that if learning set is small (in our experiment less than 1,000 of power traces), the classical template attack is practically inapplicable. It provides the success rate somewhere around 7 %. This is caused by numerical problems that are connected with covariance matrix. These numerical problems occur during the

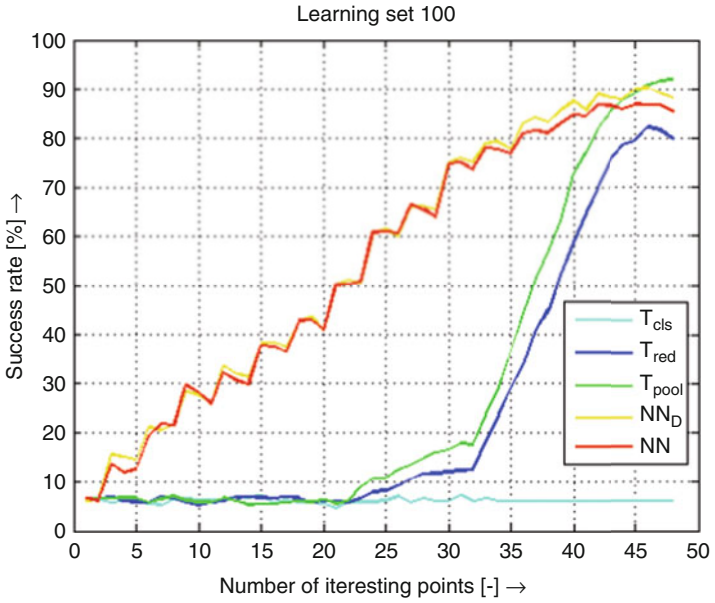


Fig. 18.8 Success rate of the secret offset revelation based on 100 power traces of DS1

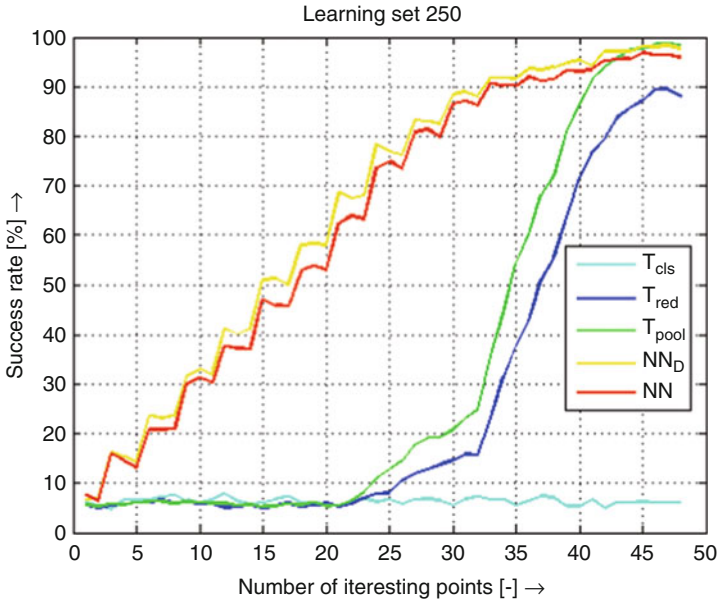


Fig. 18.9 Success rate of the secret offset revelation based on 250 power traces of DS1

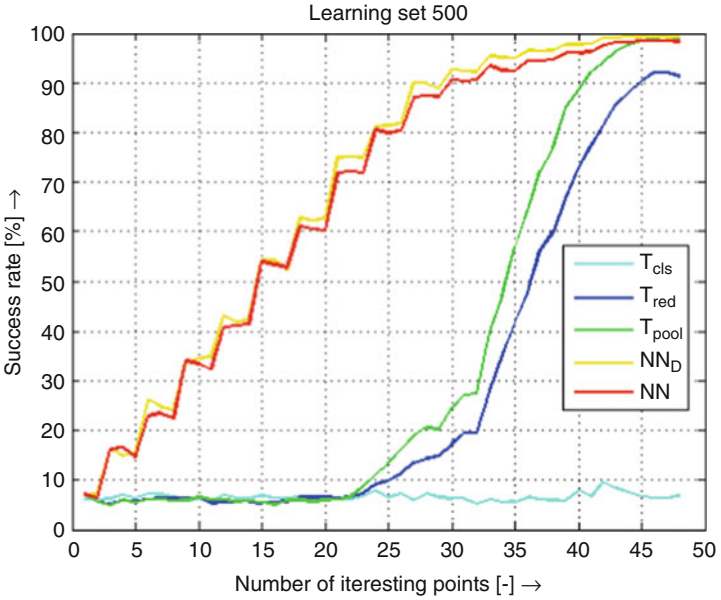


Fig. 18.10 Success rate of the secret offset revelation based on 500 power traces of DS1

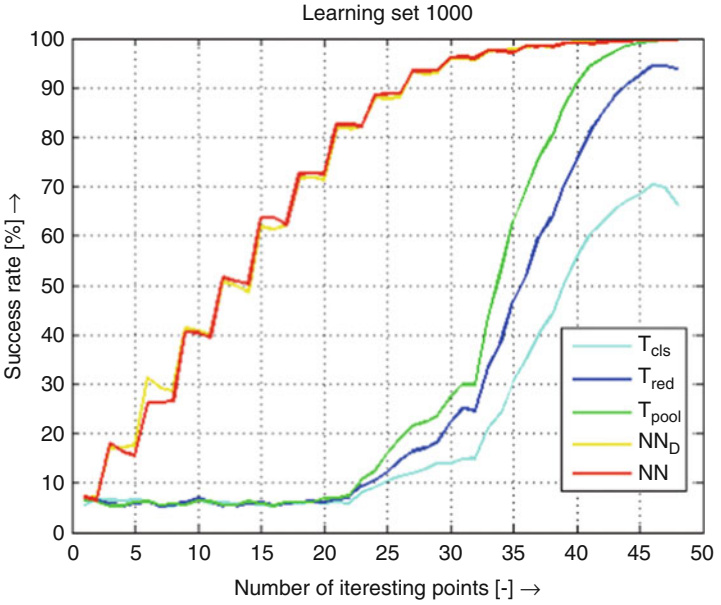


Fig. 18.11 Success rate of the secret offset revelation based on 1,000 power traces of DS1

inversion which needs to be done in Eq. (18.9). In our case, the values that were calculated were very small, what leads to bad classification results. Obtained results confirmed that the MLP is much more effective profiling power analysis attack in terms of small number of power traces and interesting points. The fact is, that the template attack based on the pooled covariance matrix and the MLP are practically the same for larger learning sets. Obtained success rates were 99.9 % and 99.6 % for T_{pool} and NN_D successively (see in Fig. 18.11). Last interesting observation was that the proposed optimization described in [33] has practically no influence on the classification results. This is presumably due to the precise method of choosing interesting points.

18.4.5 Third Experiment: Secret Offset Revelation 2

The third data set was created by Liran Lerman during the preparation of the attack for DPA Contest v4. We refer the work [31] for more information about the preparation of the original dataset. This DS3 is focused on the mask classification and we used first 1,000 traces of 1,500 available in learning set. The author chose 50 interesting points according to the computed Pearson correlation between each instance of 1,500 traces and the offset value. In other words, our DS3 represents a matrix 1000×54 where each of the last four values represents a label value. Again, the label values correspond with the offset value 0 to 15 (16 possible variants). The author created a validation set containing of 1,500 power traces for attacks verification.

As in the previous experiment, we investigate a success rate of the masks revelation depending on a number of interesting points and number of power traces to classification results of attacks implemented. Learning set contains 100, 250, 500 and 1,000 successively. Figures 18.12, 18.13, 18.14, and 18.15 report the success rate to predict the right offset value as a function of the number of interesting points selected for DS3. From results, we can extract the following observations. Firstly, as a confirmation of expected previous statement, the higher the number of traces in the learning set, the higher the accuracy. Secondly, the maximum achieved success rates are lower and differences between the attacks implemented are less pronounced. We can observe this fact focusing on maximal success rates of 99.9 % and 89.7 % of the T_{pool} attack for DS2 and DS3 successively with size of 1,000 power traces. This difference in maximal success rate is even greater for smaller learning set: inspect the success rates of 91.9 % and 57.8 % of the T_{pool} attack for DS2 and DS3 successively with size of 100 power traces. This is presumably due to the method of choosing interesting points. In this data set, interesting points are not selected as precisely as in DS2. The presumption of poor selection of interesting points is confirmed by graphs shown in Figs. 18.12, 18.13, 18.14 and 18.15. The graphs show a plateau at some constant levels for points from 10 to 40. Last finding was the advantage of the usage of efficient template attack (T_{pool}).

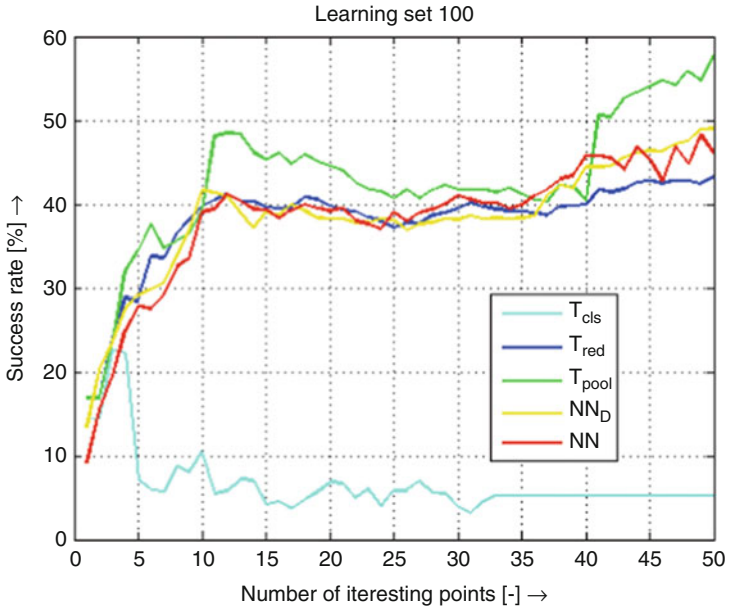


Fig. 18.12 Success rate of the secret offset revelation based on 100 power traces of DS2

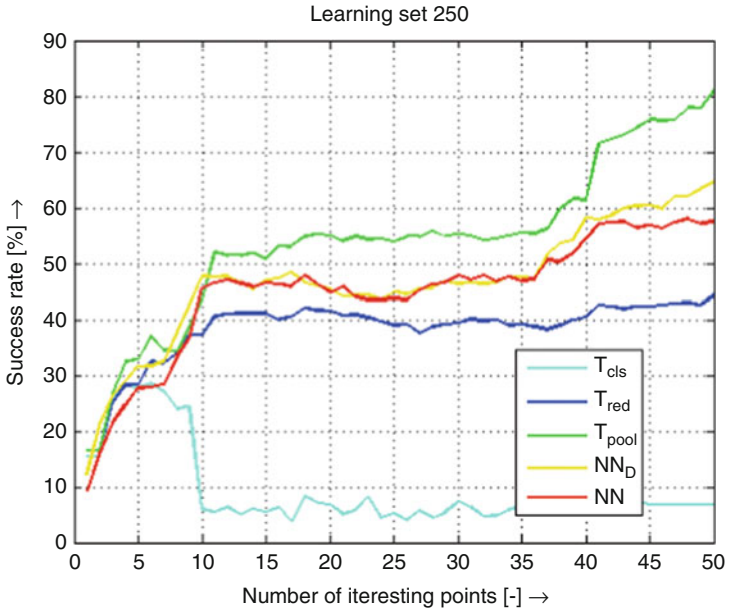


Fig. 18.13 Success rate of the secret offset revelation based on 250 power traces of DS2

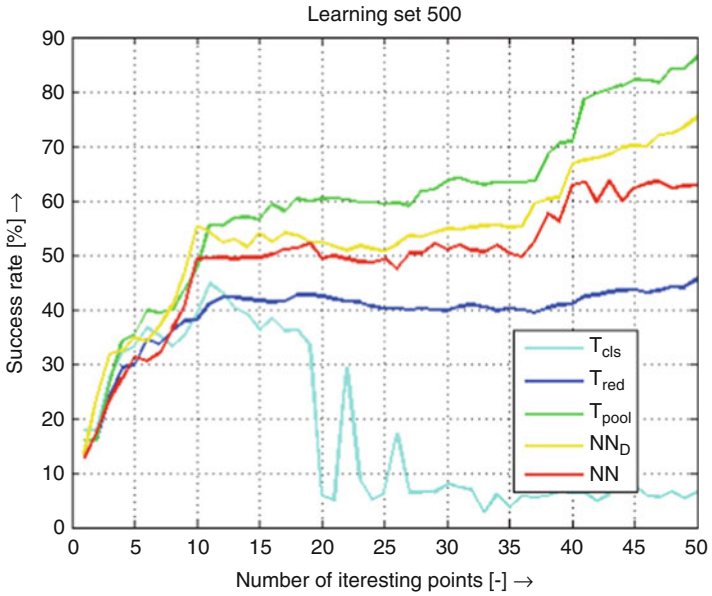


Fig. 18.14 Success rate of the secret offset revelation based on 500 power traces of DS2

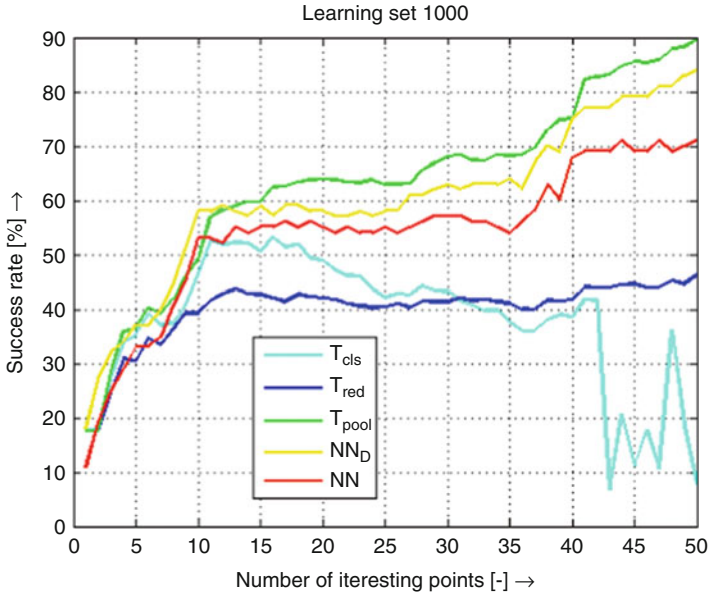


Fig. 18.15 Success rate of the secret offset revelation based on 1,000 power traces of DS2

This implementation eliminates numerical problems connected with covariance matrix and the adversary does not lose information in contrast with reduced template attack.

18.5 Conclusion

In this paper, we made the first fair comparison of power analysis based on the MLP with well-known template attacks. The first experiment realized, that was aimed on the whole secret key of the AES algorithm revelation, confirmed that the efficiency of the power analysis attack based on MLP and the template attack is comparable. The adversary needs about 18 guesses to determine the correct secret key using the original implementation of the MLP attack. This result is three times worse in comparison with the classical template attack that needs about 5.2 guesses to reveal the whole secret key. By contrast, the results of optimized method were almost identical with the best template attack implementation. Potential adversary would need in average about four guesses after the side-channel attack to determine the secret key value of the AES algorithm.

In the second experiment, we investigate a success rate of the masks revelation depending on the number of interesting points and the number of power traces. As expected, the higher the number of traces and points in the learning set, the higher the accuracy of every power analysis attacks implemented. The main finding was that the sharply rise in success rate of the MLP attack occurs much earlier than for every TA attacks. We can conclude that the MLP is much more effective profiling power analysis attack in terms of small number of power traces and interesting points. In other words, it is better to use profiling power analysis attack based on MLP if the adversary has only limited power traces measured than to realize attack based on templates. The fact is, that the most effective template attack based on pooled covariance matrix and the MLP show practically the same result for larger learning sets. Last finding was the confirmation of the efficiency of pooled template attack which eliminates numerical problems connected with covariance matrix and the adversary does not lose information in contrast with usage of reduced template attack.

Acknowledgements Research described in this paper was financed by the National Sustainability Program under grant LO1401. For the research, infrastructure of the SIX Center was used.

References

1. Federal Information Processing Standards Publication (FIPS 197). Advanced Encryption Standard (AES) (2001)
2. Oswald, M.E., Mangard, S., Herbst, C., Tillich, S.: Practical second-order dpa attacks for masked smart card implementations of block ciphers. In: Pointcheval, D. (ed.) Topics in

Cryptology - CT-RSA 2006. Lecture Notes in Computer Science, vol. 3860, pp. 192–207. Springer, Berlin (2006)
3. Raval, N., Bansod, G., Pisharoty, N.: Implementation of efficient bit permutation box for embedded security. WSEAS Trans. Comput. **13**(1), 442–451 (2014)
4. Herbst, C., Oswald, E., Mangard, S.: An AES smart card implementation resistant to power analysis attacks. In: Second International Conference on Applied Cryptography and Network Security (ACNS 2006). Lecture Notes in Computer Science, vol. 3989, 239–252. Springer, Heidelberg (2006)
5. 5. Joye, M., Olivier, F.: Side-channel analysis. In: van Tilborg, H.C.A., Jajodia, S. (eds.) Encyclopedia of Cryptography and Security, 2nd edn., pp. 1198–1204. Springer (2011). ISBN: 978-1-4419-5905-8
6. Fouque, P.A., Kunz-Jacques, S., Martinet, G., Muller, F., Valette, F.: Power attack on small rsa public exponent. In: 8th International Workshop Cryptographic Hardware and Embedded Systems - CHES 2006. Lecture Notes in Computer Science, vol. 4249, pp. 339–353. Springer, Berlin (2006)
7. Choudary, O., Kuhn, M.G.: Efficient template attacks. In: Smart Card Research and Advanced Applications - 12th International Conference, CARDIS 2013, Berlin, 27-29 November 2013, pp. 253–270. Revised Selected Papers. http://dblp.uni-trier.de/rec/bibtex/conf/cardis/ChoudaryK13 (2013)
8. Liu, M., Shien, W.: On the security of yoon and yoo's biometrics remote user authentication scheme. WSEAS Trans. Inf. Sci. Appl. **11**(1), 94–104 (2014)
9. Mangard, S., Oswald, E., Popp, T.: Power Analysis Attacks: Revealing the Secrets of Smart Cards (Advances in Information Security). Springer, New York, Secaucus (2007)
10. Kocher, P.C., Jaffe, J., Jun, B.: Differential power analysis. In: CRYPTO '99: Proceedings of the 19th Annual International Cryptology Conference on Advances in Cryptology, pp. 388–397. Springer, London (1999)
11. Coron, J.S., Goubin, L.: On boolean and arithmetic masking against differential power analysis. In: Proceedings of the Second International Workshop on Cryptographic Hardware and Embedded Systems (CHES '00), pp. 231–237. Springer, London (2000)
12. Nassar, M., Souissi, Y., Guilley, S., Danger, J.L.: RSM: A small and fast countermeasure for AES, secure against 1st and 2nd-order zero-offset scas. In: DATE, pp. 1173–1178 (2012)
13. Muresan, R., Vahedi, H., Zhanrong, Y., Gregori, S.: Power-smart system-on-chip architecture for embedded cryptosystems. In: Proceedings of the 3rd IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS '05), pp. 184–189. ACM, New York (2005)
14. Mesquita, D., Techer, J.D., Torres, L., Sassatelli, G., Cambon, G., Robert, M., Moraes, F.: Current mask generation: A transistor level security against dpa attacks. In: SBCCI, pp. 115–120 (2005)
15. Amin, A., Alsomani, T.: Elliptic curve cryptoprocessor with hierarchical security. WSEAS Trans. Circuits Syst. **13**(1), 135–145 (2014)
16. Chari, S., Rao, J.R., Rohatgi, P.: Template attacks. In: CHES, pp. 13–28 (2002)
17. Hanley, N., Tunstall, M., Marnane, W.P.: Using templates to distinguish multiplications from squaring operations. Int. J. Inf. Secur. **10**(4), 255–266 (2011)
18. Bar, M., Drexler, H., Pulkus, J.: Improved template attacks. In: COSADE 2010 - First International Workshop on Constructive Side-Channel Analysis and Secure Design, pp. 81–89 (2010)
19. Brier, E., Clavier, C., Olivier, F.: Correlation power analysis with a leakage model. In: CHES, pp. 16–29 (2004)
20. 20. Quisquater, J.J., Samyde, D.: Automatic code recognition for smart cards using a kohonen neural network. In: Proceedings of the 5th Conference on Smart Card Research and Advanced Application Conference (CARDIS'02), Berkeley, vol. 5. http://dblp.uni-trier.de/rec/bibtex/conf/cardis/QuisquaterS02 (2002)
21. Kur, J., Smolka, T., Svenda, P.: Improving resiliency of java card code against power analysis. In: Mikulaska kryptobesidka, Sbornik prispevku, pp. 29–39 (2009)

22. Martinasek, Z., Macha, T., Zeman, V.: Classifier of power side channel. In: Proceedings of NIMT2010, September 2010
23. Yang, S., Zhou, Y., Liu, J., Chen, D.: Back propagation neural network based leakage characterization for practical security analysis of cryptographic implementations. In: Proceedings of the 14th International Conference on Information Security and Cryptology (ICISC '11), pp. 169–185. Springer, Berlin (2012)
24. Lerman, L., Bontempi, G., Markowitch, O.: Side channel attack: An approach based on machine learningn. In: COSADE 2011 - Second International Workshop on Constructive Side-Channel Analysis and Secure Design, pp. 29–41 (2011)
25. Liran, L., Gianluca, B., Olivier, M.: Power analysis attack: An approach based on machine learning. Int. J. Appl. Cryptogr. **3**(2), 97–115 (2013)
26. Hospodar, G., Gierlichs, B., Mulder, E.D., Verbauwhede, I., Vandewalle, J.: Machine learning in side-channel analysis: A first study. J. Cryptogr. Eng. **1**(4), 293–302 (2011)
27. Hospodar, G., Mulder, E., Gierlichs, B., Vandewalle, J., Verbauwhede, I.: Least squares support vector machines for side-channel analysis. In: COSADE 2011 - Second International Workshop on Constructive Side-Channel Analysis and Secure Design, pp. 293–302 (2011)
28. Heuser, A., Zohner, M.: Intelligent machine homicide - breaking cryptographic devices using support vector machines. In: COSADE, pp. 249–264 (2012)
29. Bartkewitz, T., Lemke-Rust, K.: Efficient template attacks based on probabilistic multi-class support vector machines. In: Proceedings of the 11th International Conference on Smart Card Research and Advanced Applications (CARDIS '12), pp. 263–276. Springer, Berlin (2013)
30. Lerman, L., Bontempi, G., Taieb, S.B., Markowitch, O.: A time series approach for profiling attack. In: Gierlichs, B., Guilley, S., Mukhopadhyay, D. (eds.) SPACE. Lecture Notes in Computer Science, vol. 8204, pp. 75–94. Springer, Berlin (2013)
31. Lerman, L., Medeiros, S., Bontempi, G., Markowitch, O.: A machine learning approach against a masked AES. In: Francillon, A., Rohatgi, P. (eds.) Smart Card Research and Advanced Applications. Lecture Notes in Computer Science, pp. 61–75. Springer International Publishing, Berlin (2014)
32. Martinasek, Z., Zeman, V.: Innovative method of the power analysis. Radioengineering **22**(2), IF 0.687 (2013)
33. Martinasek, Z., Hajny, J., Malina, L.: Optimization of power analysis using neural network. In: Francillon, A., Rohatgi, P. (eds.) Smart Card Research and Advanced Applications. Lecture Notes in Computer Science, pp. 94–107. Springer International Publishing, Heidelberg (2014)
34. Standaert, F.X., Malkin, T., Yung, M.: A unified framework for the analysis of side-channel key recovery attacks. In: EUROCRYPT, pp. 443–461 (2009)
35. Martinasek, Z., Clupek, V., Krisztina, T.: General scheme of differential power analysis. In: 2013 36th International Conference on Telecommunications and Signal Processing (TSP), pp. 358–362 (2013)
36. Martinasek, Z., Zeman, V., Sysel, P., Trasy, K.: Near electromagnetic field measurement of microprocessor. Przegl. Elektrotechniczny **89**(2a), 203–207 (2013)
37. Guilleyho, S.: DPA contest v4. http://www.dpacontest.org/v4/index.php (2013)
38. Nabney, I.T.: NETLAB: Algorithms for Pattern Recognition. Advances in Pattern Recognition. Springer, New York (2002)
39. Kasabov, N.K.: Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering, 1st edn. MIT Press, Cambridge (1996)
40. Archambeau, C., Peeters, E., Standaert, F.X., Quisquater, J.J.: Template attacks in principal subspaces. In: CHES, pp. 1–14 (2006)
41. Jain, L.C., Martin, N.M.: Fusion of Neural Networks, Fuzzy Sets, and Genetic Algorithms: Industrial Applications, 1st edn. CRC Press, Boca Raton (1998)
42. Moradi, A., Guilley, S., Heuser, A.: Detecting hidden leakages. Cryptology ePrint Archive, Report 2013/842. http://eprint.iacr.org/ (2013)

Chapter 19
A Particular Case of Evans-Hudson Diffusion

Cristina Serbănescu

Abstract We know that the Markov processes are the solutions of certain stochastic equations. In this article we will construct a noncommutative Markov process by noncommutative stochastic calculus. We will also show that these are particular cases of Evans-Hudson diffusions. At the end we will present two examples starting from the classical theory of probabilities (the Brownian motion and the Poisson process) which lead to particular cases of the noncommutative Markov processes.

Keywords Noncommutative Markov process • C*-algebra • Brownian motion • Poisson process • Stochastic equation

19.1 Introduction

Studies in Quantum mechanics have posed the problem of completely positive applications on C*-algebra of continuous linear operators on a Hilbert space. We consider completely positive applications because they describe the evolution of a quantum system, a high-physics energy system and we assume that this evolution is not affected by the existence of other systems that do not interact with the given one. Details concerning the way that the high-physics energy has come to pose this problem may be found in [2] and [3]. Starting from a semigroup of positive operators or from its infinitesimal operator, we can construct a homogenous Markov process. The construction of these processes is done through different methods of which we emphasize on solving stochastic integral equations [16]. Hence the theory of quantum probabilities has developed as a noncommutative theory of probabilities in [1] with motivations in high-physics energy [10]. The corresponding stochastic processes were constructed only in the case of infinitesimal operators and are expressed as finite sums. These are called Evans-Hudson diffusions [5]. This article builds these processes on the antisymmetric Fock space (called fermionic) in which the infinitesimal operator is an infinite sum. The case of the symmetric Fock space

C. Serbănescu (✉)
Department of Mathematics, University "Politehnica" Bucharest, Splaiul Independentei nr. 313, Sector 6, Bucharest, Romania
e-mail: serbanescuc@hotmail.com

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_19

(called bosonic) was treated in [12]. This article shows how the obtained processes as solutions of certain stochastic integral equations are noncommutative Markov processes and appear as a particular case of certain Evans-Hudson diffusions [14] with an infinite number of components, a notion yet to be defined.

This paper shows how noncommutative Markov processes are obtained as solutions of certain stochastic equations, being particular cases of Evans-Hudson Diffusions with an infinity of components.

The case of Markov processes on symmetric Fock spaces for infinitesimal operator as an infinite sum was studied by Hudson and Parthasarathy [10]. This paper aims to build noncommutative Markov processes on antisymmetric Fock spaces where we do not have exponential commutative vectors and where the commutative property does not occur between operators describing disjoint time intervals. Unlike the symmetric Fock space, the defined operators are continuous and the integral is a particular case of Bochner integral [8].

The Brownian Motion and the Poisson Process were given as examples.

19.2 Fermion Stochastic Integrals of Simple Processes

First we construct a stochastic integral on Fermion Fock space [9, 13] by analogy with the same kind of integral on Boson Fock space [4], first of simple processes. We define the Fermion stochastic integral for square-integrable integrands. We present the infinitesimal operators like infinite sums, but we assume they are continuous. Because of the canonical anticommutation relation we have left, right and mixed stochastic integrals.

The noncommutative stochastic calculus was developed on Fermion Fock space.

Definition 2.1. *Let H be a Hilbert space. We define the antisymmetric Fock space $\Gamma_a(H)$ over H as the linear hull of all $x_1 \wedge x_2 \wedge \cdots \wedge x_n$, $n \geq 0$, $x_i \in H$ (where for $n = 0$, we have the unit element, namely 1) with the following inner product*:

$$\langle x_1 \wedge x_2 \wedge \cdots \wedge x_n, y_1 \wedge y_2 \wedge \cdots \wedge y_k \rangle = \delta_{n,k} \quad \det\left(\langle x_i, y_i \rangle\right)_{i,j=1,\ldots,n}$$

for $n = k = 0$ the determinant is considered to be 1.

About this space we mention the following:

(i) $\Gamma_a(H) = \underset{n\geq 0}{\oplus} H_{\wedge}^n$, *where $H_{\wedge}^n$ is the closed linear hull of all $x_1 \wedge x_2 \wedge \cdots \wedge x_n$, $x_i \in H$*

(ii) $x_{\gamma(1)} \wedge \cdots \wedge x_{\gamma(n)} = \varepsilon(\gamma)\, x_{\gamma(1)} \wedge \cdots \wedge x_{\gamma(n)}$ *where $\varepsilon(\gamma)$ is 1 or -1 if γ is even or odd.*

If two x_i with different indexes i are equal this product is null.

(iii) $\Gamma_a(H) = \left\{ \sum_{n\geq 0} x_n : x_n \in H_\wedge^n,\ \{n : x_n \neq 0\}\ finite \right\}$ *is an associative algebra, with unit element 1 and $x_1 \wedge \cdots \wedge x_n$ is the product of $x_1, \ldots, x_n$, $x_i \in H = H_\wedge^1$, in established order.*

(iv) *If $H \subset K$, then $\Gamma_a(H) \subset \Gamma_a(K)$.*

Definition 2.2. *Let H be a Hilbert space*

(a) *By a filtration in H we mean a family $(H_t)_{t\in[0,\infty)}$ of closed subspaces of H such that*

$$H_s \subset H_t, \quad \forall s < t.$$

(b) *We say that it is a right continuous filtration if $H_t = \bigcap_{u>t} H_u$*

(c) *We say that it is a left continuous filtration if H_t is the closure of $\bigcup_{u>s} H_s$*

(d) *We say that $(H_t)_{t>0}$ is continuous if the filtration is right and left continuous.*

The idea of defining these processes may be found in [6].

Definition 2.3. *An adapted process is a family of operators $F = (F(t); t \geq 0)$ on h such that for each $t \geq 0$:*

(a) $D(F(t)) = h_0 \underset{\sim}{\otimes} \varepsilon_t \underset{\sim}{\otimes} h^t$.

(b) *There is an operator $F^+(t) : h_0 \underset{\sim}{\otimes} \varepsilon_t \underset{\sim}{\otimes} h^t \to h_0$ such that*

$$\langle F(t)\zeta, \eta\rangle = \left\langle \zeta, F(t)^+\eta \right\rangle \text{for } \forall \zeta\eta \in h_0 \underset{\sim}{\otimes} \varepsilon_t \underset{\sim}{\otimes} h^t$$

(c) *There are operators $F_1(t)$ and $F_1^+(t)$ on $h_0 \otimes \varepsilon_t$ such that*:

$$F(t) = F_1(t) \otimes 1$$

$$F^+(t) = F_1^+(t) \otimes 1$$

(d) *For each t_0 and $x \in \varepsilon$ we have*:

$$\sup_{\|u\|\leq 1} \|(F(t_0 + h) - F(t))(u \otimes x)\| \overset{h\to 0}{\to} 0$$

hence $\forall x \in \varepsilon$ and $t \in [0, \infty)$ $\lim_{s\to t} \|F(t) - F(s)\|_x = 0$
where $\|T\|_x = \sup_{\|u\|\leq 1} \|T(u \otimes x)\|$

Definition 2.4. *A simple process is an adapted process of the form*:

$F(t) = \sum_{n=0}^{\infty} F_n \chi_{[t_n,t_{n+1})}(t); t \geq 0$ *for some sequence* $0 = t_0 < t_1 < \cdots < t_n \to \infty$

We denote by A_0 *and* A, *respectively, the sets of simple and adapted processes.*

Definition 2.5. *Let* $F, G, H \in A_0$ *and write*

$$F = \sum_{n=0}^{\infty} F_n \chi_{[t_n,t_{n+1})} \ , G = \sum_{n=0}^{\infty} G_n \chi_{[t_n,t_{n+1})} \ , H = \sum_{n=0}^{\infty} H_n \chi_{[t_n,t_{n+1})}$$

$$0 = t_0 < t_1 < \cdots < t_n \to \infty$$

The family of operators $M = (M(t), t \geq 0)$ *with* $D(M(t)) = h_0 \underset{\sim}{\otimes} \varepsilon_t \underset{\sim}{\otimes} h^t$ *defined by* $M(0) = 0$

$M(t) = M(t_n) + \left(A_L^+(t) - A_L^+(t_n)\right) F_n + G_n (A_L(t) - A_L(t_n)) + (t - t_n) H_n$ *for* $t_n < t < t_{n+1}$ *is called stochastic integral of* (F, G, H) *and are denoted by*:

$$M(t) = \int_0^t dA_L^+ F + G dA_L + H ds$$

19.3 Stochastic Integrals of Continuous Processes

Now we want to estimate the norm of $M(t)(u \otimes x)$, in order to define the stochastic integrals [11].

We consider three possibilities, where the first is:

$M(t) = \sum_{n=0}^{b} \left(A_L^+(s_{n+1}) - A_L^+(s_n)\right) F_n$ for $t_b < t < t_{b+1}$, $s_i = t_i$ for $i = 0, \cdots, b$ and $s_{b+1} = t$.

We denote $F(t) = \sum_{n=0} \chi_{[t_n,t_{n+1}]}(t) F_n$ and we write briefly

$$dM = \left(dA_L^+\right) F \text{ or } M(t) = \int_0^t \left(dA_L^+\right) F.$$

We write as follows:

$$\|M(t)(u \otimes x)\|^2 \leq \left\| \sum_p L_p * L_p \right\|^2 \int_0^t \|F(u \otimes x)\|^2 da + \int_0^t \|F(\theta u \otimes x)\|^2 da$$

$$+ \sum_k \sup_{a \leq t} \|M(a)(u \otimes x_{ck})\|^2 \|x_k\|^2$$

Now we deduce that if $F_c = 1, 2, \ldots$ are "simple integrands" like before and if every $u \in h_0$, $x \in \varepsilon$ and $t > 0$, $\int_0^t \left\| F_c\,(u \otimes x) - F_{c'}\,(u \otimes x) \right\|^2 da \to 0$ for $c, c' \to \infty$, then for every $t > 0$, $u \in h_0$ and $x \in \varepsilon$, $\sup_{a \le t} \left\| M_c\,(u \otimes x) - M_{c'}\,(u \otimes x) \right\|^2 \to 0$, where $dM_c = \left(dA_L^+\right) F_c$.

We consider $x = x_1 \wedge \cdots \wedge x_r$ and by induction on r, the term $\sum_k \sup_{a \le t} \|M(a)\,(u \otimes x_{ck})\|^2 \|x_k\|^2$ vanishes for $r = 0$.

This is the way we define $\int_0^t \left(dA_L^+\right) F$ for those F for which it is a sequence F_c of simple integrands with $\int_0^t \left\| F_c\,(u \otimes x) - F_{c'}\,(u \otimes x) \right\|^2 da \to 0$ for $c' \to \infty$ for every $t > 0$, $u \in h_0$ and $x \in \varepsilon$.

We remark that if $F(t) = F_1(t) \otimes 1$ with respect to $h = h_t \otimes h^t$ and if $F(t)\,(u \otimes x)$ is continuous in t for every $u \in h_0$, $x \in \varepsilon$, then there exists a sequence F_c namely $F_c(t) = \sum_{k \ge 0} \chi_{\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right]} F\left(\frac{k}{2^n}\right)$.

We also mention that:

$$\|M(t)\|_x^2 \le \left\| \sum_p L_p * L_p \right\|^2 \int_0^t \|F(a)\|_x^2 da + \int_0^t \|F(a)\|_x^2 da + \sum_k \sup_{a \le t} \|M(a)\|_{x_{ck}}^2 \|x_k\|^2.$$

Writing the formulas $\|M(t)\,(u \otimes x)\|^2$ for $M(t) - M(s)$, we shall obtain:

$$\|M(t) - M(s)\|_x^2 \le \left\| \sum_p L_p * L_p \right\|^2 \int_s^t \|F(a)\|_x^2 da + r \int_s^t \|F(a)\|_x^2 da + \sum_k \sup_{a \le t} \|M(a) - M(s)\|_{x_{ck}}^2 \|x_k\|^2$$

and by induction we show that F is continuous hence $\lim_{s \to t} \|F(s) - F(t)\|_x = 0$ for every x, then $M(t)$ follows continuous similarly.

Definition 3.1. *The integrals can be defined separately and we have*:

$$M(t) = \int_0^t \left(dA_L^+ F + G dA_L + H ds\right) = \int_0^t dA_L^+ F + \int_0^t G dA_L + \int_0^t H ds.$$

19.4 Stochastic Equations

Theorem 4.1. *Let be the operators $X(0)$, B and D in $L(h_0)$.*
We show that the stochastic differential equation:

$$X(t) = X(0) + \int_0^t dA_L^* (B_F X D_F) + \int_0^t (B_G X D_G)\, dA_L + \int_0^t (B_H X D_H)\, ds$$

has a unique solution which is a continuous process.

Proof. We remark that the integrands are "allowable", hence the stochastic integrals are well defined.

Unicity: if X and Y are two solutions, with $X(0) = Y(0)$, then $Z = X - Y$ will be a solution of the equation with $Z(0) = 0$.

Since $\|BFD\|_x \leq \|B\| \, \|F\|_x \, \|D\|$ for $B, D \in L(h_0)$, we have for $t \leq T$:

$$\|Z(t)\|_x^2 \leq c \int_0^t \|Z(a)\|_x^2 da + \sum_k \sup_{a \leq t} \int_0^a \|B_F Z(a) D_F\|^2 X C_k \|x_k\|^2 da$$
$$+ \left(\int_0^t \sum_k \|(B_G Z(a) D_G)\|^2 X C_k da \right) e^{\sum_k \|x_k\|^2}$$

We give the proof by induction on r, if $x = x_1 \wedge x_2 \wedge \cdots \wedge x_r$.

Knowing that

$Z(a)\,(u \otimes (x_1 \wedge x_2 \wedge \cdots \wedge x_{r-1})) = 0$ for all u and x_i, we deduce:

$\|Z(t)\|_x^2 \leq c \int_0^t \|Z(a)\|_x^2 da + 0$ (for $k = 0$ this is obvious) and using Gronwall's lemma, we obtain $\|Z(t)\|_x^2 \leq 0 e^{ct}$, hence $Z(a)\,(u \otimes (x_1 \wedge x_2 \wedge \cdots \wedge x_{r-1})) = 0$ for all u and x_i.

Existence: We establish the existence iteratively.

We fix $T > 0$ and we consider $X(0)(t) = X(0)$ for every $t \leq T$ and then inductively:

$$X_{n+1}(t) = X(0) + \int_0^t dA_L^* (B_F X_n D_F) + \int_0^t (B_G X_n D_G)\, dA_L + \int_0^t (B_H X_n D_H)\, ds$$

We have:

$$\begin{aligned}
&\|X_{n+1}(t) - Xn(t)\|_y \\
&= \left\| \int_0^t dA_L^* \left(B_F \left(X_n - X_{n-1}\right) D_F\right) + \int_0^t \left(B_G \left(X_n - X_{n-1}\right) D_G\right) dA_L \right. \\
&\quad \left. + \int_0^t \left(B_H \left(X_n - X_{n-1}\right) D_H\right) ds \right\|_y \\
&\leq d^{'} {q_{n-1}}^{(p-1)} + d^{'} {q_{n-2}}^{(p-1)} cT + \cdots + d^{'} {q_0}^{(p-1)} \left((cT)^{n-1} / (n-1)!\right) \\
&\quad + c^{'} \left(c^n T^n / n!\right)
\end{aligned}$$

If we denote with q_n the last expression which doesn't depend on t, we have, if k and $n-k$ converge to ∞ with n:

$$\begin{aligned}
q_n &= d' \sum_{j=1}^{k-1} {q_{n-j}}^{(p-1)} (cT)^{j-1} / (j-1)! \\
&\quad + d' \left({q_{n-k}}^{(p-1)} (cT)^{k-1} / (k-1)! + {q_{n-k-1}}^{(p-1)} (cT)^k / k!\right) \\
&\quad + \sum_{j=k+2}^{n} {q_{n-j}}^{(p-1)} (cT)^{j-1} / (j-1)! + c' \left(c^n T^n / n!\right)
\end{aligned}$$

Now we use $\left(\sum_k a_k\right) \Big/ \left(\sum_k b_k\right) \leq \max\left(a_k / b_k\right)$, and we have:
$\max_{k+2 \leq j \leq n} \left((cT) \Big/ (j-1), cT/n\right) =$
$\max\left(\max_{j \geq n-k} \left(q_{j+1}^{(p-1)} \Big/ q_j^{(p-1)}\right), \left(q_{n-k}^{(p-1)} / q_{n-1-k}^{(p-1)}\right) + cT/k, \max_{k+1 \leq j} \left(cT/j\right)\right)$
which converges to 0.

Hence $\sum \left(\left(X_n - X_{n-1}\right)(t)\right)$ is the solution of the equation.

Theorem 4.2. *We consider the stochastic integral equation*: $U(t) = 1 + \int_0^t \left(U\theta\left(dA_L^+\right) + U\theta\left(dA_L\right) + UXds\right)$,

where $X = -\left(\sum_p L_p + L_p\right) \Big/ 2.$

Then there exists a unique unitary process satisfying this equation.

Proof. We have $U^+(t) = 1 + \int_0^t \left(\left(dA_L^+ \right) \theta U^+ + (dA_L)\, \theta U^+ + X U^+ ds \right)$ (since $X = X^*$).

19.5 Noncommutative Markov Processes as Stochastic Equation Solutions

Definition 5.1. *A noncommutative Markov process is a system which includes*:

(i) *A Hilbert space* h_0.
(ii) *A C*-algebra* $A \subset L(h_0)$ *with* 1.
(iii) *A family of completely positive mappings*: $T_1 : A \to A,\ t \geq 0$ *with* $T_t 1 = 1$, $T_0 = 1$ *and* $T_{t+s} = T_t T_s$ (*briefly a semigroup of completely positive mappings on A with* $T_t 1 = 1$).
(iv) *Another Hilbert space h, in which* h_0 *is a closed subspace*.
(v) *A family* $(j_t)_{t \geq 0}$ *of *-homomorphisms* $j_t : A \to L(h)$, *such that*:

1. $j_0(x) = x \oplus 0$ *relatively to* $h = h_0 \oplus h_0^{\perp}$.
2. $j_s(1)\, j_{s+t}(x)\, j_s(1) = j_s(T_t x)$.

Remark.

(a) $j_s(1)$ is a projector.
(b) $j_s(1) \leq j_{s+t}(1)$ for $t \geq 0$, hence denoting $h_t = \operatorname{Im} j_t(1)$, notation which is not incompatible with h_0, we obtain a filtration (h_t).
(c) There results that T_t is completely positive:
$j_s(T_t x) = j_s(1)\, j_{s+t}(x)\, j_s(1)$ for $s = 0$ we have $j_0(T_t x) = j_0(1)\, j_t(x)\, j_0(1)$ and $\langle j_0(T_t x)\, u, v \rangle = \langle j_0(1)\, j_t(x)\, j_0(1) u, v \rangle$ and j_0 is a projector.

We also have:

$$\langle (T_t x)\, u, v \rangle = \langle j_t(x)\, (u \otimes 1), (v \otimes 1) \rangle .$$

Let be $S_i \in A$, $V_i \in A$, then we have:

$$\left\langle \sum_{i,j} S_i * T_t \left(V_i^* V_j \right) S_j u, v \right\rangle = \sum_{i,j} \left\langle T_t \left(V_i^* V_j \right) S_j u, S_i v \right\rangle$$

$$= \sum_{i,j} \left\langle j_t \left(V_i^* V_j \right) \left(S_j u \otimes 1 \right), \left(S_i v \otimes 1 \right) \right\rangle$$

$$= \sum_{i,j} \left\langle j_t * (V_i)\, j_t(V)_j \left(S_j u \otimes 1 \right), S_i v \otimes 1 \right\rangle$$

$$= \sum_{i,j} \left\langle j_t \left(V_j \right) \left(S_j u \otimes 1 \right), j_t (V_i)\, S_i v \otimes 1 \right\rangle = \left\| \sum_j \langle j_t (V_i)\, S_i v \otimes 1 \rangle \right\|^2 \geq 0$$

(d) If $S : A \to A$ is continuous and linear, then $T_t = e^{tS}$ defines a semigroup, but generally T_t are not completely positive. If $S1 = 0$ then $T_t 1 = 1$.
(e) If $U \in L(h)$ is unitary, then $T \to UTU^*$ is a *-homomorphism:

$$L(h) \to L(h).$$

We shall use the following formulas:

I. If $M(t) = M(0) + \int_0^t dA_L^+ F + (dA)_L G + (ds)H$ then

$$\begin{aligned}\langle M(t)(u \otimes x), M(t)(v \otimes y)\rangle &= \langle M(0)(u \otimes x), M(0)(v \otimes y)\rangle \\ &+ \int_0^t \Big(\sum\nolimits_p \left\langle L_p F(a)(u \otimes x), L_p F(a)(v \otimes y)\right\rangle \\ &+ \sum\nolimits_{p,j} (-1)^{j+r+w} \left(x_{jp}(a) \left\langle M(a)\left(u \otimes x_{cj}\right), \ L_p I F(a) I (v \otimes y)\right\rangle\right. \\ &+ \overline{y_{jp}}(a) \left\langle L_p I F(a) I (u \otimes x), M(a)\left(v \otimes y_{cj}\right)\right\rangle \\ &+ (-1)^{j-1} \left(\overline{y_{jp}}(a) \langle M(a)(u \otimes x), \ L_p I G(a) I \left(v \otimes y_{cj}\right)\right\rangle \\ &+ \left. x_{jp}(a) \left\langle L_p I G(a) I \left(u \otimes x_{cj}\right), M(a)(v \otimes y)\right\rangle\right) \\ &+ \langle H(a)(u \otimes x), M(a)(v \otimes y)\rangle) \, da\end{aligned}$$

II. For $S \in L(h_0)$, we have $(S \otimes 1) = \int_0^t dA_L^+ F + (dA_L) G + H ds =$ $\int_0^t \left(dA_{SL}^+\right) F + (dA_{LS^*}) G + SH ds$ as we know from the definition of the stochastic integral.

III. From $U^*(t) = 1 + \int_0^t \left(\left(dA_L^+\right)\theta U^* + (dA_L)\theta U^* + XU^* ds\right)$ we deduce that

$$(S \otimes 1) U^*(t) = (S \otimes 1) + \int_0^t \left(\left(dA_L^+\right)\theta U^* + (dA_{LS^*})\theta U^* + XU^* ds\right)$$

We consider

$$U(t) = 1 + \int_0^t \left(U\theta\left(dA_L^+\right) + U\theta\left(dA_L\right) + UXds\right)$$

we write it as follows:

$$U\left(t+s\right) = U(s) + \int_0^{s+t} \left(U\theta\left(dA_L^+\right) + U\theta\left(dA_L\right) + UXds\right)$$

The integral can be considered as $\int_0^t$ of the same integrant with h_s instead of h_0 and h^s instead of h^0. Using "III.", the equation can be written: $U(s)^{-1}U\left(t+s\right) = 1 + \int_0^{s+t} \left(U(s)^{-1}U\theta\left(dA_L^+\right) + U(s)^{-1}U\theta\left(dA_L\right) + U(s)^{-1}UXds\right)$ and then $U(s)^{-1}$ $U\left(.+s\right)$ appears as $U(.)$.

Lemma 5.2. *We consider the equation*

$U(t) = 1 + \int_0^t \left(U\theta\left(dA_L^+\right) + U\theta\left(dA_L\right) + UXds\right)$ *where* $X^* = X = -\left(\sum L_p + L_p\right)/2$.

If we define:

$$A = \left\{S; \in L\left(h_0\right), S\theta = \theta S\right\}, \quad T_t(S) = e^{ty}(S)$$

where $Y(S) = \left(\sum_p {L_p}^* S L_p\right) + XS + SX$.

Then we have $\langle T_t(S)u, v\rangle = \langle U(t)\left(S \otimes 1\right)U(t)\left(u \otimes 1\right), \left(v \otimes 1\right)\rangle$.

Proposition 5.3. *We consider the equation*

$$U(t) = 1 + \int_0^t \left(U\theta\left(dA_L^+\right) + U\theta\left(dA_L\right) + UXds\right)$$

where $X = -\left(\sum L_p + L_p\right)/2$.

Then, if we define

$$A = \left\{S; \in L\left(h_0\right), S\theta = \theta S\right\}, \quad T_t(S) = e^{ty}(S)$$

where
$Y(S) = \left(\sum_p {L_p}^* S L_p\right) + XS + SX$ *and* $j_t(S) = (U_t\,(S \otimes 1)\,{U_t}^*)\,P_t$, *where* P_t *is the projector on* h_t, *the system* $h_0, L(h_0)$, T_t *and* j_t *is a noncommutative Markov process.*

Proof. Writing the equation

$$U^+(t) \;=\; 1 + \int_0^t \left(\left(dA_L^+\right)\theta U^+ + (dA_L)\,\theta U^+ + XU^+ ds\right) \text{ and since } U^*(t)$$

appears as $\otimes 1$ relatively to $h = h_t \otimes h^t$ and $P_{s+t} = 1\otimes$ our relation becomes

$$\begin{aligned}&\left\langle (S \otimes 1)\,U(s+t)^*\,(u \otimes x)\,, U(s+t)^*\,(v \otimes y)\right\rangle = \\ &\left\langle (T_t(S) \otimes 1)\,U(s)^*\,(u \otimes x)\,, U(s)^*\,(v \otimes y)\right\rangle\end{aligned}$$

We have $U(s)^*\,(u \otimes x)\,, U(s)^*\,(v \otimes y) \;\in\; h_s$ and it suffices to show that $\left\langle (S \otimes 1)\,U(s+t)^*\,(u \otimes x)\,, U(s+t)^*\,(v \otimes y)\right\rangle = \langle (T_t(S) \otimes 1)\,u, v\rangle$.

Hence we obtain the formula from Lemma 5.2, that is

$$\langle T_t(S)u, v\rangle = \left\langle U(t)\,(S \otimes 1)\,U(t)^*\,(u \otimes 1)\,, (v \otimes 1)\right\rangle.$$

Then $T \to UTU^*$ is a *-homomorphism: $L(h) \to L(h)$.

19.5.1 The Brownian Motion as Noncommutative Markov Process

We consider H a Hilbert space, a Brownian x_t on a probability space (E, K, P), $A = A^* \in L(H)$ and $U\,(t,\omega) = e^{ix_t(\omega)A}$. Hence $U(t,\omega)$ is a unitary operator of $L(H)$. Let be $T_t : L(H) \to L(H)$ defined as $T_t(x) = \int U(t)XU(t)^* dP$.

We have $U(t) = \sum_{n\geq 0} i^n (x_t)^n A^n/n!$, hence

$$\begin{aligned}U(t)^* &= \textstyle\sum_{n\geq 0}(-1)^n (x_t)^n A^n/n!, U(t)X(t)U(t)^* \\ &= \textstyle\sum_{n,k}(-1)^{n-k}(x_t)^{n+k} A^n XA^k/n!k! \\ &= \textstyle\sum_u \sum_{n+k=u}(x_t)^{n+k} i^u i^{-2k} A^n XA^k/n!k! \\ &= \textstyle\sum_u (ix_t)^u \sum_{n+k=u}(-1)^k A^n XA^k/n!k! = \sum_u (ix_t)^u D^u(X)/u!\end{aligned}$$

where $D(X) = AX - XA$.

Indeed $D = P - Q$.

$P(X) = AX$, $Q(X) = XA$, hence

$D^u(X) = (P-Q)^u(X) = \sum_{n+k=u} C_u^n (-1)^k P^n Q^k X$, since P and Q commute.

Hence $T_t(X) = \sum_u i^u E\left((x_t)^u\right) D^u(X)/u!$.

From $E\left(e^{i\lambda x t}\right) = e^{-t\lambda^2/2} E\left(e^{i\lambda x_t}\right) = e^{-t\lambda^2/2}$ we deduce that

$$\sum_n i^n E\left((x_t)^n\right) \lambda^n/n! = \sum_n \left(-\lambda^2 t/2\right)^n/n!$$

and replacing $\lambda = D$ we obtain $T_t(X) = e^{-tD^2/2}(X)$, that is $T_t = e^{-tD^2/2}$.

We have

$-\left(D^2/2\right)(X) = -\left(A^2X - XA^2\right)/2 + AXA$, hence it is of the considerate form with only one term $L = L^* = A$.

19.5.2 The Poisson Process as Noncommutative Markov Process

We consider H a Hilbert space, a sequence U_n of unitary operators, a convergent sum with positive terms $\sum_n \lambda_n = \lambda$, $p_n = \lambda_n/\lambda$, a probability space (E, K, P) and a particular composite Poisson process on it, that is $x_t = y_{zt}$, where (z_t) is a Poisson process of parameter λ and $(y_n)_{n\geq 1}$ is a sequence of independent variables, independent of (z_t), all having the repartition $\Lambda = \Sigma\, p_n \varepsilon_n$. We consider $Y_0 = 1$.

For every t we consider $U_t(\omega) = U_{y_{zt}} \ldots U_0$ and we define for $X \in L(H)$,

$$T_t(X) = \int U(t) X U(t)^* dP.$$

We have

$$T_t(X) = \sum_k \int \chi_{(z_t=k)} U(t) X U(t)^* dP$$
$$= \sum_k \left((\lambda t)^k e^{-\lambda t}/k!\right) \int \chi_{(z_t=k)} U_{y_k} \ldots U_{y_0} X U^*{}_{y_0} dP =$$
$$\sum_k \left((\lambda t)^k e^{-\lambda t}/k!\right) \sum_{n_1,\ldots,n_k} p_{n_1} \ldots p_{n_k} U_{n_k} \ldots U_{n_1} X U^*{}_{n_1} \ldots U^*{}_{n_k}$$

We denote $L_i(X) = U_i X U_i{}^*$ and we have:

$$T_t(X) = \sum_k \left((t)^k e^{-\lambda t}/k!\right) \sum_{n_1,\ldots,n_k} \lambda_{n_1} \ldots \lambda_{n_k} L_{n_k} \ldots L_{n_1}(X)$$
$$= \sum_k \left((t)^k e^{-\lambda t}/k!\right) \left(\sum_n \lambda_n L_n\right)^k (X) = e^{t\left(-\lambda + \sum_n \lambda_n L_n\right)}(X)$$

hence $T_t = e^{t\left(-\lambda + \sum_n \lambda_n L_n\right)}$.

We remark now that

$$\left(-\lambda + \sum\nolimits_n \lambda_n L_n\right)(X) = \sum\nolimits_n \left(\lambda_n{}^{1/2} U^*{}_n\right)^* X \left(\lambda_n{}^{1/2} U^*{}_n\right) + ZX + XZ,$$

where $Z = -\lambda/2 = -\sum_n \left(-\lambda_n^{1/2} U^*{}_n\right)^* \left(\lambda_n^{1/2} U^*{}_n\right)/2$, hence T_t is a particular case of the considerate semigroups.

19.6 Conclusions

The need to build the non-commutative Markov processes was given by the evolution of probabilities in quantum mechanics. This paper aims to build these processes on antisymmetric Fock space where we do not have exponential commutative vectors and where the commutative property does not occur between operators describing disjoint time intervals. For this reason the processes are obtained as solutions of stochastic integral equations. This mathematical model creates the possibility to construct physical processes as stochastic integral equations solutions, being at the same time a new method of proving that certain processes are noncommutative. The model may be used in diffusion processes.

References

1. Lindblad, G.: On the generators of quantum dynamical semigroups. Commun. Math. Phys. **48**, 119–130 (1976)
2. Alicky, R.: Quantum Dynamic Semigroups and Applications. Lecture Notes in Physics 286, Part 1. Springer, Berlin (1987)
3. Gorini, V., Kossakowsky, A., Sudarshan, E.C.G.: Completely positive dynamical semigroups of n-level systems. J. Math. Phys. **17**, 821–825 (1976)
4. Stinespring, W.F.: Positive functions on C*-algebras. Proc. Am. Math. Soc. **6**, 211–216 (1955)
5. Evans, M.P.: Existence of quantum diffusions. Probab. Theory Related Fields **81**, 473–483 (1989)
6. Parthasarathy, K.R., Bhat, R.B.V.: Markov dilations of nonconservative dynamical semigroups and a quantum boundary theory. Annales de l'Institut Henri Poincaré, Probabilités et Statistique **30**, 601–652 (1995)
7. Cuculescu, I., Oprea, A.: Noncommutative Probabilities. Kluwer, Boston (1994)
8. Barnett, C., Streater, R.F., Wilde, I.F.: The Itô-Clifford integral II, stochastic differential equations. J. London Math. Soc. **27**, 373–384 (1983)
9. Serbanescu, C.: An Ito product formula for Fermion stochastic integrals. Scientific Bulletin Politehnica Univ. Bucharest **60**, 71–81 (1998)
10. Hudson, R.L., Parthasarathy, K.R.: Quantum Itô's formula and stochastic evolutions. Commun. Math. Phys. **93**, 301–323 (1984)
11. Applebaum, D.B., Hudson, R.L.: Fermion Itô's formula and stochastic evolutions. Commun. Math. Phys. **96**, 473–496 (1984)
12. Hudson, R.L., Parthasarathy, K.R.: Stochastic dilations of uniformly continuous completely positive semigroups. Acta Appl. Math. **2**, 353–378 (1984)

13. Serbanescu C.: Fermoin Stochastic Integrals of Continuous Processes. Analele Universităţii, Bucureşti, pp. 277–288 (2004)
14. Applebaum, D.B., Hudson, R.L.: Fermion diffusions. J. Math. Phys. **25**, 858–861 (1984)
15. Applebaum, D.: Fermion Itô's formula II the gauge process in fermion Fock space. Publications Res. Inst. Math. Sci., Kyoto Univ. **23**, 17–56 (1987)
16. Ikeda, N., Watanabe, S.: Stochastic Differential Equations and Diffusion Processes, 2nd edn. North Holland Mathematical Library, Amsterdam (1989)

Chapter 20
Basic Study on Contribution to Dynamic Stability by Large Photovoltaic Power Generation

Junichi Arai and Shingo Uchiyama

Abstract Large photovoltaic power generation facilities are installing in the world. These large power generation are required to contribute actively to ac power system operation. This paper presents basic study results for dynamic stability improvement by large photovoltaic power generation.

Keywords Photovoltaic • Dynamic stability • Mega solar • Damping control

20.1 Introduction

Photovoltaic power generation has many advantages as no CO_2 emission during power generation and no limitation for installation place, of course it does not generate power during night and lower power generation is obtained in the rain. Photovoltaic power generation is promoted in the world. In recent year large capacity of photovoltaic power generations are installed, it is called as Mega Solar. 7 MW and 13 MW [1] facilities are in operation near Tokyo, and larger facilities as 49 MW [2] and 82 MW [3] are planning in Japan. These large mega solar will affect power system on the matter of voltage stability, frequency stability and dynamic stability. This paper assumes 200 MW large mega solar located in near large generator and makes study on dynamic stability of the synchronous generator by controlling mega solar power. A damping control in mega solar is designed and its effect is demonstrated by simulation.

J. Arai (✉) • S. Uchiyama
Kogakuin University, 1-24-2, Nishi-Shinjuku, Shinjuku-ku, Tokyo, Japan
e-mail: arai@cc.kogakuin.ac.jp

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_20

20.2 Power System Model

A power system shown in Fig. 20.1 is assumed in which one synchronous generator connects to an infinite bus through two circuit transmission lines. The capacity of ac generator is 1,000 MW and transmission line is 550 kV and 250 km length. 200 MW photovoltaic power generation, mega solar, is connected to the sending side of the transmission line. Figure 20.2 shows AVR control block. Parameters of this system are listed in Table 20.1.

Figure 20.3 shows a configuration of a mega solar that consisted of an inverter, transformer and a dc voltage source [4]. The inverter has six switching arms and PWM logic with 2 kHz carrier wave is applied. In this study solar cell side is represented by a constant voltage source for simplicity.

Control circuit of the inverter is shown in Fig. 20.4. It has an ac voltage phase detector, an active power control, a reactive power control, and a pulse width modulation (PWM) logic. The voltage phase detector, PLL, is modelled referring [5]. The active power control has proportional and integral functions as shown in Fig. 20.5. The reactive power control has the same transfer functions as Fig. 20.5.

We apply a conversion block that converts outputs of active power control and reactive power control to a phase angle signal and an amplitude signal. The conversion block is represented by Fig. 20.6 that solves interaction between the active

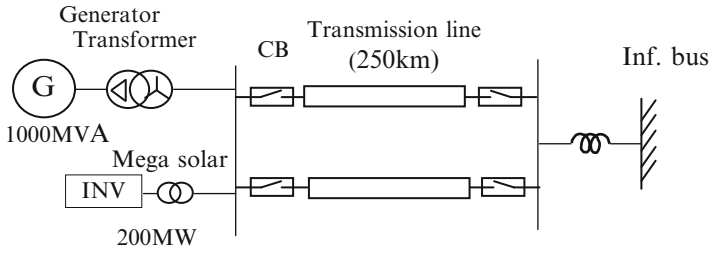


Fig. 20.1 Power system

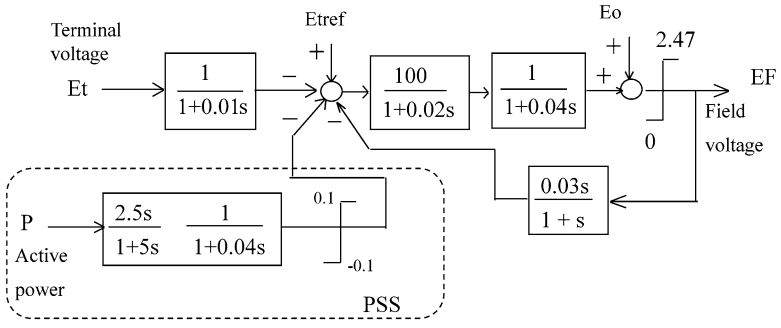


Fig. 20.2 Automatic voltage regulator for generator

Table 20.1 Parameters of power system

Components	Parameters
Generator	1,000 MVA,25 kV, 50 Hz
	Xd = 1.79 pu, Xd' = 0.169 pu, Xd" = 0.135 pu
	Xl = 0.13 pu, Xq = 1.71 pu, Xq' = 0.228 pu
	Xq" = 0.2 pu, Tdo' = 4.30 s, Tdo" = 0.032 s
	Tqo' = 0.85 s, Tqo" = 0.05 s, H = 3.59 s
Generator initial power	981.4 MW (0.981 pu), 269 Mvar
Transformer	1,000 MVA, 25 kV/550 kV, 11.65 %Z
Transmission line	250 km, 2 cct, 550 kV
	Pos. 0.01355 ohm/km, 0.823 mH/km
	0.01419 uF/km
	Zero 0.2392 ohm/km, 3.244 mH/km
	0.00526 uF/km
SCR of receiving ac system	26,800 MVA
Inverter of PV	200 MW, APR and AQR
Tr for inverter	200 MVA, 40 kV/550 kV, 5 %Z

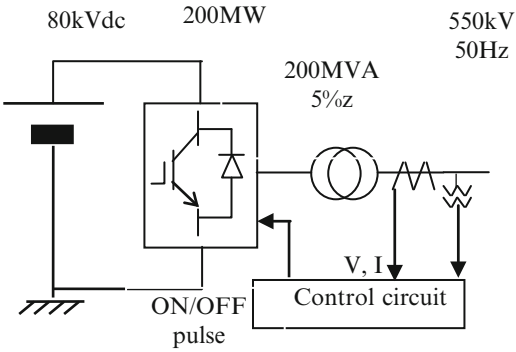


Fig. 20.3 Configuration of mega solar

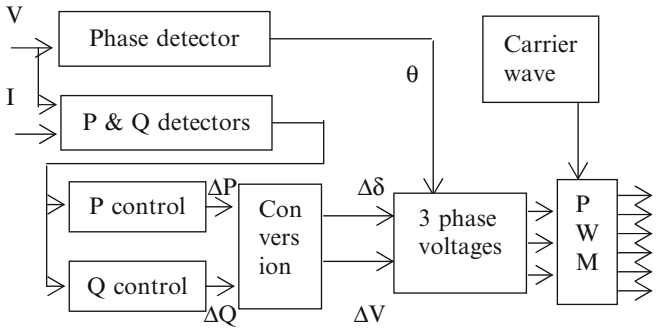


Fig. 20.4 Control block of mega solar

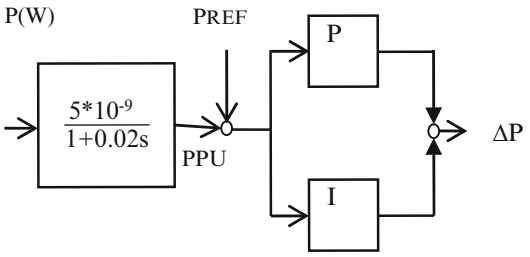


Fig. 20.5 Active power control

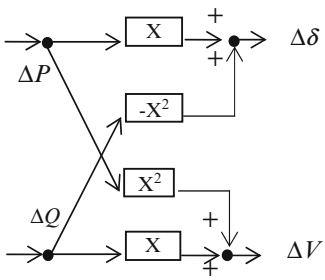


Fig. 20.6 Conversion block

power and the reactive power control loops [6, 7]. The phase angle and the amplitude signals are used to determine three phase voltage signals of the inverter output. These three voltage signals are fed to PWM logic with 2 kHz triangular carrier wave, six on/off pulses are generated and fed to switches in the inverter main circuit.

20.3 Design of Damping Control

Generator response at one transmission circuit opening without the mega-solar is show in Fig. 20.7. After disconnection of one circuit in two circuits, power swing of 1.3 Hz appears and no damping is observed. This means this system has a problem on dynamic stability. At this stage PSS in generator control is not applied.

In Fig. 20.8 a sine curve representing active power swing and a cosine curve are shown. Section A is duration of deceleration and section B is of acceleration of the generator shaft. If the shaft is accelerated in A and decelerated in B, the swing will be damped. It corresponds to the cosine curve with negative sign. An ideal damping is considered to be obtained by injection of active power that has 90° delay of 1.3 Hz power swing. A reasonable delay signal less than 90° is applicable actually.

A selected damping control is shown in Fig. 20.9. It has a high pass filter that removes dc component and a phase compensation block. This compensation function has 60° delay characteristic against 1.3 Hz, and gain 20 is selected as getting suitable damping effect. The input, P, of this damping control is the active power through the generator transformer, and output, ΔPd, is a modification signal for a reference signal, Pref, in the inverter active power control. Pref is given from MPPT control.

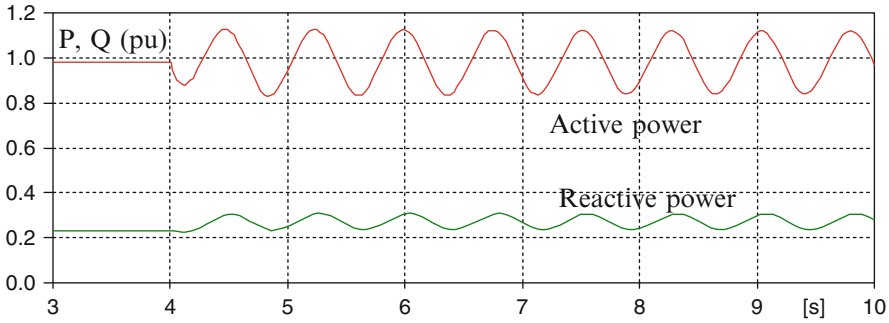


Fig. 20.7 Active power and reactive power of generator

Fig. 20.8 Active power and damping signal

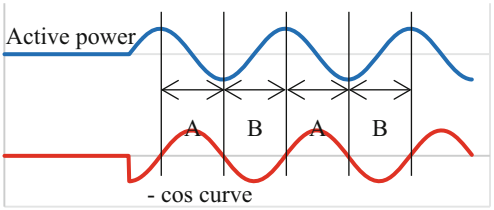
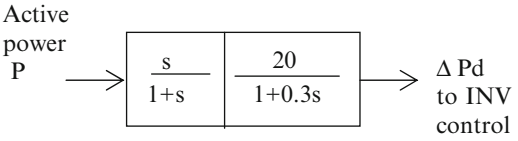


Fig. 20.9 Damping control



For the selection of the damping control time constant and gain, if 20 or 30° delay is applied enough damping effect is not obtained due to large phase deference. If 80° delay is applied, higher gain must be selected to get 1.3 Hz component and enough damping is not obtained also. As a result 60° delay is selected as mentioned above.

20.4 Simulation

Simulation for the power system is carried out by means of EMTP-ATPDraw.

20.4.1 Case 1: Pref = 1 pu with 1 pu limiter

Simulation results of one transmission circuit open are shown in Fig. 20.10, Case 1, in which the damping control is added. The active power reference, Pref = 1.0 pu based on the inverter capacity, and +1.0 pu limiter is added after summation of

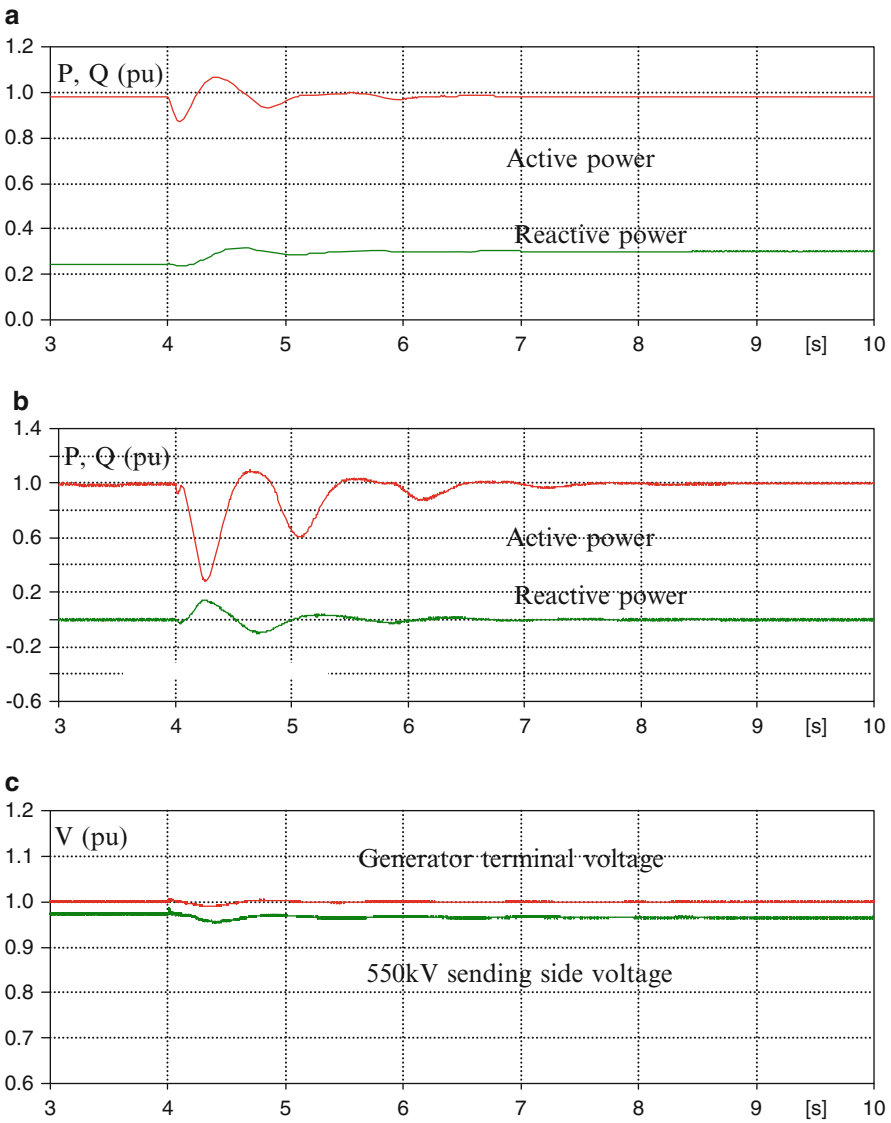


Fig. 20.10 Case 1. (**a**) Generator power. (**b**) Mega solar power. (**c**) Voltage

Pref + ΔPd. Normally a mega solar has no margin over 1.0 pu, so +1.0 limiter is placed. The reactive power is controlled to keep 0 Mvar for operation of pf = 1.

From Fig. 20.10, Case 1, one transmission circuit is opened at t = 4 s. The active power and the reactive power are well damped by the inverter control as shown in Fig. 20.10a. The inverter active power changes largely after one circuit is opened shown in Fig. 20.10b, but it stabilize the generator power sing. It has very good

performance. However the active power limit is set by 1.0 pu at the control circuit, inverter active power of 1.08 pu is observed in very short time. The ac voltages shown in Fig. 20.10c stay almost in the previous amplitude, it is also satisfactory.

20.4.2 Case 2: Pref = 0.8 pu with 1 pu Limiter

If the power level of the mega solar is 0.8 pu, more effective damping is estimated because there is 20 % margin for damping control action band. Case 2 is for this condition, results are shown in Fig. 20.11.

From Fig. 20.11, however the active power of the mega solar changes larger than Case 1, the generator power swing is almost same as Case 1. It means it is not necessary to reduce active power reference, Pref, after disturbances to get control margin. The reason why similar effects are obtained is considered that just after line opened the mega solar reduces its power, it prevent generator acceleration then effective damping is obtained. It can explain the upper limit for Pref has no large effect.

20.4.3 Case 3: PSS Application

The effect of PSS in the generator voltage controller is examined as Case 3. The mega solar is operated without damping control, i.e., it is operated by constant active power control and constant reactive power control. From Fig. 20.12 similar effect is observed as the damping control by the mega solar. The ac voltage on generator terminal changes larger than other case, the reason why is PSS suppress power variation by modifying generator internal voltage.

20.4.4 Case 4: Both of Inverter Damping Control and PSS

Case 4 is calculated to see the interaction between PSS and the damping control of the mega solar. Both controller are in active. From Fig. 20.13, active power is damped as same as other cases. However prospected mischief is not observed, better performance is not obtained. It looks Case 1 is better than Case 4 in view of active power damping and ac voltage variation.

Study of damping effect for generator power swing is performed by simulation, and it shows power swing can be damped effectively. The mega solar does not generate power during night, so this damping is not utilized during night. If the inverter of the mega solar is disconnected from solar cells and operated as STATCOM [8, 9], damping effect will be obtained during night.

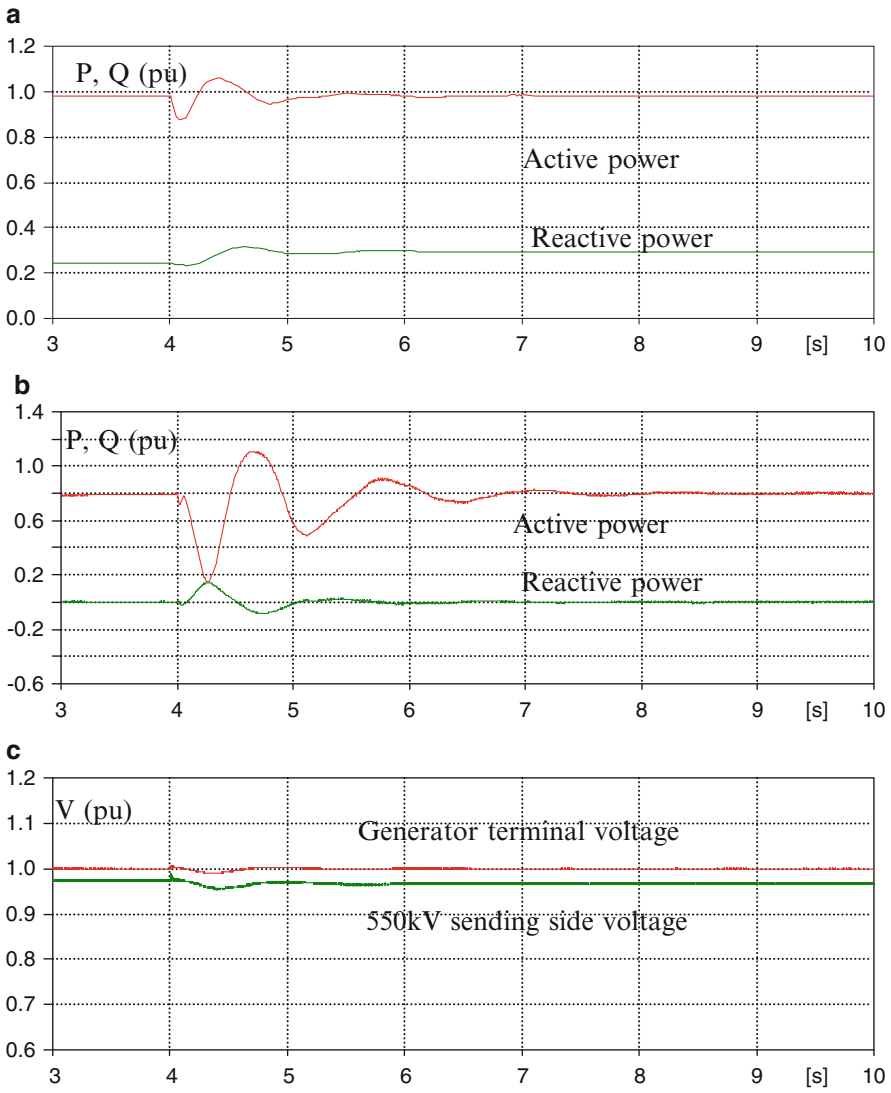


Fig. 20.11 Case 2. (**a**) Generator power. (**b**) Mega solar power. (**c**) Voltage

20.5 Conclusions

This study showed mega solar located near generator could supply damping effect on generator power swing, i.e., on dynamic stability. The damping effect is better than conventional PSS.

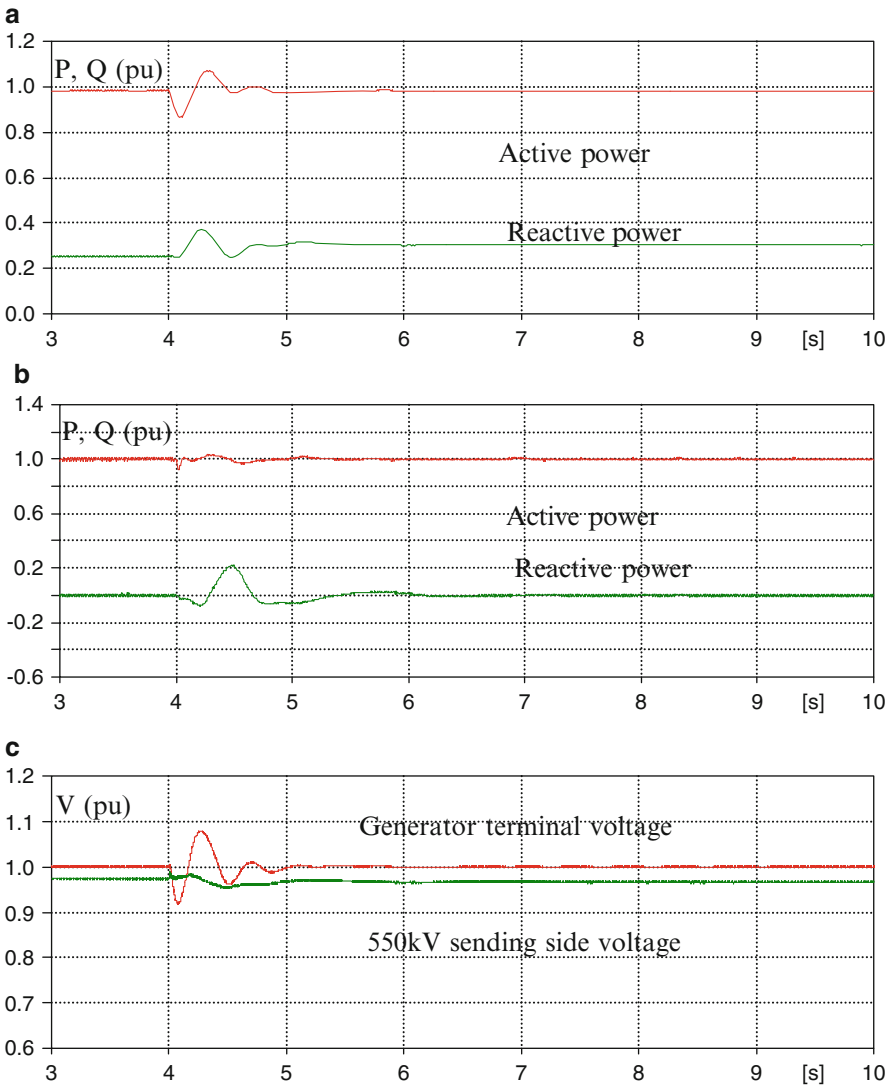


Fig. 20.12 Case 3. (**a**) Generator power. (**b**) Mega solar power. (**c**) Voltage

There is possibility the mega solar is located at middle point of transmission line not close to generator, so as future study damping control of mega solar at middle point of transmission line will be performed. Transient response after fault will be studied further.

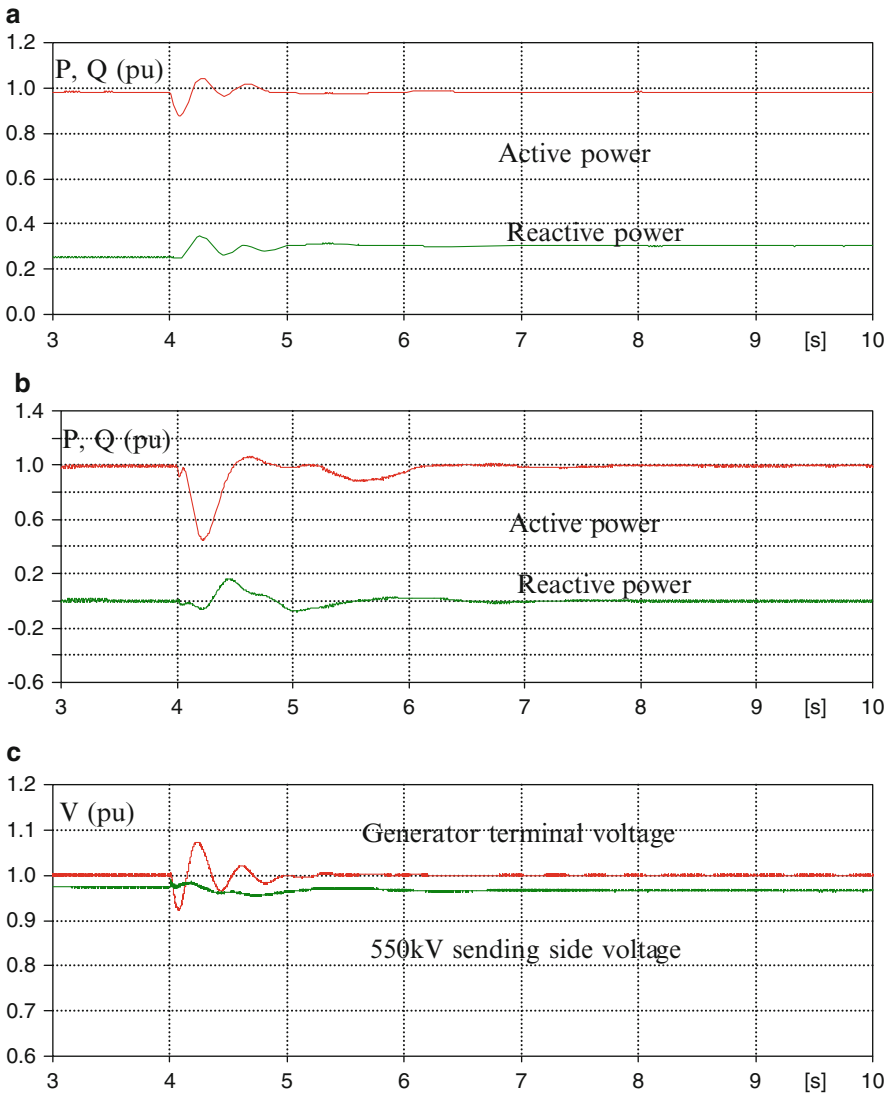


Fig. 20.13 Case 4. (**a**) Generator power. (**b**) Mega solar power. (**c**) Voltage

References

1. http://www.city.kawasaki.jp/kurashi/category/29-4-3-2-0-0-0-0-0.html, Kawasaki city mega sola
2. http://www.marubeni.co.jp/dbps_data/news/2012/121022.html, Oita city mega sola
3. http://www.marubeni.co.jp/news/2013/release/00044.html, Kisomisaki mega sola
4. Haberlin, H.: Photovoltaics System Design and Practice. Wiley, Chichester (2012)

5. Sood, V.K.: HVDC and FACTS Controllers. Kluwer Academic, Boston (2004)
6. Mori, T., Arai, J., Tsumenaga, M.: Development of stability analysis of inverter control for renewable energy. International conference on electrical engineering 2012, pp. 608–613. 8–12 July 2012
7. Mohammed Hassan, A., Arai, J.: Analysis of ac transmission from desert area large scale photovoltaic generation, EEIC2013, No. 103. Hong Kong, 24–25 Dec 2013
8. Fujii, T., Chisyaki, H., Teramoto, H., Sato, T., Matsushita, Y., Shinki, Y., Funahashi, S., Morishima, N.: Performance of the +/− 80MVA GCT STATCOM under commercial operation. IEE J. Trans. Industry Appl. **128**(4), 354–360 (2008)
9. Akedani, T., Hayashi, J., Temma, K., Morishima, N.: 450 MVA STATCOM installation plan for stability improvement, CIGRE 2010 B4-207 (2010)

Chapter 21
Exploring the Design Space of Signed-Binary Adder Cells

David Neuhäuser

Abstract Arithmetic based on signed-binary number representation is an alternative to carry-save arithmetic. Both offer adders with word-length independent latencies. Comparing both approaches requires optimized adder cells. Small and fast full adder designs have been introduced. A thorough investigation of signed-binary adder cells is still missing. We show that for an example signed-binary encoding scheme the design space consists of 2^{38} different truth tables. Each represents a bit-level signed-binary adder cell. We proposed a new method to enumerate and analyze such a huge design space to gain small area, low power, or low latency signed-binary adder cells and show the limitations of our approach.

Keywords Signed-digit • Signed-binary • Adder cell optimization • Digital design space exploration

21.1 Introduction

Arithmetic based on signed-binary number representation is an alternative to carry-save arithmetic, because both offer adders with word-length independent latencies. Not frequently implemented in state-of-the-art hardware designs, signed-binary is recurring in prototype development, with promising performance improvements under restricted constraints.

Implemented signed-binary arithmetic is based on signed-binary adder cells (SBACs). When designing signed-binary adder cells with arithmetic decomposition [1], various encoding decisions have to be made. The free combination of encodings and resulting degrees of freedom, gained by the arithmetic decomposition, yields a huge design space. Choosing the best combination of encodings and calculation schemes in terms of area, latency, and power consumption, is therefore a rather

D. Neuhäuser (✉)
Department of Computer Science and Mathematics
Friedrich Schiller University Jena
07737 Jena, Germany
e-mail: david.neuhaeuser@uni-jena.de

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_21

difficult task and can not be accomplished manually by the designer, nor can it be computed easily. We demonstrate the need of optimizing SBACs and present concepts of a systematic design space exploration of these cells.

We review signed-binary arithmetic in the following section. In Sect. 21.3, we introduce the concept of signed-binary design space exploration. In Sect. 21.4, we discuss different algorithmic approaches and present our new algorithm. Section 21.5 discussed minimization techniques for the generated adder cell truth tables. In Sect. 21.6, we introduce an evaluation concept to be applied to the enumerated adder cell design space. We conclude in Sect. 21.7 with a summation of our results and conclude.

21.2 Preliminaries

21.2.1 Signed-Digit Number Representation

The Signed-Digit Number Representation (SDNR), representing integers using signed digits, is a redundant number representation. Be T_{sd} the number range and R_{sd} the set of representations with the interpretation function I_{sd}.

$$I_{sd} : R_{sd} \rightarrow T_{sd} \tag{21.1}$$

A number $s \in T_{sd}$ can have multiple representations $r \in R_{sd}$.

SDNR uses a denominational number system with fixed base $\mathbb{B}$. Every digit of the representations has its own sign. The set of possible digits Z_{sd} of the representations is consecutive.

$$Z_{sd} = \{z^* \in \mathbb{Z} : -\alpha \leq z^* \leq \beta\} \tag{21.2}$$

$\mathbb{Z}$ denotes the set of integers, $\alpha, \beta > 0$. Redundancy exists, when $\alpha + \beta + 1 > \mathbb{B}$. Has $r \in R_{sd}$ base $\mathbb{B}$ and at most l digits, the number range described through I is T_{sd}.

$$T_{sd} = [-\alpha \times \frac{\mathbb{B}^l - 1}{\mathbb{B} - 1}, \beta \times \frac{\mathbb{B}^l - 1}{\mathbb{B} - 1}] \cap \mathbb{Z} \tag{21.3}$$

21.2.2 Signed-Binary Number Representation

The special case of SDNR with $\mathbb{B} = 2$ and $\alpha = \beta = 1$ is known as Signed-Binary Number Representation (SBNR) [2]. For $\mathbb{B} = 2$ and $\alpha = \beta = 1$ we get $Z_{sd} = \{-1, 0, 1\}$. Every $z^* \in Z_{sd}$ can be interpreted as signed-binary digit.

A signed-binary number $A_{sb} \in R_{sb}$ is defined as

$$A_{sb} = (z^*_{l-1}, \ldots, z^*_0), z^*_i \in \{-1, 0, 1\}, 0 \leq i < l \tag{21.4}$$

$$I_{sb}(A_{sb}) = \sum_{i=0}^{l-1} 2^i \times z^*_i \tag{21.5}$$

therefore

$$R_{sb} = \{A_{sb} : A_{sb} = (z^*_{l-1}, \ldots, z^*_0), z^*_i \in \{-1, 0, 1\}, 0 \leq i < l\} \tag{21.6}$$

and by Eq. (21.3)

$$T_{sb} = [-(2^l - 1), 2^l - 1] \cap \mathbb{Z} \tag{21.7}$$

Be Z_{sb} the set of signed digits and Q their bit-level encoding with the interpretation function J_{sb}.

$$J_{sb} : Q \rightarrow Z_{sb} \tag{21.8}$$

At least two bits are needed to encode $\{-1, 0, 1\}$, because

$$|Z_{sb}| = 3 \text{ and } 2^1 < 3 < 2^2 \tag{21.9}$$

Let us fix our encoding scheme for now to two bits. We get $Q = \{q \in \{0, 1\} \times \{0, 1\}\}$. A digit $z^* \in Z_{sb}$ can have multiple bit-level encodings $q \in Q$. The set of all interpretation functions be

$$\mathbb{J}_{sb} = \{J_{sb} : Q \rightarrow Z_{sb}\} \tag{21.10}$$

Let q^1 and q^0 be the two bits with $q = (q^1, q^0), q \in Q$. J_{sb} be the function to decode q to z^*.

$$z^* = J_{sb}(q) = J_{sb}(q^1, q^0) \tag{21.11}$$

The number of encodings schemes $J_{sb} \in \mathbb{J}_{sb}$ is equal to $|\mathbb{J}_{sb}|$. The two bits q^1 and q^0 can have four different bit combinations:

$$q = (q^1, q^0) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\} \tag{21.12}$$

Since $|Z_{sb}| = 3$ and $|Q| = 4$, the fourth bit combination can be left unused ($\mathbb{J}_{sb3}$) or can be used to encode an already encoded digit ($\mathbb{J}_{sb4}$).

$$\mathbb{J}_{sb3} = \{J_{sb} : Q \rightarrow Z_{sb}\}, |J_{sb}| = 3 \tag{21.13}$$

$$\mathbb{J}_{sb4} = \{J_{sb} : Q \rightarrow Z_{sb}\}, |J_{sb}| = 4 \tag{21.14}$$

$$\oslash = \mathbb{J}_{sb3} \cap \mathbb{J}_{sb4} \tag{21.15}$$

$$\mathbb{J}_{sb} = \mathbb{J}_{sb3} \cup \mathbb{J}_{sb4} \tag{21.16}$$

Leaving the fourth bit combination unused, the amount of possible encoding schemes $|\mathbb{J}_{sb3}|$ is:

$$|\mathbb{J}_{sb3}| = \frac{n!}{(n-k)!} = \frac{4!}{(4-3)!} = 4! = 24 \tag{21.17}$$

Be $\mathbb{P}_{-1}$ the amount of permutations for $\{-1, -1, 0, 1\}$, $\mathbb{P}_0$ for $\{-1, 0, 0, 1\}$ and $\mathbb{P}_1$ for $\{-1, 0, 1, 1\}$.

$$\mathbb{P}_{-1} = \mathbb{P}_0 = \mathbb{P}_1 = \frac{4!}{2!} = 12 \tag{21.18}$$

$$\mathbb{P}_{-1} + \mathbb{P}_0 + \mathbb{P}_1 = 36 \tag{21.19}$$

Therefor $|\mathbb{J}_{sb4}| = 36$. The total amount of all possible encoding schemes is:

$$|\mathbb{J}_{sb}| = |\mathbb{J}_{sb3}| + |\mathbb{J}_{sb4}| = 24 + 36 = 60 \tag{21.20}$$

At bit-level, a signed-binary number is defined as A_{bl}

$$A_{bl} = (q_{l-1}, \ldots, q_0), q_i \in Q, 0 \leq i < l \tag{21.21}$$

$$K_{sb} : Q^l \rightarrow R_{sb} \tag{21.22}$$

$$A_{sb} = K_{sb}(A_{bl}) = (z^*_{l-1}, \ldots, z^*_0), z^*_i = J(q_i), 0 \leq i < l \tag{21.23}$$

21.2.3 Signed-Binary Adder

A signed-binary adder reduces two bit-level encoded (A_{bl}, B_{bl}) signed binary numbers (A_{sb}, B_{sb}) to one bit-level encoded (S_{bl}) signed binary number (S_{sb}) in a way, that the integer (S) represented by the output signed binary number is the numerical sum of the integers (A, B) represented by the input signed binary numbers.

The signed binary adder operation is defined as:

$$A_{sb} \square B_{sb} = S_{bl} \tag{21.24}$$

$$I_{sb}(A_{sb}) = A \tag{21.25}$$

$$I_{sb}(B_{sb}) = B \tag{21.26}$$

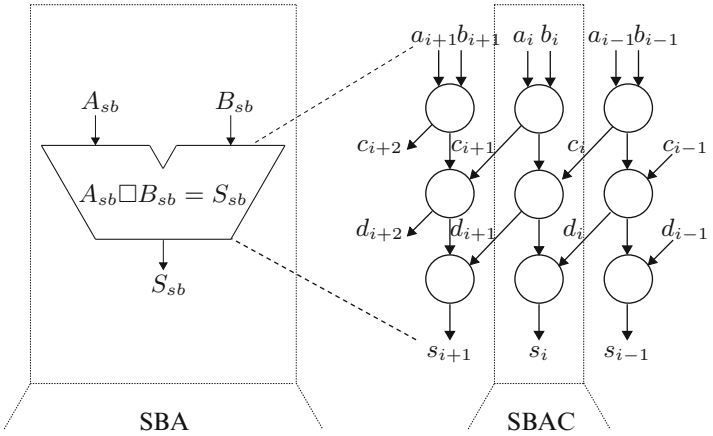


Fig. 21.1 SBA consisting of three-level SBACs shown at numerical-level [3, p. 111]

$$I_{sb}(S_{sb}) = S \tag{21.27}$$

$$A + B = S \tag{21.28}$$

Figure 21.1 shows the signed-binary adder operation and its decomposition.

The bit-level adder operation, conducted by a signed-binary adder, is defined as:

$$A_{bl} \diamond B_{bl} = S_{bl} \tag{21.29}$$

$$K_{sb}(A_{bl}) = A_{sb} \tag{21.30}$$

$$K_{sb}(B_{bl}) = B_{sb} \tag{21.31}$$

$$K_{sb}(S_{bl}) = S_{sb} \tag{21.32}$$

$$A_{sb}\square B_{sb} = S_{sb} \tag{21.33}$$

Figure 21.2 shows the bit-level decomposition.

By using a signed-binary adder at bit-level we can calculate an addition of integer values:

$$A_{bl} \diamond B_{bl} = S_{bl} \tag{21.34}$$

$$I_{sb}(K_{sb}(A_{bl})) = A \tag{21.35}$$

$$I_{sb}(K_{sb}(B_{bl})) = B \tag{21.36}$$

$$I_{sb}(K_{sb}(S_{bl})) = S \tag{21.37}$$

$$A + B = S \tag{21.38}$$

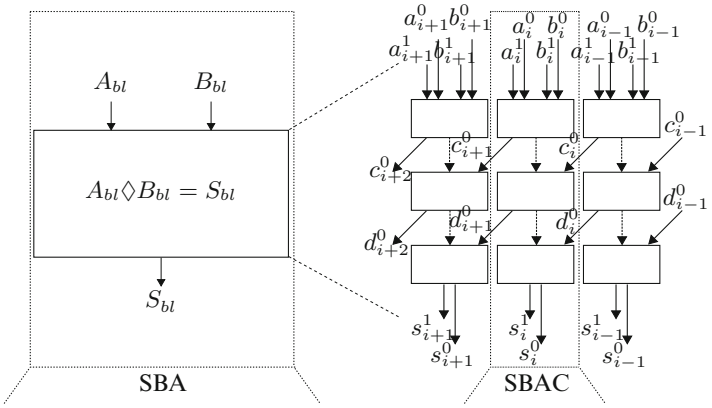


Fig. 21.2 SBA consisting of three-level SBACs [3, p. 111], bit-level

21.2.4 Signed-Binary Adder Cell

To actually calculate $A_{sb} \square B_{sb} = S_{sb}$ we decompose this operation into digit operations, recall Fig. 21.1.

$$A_{sb} = (a^*_{l-1}, \ldots, a^*_0) \tag{21.39}$$

$$B_{sb} = (b^*_{l-1}, \ldots, b^*_0) \tag{21.40}$$

$$S_{sb} = (s^*_{l-1}, \ldots, s^*_0) \tag{21.41}$$

$$S_{sb} = A_{sb} \square B_{sb} \tag{21.42}$$

One operation at digit i calculates s^*_i. $a_i, b_i \in \{-1, 0, 1\}$, $a_i + b_i \in \{-2, -1, 0, 1, 2\}$, and $s_i \in \{-1, 0, 1\}$. We need some "carry" to propagate $\{-2, 2\}$ to the digit at $i + 1$. Focusing on a third-level design as in Chow and Robertson [3], we introduce $c^*_i \in \{-1, 0\}$ and $d^*_i \in \{0, 1\}$ as a solution. We include c^*_i and d^*_i in Eq. (21.42):

$$C_{sb} = (c^*_l, \ldots, c^*_0) \tag{21.43}$$

$$D_{sb} = (d^*_l, \ldots, d^*_0) \tag{21.44}$$

$$S_{sb} = A_{sb} \square B_{sb} \square C_{sb} \square D_{sb} \tag{21.45}$$

We get

$$\sum_{i=0}^{l-1} 2^i \times s^*_i =$$

$$\sum_{i=0}^{l-1} 2^i \times (a^*_i + b^*_i + c^*_i + d^*_i - 2 \times c^*_{i+1} - 2 \times d^*_{i+1}) \tag{21.46}$$

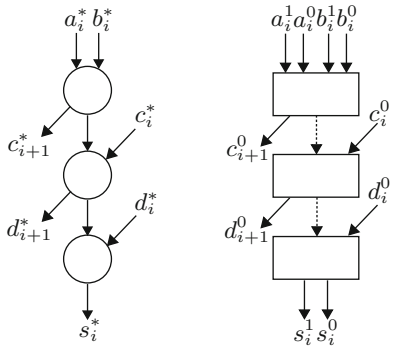


Fig. 21.3 SBAC, bit- and numerical-level

c_0^* and d_0^* are the carry-ins of the whole adder, set to 0 in normal operation. The carry-outs of the whole adder are c_l^* and d_l^*.

The signed-binary adder cell (SBAC) calculates the arithmetic decomposition of Eq. (21.46) at digit i, see Fig. 21.3.

$$s_i^* + 2 * c_{i+1}^* + 2 * d_{i+1}^* = a_i^* + b_i^* + c_i^* + d_i^* \tag{21.47}$$

The calculation of c_{i+1}^* must be independent from c_i^* and d_i^*. The calculation of d_{i+1}^* must be independent from d_i^*, see again Fig. 21.1. This guarantees locally constraint carries and allows fully digit parallel addition.

We denote calculation independence as $\perp$. The independence requirement of c_{i+1}^* and d_{i+1}^* can now be described as

$$c_{i+1}^* \perp c_i^* \tag{21.48}$$

$$c_{i+1}^* \perp d_i^* \tag{21.49}$$

$$d_{i+1}^* \perp d_i^* \tag{21.50}$$

By enforcing these independencies, the remaining carry chain is locally constraint, the calculation of any s_i^* depends only on $a_i^*, b_i^*, a_{i-1}^*, b_{i-1}^*, a_{i-2}^*, b_{i-2}^*$ [4].

We have to describe a SBAC at bit-level to implement it. c_i^* and d_i^* have each two states, either to propagate or not-propagate a carry. Therefore is one bit for each sufficient to encode c_i^* and d_i^*. Either one of c^* and d^* is $\in \{-1, 0\}$, the other one $\in \{0, 1\}$.

Let $P = \{0, 1\}$ be the bit-level set, and $Z = \{-1, 0, 1\}$ the signed binary level set of possible carries. Let L_{sb} be the according bit-level interpretation function of c^0 for carry c^* and M_{sb} respective of d^0 for carry d^*.

$$L_{sb} : P \rightarrow Z_{sb} \tag{21.51}$$

$$c^* = L_{sb}(c^0), c^0 \in P \tag{21.52}$$

$$M_{sb} : P \rightarrow Z_{sb} \tag{21.53}$$

$$d^* = M_{sb}(d^0), d^0 \in P \tag{21.54}$$

The bit-level description of a SBAC at digit i is in accordance to Eqs. (21.47) and (21.51)–(21.54):

$$\begin{aligned} &J_{sb}(s_i^1, s_i^0) + 2 * L_{sb}(c_{i+1}^0) + 2 * M_{sb}(d_{i+1}^0) = \\ &J_{sb}(a_i^1, a_i^0) + J_{sb}(b_i^1, b_i^0) + L_{sb}(c_i^0) + M_{sb}(d_i^0) \end{aligned} \tag{21.55}$$

The signed-digit carry independence of Eqs. (21.48)–(21.50) has to be enforced at bit-level. At bit-level, this independence is

$$c_{i+1}^0 \perp c_i^0 \tag{21.56}$$

$$c_{i+1}^0 \perp d_i^0 \tag{21.57}$$

$$d_{i+1}^0 \perp d_i^0 \tag{21.58}$$

21.3 Signed-Binary Design Space Exploration

Full adder CMOS designs have been extensively investigated [5]. All different FA designs are based on the same truth table (TT), interpreting any bit set to one as a numerical one and a bit set to zero as a numerical zero.

As for SBACs, the situation is different. We do not have one valid TT, rather a huge amount. We design a SBAC at the bit-level, recall Eq. (21.55) and Fig. 21.3. On one hand, we have the choice of the encoding schemes for all inputs and outputs. On the other hand, if we have fixed the encoding schemes, we have the choice of how to calculate any output bit. For example, do we actually need a^1 to calculate s^1?

J_{sb} of Eq. (21.8) is used to decode the bit-level to the numerical signed-binary level. We assume the same function J_{sb} to be used for the inputs $a_i^* = J_{sb}(a_i^1, a_i^0)$, $s_i^* = J_{sb}(s_i^1, s_i^0)$, and $s_i^* = J_{sb}(s_i^1, s_i^0)$. There are 60 choices for J_{sb}, and six L_{sb}, M_{sb}, recall Eqs. (21.51)–(21.54). Since some choices for L_{sb} and M_{sb} are mutual exclusive, we gain eight combination of choices for L_{sb} and M_{sb}. There are 60 choices for J_{sb}. In total, we have 480 choices of encoding function combination.

For each of these encoding schemes we have an enormous amount of possibilities on deciding, how to calculate each output digit. If we fix on one encoding scheme and on one calculation scheme, we get one TT to represent one SBAC. Stating such a TT for a SBAC is the same as stating the only one TT for a FA. For the FA are several implementations, meaning transistor netlists, possible. The same holds for any of the TTs representing a SBAC . This complexity is stated in Table 21.1.

Table 21.1 Complexity of adder cells

	Bits				Complexity	
	In	Out	Propagated	Reduced	TTs	Impl.
FA	3	2	1	1	1	≈ 6
SBAC	6	4	2	2	$\approx 2^{40}$	?

21.4 Algorithms for Truth Table Generation

Our approach is to generate all TTs representing a SBAC, without fixing the encoding scheme of intermediate data. Such a TT has 64 rows with 10 columns, as sketched in Table 21.2. Every row has to be correct with respect to the arithmetic decomposition at bit-level, recall Eq. (21.55). We have to guarantee that the generated values for c^0_{i+1} do not depend on c^0_i and d^0_i and the generated values for d^0_{i+1} do not depend on d^0_i.

Any TT structured as in Table 21.2 can be divided into blocks of four successive rows (indicated by thin lines in Table 21.2). Rows in a block have identical values for a^1_i, a^0_i, b^1_i, and b^0_i. Thus conditions (21.56) and (21.57) together are equivalent to the condition that c^0_{i+1} must be the same in all four rows of each block.

A block can be further subdivided into two sub-blocks of size two (indicated by dashed lines in Table 21.2). Both rows of a sub-block have also identical values for c^0_i. Thus condition (21.58) is equivalent to the condition that d^0_{i+1} is the same in both rows of each sub-block.

The TT shown in Table 21.2 is numerically correct w.r.t. the encoding functions from Table 21.3; moreover, all independence conditions are satisfied.

To generate all TTs representing a SBAC, we tried several algorithmic approaches. Straightforward algorithms proved not feasible, due to computing complexity. But we successfully implemented a sophisticated algorithm. To obtain an "optimal" SBAC, we have to evaluate all valid TTs according to a measure. This is done by the function *evaluate()* in the following algorithms and discussed in detail in Sect. 21.6.

Our first idea was a brute force enumeration of all truth tables. But since a truth table has 4×64 boolean degrees of freedom, we would have to check more than 10^{77} truth tables for validity, what is obviously beyond computational capacity.

Thus we switched to row-by-row truth table generation. Any specific (10-bit) instance of a row repeats in a huge number of truth tables. By checking each of the 16 possible outputs in a row for numerical correctness only once, we reduced the number of checks to only 1,024. But we observed that many rows still allowed more than one solution. Thus the number of valid truth tables grew exponentially in the row number, and we would have needed about 4 TB of main memory to hold them all. So we decided to keep, separately for each row, only the set of valid output vectors in memory and to generate the truth tables on the fly.

Table 21.2 SBAC example truth table

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0	j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0
...											
00	0 0 0 0 0 0	0	1	0	1	56	1 1 1 0 0 0	1	1	1	1
01	0 0 0 0 0 1	0	1	0	0	57	1 1 1 0 0 1	1	1	0	1
02	0 0 0 0 1 0	0	1	1	1	58	1 1 1 0 1 0	1	0	0	1
03	0 0 0 0 1 1	0	1	0	1	59	1 1 1 0 1 1	1	0	0	0
04	0 0 0 1 0 0	0	1	1	1	60	1 1 1 1 0 0	1	0	0	1
05	0 0 0 1 0 1	0	1	0	1	61	1 1 1 1 0 1	1	0	0	0
06	0 0 0 1 1 0	0	0	0	1	62	1 1 1 1 1 0	1	0	1	1
07	0 0 0 1 1 1	0	0	0	0	63	1 1 1 1 1 1	1	0	0	1
...											

Table 21.3 Example decoding function J_{sb} to encode a_i^*,b_i^*, and s_i^*, L_{sb} to encode c^*, and M_{sb} to encode d^*

q^1	q^0	$J_{sb}(q^1, q^0)$
0	0	−1
0	1	0
1	0	0
1	1	1

c_i^0	$L_{sb}(c_i^0)$
0	0
1	1

d_i^0	$M_{sb}(d_i^0)$
0	0
1	−1

Algorithm 1 Line-by-row SBAC truth table dependence graph generation

Require: test() to test numerical correctness
Require: .add() to add a row solution
Require: .enforce_independence() to enforce output-input independence if required
Require: evaluate() to recursively grade truth table

```
solutions ← clear
for input = 0 → 2^6 − 1 do
    newsolutions ← clear
    for output = 0 → 2^4 − 1 do
        if test(input, output) then
            newsolutions.add(output)
        end if
    end for
    solutions.enforceindependence(input, newsolutions)
end for
for i ∈ solutions do
    evaluate(i)
end for
```

As our final improvement, for each block of the truth table, we merge the output information for all rows in the block and remove those combinations that do not satisfy the independence conditions. We now describe this part of our method in detail.

We take as an example the encoding functions shown in Table 21.3. The first row $j = 0, a_i^1 = a_i^0 = b_i^1 = b_i^0 = c_i^0 = d_i^0 = 0$ of our TT represents $a_i^* = b_i^* = -1, c_i^* = d_i^* = 0$. For every $c_{i+1}^0, d_{i+1}^0, s_i^1$, and s_i^0 we calculate c_{i+1}^*, d_{i+1}^*, and s_i^*, according to Table 21.3 and test Eq. (21.55). Be $\sum \equiv 2c_{i+1}^* + 2d_{i+1}^* + s_i^*$.

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	$c_{i+1}^0 d_{i+1}^0 s_i^1 s_i^0$	$c_{i+1}^* d_{i+1}^* s_i^*$	$\sum$
0	0 0 0 0 0 0	0 0 0 0	0 0 −1	−1
0	0 0 0 0 0 0	0 0 0 1	0 0 0	0
0	0 0 0 0 0 0	0 0 1 0	0 0 0	0
0	0 0 0 0 0 0	0 0 1 1	0 0 1	1
0	0 0 0 0 0 0	0 1 0 0	0 −1 −1	−3
0	0 0 0 0 0 0	0 1 0 1	0 −1 0	−2

(continued)

(continued)

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	$c_{i+1}^0 d_{i+1}^0 s_i^1 s_i^0$	$c_{i+1}^* d_{i+1}^* s_i^*$	$\sum$
0	0 0 0 0 0 0	0 1 1 0	0 −1 0	−2
0	0 0 0 0 0 0	0 1 1 1	0 −1 1	−1
0	0 0 0 0 0 0	1 0 0 0	1 0 −1	1
0	0 0 0 0 0 0	1 0 0 1	1 0 0	2
0	0 0 0 0 0 0	1 0 1 0	1 0 0	2
0	0 0 0 0 0 0	1 0 1 1	1 0 1	3
0	0 0 0 0 0 0	1 1 0 0	1 −1 −1	−1
0	0 0 0 0 0 0	1 1 0 1	1 −1 0	0
0	0 0 0 0 0 0	1 1 1 0	1 −1 0	0
0	0 0 0 0 0 0	1 1 1 1	1 −1 1	1

For $j = 0$, $a_i^* + b_i^* + c_i^* + d_i^* + 0 = -2$. By Eq. (21.55), the TT is only numerically correct for $2c_{i+1}^* + 2d_{i+1}^* + s_i^* = -2$. We get two numerical correct rows for $j = 0$.

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0
0	0 0 0 0 0 0	0	1	0	1
0	0 0 0 0 0 0	0	1	1	0

Note that more than one bit combination c_{i+1}^0, d_{i+1}^0, s_i^1, and s_i^0 can be numerically correct. Let us look at all numerically correct solutions of the first four rows.

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0
0	0 0 0 0 0 0	0	1	0	1
		0	1	1	0
1	0 0 0 0 0 1	0	1	0	0
2	0 0 0 0 1 0	0	0	0	0
		0	1	1	1
		1	1	0	0
3	0 0 0 0 1 1	0	1	0	1
		0	1	1	0

Any combination of numerically correct rows forms a numerically correct TT. We can combine the first solution of row $j = 0$, the only solution of row $j = 1$, the first solution of row $j = 2$, and the first solution of row $j = 3$.

We could also combine the second solution of row $j = 0$, row $j = 1$, the third solution of row $j = 2$, and the second solution of row $j = 3$. Now we get the following first four rows of a numerical correct TT.

In total we get $2 * 1 * 3 * 2 = 12$ combinations looking at only the first four rows. These have to be reduced further by enforcing the independence rules of Eqs. (21.56)–(21.58). Equation (21.56) does not hold for the first solution of

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0
0	0 0 0 0 0 0	0	1	0	1
1	0 0 0 0 0 1	0	1	0	0
2	0 0 0 0 1 0	0	0	0	0
3	0 0 0 0 1 1	0	1	0	1

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0
0	0 0 0 0 0 0	0	1	1	0
1	0 0 0 0 0 1	0	1	0	0
2	0 0 0 0 1 0	1	1	0	0
3	0 0 0 0 1 1	0	1	1	0

row $j = 2$ with respect to any solution of row $j - 2 = 2 - 2 = 0$, because $d_{i+1,2}^0 = 0 \neq d_{i+1,0}^0 = 1$. We have to remove the second solution of row $j = 2$. Enforcing Eq. (21.56) we get the following solutions.

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0
0	0 0 0 0 0 0	0	1	0	1
		0	1	1	0
1	0 0 0 0 0 1	0	1	0	0
2	0 0 0 0 1 0	0	1	1	1
		1	1	0	0
3	0 0 0 0 1 1	0	1	0	1
		0	1	1	0

Equation (21.57) does not hold for any solution of row $j = 3$ with respect to the second solution of row $j - 1 = 3 - 1 = 2$, because $c_{i+1,3}^0 = 0 \neq c_{i+1,2}^0 = 1$. We have to remove the second solution (former third solution) of row $j = 2$.

j	$a_i^1 a_i^0 b_i^1 b_i^0 c_i^0 d_i^0$	c_{i+1}^0	d_{i+1}^0	s_i^1	s_i^0
0	0 0 0 0 0 0	0	1	0	1
		0	1	1	0
1	0 0 0 0 0 1	0	1	0	0
2	0 0 0 0 1 0	0	1	1	1
3	0 0 0 0 1 1	0	1	0	1
		0	1	1	0

Equation (21.58) does hold for any solution in any j, because $d_{i+1,j}^0 = 1$. We reduced the amount of solutions to $2 * 1 * 1 * 2 = 4$.

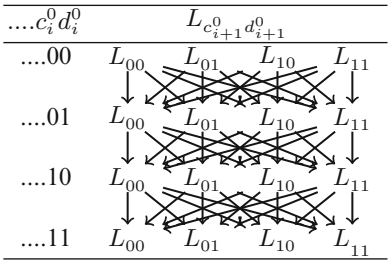


Fig. 21.4 No numerical correctness enforced. Maximum $4^4 = 256$ paths

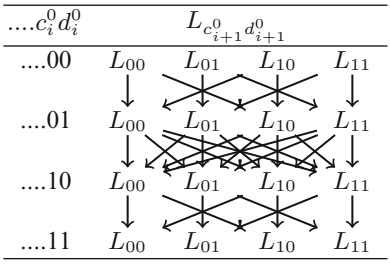


Fig. 21.5 No numerical correctness, but $d_{i+1}^0 \perp d_i^0$ enforced. With $d_{i+1}^0 \perp d_i^0 \Leftrightarrow (d_{i+1}(j) = 1 \Rightarrow d_i(j) = d_i(j-1))$, $4 \cdot 2 \cdot 4 \cdot 2 = 64$ paths

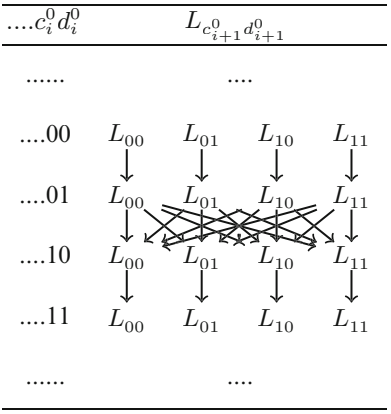


Fig. 21.6 No numerical correctness, but $d_{i+1}^0 \perp d_i^0$ and $c_{i+1}^0 \perp d_i^0$ enforced. With $c_{i+1}^0 \perp d_i^0 \Leftrightarrow (c_{i+1}(j) = 1 \Rightarrow d_i(j) = d_i(j-1))$, maximum $4 \cdot 1 \cdot 4 \cdot 1 = 16$ paths are left

We leave this example for now and take a more general approach. We divide our 64 rows into 16 blocks of four consecutive rows. In any block, we have to enforce Eqs. (21.56)–(21.58). Every four rows with Eqs. (21.56)–(21.58) form a *block result*. Between two adjacent rows, we do not have to enforce any independence constraints, because Eqs. (21.56)–(21.58) do not apply.

With the constraints of independence of Eqs. (21.56)–(21.58) we define possible traversing paths out of a solution set $L_{c_{i+1}^0 d_{i+1}^0}$ of row j of a previous row into a solution set $L_{c_{i+1}^0 d_{i+1}^0}$ of row $j+1$. Figures 21.4, 21.5, 21.6, 21.7 and 21.8 illustrate the generation of traversing paths.

We gain a very small memory usage. A valid TT is defined by a path from any member of the solution set of row 0 to any member of the solution set of row

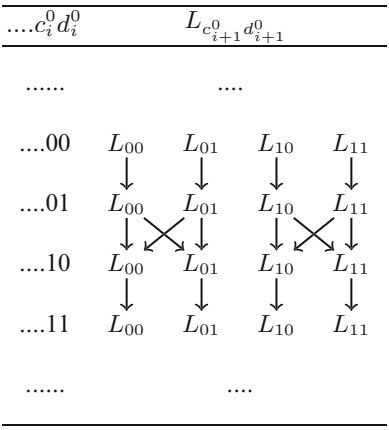


Fig. 21.7 No numerical correctness, but $d_{i+1}^0 \perp d_i^0$, $c_{i+1}^0 \perp d_i^0$, and $c_{i+1}^0 \perp c_i^0$ enforced. With $c_{i+1}^0 \perp c_i^0 \Leftrightarrow (c_{i+1}(j) = 1 \Rightarrow c_i(j) = c_i(j-1))$, maximum $4 \cdot 1 \cdot 2 \cdot 1 = 8$ paths are left

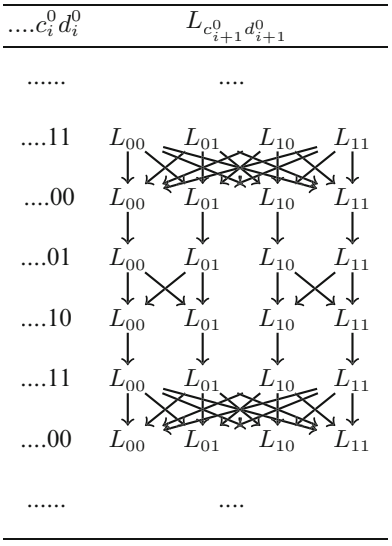


Fig. 21.8 In context to previous and following block

63, enforcing numerical correctness of Eq. (21.55). Traversing all paths gives all solutions.

For the given example functions of Table 21.3, we get the block results and the total amount of TTs in Fig. 21.9. Depending on the functions J_{sb},L_{sb}, and M_{sb}, the valid paths through the blocks and the total amount of solutions differ.

21.5 Minimization Techniques

Any valid TT can be expressed by VHDL source code, which can be synthesized. Since the synthesis process needs some time, a quick evaluation method is needed. TTs of SBACs consist of four columns, one for each of s_i^1,s_i^0,c_{i+1}^0, and d_{i+1}^0.

Fig. 21.9 Block solution statistics and total solutions for example encoding functions of Table 21.3

block	valid paths
0	4
1	4
2	4
3	8
4	4
5	8
6	8
7	4
8	4
9	8
10	8
11	4
12	8
13	4
14	4
15	4
total (product)	2^{38} =274877906944

We can create a disjunctive normal form for each of the four and optimize them independently.

21.5.1 Minimization of Boolean Functions

Several algorithms exist to minimize a disjunctive normal form. Boolean transformation rules are hard to formalize and implement because they are step by step rules, resulting in a chain of equivalent transformations, hence for a computer they are not handy to use. They are useful as a pen and paper approach.

Karnaugh-Veitch-diagrams [6, 7] are hard to formalize and implement. They make increased use of trained human pattern recognition capabilities. For a computer they are also not very suitable and rather useful as a pen and paper approach, too. Their practicability is restricted to four, at most five literals, whereas we are dealing with six literals.

Trying to formalize Karnaugh-Veitch-diagrams leads to the Quine–McCluskey algorithm. The Quine–McCluskey algorithm is easy to formalize and implement and useful for many input literals. The algorithm has a high runtime complexity but for at most six literals it is sufficient.

The Espresso algorithm [8] is an improved Quine–McCluskey algorithm in terms of complexity with the disadvantage of being a heuristic approach. This makes it suitable for complex boolean expressions, where the Quine–McCluskey algorithm is runtime insufficient and where obtaining optimal results is a secondary goal.

Since we are looking for an automated processing of $\approx 2^{38}$ TTs having at most six input literals, Quine–McCluskey algorithm suits best.

21.5.2 Minimization of Multiple-Output Boolean Functions

We want to derive optimal gate netlists from our TTs, by merging equal parts of the four minimized boolean expressions of s_i^1, s_i^0, c_{i+1}^0, and d_{i+1}^0. This technique is used e.g. in optimal CMOS implementations of FAs [5].

This task is known as minimizing multiple-output boolean functions. The Quine–McCluskey algorithm has to be improved to be able to minimize multiple-output boolean functions. Multiple-output minimization based on a modified Quine–McCluskey algorithm has been discussed [9]. The use of decision trees has been suggested [10, 11] for the minimization of multiple-valued functions. A mapping procedure for two-level multiple-output logic minimization has been developed [12].

Our final aim is to create optimal SBAC CMOS supergates. Optimization of supergates described at transistor level has been thoroughly discussed [13] and refined [14–16].

21.6 Evaluation of Signed-Binary Adder Cell Truth-Tables

We can use minimized boolean functions and count boolean operators as a measure. The amount of operator reflects the expected area A and power consumption P, the amount of hierarchy levels of the operators reflects the expected critical path latency T of implemented SBACs. More exact characteristics can be measured by deriving A, T, and P from synthesis.

Both approaches take too much time to explore the whole design space. Instead of evaluating all available TTs, we suggest a two level randomized approach in 2.

This approach has *time* to pick some solutions at random and to quickly measure them, e.g. by counting logic operators. The best *maxamount* solutions are kept. After *time*, these solutions are synthesized and a more exact measure is applied. Both measures can now be compared and possible correlations can be observed. The best solution can be compared to available SBACs [3].

21.7 Results and Conclusion

To compare carry-save and signed-binary arithmetic designs, we need optimal implementations of signed-binary adder cells. Methods of systematic optimization

Algorithm 2 SBAC truth table evaluation approach

```
Require: time the experiment has to run
Require: maxamount of second level truth tables
Require: qickmeasure() function to quickly measure truth table
  firstlevel ← all truth tables
  secondlevel.clear(); limit ← 0; amount ← 0
  while time left do
    pickedtruthtable ← random truth table ∈ firstlevel
    while pickedtruthtable is tagged do
      pickedtruthtable ← pickedtruthtable.next
      while pickedtruthtable ∈ secondlevel do
        pickedtruthtable ← pickedtruthtable.next
      end while
    end while
    tag pickedtruthtable
    pickedmeasure ← quickmeasure(pickedtruthtable)
    if amount < maxamount then
      secondlevel.insertsorted(pickedtruthtable,
        pickedmeasure)
      maxamount + +
      limit ← max(limit, pickedmeasure)
    else
      if pickedmeasure < limit then
        secondlevel.removelast()
        secondlevel.insertsorted(pickedtruthtable,
          pickedmeasure)
        limit ← secondlevel.last.measure
      end if
    end if
  end while
  synthesize and measure secondlevel
```

are still lacking. By developing an enumeration algorithm and providing an adder cell evaluation concept, we narrowed this gap.

For an example signed-binary encoding scheme, the design space consist of 2^{38} different truth tables. To find optimal signed-binary adder cell implementations, the design space has to be explored.

We showed different approaches and pointed out, why some of them should not be implemented. We showed that the design space of SBACs can be (partially) traversed and evaluated to choose latency, area, or power optimal SBAC implementations. We presented and implemented an algorithm to traverse the design space of an SBAC. We also pointed out some methods to optimize the boolean function represented by the according TTs. We suggested an algorithm to be applied to our traversing approach to finally derive better SBACs as provided by literature.

This is the first step in optimizing signed-binary adder cell implementations, needed to provide competitive signed-binary computer arithmetic designs.

References

1. Carter, T.M., Robertson, J.E.: The set theory of arithmetic decomposition. IEEE Trans. Comput. **39**, 993–1005 (1990)
2. Avizienis, A.A. Signed-digit number representations for fast parallel arithmetic. IRE Trans. Electron. Comput. **10**(3), 389–400 (1961)
3. Chow, C.Y., Robertson, J.E.: Logical design of a redundant binary adder. In: Proceedings of the 4th Symposium on Computer Arithmetic, pp. 109–115 (1978)
4. Zehendner, E.: Reguläre parallele Addierer für redundante binäre Zahlsysteme. Technical Report, Report 255, Institut für Mathematik der University ät Augsburg (1992)
5. Alioto, M., Palumbo, G.: Analysis and comparison on full adder block in submicron technology. IEEE Trans. Very Large Scale Integr. VLSI Syst. **10**(6), 806–823 (2002)
6. Karnaugh, M.: The map method for synthesis of combinational logic circuits. Trans. AIEE **72**(9), 593–599 (1953)
7. Veitch, E.W.: A chart method for simplifying truth functions. In: Proceedings of the Association for Computing Machinery, Pittsburgh, May 1952, pp. 127–133
8. Rudell, R.L.: Multiple-Valued Logic Minimization for PLA Syntheis. Technical Report, UCB/ERL M86/65, Electrical Engineering and Computer Science Department, University of California (1986)
9. Agrawal, P., Agrawal, V.D., Biswas, N.N.: Multiple output minimization. In: Proceedings of the 22nd Design Automation Conference, June 1985, pp. 674–680
10. Cerny, E., Mange, D., Sanchez, E.: Synthesis of minimal binary decision trees. IEEE Trans. Comput. **C-28**(7), 472–482 (1979)
11. Lloris, A., Gomez, J.F., Roman, R.: Using decision trees for the minimization of multiple-valued functions. Int. J. Electron. **75**(6), 1035–1041 (1993)
12. Rushdi, A.M., Ba-Rukab, O.M.: A purely map procedure for two-level multiple-output logic minimization. Int. J. Comput. Math. **84**, 1–10 (2007)
13. Malik, A.A.: Optimization of primitive gate networks using multiple output two-level minimization. In: Proceedings of the 29th ACM/IEEE Design Automation Conference (DAC '92), pp. 449–453. IEEE Computer Society Press, Los Alamitos (1992)
14. Kagaris, D., Haniotakis, T.: Transistor-level optimization of supergates. In: Proceedings of 7th International Symposium on Quality Electronic Design (ISQED '06), March 2006, pp. 685–690
15. Kagaris, D., Haniotakis, T.: Transistor-level synthesis for low-power applications. In: Proceedings of the 8th International Symposium on Quality Electronic Design (ISQED '07), March 2007, pp. 607–612
16. Liu, C.-P.L., Abraham, J.A.: Transistor level synthesis for static cmos combinational circuits. In: Proceedings of the 9th Great Lakes Symposium on VLSI, March 1999, pp. 172–175

Chapter 22
Green Element Solutions of the Source Identification and Concentration in Groundwater Transport Problems

Ednah Onyari and Akpofure Taigbenu

Abstract The inverse problem of source identification in groundwater contaminant transport poses challenges of non-uniqueness and instability of the numerical solutions. In this work, a methodology based on the Green element method (GEM), is presented for simultaneous recovery of the release history of pollution sources and concentration plume from available concentration measurements. The ill-conditioned, overdetermined system of equations that arises from the Green element discretization is solved by the least square method with Tikhonov regularization and aided by the singular value decomposition technique. The performance of the methodology is illustrated using two cases. The influence of the number of pollutant sources and their magnitudes are also examined. It is found that GEM is capable of correctly predicting the source strengths of pollutants instantaneously introduced into the aquifer and as well the concentration plume arising therefrom.

Keywords Green element method • Instantaneous pollution sources • Singular value decomposition • Tikhonov regularization

22.1 Introduction

The quality of groundwater can degrade due to pollutant sources that may be released continuously or instantaneously into the subsurface. The former type of pollution can arise from landfill sites, mine dumps, septic tanks, salt water

E. Onyari (✉)
Civil and Chemical Engineering, University of South Africa, Pretoria, South Africa

Civil and Environmental Engineering, University of the Witwatersrand, Johannesburg, South Africa
e-mail: onyarek@unisa.ac.za

A. Taigbenu
Civil and Environmental Engineering, University of the Witwatersrand, Johannesburg, South Africa
e-mail: akpofure.taigbenu@wits.ac.za

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_22

intrusion from the sea, and seepage from polluted streams. The latter may arise from unintentional spills of toxic substances into groundwater. Once an aquifer is contaminated it is important to monitor, manage and develop a remediation strategy for clean-up, and this requires concise quantitative understanding of the contaminant characteristics in terms of its historical distribution, and the location and strength of the sources.

The inverse groundwater contaminant transport problem has received much attention in the last few decades [1–7]. Broadly categorizing this problem, the first deals with identifying the source in terms of its location, release history, duration, start time, and the second relates to reconstructing the distribution of the plume from concentration observations. In this work, the Green element method (GEM), originally proposed by Taigbenu [8], is used in conjunction with the Tikhonov regularization method to calculate the strengths of pollution sources accidentally introduced into the aquifer and to reconstruct the historical contaminant distribution. We demonstrate the applicability of the methodology for single and multiple sources.

22.2 Governing Equation

The inverse contaminant transport problem addressed in this work is governed by the advection dispersion differential equation

$$R\frac{\partial C}{\partial t} = D\nabla^2 C - \nabla\cdot(\mathbf{V}C) \text{ on } \Omega. \tag{22.1}$$

where $\nabla = i\partial/\partial x + j\partial/\partial y$ is the two dimensional gradient operator in x and y, C is the contaminant concentration, $\mathbf{V} = \mathbf{i}u + \mathbf{j}v$ is the pore velocity vector with components u and v in the x and y directions, D is the hydrodynamic dispersion coefficient and R is the retardation factor. The solution to this inverse problem requires solving Eq. (22.1) subject to the following boundary conditions;

$$C(x, y, t) = f_1 \text{ on } \Gamma_1. \tag{22.2a}$$

$$-D\nabla C\cdot\mathbf{n} = q_1 \text{ on } \Gamma_2. \tag{22.2b}$$

$$\beta_1 C + \beta_2 D\nabla C\cdot\mathbf{n} = f_3 \text{ on } \Gamma_3. \tag{22.2c}$$

where $\mathbf{n}$ is the unit outward pointing normal vector on the boundary, and β_1 and β_2 are constants. The inverse problem is presented by F instantaneous sources of unknown strengths S_m at positions (x_m,y_m) $(m = 1, 2, \ldots, F)$, and a part of the boundary Γ_4 where neither the concentration nor its flux is specified. Furthermore, there are P internal points (x_j,y_j) $(j = 1, 2, \ldots, P)$ where concentration measure-

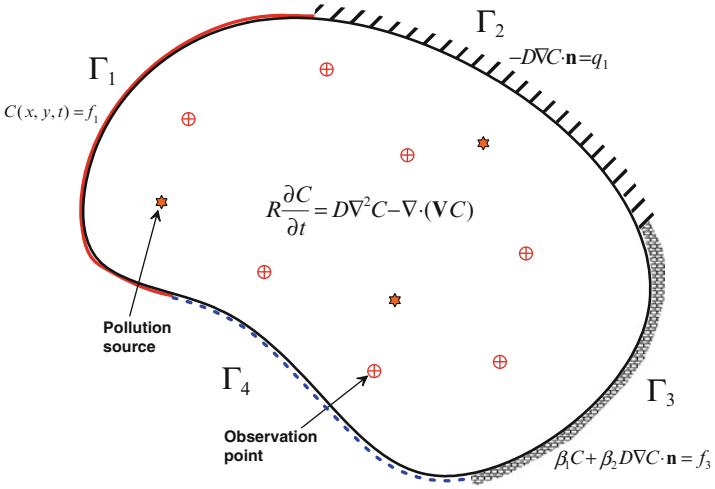


Fig. 22.1 Schematic of the problem statement

ments $C_j = C(x_j,y_j,t)$ are available. Figure 22.1 shows a schematic representation of the problem statement in a domain Ω with boundary $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$.

22.2.1 GEM Implementation

Green's second identity is applied to Eq. (22.1), using the fundamental solution of $\nabla^2 G = \delta(r - r_i)$ in the infinite space, to obtain its integral representation.

$$D\left[-\lambda C_{(r_i)} + \int_\Gamma C\frac{\partial G}{\partial n}ds\right] + \int_\Gamma Gq\ ds + \iint_\Omega G\left[\left(R\frac{\partial C}{\partial t} + \mathbf{V}\cdot\nabla C\right)\right]dA = 0. \tag{22.3}$$

where $r_i =$ source node and $\lambda =$ nodal angle at r_i and $q = -D\nabla C\cdot\mathbf{n}$. Equation (22.3) is implemented in the Green element sense by discretizing the computational domain into elements. Using rectangular elements, the quantities C, $\mathbf{V}$ and q are interpolated by linear interpolation functions in space ($C \approx N_j C_j$) so that Eq. (22.3) becomes.

$$V_{ij}C_j + L_{ij}q_j + W_{ij}\frac{\partial C_j}{\partial t} + X_{ikj}u_k C_j + Y_{ikj}v_k C_j = 0 \tag{22.4}$$

$$V_{ij} = D\left(\int_{\Gamma^e} N_j \nabla G_i \cdot \mathbf{n}\, ds - \delta_{ij}\lambda\right), \quad L_{ij} = \int_{\Gamma^e} N_j G_i\, ds,$$
$$\mathrm{X}_{ikj} = \iint_{\Omega^e} G_i N_k \frac{\partial N_j}{\partial x} dA, \quad \mathrm{Y}_{ikj} = \iint_{\Omega^e} G_i N_k \frac{\partial N_j}{\partial y} dA, \quad W_{ij} = R\iint_{\Omega^e} G_i N_j\, dA \tag{22.5}$$

where Ω^e and Γ^e are respectively the element domain and boundary. Aggregating the discrete element Eq. (22.5) for all elements gives a matrix equation of the form

$$E_{ij}C_j + L_{ij}q_j + W_{ij}\frac{dC_j}{dt} = 0. \tag{22.6}$$

where $E_{ij} = V_{ij} + \mathrm{X}_{ikj}u_k + \mathrm{Y}_{ikj}v_k$. A finite differencing of the temporal derivative is $dC/dt \approx [C^{(2)} - C^{(1)}]/\triangle t$ at the time $t = t_1 + \theta\triangle t$, where $0 \le \theta \le 1$ is the difference weighting factor, and $\triangle t$ is the time step between the current time t_2 and the previous one t_1. Introducing this approximation into Eq. (22.6) and weighting the other terms by θ gives

$$\left(\theta E_{ij} + \frac{W_{ij}}{\Delta t}\right)C_j^{(2)} + \theta L_{ij}q_j^{(2)} - \left(\omega E_{ij} + \frac{W_{ij}}{\Delta t}\right)C_j^{(1)} - \omega L_{ij}q_j^{(1)} = 0. \tag{22.7}$$

where $\omega = \theta - 1$ and the bracketed superscripts represent the times at which the quantities are evaluated. The third term in Eq. (22.7) with $C_j^{(1)}$ is usually taken in the direct problem to be known but for this inverse problem, there are F locations where it has to be calculated at $t = 0$. All known data from Eqs. (22.2a)–(22.2c), concentration measurements at P observation points and known initial data are incorporated into Eq. (22.7) to give

$$\mathbf{A}\mathbf{w} = \mathbf{b}. \tag{22.8}$$

$$\mathbf{A} = \begin{bmatrix} \theta E_{ij} + \frac{W_{ij}}{\Delta t} \\ \theta L_{ij} \\ -\left(\theta E_{ij} + \frac{W_{ij}}{\Delta t}\right) \end{bmatrix} \quad \text{and} \quad \mathbf{w} = \begin{Bmatrix} C_j^{(2)} \\ q_j^{(2)} \\ C_j^{(0)} \end{Bmatrix}. \tag{22.9}$$

where $\mathbf{w}$ is an $N \times 1$ vector of unknowns which include the F pollutant source concentrations. The matrix $\mathbf{A}$ is an $M \times N$ matrix, where M is the number of nodes in the computational domain and $M \ge N$. Equation (22.8) is over-determined and its solution is amenable to the least square method, while the matrix $\mathbf{A}$ is usually ill-conditioned and it is regularized by the Tikhonov regularization method. The decomposition of $\mathbf{A}$ is facilitated by the singular value decomposition (SVD) method.

$$\mathbf{A} = \mathbf{UDV}^t = \sum_{i=1}^{N} \psi_i u_i v_i^{tr} \tag{22.10}$$

where **U** and **V** are, respectively, $M \times M$ and $N \times N$ orthogonal matrices and **D** is an $M \times N$ diagonal matrix with N non-negative diagonal elements $\psi_1, \psi_2, \ldots, \psi_N$. The least square solution of Eq. (22.8) with the Tikhonov regularization minimizes the Euclidian norm $\| \mathbf{Aw} - \mathbf{b} \|^2 + \alpha^2 \| \mathbf{Iw} \|^2$ in calculating the solution for **w** that is given as

$$w(\alpha) = \sum_{i=1}^{N} \frac{\psi_i}{\alpha^2 + \psi_i^2} u_i^{tr} \mathbf{b} v_i \tag{22.11}$$

where u_i and v_i are the ith column of the matrices **U** and **V**, respectively and α is the regularization parameter whose choice is carefully made so that it is not too small to retain the instability of the numerical solution or too large to have smooth unrealistic solutions.

22.3 Results and Discussion

The problem of contaminant transport due to multiple instantaneous point pollution sources in a 2D homogeneous aquifer under a uniform flow velocity in the x-direction is solved by the inverse GEM formulation earlier described. The analytical solution for the historical concentration distribution in two dimensions from instantaneous sources is known and given in Bear [9] as:

$$C(x, y, t) = \sum_{m=1}^{F} \frac{S_m}{4\pi t D} \exp\left[-\frac{(x - x_m - ut)^2}{4Dt} - \frac{(y - y_m)^2}{4Dt} \right] \tag{22.12}$$

where S_m is the concentration of the pollution source that is instantaneously injected into the aquifer at (x_m, y_m). Two examples of this problem with single and double pollution sources are examined.

Table 22.1 Aquifer, pollutant and simulation parameters of Examples 1 and 2

Parameter	Example 1	Example 2
Spatial discretization: Δx, Δy	1.25, 1.0	1.25, 1.0
Source concentration, S_1	100	1,500
Source concentration, S_2	–	2,500
Source 1 location	(5.0,4.0)	(5.0,4.0)
Source 2 location	–	(5.0,7.0)

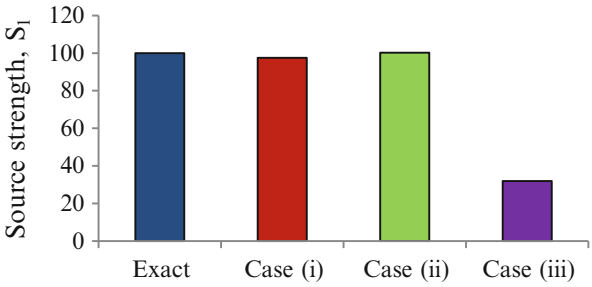


Fig. 22.2 Exact and computed pollution source strength of Example 1

22.3.1 Example 1

In this example a conservative pollutant is instantaneously injected at time $t = 0$ at the position ($x = 5.0$, $y = 4.0$). The GEM simulation is carried out in a 2-D rectangular domain [50×10] with concentration specified on all boundaries. Unit values are specified for u and D. Table 22.1 presents the pollutant and GEM simulation parameters that are used for this example. Three cases of observation points, located downstream of the pollutant source, are examined, namely case (i) along $x = 6.25$ and $x = 20$, case (ii) along $x = 7.5$ and $x = 20$, and case (iii) along $x = 8.25$ and $x = 20$. A uniform time step $\Delta t = 0.625$ and the fully implicit scheme with $\theta = 1$ are used in the GEM simulations. A value of 10^{-4} is used for the regularization parameter. Figure 22.2 shows the exact and computed pollutant source strengths. The result shows that GEM correctly predicts the pollution source strength when the lead observation points are in close proximity (Cases (i) and (ii)). This is reasonable, considering that sufficient information on the concentration should be available at observation points to support the prediction capability of the numerical scheme. The distribution of the concentration at $t = 10$ are presented as contour plots for both the exact and GEM solution in Fig. 22.3. It is observed that the exact concentration distribution is correctly predicted by GEM.

22.3.2 Example 2

This example has two conservative pollutants positioned at ($x = 5.0$, $y = 4.0$) and ($x = 5.0$, $y = 7.0$) that are instantaneously injected at time $t = 0$. Unit values are taken for u and D. The pollutant and GEM simulation parameters are given in Table 22.1. As in the first example, the GEM simulation is carried out in a 2-D rectangular domain [50×10] in which the concentration is specified on all boundaries. Observation points are located downstream of the pollutant source along

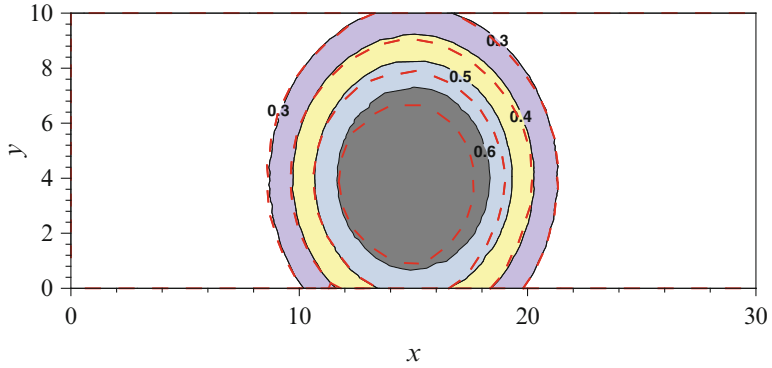


Fig. 22.3 Distributions of the concentration at $t = 10$ for Example 1: *graded colour* shows the exact, and the *red dashed line* shows GEM solution

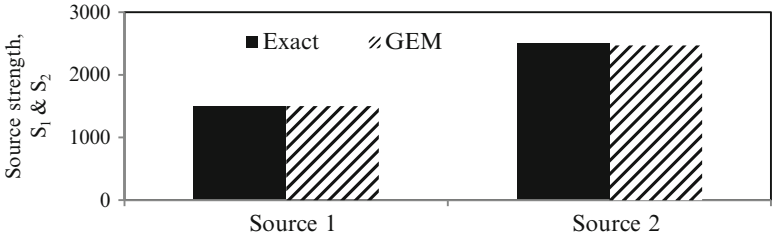


Fig. 22.4 Exact and computed source strength of Example 2

$x = 6.25$, and $x = 20$. A uniform time step $\Delta t = 0.625$, the fully implicit scheme with $\theta = 1$, and a regularization parameter value of 10^{-3} are used in the GEM simulation.

With the doubling of the number of pollutant sources whose strengths are more than tenfold that of Example 1, the excellent prediction of the source strengths by GEM are shown in Fig. 22.4. The contour plots of the concentration distribution at times $t = 5$, and $t = 15$ are presented in Fig. 22.5a, b. The results indicate improved estimation of the peak concentration with increase in time.

22.4 Conclusion

The Green element method has been used to simultaneously recover the source release history and to reconstruct the plume's historical distribution. The prediction capability of GEM is enhanced by having observation points located in close proximity of the sources and as well as ensuring that they are fairly well distributed over the computational domain. Using two illustrative examples, the promise of

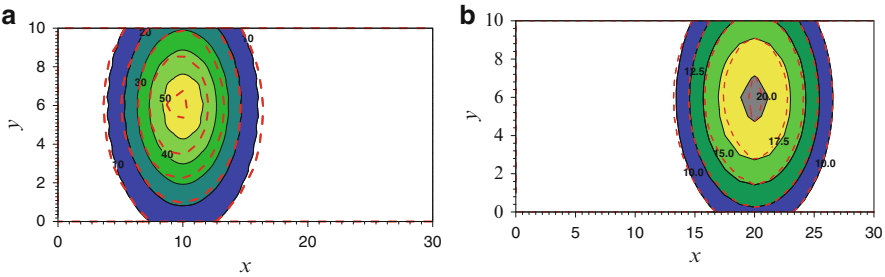


Fig. 22.5 Spatial distribution of contaminant plume (**a**) time = 5, (**b**) time = 15

GEM in solving the inverse instantaneous pollution source problem has been demonstrated.

References

1. Gorelick, S.M., Evans, B., Remson, I.: Identifying sources of groundwater pollution: an optimization approach. Water Resour. Res. **19**, 779–790 (1983)
2. Wagner, B.J.: Simultaneous parameter estimation and contaminant source characterisation for coupled groundwater flow and contaminant flow and contaminant transport modelling. J. Hydrol. **135**, 275–303 (1992)
3. Skaggs, T.H., Kabala, Z.J.: Recovering the release history of a groundwater contaminant. Water Resour. Res. **30**, 71–79 (1994)
4. Skaggs, T.H., Kabala, Z.J.: Recovering the history of a groundwater contaminant plume: method of quasi-reversibility. Water Resour. Res. **31**, 2669–2673 (1995)
5. Mahar, P.S., Datta, B.: Identification of pollution sources in transient groundwater system. Water Resour. Manag. **14**, 209–227 (2000)
6. Michalak, A.M., Kitanidis, P.K.: Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling. Water Resour. Res. **40**, W08302 (2004)
7. Prakash, O., Datta, B.: Characterization of groundwater pollution sources with unknown release time history. J. Water Resour. Protect. **6**, 337–350 (2014)
8. Taigbenu, A.E.: The Green Element Method. Kluwer, The Netherlands (1999)
9. Bear, J.: Dynamics of Fluids in Porous Media. Elsevier, New York (1972)

Chapter 23
First Time Electronic Structure Calculation of Poly[μ_2-*L*-Alanine-μ_3-Sodium Nitrate (I)] Crystals with Non-linear Optical Properties

A. Duarte-Moller, E. Gallegos-Loya, and E. Orrantia Borunda

Abstract The abstract should summarize the contents of the paper and should Poly[μ_2-*L*-alanine-μ_3-nitrato-sodium(I)], *p*-LASN, crystals were grown by the slow evaporation at room temperature technique. The nominal size of the crystals obtained by the method was of 500 nm. Single Crystal Diffraction was carried out in order to determine atomic structure and refine its lattice parameter. The electronic structure was obtained by using the Becke-Lee-Yang-Part and Hartree-Fock approximations with hybrid exchange-correlation three-parameter functional and G-311**G(*dp*) basis set. After calculations the band gap obtained directly from the density of states was 2.72 eV. The total polarizability obtained was 70.7390, the value for the total hyperpolarizability is 56.0243 and the dipolar moment was 10.6364.

Keywords Electronic structure • Alanine • Second harmonic • Non-linear optic

23.1 Introduction

Some organic compounds exhibit large NLO responses and, in many cases, orders of magnitude larger than widely known inorganic materials. They also offer molecular

A. Duarte-Moller (✉)
Centro de Investigación en Materiales Avanzados, S. C., Miguel de Cervantes 120, Complejo Industrial Chihuahua, Chihuahua, Chih. 31109, Mexico

Universidad Tecnológica de Querétaro, Av. Pie de la Cuesta 2508, Col. Unidad Nacional, Querétaro, Qro. 76148, Mexico
e-mail: alberto.duarte@cimav.edu.mx

E.O. Borunda
Centro de Investigación en Materiales Avanzados, S. C., Miguel de Cervantes 120, Complejo Industrial Chihuahua, Chihuahua, Chih. 31109, Mexico

E. Gallegos-Loya
UVM Educación, S.C., Campus Guadalajara Sur, Periférico Sur # 8100, Col. Sta. Ma. Tequepexpan, Tlaquepaque, Jalisco. C.P. 45601, México

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_23

design flexibility and the possibility of a virtually unlimited number of crystalline structures [1, 2]. A number of such crystals, especially from the amino acid family, recently have been reported [3–6]. Some amino acid crystal with simple inorganic salts appear to be promising materials for second harmonic generation (SHG) [7].

Amino acids exhibit specific features such as (I) molecular chirality, which secures acentric crystallographic structures [8]; (II) absence of strongly conjugated bonds, leading to wide transparency ranges in the visible and UV spectral regions; (III) zwitterionic nature of molecules, which favors crystal hardness; (IV) Amino acids can be used as chiral auxiliaries for nitro-aromatics and other donor–acceptor molecules with large hyperpolarizabilities and (V) as a basis for synthesizing organic and inorganic compounds.

A series of studies on semi-organic amino acid compounds such as *L*-arginine phosphate, *L*-arginine hydrobromide, *L*-histidine tetrafluoroborate, *L*-arginine hydrochloride, *L*-alanine acetate [8] and glycine sodium nitrate [9] as potential NLO crystals have been reported. *L*-Alanine is an amino acid, and it forms a number of complexes when reacted with inorganic acid and salts to produce an outstanding material for NLO applications. It belongs to the orthorhombic crystal system (space group *P*212121) with a molecular weight of 89.09 and has a melting point of 297 °C.

All of the NLO molecular materials show a wide transparent window in an UV-vis spectrum and a non-centrosymmetric geometry. However these materials need to have an absolute value of the susceptibility, $\chi^{(2)}$, which is basically associated with the non-centrosymmetric crystal structure. This property is the analogous to the molecular property called first polarizability, β. Materials like GSN, has been shown the SHG signal when it was excited by an intense IR radiation of 1,064 nm, commonly obtained from a pulsed Nd-YAG laser.

This work reports the DOS, polarizability, dipolar moment and the first hyperpolarizability of the *p*-LASN.

23.2 Experimental Details

In order to begin with the computational calculations we sintetise a sample by the low evaporation at room temperature. As follow step a single crystal X-ray diffraction experiment was carried out an after it the unit cell was refined by using the following parameters:

$a = 5.388(9)$ Å, $b = 9.315(15)$ Å, $c = 13.63(2)$ Å,

$\alpha = \beta = \gamma = 90°$

This unit cell was used to conduct a search in the Cambridge Structural Database (version 5.30 plus four updates). A positive match was found in a work by Van Hecke et al. [10].

In this experiment, a single crystal of *L*-alanine sodium nitrate which measured approximately $0.3 \times 0.1 \times 0.1$ mm was mounted on a Bruker Kappa APEXII DUO diffractometer. With the crystal at 298 K, a small set of 36 frames were collected in

order to determine the unit cell. One hundred reflections from these 36 frames were harvested and were used to index and refine the unit cell.

In order to quantify the β parameter is convenient to use an ab-initio calculations, which is highly recommended as an excellent alternative method to design NLO molecules and also predict its electronic structure. Nevertheless the correct choice of the basis is the goal in this kind of calculations. The electronic structure was obtained by using the Becke-Lee-Yang-Part and Hartree-Fock approximations with hybrid exchange-correlation three-parameter functional and G-311**G(*dp*) basis set.

23.3 Results and Discussion

In the *p*-LASN structure the asymmetric unit consisted of one sodium and one nitrate ion and one *L*-alanine molecule.

The coordination geometry around the sodium atom was trigonal bipyramidal with three bidentate nitrate anions coordinating through their oxygen atoms and two *L*-alanine molecules, each coordinating through one carboxyl oxygen atom. Three nitrate anions were bidentate coordinating to the sodium atom (2.612 (2)–2.771 (2) Å) and form one plane which is parallel to the (1 1 0) plane. The third nitrate oxygen atoms were coordinating to other symmetry equivalent sodium atoms and extend the plane formed. Almost perpendicular to this plane, two *L*-alanine molecules are coordinating to the sodium atom, each through one carboxyl oxygen atom (2.3651 (16) and 2.3891 (17) Å). The other carboxyl oxygen atoms were coordinated to sodium atoms in the upper and lower planes, respectively. Hence, an infinite amount of planes parallel to (1 1 0) are formed by nitrate anions and sodium atoms. These planes are perpendicularly linked to each other by *L*-alanine molecules.

Intermolecular hydrogen bonds are observed between N1(H1A)···O(1)[1/2 + *x*,−1/2 − *y*,2 − *z*] (1.92 (4) Å), N1(H1B)···O(5)[1/2 + *x*,1/2 − *y*,2 − *z*] (2.10 (3) Å) and N1(H1C)···O(2)[1 + *x*,*y*,*z*] (1.87 (4) Å) and an intramolecular hydrogen bond is found for N1(H1B)···O(2) (2.44 (3) Å).

Figure 23.1 shows the expected molecule of *p*-LASN and in Fig. 23.2 appears the *p*-LASN crystal after geometry optimization. The geometry optimization was completed by the Becke–Lee–Yang–Parr hybrid exchange-correlation three-parameter functional (B3LYP) [11] and 6-311++G(d,p) basis set. Energy gap, total dipole moment, polarizability and the first hyperpolarizability were calculated at the HF, DFT (B3LYP) and MP2 levels. In MP2 method, the approximation known as frozen core electron correlation was employed.

As an initial step the calculation of the HOMO and LUMO orbitals in the molecule was done. These representations appears in Fig. 23.3a, b. These figures shown the orbitals distributions on each case, higher and lower orbitals. This property shows the high polarizability of the *p*-LASN molecule.

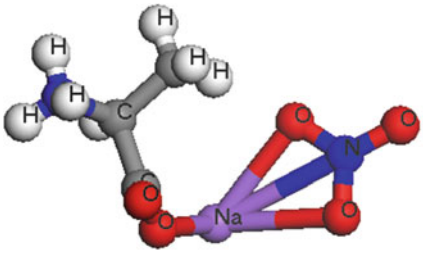


Fig. 23.1 *p*-LASN molecule. As we appreciate a bipyramidal structure is observed

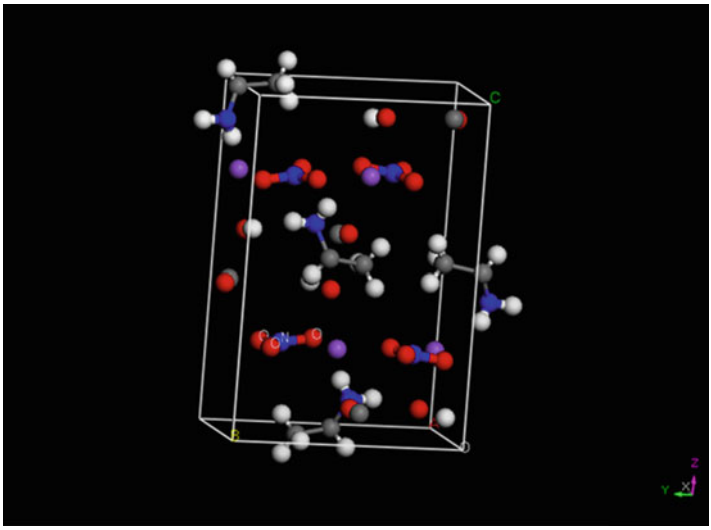


Fig. 23.2 *p*-LASN crystal generated by using the molecule from Fig. 23.1. *White* atoms are H, *red* atoms are O, *blue* atoms are N, and *purple* atoms are Na

The density of states calculation (Fig. 23.4) shows the distribution of s and d electrons in the energy bands. The overall distributions of states across the energy range of DOS are similar to that of GSN molecular crystals [12]. In that calculation it is observed that the conduction band above the Fermi level is occupied almost 80 % by *p* type electrons associated mainly with the NO_3 ion. The unoccupied states are principally due a mixture of *s* and *p* characters in the alanine molecule. An energy bandgap is observed at 2.2 eV above the Fermi level placed at 0 eV.

The optical activity is well determinated by the knowledge parameters α, β and μ [13]. At a molecular level the response for an isolated molecule under action of an electric field is given by

$$\mu_i = \mu_i^{(0)} + \sum_j \alpha_{ij} E_j + \sum_{jk} \beta_{ijk} E_j E_k + \sum_{jkl} \gamma_{ijkl} E_j E_k E_l$$

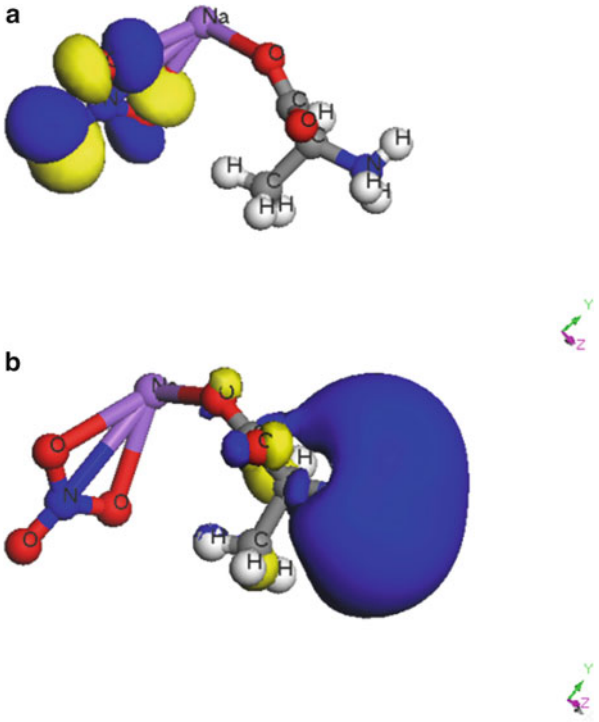


Fig. 23.3 (**a**) Representative HOMO for *p*-LASN. (**b**) Representative LUMO for *p*-LASN

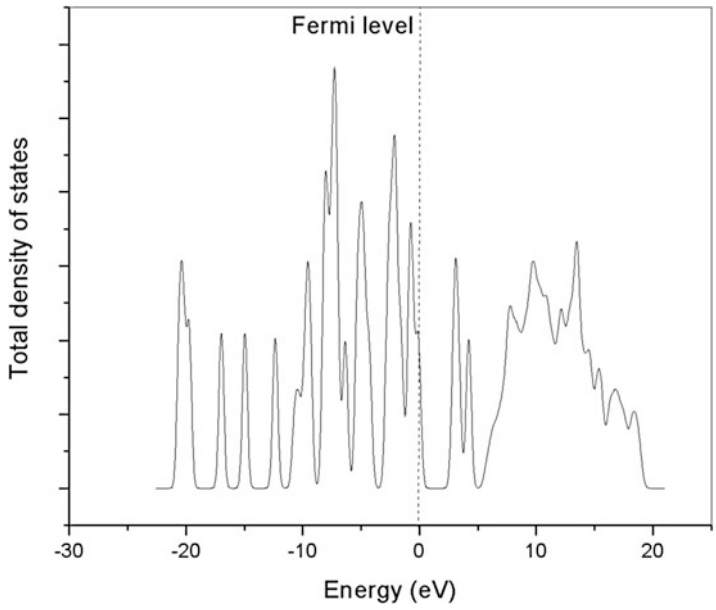


Fig. 23.4 Total DOS for *p*-LASN showing the distribution of s and d electrons in the energy bands

Table 23.1 Calculated values for α, β and μ

Polarizability α	Hyperpolarizability β	Dipolar moment μ
$\alpha_x = 66.7560$	$\beta_x = -180.1879$	$\mu_x = -10.5732$
$\alpha_y = 70.1644$	$\beta_y = -73.2246$	$\mu_y = 0.1100$
$\alpha_z = 75.2966$	$\beta_z = -201.4381$	$\mu_z = 1.1532$
$\alpha_{total} = 70.7390$	$\beta_{total} = 56.0243$	$\mu_{total} = 10.6364$

The correct value of the polarizability α is described by the second rank tensor, however the average value α_{total} can be obtained form

$$\alpha_{total} = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right)$$

and the magnitude of the first hyperpolarizability can be calculated from

$$\beta_{total} = \frac{1}{5}\left(\beta_x + \beta_y + \beta_z\right)$$

where the x, y and z components of β have been extracted from the 3D matrix generated by Gaussian 09 software.

Finally the results of calculation for the polarizability α, first hyperpolarizability β and the dipolar moment μ appear in Table 23.1. Unfortunately for *p*-LASN the values for this parameters have not been reported and is not possible to do a comparison, however the high value of the hyperpolarizability is enough to have a NLO phenomenon.

23.4 Conclusions

The structure of *p*-LASN was confirmed by using single crystal diffraction. Its lattice parameters were found to be:

$a = 5.388(9)$ Å, $b = 9.315(15)$ Å, $c = 13.63(2)$ Å, $\alpha = \beta = \gamma = 90°$

This unit cell was used to conduct a search in the Cambridge Structural Database (version 5.30 plus four updates). A positive match was found in a work by K. Van Hecke, E. Cartuyvels, T. N. Parac-Vogt, C. Gorller-Walrand, and L. Van Meervelt.

The coordination geometry around the Na atom was shown to be as trigonal–bipyramidal, with three bidentate nitrate anions coordinating through their O atoms and two *L*-alanine molecules each coordinating through one carboxylate O atom.

The use of the B3LYP (Becke-Lee-Yang-Part) supported by Gaussian 09 software with hybrid exchange-correlation three-parameter functional and G-311**G(dp) basis set and HF approximation were a good option to reproduce some electronic properties of the *p*-LASN sample.

Acknowledgments The authors thank the National Council of Science and Technology of Mexico for its financial support. Also, they thank the National Laboratory of Nanotechnology of CIMAV, S.C., Chihuahua, Mexico. The authors are very grateful to acknowledge M. Sci. Enrique Torres Moye (X-ray laboratory), M. Sci. Daniel Lardizabal (thermal analysis laboratory).

References

1. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. **147**, 195–197 (1981)
2. Vijayan, N., Rajasekaran, S., Bhagavannarayana, G., Ramesh Babu, R., Gopalakrishnan, R., Palanichamy, M., Ramasamy, P.: Cryst. Growth Des. **6**(11), 2441 (2006)
3. Rodrigues, J., Misoguti, L., Nunes, F.D., Mendonca, C.R., Zilo, S.C.: Opt. Mater. **22**, 235 (2003)
4. Ambujam, K., Selvakumar, S., Prem, A.D., Mohamed, G., Sagayaraj, P.: Cryst. Res. Tech. **41**, 671 (2006)
5. Ramesh Kumar, G., Gokul Raj, S., Mohan, R., Jeyavel, R.: Cryst. Growth Des. **6**, 1308 (2006)
6. Sethuraman, K., Ramesh Babu, R., Gopalakrishnan, R., Ramasamy, P.: Cryst. Growth Des. **8**(6), 1863 (2008)
7. Meera, K., Muralidharan, R., Dhanasekaran, R., Prapun, M., Ramasamy, P.: J. Cryst. Growth **263**, 510 (2004)
8. Mohankumar, R., Rajanbabu, D., Jayaraman, D., Jayavel, R., Kitamura, K.: J. Cryst. Growth **275**, 1935 (2005)
9. Narayan Bhat, M., Dharmaprakash, S.: J. Cryst. Growth **236**, 376 (2002)
10. Van Kristof, H., Els, C., Tatjana Parac-Vogt, N., Christiane, G.W., Luc, V.M.: Acta Cryst. E **63**, m2354 (2007)
11. Foresman, J.B.: Frisch, Exploring Chemistry with Electronic Structure Methods, 2nd edn. Gaussian Inc., Pittsburgh (1996)
12. Hernández-Paredes, J., Glossman-Mitnik, D., Esparza-Ponce, H.E., Alváres-Ramos, M.E., Duarte-Moller, A.: J. Mol. Struct. **875**(1), 295 (2008)
13. Ostroverkhov, V., Ostroverkhova, O., Petschek, R.G., Singer, K.D., Sukhomlinova, L., Twieg, R.J., Wang, S.-X., Chien, L.C.: Chem. Phys. **257**, 263 (2000)

Chapter 24
Aspects of Designing the Tracking Systems for Photovoltaic Panels with Low Concentration of Solar Radiation

Ionel Laurentiu Alboteanu, Florin Ravigan, Sonia Degeratu, and Constantin Şulea

Abstract The paper aims make contributions to the optimization of the photovoltaic conversion process by concentrating solar radiation and orientation of photovoltaic modules. The photovoltaic concentrating system aims to reduce expenses regarding photovoltaic surface and replace it with optical materials. Geometric design issues are presented for these systems adapted to the conditions of a certain geographical location.

Keywords Photovoltaic • Tracking system • Low concentrating photovoltaic

24.1 Introduction

Ideally, a PV panel should follow the sun so that the sun rays fall perpendicular to its surface, thus maximizing solar energy capture and thus we obtain the maximum output power. The tracking systems using controlled mechanisms that allow maximization of direct normal radiation received on PV panel [1].

The idea of concentrating solar radiation to generate electricity photovoltaic appeared almost simultaneously with photovoltaic science.

Operating principle of concentrating photovoltaic systems is based on using optics materials to focus sunlight on a photovoltaic surface, thus increasing the amount of energy captured and converted. To focus sunlight on photovoltaic surface and increasing thus the energy potential, the concentrating solar PV systems use either optical elements retractable (usually Fresnel lens) or reflective elements (usually mirrors) [2].

I.L. Alboteanu (✉) • F. Ravigan • S. Degeratu • C. Şulea
Faculty of Electrical Engineering, University of Craiova, 107, Decebal Bvd., Craiova 200440, Romania
e-mail: lalboteanu@em.ucv.ro; ravigan.florin@gmail.com; sdegeratu@em.ucv.ro; constantin.sulea@gmail.com

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_24

In the literature there are three types of solar concentrators: high, medium or low concentration [3]. This classification depends on the concentration ratio, defined as the ratio of the amount of radiation incident on the photovoltaic module and the amount of radiation available. In the paper is approached low concentrating photovoltaic system (LCPV).

24.2 Aspects of Celestial Mechanics and Modelling of Angular System

It is known that the Earth behaves a complete rotation in a year, around the Sun in an elliptical orbit and a complete rotation around its own axis during 24 h. Earth's rotation axis has a fixed direction in space and inclined angle $\delta_0 = 23.5°$ to the perpendicular plane of the orbit (Fig. 24.1). The angle between the direction to the Sun and the equatorial plane, δ is named declination and varies during the year from 23.4°, at the moment of the summer solstice (June 21) to −23.5°, at the winter solstice (December 21).

On March 21, respectively—September 21 declination $\delta = 0$ and the length of day and night are equal.

Declination can be calculated with the Copper formula [1]:

$$\delta = 23.45 \cdot \sin\left(360^\circ \frac{284 + n}{365}\right) \tag{24.1}$$

where n is the number of days in a year, the first day considering January 1.

Using monthly average of 'n' values is calculated declination of Earth during a year, and the resulting graph is represented in Fig. 24.2.

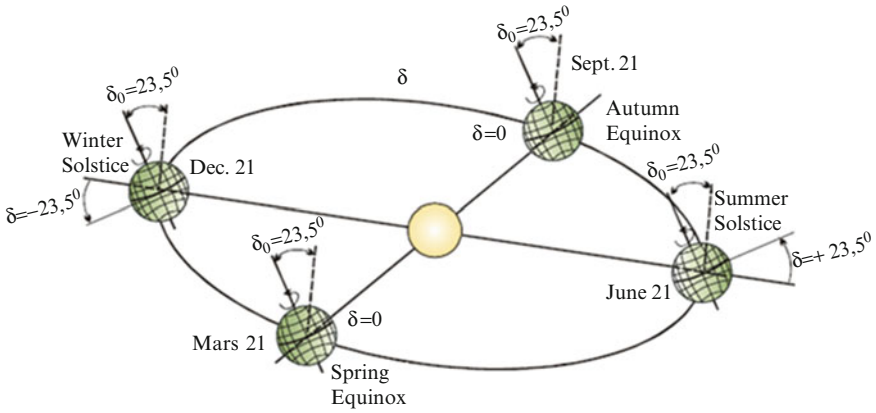


Fig. 24.1 Earth's orbit and the angle of declination, δ

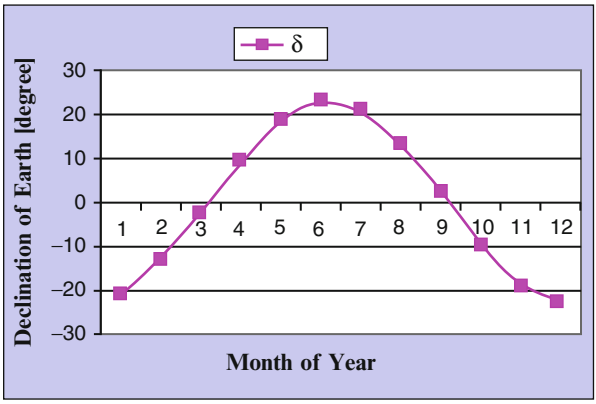


Fig. 24.2 Annual graphical evolution of declination angle of the Earth

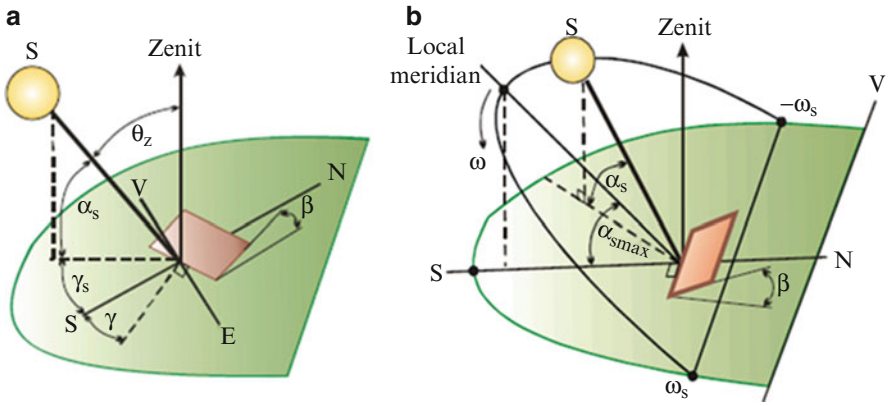


Fig. 24.3 Explanation regarding to sun's angles

Geometric relations between an arbitrarily oriented plane to the horizontal and direct sunlight that falls on the plan at any point of time, the position of the sun to this plan can be described in terms of several angles.

Latitude, φ—the angle measured from the equator to the point of interest on the earth's surface, is considered positive for the northern hemisphere and negative—to the south.

The inclination angle of plane β is the angle between the plane and the horizontal surface; $0 \leq \beta \leq 180$, (Fig. 24.3). For normal solar installations, maximum angle does not exceed 90°.

Azimuthally angle, γ—the angle between the projection on the horizontal plane perpendicular to the surface of the plan and the local meridian (Fig. 24.3); equal to zero for that plan south facing; negative—to east, positive—to west; $-180 \leq \gamma \leq 180$.

Solar azimuthally angle, γs—the angle between the south and the projection on the horizontal direct radiation (sunlight) (Fig. 24.3b); angles measured from south east direction are negative, the measured westward—positive.

The angle of elevation of the sun, αs—the angle between the horizon and the sun line linking point of interest, or the incident solar beam at the point of interest (Fig. 24.3).

Zenith angle, θz—the angle between the vertical and the line connecting the sun and the point of interest, or αs complementary angle (Fig. 24.3).

Hour angle, ω—determines the position of the sun in the sky at a given moment. Equals zero when crossing the local meridian sun, i.e. when midday positive and negative east–west (Fig. 24.3b). Accordingly, ωs corresponds to the angle of sunrise, and (−ωs), the angle of the sun dusk.

It is obvious that in an hour the sun across the sky at an angle equal to 15°, and his position at any time T is determined by the expression:

$$\omega = 15 \cdot (12 - T) \tag{24.2}$$

If you know the angles δ, φ and ω, then easily determine the position of the sun in the sky for the point of interest for any time, any day, using the expressions [1]:

$$\sin \alpha_s = \sin \delta \sin \varphi + \cos \delta \cos \varphi \cos \omega = \cos \theta_z \tag{24.3}$$

$$\cos \gamma_s = \frac{\sin \alpha_s \sin \varphi - \sin \delta}{\cos \alpha_s \cos \varphi} \tag{24.4}$$

In Eq. (24.3) by imposing the condition $\alpha_s = 0$, calculate east respectively west angle hourly of the sun from the relationship:

$$\omega_s = \pm \cos^{-1} (-\tan \varphi \cdot \tan \delta) \tag{24.5}$$

For every day of the year in (24.4) with declination δ previously determined from (24.1) for a time T is determined the hour angle ω and knowing the latitude φ is determined the sun elevation angle αs.

In Fig. 24.4 is presented photovoltaic panel, P directed to the south. Surface of panel P is inclined to the horizontal with β angle.

Solar radiation on the PV panel will be highest when the afternoon when the sun elevation angle, α_s is the maximum distance and sunlight will be minimal time and angle $\omega = 0$. This situation will occur where direct radiation is perpendicular to the surface of the PV panel, P.

Figure 24.4 shows that $\theta z = \beta$, and angle of the panel on S-N direction (elevation), from the horizontal plane is determined by the relationship:

$$\cos \theta_z = \sin \delta \sin \varphi + \cos \delta \cos \varphi = \cos (\varphi - \delta) \tag{24.6}$$

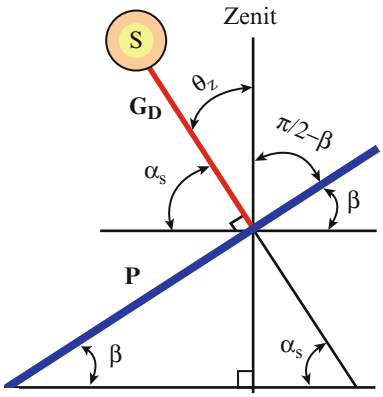


Fig. 24.4 Direct solar radiation on an inclined plane in midday: $\omega = 0$; $\gamma = 0$

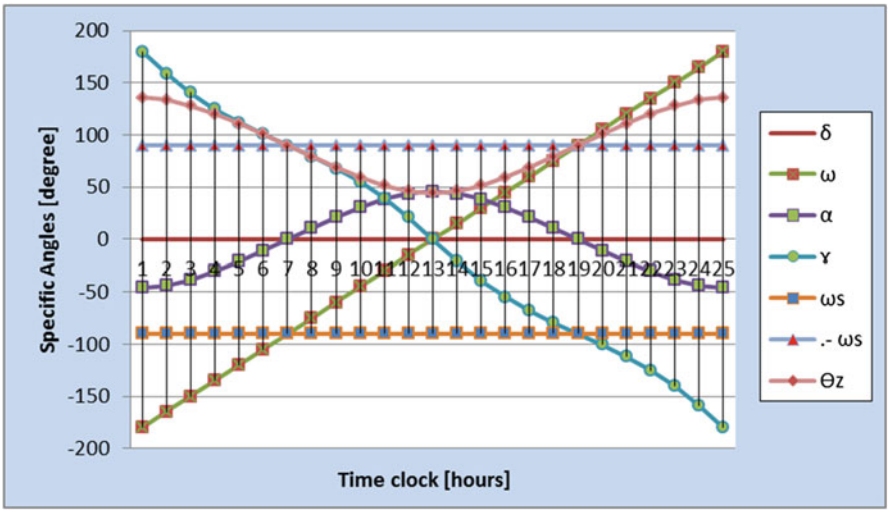


Fig. 24.5 Simulation results of angular system at equinox (March 21, September 21), for Craiova City

with:

$$\beta = \varphi - \delta \tag{24.7}$$

Based on present relationships above, customizing for city of Craiova was obtained graphical representation of specific angles describing the position of the sun in the sky.

In Fig. 24.5 is presented simulation results for the angular system at equinoxes. The same graphs that describing the angular system is shown in Fig. 24.6, for summer solstice and, for the winter solstice in Fig. 24.7.

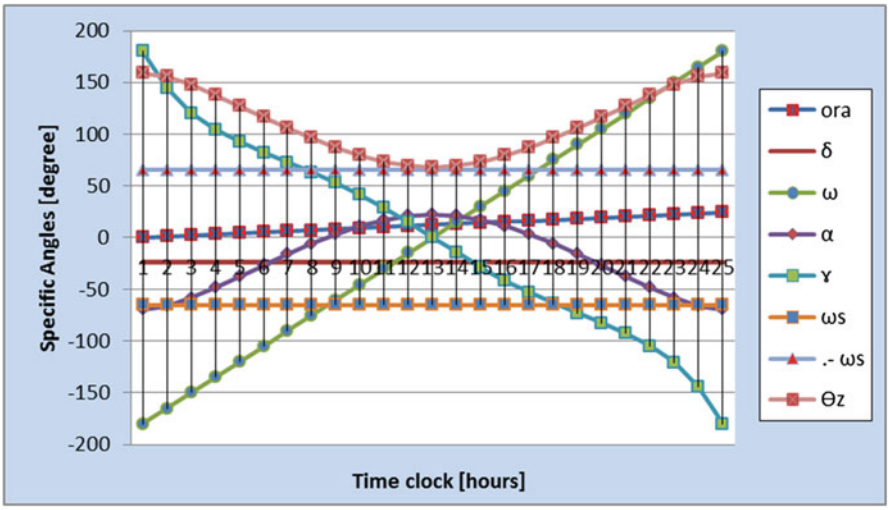


Fig. 24.6 Simulation results of angular system at summer solstice (June 21), for Craiova City

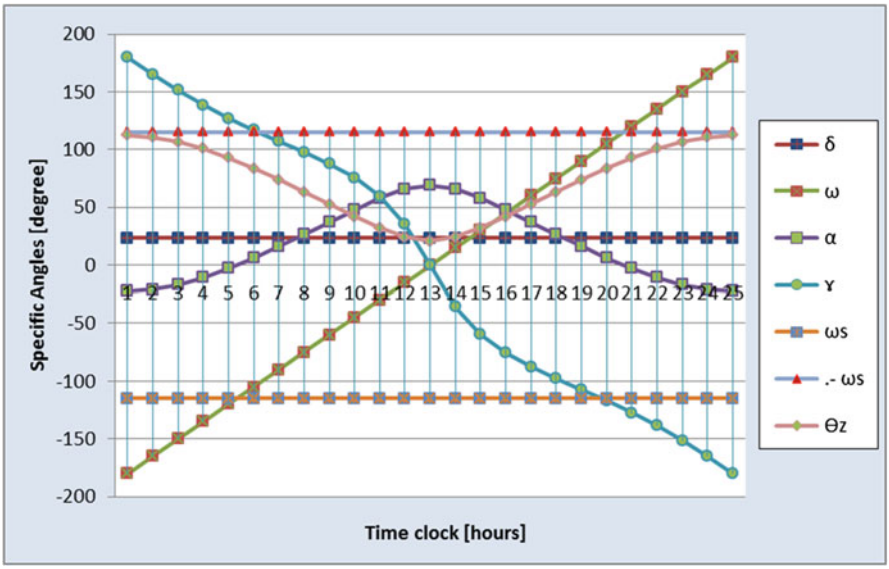


Fig. 24.7 Simulation results of angular system at winter solstice (December 21), for Craiova City

Analyzing the graphs resulted from the simulation of angular system customized for location Craiova, in all three cases, we can say that:

– sun altitude or elevation angle (αs) shows maximum values at midday 48° at the equinoxes, 68° at summer solstice and 25° at winter solstice;

– Sunset hour angle is 98° at equinox, at summer solstice is 110° and 60° at winter solstice;
– Sunrise hour angle is symmetric with the sunset hour angle having the same values but with a negative sign;
– Azimuth angle presents positive in the first half of the day dropping to zero during the midday, then afternoon is symmetrical values from the first half of the day;
– Hour angle shows negative peaks in the morning and will then drop to zero during the midday and afternoon shows symmetrical values of in the morning.

24.3 Modelling of Tracking System for Photovoltaic Panel

Electricity production of a PV system depends largely on solar radiation absorbed by the photovoltaic panels. As the sun changes with the seasons and over a day, the amount of radiation available for the conversion process depends on the panel tracking.

Since PV modules have a relatively low yield (up to 30 % in laboratory conditions) the aim is to optimize their energy [4, 5].

A method of optimizing available solar energy conversion with real possibilities of implementation is the use of tracking systems. Literature shows that the uses of tracking systems increase from 20 % to 40 % the amount of energy produced by converting [6].

In practice are two kinds of tracking systems: single axis and double axes tracking systems (Fig. 24.8).

It must be modelled and simulated the tracking system to determine the energy absorbed by the PV panel. Modelling and simulation was performed for double axes tracking system.

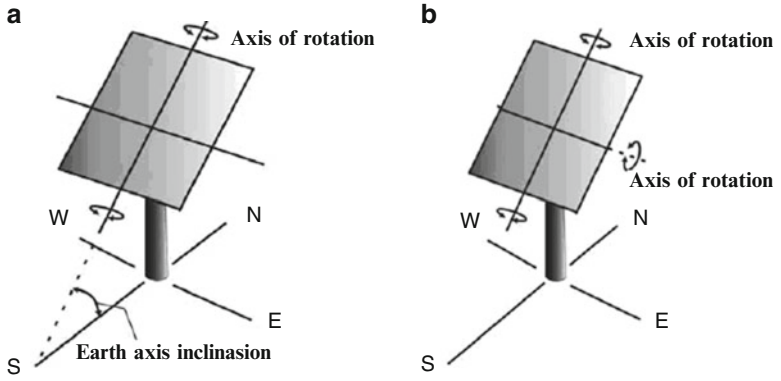


Fig. 24.8 Tracking systems of PV panel: (**a**) single axis and (**b**) double axes according to [7]

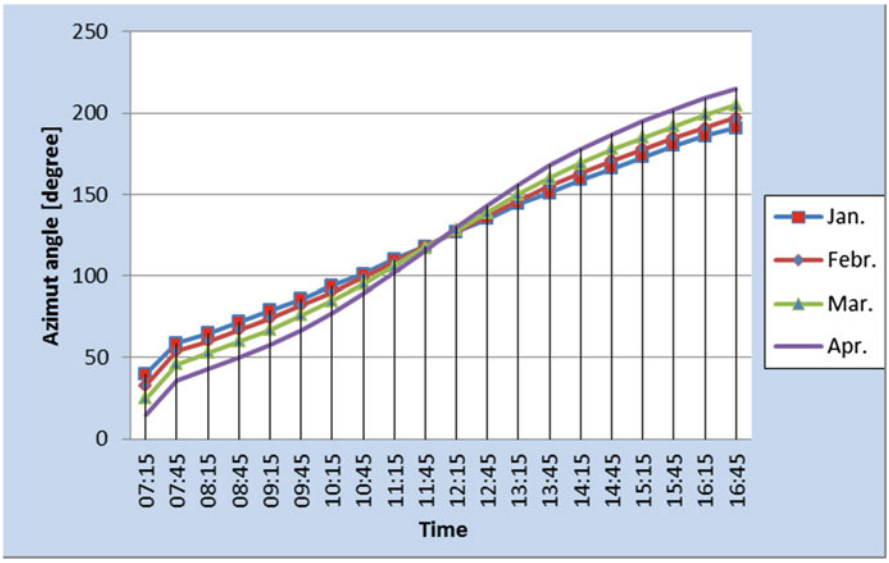


Fig. 24.9 Azimuth angle of photovoltaic panel

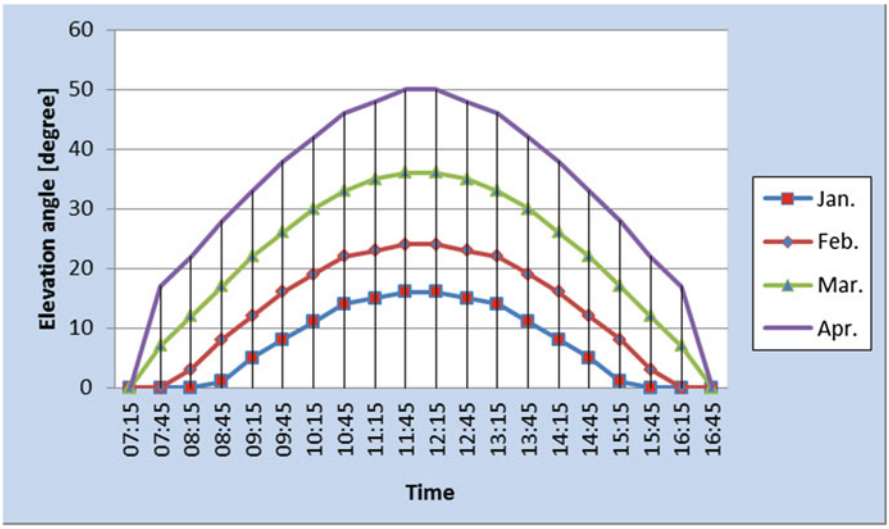


Fig. 24.10 Elevation angle of photovoltaic panel

The relations (24.4) and (24.7) describe the position of the PV panel at any time of the year so that sunlight falling perpendicular to it. For location Craiova was resulted both graphs presented in Figs. 24.9 and 24.10 that describe the two angles corresponding to the PV panel for first four months of the year.

24.4 Modelling of the Concentrating System for the Photovoltaic Panel (LCPV)

If we ignore tracking, the analysed low concentration photovoltaic system (LCPV) contains: a photovoltaic module and two mirrors arranged symmetrically on large sides of photovoltaic module [2, 8].

Proper functioning of this low concentration photovoltaic system depends by, in particular, the following two requirements [2, 9]:

(a) PV module must be completely crossed by direct radiation (excluded initially the effect of shading and diffuse radiation is neglect);
(b) PV module must be completely swept for the direct radiation reflected by mirrors.

To identify the geometric parameters that define the system LCPV in Fig. 24.11 is represented the geometric diagram describing sweeping photovoltaic module by direct and the reflected radiation during a step of tracking: in Fig. 24.11 the red rays represent the start of the step, the rays mid of step is black and the blue is end-of-step rays.

To describe the geometry of the LCPV system considered, were introduced the following parameters [2, 10–12](Fig. 24.11):

– *x*—the angle formed between the mirror and photovoltaic module—constant parameter;
– *h*—the maximum incidence angle formed by the PV normal surface and sun's ray;

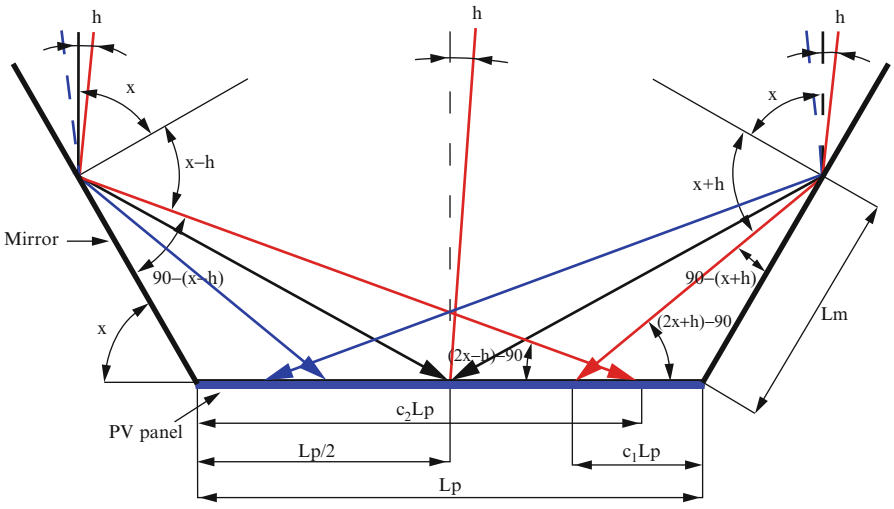


Fig. 24.11 Scheme for geometric modeling of LCVP system

- k—the ratio of the (Lm) width mirrors and (Lp) PV module width;
- c_1, c_2—PV module with coefficients swept rays reflected by the mirrors;
- k_l—the longitudinal deflection coefficient, defined as ratio of the additional length mirror (necessary to compensate for reflected sunlight deviation caused by deviations elevation to elevation solar PV module) and PV module width.

Sweeping the entire surface of the PV module is determined by the median rays (black color line Fig. 24.11), which designates the middle of a tracking step; as a result, the ratio k can be modeled analytically by applying sine theorem in the triangles formed by the Lp and Lm sides with the median rays.

$$k = \frac{L_m}{L_p} = \frac{-\cos(2 \cdot x)}{\cos(x)} \tag{24.8}$$

By means of the analytical expression obtained in Fig. 24.12 is shown k reports variation according to the x angle for different values of the h angle of incidence.

In a graphics evolution it can seen as a gauge of system increases with increasing angle x.

Partially reflected beam sweeps photovoltaic module width. To determine the ratio of the width swept is introduced c coefficient, described as the ratio between the share swept width (cLm) and PV module width (Lp).

Considering the known Lp width of the PV module, Lm width of the two mirrors and x tilt angle mirror-photovoltaic module (Fig. 24.11) can determine the coefficients c_1 and c_2, which describe what part of PV module width (Lp) is swept by rays reflected from each mirror.

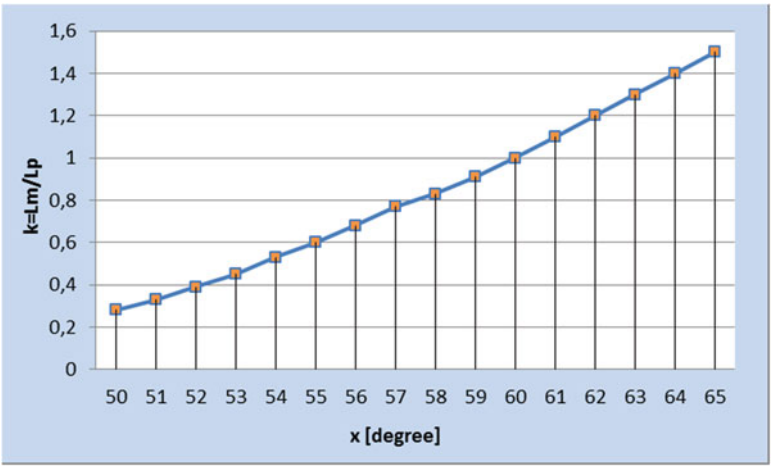


Fig. 24.12 The evolution of ratio k for different values of the angle of incidence h

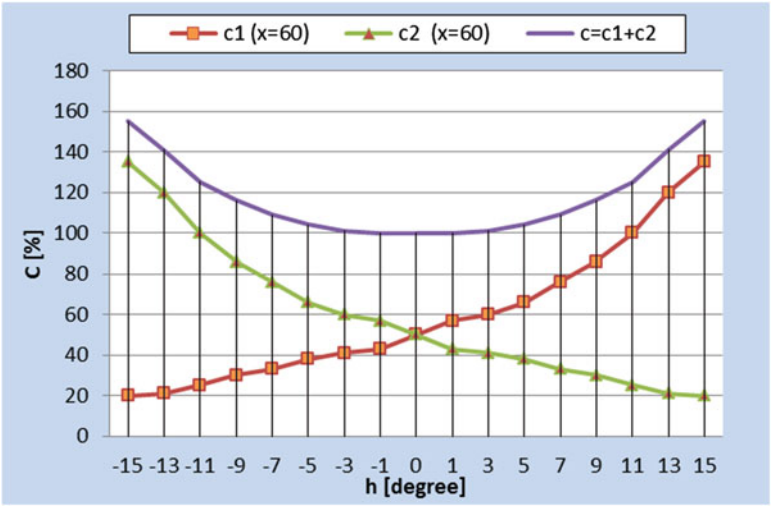


Fig. 24.13 Evolution of the coefficient c for different values of h at ($x = 60°$)

$$c_1 = \frac{L_m \cdot \cos(x + h)}{-L_p \cdot \cos(2 \cdot x + h)} \tag{24.9}$$

$$c_2 = \frac{L_m \cdot \cos(x - h)}{-L_p \cdot \cos(2 \cdot x - h)} \tag{24.10}$$

Evolution of coefficients c_1, c_2, ie c for different values of the h angle of incidence of sunlight for the best case of the angle of the mirror ($x = 60°$) is shown in Fig. 24.13.

Complete sweep PV surface condition is satisfied only if $c_1 + c_2 \geq 100$ %; in Fig. 24.13, $c_1 + c_2$ curve show that for the $\theta = 60°$ is equal to 100 % in the case of continuous orientation (angle of incidence corresponding to zero) and increases with the value of the maximum angle of incidence (i.e., the length tracking step in the case to steps tracking). By multiplying the coefficients c_1 and c_2 with the width and length of photovoltaic module is obtained surface covered by reflected radiation to PV panel.

24.5 Conclusions

In this work are presented some geometric aspects to be considered when designing the tracking systems for PV panels with low concentration of radiation.

Developed mathematical models have been custom for a specific location in order to study the efficiency of PV panels tracking with low concentration of solar radiation in specific conditions of a particular geographical area.

Implementation models on physical prototypes, interpretation of the results on the effectiveness of these systems will be subject to future work.

Acknowledgment This work was partially supported by the grant number 29C/2014, awarded in the internal grant competition of the University of Craiova.

References

1. Messenger, R., Ventre, J.: Photovoltaic System Engineering, 2nd edn. CRC Press, Boca Raton (2004)
2. Hermenean, I., Vişa, I., Diaconescu, D.: On the geometric modelling of a concentrating PV-mirror system. Bull Transilvania Univ Braşov Ser I Eng Sci **2**(51), 73–80 (2009)
3. Luque, A., Andreev, V.M.: Concentrator Photovoltaics. Springer, Berlin (2007)
4. Luque, A., Hegedus, S.: Handbook of Photovoltaic Science and Engineering. Wiley-VCH, New York (2003)
5. Vişa, I., Diaconescu, D., Popa, V., Burduhos, B.: On the incidence angle optimization of the dual-axis solar trackers,. In: 11th International Research/Expert Conference TMT—Trends in the Development of Machinery and Associated Technology, Hamammet, Tunisia, 04–09 Septembrie 2007, pp. 1111–1113. ISBN 994861733-X
6. Comşiţ, M.: Specific orientation mechanisms of solar energy conversion systems. PhD thesis, Transilvania University of Brasov (2007)
7. Alboteanu, L., Manolea, G.H., Ravigan, F.: Positioning systems for solar panels placed in isolated areas. Ann. Univ. Craiova Electr. Eng. Ser. **30**, 163–168 (2006). ISSN 1832-3804
8. Benecke, M.A., Van Dyk, E.E., Vorster, F.J.: Investigation of the design aspects on the performance of a LCPV system. Nelson Mandela Metropolitan University, Centre for Renewable and Sustainable Energy Studies
9. Benecke, M.A., Van Dyk, E.E., Vorster, F.J.: The design and analysis of a vertical receiver LCPV system. Nelson Mandela Metropolitan University, Centre for Renewable and Sustainable Energy Studies
10. Bett, A.W., Lerchenmüller, H.: The FLATCON system from concentrix solar. In: Concentrator Photovoltaics. Springer Series in Optical Science. Springer, Berlin (2007).
11. Grasso, G., Morichetti, F., Righetti, A., Pietralunga, S.M.: Trackless LCPV modules: a competitive solution. In: 38th IEEE Photovoltaic Specialists Conference (PVSC), pp. 2048–2051 (2012). ISSN: 0160-8371
12. Palmer, T, et al.: Tracking Systems for CPV: Challenges and Opportunities. In: Solar Power International (2007)

Chapter 25
Systolic Approach for QR Decomposition

Halil Snopce and Azir Aliu

Abstract In this paper we discuss the parallelization of the QR decomposition of matrices based on Given's rotation using the iterative algorithm. For this purpose we have used the systolic approach. The mathematical background of the problem is followed by the parallelization which continues step by step as it is shown at Figs. 25.5 and 25.6. The output values of Fig. 25.5 become the input for Fig. 25.6 and vice versa, the output values of Fig. 25.6 become the input for Fig. 25.5. This kind of iteration is repeated until achieving the convergence.

Keywords QR decomposition • Parallelization of QR decomposition • Systolic array • Given's rotations • Computing the orthonormal matrix • Computing the upper triangular matrix

25.1 Introduction

An important matrix problem that arises in many applications, like signal processing, image processing, solution of differential equations, etc., is the problem of solving a set of simultaneous linear equations. The usual numerical method for solving such problems is the triangularization of the coefficient matrix, followed by the use of back substitution. QR decomposition is one of the best methods for matrix triangularization. Most of the QR-decomposition implementations are based on three methods. The Given's rotation method, the Gram-Schmidt method and the method with the Hausholder transformations. The Hausholder transformation is one of the most computationally efficient methods to compute the QR-decomposition of a matrix. The error analysis carried out by Wilkinson [1, 2], showed that the Hausholder transformation outperforms the Given's method under finite precision computation. Due to the vector processing nature of the Hausholder transformation, no local connections in the implementation of the array are necessary. Therefore QR decomposition by the method of Hausholder transformation is more difficult.

H. Snopce (✉) • A. Aliu
SEE-University, CST Faculty, Ilindenska 335, Tetovo 1200, Macedonia (FYROM)
e-mail: h.snopce@seeu.edu.mk

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_25

Especially, the systolic approach is difficult because we have to find only local connections. This is the reason why the systolic approach shown in this paper is based on Given's rotations.

The QR decomposition is a form of orthogonal triangularization which is particularly useful in least squares computations and simultaneous equations. It is one of the most stable numerical algorithms.

The basic idea of the QR-decomposition of a matrix is to express a given $m \times n$ matrix A in the form $A = QR$, where Q is an orthonormal $m \times n$ matrix and R is an $n \times n$ upper triangular matrix with nonzero diagonal entries.

The QR decomposition is a technique which is used in the architectures which require parallel computations using a triangular array of relatively simple processing elements. For example, it is applied at the adaptive filtering and beam-forming problems, where there is no special structure of the data matrix.

A parallel version of Given's rotation was proposed in [3]. In [4] it is proposed an alternative way for parallelization of Given's rotation which is more efficient for larger matrices. In [5] it is given a parallel pipeline version of Given's rotation for the QR decomposition. In [6] one can find the block version of the QR decomposition, which first transforms the matrix into the Hassenberg form and then applies Given's rotation to it.

The analysis in this paper uses the givens rotation method [7]. In [7] and [8] are proposed two systolic arrays for the QR decomposition with hardware complexity $O(n^2)$ and time complexity O(n) which are based on the method of Given's rotation. The design which is based on the method of Householder transformation is given in [9].

In this paper we give first the mathematical background of the QR decomposition method based on Given's rotation, and then we analyze the corresponding systolic array for processing with this method.

25.2 The Systolic Array Based on Given's Rotation

The upper triangular matrix is obtained using sequences of Given's rotations [10] such that the subdiagonal elements of the first column are nullified first, followed by those of the second column and so forth, until an upper triangular matrix is reached. The procedure can be written in the form given below:

$$Q^T A = R$$

where

$$Q^T = Q_{n-1} Q_{n-2} \dots Q_1 \tag{25.1}$$

and

$$Q_p = Q^{p,p} Q^{p+1,p} \dots Q^{n-1,p}$$

where $Q^{p,q}$ is the Given's rotation operator used to annihilate the matrix element located at row $q + 1$ and column p. When we work with 2×2 matrices, an elementary Given's transformation has the form:

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \cdot \begin{bmatrix} 0 \dots 0 & r_i & r \dots r_k \\ 0 \dots 0 & x_i & x_{i+1} \dots x_k \end{bmatrix} = \begin{bmatrix} 0 \dots 0 & r_i' & r_{i+1}' \dots r_k' \\ 0 \dots 0 & 0 & x_{i+1}' \dots x_k' \end{bmatrix} \tag{25.2}$$

where c and s are the cosine and the sine of the annihilation angle, such that:

$$c = \frac{r_i}{\sqrt{r_i^2 + x_i^2}}, \quad s = \frac{x_i}{\sqrt{r_i^2 + x_i^2}}.$$

In [11] it is shown that a triangular systolic array can be used to obtain the upper triangular matrix R based on sequences of Given's rotations. This systolic array is shown in Fig. 25.1.

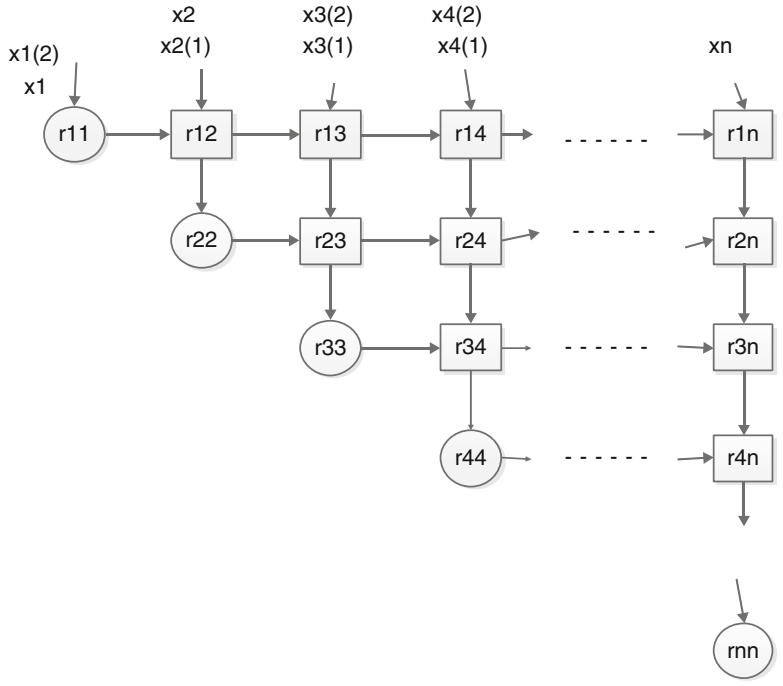


Fig. 25.1 Triangular systolic array for computing the upper triangular matrix R

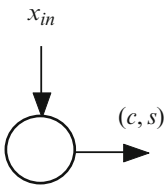


Fig. 25.2 Input and output of the circle cell of the array in Fig. 25.1

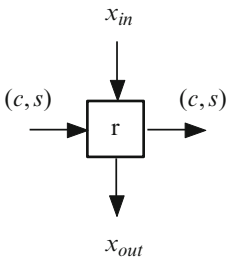


Fig. 25.3 Input and output of the quadratic cell of the array in Fig. 25.1

As we can see, the array consists of two different shapes of cells. The cells in the shape of a circle (Fig. 25.2), and the cells in quadratic shape (as in Fig. 25.3).

The cells of Fig. 25.2 perform according to Algorithm 1:

Algorithm 1

$$
\begin{aligned}
&If \;\; x_{in} = 0 \;\; then\\
&\quad c = 1; \;\; s = 0\\
&otherwise\\
&\quad r' = \sqrt{r^2 + x_{in}^2};\\
&\quad c = r/r'; \;\; s = x_{in}/r'\\
&\quad r = r';\\
&end
\end{aligned}
$$

The calculations of quadratic cells are given by the relations below:

$$
\begin{aligned}
x_{out} &= cx_{in} - sr\\
r &= sx_{in} + cr
\end{aligned}
$$

According to the relations in (25.1), the Q matrix cannot be obtained by multiplying cumulatively the rotation parameters propagated to the right. Accumulation of the rotation parameters is possible by using an additional rectangular systolic array.

Before giving an explanation about the systolic array for thr QR decomposition, we will introduce the methodology of computing $R^{-T}x$. This computation will be used in the general design of the systolic array for a QR decomposition.

25.2.1 The Computation of $R^{-T}x$

We present a brief derivation of the result presented in [12], about the property that a triangular array can compute $R^{-T}x$ in one phase with the matrix R situated in that array.

Let $r_{ij} = [R]_{ij}$ and $r'_{ij} = \left[R^{-1}\right]$, where $r_{ij} = 0$ and $r'_{ij} = 0$ for $i > j$. It can be shown that:

$$r'_{ij} = \begin{cases} \frac{1}{r_{ii}}; & i = j \\ -\sum_{k=i}^{j-1} \frac{r'_{ik} r_{kj}}{r_{jj}}; & i < j \leq n \end{cases} \tag{25.3}$$

Let

$$[y_1, \ldots, y_n]^T = R^{-T} X \tag{25.4}$$

Then the recursive computation of (25.4), where R^{-T} is a n × n matrix and X is an n × m matrix is:

$$y_j = \sum_{i=1}^{j} x_i r'_{ij}, \quad i = 1, \ldots, n \tag{25.5}$$

In particular (because we want to use R and X to compute $R^{-T}X$), y_j can be expressed in terms of r_{ij} and x_i. By substituting Eq. (25.4) into Eq. (25.5) we have:

$$y_j = \sum_{i=1}^{j} x_i r'_{ij} = y_j = \sum_{i=1}^{j-1} x_i r'_{ij} + x_j r'_{jj} = \sum_{i=1}^{j-1} x_i r'_{ij} + \frac{x_j}{r_{jj}} \tag{25.6}$$

If we continue, by transforming the relation (25.6), we will have:

$$y_j = \frac{x_j}{r_{jj}} + \sum_{i=1}^{j-1} x_i r'_{ij} = \frac{x_j}{r_{jj}} - \sum_{i=1}^{j-1} x_i \sum_{k=i}^{j-1} \frac{r'_{ik} r_{kj}}{r_{jj}}$$

And finally we get:

$$y_j = \frac{1}{r_{jj}} \cdot \left(x_j - \sum_{i=1}^{j-1} x_i \sum_{k=i}^{j-1} r'_{ik} r_{kj} \right) = \frac{1}{r_{jj}} \cdot \left(x_j - \sum_{k=1}^{j-1} \sum_{i=1}^{k} x_i r'_{ik} r_{kj} \right) \tag{25.7}$$

Using the relation (25.5), for the final form of y_j, we get

$$y_j = \frac{1}{r_{jj}} \left(x_j - \sum_{k=1}^{j-1} y_k r_{kj} \right) \tag{25.8}$$

Finally, using the relations obtained above (where Y is the n × m matrix, R is n × n upper triangular matrix and X is an n × m matrix), the algorithm for computing $R^{-T}x$ is given:

Algorithm 2

$$\begin{array}{l} for\ \ i = 1\ \ to\ \ n \\ y_1 = \frac{1}{r_{11}} \cdot x_1 \\ for\ \ j = 2\ \ to\ \ n \\ \quad begin \\ \quad z_j = x_j \\ \quad for\ \ k = 1\ \ to\ \ j-1 \\ \quad\quad z_j = z_j - y_k r_{kj} \\ \quad y_j = \frac{z_j}{r_{jj}} \\ end \end{array}$$

The corresponding systolic array is similar as the array in Fig. 25.1. The data movement of input values x and output values y is presented in Fig. 25.4.

In the case presented above, the elements of the matrix R are stored in the triangular array. The cells of Fig. 25.2 (circle cells) perform the division part of Eq. (25.8) (the part $1/r_{jj}$). The second part of Eq. (25.8) (the part $x_j - \sum_{k=1}^{j-1} y_k r_{kj}$) is performed by the quadratic cells shown in Fig. 25.3.

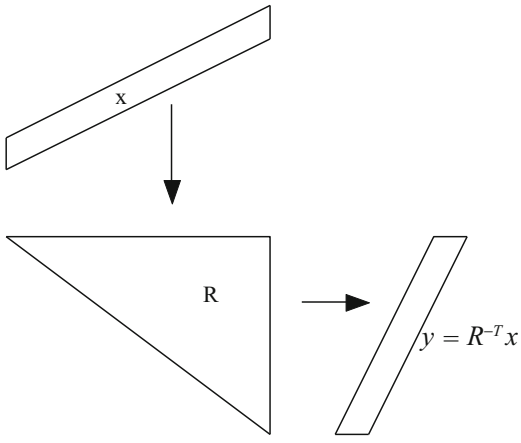


Fig. 25.4 Data movement of x and y in the computation of $R^{-T}x$

25.3 The QR Systolic Array

The design of the systolic array for a QR-decomposition of a matrix A will be based on an iterative algorithm which consists of two basic steps. Initially we set $A_1 = A$. The first step is to compute $A_k = Q_k R_k$. The process has to be continued until the convergence. To compute the next iteration A_{k+1} we start from the relation (25.1) and taking into the consideration that Q is orthonormal $Q^T Q = I$, we have:

$$Q_k^T A_k = R_k \Rightarrow A_k = Q_k R_k \Rightarrow A_k R_k^{-1} = Q_k \tag{25.9}$$

$$A_{k+1} = R_k Q_k = Q_k^T A_k Q_k = Q_k^T Q_k R_k Q_k = R_k Q_k \tag{25.10}$$

So, this can be expressed as follows:

Algorithm 3

Set $A = A_1$

Step 1: For $k = 1,\ 2, \ldots,$ compute $A_k = Q_k R_k$.

Step 2: Compute $A_{k+1} = R_k Q_k$. If A_{k+1} converges, then stop. Otherwise go back to step 1.

From $A_k = Q_k R_k$ we have that

$$A_k{}^T = R_k{}^T Q_k{}^T \Rightarrow R_k^{-T} A_k{}^T = Q_k{}^T$$

If the ith column of the matrices $A_k{}^T$ and $Q_k{}^T$ is denoted by a_i and q_i respectively, then

$$R_k^{-T} [a_1\ \ a_2 \ldots a_n] = [q_1,\ \ q_2, \ldots q_n] \tag{25.11}$$

We already have shown how to compute $R^{-T} x$. So, the systolic array is similar to that one shown in Fig. 25.4. Since the ith column of $A_k{}^T$ is the same with the ith row of A_k, the elements of the matrix A_k will be inputted row by row. The corresponding systolic array for computing the elements of Q_k as output elements is given in Fig. 25.5.

Of course, the triangular array which contains the elements of the matrix R (presented as right angle triangle) is the same with the array in Fig. 25.1.

Figure 25.5 in fact is the design for systolic computing of step 1 of Algorithm 2. To do the second step, which consists in computing $A_{k+1} = R_k Q_k$, the output elements of Fig. 25.5 become row by row the input elements for the new computation. It is illustratively shown in Fig. 25.5. So, the new array, which computes the element of the matrix A_{k+1} using as an input elements the computed ones illustrated in Fig. 25.5, is shown in Fig. 25.6. The elements of A_{k+1} come out column by column.

If the obtained result is not convergent, then a new iteration will be repeated until achieving a convergence.

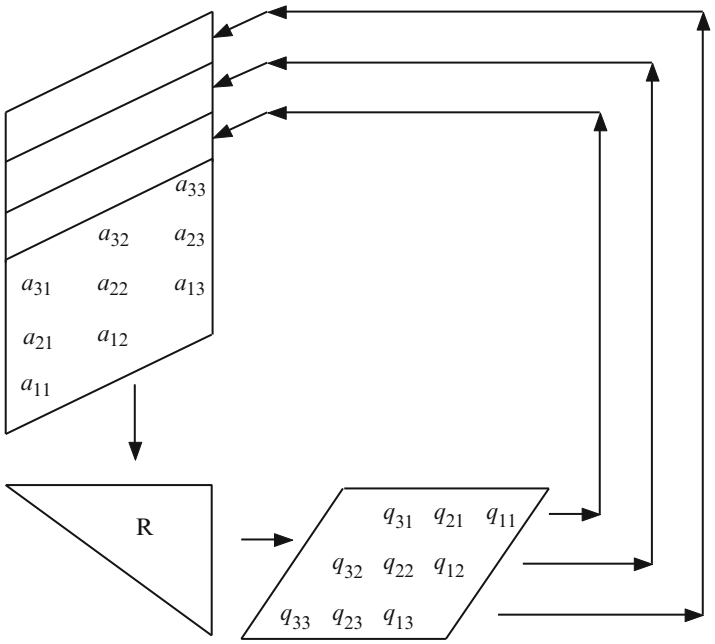


Fig. 25.5 Systolic model for computing the Q matrix

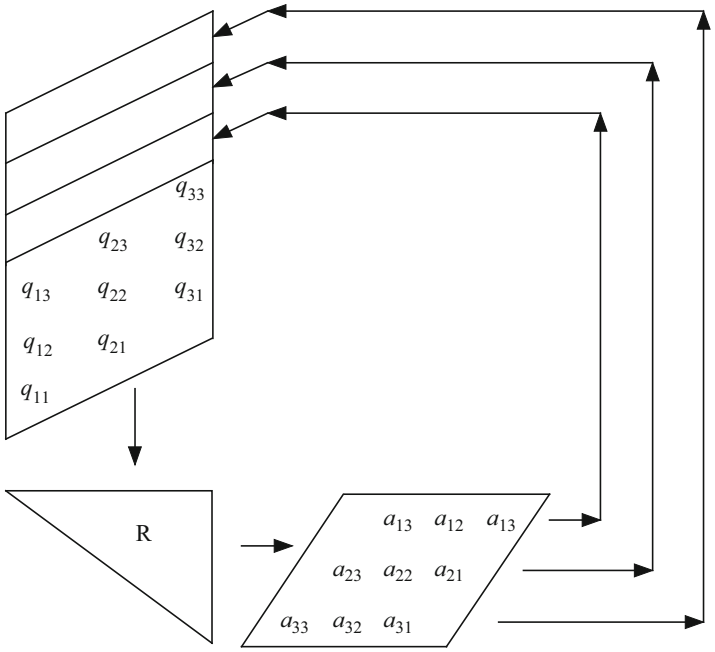


Fig. 25.6 Systolic computing of the product RQ

References

1. Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Clarendon Press, Oxford (1965)
2. Johnsson, L.: A computational array for the QR method. In: Proceedings of 1982 Conference on Advanced Research in VLSI, pp. 123–129. MIT, Cambridge, MA
3. Sameh, A.H., Kuck, D.J.: On stable parallel linear system solvers. J. ACM **25**, 81–95 (1978)
4. Modi, J., Clarke, M.: An alternative givens ordering. Numer. Math. **43**(1), 83–90 (1984)
5. Hofmann, M., Kontoghiorghes, E.: Pipeline Givens sequences for computing the QR decomposition on a EREW PRAM. Parallel Comput **32**(3), 222–230 (2006)
6. Berry, M., Dongara, J., Kim, Y.: A parallel algorithm for the reduction of a nonsymmetric matrix to block upper-Hessenberg form. Parallel Comput **21**(8), 1189–1211 (1995)
7. Kung, S.Y.: VLSI array processors. Englewood Cliffs, Prentice Hall (1988)
8. Swartzlander EE. (ed.) Systolic Signal Processing Systems. Marcel Decker, New York (1987)
9. Kahaner, D., Moler, C., Nash, S.: Numerical Methods and Software. Prentice Hall, Englewood Cliffs (1989)
10. Jacobi, C.G.J.: Uber eine neue Auflosungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen. Astronomishe Nachricten, 22, 1845. English translation by G.W. Stewart, Technical Report 2877, Department of Computer Science, University of Maryland (1992)
11. Gentleman, W.M., Kung, H.T.: Matrix triangularization by systolic array. Proc. SPIE Int. Soc. Opt. Eng. **298**, 298 (1981)
12. McWhirter, J.G., Shepherd, T.J.: An efficient systolic array for MVDR beamforming. In: Proceedings of International Conference on Systolic Arrays, pp. 11–20 (1988)

Chapter 26
A Geometric Approach for the Model Parameter Estimation in a Permanent Magnet Synchronous Motor

Paolo Mercorelli

Abstract Control of permanent magnetic motors is not an easy task because of the presence of unknown parameters. Techniques are needed in order to achieve a suitable controlled dynamics identification. The proposed strategy uses the geometric approach to realised a decoupling of the system. The estimation of the parameters of a Permanent Magnet Synchronous Motor (PMSM) is simplified through a decoupling. The decoupling is realised using a feedback controller combined with a feedforward one. The feedforward controller is conceived through an input partition matrix. This technique can be applied to a large variety of motors or to any system for which the decoupling conditions are satisfied. Simulation and measured results are reported to validate the proposed strategy.

Keywords Geometric approach • Permanent magnet synchronous motor • Identification

26.1 Introduction and Motivations

Recently the interest in the topic of geometric control has increased in theoretical aspects and applications as well, see for instance [1], particulary in control problems like Non-interaction and Model Predictive Control, see [2]. It is known that, an accurate knowledge of the model and its parameters is necessary for realising an effective control. For achieving a desired system performance, advanced control systems are usually required to provide fast and accurate response, quick disturbance recovery and parameter variations insensitivity [3]. Acquiring accurate models for systems under investigation is usually the fundamental part in advanced control system designs, see [4]. The most common parameters required for the implementation of such advanced control algorithms are the classical simplified model parameters:

P. Mercorelli (✉)
Institute of Product and Process Innovation, Leuphana University of Lueneburg, Volgershall 1, 21339 Lueneburg, Germany,
e-mail: mercorelli@uni.leuphana.de

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_26

L_d—the direct axis self-inductance, L_q—the quadrature axis self-inductance, and Φ—the permanent magnet flux linkage. Techniques have been proposed for the parameters' identification of a Permanent Magnet Synchronous Motor (PMSM) from different perspectives, such as offline [5, 6] and online identification of PMSM electrical parameters, [7]. These techniques are based on the decoupled control of linear systems when the motor's mechanical dynamics are ignored. Using a decoupling control strategy, internal dynamics may be almost obscured, but it is useful to remember that there are no limitations in the controllability and observability of the system. In the report by [8] a decoupling technique is used to control a permanent magnets machine more efficiently in a sensorless way using an observer. Despite limitations on the frequency range of identification, this paper proposes a dynamic observer based on a geometric decoupling technique to estimate parameter Φ. The proposed identification technique, similar to that presented in [9], applies a procedure based on the work in [10]. In the meantime, the paper proposes a particular observer that identifies the permanent magnet flux using the estimated L_{dq} and R_s parameters from an ARMA identification structure as presented in [10]. The paper is organised in the following way: a sketch of the model of the synchronous motor and its behaviour are given in Sects. 26.2 and 26.3 is devoted to deriving, proposing and discussing the dynamic estimator, and Sect. 26.4 shows the simulation results using real data for a three-phase PMSM.

The main nomenclature

$\mathbf{u}_{in}(t) = [u_a(t), u_b(t), u_0(t)]^T$: three phase input voltage vector
$\mathbf{i}(t) = [i_a(t), i_b(t), i_0(t)]^T$: three phase input current vector
$\mathbf{u}_q(t)$: induced voltage vector
ω_{el}: electrical pulsation
R_s: coil resistance
L_{dq}: dq coil inductance
$\mathbf{A}$: state matrix of the electrical model
$\mathbf{B}$: input matrix of the electrical model
$\mathscr{B} = \mathsf{im}\mathbf{B}$: image of matrix $\mathbf{B}$ (subspace spanned by the columns of matrix $\mathbf{B}$)
$\mathsf{min}\mathscr{I}(\mathbf{A}, \mathscr{B}) = \sum_{i=0}^{n-1} \mathbf{A}^i \mathsf{im}\mathbf{B}$: minimum $\mathbf{A}$–invariant subspace containing $\mathsf{im}(\mathbf{B})$
$\mathbf{F}$: decoupling feedback matrix field
$\mathbf{g}(\omega_{el})$: Park transformation
$\mathbf{T}(\omega_{el})$: decoupling feedforward matrix field
$\mathscr{I}$: invariant subspace
$\mathscr{C}_d$: kernel of output matrix $\mathbf{C}_d$ (d component of the current)
$\mathscr{C}_q$: kernel of output matrix $\mathbf{C}_q$ (q component of the current)
$\mathscr{C}_0$: kernel of output matrix $\mathbf{C}_0$ (0 component of the current)

26.2 Model of a Synchronous Motor

For aiding advanced controller design for PMSM, it is very important to obtain an appropriate model of the motor. A good model should not only be an accurate representation of system dynamics but it should also facilitate the application of the existing control techniques. Among a variety of models presented in the literature since the introduction of PMSM, the two-axis dq-model, obtained using Park dq-transformation is the most widely used in variable speed PMSM drive control applications [3, 7]. The Park dq-transformation is a coordinate transformation that converts the three-phase stationary variables into variables in a rotating coordinate system. In dq-transformation, the rotating coordinate is defined relative to a stationary reference angle as illustrated in Fig. 26.1. The dq-model is considered in this work.

$$\begin{bmatrix} u_d(t) \\ u_q(t) \\ u_0(t) \end{bmatrix} = \begin{bmatrix} \frac{2\sin(\omega_{el}t)}{3} & \frac{2\sin(\omega_{el}-2\pi/3)}{3} & \frac{2\sin(\omega_{el}+2\pi/3)}{3} \\ \frac{2\cos(\omega_{el}t)}{3} & \frac{2\cos(\omega_{el}-2\pi/3)}{3} & \frac{2\cos(\omega_{el}+2\pi/3)}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} u_a(t) \\ u_b(t) \\ u_c(t) \end{bmatrix}, \tag{26.1}$$

$$\begin{bmatrix} i_d(t) \\ i_q(t) \\ i_0(t) \end{bmatrix} = \begin{bmatrix} \frac{2\cos(\omega_{el}t)}{3} & \frac{2\cos(\omega_{el}-2\pi/3)}{3} & \frac{2\cos(\omega_{el}+2\pi/3)}{3} \\ \frac{-2\sin(\omega_{el}t)}{3} & \frac{-2\sin(\omega_{el}-2\pi/3)}{3} & \frac{-2\sin(\omega_{el}+2\pi/3)}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} i_a(t) \\ i_b(t) \\ i_c(t) \end{bmatrix}. \tag{26.2}$$

The dynamic model of the synchronous motor in dq-coordinates can be represented as follows:

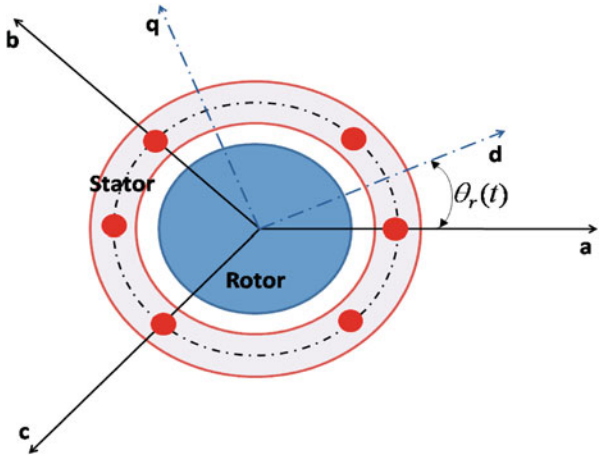


Fig. 26.1 Park transformation for the motor

$$\begin{bmatrix} \frac{di_d(t)}{dt} \\ \frac{di_q(t)}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{R_s}{L_d} & \frac{L_q}{L_d}\omega_{el}(t) \\ -\frac{R_s}{L_q} & -\frac{L_d}{L_q}\omega_{el}(t) \end{bmatrix} \begin{bmatrix} i_d(t) \\ i_q(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{L_d} & 0 \\ 0 & \frac{1}{L_q} \end{bmatrix} \begin{bmatrix} u_d(t) \\ u_q(t) \end{bmatrix} - \begin{bmatrix} 0 \\ \frac{\Phi\omega_{el}(t)}{L_q} \end{bmatrix}, \tag{26.3}$$

and

$$M_m = \frac{3}{2} p \Big\{ \Phi i_q(t) + (L_d - L_q) i_d(t) i_q(t) \Big\}. \tag{26.4}$$

In (26.3) and (26.4), $i_d(t)$, $i_q(t)$, $u_d(t)$ and $u_q(t)$ are the dq-components of the stator currents and voltages in synchronously rotating rotor reference frame, $\omega_{el}(t)$ is the rotor electrical angular speed, the parameters R_s, L_d, L_q, Φ and p are the stator resistance, d-axis and q-axis inductance, the amplitude of the permanent magnet flux linkage, and p the number of couples of permanent magnets, respectively. At the end, M_m indicates the motor torque. Considering an isotropic motor with $L_d \simeq L_q = L_{dq}$, it follows:

$$\begin{bmatrix} \frac{di_d(t)}{dt} \\ \frac{di_q(t)}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{R_s}{L_{dq}} & \omega_{el}(t) \\ -\frac{R_s}{L_{dq}} & \omega_{el}(t) \end{bmatrix} \begin{bmatrix} i_d(t) \\ i_q(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{L_{dq}} & 0 \\ 0 & \frac{1}{L_{dq}} \end{bmatrix} \begin{bmatrix} u_d(t) \\ u_q(t) \end{bmatrix} - \begin{bmatrix} 0 \\ \frac{\Phi\omega_{el}(t)}{L_{dq}} \end{bmatrix}, \tag{26.5}$$

and

$$M_m = \frac{3}{2} p \Phi i_q(t), \tag{26.6}$$

with the following movement equation:

$$M_m - M_w = J \frac{d\omega_{mec}(t)}{dt}, \tag{26.7}$$

where $p\omega_{mech}(t) = \omega_{el}(t)$ and M_w is an unknown mechanical load.

26.3 Design of a Decoupling Control Strategy

The present estimator uses the measurements of input voltages, currents and angular velocity of the motor to estimate the "dq" winding inductance, the rotor resistance and amplitude of the linkage flux. The structure of the estimator is described in Fig. 26.2. This diagram shows how the estimator works. In particular, after having decoupled the system described in (26.5), the stator resistance R_s and the inductance L_{dq} are estimated through a minimum error variance approach. The estimated values $\hat{R}_s$ and $\hat{L}_{dq}$ are used to estimate the amplitude of the linkage flux ($\hat{\Phi}$). The earliest geometric approaches to decoupling control were due to [11] and to [12, 13]. The following definition taken from [11] recalls the concept of decoupling.

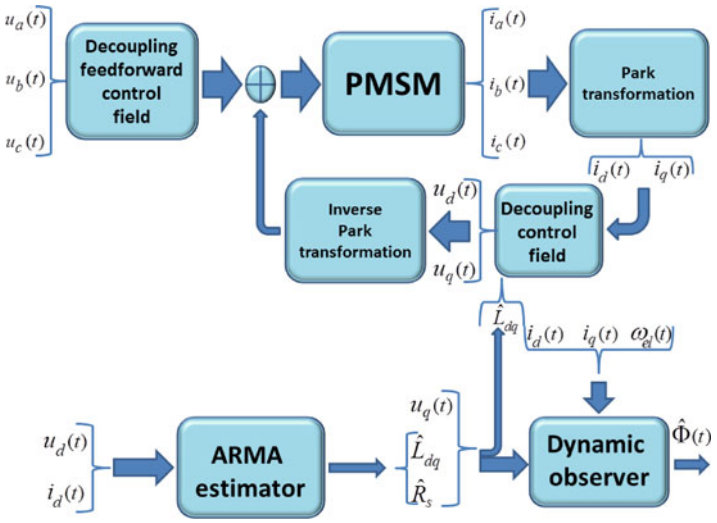


Fig. 26.2 Conceptual structure of the whole estimator

Definition 1. A control law for the dynamic system described by (26.1)–(26.3) is *decoupling* with respect to the regulated outputs $i_d(t)$, $i_q(t)$, and $i_o(t)$, if there exist a feedback matrix field $\mathbf{F}(\omega_{el})$ and input partition matrix field $\mathbf{T}(\omega_{el})) = [\mathbf{T}_d, \mathbf{T}_q, \mathbf{T}_0]^T$ of the input voltage vector such that for zero initial conditions, each input $u_{(\cdot)}(t)$ (with all other inputs, identically zero) only affects the corresponding output $i_d(t)$, $i_q(t)$, or $i_o(t)$. □

For achieving a decoupled structure of the system described in Eq. (26.5), a matrix field $\mathbf{F}(\omega_{el})$ is to be calculated such that:

$$\big(\mathbf{A} + \mathbf{BF}(\omega_{el})\big)\mathscr{V} \subseteq \mathscr{V}, \tag{26.8}$$

where $\mathbf{u}(t) = \mathbf{F}(\omega_{el})\mathbf{x}(t)$ is a state feedback with $\mathbf{u}(t) = [u_d(t), u_q(t)]^T$ and $\mathbf{x}(t) = [i_d(t), i_q(t)]^T$,

$$\mathbf{A} = \begin{bmatrix} -\frac{R_s}{L_{dq}} & \omega_{el}(t) \\ -\frac{R_s}{L_{dq}} & \omega_{el}(t) \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \frac{1}{L_{dq}} & 0 \\ 0 & \frac{1}{L_{dq}} \end{bmatrix}, \tag{26.9}$$

and $\mathscr{V} = im([0,1]^T)$ of Eq. (26.8), according to [11], is a controlled invariant subspace. More explicitly it follows:

$$\mathbf{F}(\omega_{el}) = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}, \text{and} \begin{bmatrix} u_d(t) \\ u_q(t) \end{bmatrix} = \mathbf{F}(\omega_{el}) \begin{bmatrix} i_d(t) \\ i_q(t) \end{bmatrix},$$

then the decoupling of the dynamics is obtained via the following relationship:

$$\mathsf{im}\left(\begin{bmatrix} -\frac{R_s}{L_{dq}}\,\omega_{el}(t) \\ -\frac{R_s}{L_{dq}}\,\omega_{el}(t) \end{bmatrix}\right) + \mathsf{im}\left(\begin{bmatrix} \frac{1}{L_{dq}} & 0 \\ 0 & \frac{1}{L_{dq}} \end{bmatrix}\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \subseteq \mathsf{im}\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad (26.10)$$

where parameters F_{11}, F_{12}, F_{21}, and F_{22} are to be calculated in order to guarantee condition (26.10) and a suitable dynamics for sake of estimation. Condition (26.10) is guaranteed if:

$$F_{12} = -\omega_{el}(t) L_{dq}. \quad (26.11)$$

$$\frac{di_d(t)}{dt} = -\frac{R_s}{L_{dq}} i_d(t) + \frac{u_d(t)}{L_{dq}}, \quad (26.12)$$

Considering now the following output matrix:

$$\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{C}_d \\ \mathbf{C}_q \end{bmatrix}. \quad (26.13)$$

It is to be shown that, if:

$$\mathbf{g}(\omega_{el}) = \begin{bmatrix} \frac{2\sin(\omega_{el}t)}{3} & \frac{2\sin(\omega_{el}-2\pi/3)}{3} & \frac{2\sin(\omega_{el}+2\pi/3)}{3} \\ \frac{2\cos(\omega_{el}t)}{3} & \frac{2\cos(\omega_{el}-2\pi/3)}{3} & \frac{2\cos(\omega_{el}+2\pi/3)}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, \quad (26.14)$$

then there exists a decoupling and stabilizing state feedback matrix field $\mathbf{F}(\omega_{el})$, along with two input partition matrix fields $\mathbf{T}_d(\omega_{el})$, $\mathbf{T}_q(\omega_{el})$, and $\mathbf{T}_c(\omega_{el})$ such that, for the dynamic triples

$$\begin{aligned} &\big(\mathbf{C}_d,\ \mathbf{A} + \mathbf{BF}(\omega_{el}),\ \mathbf{g}(\omega_{el})\mathbf{T}_d\big), \\ &\big(\mathbf{C}_q,\ \mathbf{A} + \mathbf{BF}(\omega_{el}),\ \mathbf{g}(\omega_{el})\mathbf{T}_q\big), \end{aligned} \quad (26.15)$$

it holds the following conditions:

$$\mathscr{R}_d(\omega_{el}) = \min \mathscr{I}\Big(\mathbf{A} + \mathbf{BF}(\omega_{el}),\ \mathbf{g}(\omega_{el})\mathbf{T}_d(\omega_{el})\Big) \subseteq \mathscr{C}_q \quad \forall \omega_{el}, \quad (26.16)$$

and

$$\mathbf{C}_d \mathscr{R}_d(\omega_{el}) = \mathsf{im}(\mathbf{C}_d), \quad \forall \omega_{el}. \quad (26.17)$$

$$\mathscr{R}_q(\omega_{el}) = \min \mathscr{I}\Big(\mathbf{A} + \mathbf{BF}(\omega_{el}),\ \mathbf{g}(\omega_{el})\mathbf{T}_q(\omega_{el})\Big) \subseteq \mathscr{C}_d \quad \forall \omega_{el}, \quad (26.18)$$

and

$$\mathbf{C}_q \mathscr{R}_q(\omega_{el}) = \mathsf{im}(\mathbf{C}_q), \quad \forall \omega_{el}. \quad (26.19)$$

Here,

$$\min \mathscr{I}\left(\mathbf{A}, \mathsf{im}(\mathbf{BF})\right) = \sum_{i=0}^{n-1} \mathbf{A}^i \mathsf{im}(\mathbf{B})$$

is a minimum $\mathbf{A}$–invariant subspace containing $\mathsf{im}(\mathbf{B})$. Moreover, the partition matrix fields $\mathbf{T}_d(\omega_{el})$, $\mathbf{T}_q(\omega_{el})$ and $\mathbf{T}_0(\omega_{el})$ satisfy the following relationships:

$$\begin{aligned} \mathsf{im}\big(\mathbf{g}(\omega_{el}) \cdot \mathbf{T}_d(\omega_{el})\big) &= \mathsf{im}\big(\mathbf{g}(\omega_{el})\big) \cap \mathscr{R}_d(\omega_{el}), \\ \mathsf{im}\big(\mathbf{g}(\omega_{el}) \cdot \mathbf{T}_q(\omega_{el})\big) &= \mathsf{im}\big(\mathbf{g}(\omega_{el})\big) \cap \mathscr{R}_q(\omega_{el}). \end{aligned} \tag{26.20}$$

The stabilizing matrix field $\mathbf{F}(\omega_{el})$ is such that:

$$\big(\mathbf{A} + \mathbf{BF}(\omega_{el})\big)\mathscr{R}_d(\omega_{el}) \subseteq \mathscr{R}_d(\omega_{el}), \tag{26.21}$$

and

$$\big(\mathbf{A} + \mathbf{BF}(\omega_{el})\big)\mathscr{R}_q(\omega_{el}) \subseteq \mathscr{R}_q(\omega_{el}). \tag{26.22}$$

Considering

$$\mathbf{T}(\omega_{el}) = \Big[\mathbf{T}_d(\omega_{el}), \mathbf{T}_q(\omega_{el}), \mathbf{T}_0(\omega_{el}), \mathbf{T}_c(\omega_{el})\Big],$$

where $\mathbf{T}_c(\omega_{el})$ is defined in a complementary fashion and it is straightforward to show that matrix field $\mathbf{T}_c = \mathbf{0}$. In particular, matrix field $\mathbf{T}_c$ represents the complementary matrix field partition to the subspaces of d-coordinate, q-coordinate and 0-coordinate. The system is described using just three variables, therefore partition fields $\mathbf{T}_d(\omega_{el})$ and $\mathbf{T}_q(\omega_{el})$ complete the transformation and thus $\mathbf{T}_c = \mathbf{0}$.

$$\begin{aligned} \mathsf{im}\mathbf{T}(\omega_{el}) &= \mathsf{im}\Big[\mathbf{T}_d(\omega_{el}), \mathbf{T}_q(\omega_{el}), \mathbf{T}_0(\omega_{el})\Big] \\ &= \mathsf{im}\mathbf{T}_d(\omega_{el}) \oplus \mathsf{im}\mathbf{T}_q(\omega_{el}) \oplus \mathsf{im}\mathbf{T}_0(\omega_{el}). \end{aligned} \tag{26.23}$$

Considering the output matrix (26.13) corresponding to d-coordinate, q-coordinate and 0-coordinate, their respective kernels are as follows:

$$\mathscr{C}_d = \mathsf{im}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \mathscr{C}_q = \mathsf{im}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{26.24}$$

According to definition $\mathbf{B}$ of Eqs. (26.9) it is straightforward to observe that the following three equations hold $\forall\ \omega_{el}$:

$$\mathsf{im}(\mathbf{B}) \cap \mathscr{C}_q \neq \mathbf{0}, \tag{26.25}$$

$$\mathsf{im}(\mathbf{B}) \cap \mathscr{C}_d \neq \mathbf{0}. \tag{26.26}$$

The following calculations allow to get the required fields for the decoupling of the system:

$$\mathbf{T}_d(\omega_{el}) = \big(\mathbf{g}(\omega_{el})\big)^{\dagger} \cdot \mathsf{im}(\mathbf{B}) \cap \mathscr{C}_q, \tag{26.27}$$

$$\mathbf{T}_q(\omega_{el}) = \big(\mathbf{g}(\omega_{el})\big)^{\dagger} \cdot \mathsf{im}(\mathbf{B}) \cap \mathscr{C}_d. \tag{26.28}$$

Field $\mathbf{g}(\omega_{el})$ is a function of ω_{el} without singularities if $\omega_{el}t \neq k\pi$ with $k \in \mathbb{N}$, where with $(\mathbf{g}(\omega_{el}))^{\dagger}$ the pseudo inverse of field $\mathbf{g}(\omega_{el})$ is indicated. Adding all 3 T-Fields together, we get a new field $\mathbf{T}(\omega_{el})$:

$$\mathbf{T}(\omega_{el}) = \mathbf{T}_d(\omega_{el}) + \mathbf{T}_q(\omega_{el}). \tag{26.29}$$

Field $\mathbf{T}(\omega_{el})$ can be seen as a preselecting field and the following product realises the mechanical decoupling:

$$\mathscr{B} = \mathsf{im}\big(\mathbf{g}(\omega_{el})\mathbf{T}(\omega_{el})\big) = \mathsf{im}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{26.30}$$

in which matrix $\mathbf{B}$ can be seen as a resulting input matrix.

26.3.1 The Dynamic Estimator of Φ

As it is shown in Fig. 26.2, parameters R_s and L_{dq} can be estimated by using an ARMA identification structure. These two values are needed to estimate flux Φ. If the electrical part of the system "q" and "d" axes is considered, then, assuming that $\omega_{el}(t) \neq 0$, $i_q(t) \neq 0$, and $i_d(t) \neq 0$, the following equation can be considered:

$$\Phi(t) = -\frac{L_{dq}\frac{di_q(t)}{dt} + R_s i_d(t) + L_{dq}\omega_{el}(t) i_q(t) - u_q(t)}{\omega_{el}(t)}. \tag{26.31}$$

Consider the following dynamic system:

$$\frac{d\hat{\Phi}(t)}{dt} = -\mathscr{K}\hat{\Phi}(t) - \mathscr{K}\Big(\frac{\hat{L}_{dq}\frac{di_q(t)}{dt} + \hat{R}_s i_d(t)}{\omega_{el}(t)} + \frac{\hat{L}_{dq}\omega_{el}(t) i_q(t) + u_q(t)}{\omega_{el}(t)}\Big), \tag{26.32}$$

where $\mathscr{K}$ is a function to be calculated. Eq. (26.32) represents the estimators of Φ and $\hat{L}_{dq}$ and $\hat{R}_s$ represent the estimated inductance and resistance respectively by an ARMA procedures in [10]. If the error functions are defined as the differences between the true and the observed values, then:

$$e_{\Phi}(t) = \Phi(t) - \hat{\Phi}(t), \tag{26.33}$$

and

$$\frac{de_{\Phi}(t)}{dt} = \frac{d\Phi(t)}{dt} - \frac{d\hat{\Phi}(t)}{dt}. \tag{26.34}$$

If the following assumption is given:

$$\|\frac{d\Phi(t)}{dt}\| << \|\frac{d\hat{\Phi}(t)}{dt}\|, \tag{26.35}$$

then in Eq. (26.34), the term $\frac{d\Phi(t)}{dt}$ is negligible. Using Eqs. (26.32), (26.34) becomes

$$\frac{de_{\Phi}(t)}{dt} = \mathscr{K}\hat{\Phi}(t) + \mathscr{K}\Big(\frac{\hat{L}_{dq}\frac{di_q(t)}{dt} + \hat{R}_s i_d(t)}{\omega_{el}(t)} + \frac{\hat{L}_{dq}\omega_{el}(t)i_q(t) + u_q(t)}{\omega_{el}(t)}\Big). \tag{26.36}$$

Because of Eqs. (26.31), (26.36) being able to be written as follows:

$$\frac{de_{\Phi(t)}}{dt} = \mathscr{K}\hat{\Phi}(t) - \mathscr{K}\Phi(t),$$

and considering (26.33), then:

$$\frac{de_{\Phi}(t)}{dt} + \mathscr{K}\Phi(t) = 0. \tag{26.37}$$

$\mathscr{K}$ can be chosen to make Eq. (26.37) exponentially stable. To guarantee exponential stability, $\mathscr{K}$ must be

$$\mathscr{K} > 0.$$

To guarantee $\|\frac{d\Phi(t)}{dt}\| << \|\frac{d\hat{\Phi}(t)}{dt}\|$, then $\mathscr{K} >> 0$. The observer defined in (26.32) suffers from the presence of the derivative of the measured current. In fact, if measurement noise is present in the measured current, then undesirable spikes are generated by the differentiation. The proposed algorithm must cancel the contribution from the measured current derivative. This is possible by correcting the observed velocity with a function of the measured current, using a supplementary variable defined as:

$$\eta(t) = \hat{\Phi}(t) + \mathscr{N}(i_q(t)), \tag{26.38}$$

where $\mathscr{N}(i_q(t))$ is the function to be designed.

Consider

$$\frac{d\eta(t)}{dt} = \frac{d\Phi(t)}{dt} + \frac{d\mathscr{N}(i_q(t))}{dt} \tag{26.39}$$

and let

$$\frac{d\mathcal{N}(i_q(t))}{dt} = \frac{d\mathcal{N}(i_q)}{di_q(t)}\frac{di_q(t)}{dt} = \frac{\mathcal{K}\hat{L}_{dq}}{\omega_{el}(t)}\frac{di_q(t)}{dt}. \tag{26.40}$$

The purpose of (26.40) is to cancel the differential contribution from (26.32). In fact, (26.38) and (26.39) yield, respectively:

$$\hat{\Phi}(t) = \eta(t) - \mathcal{N}(i_q(t)), \quad \text{and} \tag{26.41}$$

$$\frac{d\hat{\Phi}(t)}{dt} = \frac{d\eta(t)}{dt} - \frac{d\mathcal{N}(i_q(t))}{dt}. \tag{26.42}$$

Substituting (26.40) in (26.42) results in:

$$\frac{d\hat{\Phi}(t)}{dt} = \frac{d\eta(t)}{dt} - \frac{\mathcal{K}\hat{L}_{dq}}{\omega_{el}(t)}\frac{di_q(t)}{dt}. \tag{26.43}$$

Inserting Eq. (26.43) into Eq. (26.32), the following expression is obtained:

$$\frac{d\eta(t)}{dt} - \frac{\mathcal{K}\hat{L}_{dq}}{\omega_{el}(t)}\frac{di_q(t)}{dt} = -\mathcal{K}\hat{\Phi}(t) - \mathcal{K}\Big(\frac{\hat{L}_{dq}\frac{di_q(t)}{dt} + \hat{R}_s i_d(t)}{\omega_{el}(t)} + \frac{\hat{L}_{dq}\omega_{el}(t)i_q(t) + u_q(t)}{\omega_{el}(t)}\Big), \tag{26.44}$$

then:

$$\frac{d\eta(t)}{dt} = -\mathcal{K}\hat{\Phi}(t) - \mathcal{K}\frac{\Big(\hat{R}_s i_d(t) + \hat{L}_{dq}\omega_{el}(t)i_q(t) + u_q(t)\Big)}{\omega_{el}(t)}. \tag{26.45}$$

Letting $\mathcal{N}(i_q(t)) = k_{app}i_q(t)$, where a parameter has been indicated with k_{app}, then from (26.40) $\Rightarrow \mathcal{K} = \frac{k_{app}\omega_{el}(t)}{\hat{L}_{dq}}$, and Eq. (26.41) becomes:

$$\hat{\Phi}(t) = \eta(t) - k_{app}i_q(t). \tag{26.46}$$

Finally, substituting (26.46) in (26.45) results in the following equation:

$$\frac{d\eta(t)}{dt} = -\frac{k_{app}\omega_{el}(t)}{\hat{L}_{dq}}\big(\eta(t) - k_{app}i_q(t)\big) + \frac{k_{app}}{\hat{L}_{dq}}\Big(\hat{R}_s i_d(t) + \hat{L}_{dq}\omega_{el}(t)i_q(t) + u_q(t)\Big),$$

$$\hat{\Phi}(t) = \eta(t) - k_{app}i_q(t). \tag{26.47}$$

Using the implicit Euler method, the following velocity observer structure is obtained:

$$\eta(k) = \frac{\eta(k-1)}{1 + t_s \frac{k_{app}\omega_{el}(k)}{\hat{L}_{dq}}} + \frac{t_s \frac{k_{app}^2 \omega_{el}(k) i_q(k)}{\hat{L}_{dq}} + k_{app}\omega_{el}(k) i_q(k) + \frac{t_s \hat{R}_s k_{app} i_d(k)}{\hat{L}_{dq}}}{1 + t_s \frac{k_{app}\omega_{el}(k)}{\hat{L}_{dq}}} i_q(k)$$

$$+ \frac{t_s \frac{k_{app}}{\hat{L}_{dq}}}{1 + t_s \frac{k_{app}\omega_{el}(k)}{\hat{L}_{dq}}} u_q(k),$$

$$\hat{\Phi}(k) = \eta(k) - k_{app} i_q(k), \tag{26.48}$$

where t_s is the sampling period.

Remark 1. Assumption (26.35) states that the dynamics of the approximating observer should be faster than the dynamics of the physical system. This assumption is typical for the design of observers. □

Remark 2. The estimator of Eq. (26.48) presents the following limitations: for low velocity of the motor ($\omega_{mec.}(t) << \omega_{mec_n}(t)$), where $\omega_{mec_n}(t)$ represents the nominal velocity of the motor), the estimation of Φ becomes inaccurate. Because of $\omega_{el}(t)$ dividing the state variable η, the observer described by (26.48) becomes hyperdynamic. Critical phases of the estimation are the starting and ending of the movement. Another critical phase is represented by a high velocity regime. In fact, it has been proven through simulations, that if $\omega_{mec}(t) >> \omega_{mec_n}(t)$, then the observer described by (26.48) becomes hypodynamic. According to the simulation results, within some range of frequency, this hypo-dynamicity can be compensated by a suitable choice of k_{app}. □

Remark 3. The Implicit Euler method guarantees the finite time convergence of the observer for any choice of k_{app}. Nevertheless, any other method can demonstrate the validity of the presented results. Implicit Euler method is a straightforward one. □

26.4 Simulation and Measured Results

Results have been performed using a special stand with a 58-kW traction PMSM. The stand consists of a PMSM, a tram wheel and a continuous rail. The PMSM is a prototype for low floor trams. The PMSM parameters are: nominal power 58 kW, nominal torque 852 Nm, nominal speed 650 rpm, nominal phase current 122 A, nominal input voltage 230 V and number of poles 44. The model parameters are: $R = 0.08723$ Ohm, $L_{dq} = Ld = Lq = 0.8$ mH, $\Phi = 0.167$ Wb. The engine has a nominal power 55 kW, a nominal voltage 380 V and nominal speed 589 rpm.

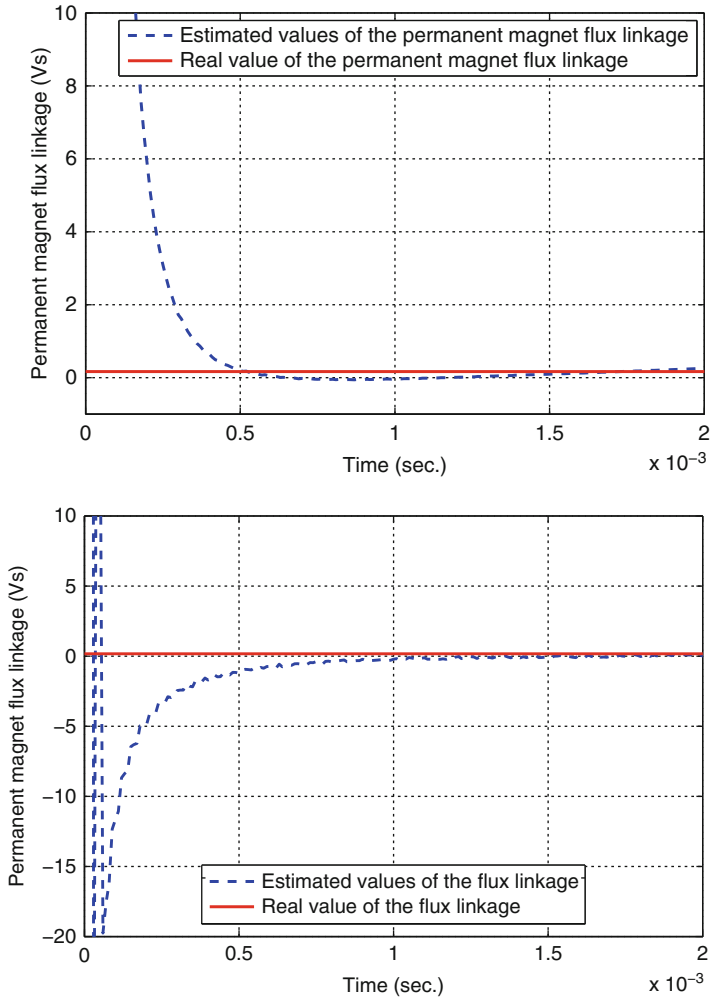


Fig. 26.3 Simulated results: estimated and real values of the permanent magnet flux linkage for $k_{app} = 20$. Simulation results (on the *top*) and measured results (on the *bottom*)

Figure 26.3 shows the estimation of Φ magnet flux starting from R_s stator resistance and L_{dq} inductance. These simulation and measured results are obtained using values of k_{app} equal to 20 and at $t = 0$ it corresponds $\omega_{el}(t) = 0$. From these figures, the effect of the limit of the procedure discussed in remark 2 is visible at the beginning of the estimation. In particular this effect is visible in the real measured results. Figure 26.4 shows a detail of the estimation of the measured magnetic flux of the motor.

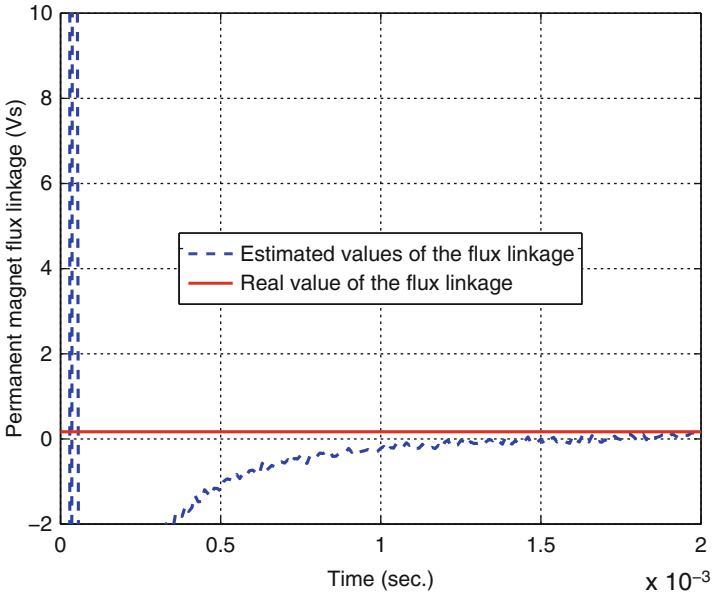


Fig. 26.4 Detail of the estimation of the measured magnetic flux

26.5 Conclusions and Future Work

This paper considers a decoupling dynamic estimator for fully automated parameters identification for three-phase synchronous motors. The proposed strategy uses the geometric approach to realise a decoupling of the system. The estimation of the parameters of the motor is simplified through a decoupling. The decoupling is realised using a feedback controller combined with a feedforward one. The feedforward controller is conceived through an input partition matrix. The proposed dynamic estimator is shown to identify the amplitude of the linkage flux using the estimated inductance and resistance. Through simulations and measured results on a synchronous motor used in automotive applications, this paper verifies the effectiveness of the proposed method in identification of PMSM model parameters and discusses the limits of the proposed procedure. Simulation and measured results are reported to validate the proposed strategy. Future work includes the estimation of a mechanical load and the general test of the present algorithm using a real motor.

References

1. Mercorelli, P.: Invariant subspace for grasping internal forces and non-interacting force-motion control in robotic manipulation. Kybernetika **48**(6), 1229–1249 (2012)
2. Mercorelli, P.: Geometric structures using model predictive control for an electromagnetic actuator. WSEAS Trans. Syst. Control **9**, 140–149 (2014)
3. Rahman, M.A., Vilathgamuwa, D.M., Uddin, M.N., King-Jet, T.: Nonlinear control of interior permanent magnet synchronous motor. IEEE Trans. Ind. Appl. **39**(2), 408–416 (2003)
4. Mercorelli, P.: A Lyapunov approach for a pi-controller with anti-windup in a permanent magnet synchronous motor using chopper control. Int. J. Math. Models Methods Appl. Sci. **8**, 44–51 (2014)
5. Kilthau, A., Pacas J.: Appropriate models for the controls of the synchronous reluctance machine. In: Proceedings IEEE IAS Annual Meeting, pp. 2289–2295 (2002)
6. Weisgerber, S., Proca, A., Keyhani, A.: Estimation of permanent magnet motor parameters. In: Proceedings of the 32^{nd} IEEE Industrial Applications Society Annual Meeting, New Orleans, pp. 29–34 (1997)
7. Khaburi, D.A., Shahnazari, M.: Parameters identification of permanent magnet synchronous machine in vector control. In: Proceedings of the 10th European Conference on Power Electronics and Applications (EPE 2003), Toulouse, 2–4 September 2003
8. Mercorelli, P., Lehmann, K., Liu, S.: On robustness properties in permanent magnet machine control using decoupling controller. In: Proceedings of the 4th IFAC International Symposium on Robust Control Design, Milan, 25–27 June 2003 (2003)
9. Mercorelli, P.: Robust feedback linearization using an adaptive pd regulator for a sensorless control of a throttle valve. Mechatronics. J. IFAC. **19**(8), 1334–1345 (2009). doi:10.1016/j.mechatronics.2009.08.008
10. Mercorelli, P.: A decoupling dynamic estimator for online parameters indentification of permanent magnet three-phase synchronous motors. In: Proceedings of the 16th IFAC Symposium on System Identification, pp. 757–762 (2012)
11. Basile, G., Marro, G.: Controlled and Conditioned Invariants in Linear System Theory. Prentice Hall, New Jersey (1992)
12. Wonham, W.M., Morse, A.S.: Decoupling and pole assignment in linear multivariable systems: A geometric approach. SIAM J. Control **8**(1), 1–18 (1970)
13. Bhattacharyya, S.P.: Generalized controllability, controlled invariant subspace and parameter invariant control. SIAM J. Algebr. Discrete Methods **4**(4), 529–533 (1983)

Chapter 27
Application of the Monte Carlo Method for the Determination of Physical Parameters of an Electrical Discharge

Leyla Zeghichi, Leïla Mokhnache, and Mebarek Djebabra

Abstract The aim of this paper is to use of the Monte Carlo method to try to reproduce an electrical discharge in the oxygen gas; by following the random histories of free electrons and using the sampling laws, we can determine some electrical discharge parameters. Additionally we use the simulation results to verify the electrical breakdown criteria under an homogenous field and for small distances. The obtained results are compared with those collected from literature.

Keywords Attachment • Collision probability • Electrical discharge • Ionization • Mean free flight time • Monte Carlo method

27.1 Introduction

In their normal state of temperature and pressure, the gases are perfect insulations. The reason for this, that they contain only neutral species (molecules and atoms) [1–5]. The conduction in air at low field is in the interval 10^{-16}–10^{-17} A/cm^2. This current, results from cosmic radiations and radioactive substances present in earth and the atmosphere. However, when applying a sufficiently strong electric field between two electrodes immersed in a gaseous medium, it becomes more or less conductive and an electrical breakdown occurs [2]. So, the complex phenomena that occur are called electrical discharge in gases [4].

L. Zeghichi (✉)
Department of Physics, Ouargla University, P.O. Box 511, Ouargla, 30000 Algeria
e-mail: zeghichi.leyla@univ-ouargla.dz

L. Mokhnache
Department of Electrical Engineering, Batna University, Batna, Algeria
e-mail: mokhnache@yahoo.fr

M. Djebabra
Institute of Health and Safety, Batna University, Batna, Algeria
e-mail: mebarek_djebabra@yahoo.fr

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_27

As a rule, an electrical discharge in gases is produced mainly via ionization by collision, photo-ionization, and the secondary ionization processes. In insulating gases (also called electron-attaching gases) the process of attachment also plays an important role [6]. It follows the generation of new electrons and ions in the Townsend avalanches that grow up until a maintenance state is established. The discharge becomes then independent of the external sources which produce free electric charges in the gas.

Several studies [6–11] have been made in the framework of the discharge modeling to improve the understanding of the fundamental processes of electrical discharges in gases, with the aim is to obtain good results.

A model must reproduce as finely as possible the physical phenomena involved in the studied system. The task of the modeler is first to identify the main characteristics of the physical problem, and formulate them mathematically. Because of the complexity of the studied systems, the mathematical representation is related to the choice of approximations and hypotheses that make the problem resolvable.

The Monte Carlo Method (MCM), based on stochastic laws, is used in our computation because that it considers several stochastic events during the process of the electrical discharge; it consists in simulation of a large number of events (charged particles) by another set easily achievable (random variables). The simulation of electron motion has been performed in order to accurately calculate the ionization and the attachment rates as well the mean kinetic energy.

In this paper we will study, in the oxygen gas, the temporal evolution of an electrical discharge at an atmospheric pressure; moreover the influence of pressure and applied voltage on the breakdown inception is to be considered.

27.2 The Monte Carlo Method

Physical concepts of MCM applied to the study of an electrical discharge are [9]: the mean free flight time, the mean free path or null-collision Monte Carlo Method.

The application of the constant step MCM version for the study of electrons' motion, under the effect of an applied electric field, requires the evaluation process of the different physical parameters after experiencing energy loss and gain (taking into account the different processes of atomic collisions elastic or inelastic).

27.2.1 The Model

At time $t = 0$, the initial electrons are emitted from the cathode according to a cosine distribution. The energy gain of the electrons in a small time interval 'dt' is governed by the equation of motion.

Before collision; the velocity and position components', for each electron, are given by [7]:

$$\begin{aligned} v_{y0} &= v_0 \sin(\theta)\cos(\varphi) \\ v_{y0} &= v_0 \cos(\theta)\sin(\varphi) \\ v_{z0} &= v_0 \cos(\theta) \end{aligned} \tag{27.1}$$

where v_0 is given by:

$$(2\varepsilon_0 (e/m))^{1/2} \tag{27.2}$$

And the θ and φ angles are defined as functions of a random number r such as:

$$\cos(\theta) = 1 - 2r \tag{27.3}$$

$$\varphi = 2\pi r \tag{27.4}$$

The collision probability P_1, that follows the Poisson's distribution, is given by:

$$P_1 = 1 - \exp(dt/Tm) \tag{27.5}$$

Tm is the electron mean free flight time between two successive collisions; it is determined by the electron collision total cross section $Q(\varepsilon)$ as:

$$Tm = [N \cdot Q(\varepsilon) \cdot v(\varepsilon)]^{-1} \tag{27.6}$$

where $v(\varepsilon)$ is the velocity of electrons and N the gas number density.

Before a collision process; the velocity and position components',

$$\begin{aligned} v_x &= v_{x0} \\ v_y &= v_{y0} \\ v_y &= v_{y0} + a \cdot dt \end{aligned} \tag{27.7}$$

$$\begin{aligned} x &= x_0 + v_x \cdot dt \\ y &= y_0 + v_y \cdot dt \\ z &= z_0 + v_{z0} \cdot dt + a \cdot dt^2 \end{aligned} \tag{27.8}$$

27.2.2 Simulation Implementation

The flight time is divided into a number of smaller elements according to:

$$dt = Tm_0/K \tag{27.9}$$

where K is a sufficiently large integer and T_{m0} is the mean free flight time.

The occurrence of collision between an electron and a gas molecule and its kind are determined by comparison of the collision probability P_1 with computer generated random numbers r.

The interval [0, P_1] is divided into segments of lengths that correspond to the probabilities of different types of collision after increasing scheduling of these probabilities. The remaining portion of the interval [0, 1] is for the case where no collision is possible (Fig. 27.1).

The electron energy is described for the different processes of collision:

– For the elastic collision the energy is given by:

$$\varepsilon_1 = [1 - 2\,(M/m)\cos(\delta)]\,\varepsilon_0 \tag{27.10}$$

where δ is the scattering angle of the electron after the collision, m and M are, respectively, the mass of electron and an O_2 molecule and ε_0 is the electron's energy before collision.

Next to the processes of attachment, excitation and ionization, the onset energy "los" is subtracted from the electron energy as follow:

– For an attachment of the electron, all its energy is to be lost, and therefore this electron is lost in the swarm

$$\varepsilon_1 = 0 \tag{27.11}$$

– For an exciting process of a molecule to a higher state (different rotations, vibrations and electronic excited states), the energy of the electron is reduced with the energy needed to excite the molecule and the resulting energy is given by:

$$\varepsilon_1 = \varepsilon_0 - \text{los} \tag{27.12}$$

– Finally, for an ionizing process, the remaining energy is shared between the primary and the ejected electrons with the ratios r and (1–r) as:

$$\varepsilon_{1\text{primary}} = (\varepsilon_0\text{–los})\,r \tag{27.13}$$

$$\varepsilon_{1\text{secondary}} = (\varepsilon_0\text{–los})\,(1\text{–r}) \tag{27.14}$$

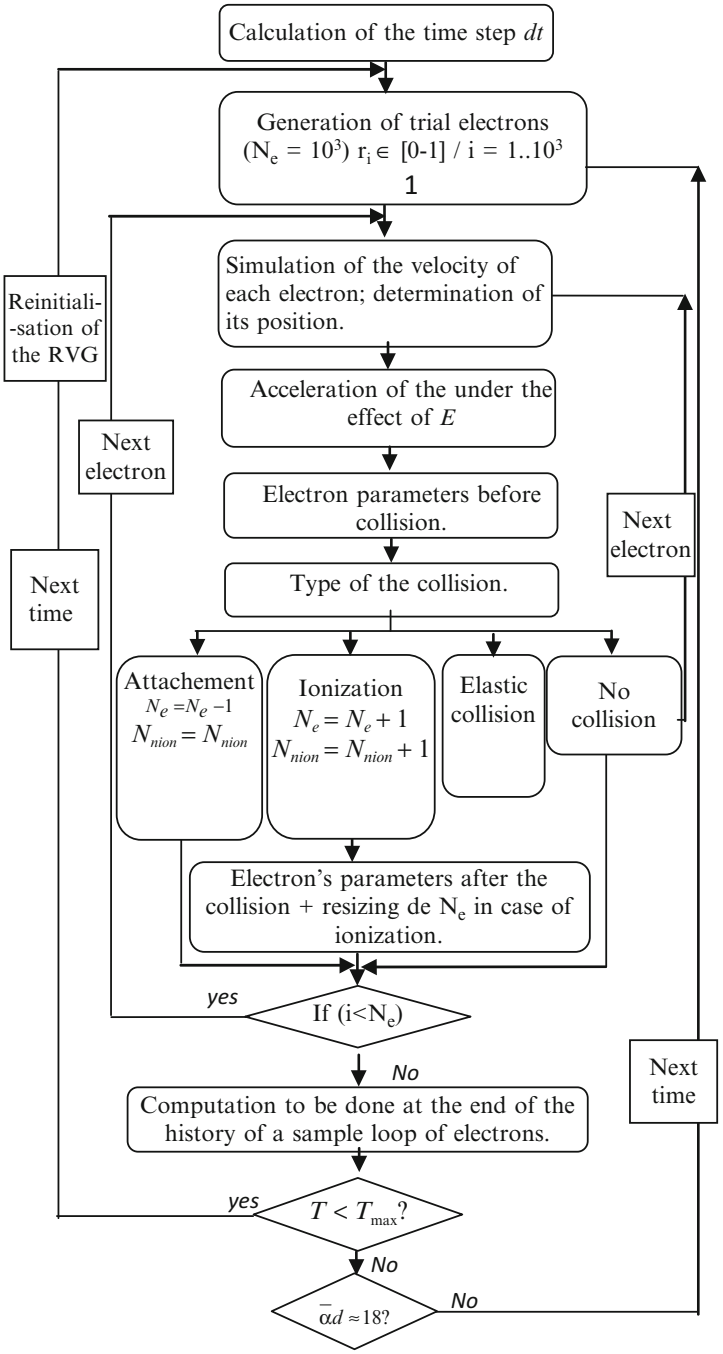


Fig. 27.1 MC simulation flow chart

27.2.3 Exploitation of the Simulation Results

By means of tracking the history of each individual in a population upon appearance until its disappearance [8]; several values (the electron's energy, its position, the number of positive and negative ions, etc.) can be stored. Through averaging on the histories, different properties can be determined.

$$\bar{z} = (1/N)\,\mathrm{sum}\,(z_i) \tag{27.15}$$

$$\bar{\varepsilon} = (1/N)\,\mathrm{sum}\,(\varepsilon_i) \tag{27.16}$$

27.3 Simulation and Results

In this paper we describe the development of an electric discharge in the oxygen O_2 gas within plane-plane geometry. At time $t = 0$, a number of electrons are released from the cathode with small energy 0.1 eV. The calculations of the physical parameters are performed at atmospheric pressure for an applied voltage of 57 kV. The electrical breakdown criteria are checked under different gas pressures for several voltages values. The cross section set of the O_2 molecule used is that referred in [12].

27.3.1 The Physical Parameters

Figures 27.2, 27.3, and 27.4 represent, respectively, the temporal variation of the ionization (α), attachment (η) rates and of the mean kinetic energy. As much as the space charge accumulates in the time, the induced field increases.

In Figs. 27.2 and 27.3, we note that: the ionization coefficient, in the case of the atmospheric pressure under 57 kV and a temperature of 293 K (20 °C), increases and reaches high values in a short delay; and that the electrons in the beginning of the discharge are attached because they have, initially, low energies and the attachment rate increases and then stabilizes at a certain value.

Figure 27.4 shows that the electrons have a short free path (in 2 cm interval) and as the energy gain depends on the travelled distance, the mean kinetic energy of the electrons does not vary significantly.

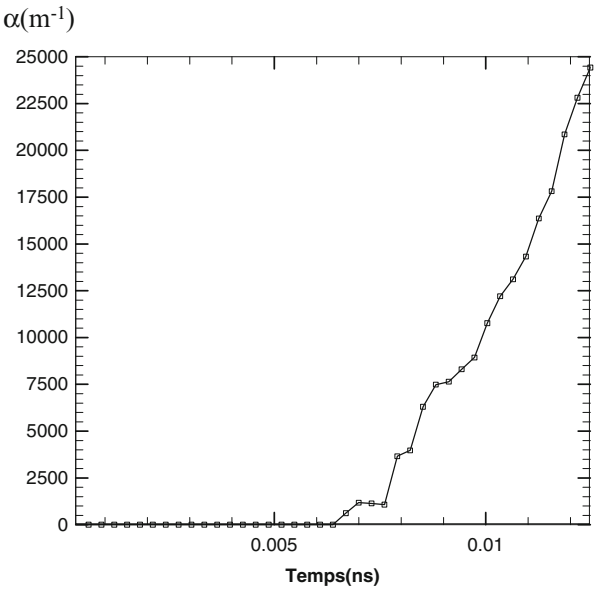
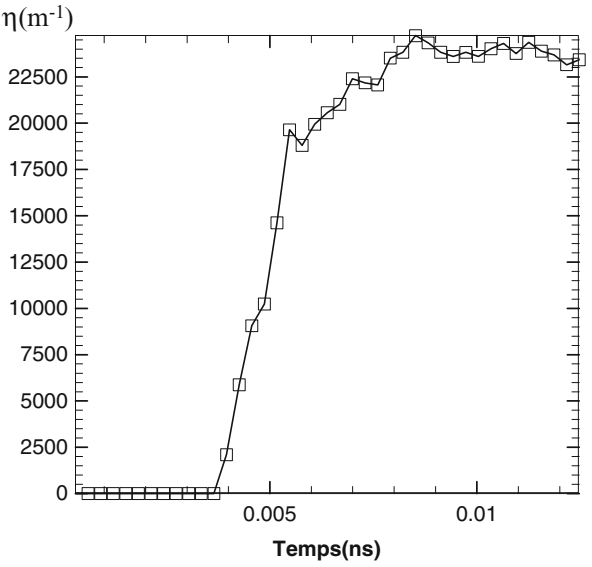


Fig. 27.2 Temporal variation of the ionization coefficient (P = 1 atm, V = 57 kV)

Fig. 27.3 Temporal variation of the attachment coefficient (P = 1 atm, V = 57 kV)



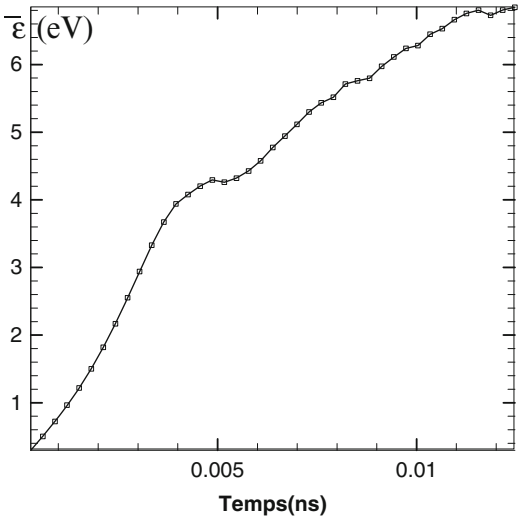


Fig. 27.4 Temporal variation of the mean kinetic energy (P = 1 atm, V = 57 kV).

Table 27.1 Effect of voltage and pressure on the electrical breakdown onset

P	V (V)	α (m^{-1})	η (m^{-1})	$(\alpha - \eta) \times d$
1 Torr	10	0.00	4.68E−01	−9.37E−03
	200	116.46	4.845873	2.23
	600	412.38	9.65E−01	8.22
1 atm	40,000	737.36	2,825.32	−41.76
	50,000	3,108.40	2,679.79	8.57
	55,000	26,693.99	25,623.02	21.41
5 atm	2,000	0.00	0.00	0.00
	20,000	0.00	5.11E−02	−1.02E−03
	60,000	0.00	91.477350	−1.80

27.3.2 Check of the Breakdown Criteria

For different pressures and different voltages with an inter-electrode distance of 2 cm, the values of the ionization and attachment coefficients, and values of the mean kinetic energy are collected in Table 27.1.

For the voltages 10, 40,000, 20,000 and 60,000 V, which correspond, respectively, to the pressures 1 Torr, 1 atm and 5 atm, the values of $(\alpha - \eta) \times d$ are negative that means that $\alpha < \eta$ and there is not effective ionization.

For the voltages 200 and 50,000 V which correspond, respectively, to the pressures 1 Torr and 1 atm, the values of $(\alpha - \eta) \times d$ are positive but small because there are ionizing collisions between the electrons and molecules which are not enough to ensure that the discharge is maintained.

Electrical discharges in gases are of two types (1) non-self sustaining, (2) self sustaining. The breakdown in a gas is the transition from (1) to (2). The increase of current in the breakdown is due to the ionization process in which ion-electron pairs

are created by collisions with neutral atoms and their drift towards the anode and the cathode. For 1 Torr and under 600 V, we have a self-sustaining discharge of the Townsend type and for 1 atm under 55,000 V the streamer breakdown criterion is checked.

27.4 Conclusion

The idea of this paper was to use the Monte Carlo method to study and follow the fundamental processes of an electrical discharge, which have a random aspect, through computing its physical parameters.

The simulation results give ionization and attachment coefficients values and electrons mean kinetic energy, as functions of time. When the voltage is sufficiently high, the ionization coefficient increases, and the gas becomes conductor, and therefore the onset of the electrical breakdown.

By means of the simulation results we have verified the breakdown criteria for different values of applied electric field (applied voltages) and pressure. For the low pressures, the discharge phenomenon is a Townsend one and for the high pressures the discharge happened by Streamer).

References

1. Meek, J.M., Craggs, J.D.: Electrical Breakdown of Gases. Clarendon, Oxford (1953)
2. Kuffel, E., Zaengl, W.S., Kuffel, J.: High Voltage Engineering Fundamentals, 2nd edn, p. 534. Butterworth-Heinemann, Oxford (2000)
3. Raju, G.G.: Dielectrics in Electric Field. Marcel Dekker, New York (2003)
4. Raizer, Y.R.: Gas Discharge Physics. Springer, Berlin (1991)
5. Naidu, M.S.: High voltage engineering, 2nd edn. Quebecor/Book Press, New York (1995)
6. Settaouti, A., Settaouti, L.: Monte Carlo simulation of electron swarm parameters in O_2. Eur. Phys. J. Appl. Phys. **37**, 335–341 (2007)
7. Govinda, G.R., Liu, J.: Simulation of electrical discharges in gases—uniform electric fields. IEEE Trans. Dielectr. Electr. Insul. **2**(5), 1004–1015 (1995)
8. Djillali, B., Bachir, B.: Etude d'une décharge luminescente continue dans l'argon Par la méthode de Monte Carlo. In: CNHT'2003—5éme Conférence sur la Haute Tension—USTMB, Oran, pp. 61–66 (2003)
9. Raju, G.G.: Dielectrics in Electric Fields. Power Engineering, vol. 19, 1st edn. CRC Press, New York (2003)
10. Zeghichi, L., Mokhnache, L., Djebabra, M.: The Monte Carlo Method for the study of an electrical discharge. In: 2013 IEEE International Conference on Solid Dielectrics (ICSD), Bologna, Italy, June 30–July 4, pp. 636–639 (2013)
11. Zeghichi, L., Mokhnache, L., Djebabra, M.: Monte Carlo simulation for an electrical discharge in O_2. Adv. Mater. Res. **227**, 211–214 (2011)
12. Phelps, A.V.: Atomic & molecular physics. JILA NIST-CU website. [Online] (2005). ftp://jila.colorado.edu/collision_data/electronneutral/ELECTRON.TXT

Chapter 28
Intersection Management Based on V2I Velocity Synchronization

Xuguang Hao, Abdeljalil Abbas-Turki, Florent Perronnet, Rachid Bouyekhf, and Abdellah El Moudni

Abstract In this last decade, new approaches for managing intersections have emerged. Instead of relying on the traffic lights, vehicles negotiate together the better way for sharing the junction. They explore the opportunities of new advances in terms of cooperative driving and unmanned vehicles. Vehicles are than able not only to communicate with the surrounding environment but also to control their speed. Many protocols are proposed for sharing the conflicting space, such as time reservation, priorities between vehicles and sequences. In this paper we consider the speed synchronization through V2I. Vehicles adapt their speed according to their position. This paper firstly reviews some proposed protocols so as to introduce the synchronization of vehicles' velocity. The synchronization is based on the Sequence-Based Protocol (SBP). So the discussion mainly focus on it. Before proposing the approach, some practical and theoretical problems are highlighted for clarity. Finally, we perform and discuss the simulation of vehicles in to a junction loop. The simulation shows a high level of stability even with several vehicles.

Keywords Advanced cruise control • V2I • Velocity synchronization

28.1 Introduction

In modern cities, traditional solutions of transportation congestions, such as traffic lights [1, 2] and road planning, are encountering more and more different kinds of difficulties, especially in big cites. At present, many researchers focus on improving the urban intersection management with the latest technologies, for example wireless communication [3, 4], positioning, advanced cruise control [5] and so on.

With these informational technologies, people could be able to consider the transportation problems in different visions. The application of these modern tools

X. Hao (✉) • A. Abbas-Turki • F. Perronnet • R. Bouyekhf • A. El Moudni
Université de Technologie de Belfort-Montbéliard
90010 Belfort cedex, France
e-mail: xuguang.hao@utbm.fr; abdeljalil.abbas-turki@utbm.fr; florent.perronnet@utbm.fr; rachid.bouyekhf@utbm.fr; abdellah.el-moudni@utbm.fr; http://www.utbm.fr

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_28

to the planning and management of intersection is one of the core solutions of urban traffic congestion so as to improve the traffic capacity and efficiency in future. Some works are focusing on Cooperative Intersection Management (CIM) in order to control the passage of vehicles at urban intersection without traffic lights. They are based on the rapid progressing of vehicles equipped with on-board computer, wireless modules, sensors and so on. More precisely, the intelligent vehicle's function includes positioning, wireless communications between vehicles and infrastructure (V2I) as well as controllable motion. The intersection infrastructure and vehicle, as fundamental components of intersection management, all play important roles.

We introduce a new approach for synchronizing vehicles' velocity so as to safely and efficiently traverse the intersection.

This paper is organized as follows. Firstly, it gives an overview of last works on cooperative intersection management. Then, we present the components of Transparent Intersection Management (TIM) and the sequence policy that we adopt. Some important characteristics will be presented and will be used in the fourth section. In the fourth part, basing on sequence-based protocol, we propose a new approach for synchronizing the speed of vehicles. A simple simulation that applying the new approach at a intersection will be shown later. After that we make a more complex simulation to show the effect of applying the approach more than one time on same vehicles at the same intersection. Finally, we conclude on the approach advantages.

28.2 Literature Overview

28.2.1 Reservation-Based Protocol

Based on a multi-agent model, Dresner and Stone have presented the Reservation-Based Protocol in [6]. Vehicles and intersections are implemented as intelligent agents able to communicate together. When the vehicle approaching the intersection, it sends a request for the right-of-way that is kinetic parameters of the vehicle as well as its destination. Accordingly, the intersection manager simulates the journey of the vehicle in the gridded intersection map, in order to reserve space and time of the greeds. It reject the requests of others until the end of reservation. It shows that it is possible to make intersection control much more efficient than traditional control mechanisms.

In [7], the authors implemented a mixed reality platform with a real autonomous vehicle 'Marvin' which could interact with multiple virtual vehicles in a simulation at a real intersection. Its experiment shows that, with several techniques, the Autonomous Intersection Management (AIM) protocol outperform traffic signals. The test result of [8] shows that the protocol has potential to decrease vehicular delay. The more recent work [9] explored the possibility of applying autonomous vehicle auctions at each intersection to determine the order using autonomous reservation protocols with a microscopic simulator performing on city-scale maps.

The mixed reality platform [7] has shown that vehicles are not so controllable as it has been assumed. However, the collision avoidance is strongly dependent on the speed and on the time of traversing the intersection. Thus, there is a high collision risk if there is a tight timing between two vehicles whereas a high idle time between two vehicles will significantly compromises the traffic efficiency without entirely eliminate collision risk.

28.2.2 Intersection-Based Cooperative Adaptive Cruise Control Protocol

Another team has proposed a heuristic optimization algorithm for controlling the automated vehicles at traditional intersections with a game theory framework entitled CACC-CG [10, 11]. The framework is considered as a decision process that repeats at each time step of simulation to optimize the movement of automated vehicles. The protocol controls trajectories of vehicles which are equipped with Cooperative Adaptive Cruise Control (CACC) to avoid collisions and minimize vehicle's delay and consequently reduce the total delay of intersection.

In [11], the authors clearly proposed the Intersection-based Cooperative Adaptive Cruise Control Protocol (iCACC). In order to optimize the movement of autonomous vehicles, three zones are assumed that they fall in the vehicle trajectory. The "smart" intersection takes into account the physical characteristics that may affect the motion of vehicles to simulate and to optimize the speed of the vehicles. As being fully equipped, the vehicle must adapts itself in Zone II in order to control the point of time that it arrives at the conflict zone. The speed adaptation makes sure the vehicle will reach its maximum velocity when it cross the Conflict Point. And consequently, the vehicle will pass the intersection box at the maximum speed. In iCACC protocol, the fundamental is that, basing on gridded intersection zone, the manager simulates and assumes the journey of vehicle based on the current situation and make a precise reservation that the intelligent vehicle must obey.

As for the reservation protocol, it is hard to guarantee that the vehicle will traverse the intersection at exactly the mentioned high speed. Moreover, the intersection sends messages for slowing down vehicles. However, because of message drop and loss there is a high collision risk.

28.2.3 Sequence-Based Protocol

The intersection and vehicle all play important roles in decision process of intersection management. At this point, there are a lot of works could be done to improve current traffic efficiency. In order to define roles in a better way, the Sequence-Based

Protocol (SBP) was proposed [12–15]. As named as Sequence-Based Protocol, one fundamental of this protocol is sequencing all the "full-automated" vehicles that are waiting to pass the intersection.

In "this" centralized protocol, basing on the informations that collected from these vehicles, the intersection manager could apply optimization methods to form the passing sequence of vehicles. The sequence means deciding explicitly which vehicle will traverse the intersection first, which is the second vehicle and so on. There is no conditions on times and speeds. Hence, the protocol can be applied for automated vehicles or manned vehicles. For safety reasons, the vehicle cannot traverse the intersection without a consent from the intersection manager. Hence, a vehicle that has not received a message from the intersection automatically stop before the junction box that we will call later conflict zone. Hence, the principle of the default deny is chosen.

For manned vehicles, the intersection manager assigns only "right-of-way" to vehicles. The right-of-way is a green that allows a vehicle to safely traverse the intersection. The right-of-way can be distributed to several vehicles simultaneously if there is no conflict. Hence the intersection manager permanently broadcasts message with a list of vehicles that have the right-of-way. The vehicle checks if it is in the list.

The unmanned vehicles considers the "right-of-way" more complex than a simple green. As in the case of manned vehicles, the intersection manager sends a list of allowed vehicles with their movement parameters that are speed, position, movement direction, destination, etc. The list of vehicles is ordered according to the decided sequence. The main difference between the intersections of manned and of unmanned vehicles is that two unmanned vehicles with conflicting movements are included in the same list. This means that both one must synchronizes its speed to do not collide with the other. In general case, an unmanned vehicles should observe all precedent vehicles in the list to avoid collision. So the main raised issue by the sequence based protocol is how to synchronize the speed between all unmanned vehicles. This issue is treated as the core of this paper.

Currently, there are two approaches for synchronizing speeds. The first one assumes that the vehicle immediately slow down until it gets enough space to traverse safely the intersection. The second one assumes that the vehicle slows down to completely stop before the conflict zone but if there is enough space to traverse safely the intersection during the slowing down, the vehicle speed up and traverse the intersection. The main issue with both approaches is that vehicles slow down near the extremity of each side of the lane. This can cause congestions at near intersections if the traffic flow is high. Moreover, both approaches have been developed, simulated and tested for only mini-robots. Since test of both approaches have shown that they are safe and they allow a good performance, then they deserve to be improved in terms of lane occupancy.

28.3 Studied Intersection

Section 28.2.3 has given a simple description of the two existing approach for synchronizing speed of vehicles that listed in one sequence. To express the work of this paper more clearly, Fig. 28.1 shows the two approaches controlling the vehicle coloured yellow .

Controlling by using the first approach, the vehicle will immediately slow down at the beginning of lane to get enough space. Its lower speed even stop will cause the access of the lane that it just has entered be locked. This occupancy of the access of lane will consequently influence the motion of vehicle listed in the sequence but behind it and will also run into the same lane. The sequence is the one maintained by the left one in Fig. 28.1. The follower will slow down or stop before the conflict zone of the same intersection. As all the vehicles are maintained in one sequence, slowing down of one vehicle may leads to a high congestion in high flow, just as shown in Fig. 28.1. Another possible consequence of first approach is the lower usage of lane.

The second approach doesn't have the usage problem, but will influence the fluidness of its downstream intersection shown as right part of Fig. 28.1. Because of sequence of vehicles, the stop of yellow vehicle before stop line will also finally lead to the stop of the vehicles that listed in the sequence but behind it. Then the congestion will occurs in high traffic flow. Another problem of this approach comes from the speeding up of yellow vehicle when it is traversing the conflict zone. Especially in high flow, the low speed may lead to the slowing down of the vehicles which running on the other lanes and preparing to traverse the same conflict zone.

The most basic reasons of issue of the two approaches is the juncture and strategy of speed adjustment. At extreme situation, the stop of yellow vehicle could not be avoided. So in order to improve the performance of the two approaches, this paper focus on when and where to slow down the vehicle. The target becomes releasing the occupancy of the access of lane to increase the usage of lane and getting velocity as high as possible when the vehicle traversing the conflict zone. This will also have influence as low as possible to the vehicles which are running on other lanes and preparing to pass the same conflict zone.

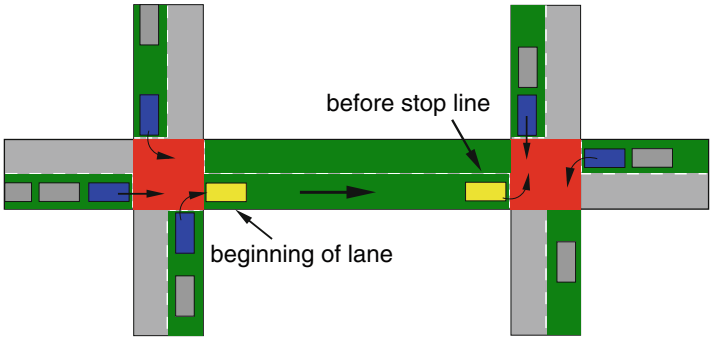


Fig. 28.1 Current two approaches for synchronizing speeds

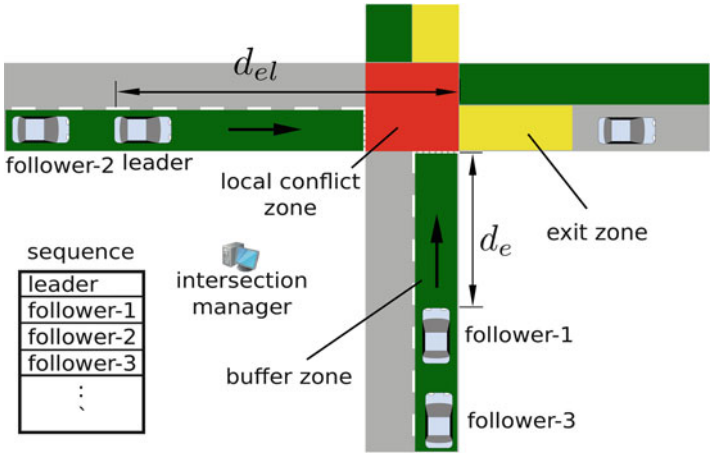


Fig. 28.2 Structure of intersection

28.3.1 Structure of Intersection

This paper focus on the intersection that the vehicle we will try to control is approaching. The intersection is a 4-leg one which has 4 input lanes and 4 output lanes as shown in Fig. 28.2. The protocol of intersection management is the Transparent Intersection Management (TIM) [14]. It uses client/server architecture to organize the management of intersection.

The necessary components of TIM are as follows:

- Intersection Manger (server) maintains the sequence of vehicles and could optimize it as needed.
- The autonomous vehicle (client) could communicate with the intersection manager. It is equipped with the new control protocol.
- The green ones are the buffer zone. All clients moving on them will communicate with server for negotiating its right-of-way and exchanging other informations which are necessary in the management process.
- The red zone is conflict zone. The manager will make sure that, with right-of-way, at any time, all the vehicles running in it have no conflict of movement route with each other.
- The yellow region is called exit zone. The vehicle that has entered this region will report its exit of conflict zone to manager as quickly as possible.

In the following, we consider that the sampling time of the vehicle is bounded to τ_v. We assume that communication (V2I) is not perfect. That means message drop and loss are possible. Nevertheless, we assume that there exist a stable value of roundtrip time, that is RTT_{v2i}. The positioning system is assumed precise because it is a prerequisite of autonomous vehicle.

28.3.2 Obstacles

In TIM there are two kinds of factors will influence the motion of vehicle. The measurement of frontal sensor will be used to avoid collision with precedent vehicle. The information received from manager will be used to adjust the motion in order to achieve the goal of the new protocol. Corresponding to the two kinds of factors, we classify the objects that will influence the motion into Real Obstacle and Virtual Obstacles.

28.3.2.1 Virtual Obstacles

The Virtual Obstacles (VOs) are the objects that do not have direct influence on the motion of vehicle in the new protocol. For example, as listed in the 'sequence' shown in Fig. 28.2, vehicle 'leader' and 'follower-1' running on different lanes. As to 'follower-1', it does not need to immediately react to the situation of 'leader'. This characteristic plays a fundamental role in this paper. It enable us to do more works than with the real preceding vehicle. This protocol take into account two types of virtual obstacles.

The first type is the Virtual Preceding Vehicle (VPV) that listed in the sequence. Normally, virtual preceding vehicle has higher priority to pass the intersection than its follower. As it is running on other lane, the follower doesn't need to react immediately to its motions, even if the follower is closer to the conflict zone than its virtual preceding. In other words, the follower could smoothly adjust its motion as it desires before it enters the conflict zone. It just needs to make sure that when entered the conflict zone with right-of-way, it has got desired motion. This characteristic is interesting, because we could take the best advantage of it to do some thing we would like.

The other type of virtual obstacle is the conflict zone shown in Fig. 28.2, the red area. The conflict zone plays a very important role in this paper for the adjustment of motion of vehicle. It's the safe stone that the vehicle will stop before it if the vehicle has not got the right-of-way but has reached the beginning of conflict zone. It has been introduced in one of the control policies of [13]. This paper considers it as a virtual preceding obstacle that has no speed and stay at its position for ever.

28.3.2.2 Real Obstacle

The Real Obstacle is the real preceding vehicle that the traditional cruise control only takes into account. It also be listed in sequence maintained by manager. So the new policy needs to distinguish whether its virtual preceding vehicle is also the real preceding one. Its motion information will comes from the frontal equipped sensors.

Corresponding to the two kinds of obstacles, we could get two accelerations a_r and a_v respectively come from dealing with real obstacle and virtual obstacles. As the safety is the primary standard, the final acceleration a that will be taken by follower is given by:

$$a = \min(a_r,\ a_v) \tag{28.1}$$

28.3.3 Sequence Policy

There exist many policies for sequencing the incoming vehicles, for example, First Come First Served (FCFS), Distributed Clearing Policy (DCP) and Autonomous Distributed Clearing Policy (ADCP) [13] etc. As the adjustment of vehicle's motion is concerned in this paper, FCFS is chosen as sequence policy for simplicity.

28.4 The Sequence-Based Cooperative Adaptive Cruise Control Policy

During more than half century, a lot of attentions have been paid on finding some methods that could perfectly reflect driver's behaviours in transportation. Some of their works are Intelligent Driver Model (IDM), enhanced IDM (EIDM) and a special model based on IDM for TIM which named as Cooperative Intelligent Driver Model (CIDM) [14].

In this section, a new Advanced Cruise Control (ACC) policy for real preceding vehicle is firstly proposed. We name it as ExACC policy. We also use it to treat the two kinds of virtual obstacles at same time for safety. After that a simple strategy is introduced for adjusting the speed of vehicle with maximal capacity for traversing the intersection with higher speed.

For clarity, some general symbols are shown in Table 28.1.

Table 28.1 Parameters of advanced cruise control

Parameter	Meaning	Value/unit
v_0	Desired speed of vehicle	$15\ \mathrm{ms}^{-1}$
s_0	Minimum distance headway	2 m
a	Maximum acceleration	$4\ \mathrm{ms}^{-2}$
b	Minimum deceleration	$-4\ \mathrm{ms}^{-2}$
τ	Sampling time	s
s	Distance headway without s_0	m
v	Speed of vehicle	ms^{-1}

28.4.1 Advanced Cruise Control Policy for Real Obstacle

Generally, one of the most important characteristics of traditional adaptive cruise control is to react as quickly as possible to abrupt brake of real preceding vehicle for safety. After getting enough space, it will speed up with an acceptable and comfortable acceleration to achieve an equilibrium situation.

Normally, ACC make action decision for next sampling time basing on the situation of current point of time. They could not give any information about future action of leader vehicle. With this characteristic, they normally could not immediately react to extreme case, for example the leader brakes at maximal deceleration which is bigger too much than that of follower vehicle. So if some assumptions of extreme situation of leader could be introduced into the decision process, the follower may be able to react better so as to improve safety.

We name the policy that we will introduce as ExACC policy. It originally introduces a prediction of leader's motion into decision process. The prediction assumes that the leader will brake at its maximal capacity until stop from the beginning of next sampling time. Follower will take an acceleration during the next reaction step. And as reaction to leader, the follower will also take its maximal capacity, from the end of next reaction step, to try to stop and make sure that the final distance headway is greater than or is equal to s_0. In other words, follower will react to the assumption with a delay, one sampling time.

Figure 28.3 shows the strategy of ExACC policy. v_l, v_f, v_t are respectively the velocity of current point of time of leader and follower, the velocity of follower at the end of next τ_v. l the length of follower. x_t is the distance the follower will move in the next τ_v. h_l, h_f respectively the movement distance during the brake of leader and follower.

The above figure clearly shows the following relations:

$$v_t = v_f + a_r \tau \tag{28.2}$$

$$x_t = \frac{\tau}{2}(v_f + v_t) \tag{28.3}$$

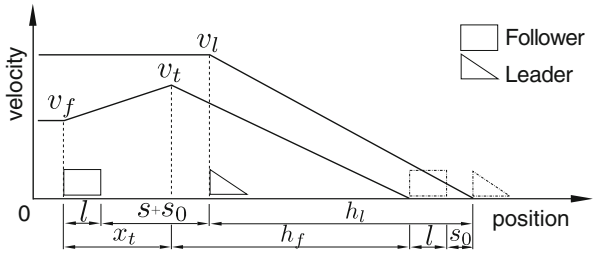


Fig. 28.3 Strategy of ExACC function

$$h_l = -\frac{v_l^2}{2b_l} \tag{28.4}$$

$$h_f = -\frac{v_t^2}{2b_f} \tag{28.5}$$

$$h = (l + s + s_0 + h_l) - (x_t + h_f + l + s_0) = 0 \tag{28.6}$$

where, b_l, b_f are respectively the maximal deceleration of leader and follower, a_r the acceleration that follower will take during the next step τ_v. We draw the reader attention to the fact that for safety reasons, τ is bigger than τ_v and RTT_{v2i}. In the following we consider that $\tau = 1,2\max(\tau_v, RTT_{v2i})$.

Then we have the acceleration, basing on Eq. (28.2), that the follower should take during the next τ_v

$$a_r = \frac{b_f\tau - 2v_f \pm 2b_f\sqrt{\frac{b_f b_l\tau^2 + 4b_l v_f\tau + 4v_l^2 - 8b_l s}{4b_f b_l}}}{2\tau} \tag{28.7}$$

Figure 28.3 shows the case that the follower is behind the leader at initial state. Actually, in the scenario that the follower is running before the leader, if we also define the headway $s + s_0$ is from rear bumper of leader to front of follower, we could also have the same result as well as function (28.7). This scenario could occurs if the leader is a virtual preceding vehicle of follower as shown in Fig. 28.2.

As we should take into account the extreme situation that the bumper to bumper distance is less than s_0, the final ExACC policy is:

$$a_r(v_f, v_l, s) = \begin{cases} b_f, & \text{for } s < 0 \\ \frac{b_f\tau - 2v_f - 2b_f a^*}{2\tau}, & \text{for } s \geq 0 \end{cases} \tag{28.8}$$

where,

$$a^*(v_f, v_l, s) = \sqrt{\frac{b_f b_l\tau^2 + 4b_l v_f\tau + 4v_l^2 - 8b_l s}{4b_f b_l}}$$

The following Fig. 28.4 shows a extreme scenario that the leader brakes frequently with maximal capacity $-15\,\mathrm{ms}^{-2}$. The acceleration capacity of follower is $[-4, +4]\,\mathrm{ms}^{-2}$.

It shows that the follower equipped with ExACC policy could react immediately to abrupt brake of leader without any collision. We take this policy for dealing with the real preceding vehicle in order to get the acceleration a_r in Eq. (28.1).

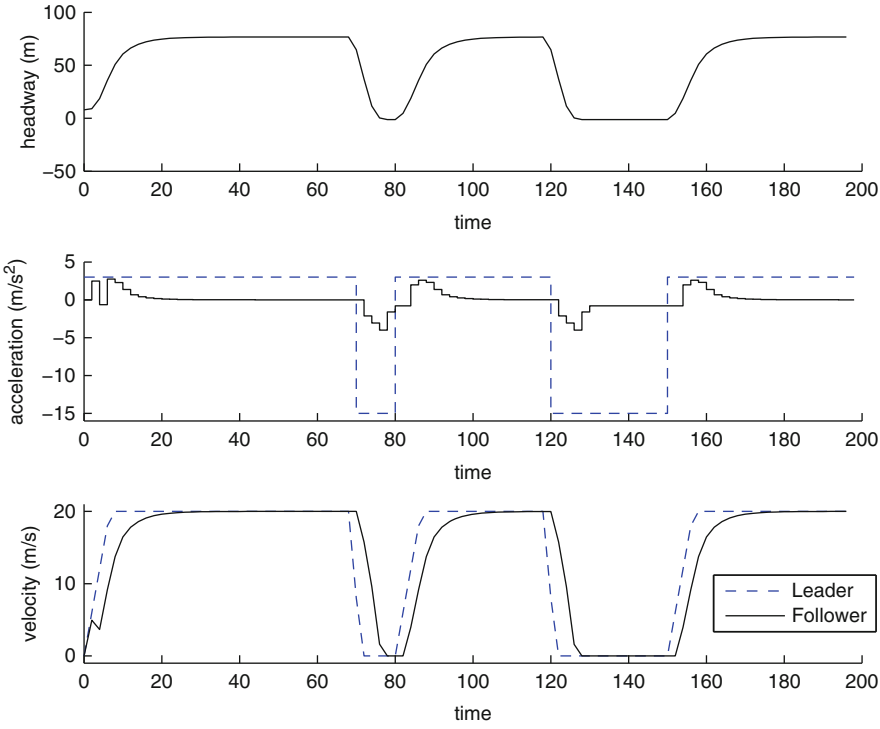


Fig. 28.4 ACC function: scenario multi brake

28.4.2 Strategy for Virtual Obstacles

As discussed in Sect. 28.3, the most important characteristic of treating virtual obstacles is that the follower only need to make sure that when passing the stop line with right-of-way it has a higher speed for traverse the conflict zone as quickly as possible. So the point of time and position that follower start to adapt its velocity are the key points of the strategy. With the capacity of reacting to extreme situation, we also apply the policy (28.8), with a adaptation, on dealing with the two types of virtual obstacles.

Table 28.2 shows the symbols that will be used in following strategy for virtual obstacles. Some of the symbols could also be found in Fig. 28.2. Same with previous policy for real preceding vehicle, the headway doesn't include s_0.

In order to deal with the two types of virtual obstacles, we established a control strategy, function (28.9). It bases on function (28.8) and considers the conflict zone as another type of virtual obstacle which has same parameters with virtual preceding

Table 28.2 Parameters of ACC function for virtual obstacles

Parameter	Meaning	Unit
d_{el}	Escape distance of leader	m
d_e	Escape distance of follower	m
s_e	Equilibrium distance headway	m
t_e	Escape time of follower	s
v_t	Target velocity of follower	ms^{-1}

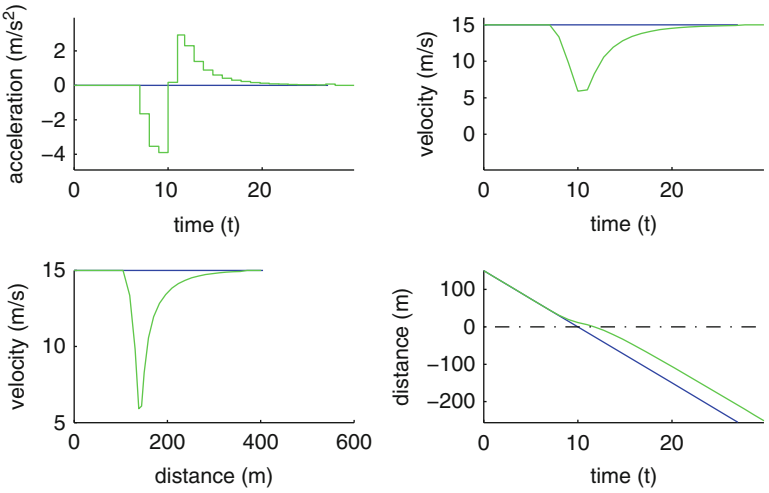


Fig. 28.5 Application of ExACC on virtual obstacles

vehicle except the velocity and position. As discussed before, this obstacle has speed $0\,\mathrm{ms}^{-1}$ and stay at its position for ever.

$$a_{v1}(v_f, v_l, s_v, s_c) = \begin{cases} a_c, & \text{for } s_v < 0 \text{ or } a_p \leq a_c \\ a_p, & \text{for other cases} \end{cases} \tag{28.9}$$

where,

$$a_c = a_r(v_f, 0, s_c)$$
$$a_p = a_r(v_f, v_l, s_v)$$

s_c, s_v respectively the distance from beginning of local conflict zone or virtual preceding vehicle to follower.

Figure 28.5 shows a normal simple scenario with two vehicles running on different lanes at an intersection. They have same initial velocity ($15\,\mathrm{ms}^{-1}$) and same distance (150 m) to the conflict zone. The virtual preceding leader runs at constant velocity while the follower is controlled by policy (28.8). The dash point line indicates the conflict zone.

In above scenario, at the beginning of adjustment, the vehicle brakes sharply. It's not a good experience. In TIM, if the follower does not obtain the right-of-way, it needs to stop before conflict zone. In addition, at some extreme cases, the follower needs to brake sharply even stop for avoiding collision with virtual preceding. One way of getting a smoother motion of vehicle and avoiding sudden velocity adjustment is to make a light deceleration when approaching the stop line.

28.4.3 Smoothing Strategy (SS)

For the sake of smoothness, we introduce a simple velocity adjustment strategy. It takes into account a specified distance that from vehicle to conflict zone. During the whole speed adjustment, it will takes a constant deceleration and a constant acceleration. The absolute value of the two acceleration is same. When arrived at the stop line, the vehicle should get the desired velocity v_t, shown in Fig. 28.6.

Where, v_m, t_1 and t_2 are respectively the minimal speed during the adjustment, the brake time and the speeding up time. Then we have

$$\begin{aligned} v_m &= v_f - at_1^* \\ v_t &= v_m + at_2^* \\ t_e &= t_1^* + t_2^* \\ (v_f + v_m)t_1^* + (v_m + v_t)t_2^* &= 2d_e \end{aligned} \tag{28.10}$$

For simplicity, we assume that the virtual leader will run at current velocity during its escape journey d_{el} as shown as Fig. 28.2. So we could have the escape time of leader $t_{el} = \frac{d_{el}}{v_l}$. That's also the escape time t_e that follower will cost when it arrives at conflict zone.

Because that normally the follower will not stop and has a sampling time τ_v, the actual brake time should be divisible by τ_v. With same reason the acceleration time should also be divisible by τ_v, if the target velocity is not the maximal speed of follower. Then we finally have adaptation relation:

$$t_1 = \lceil \frac{t_1^*}{\tau_v} \rceil \tau_v, \; v_m = v_f - bt_1$$

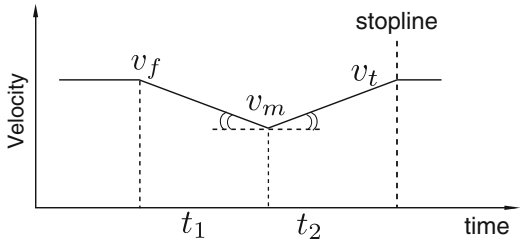


Fig. 28.6 Smoothing strategy

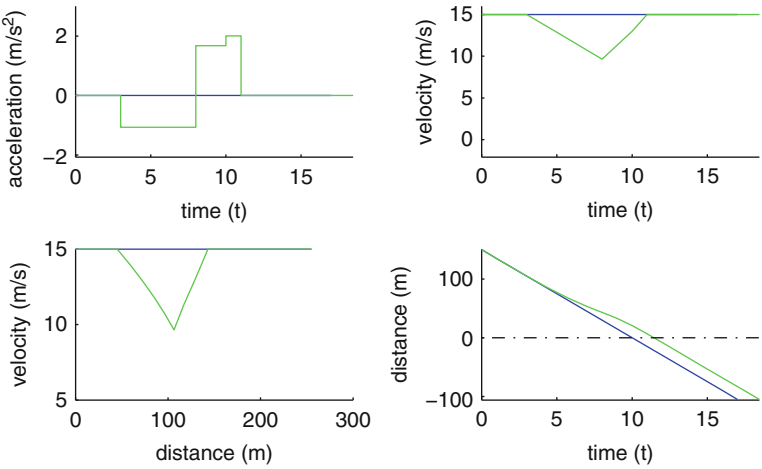


Fig. 28.7 Simple scenario of smoothing strategy

and

$$t_2 = \lceil \frac{t_e - t_1}{\tau_v} \rceil \tau_v, \; a = \frac{v_l - v_m}{t_2}$$

Then it is easy to get the acceleration a_{v2} of vehicle with smoothing strategy. Figure 28.7 shows the simulation in same scenario as Fig. 28.5.

Combining function (28.9) and Smoothing Strategy we get the final strategy for dealing with virtual obstacles:

$$a_v = \min(a_{v1}, \; a_{v2}) \tag{28.11}$$

28.4.4 Simulation

We make simulation under MATLAB to test the new policy (28.1) at the intersection Fig. 28.2.

Initial escape distance of vehicles from beginning of local conflict zone is 150 m. Initial speed is 15 ms^{-1}. The sampling time of vehicle is 1 s. Every two vehicles are generated at same time but are located at two different lanes. The delay between two successive generations of vehicles is 2 s. The sequence of vehicles is shown as 'sequence' in Fig. 28.2.

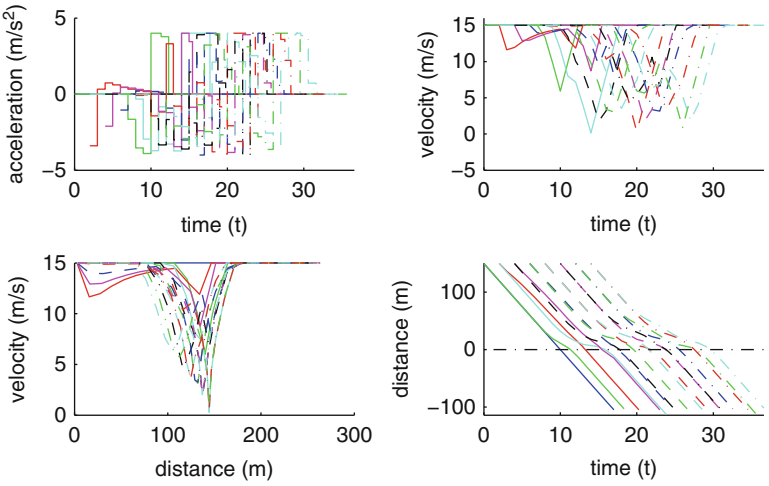


Fig. 28.8 Combination of function (28.7) and (28.9)

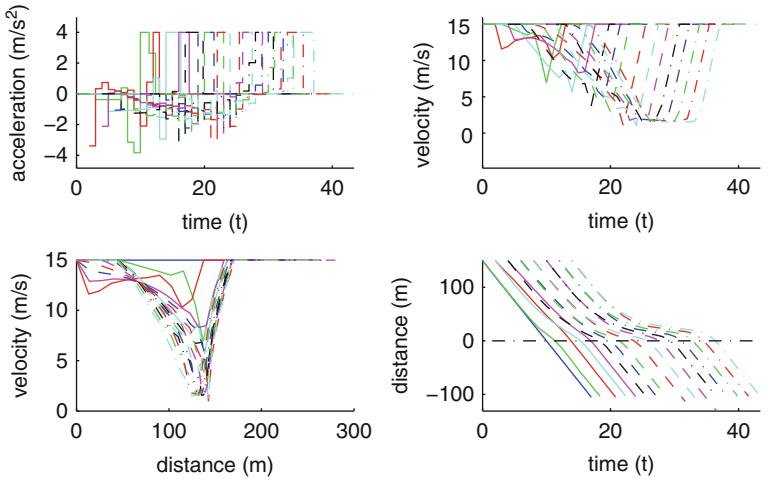


Fig. 28.9 Combination of the three strategies

Figure 28.8 shows that the new approach has successfully synchronized the vehicles' speed before the vehicles enter the conflict zone. And all the vehicles reach their minimal speed during the velocity adjustment. This could help to improve the usage of lane. But the vehicles takes abrupt brake at the beginning of adjustment. After combining smoothing strategy, Fig. 28.9 shows that the vehicles make a smoother adjustment and get a higher speed when they enter the conflict zone. It is helpful for the efficiency of intersection.

28.5 Simulation

From previous simulation, we could see that the approach has successfully synchronized the speeds of vehicle before they enter the conflict region. Then the vehicles could be able to enter the conflict zone of intersection under the order that specified by the intersection manager. From the following simulation, we could also see this characteristic, even if, before the vehicles enter conflict zone, the vehicles' order is not same with the manager's sequence.

Actually, in real world, there is no transportation system like Fig. 28.2. Obviously, it's not enough that applying the approach on only one intersection which belongs to an open environment to show its characteristics.

In order to show the effect of synchronizing speed of same vehicles at same intersection, we have designed a special closed transportation system, Fig. 28.10.

The transportation system is comprised of one intersection and two lanes: lane 1 and lane 2. The red zone is the conflict zone of intersection. The intersection connects the output of one lane to the other's input. That means after the vehicle has passed the intersection twice, it will renter the lane that it had moved on. With this feature, we could be able to observe the improvement of traffic flow when applying the approach on same vehicles which are running on a constant environment. With this we could be able to get its fundamental characteristic.

The basic ACC parameters of the vehicles involved in the following simulations is shown in Table 28.1. The sample interval of vehicle is 1 s. The initial speed of every vehicle is its desired 15 m/s. The length of the two lanes is same, 150 m. The start position of vehicle is the beginning of lane as shown in Fig. 28.10.

Every time, we append two new vehicles into the system. The new vehicles will be located at the two lanes respectively and will be ordered with the rule that the vehicle on lane 1 is located in the sequence before the other one. It's obvious that the two new vehicles will arrive at the intersection at same time, if there is no speed adaptation. The interval time between two successive appending of vehicles is 2 s. This indicates that the initial distance between two vehicles that running on same lane is 30 m.

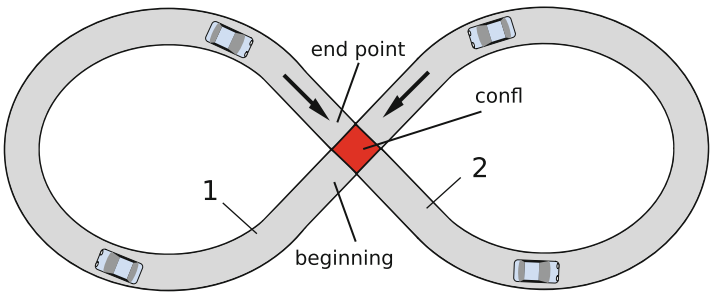


Fig. 28.10 Closed transportation environment

Because the vehicle will repeatedly pass the same intersection until the simulation stops, the movement that consider the start position of vehicle as origin is meaningless. As mentioned before, after one vehicle has passed the intersection twice, it will arrive at the position that it was appended into the system the first time. We regard the interval between the two successive times that the vehicle arrive at its beginning position as a cycle. So the vehicle's journey within one cycle is 300 m. Within one loop, we consider the position of intersection that it will pass at first time as origin. In other words, the middle point of voyage of vehicle in one cycle is the position 0. Before the origin, the distance between vehicle and intersection is positive. Correspondingly, after the vehicle has passed the intersection, the distance is negative. So at the beginning of one loop, the position of vehicle is 150 m.

In the above Fig. 28.11, when the first two vehicles approaching the intersection, because there is no preceding obstacles in its detecting distance, the first one runs at desired velocity. The second considers the first one as a virtual preceding obstacle, so it brakes sharply under the control of function (28.1) in order to get a safe distance as shown in red region 2 in Fig. 28.11. After the first one has passed the conflict zone, the second one loses its preceding obstacle and speeds up with maximal capacity

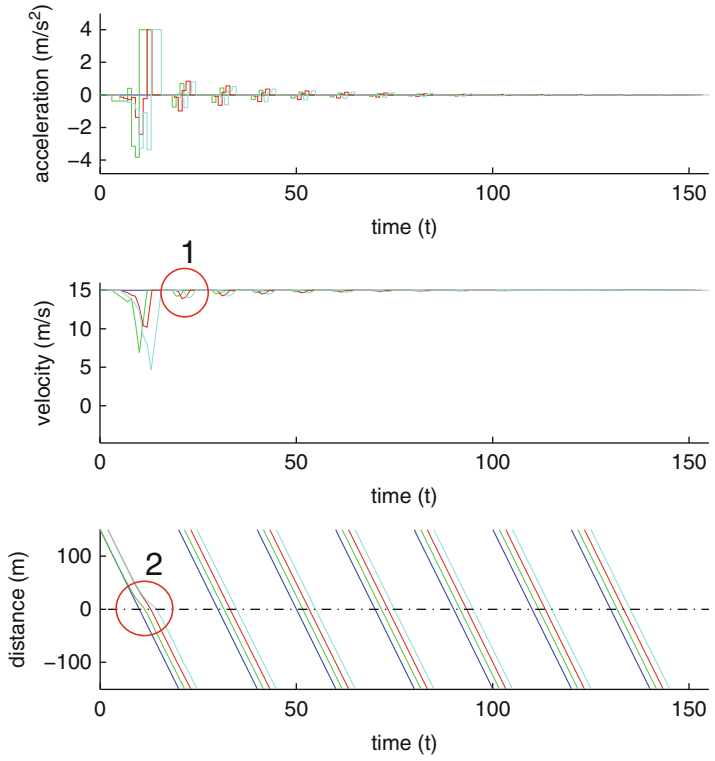


Fig. 28.11 Scenario with four vehicles and seven cycles

until its desired velocity so as to pass the conflict zone as quickly as possible. At the end of its acceleration, the distance between the two vehicles, corresponding to the position of intersection that they are approaching again, may be shorter than the minimal safe headway. So when the two vehicles approaching the intersection, the second one needs to adapt its motion again for safe reason, shown in red region 1 in Fig. 28.11. But the adaptation process is shorter than the last time. The following vehicles will consequently take similar actions. After six loops, the traffic flow achieves a state that every vehicle runs at desired velocity.

Figure 28.11 also shows that the vehicle adapt its speed only when it is approaching the intersection. This could help to increase the capacity of lane.

Figure 28.12 shows that with the increase of number of vehicles the damping process will correspondingly raise. After almost 16 cycles, the traffic flow achieves the equilibrium too.

Figure 28.13 shows that with the increase of density of vehicles the traffic flow will could only achieve an dynamic stable situation, shown in Fig. 28.14b, where constant brake waves spread in the flow cycle. But if the sample interval of vehicle decreases, the traffic flow could achieve an equilibrium too as shown in Fig. 28.15.

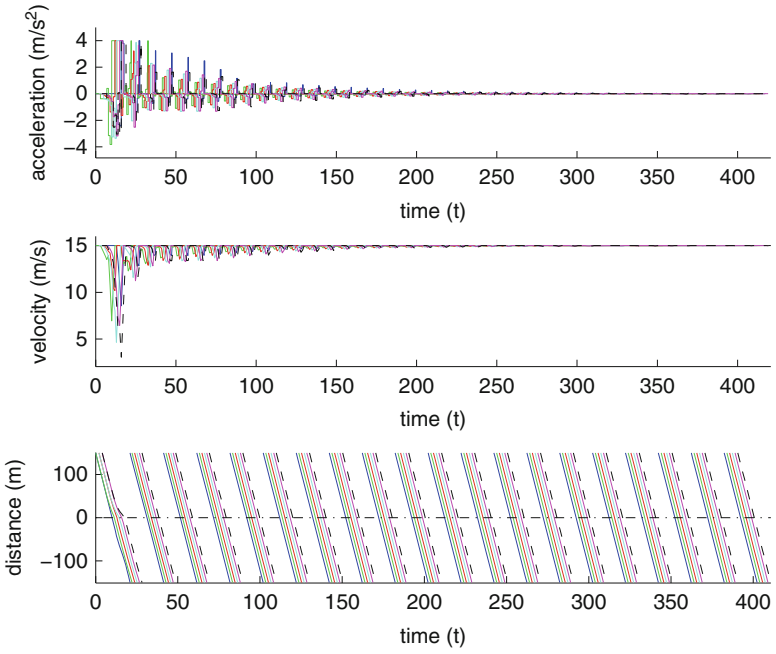


Fig. 28.12 Scenario: six vehicles

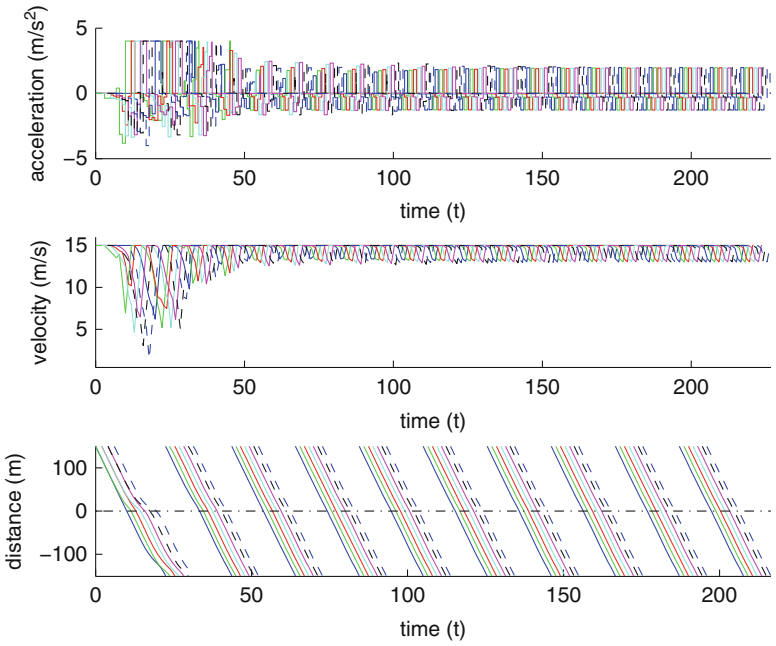


Fig. 28.13 Scenario: seven vehicles, sample interval 1 s

In other side, with the increase of vehicles, when the first vehicle approach the intersection, it will also have preceding obstacles. For example in Fig. 28.13, the 7th vehicle is its real preceding vehicle. In this case, it brakes sharply even if this will cause its distance to intersection is larger than that of its virtual preceding one, shown in red region of Fig. 28.14a. The blue continue line, green continue line and the blue dash one represent respectively the distance from the first vehicle, the second vehicle (virtual preceding vehicle of the first one) and the 7th vehicle to the intersection.

28.6 Conclusion and Future Work

In this paper we have introduced a new approach for controlling the autonomous vehicles that are waiting traverse intersection. The first advantage of this approach is synchronizing the speed of vehicle before it enters the conflict zone soon. This method also helps to improve the efficiency of intersection with a higher speed when

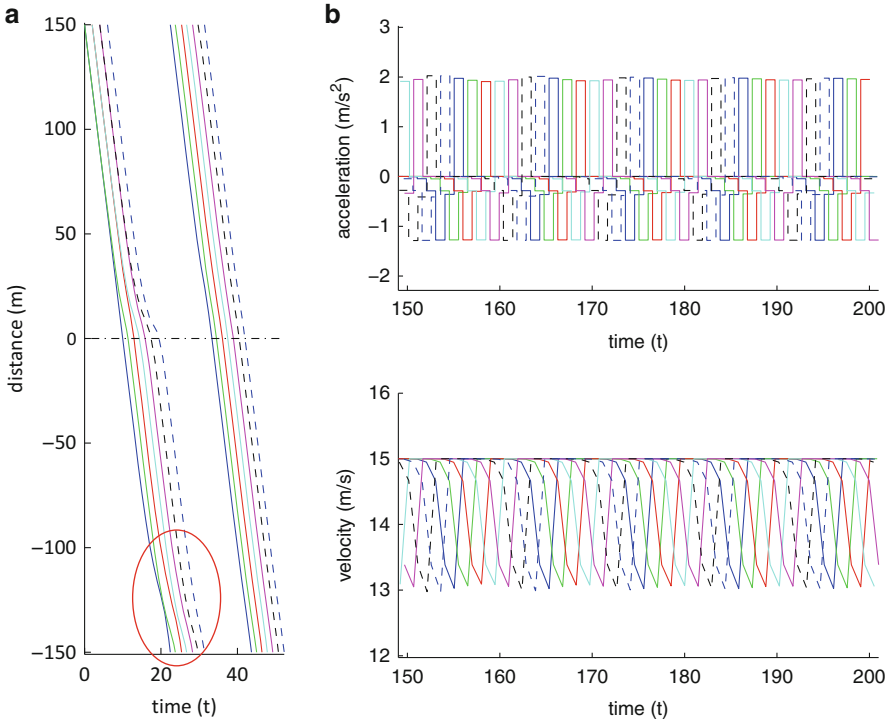


Fig. 28.14 Part of scenario: seven vehicles

they traverse the conflict zone. In addition, because the vehicle adjust its speed when they approach the conflict zone, this strategy could help to improve the use of lane, especially in the case that the vehicle needs to stop during its speed adjustment. With the help of smoothing strategy, the adjustment could be done smoothly. This will be able to have a better ride experience.

In future works, the new approach needs to be tested in the transportation system that compose of more than one intersection. Further more, as the approach could only deal with the case that the order of vehicle is specified by intersection manager, the deadlock may occurs easily. So the more complex strategy which could synchronize the speed of vehicles effectively meanwhile could avoid deadlock of intersection is more interesting.

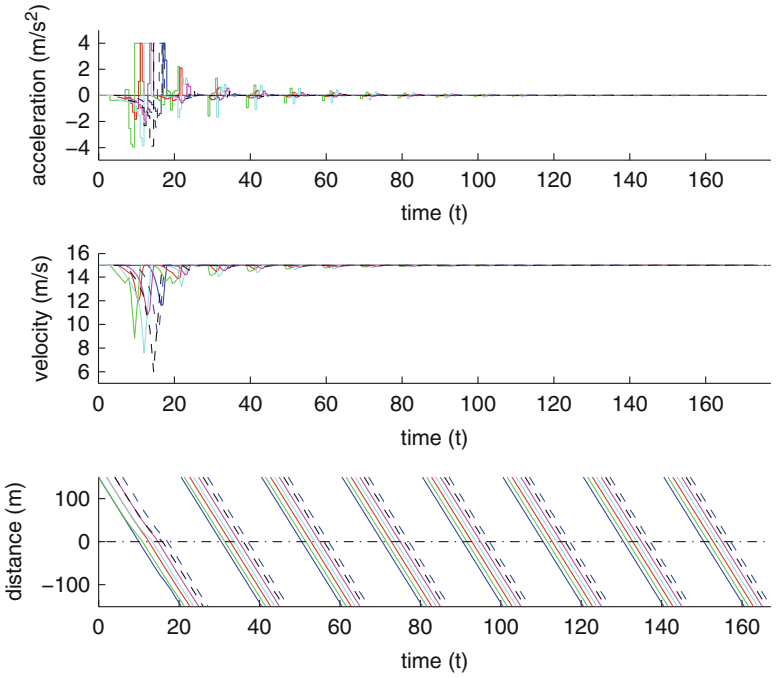


Fig. 28.15 Scenario: seven vehicles, sample interval 0.5 s

References

1. Tan, S.T., Noraini, N.H., Tan, K.L.: Improvement of conventional traffic signaling device-self routing traffic light system. WSEAS Trans. Syst. Control **9**, 309-317 (2014)
2. Subramaniam, S.K., Esro, M., Aw, F.L.: Self-algorithm traffic light controllers for heavily congested urban route. WSEAS Trans. Circuits Syst. **4**(11), 115–124 (2012)
3. Ben Jemaa, I., Shagdar, O., Muhlethaler, P., De La Fortelle, A.: Analysing impact of mobility dynamics on multicast routing in vehicular networks. In: EMERGING 2013: The Fifth International Conference on Emerging Network Intelligence, Porto, September 2013. IARIA (2013)
4. Jemaa, I.B., Shagdar, O., Muhlethaler, P., De La Fortelle, A.: Analysing impact of mobility dynamics on multicast routing in vehicular networks. In: JNCT 2013 - Journées Nationales des Communications dans les Transports, Nevers, May 2013. IEEE (2013)
5. Souza-De-Assis, ítalo, A., Oliveira, R., Fernandes, M.A.: Speed fuzzy control applied to autonomous electric vehicles. WSEAS Trans. Syst. Control **9**, 640–651 (2014)
6. Dresner, K., Stone, P.: Multiagent traffic management: A reservation-based intersection control mechanism. In: The Third International Joint Conference on Autonomous Agents and Multiagent Systems, July 2004, pp. 530–537
7. Quinlan, M., Tsz-Chiu, A., Zhu, J., Stiurca, N., Stone, P.: Bringing simulation to life: A mixed reality autonomous intersection. In: IEEE/RSJ International Conference on IROS (2010)
8. Fok, C.-L., Hanna, M., Gee, S., Au, T.-C., Stone, P., Julien, C., Vishwanath, S.: A platform for evaluating autonomous intersection management policies. In: ACM/IEEE International Conference on ICCPS (2012)

9. Carlino, D., Boyles, S.D., Stone, P.: Auction-based autonomous intersection management. In: Proceedings of the 16th IEEE Intelligent Transportation Systems Conference (ITSC), October 2013
10. Zohdy, I.H., Rakha, H.: Game theory algorithm for intersection-based cooperative adaptive cruise control (CACC) systems. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), September 2012, pp. 1097–1102
11. Zohdy, I.H., Kamalanathsharma, R.K., Rakha, H.: Intersection management for autonomous vehicles using iCACC. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), September 2012, pp. 1109–1114
12. Perronnet, F., Abbas-Turki, A., Buisson, J., El Moudni, A., Zeo, R., Ahmane, M.: Cooperative intersection management: Real implementation and feasibility study of a sequence based p rotocol for urban applications. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), September 2012, pp. 42–47
13. Perronnet, F., Abbas-Turki, A., El-Moudni, A., Buisson, J., Zéo, R.: Cooperative vehicle-actuator system: A sequence-based optimal solution algorithm as tool for evaluating policies. In: 2013 International Conference on Advanced Logistics and Transport (ICALT), May 2013, pp. 19–24
14. Perronnet, F., Abbas-Turki, A., El Moudni, A.: A sequenced-based protocol to manage autonomous vehicles at isolated intersections. In: 2013 16th International IEEE Conference on Intelligent Transportation Systems (ITSC), October 2013, pp. 1811–1816
15. Perronnet, F., Abbas-Turki, A., El Moudni, A.: Vehicle routing through deadlock-free policy for cooperative traffic control in a network of intersections: Reservation and congestion. In: 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), Qingdao, October 2014. IEEE, New York (2014)

Chapter 29
Innovation for Failure Detection and Correction in Safety-Related Systems Which Based on a New Estimator

Ossmane Krini, Jamal Krini, Abderrahim Krini, and Josef Börcsök

Abstract This scientific work presents a new method allowing to make a realistic prediction about reliability of safety related systems. The main feature of this method enables the prediction of an estimate of the remaining critical number of faults in systems. Stochastic play a very important role in safety technology. With the help of it, safety systems may be released reliably after an assessment. With the help of the probability theory meaningful statements are achieved and based on them, realistic forecasts may be given. However, in order that reliable forecasts can be conducted, new approaches in thinking need to be developed. The algorithm can provide an even more reliable prognosis than the conventional methods. Furthermore, the new method describes two processes for critical failures (detection and correction process). This contribution serves to give a short synopsis about the actual problem of the probabilistic safety technology on the base of stochastic. In that, the test methods, however, plays the most important role as the test results are source vectors for probabilistic models. However, this article tries to describe a suitable, innovative method that will correctly estimate the safety parameters.

Keywords Stochastic • Mathematical models • Safety • Probability • Reliability • Failure/Error • Matrix-calculations

O. Krini (✉) • J. Krini • A. Krini • J. Börcsök
Computer Architecture and System Programming, University of Kassel, Willhelmshöher Allee 71, Kassel 34131, Germany
e-mail: o.krini@uni-kassel.de; a.krini@uni-kassel.de; j.krini@uni-kassel.de; j.boercsoek@uni-kassel.de

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*, Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8_29

29.1 Introduction

To a greater extent than previously, safety related systems have been developed, produced and released to the market. For this reason it is essential, to know and correctly and reasonably apply the current international norms for functional safety as a basis for systems that are used in safety-critical applications.

The functional safety is part of the overall safety in terms of the EUC standards and the EUC control system. It is subjected to the correct function of the E/E/PRE safety-related systems, safety-related systems of other technologies and external devices for risk reduction. In this process it is unimportant whether it refers to a control system or the complete installation. Concerning the safety of a system, the default rate plays an important role. It describes the amount of default per unit of time and has the unit "FIT". On principle, when examining errors, it can be differentiated between safe (λS) and dangerous errors (λD). Safe errors, whether they have been found or remain unfound, normally have no influence on the safety-function of a system. However, concerning dangerous mistakes, this is not true. If such errors occur, the system will be transferred into a dangerous state, which under certain circumstances may lead to the massive endangerment of human lives. These errors too are differentiated in dangerous and traceable (λDD) or dangerous and non traceable (λDU) errors [1].

Concerning dangerous and traceable errors, if accordingly designed, the safety system may bring the overall system or the installation in a safe state. The critical state, however, is given through the non traceable, dangerous errors. If such errors occur in the safety system, there is no possibility to detect it. In the system they may lead to its switch off or, in the worst case, to its dangerous breakdown. In order to be able to run systems or installations that can be applied in safety related areas, comprehensive measures for development and certification are necessary. These serve to prevent these described dangerous situations from happening and to bring the safety system or the installation into a safe state.

On the base of the default rates the reliability functions and the default probabilities are determined. The distribution of cumulative frequencies plays a central role in this. The challenge is to choose the right density function. Afterwards the model parameters and the safety parameters need to be estimated.

The following chapter will show the new mathematic approach of how an estimator can be constructed in a structured way.

29.2 Safety Technology Based on Probabilistic Approach

According to the norm, the functions of all safety related systems form the functional safety of the overall system. Determining a level of safety integrity (SIL) forms the central element. The SIL is one of four discrete steps towards specification of the requirement for safety integrity of the safety functions related to the E/E/PE

Table 29.1 SIL at low and high demand rate according to IEC

SIL	Operation with low demand rate PFD_{avg}	Operation with high demand rate PFH [1/h]
4	$10^{-5} \leq PFD_{avg} < 10^{-4}$	$10^{-9} \leq PFH < 10^{-8}$
3	$10^{-4} \leq PFD_{avg} < 10^{-3}$	$10^{-8} \leq PFH < 10^{-7}$
2	$10^{-3} \leq PFD_{avg} < 10^{-2}$	$10^{-7} \leq PFH < 10^{-6}$
1	$10^{-2} \leq PFD_{avg} < 10^{-1}$	$10^{-6} \leq PFH < 10^{-5}$

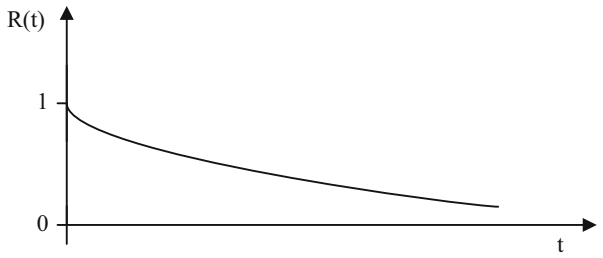


Fig. 29.1 Reliability

safety related system, with level 4 being the highest level of safety integrity, level 1 the lowest. Therefore the IEC-standard 61,508 consists of four safety levels SIL 1 through 4.

Each of these appears in a confidence interval, Table 29.1 showing the distribution of the probability.

29.2.1 Effective Distribution in Safety Theory

The reliability R(t) is the probability that a unit is functional in one view period (0, t). Figure 29.1 shows R (t) as function of time [2, 3].

The probability that the operational time T is within the considered time interval (0 . . . t) is for small t almost equal to one. For larger values of t the probability decreases more and more.

$$R(t) = e^{-\int_0^\infty \lambda(t)dt} \tag{29.1}$$

The exponential distribution is useful in many applications in engineering, for example, to describe the lifetime X of a transistor. The most known and most favorite probability model for the reliability analysis of safety systems is the exponential distribution. With this distribution it is possible to represent the time dependent probability $F(t)$ of components for which it is necessary to obtain observed data to determine X.

The failure probability is defined by the exponential distribution as

$$F(t) = 1 - e^{-\lambda \cdot t} \tag{29.2}$$

where λ is the failure rate. Respectively with failure density

$$f(t) = \begin{cases} \lambda \cdot e^{-\lambda \cdot t} & for \quad t \geq 0 \\ 0 & otherwise \end{cases} \tag{29.3}$$

If an exponential distribution for the reliability is valid, then the failure rate is constant:

$$\lambda(t) = \lambda \tag{29.4}$$

Then the equation can be rewritten as:

$$R(t) = e^{-\lambda \cdot t} \tag{29.5}$$

An important reliability parameter is the MTTF value (Mean Time To Failure).

$$MTTF = \int_0^\infty R(t)dt = \frac{1}{\lambda} \tag{29.6}$$

If an exponential distribution is suitable equation [4] can be rewritten as:

$$MTTF = \frac{1}{\lambda} \tag{29.7}$$

Within the interval (0, t] the probability of failure P(t) is calculated applying the reliability function R(t).

$$\begin{aligned} P(t) &= 1 - R(t) \\ P(t) &= 1 - e^{-\lambda \cdot t} \\ P(t) &\approx \lambda \cdot t \quad for \quad \lambda \cdot t << 1 \end{aligned} \tag{29.8}$$

Generally, the time t is applied by T1. The time from point in time zero to time T1 is characterized as proof test interval. At time T1 a periodical test or the maintenance of a safety system is taking place. Tests are carried out to allocate undetected, dangerous failures. After a proof test, the system is regarded as new. The calculated PFD-valued depends on the value T1, [2, 3, 5]. In order to be able to make probabilistic statements about possible values of safety parameters, according to the architecture [6, 7].

Different models for analysis can be drawn. In the following these will be introduced.

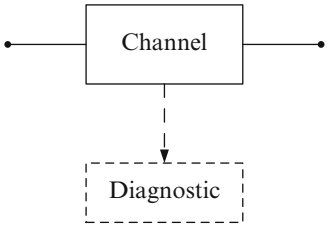


Fig. 29.2 Channel with diagnostic

29.2.2 One Out of One Architecture (1oo1)

The 1oo1 architecture is the simplest safety system around and consists of only one channel. Every dangerous fault can lead towards a failure of the safety function [2, 5, 8]. The 1oo1 architecture is presented in Fig. 29.2.

If $\lambda = \lambda_D$ is applied to equation [4] then the result in the following equation is for the 1oo1 system:

$$P(t) = 1 - e^{-\lambda_D \cdot t} \tag{29.9}$$

P (t) is developed by the MacLaurin series. For the 1oo1 system the first three terms are needed to be developed. The first three terms plus the remaining term R3 are sufficient for the calculation of the PFDavg values.

$$e^{-\lambda_D \cdot T} = 1 - \lambda_D \cdot T + \frac{{\lambda_D}^2 \cdot T^2}{2!} + R_3. \tag{29.10}$$

The description of the remaining term R_3 is chosen as follows:

R3 is the remaining term to the third order, which belongs to the exponential function with failure rate λ_D. The remaining term R_3 converges for T = 0 to the value 0 and can be neglected compared to the third term when developed towards the limit value at T = 0 [2]. Equation 29.11 is applied for a 1oo1 system. The PFD_{avg} is:

$$\begin{aligned} PFD_{avg} &= 1 + \tfrac{1}{\lambda_D \cdot T}\left[1 - \lambda_D \cdot T + \tfrac{{\lambda_D}^2 \cdot T^2}{2!} - 1\right] \\ &= \frac{\lambda_D \cdot T}{2} \end{aligned} \tag{29.11}$$

with,

$$\frac{T}{2} = t_{CE} = \frac{\lambda_{DU}}{\lambda_D}\left(\frac{T_1}{2} + MTTR\right) + \frac{\lambda_{DU}}{\lambda_D} \cdot MTTR \tag{29.12}$$

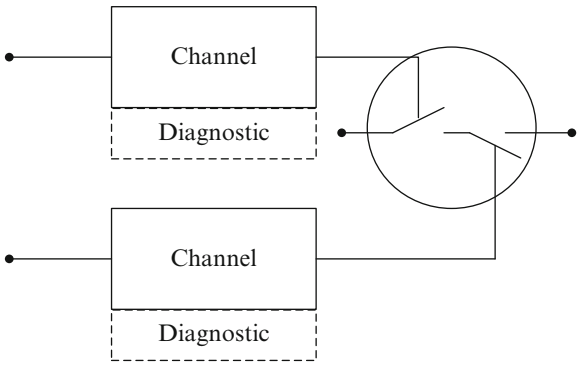


Fig. 29.3 Two Channel Architecture

Here, t_{CE} is the mean repair time of a channel. The Equation can be presented simplified as follows:

$$\begin{aligned} PFD_{avg,1oo1} &= \lambda_{DU}\left(\frac{T_1}{2} + MTTR\right) + \lambda_{DD} \cdot MTTR \\ &= \lambda_D \cdot t_{CE} \\ PFH_{1oo1} &= \lambda_{DU} \end{aligned} \tag{29.13}$$

29.2.3 One Out of Two Architecture (1oo2)

The 1oo2 architecture, see Fig. 29.3, possesses two channels in parallel, where each channel can execute the safety function by itself.

$$\begin{aligned} PFD_{avg\,1oo2} = 2[(1-\beta_D)\,\lambda_{DD} + (1-\beta)\,\lambda_{DU}]^2 t_{CE} t_{GE} \\ +\beta_D \lambda_{DD} MTTR + \beta \lambda_{DU}\left(\tfrac{T_1}{2} + MTTR\right) \end{aligned} \tag{29.14}$$

with:

$$t_{CE} = \frac{\lambda_{DU}}{\lambda_D}\left(\frac{T_1}{2} + MTTR\right) + \frac{\lambda_{DD}}{\lambda_D} MTTR \tag{29.15}$$

and

$$t_{GE} = \frac{\lambda_{DU}}{\lambda_D}\left(\frac{T_1}{3} + MTTR\right) + \frac{\lambda_{DD}}{\lambda_D} MTTR \tag{29.16}$$

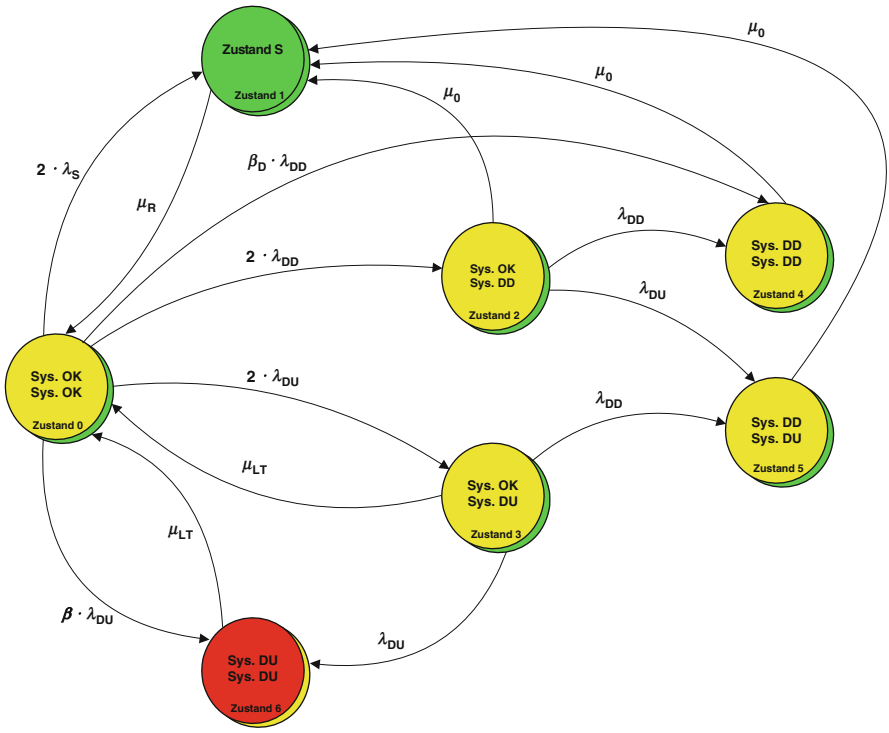


Fig. 29.4 Markov Chain for the One-out-of-two systems 1oo2

And the PFH-Value is determined by:

$$PFH_{avg\,1oo2} = 2[(1-\beta_D)\,\lambda_{DD} + (1-\beta)\,\lambda_{DU}]^2 t_{CE} + \beta_D \lambda_{DD} + \beta \lambda_{DU} \tag{29.17}$$

The average time MTTF can be the time estimated between the occurrences of two errors. For this it can be very helpful to develop a Markov-model. Figure 29.4 shows a possible approach for the One-out-of-two systems 1oo2.

The Markov-Model for a 1oo2 "Single-Board-System" is shown in Fig. 29.4. In the condition 0 both controllers are working error-free. Condition 1 represents the safe condition in which a systems fades after a safe error. The system stands in a condition with no energy. In the condition number 2 one of two channels works incorrect.

The occurred error is dangerous, but is not detected through error diagnostics. Condition 4 is characterized by two dangerous traceable errors, with one of each of them being in one of the two channels. In condition 5, however, there is a dangerous

traceable error in one channel, while at the same time there is a dangerous not traceable error occurring in the other channel. In condition 3 one the two channels operated incorrectly.

The occurring errors is dangerous and is not detected in the error analysis. In condition 3, when the error occurs in up until them the error-free channel, there is a fade of the system into the condition 5 or 6. If, however, there is no further error within the whole life span of the system in condition 3, the system may get back to the condition 0, where it is error-free.

This practically means: After this the whole system will be exchanged. If common-cause errors occur in 1oo2 systems, the following two cases are to be distinguished:

1. The joint error source leads to dangerous traceable errors. Then a fade occurs directly from system 0 to condition 4. The transmission rate is $\beta_D \cdot \lambda_{DD}$.
2. The joint error source leads to dangerous traceable errors. Then a fade occurs directly from system 0 to condition 6. The transmission rate is $\beta \cdot \lambda_{DU}$.

In the conditions 0, 2 and 3 the system is running. This must be taken into account when calculating the MTTF of the 1oo2 system.

The probability matrix P for the 1oo2 approach is:

$$P_{1oo2} = \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix} \tag{29.18}$$

where P_i

$$P_1 = \begin{bmatrix} 1 - A_1 \cdot dt & 2 \cdot \lambda_S \cdot dt & 2 \cdot \lambda_{DD} \cdot dt & 2 \cdot \lambda_{DU} \cdot dt \\ \mu_R \cdot dt & 1 - \mu_R \cdot dt & 0 & 0 \\ 0 & \mu_0 \cdot dt & 1 - A_2 \cdot dt & 0 \end{bmatrix} \tag{29.19}$$

$$P_2 = \begin{bmatrix} \beta_D \cdot \lambda_{DD} \cdot dt & 0 & \beta \cdot \lambda_{DU} \cdot dt \\ 0 & 0 & 0 \\ \lambda_{DD} \cdot dt & \lambda_{DU} \cdot dt & 0 \end{bmatrix} \tag{29.20}$$

$$P_3 = \begin{bmatrix} \mu_{LT} \cdot dt & 0 & 0 & 1 - A_3 \cdot dt \\ 0 & \mu_0 \cdot dt & 0 & 0 \\ 0 & \mu_0 \cdot dt & 0 & 0 \\ \mu_{LT} \cdot dt & 0 & 0 & 0 \end{bmatrix} \tag{29.21}$$

$$P_4 = \begin{bmatrix} 0 & \lambda_{DD} \cdot dt & \lambda_{DU} \cdot dt \\ 1 - \mu_0 \cdot dt & 0 & 0 \\ 0 & 1 - \mu_0 \cdot dt & 0 \\ 0 & 0 & 1 - \mu_{LT} \cdot dt \end{bmatrix} \tag{29.22}$$

From the probability matrix p the Q-matrix is determined. To form the Q-matrix from the P-Matrix one has to mind some criteria. The systems needs to running and the conditions may not be absorbing.

Furthermore it should be ensured that there is no secure condition or conditions showing dangerous untraceable errors. The absorbing conditions means the condition, where there is no further fade except the fade into a secure condition and an error-free condition. After the Q matrix is formed, the M-Matrix is needed for further estimations. In order to calculate the M-Matrix, the Q-matrix needs to be subtracted from the I-Matrix (unit matrix).

$$I_n = \begin{bmatrix} 1\,0\,0 \cdots 0 \\ 0\,1\,0 \cdots 0 \\ 0\,0\,1 \cdots 0 \\ \vdots\,\vdots\,\vdots\,\ddots\,\vdots \\ 0\,0\,0 \cdots 1 \end{bmatrix} \tag{29.23}$$

$$M = I - Q \tag{29.24}$$

$$M_{1oo2} = \begin{bmatrix} A_1 \cdot dt & -2 \cdot \lambda_{DD} \cdot dt & -2 \cdot \lambda_{DU} \cdot dt \\ 0 & A_2 \cdot dt & 0 \\ \mu_{LT} \cdot dt & 0 & A_3 \cdot dt \end{bmatrix}_{\tau_{LT}=\infty} \tag{29.25}$$

if $_{\tau_{LT}=\infty}$, then

$$M_{1oo2} = \begin{bmatrix} A_1 \cdot dt & -2 \cdot \lambda_{DD} \cdot dt & -2 \cdot \lambda_{DU} \cdot dt \\ 0 & A_2 \cdot dt & 0 \\ 0 & 0 & A_3 \cdot dt \end{bmatrix} \tag{29.26}$$

In order to calculate the MTTF-value, the elements of the first line need to be added to the N-Matrix. The N-Matrix is determined by the inverse of the M-Matrix.

$$N_{1oo2} = M_{1oo2}^{-1} \tag{29.27}$$

$$N_{1oo2} = \begin{bmatrix} \frac{1}{A_1} & \frac{2 \cdot \lambda_{DD}}{A_1 \cdot A_2} & \frac{2 \cdot \lambda_{DU}}{A_1 \cdot A_3} \\ 0 & \frac{1}{A_2} & 0 \\ 0 & 0 & \frac{1}{A_3} \end{bmatrix} \tag{29.28}$$

The MTTF-value can be determined by adding the elements from the first line to the N-Matrix.

$$MTTF_{1oo2} = \frac{1}{A_1} + \frac{2 \cdot \lambda_{DD}}{A_1 \cdot A_2} + \frac{2 \cdot \lambda_{DU}}{A_1 \cdot A_3} \tag{29.29}$$

Now the safety parameter MTTF can be estimated. With this value, now the reliability and default probability can be determined. The model parameters are of highest importance. In safety technology, certain distribution functions form the basis for the estimation.

In the following chapter an approach will be shown how to determine the model parameters using of estimation-algorithms.

29.2.4 Approach for the Construction of Estimators for the Safety Theory

The basis for estimating an unknown parameter ϖ is the assumption that a random variable X belongs to a certain parametric family (f.e. exponential distributed, normally distributed, poisson distribution, ...). With the help of this model it is tried to determine this parameter for which the results are the most probable. This is done with events $X_1, \ldots, X_n$ that have already taken place (sample $x_1, \ldots, x_n$ with the values n from the scope n, variables are independent and identically distributed).

Hereby, the idea of the approach is the mapping of the mathematics on the safety technology. This approach is based on the necessary distribution function which is normally applied in safety technology. Only then a realistic statement about the probability functions of the reliability and density can be made.

The expected value $E_H(t)$ of a hardware component has been reached with the following formula:

$$E_H(t) = \lambda_0 \cdot \left(1 - e^{-\int_0^t \lambda(\zeta)\, d\zeta} \right) \Bigg|_{\lambda(\zeta)=\lambda} \tag{29.30}$$

where λ_0 is the maximum failure rate. This approach implies that from this point of time $t = 0$ (start of the reliability analysis), a constant failure rate λ_0 exists in the affected hardware system. As a Weibull-distribution has been deemed, the following applies:

$$E_H(t) = \lambda_0 \cdot \left[1 - e^{(-\lambda \cdot t)}\right] \tag{29.31}$$

If the hypothesis is true that the probability of default of a safety related system is exponentially distributed, the density function will be the following:

$$f(t) = \partial F(t)/\partial t = {\partial \left(1 - e^{\lambda \cdot t}\right)}\Big/{\partial t} = \lambda \cdot e^{-\lambda \cdot t} \tag{29.32}$$

With the help of the Eq. (29.32), the time sequence of the failure rate $\lambda_H(t)$ can be determined

$$\lambda_H(t) = \frac{\partial E_H(t)}{\partial t} = \lambda_0 \cdot \lambda \cdot e^{(-\lambda \cdot t)} \tag{29.33}$$

In this connection the new percept is that the maximum failure rate λ_0 is divided into systematic and random hardware errors. This is why the equation needs to be changed into:

$$\lambda_H(t) = \frac{\partial E_H(t)}{\partial t} = (\lambda_{0_{SE}} + \lambda_{0_{RE}}) \cdot \lambda \cdot e^{(-\lambda \cdot t)} \tag{29.34}$$

If an optimized algorithm is applied to the new approach, then a forecasting for the hardware system in safety related applications—concerning the hardware error—can be made. This is done so that it can be predicted how many remaining errors $\lambda_{0_{RF}}$ as well as systematic errors $\lambda_{0_{SF}}$ can be found at certain point of time.

The estimations of the default rates λ and λ_0 are necessary. As we do not know of the distribution of the basic population (that is the probability function and the density function) and as we have the result of a sampling procedure, we can now look for the parameters $\tilde{\lambda}$, $\tilde{\lambda}_0$, $\tilde{\lambda}_{0_{SF}}$ and $\tilde{\lambda}_{0_{RF}}$ for which the realization of the precise sample is most probable. This, of course, means nothing else than a maximisation task. In calculating it the density function f (t) is needed from the Eq. (29.33) [9].

$$\begin{aligned} \Delta(\lambda_0, \lambda) &= f(x_1, \ldots, x_n, q) = \prod_{i=1}^{n} f(x_i, q) = L(\lambda_0, \lambda) \\ &= [1 - F(t_e)]^{\lambda_0 - m_e} \prod_{i=1}^{m_e} (\lambda_0 - i + 1) f(t_i) \end{aligned} \tag{29.35}$$

Whereby m_e stands for the amount of the overall errors at this point of time t_e. Here $f(t_i)$ is the failure density function and $F(t_e)$ the default probability function. If the natural logarithm is taken from the Eq. (29.36), then the following applies:

$$\begin{aligned}\ln \Delta\left(\lambda_0, \lambda\right) = \left(\lambda_0 - m_e\right) \ln\left[1 - F\left(t_e\right)\right] \\ + \sum_{i=1}^{m_e} \ln\left(\lambda_0 - i + 1\right) + \sum_{i=1}^{m_e} \ln f\left(t_i\right)\end{aligned} \tag{29.36}$$

Now after that, the Δ-function is to be maximized. This is done with the following approach:

$$\frac{\partial \ln \Delta\left(\lambda_0, \lambda\right)}{\partial \lambda_0} = \ln\left[1 - F\left(t_e\right)\right] + \sum_{i=1}^{m_e} \frac{1}{\lambda_0 - i + 1} = 0 \tag{29.37}$$

$$\frac{\partial \ln \Delta\left(\lambda_0, \lambda\right)}{\partial \lambda_0} = \ln\left[1 - F\left(t_e\right)\right] + \sum_{i=1}^{m_e} \frac{1}{\lambda_0 - i + 1} = 0 \tag{29.38}$$

The partial derivations are replaced by the expected value $E_H(t)$ and the default rate $\lambda_H(t)$. The following applies:

$$\begin{aligned}E_H(t) = \lambda_0 \cdot F(t) \\ \lambda(t) = \lambda_0 \cdot f(t)\end{aligned} \tag{29.39}$$

This leads to:

$$\frac{\partial \ln \Delta\left(\lambda_0, \lambda\right)}{\partial \lambda_0} = \ln\left[1 - \frac{E_H\left(t_e\right)}{\lambda_0}\right] + \sum_{i=1}^{m_e} \frac{1}{\lambda_0 - i + 1} = 0 \tag{29.40}$$

$$\begin{aligned}\frac{\partial \ln \Delta\left(\lambda_0, \lambda\right)}{\partial \lambda_0} = -\frac{\lambda_0 - m_e}{1 - \frac{E_H\left(t_e\right)}{\lambda_0}} \cdot \frac{\partial \frac{E_H\left(t_e\right)}{\lambda_0}}{\partial \lambda} \\ + \sum_{i=1}^{m_e} \frac{1}{\frac{\lambda\left(t_i\right)}{\lambda_0}} \cdot \frac{\partial \frac{\lambda\left(t_i\right)}{\lambda_0}}{\partial \lambda} = 0\end{aligned} \tag{29.41}$$

$$\begin{aligned}\frac{\partial \ln \Delta\left(\lambda_0, b\right)}{\partial \lambda} = -\lambda_0 \left[\frac{\lambda_0 - m_e}{\lambda_0 - {}^{s}\mu\left(t_e\right)}\right] \cdot \frac{1}{\lambda_0} \cdot \frac{\partial E_H\left(t_e\right)}{\partial \lambda} \\ + \lambda_0 \sum_{i=1}^{m_e} \frac{1}{\lambda\left(t_i\right)} \cdot \frac{1}{\lambda_0} \cdot \frac{\partial \lambda\left(t_i\right)}{\partial \lambda} = 0\end{aligned} \tag{29.42}$$

$$\frac{\partial \ln \Delta\left(\lambda_0, \lambda\right)}{\partial b} = -\left[\frac{\lambda_0 - m_e}{\lambda_0 - E_H\left(t_e\right)}\right] \cdot \frac{\partial E_H\left(t_e\right)}{\partial \lambda_0} + \sum_{i=1}^{m_e} \frac{1}{\lambda\left(t_i\right)} \cdot \frac{\partial \lambda\left(t_i\right)}{\partial \lambda} = 0 \tag{29.43}$$

With the Eqs. (29.42) and (29.43) the requested model parameters can be estimated. Therefore the expected value and the default rate may be inserted into the estimated equation and me be written for the summation "$\Sigma = \lambda_{0_{RF}} + \lambda_{0_{SF}}$". Hence, the result is:

$$\begin{aligned} &\frac{\partial \ln \Delta((\Sigma),\lambda)}{\partial \Sigma} = \ln\left[1 - \frac{\Sigma \cdot \left(1 - e^{-(\lambda)t_e}\right)}{\Sigma}\right] + \sum_{i=1}^{m_e} \frac{1}{\lambda_0 - i + 1} = 0 \\ &\frac{\partial \ln L(\Sigma,\lambda)}{\partial \lambda} = -\left[\frac{\Sigma - m_e}{\Sigma - \Sigma \cdot \left(1 - e^{-\lambda t_e}\right)}\right] \cdot \frac{\partial\left(\Sigma \cdot \left(1 - e^{-\lambda t_e}\right)\right)}{\partial \lambda} + \\ &+ \sum_{i=1}^{m_e} \frac{1}{\lambda \cdot \Sigma \cdot e^{-\lambda t_e}} \cdot \frac{\partial\left(\Sigma \cdot e^{-\lambda t_e}\right)}{\partial \lambda} = 0. \end{aligned} \tag{29.44}$$

Consequently with the substitution of the summation the following estimated equations apply:

$$-\tilde{\lambda} t_e + \sum_{i=1}^{m_e} \frac{1}{\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) - i + 1} = 0 \tag{29.45}$$

$$\begin{aligned} &- \left[\frac{\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) - m_e}{\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) - \left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot \left(1 - e^{-(\lambda)t_e}\right)}\right] \\ &\cdot \frac{\partial\left(\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot \left(1 - e^{-(\lambda)t_e}\right)\right)}{\partial \lambda} + \sum_{i=1}^{m_e} \frac{1}{\lambda \cdot \left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot e^{-\lambda t_e}} \\ &\cdot \frac{\partial\left(\lambda \cdot \left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot e^{-\lambda t_e}\right)}{\partial \lambda} = 0. \end{aligned} \tag{29.46}$$

If the Eqs. (29.45) and (29.46) are solved to the total rate $\tilde{\lambda}$, the equation will be demonstrated as the following:

$$\tilde{\lambda} = \frac{m_e}{\sum_{i=1}^{m_e} t_i + t_e \cdot \left(\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}} \right) - m_e \right)} . \tag{29.47}$$

In order to get to the estimated parameter $\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}} \right)$ the result of the Eq. (29.47) needs to be inserted into the estimated Eq. (29.45). Hence, the result is:

$$- \left(\frac{m_e}{\sum_{i=1}^{m_e} t_i + t_e \cdot \left(\left(\sum_{i=0}^{n} \tilde{u}_{0_{c_i}} + \sum_{i=0}^{n} \tilde{u}_{0_{nc_i}} \right) - m_e \right)} \right) \cdot t_e + \sum_{i=1}^{m_e} \frac{1}{\left(\sum_{i=0}^{n} \tilde{u}_{0_{c_i}} + \sum_{i=0}^{n} \tilde{u}_{0_{nc_i}} \right) - i + 1} = 0 \tag{29.48}$$

With the help of the Siemens standard SN 295500 the default rates may be taken from the tables of the standard norm. The Eqs. (29.47) and (29.48) may be solved with the help of the Cram'sche theory. Therefore both estimated parameters are determined with the Δ-function.

29.3 Contribution for Predict the Critical Failure Probability Depending on a Estimator-Frequency Ω

In order to describe a detection and correction process, different requirements have to be applied. Both a detection and a correction process run in a non-homogenous Poisson process. The critical errors are independent from each other. The mean number of critical errors in the time interval $\Delta\tau$ is proportional to the mean number of remaining not corrected critical errors. Every critical error that occurs is therefore corrected, with no new errors added. The delay frequency Ω until the correction process can be assumed to be constant. Therefore, the correction process is considered to be a detection process delayed by Ω. The following therefore applies:

$$_{C}^{s}\mu_{c/nc}(t) = {}_{D}^{s}\mu_{c/nc}\,(t - \Omega) \tag{29.49}$$

where ${}^{s}_{C}\mu_c(t)$ is the cumulative function of a critical error and ${}^{s}_{C}\mu(t)_{nc}$ is the cumulative function of a non-critical error. It can be seen that the correction curve is shifted by a constant factor Ω.

At point in time τ_x a certain number of errors N is detected, but corrected at point in time τ_y. It is very unrealistic that a delay frequency Ω can be assumed to be constant, since every critical error requires a different amount of time to be corrected. For a new Idea, though, a not constant delay frequency $\Omega(t)$ is chosen.

There is no constant delay frequency $\Omega(t)$ between the detection and correction processes. Now it remains to be clarified how the delay time $\Delta\tau(t)$ is described mathematically. The delay time $\Delta\tau(t)$ can be described at will, however, the delay time depends on the critical-systems that is going to be analyzed. Hence, the time shift $\Delta\tau(t)$, for instance, can be assumed that the frequency is linear:

$$\Omega(t) = t \tag{29.50}$$

However, this would mean, however, that proportionality exists between the abscissa, the time axis, and the ordinate, the number of errors. That is why a linear delay time is not recommended for this approach.

Another consideration, which arose during the work on this article, is to consider the delay time $\Delta\tau(t)$ as a exponential function with negative exponent.

$$\Delta\tau(t) = a \cdot e^{-t} \tag{29.51}$$

The goal is to obtain a cumulative function as a result for the correction process that reflects a relatively approximation of reality. Another promising solution seems to be the selection of a logarithmic function, since the "ln-func" can only adopt positive values. This could lead to a so-called "learning process". In order to ultimately examine the effect of the exponential or logarithmic delay time on the correction process, the detection and correction process is do be described mathematically.

Following the first order differential equation applies for the critical detection process:

$$\frac{\partial {}^{s}_{D}\mu_c(t)}{\partial t} = d_{\psi}(t) \cdot \left(\sum_{i=0}^{n} u_{0_c i} - {}^{s}_{D}\mu_c(t) \right) \tag{29.52}$$

Function $d_{\psi}(t)$ describes the detection rate for the critical errors and shall be assumed to be constant in this work. Index "ψ" indicates that the detection rate refers to the new approach and the newly introduced constant Ψ. Ψ represents the ratio between the critical error and the total number of errors:

The rate $d_{\psi}(t)$ has always to be seen as unknown and therefore needs to be estimated using the maximum-likelihood-method. The detection process runs a non-homogeneous Poisson form. Further, all critical errors shall be independent from

each other. $\sum_{i=0}^{n} u_{0_{c_i}}$ represents the total number of critical errors. ${}^{s}_{D}\mu_c(t)$ describes the expectation value of the critical errors in the detection process.

Since according to the "postulate" at the beginning of the prognosis, $t = 0$ there are no critical errors, the side condition follows the differential condition [10]:

$$ {}_{D}{}^{s}\mu_c(0) = 0 \tag{29.53} $$

Thus, the inhomogeneous differential equation of the first order can be solved completely.

$$ \begin{aligned} \frac{\partial\, {}_{D}{}^{s}\mu_c(t)}{\partial t} &= d_\psi \cdot \sum_{i=0}^{n} u_{0_{c_i}} - d_\psi \cdot {}_{D}{}^{s}\mu_c(t) \\ \frac{\partial\, {}_{D}{}^{s}\mu_c(t)}{\partial t} + d_\psi \cdot {}_{D}{}^{s}\mu_c(t) &= d_\psi \cdot \sum_{i=0}^{n} u_{0_{c_i}} \end{aligned} \tag{29.54} $$

The delay time $\Delta\tau(t)$ is irrelevant, because although the critical error has been detected, but not yet corrected. In the most of the cases, the error cannot be corrected immediately because the system is in operation.

Now the correcting process of the critical error is described mathematically as follows:

$$ \frac{\partial\, {}_{C}{}^{s}\mu_c(t)}{\partial t} = c_\psi(t) \cdot ({}_{D}{}^{s}\mu_c(t) - {}_{C}{}^{s}\mu_c(t)) \tag{29.55} $$

${}^{s}_{C}\mu_c(t)$ describes the expectation value of the critical errors. The correction rate can also be interpreted using Eq. (29.55):

$$ c_\psi(t) = \frac{\frac{\partial\, {}_{C}{}^{s}\mu_c(t)}{\partial t}}{({}_{D}{}^{s}\mu_c(t) - {}_{C}{}^{s}\mu_c(t))} = \frac{{}_{C}{}^{s}\mu_c'(t)}{{}_{D}{}^{s}\mu_c(t) - {}_{C}{}^{s}\mu_c(t)} \tag{29.56} $$

Equation 29.56 shows, that the correction rate $c_\psi(t)$ can be seen as error correction per detected, but not corrected "critical" errors. In reality, the correction rate depends on complexity of the problem to be analyzed, the abilities of the test team and the time restrictions for the handover of the finished software to the customer.

Equation 29.9 can now be solved with the following side condition:

$$ {}_{C}{}^{s}\mu_c(0) = 0 \tag{29.57} $$

The solution of the differential equation for the correction process then leads to the following expectation value:

$$ {}^{c^s}\mu_c(t) = e^{-C(t)} \cdot \left(\int_0^t \left[\left(\sum_{i=0}^{n} u_{0_{c_i}} + \sum_{i=0}^{n} u_{0_{nc_i}} \right) \cdot c(s) \cdot e^{C(s)} \cdot \left(1 - e^{-D(t)}\right) \right] ds \right) \tag{29.58} $$

where

$$ \begin{aligned} D(t) &= \int_0^t d(s)\, ds \\ C(t) &= \int_0^t c(s)\, ds \end{aligned} \tag{29.59} $$

In order to maintain the mathematical overview and for the sake of simplicity, also the correction rate $c_\psi(t)$ shall be assumed to be time-independent in the scope of this work. When the differential equation is now solved, following expectation value is obtained for the correction process:

$$ \begin{aligned} {}^{c^s}\mu_c(t) &= \left(\sum_{i=0}^{n} u_{0_{c_i}} \right) \cdot \left(1 - \left(1 + c_\psi \cdot t\right) \cdot e^{-c_\psi \cdot t}\right) \\ {}^{c^s}\mu_c(t) &= \left(\sum_{i=0}^{n} u_{0_{c_i}} \right) \cdot \left(1 - e^{-c_\psi \cdot \left(t - \frac{\ln\left(1 + c_\psi \cdot t\right)}{c_\psi}\right)}\right) \end{aligned} \tag{29.60} $$

Consequently, the delay time $\Delta\tau(t)$ of the correction process can be seen very well from the Equation.

29.4 Conclusion

In this article a new approach has been set up in order to generate a better estimation of the safety parameters. Hereby, the focus was laid on the different default rates. When estimating the probabilities, traditional methods have ignored the differentiation into systematic and random hardware.

It is tremendously important for the safety technology that all error possibilities are taken into account through a stochastic model. Here, too, the work of this contribution shows that it may be possible to insert distribution functions other than the exponential distribution.

With this new approach it is possible, too, to minimize or predict systematic hardware errors. Further, a realistic prediction about the probability of reliability as well as the probability of default can be made.

References

1. Börcsök, J.: Functional Safety Computer Architecture Part 1 and Part 2. Lecture Notes. University of Kassel, Kassel (2001)
2. Börcsök, J.: Functional Safety Systems. Hüthig, Heidelberg (2004)
3. IEC 61508, International Standard: 61508 Functional Safety of Electrical Electronic Programmable Electronic Safety Related Systems Part 1–Part 7. International Electro Technical Commission, Geneva (1999–2000)
4. Robert, C.P., Casella, G.: Monte Carlo, Statistical Methods. Springer, Berlin (1999)
5. Storey, N.: Safety Critical Computer Systems. Addison Wesley, Harlow (1996)
6. Lewis, E.E.: Introduction to Reliability Engineering, 2nd edn. Wiley, New York (1996)
7. Velten-Philipp, W., Houtermans, M.J.M.: The Effect of Diagnostic and Periodic Testing on the Reliability of Safety Systems. TÜV, Köln (2006)
8. Health & Safety Executive (HSE) UK, Programmable Electronic Systems in Safety-Related Applications, Part I. HM Stationery Office, London (1995)
9. Börcsök, J., Holub, P., Schwarz, M.H., Dang Pham, N.T.: Determine PFD-Values for Safety Related Systems Overview. ESREL, Stavanger (2007)
10. Goble, W.M.: Safety of Programmable Electronic Systems—Critical Issues, Diagnostic and Common Cause Strength. Proceedings of the IchemE Symposium. Institution of Chemical Engineers, Rugby (1995)

Ossmane Krini Head of Functional Safety, Bosch GmbH und ZF-Friedrichshafen AG and Postdoctoral at the University of Kassel. Certified Safety-Manager for IEC 61508/ISO26262

Jamal Krini Ph.D. Student (Research and Development) for Functional Safety and Department for Computer Architecture and System Programming at the University of Kassel

Abderrahim Krini Ph.D. Student (Research and Development) for Functional Safety and Department for Computer Architecture and System Programming at the University of Kassel and Bosch GmbH und ZF-Friedrichshafen AG, Germany

Josef Börcsök Head of the Department for Computer Architecture and System Programming at the University of Kassel, Germany and Certified safety-expert of Functional Safety

Index

N. Mastorakis et al. (eds.), *Computational Problems in Science and Engineering*,
Lecture Notes in Electrical Engineering 343, DOI 10.1007/978-3-319-15765-8